UNIVERSIDADE DA CORUÑA

TESIS DOCTORAL

Contribuciones al análisis estadístico del riesgo de crédito

Autor:

ANDRÉS E. DEVIA RIVERA

Directores:

RICARDO CAO ABAD JUAN M. VILAR FERNÁNDEZ

Departamento de Matemáticas

Facultade de Informática

Universidade da Coruña

España

2015

UNIVERSIDADE DA CORUÑA

Los abajo firmantes hacen constar que son los directores de la Tesis Doctoral titulada **Contribuciones al análisis estadístico del riesgo de crédito**, realizada por Andrés E. Devia Rivera en el marco del programa de doctorado en Estadística e Investigación Operativa impartido por el Departamento de Matemáticas de la Universidade da Coruña, dando su consentimiento para que su autor, cuya firma también se incluye, proceda a su presentación y posterior defensa.

Os abaixo asinantes fan constar que son os directores da Tese Doutoral titulada **Contribucións á análise estatística do risco de crédito**, desenvolta por Andrés E. Devia Rivera no marco do programa de doutoramento de Estatística e Investigación de Operacións ofertado polo Departamento de Matemáticas da Universidade da Coruña, dando o seu consentimento para que o seu autor, estando a súa sinatura tamén incluida, proceda a súa presentación e posterior defensa.

24 de septiembre, 2015

Directores:

Ricardo Cao Abad

Juan M. Vilar Fernández

Doctorando:

Andrés E. Devia Rivera

Índice general

Para mi madre Ana,
mis hermanas Lorena, Pamela y Solange,
y para mi mujer Covadonga.

Agradecimientos

En primer lugar deseo agradecer a mis directores, Ricardo Cao y Juan Vilar, por la oportunidad que me dieron de poder realizar mis estudios de doctorado en esta universidad. Durante los años en los que he trabajado en esta línea de investigación sobre modelización estadística del riesgo de crédito, he aprendido mucho de ellos, tanto en lo académico como en lo humano. Ambos profesores dedicaron inumerables horas y recursos en mi formación como investigador durante mi período como becario FPI del Departamento de Matemáticas entre los años 2006 y 2010, y posteriormente, estando yo en Madrid, de forma remota siguieron apoyándome para poder finalizar mi doctorado. Por toda la ayuda que me han brindado y por el tiempo que han invertido en dirigir mi tesis, quisiera expresar mi más profundo agradecimiento a los profesores Ricardo Cao y Juan Vilar del Depratamento de Matemáticas de la Facultade de Informática de la UDC.

Quisera también agradecer muy especialmente a Covadonga, mi mujer, quien ha sido mi mejor compañera y mi mayor soporte en este proyecto personal, y en cierto modo de ambos, de retomar la finalización de mi tesis. Sólo puedo expresar aquí lo importante que ella es para mí, y decirle que sin su apoyo no habría podido llegar hasta aquí.

También quisiera expresar mi agradecimiento a mi familia en Chile, a mi madre, Ana, a mis hermanas, Lorena, Pamela y Solange, a mi sobrinita Valentina, a mi padre, Patricio, a mis primos Rodrigo y Aarón, a mis cuñados Claudio y Francisco, a mis tíos, Amalia y Jaime, y mi primita María José, y también, como homenaje especial, a mis abuelas María y Eliana, quienes nos dejaron recientemente y a quienes extraño muchísimo. A todas y a todos en Chile, a mis amigos y a mis compañeros de la USACH, os agradezco mucho por el apoyo que me habéis dado desde que vine a España.

Agradezco también a los profesores del Departamento de Matemáticas de la Universidade da Coruña por acogerme y proporcionarme toda clase de facilidades durante mi estancia como becario FPI. En particular, agradezco muy especialmente a los profesores Salvador Naya, Germán Aneiros, Mario Francisco y Julián Costa, y a los demás miembros del Grupo de investigación *MODES*, por su amabilidad y por haberme hecho sentir uno más del grupo.

También quisiera agradecer muy especialmente a Miguel Clemente y a Javier Tarrío por su amistad, su apoyo y su compañía durante todos estos años, que tan necesaria es cuando uno se encuentra lejos de casa.

Agradezco a Gonzalo Barros, a Fernando González y a Cristina Fragueiro, equipo de analistas en la entidad financiera colaboradora en los temas de investigación que dieron lugar a los Capítulos 2 y 3 de esta memoria. Los tres se encargaron de acogerme en la entidad financiera durante mi primera estancia de investigación como becario FPI en el marco de un convenio de colaboración entre la Universidade da Coruña y la citada entidad. Tanto Gonzalo como Fernando y Cristina, procuraron siempre proporcionarme todo el material y la ayuda necesaria para realizar mi trabajo de la forma más cómoda y eficaz.

También agradezco al profesor José Luis Fernández, socio de Analistas Financieros Internacionales (AFI), por acogerme en las oficinas de AFI en Madrid durante una estancia breve de investigación realizada en el año 2009 y cuyo fruto es el estudio realizado en Capítulo 5 de esta memoria.

Quisiera expresar también mi más sincero agradecimiento a mi supervisor en mi trabajo, Jaime Domínguez de Avon Cosmetics España, por su constante ayuda para organizar mis tiempos, prácticamente sin pedirme explicaciones ni ponerme dificultades. Sin la ayuda de Jaime todo habría sido mucho más difícil, o incluso imposible, de sacar adelante.

Finalmente, deseo expresar mi agradecimiento al antiguo Ministerio de Ciencia e Innovación de España (MICINN), actual Ministerio de Economía y Competitividad (MINECO), por permitirme disfrutar de una beca FPI (referencia BES-2006-12240) en el marco de apoyo a la formación de personal investigador del MICINN, asociada a los proyectos de investigación MTM2005-00429 y MTM2008-00166. La realización de esta tesis doctoral fue financiada gracias a los fondos destinados a estos dos proyectos de investigación.

Resumen

En esta memoria se aborda el problema de la modelización estadística del riesgo de crédito. Aquí se estudian tres modelos para la predicción de la morosidad en créditos personales. El primero corresponde a un modelo de puntuación crediticia construido con técnicas de regresión logística. Como resultado, se obtiene un modelo de estimación de la propensión a la morosidad del cliente, dado un conjunto de variables explicativas sobre las que se realiza la regresión logística. El segundo modelo corresponde a un modelo de probabilidad de mora (PD) construido con técnicas de análisis de supervivencia. Este modelo de PD se obtiene a partir de la función de distribución condicional del tiempo hasta la mora del crédito y se calcula utilizando tres clases de estimadores (paramétrico, semiparamétrico y no paramétrico). El estimador no paramético de la PD, φ_n^{PLG}, se obtiene a partir del estimador de Beran (1981) para la función de supervivencia condicional con datos censurados. Resultados asintóticos como la consistencia fuerte uniforme y la normalidad asintótica del estimador φ_n^{PLG}, se exponen en el Capítulo 4. Finalmente, un tercer modelo de predicción de la morosidad en créditos personales se propone en el Capítulo 5. Allí se busca dar respuesta a un problema con escaso tratamiento en la literatura sobre riesgo de crédito, denominado reincidencia de la morosidad crediticia. El estudio se centra en obtener un modelo de probabilidad análogo al modelo propuesto en el Capítulo 3, donde se utiliza la perspectiva del análisis de supervivencia con sucesos recurrentes bajo censura y dependencia. Como resultado, se obtienen fórmulas para la probabilidad condicional de reincidencia de los impagos de un mismo sujeto conociendo su puntuación crediticia y los tiempos en los que se han producido los impagos anteriores. Este modelo puede ser utilizado por las entidades como herramienta de seguimiento y evaluación del perfil crediticio de los clientes con más propensión a comenter algún tipo de incumplimiento.

Resumo

Nesta memoria abórdase ó problema da modelización estatística do risco de crédito. Aquí estúdanse tres modelos para a predicción da morosidade en créditos persoais. O primeiro corresponde cun modelo de puntuación crediticia construido con técnicas de regresión loxística. Como resultado, obtense un modelo de estimación da propensión á morosidade do cliente, dado un conxunto de variables explicativas sobre as que se realiza a regresión loxística. O segundo modelo corresponde cun modelo de probabilidade de incumprimento (PD), construido con técnicas de análise de supervivencia. Este modelo de PD obtense a partir da función de distribución condicional do tempo ata a mora do crédito e calcúlase utilizando tres clases de estimadores (paramétrico, semi paramétrico e non paramétrico). O estimador non paramétrico da PD, φ_n^{PLG}, obténse a partir do estimador de Beran (1981) para a función de supervivencia condicional con datos censurados. Resultados asintóticos como a consistencia forte uniforme e a normalidade asintótica do estimador φ_n^{PLG} expónse no Capítulo 4 desta memoria. Finalmente, un terceiro modelo de predicción da morosidade en créditos persoais propónse no Capítulo 5. Alí búscase dar resposta a un problema con escaso tratamento na literatura actual sobre o risco de crédito, chamado reincidencia da morosidade crediticia. O estudo deste fenómeno céntrase en obter un modelo de probabilidade análogo ó modelo proposto no Capítulo 3, onde se utiliza a perspectiva do análise de supervivencia con sucesos recorrentes baixo censura e dependencia. Como resultado, obtéñense fórmulas para a probabilidade condicional de reincidencia dos impagos dun mesmo suxeito coñecendo a súa puntuación crediticia e os tempos nos que se produciron os impagos anteriores. Este modelo pode ser utilizado polas entidades como ferramenta de seguimento e avaliación do perfil de crédito dos clientes máis propensos a cometer algunha clase de incumprimento.

Abstract

In this thesis, the problem of statistical modeling of credit risk is considered. Three models for predicting defaults on personal loans are studied. Firstly, a credit scoring model is obtained via logistic regression techniques. As a result, a model to estimate the propensity to credit default is obtained given a set of explanatory variables on which the logistic regression is performed. Secondly, a model for the probability of default (PD) is built using survival analysis techniques. This model for PD is obtained from the conditional distribution function of time to credit arrears and is calculated using three different estimators (parametric, semiparametric and nonparametric). The nonparametric estimator, φ_n^{PLG}, is derived from the estimator by Beran (1981) for the conditional survival function with censored data. Asymptotic results like strong uniform consistency and asymptotic normality of φ_n^{PLG}, are obtained in Chapter 4 of this thesis. Finally, a third model to predict default in personal loans is proposed in Chapter 5. There, we try to respond to a problem with poor treatment in the current literature on credit risk, called credit default recidivism. The study of this phenomenon focuses on obtaining a probability model similar to that proposed in Chapter 3, where the approach of survival analysis with recurrent events under censoring and dependence model is used. As a result, some formulae are obtained for conditional probability of recurrent defaults for a single debtor given a known credit scoring value and the times where previous defaults have been observed. This model can be used by financial institutions as a tool for monitoring and evaluating the credit profile of their customers.

Prólogo

En esta memoria se aborda el problema de la *modelización estadística del riesgo de crédito*. Su estudio ha sido fruto de la colaboración entre el Grupo de *Modelización, Optimización e Inferencia Estadística* (*MODES*) de la *Universidade da Coruña* y una entidad financiera española que aportó los datos reales con los que se trabajó. Dicha entidad, tal como ha ocurrido con otras entidades financieras en todo el mundo, comenzó en 2006 un proceso de adaptación regulatoria que culminaría con la implantación de un nuevo sistema de administración y medición del riesgo de crédito llamado *Nuevo Acuerdo de Capital de Basilea* [1].

El Nuevo Acuerdo de Capital de Basilea, al que también se suele hacer referencia como *Basilea II*, es el segundo de los Acuerdos de Basilea destinados a establecer un conjunto de mecanismos estandarizados para la medición integral del riesgo que afecta a la asignación de capital de las entidades de crédito. Tales mecanismos consisten fundamentalmente en recomendaciones sobre leyes y regulaciones financieras elaboradas por el *Comité de Supervisión financiera de Basilea*[2](CSBB).

El Nuevo Acuerdo de Basilea se compone de tres partes fundamentales llamadas pilares y sus principales objetivos son:

[1] Basel Committee on Banking Supervision (2001a). The new Basel Capital Accord. Bank for International Settlements.

[2] Organismo formado por gobernadores de los bancos centrales de Bélgica, Canadá, Francia, Alemania, Italia, España, Holanda, Suecia, Suiza, Reino Unido, Luxemburgo, Japón, Estados Unidos, y por representantes del Banco Central Europeo. Este comité celebra reuniones periódicas en el Banco de Pagos Internacionales (o BIS, por su sigla en inglés), en Basilea, Suiza.

1. Garantizar que la asignación de capital sea más sensible al riesgo

2. Separar el riesgo operacional del riesgo de crédito, y la cuantificación de ambos

3. Intentar alinear el capital económico y regulatorio más estrechamente para reducir las posibilidades de arbitraje regulatorio.

Si bien, este material no pretende ser en sí mismo una memoria sobre los contenidos de Basilea II, las técnicas que aquí se tratan tienen por objetivo el de contribuir al estudio de nuevas metodologías cuantitativas que pueden aplicarse en la *medición del riesgo de crédito*. Por ello, como se verá más adelante, las técnicas utilizadas y los resultados empíricos obtenidos durante el desarrollo de esta memoria se enmarcan precisamente dentro del ámbito de discusión del Pilar I de Basilea II.

Según se explica en el Pilar I de Basilea II, los bancos están obligados a cumplir los requerimientos mínimos de capital que se determinan a partir de la relación entre el capital en exposición y los activos ponderados por riesgo. En el Pilar I se describen los enfoques que se pueden adoptar para el cálculo de las reservas mínimas de capital por riesgo de insolvencia. Además, se describen varias metodologías cuantitativas para la construcción y validación de los modelos predictivos que pueden utilizar las entidades para obtener estimaciones de sus parámetros fundamentales de riesgo.

De la extensa literatura disponible actualmente sobre medición del riesgo de crédito, se desprende que el primero y más importante de los parámetros del riesgo de crédito es la *probabilidad de mora* (PD). Por este motivo, el objetivo principal de la colaboración entre el grupo de investigación *MODES* y la entidad de crédito citada fue la investigación de técnicas estadísticas para el estudio y modelización de la PD en créditos personales. Como resultado de la investigación realizada, en esta memoria se proponen cinco modelos de estimación de la PD para los que se han utilizado cuatro metodologías estadísticas distintas. La teoría desarrollada y los resultados obtenidos se exponen a partir del Capítulo 2.

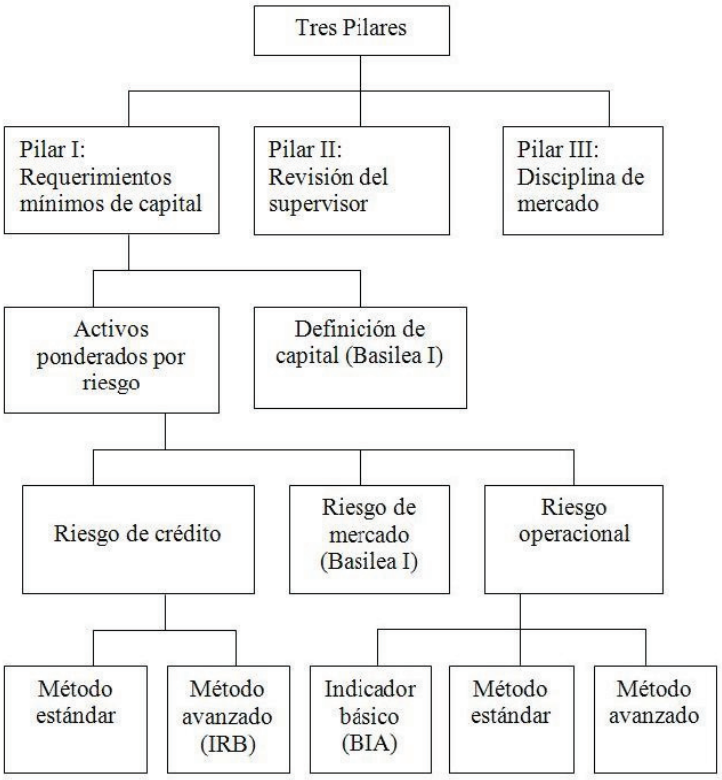

Figura 0.1. Estructura básica de Basilea II.

La Figura 0.1 muestra un esquema simplificado de la estructura del Nuevo Acuerdo de Basilea. Allí se explica el grado de dependencia que existe entre el cálculo de *requerimientos mínimos de capital* y las tecnologías previstas en Basilea II para medir cada una de las fuentes del riesgo bancario.

En el Pilar I de Basilea II se establece que existen dos mecanismos para el cálculo de los parámetros del riesgo de crédito. El primero de ellos se denomina *método estándar* y consiste básicamente en que las entidades de crédito podrán incorporar en su metodología de cálculo del capital regulatorio, estimaciones de las ponderaciones por riesgo obtenidas con modelos cuantitativos desarrollados por agencias de calificación externa. Algunas de las agencias más conocidas a nivel mundial son, por ejemplo, Moody's KMV, J. P. Morgan, Standards & Poor's y Fitch Ratings, entre otras. El segundo mecanismo se denomina *método basado en calificaciones internas*, o simple-

mente *método IRB* (del acrónimo en inglés *internal rating based method*).

Las técnicas estadísticas que se estudian en esta memoria tienen por objetivo el desarrollo de metodologías que permitan elaborar modelos de riesgo de crédito capaces de ser integrados en los sistemas de medición y estandarización del cálculo de requerimientos de capital. Sobre este punto se profundizará más adelante en el Capítulo 1, donde se exponen algunas definiciones básicas para entender el problema de la medición del riesgo de crédito como son, la morosidad, los parámetros fundamentales del riesgo de crédito y en qué consisten los modelos basados en calificaciones internas (IRB), entre otras. Finalmente, se complementa la discusión expuesta con la revisión de algunos de los modelos de riesgo de crédito más conocidos en la literatura.

El problema de la medición del riesgo de crédito ha ido incrementando su importancia en la última década, especialmente en lo relativo a la complejidad de las técnicas que se han ido desarrollando para hacer más eficiente el control del riesgo por parte de las entidades financieras. Este problema, que en principio podría ser considerado como de exclusiva competencia de los profesionales del mundo de las finanzas, se ha convertido en un problema de interés mucho más general, en el que hoy también participan especialistas provenientes de diversas áreas del conocimiento de corte cuantitativo como son, por ejemplo, los matemáticos, los estadísticos, los ingenieros y los físicos, entre otros profesionales.

En la actualidad existe una extensa literatura dedicada a la medición del riesgo de crédito. Esto se debe fundamentalmente a dos razones, la primera tiene que ver con la finalidad operacional del modelo, es decir, qué características se quieren medir exactamemente y qué ingredientes necesita el modelo para funcionar. Un ejemplo de esto son los modelos de puntuación crediticia (también conocidos como *modelos de credit scoring*). Estos modelos son herramientas estadísticas que las entidades de crédito utilizan para analizar la calidad crediticia de sus clientes dependiendo del tipo de exposición al que corresponden. En este caso, el resultado numérico del modelo, la puntuación o scoring, no sólo simplifica y resume la enorme cantidad de información que posee el banco acerca de sus clientes, sino que además, este resultado es empleado en múltiples decisiones relacionadas con la asignación de crédito a los mismos. Así, un modelo de puntuación crediticia es un importante instrumento de toma de decisiones y permite al banco contar con un indicador de solvencia fiable y fácil de interpretar. Debido a la importancia de este tipo

de modelos para las entidades financieras, en el Capítulo 2 se presenta la construcción de un modelo de puntuación crediticia basado en modelos de regresión logística, una metodología que viene siendo cada vez más utilizada en el estudio de perfiles crediticios como herramienta de análisis del riesgo bancario. El capítulo finaliza con los resultados de algunas aplicaciones del modelo propuesto.

Por otra parte, también existen modelos diseñados para predecir siniestros y conductas de riesgo asociadas a los clientes. Tal es el caso de los créditos a particulares, donde el interés de la entidad está en poder predecir la pérdida de solvencia del acreditado debido, por ejemplo, a un endeudamiento excesivo, a un posible divorcio, a una enfermedad catastrófica, a un accidente inesperado, e incluso al fallecimiento del acreditado o del avalista de la deuda, etc. En el caso de los créditos a empresas, es bien sabido que la pérdida de solvencia se puede deber a múltiples causas, como pueden ser la pérdida de valor de sus activos, la disminución de su cuota de mercado, la pérdida de competitividad, el impago a otros acreedores, las huelgas y pleitos con los trabajadores, y así, un sin número de situaciones que aumentan la incertidumbre sobre la capacidad de pago de la empresa. Analizar y controlar el impacto que pueden ocasionar en la solvencia del banco situaciones como las mencionadas, constituye el principal objetivo de estudio de los distintos enfoques de medición del riesgo de crédito.

En relación con lo anterior, en Basilea II se han establecido criterios específicos para caracterizar situaciones como las descritas en una misma definición estandarizada y medible. Una de ellas es *la mora* o *incumplimiento* del crédito. Así, la probabilidad de que ocurra dicho suceso, es decir, *la probabilidad de mora*, se convierte en el parámetro fundamental del riesgo de crédito y su medición es el principal objetivo que se persigue con los modelos basados en métodos IRB. La definición formal de esta probabilidad es parte de los contenidos que se presentan en el Capítulo 1.

La mayor parte de esta memoria tiene que ver con el estudio de la morosidad en créditos personales, y su modelización por la vía de técnicas estadísticas constituye el eje central de los contenidos que aquí se tratan. En tal sentido, en el Capítulo 3 se presenta una metodología para la modelización de la probabilidad de mora basada en el análisis de supervivencia. Dicha metodología se compone de un modelo de probabilidad que se obtiene a partir de la función de distribución condicional del tiempo hasta el incumplimiento

y de las técnicas de estimación que se proponen para su estudio bajo ciertos supuestos e hipótesis en los que se entrará de lleno más adelante.

Uno de los supuestos fundamentales que sustentan este modelo de probabilidad, es que el tiempo hasta que se produce la mora del crédito está relacionado con el perfil crediticio del cliente. Esto implica que la probabilidad de mora se obtiene condicionando sobre los valores que toman, por ejemplo, las variables estudiadas en el Capítulo 2. Sin embargo, debido a los inconvenientes que padecen los métodos de regresión no paramétrica cuando la dimensión de la covariable es mayor que uno, es preferible reducir su dimensión empleando como variable regresora la puntuación crediticia del cliente. Este problema se conoce como la *maldición de la dimensionalidad* (ver Stone (1980, 1982)). El Capítulo 3 finaliza con la aplicación del modelo a un conjunto de datos reales.

El Capítulo 4 está dedicado al estudio de las propiedades asintóticas del estimador no paramétrico de la probabilidad de mora condicional propuesto en el Capítulo 3. Este estimador se obtiene a partir del estimador producto límite generalizado (PLG) de la distribución condicional con datos censurados propuesto por Beran (1981). Como resultado del análisis expuesto, se demuestra la propiedad de *consistencia fuerte uniforme*, se obtienen expresiones analíticas para el *sesgo* y la *varianza* asintóticos y se obtiene la *normalidad asintótica*. Como producto de los resultados anteriores, se obtiene el valor asintóticamente óptimo del parámetro de suavizado del estimador núcleo de la probabilidad de mora condicional.

Finalmente, el Capítulo 5 está dedicado al estudio de un problema que prácticamente no ha sido tratado en la literatura especializada en riesgo de crédito. Allí se aborda el problema de la reincidencia de la morosidad en créditos personales. Para abordar este problema se adapta el enfoque de análisis de supervivencia visto en el Capítulo 3 para medir el riesgo de morosidad debido al impago recurrente de los créditos. Este fenómeno ha sido denominado como *reincidencia de la morosidad* y su estudio se basa en la modelización de la distribución de probabilidad condicional de los tiempos hasta el j-ésimo impago dados los $j-1$ impagos anteriores y dado el valor que toma la covariable (unidimensional) del modelo. El objetivo que se pretende con el modelo de reincidencia de los impagos es estudiar la probabilidad de que un cliente incumpla sucesivamente sus obligaciones con la entidad financiera sabiendo que ha incumplido al menos una vez en el pasado.

Capítulo 1

Introducción al riesgo de crédito

En esta memoria se aborda el problema de la medición del riesgo de crédito y el material de que se compone es el resultado de la investigación llevada a cabo en este terreno teniendo como punto de partida la colaboración entre el Grupo de *Modelización, Optimización e Inferencia Estadística* (*MODES*) de la *Universidade da Coruña* y una entidad financiera española que contribuyó con datos y experiencia empíricos.

El propósito de esta memoria es contribuir al estudio de las técnicas estadísticas que pueden implementarse como herramientas de medición del riesgo de crédito de acuerdo con las recomendaciones recogidas en el *Pilar I de Basilea II*. Para ello, se ha seguido un enfoque metodológico de la estadística aplicada a las finanzas denominado *modelización del riesgo de crédito vía análisis de supervivencia*.

En los capítulos siguientes, se da a conocer el eje central de la investigación llevada a cabo en esta memoria. Se aborda el problema de la construcción de modelos de riesgo de crédito en todas sus fases, desde el tratamiento previo de las variables involucradas hasta el ajuste y validación estadística de los mismos. Finalmente, los modelos propuestos son aplicados a bases de datos reales, cuyos resultados son contrastados con la evidencia empírica observada por la entidad financiera colaboradora.

La abundante literatura existente sobre modelos de riesgo de crédito, muestra que la mayoría de las técnicas ya implementadas corresponden a *modelos paramétricos*, es decir, modelos en los que las distribuciones de probabilidad son conocidas y donde se cumplen determinados supuestos e hipótesis que justifican su validez. Sin duda, el ejemplo más importante es el actual modelo regulatorio de Basilea II, construido a partir del modelo fatorial de Vasicek (1987, 1997), en el que supone la *normalidad* de los factores.

Dentro del contexto del análisis de supervivencia, es posible encontrar trabajos en los que se estudia el riesgo de crédito por medio de *modelos semiparamétricos*. En este caso, uno de los supuestos fundamentales es el de riesgos proporcionales, lo que conduce a utilizar, por ejemplo, el modelo de regresión de Cox (1972) para estimar la distribución condicional de la variable que controla el tiempo hasta el incumplimiento del crédito. En general, ya sea mediante modelos paramétricos o semiparamétricos, es necesario utilizar supuestos acerca de las distribuciones de probabilidad subyascentes en los procesos que dan lugar al riesgo de crédito.

En esta memoria, además de estudiar algunos modelos como los descritos, también se estudian otras técnicas de modelización basadas en *métodos de estimación no paramétrica*. Esto último es parte de los contenidos incluidos a partir del Capítulos 3 en adelante.

Como resultado de lo anterior, en el presente trabajo se propone un modelo de medición del riesgo de crédito aplicable al caso de exposiones de carácter minorista (créditos personales). Se volverá sobre este tema más en adelante en el Capítulo 3. El grado de ajuste obtenido con los modelos estudiados y los métodos de estimación propuestos fue evaluado a partir de su aplicación a dos conjuntos de datos reales de créditos personales pertenecientes a una entidad financiera española. Los resultados obtenidos son analizados en las secciones finales de los Capítulos 2, 3, 5 y 6.

1.1. Riesgo bancario y riesgo de crédito

La principal actividad comercial de la industria bancaria, aquella a la que dedica la mayor parte de sus esfuerzos, la que genera sus mayores beneficios y la expone a los mayores riesgos, es la comercialización crediticia. Esta actividad está expuesta a una serie de riesgos muy diversos y la entidades de crédito

deben enfrentarse a ellos continuamente. Esto las convierte básicamente en instituciones administradoras del riesgo en busca de una rentabilidad que las compense adecuadamente.

La literatura acerca del riesgo bancario es extensa y su revisión exhaustiva no es parte de los objetivos de esta memoria. Sin embargo, informalmente, se puede decir que el riesgo bancario consiste en la unión de todas aquellas situaciones que afectan negativamente tanto a los beneficios como al capital del banco. Cada una de esas situaciones por separado, y las interacciones que puedan producirse entre ellas, pueden vulnerar la estabilidad financiera de la entidad, que, llegado a un caso extremo, puede ocasionar el deterioro del valor de la entidad poniendo en peligro su solvencia.

Si bien, no todas las entidades de crédito enfrentan exactamente los mismos tipos de riesgo, debido a la diversidad de sus líneas de negocio, y por tratarse de entidades financieras, la mayoría de ellos están afectos a determinadas fuentes de riesgo comunes. Los tipos de riesgo financiero más comunes en la literatura son el *riesgo de crédito*, de *mercado*, *operacional*, de *tipos de interés*, de *liquidez*, de *tipo de cambio* y *soberano*, entre otros (ver, por ejemplo, Pyle (1997), Bessis (2002) y Saunders y Cornett (2008)).

En particular, el riesgo de crédito es el tipo de riesgo bancario que más puede afectar a la solvencia del banco. Prueba de ello es la extensa cronología[1] de acontecimientos que han afectado al sistema financiero mundial, la mayoría de ellos provocados, directa o indirectamente, por múltiples crisis crediticias internacionales. Dicha experiencia fué la principal razón por la que el Comité de Basilea decidió poner en marcha un nuevo sistema de regulación financiera mundial, proceso que comenzó en los albores de los años 90 con el primer Acuerdo de Capital de Basilea (Basilea I) y que culminó en 2004 con el marco revisado de Basilea II (Basel Committee on Banking Supervision (2004)).

1.1.1. Morosidad crediticia y riesgo por insolvencia

El riesgo de crédito surge como consecuencia de operaciones crediticias fallidas, es decir, créditos que al caer en el incumplimiento de pagos pueden

[1] Algunos trabajos publicados sobre esta materia y que contienen información a partir de las últimas dos décadas del siglo XX hasta nuestros días son, por ejemplo, los debidos a Girón (1998), Bank for International Settlements (2009) y Cecchetti et al. (2009), entre otros.

deteriorar o comprometer el patrimonio de la entidad financiera (el capital). En ese sentido, la *morosidad crediticia* surge cuando los créditos pasan de estar en situación de *riesgo de incumplimiento* a convertirse en créditos morosos, lo que en la terminología financiera internacional se conoce como *entrar o caer en default*. En adelante, se hablará indistintamente de *mora*, *impago* o *incumplimiento* para hacer referencia al estado en el que se encuentra un crédito cuando el titular de la deuda es incapaz de cumplir sus compromisos de pago con la entidad.

Debido a que la mayor parte del negocio de una entidad financiera tradicional es la comercialización de productos crediticios, cuando la proporción de créditos morosos, o tasa de mora (TM), alcanza un valor crítico, las pérdidas debidas a la morosidad pueden comprometer el capital de la entidad generando una situación de *riesgo por insolvencia*. En ese sentido, el Acuerdo de Basilea II establece la obligación que tienen las entidades financiera de contar con herramientas que les permitan predecir y mitigar las fuentes del *riesgo por insolvencia*, y en particular las que se deben al riesgo de crédito.

Desde el punto de vista operacional, dichas herramientas, o modelos, deben cumplir ciertas características esenciales. Estas características forman parte de las directrices establecidas por el Comité de Basilea en el Pilar I del nuevo acuerdo. Autores como Duffie y Singleton (2003), Schönbucher (2003) y Turnbull (2003) exponen que algunas de las características deseables que debe cumplir un modelo integral para el riesgo de crédito son:

- Que el modelo cumpla con el principio de parsimonia, es decir, que el número de parámetros necesarios para definir la estructura de morosidad de la cartera de créditos no sea demasiado grande.
- Que el modelo posea un poder predictivo comparable al que posee el modelo regulatorio de Basilea II, utilizado actualmente por bancos y entidades de crédito tradicionales.
- Que el modelo utilice la mayor cantidad posible de información contenida en las bases de datos de la entidad y que los datos sean accesibles.
- Que el coste computacional derivado de la complejidad matemática del modelo y del mecanismo de estimación empleado sea mínimo.
- Que el algoritmo de cálculo del modelo pueda ser implementado en el soporte informático de la entidad.

- Por último, que sea posible la documentación del modelo por parte de la entidad y la replicación de éste por parte del regulador, que en el caso de España corresponde al Banco de España (BDE).

En relación con algunos de los objetivos específicos de esta memoria que han sido comentados anteriormente, se han estudiado técnicas de modelización estadística para medir el riesgo de crédito a particulares, de modo que es conveniente definir apropiadamente lo que se entenderá por dicho concepto.

En adelante, se hablará de *créditos a particulares* cuando se haga referencia a exposiciones minoristas (ver §156, Basel Committee on Banking Supervision (2001a)), es decir, créditos solicitados por personas físicas de forma individual. En este punto es conveniente aclarar que el estudio con datos reales que será presentado en capítulos posteriores no discrimina entre *asalariados* y *autónomos*, y por tanto, el modelo estadístisco que se estudia más adelante permite tratar el problema de la morosidad en *créditos para autónomos* como un caso especial de los *créditos a particulares*.

En segundo lugar, supondremos que estamos en el escenario crediticio más simple, es decir, aquel en que cada acreditado posee una probabilidad de cometer incumplimiento independiente del resto de la cartera. Este fuerte supuesto será justificado más adelante en el Capítulo 3, y permitirá trabajar sobre la base de un modelo en el que no existe correlación entre los tiempos hasta el incumplimiento.

Además, en adelante supondremos que las operaciones de crédito son concedidas independiente de la finalidad del mismo e independiente del tipo de garantía que las entidades exijan al deudor. Así, para los objetivos de esta memoria, se considerán créditos a particulares los siguientes tipos de exposiciones: *préstamos personales*, *créditos de consumo*, *tarjetas de crédito* y *créditos hipotecarios*, entre otros. Aunque estos últimos no serán estudiados en esta memoria, los resultados obtenidos con los modelos propuestos permiten suponer que éstos pueden ser adaptados a este tipo de operaciones crediticias siguiendo, por ejemplo, las ideas de Beran y Djaïdja (2007). Por último, en adelante se entenderá por *acreditado* aquel particular al que se le ha otorgado un crédito personal, y que por tanto es considerado prestatario o deudor del banco. Asimismo, se entenderá por *solicitante* aquel particular que, habiendo solicitado un crédito, aún no ha formalizado un compromiso de pago con la entidad.

1.1.2. Enfoque basado en calificaciones internas (IRB)

El enfoque de medición del riesgo de crédito basado en calificaciones internas, más conocido como enfoque de modelos IRB (iniciales en inglés de *Internal Rating-Based approach*), reemplaza los métodos vigentes hasta antes de 1999 como en el primer Acuerdo de Basilea (Basilea I). Hasta entonces, el capital regulatorio del los bancos se calculaba empleando las ponderaciones por riesgo proporcionadas por el ente regulador, es decir, por los bancos centrales correspondientes.

Tan pronto se dió a conocer el Nuevo Acuerdo de Basilea, se ponen en marcha una serie de cambios referentes a los mecanismos de calificación de los acreditados entre los que destaca la posibilidad de que los bancos utilicen *medidas internas de calificación*, es decir, modelos propios que sirvan para evaluar la calidad de sus acreditados y que por medio de ellos se puedan obtener estimaciones de lo que se conoce como *componentes* o *parámetros del riesgo de crédito*. Actualmente existe una extensa literatura dedicada al estudio y al desarrollo de modelos de riesgo de crédito basados en el enfoque de métodos IRB. Ver por ejemplo los trabajos de Crouhy et al. (2000), Gordy (2000), Basel Committee on Banking Supervision (1999, 2001b), Saunders y Allen (2002), Bluhm et al. (2003), Basel Committee on Banking Supervision (2005b), Chan-Lau (2006) y Engelmann y Rauhmeier (2006), entre otros autores.

Según se establece en Basilea II, es posible medir el riesgo de crédito a partir del estudio y de la estimación de un conjunto de parámetros fundamentales: la *probabilidad de mora*, la *exposición a la mora*, la *pérdida dada la mora* y la *madurez o vencimiento del crédito*.

En lo concerniente específicamente al riesgo de crédito en exposiciones de carácter minorista, es decir, en créditos a particulares, el Comité de Basilea introduce dos modificaciones fundamentales sobre la medición y el control del riesgo de crédito con respecto a otros tipos de exposiciones. La primera modificación tiene que ver con los parámetros fundamentales del riesgo de crédito ya mencionados. Según ésta, en el caso de las exposiciones minoristas sólo interesan tres tipos de parámetros de riesgo: la *probabilidad de mora*, PD, la *pérdida dada la mora*, LGD, y la *pérdida esperada*, EL.

La segunda modificación tiene que ver con las variantes del método IRB. En el caso de exposiciones minoristas, sólo se aplica el método IRB avanzado.

1.1.3. Los parámetros del riesgo de crédito

Los factores o parámetros de riesgo que pueden calcular internamente las entidades a partir del enfoque de modelos IRB son los que se describen a continuación:

La probabilidad de mora

La probabilidad de mora o de incumplimiento, denotada como PD (sigla en inglés de *probability of default*), corresponde a la probabilidad de que un acreditado se declare incapaz de hacer frente a sus compromisos con la entidad al cabo de 1 año de formalizada la deuda. Como consecuencia, se considera que el cliente es moroso cuando se produce un impago por un período de entre 90 y 180 días.

La exposición en caso de mora

La exposición en caso de mora, denotada como EAD (sigla en inglés de *exposure at default*), es la parte de la deuda que queda expuesta al riesgo de pérdida cuando se produce la mora.

La pérdida dada la mora o severidad

La pérdida dada la mora o severidad, denotada como LGD (sigla en inglés de *loss given default*), es la proporción de la deuda que la entidad espera perder si el acreditado incumple sus obligaciones y se expresa como porcentaje o proporción sobre la EAD.

La Madurez

La madurez del crédito, denotada como M (de su nombre en inglés *maturity*) corresponde al período de vencimiento de la operación. Este parámetro es muy importante para validar los datos utilizados por las entidades evitando que se produzcan fechas incoherentes (fechas de incumplimiento anteriores a la concesión del crédito o posteriores al vencimiento del mismo).

Tabla 1.1. Estructura básica del enfoque basado en calificaciones internas

	IRB Fundamental		IRB Avanzado	
Tipo de exposición	Estimaciones internas	Estimaciones hechas por el supervisor	Estimaciones internas	Estimaciones hechas por el supervisor
Empresas, Soberana, y Bancos	PD	LGD, EAD M	PD, LGD EAD, M	
Crédito a particulares	PD, LGD EL	M	PD, LGD EL	M

En la Tabla 1.1 se puede observar un esquema básico con la estructura del enfoque de medición del riesgo de crédito basado en métodos IRB y su relación con las fuentes de estimación de los parámetros del riesgo de crédito según el tipo de exposición.

1.2. Medición del riesgo de crédito vía análisis de supervivencia

Como ya se ha mencionado, en esta memoria se estudia el problema de la medición del riesgo de crédito por medio de técnicas de modelización estadística bajo el enfoque del análisis de supervivencia.

La aplicación del análisis de supervivencia a la medición del riesgo de crédito tiene su orígen en el artículo de Narain (1992). En él se utiliza un modelo de regresión de Cox para estimar la función de distribución del tiempo hasta que se produce un crédito moroso, siendo éste el primer trabajo donde se emplea este tipo de metodología en el contexto del riesgo bancario. Desde entonces numerosos autores han contribuido a desarrollar este enfoque de medición del riesgo de crédito, avanzando en la teoría de la estimación y la modelización de la PD por medio del estudio del tiempo de supervivencia de los créditos. Ver por ejemplo los trabajos de Carling et al. (1998), Stepanova y Thomas (2002), Hanson y Schuermann (2004), Roszbach (2004), Glennon y Nigro (2005), Allen y Rose (2006) y Malik y Thomas (2006), donde se analiza

el caso de préstamos personales. Beran y Djaïdja (2007) consideran el caso de créditos hipotecarios y proponen un modelo para la probabilidad de mora condicional que toma en consideración el bajo porcentaje de morosidad que normalmente afecta a ese tipo de carteras. Este modelo intenta corregir el problema de alta censura a través del concepto de *créditos inmunes,* es decir, acreditados que, por su nivel de solvencia, tienen una probabilidad de mora nula.

1.2.1. Tiempo de vida de una operación crediticia

En el estudio de datos de tiempos de vida del tipo *tiempo hasta que ocurre un suceso*, existen diversas estructuras de observación de los datos, lo que habitualmente crea ciertos problemas para analizarlos. Uno de estos problemas es el de la *censura* que, en un sentido informal, significa que los tiempos de vida sólo pueden ser observados dentro de un determinado intervalo de tiempo posiblemente aleatorio. En ese sentido, en la medición del riesgo de crédito resulta natural relacionar el tiempo hasta que un acreditado deja de pagar, es decir, *el tiempo hasta el incumplimiento*, con la vigencia o *tiempo de vida del crédito*. Como se verá más adelante, en el caso de exposiciones de carácter minorista, es frecuente encontrarse con un tipo particular de censura conocida como *censura aleatoria por la derecha*, lo que justifica el estudio de un modelo de supervivencia para estimar la PD en este tipo de carteras.

A continuación, en la Figura 1.1, se representa el mecanismo de censura que puede afectar a los tiempos de vida de una cartera de créditos personales. Para ello, en la Figura 1.1 se representan las variables aleatorias que serán estudiadas en esta memoria bajo la perspectiva del análisis de supervivencia. Como se puede apreciar en dicha figura, en este modelo se considera que el tiempo de vida de una operación crediticia es aleatorio y sólo se puede observar dentro del intervalo de tiempo $[0, \tau]$.

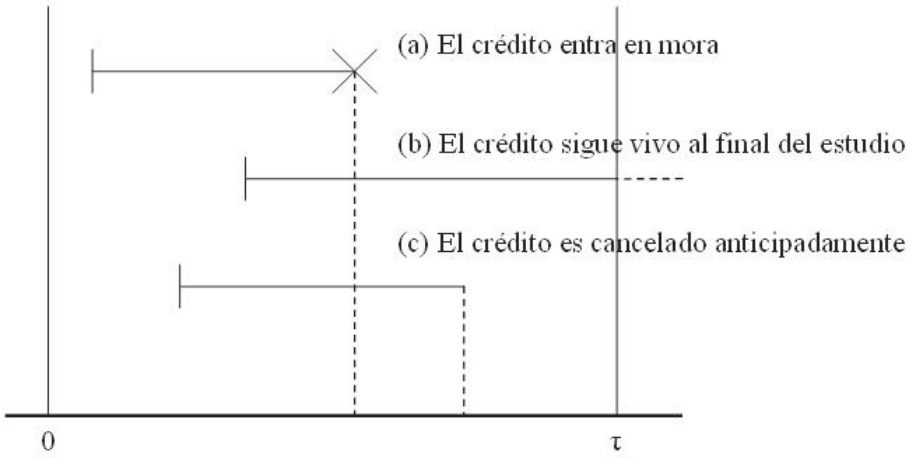

Figura 1.1 Tiempo de vida de un crédito observado en el intervalo $[0, \tau]$.

Las variables aleatorias relacionadas com la Figura 1.1 son el *tiempo hasta la entrada en mora*, T, el *tiempo de censura del crédito*, C, y la variable *indicadora de no censura*, o lo que es lo mismo, la variable *indicadora de mora*, δ. A continuación, se explican las situaciones representadas en la Figura 1.1.

(a) En la primera situación, se observa que el crédito entra en mora en un instante del tiempo entre 0 y τ. Esto implica que el tiempo en que se produce la mora del crédito es registrado (observado) por la entidad. Según esto, se tiene que $T \leq C$, o bien, que $\delta = 1$. Desde el punto de vista estadístico, debido a que el suceso de entrada en mora ocurre antes del fin del estudio, se dice que el tiempo de vida del crédito es *no censurado*.

(b) En este caso se aprecia que, de ocurrir un impago, el tiempo que tarda éste en producirse es superior a τ, es decir, es superior al máximo tiempo observable por la entidad. En este caso, sólo se observa que $T > C$, o bien, que $\delta = 0$, y por tanto, se dice que el tiempo de vida del crédito es *censurado por la derecha*.

(c) La tercera situación merece especial atención, pues de ella depende gran parte del planteamiento del modelo que se estudia a patir del Capítulo 3 en adelante. En la Figura 1.1 se observa que el tiempo de vida del crédito finaliza antes de que acabe el estudio en el tiempo τ, y que la mora del crédito no ha sido observada. Esta situación puede deberse a dos motivos. El primero de ellos es la *cancelación anticipada* del crédito, y aunque su ocurrencia es poco frecuente debido a que típicamente esta opción conlleva una

penalización económica para el acreditado, en esta memoria se ha considerado como una segunda clase de *censura aleatoria por la derecha*, y por tanto, la entidad sólo observa que $T > C$ ($\delta = 0$). La segunda causa se debe a que el plazo del crédito es inferior a la longitud del intervalo $[0, \tau]$, y por tanto, el acreditado cumple íntegramente con el pago de la deuda antes de concluir el estudio, es decir, en vencimiento.

1.2.2. Enfoque basado en el análisis de supervivencia

Siguiendo la idea original de Narain (1992), según el enfoque basado en el análisis de supervivencia, la información disponible para el cálculo de la probabilidad de mora condicional en una cartera de créditos personales, consiste en una muestra aleatoria de n observaciones independientes e idénticamente distribuidas (*i.i.d.*), $\{(\xi_1, \boldsymbol{X}_1, \delta_1), \ldots, (\xi_n, \boldsymbol{X}_n, \delta_n)\}$, con la misma distribución del vector aleatorio, $(\xi, \boldsymbol{X}, \delta)$, donde $\xi = \min\{T, C\}$ es el tiempo de vida observado del crédito, T es el tiempo hasta que se produce la mora del crédito, C es el tiempo de censura, $\delta = I(T \leq C)$ es la variable indicadora de que se ha producido la mora y $\boldsymbol{X}$ es un vector de covariables explicativas. En este enfoque se cumplen los siguientes supuestos:

H 1.1 *Existe una relación de dependencia entre las variables T y $\boldsymbol{X}$.*

H 1.2 *Cada acreditado tiene una probabilidad de mora que depende del valor del vector de covariables, $\boldsymbol{X}$, asociada al mismo.*

H 1.3 *Las variables aleatorias T y C son condicionalmente independientes dada la covariable $\boldsymbol{X}$.*

Según lo anterior, es posible caracterizar la función de distribución condicional de T dada $\boldsymbol{X} = \boldsymbol{x}$ usando algunas funciones usuales del análisis de supervivencia. Tales funciones son: la función de supervivencia condicional, $S(t|\boldsymbol{x})$, la función razón de fallo condicional, $\lambda(t|\boldsymbol{x})$, la función de fallo acumulativa condicional, $\Lambda(t|\boldsymbol{x})$ y la función de distribución condicional, $F(t|\boldsymbol{x})$, definidas respectivamente por:

$$S\left(t|\boldsymbol{x}\right) = P\left(T > t|\boldsymbol{X} = \boldsymbol{x}\right) = \int_t^{\infty} f(u|\boldsymbol{x})du, \tag{1.1}$$

$$\lambda(t|\boldsymbol{x}) = \lim_{\Delta t \to 0} \frac{P\left(t \leq T < t + \Delta t | T \geq t, \boldsymbol{X} = \boldsymbol{x}\right)}{\Delta t},$$

$$\Lambda(t|\boldsymbol{x}) = \int_0^t \lambda(u|\boldsymbol{x}) du,$$

$$F(t|\boldsymbol{x}) = 1 - S\left(t|\boldsymbol{x}\right) = 1 - e^{-\Lambda(t|\boldsymbol{x})}. \tag{1.2}$$

Debido a que el objetivo principal de esta memoria es modelizar la probabilidad de mora en créditos personales, a continuación se presenta la relación existente entre la probabilidad de mora condicional, $PD\left(t|\boldsymbol{x}\right)$, y la función de distribución condicional del tiempo hasta la entrada en mora, $F(t|\boldsymbol{x})$.

1.2.3. Cálculo de la PD por el método de la función de distribución condicional

Siguiendo la definición de probabilidad de mora que ha sido descrita en la sección 1.1.3, Cao et al. (2009) proponen una función para calcular la probabilidad de mora condicional y estudian su aplicación a un caso con datos reales de créditos al consumo. Según han determinado estos autores, la fórmula para calcular la probabilidad de mora de un crédito en el instante $t + b$ condicionado a un valor fijo de madurez, t, y a un valor observado, $\boldsymbol{x}$, del vector de covariables, $\boldsymbol{X}$, se obtiene a partir de la expresión:

$$\begin{aligned} PD\left(t|\boldsymbol{x}\right) &= \varphi\left(t|\boldsymbol{x}\right) = P\left(t \leq T < t + b | T \geq t, \boldsymbol{X} = \boldsymbol{x}\right) \\ &= \frac{P\left(T \leq t + b | \boldsymbol{X} = \boldsymbol{x}\right) - P\left(T \leq t | \boldsymbol{X} = \boldsymbol{x}\right)}{1 - P\left(T \leq t | \boldsymbol{X} = \boldsymbol{x}\right)} \\ &= \frac{F(t + b|\boldsymbol{x}) - F\left(t|\boldsymbol{x}\right)}{1 - F\left(t|\boldsymbol{x}\right)} \\ &= \frac{1 - F\left(t|\boldsymbol{x}\right) - (1 - F(t + b|\boldsymbol{x}))}{1 - F\left(t|\boldsymbol{x}\right)} \\ &= 1 - \frac{S\left(t + b|\boldsymbol{x}\right)}{S\left(t|\boldsymbol{x}\right)}, \end{aligned} \tag{1.3}$$

donde t es la madurez observada del crédito y b es el horizonte de predicción de la entrada en mora, cuyo valor es típicamente igual o mayor a un año, según lo recomendado por Basilea II.

Como se puede apreciar en la fórmula (1.3), una vez que se obtiene un estimador de la función de supervivencia condicional, $S(t|\boldsymbol{x})$, definida en (1.1), reemplazando dicha expresión en (1.3) se obtiene como resultado un estimador de la función de probabilidad de mora condicionada al perfil crediticio $\boldsymbol{x}$, $PD(t|\boldsymbol{x})$.

En los siguientes capítulos de esta memoria se estudia la relación existente entre T y $\boldsymbol{X}$ siguiendo la metodología propuesta por Cao et al. (2009), donde los autores han investigado diferentes métodos para estimar las funciones definidas en (1.2) y en (1.3).

1.3. Otros enfoques de medición del riesgo de crédito

La creciente necesidad de los bancos y agencias calificadoras por contar con modelos eficientes de calificación de los acreditados y, a partir de ello, obtener estimaciones fiables de los parámetros del riesgo de crédito, ha permitido que surja un amplio abanico de enfoques metodológicos para obtener dichas estimaciones. Uno de los enfoques más importantes que existe en la actualidad es el modelo regulatorio de Basilea, que como su nombre lo indica, es el modelo establecido en Basilea II para proveer de un método de medición y estimación de los parámetros del riesgo de crédito a todas aquellas entidades que han decido adoptar el *método estándar* de calificación crediticia. Además, este modelo también puede ser utilizado en el caso de las entidades que han optado por la implantación de *métodos IRB* para administrar su riesgo de crédito. A continuación se explicará brevemente en qué consiste el enfoque unifactorial de Basilea para la medición del riesgo de crédito.

1.3.1. El modelo factorial de Basilea II

La utilización de modelos factoriales para la obtención de la distribución de pérdidas asociada a una cartera de créditos permiten replicar comportamientos de impago de distintos tipos de acreditados. En particular, los modelos factoriales han sido diseñados para ser empleados en el control y en la administración del riesgo de crédito debido a exposiciones de carácter corporativo, es decir, cuando los acreditados son industrias, empresas,

bancos, naciones, etc. En este tipo de modelos es crucial tener en cuenta la estructura de correlación existente entre los acreditados, y al mismo tiempo respetar el principio de parsimonia del modelo, con el fin de obtener resultados asequibles desde el punto de vista analítico. En términos generales, los modelos factoriales surgen de la idea de Vasicek (1997) de valorar las pérdidas de una cartera de créditos adaptando para ello las técnicas para determinar el *valor-en-riesgo* (VaR) de la firma, desarrolladas por Black y Sholes (1973) y Merton (1974). En dicho modelo se determina que el suceso de impago de la empresa es un proceso latente que se produce en función del valor que toman sus activos. En particular lo que observa el banco no es el impago en sí mismo sino un indicador indirecto de ello, esto es, una empresa se considerará que entrará en mora si el valor de sus activos está por debajo del valor de la deuda contraída. De esta forma, la probabilidad de mora es la probabilidad de que el valor del activo de una empresa sea inferior al valor de su deuda.

En la literatura existen numerosas referencias a los modelos factoriales de riesgo de crédito. En concreto Finger (1999), Gordy (2000), Schönbucher (2000), Finger (2001) y Lucas et al. (2001) estudian el modelo unifactorial para la obtención analítica de la distribución de pérdidas de la cartera bajo el principio de *granularidad* (aproximación cuando el número de exposiciones tiende a infinito). Además, Trucharte y Antuña (2001), Gordy (2003) y Hamerle et al. (2003) estudian resultados fundamentales para los modelos asintóticos unifactoriales de riesgo de crédito, como el modelo de Basilea II.

1.3.2. Formulación del modelo unifactorial

A continuación se describe brevemente la formulación del modelo de incumplimiento en exposiciones corporativas para el cual se deriva la distribución del número de morosos en función de la probabilidad de mora condicional con un único factor estructural.

Debido a que el modelo elegido por el Comité de Basilea como propuesta metodológica para la reforma de Basilea I, es un modelo factorial, es interesante exponer aquí la versión simplificada del modelo regulatorio basado en un sólo factor.

Por definición, el suceso de entrada en mora de una compañía es una variable aleatoria latente y el estudio de un modelo para determinar el número

de empresas que caerán en dicha situación requiere de la definición de otra variable aleatoria auxiliar que sirva para cuantificar dicha morosidad. Para esto, se define la variable aleatoria auxiliar, V, que servirá para medir el valor de los activos de la firma. La conceptualización de esta variable aleatoria se basa en la siguiente hipótesis:

H 1.4 *El valor de la firma, V, depende de un factor aleatorio, Z, único y común para todas las empresas (los acreditados). V depende además, de un componente idiosincrático, también aleatorio, ε, que recoge las características propias e individuales de cada acreditado.*

El concepto anterior se puede representar de la siguiente manera:

$$V = \sqrt{\rho}\, Z + \sqrt{1-\rho}\, \varepsilon, \tag{1.4}$$

donde las variables Z y ε, se suponen independientes y con la misma distribución de probabilidad, la distribución normal estándar.

El modelo definido en (1.4) asume que los valores de los activos de dos acreditados cualesquiera tienen un coeficiente de correlación igual a ρ.

El principal problema de la entidad es poder determinar las pérdidas de la cartera sabiendo que no existe independencia entre los acreditados en el momento de convertirse en morosos. Como ya se ha dicho, el valor de los activos de todos los acreditados depende de un factor que es común para todos (por ejemplo, el ciclo económico) y suponemos, bajo este enfoque, que es el responsable de que los impagos tiendan a ocurrir al mismo tiempo (por ejemplo una recesión). También se ha establecido que existe otro componente, ε, propio de cada acreditado, que también influye en el comportamiento del valor de sus activos. Aquí radica entonces la importancia de los modelos factoriales. En efecto, si el factor común queda fijado en un determinado valor (es decir, la economía se encuentra en una determinada fase del ciclo), el único condicionante del valor de los activos, y en consecuencia de la probabilidad de impago, es el componente idiosincrásico, ε, que es propio de cada acreditado y, por tanto, proporciona la *independencia condicional entre acreditados*. De esta forma, se obtiene que la probabilidad de mora condicionada a que el factor sistemático Z tome el valor z viene dada por la siguiente expresión:

$$PD(z) = P(V < K | Z = z), \tag{1.5}$$

donde K es un determinado valor umbral de los activos bajo el cual, según Merton (1974), el acreditado se convierte en moroso. Así, de la ecuación (1.5) se desprende, por tanto, que la probabilidad de impago del acreditado es equivalente a la probabilidad de que V sea menor que el umbral K, para cada acreditado.

A continuación obtendremos la distribución de probabilidad condicional para el número de créditos morosos, D, en el caso del modelo unifactorial. Para esto seguiremos el procedimiento empleado por Schönbucher (2000, págs. 8-9).

Empleando la ley de esperanzas iteradas, se puede demostrar que la variable aleatoria D condicionada al factor Z sigue una ley de probabilidad binomial. Así, la probabilidad de que en una cartera de tamaño N existan exactamente n créditos morosos viene dada por la siguiente fórmula:

$$P\left(D=n\right)=\int_{-\infty}^{\infty}P\left(D=n|Z=z\right)\phi(z)dz, \tag{1.6}$$

donde ϕ es la función de densidad normal estándar.

Condicionalmente a $Z=z$ y fijado un valor umbral, K, idéntico para toda la cartera, la probabilidad condicional de que existan exactamente n créditos morosos viene dada por la siguiente fórmula:

$$P\left(D=n|Z=z\right)=\binom{N}{n}\left(PD(z)\right)^{n}\left(1-PD(z)\right)^{N-n}, \tag{1.7}$$

donde $PD(z)$ es la función de probabilidad de mora condicional de la firma cuando el factor sistemático Z toma el valor z y N es el número total de créditos. Por su parte, la función $PD(z)$ se calcula a partir del siguiente razonamiento:

$$\begin{aligned} PD(z) &= P\left(V<K|Z=z\right) \\ &= P\left(\sqrt{\rho}\,Z+\sqrt{1-\rho}\,\varepsilon<K|Z=z\right) \\ &= P\left(\varepsilon<\frac{K-\sqrt{\rho}\,Z}{\sqrt{1-\rho}}|Z=z\right) \\ &= \Phi\left(\frac{K-\sqrt{\rho}\,z}{\sqrt{1-\rho}}\right). \end{aligned} \tag{1.8}$$

Llamando $K_{\rho,z}$ al valor $\frac{K-\sqrt{\rho}\,z}{\sqrt{1-\rho}}$ y sustituyendo (1.7) y (1.8) en (1.6) se obtiene que:

$$P(D=n)=\int_{-\infty}^{\infty}\binom{N}{n}\left(\Phi\left(K_{\rho,z}\right)\right)^{n}\left(1-\Phi\left(K_{\rho,z}\right)\right)^{N-n}\phi(z)dz, \tag{1.9}$$

donde Φ es la función de distribución acumulativa normal estándar.

De este modo, a partir del resultado (1.9), se obtiene que la distribución de probabilidad acumulativa del número de créditos morosos en una cartera de exposiciones corporativas viene dada por siguiente la fórmula:

$$P(D\leq m)=\sum_{n=0}^{m}\binom{N}{n}\int_{-\infty}^{\infty}\left(\Phi\left(K_{\rho,z}\right)\right)^{n}\left(1-\Phi\left(K_{\rho,z}\right)\right)^{N-n}\phi(z)dz.$$

Finalmente, el lector interesado en profundizar sobre el enfoque de modelización factorial del riesgo de crédito, puede revisar algunas publicaciones referidas a versiones comerciales de estos modelos que han sido desarrollados por agencias de calificación internacionales. Ver por ejemplo los trabajos de Morgan (1997), Credit Suisse Financial Products (1997) y Wilson (1997a, 1997b), entre otros. Además de lo anterior, también existe una amplia colección de trabajos de carácter recopilatorio ampliamente difundidos en la literatura sobre riesgo de crédito, tal es el caso de Crouhy et al. (2000), Gordy (2000), Saunders y Allen (2002), Altman et al. (2003) y Bluhm et al. (2003), entre otros.

Capítulo 2

Construcción de un modelo de puntuación crediticia

2.1. Introducción

El estudio del perfil crediticio por parte de las entidades financieras e instituciones financieras afines tiene su origen a comienzos de la segunda mitad del siglo XX. En esa época, la evaluación del perfil crediticio se realizaba principalmente mediante opiniones subjetivas basadas en la experiencia empírica adquirida por los analistas, lo que se conoce como *criterio de expertos*. Particularmente, las decisiones sobre la concesión de los créditos se basaban en la información aportada por cuatro características fundamentales de los clientes, método denominado como *las 4 C*:

- *El carácter*: carcaterística que mide la reputación del cliente.
- *El capital*: característica que mide la capacidad de retorno de la inversión.
- *La capacidad*: característica que mide la capacidad para generar ganancias.
- *El colateral*: característica que cuantifica el valor del bien en garantía.

Si bien, esta forma de administrar la comercialización de los créditos bancarios predominó en gran parte del mundo hasta los años 70, en países

como Estados Unidos, Canadá, Japón, Suiza y otros países industrializados, se hizo cada vez más frecuente la combinación del análisis subjetivo con el estudio de los datos por medio de modelos matemáticos de predicción. El objetivo era mejorar la capacidad para predecir el impago de los créditos y reducir el riesgo de pérdida por insolvencia para la entidad.

Inicialmente, las herramientas estadísticas empleadas por las entidades financieras para estudiar el perfil de sus clientes se reducían a unos cuantos métodos de discriminación y clasificación inspirados en el *análisis discriminante lineal* de Fisher (1936). Según algunos autores (Hand y Henley (1997) y Crook et al. (2007)), la primera experiencia formal con un modelo de análisis crediticio basado en análisis discriminante lineal (ADL) se le atribuye al trabajo realizado por Durand (1941), quien obtuvo buenos resultados prediciendo el comportamiento de impagos de un grupo de 37 compañías en Estados Unidos a comienzos del siglo XX.

Posteriormente, aparece una nueva clase de modelos ampliamente utilizados en el contexto de los *modelos de puntuación creditica*, también llamados *modelos de credit scoring*, se trata del modelo de análisis discriminante de Altman (1968), cuyo modelo denominado Z-score[1] es considerado una herramienta pionera en el contexto del riesgo de crédito corporativo. Más tarde, Altman et al. (1977) introducen modificaciones en el modelo Z-score original dando lugar a un nuevo modelo, el modelo $ZETA$. El objetivo de esta clase de modelos de riesgo de crédito es el de predecir la quiebra de la compañía en base a una puntuación proveniente de la combinación lineal de los factores que explican la solvencia y la calidad de la gestión financiera de esta. Otros autores que han estudiado y comparado la aplicación del ADL con otras técnicas de modelización en el contexto de la puntuación crediticia son Myers y Forgy (1963), Lane (1972), Apilado et al. (1974) y Moses y Liao (1987), entre otros.

A partir de trabajos como los de Edward Altman, muchos autores han estudiado las ventajas y desventajas de utilizar modelos de análisis discriminante en el ámbito de los negocios, en finanzas y en otras áreas de la economía.

[1] El modelo Z-score del profesor Edward Altman (1968) fue desarollado como una herramienta de predicción de la quiebra de una compañía. Este modelo ha gozado durante décadas de una gran aceptación entre los analistas financieros de todo el mundo y su capacidad predictiva ha sido ampliamente estudiada. Véanse, por ejemplo, los trabajos de Scott (1981), Begley et al. (1996), Altman (2000), Grice e Ingram (2001), Altman (2002) y Alexakis (2008), entre otros.

Ver, por ejemplo, las críticas que se han hecho sobre este tema en los trabajos de Eisenbeis (1977, 1978) y de Rosenberg y Gleit (1994). Eisenbeis (1977) señala que para que una regla de clasificación basada en ADL sea óptima se requiere que las poblaciones a las que pertenecen los sujetos, es decir, los acreditados o solicitantes de los créditos, estén caracterizadas por distribuciones multivariantes elipsoidales, de las que la normal multivariante es un caso particular. Sin embargo, es conocido por quienes se dedican al estudio del riesgo financiero, que la normalidad de las variables crediticias es un supuesto fuertemente restrictivo que en muchos casos no es posible verificar. Este problema ha sido tratado en trabajos como los debidos a Reichert et al. (1983), Moses y Liao (1987) y Hand y Henley (1997), entre otros.

Como ya se ha mencionado, los sistemas de evaluación y estudio del perfil crediticio han ido evolucionando en relación con su poder predictivo desde hace décadas. El buen desempeño de estos modelos se ha convertido en un objetivo fundamental para garantizar la eficiencia de las políticas de administración y control del riesgo de crédito. Esto se refleja claramente en el Pilar I de Basilea II (Basel Committee on Banking Supervision (1999)). Allí se han establecido normas y patrones acerca de los métodos de calibración de los modelos de puntuación crediticia bajo el enfoque de *métodos basados en calificaciones internas* (ver Sección 1.1.2). En la actualidad existe abundante literatura sobre la utilización de técnicas matemáticas para la construcción de modelos de puntuación crediticia. Estas técnicas se agrupan en 3 enfoques metodológicos. El primero corresponde a técnicas clásicas de la estadística paramétrica y no paramétrica. De entre ellas, las técnicas más utilizados son:

- El análisis de regresión lineal multivariante.
- Los modelos lineales generalizados (modelos *probit*, modelos *logit* y modelos *tobit*, entre otros).
- El análisis discriminante no paramétrico (el método de vecinos más cercanos y el método de árboles de decisión, entre otros).

Algunos autores que han estudiado los modelos de puntuación crediticia bajo este enfoque son, entre otros, Beaver (1967), Platt y Platt (1991), Greene (1992), Lawrence et al. (1992), Henley y Hand (1996) y Hand y Henley (1997).

El segundo enfoque, más reciente que el primero, ha ido emergiendo con fuerza en los últimos años desde que se han incorporado técnicas de automatización y de *data mining* (minería de datos) al análisis del riesgo de crédito. Este enfoque está compuesto por técnicas tales como:

- Los métodos de programación matemática y optimización.
- Los sistemas expertos y de inteligencia artificial (redes de neuronas y máquinas con soporte vectorial).
- Los métodos basados en técnicas evolutivas y en algoritmos genéticos.
- Los métodos híbridos basados en la combinación de técnicas correspondientes al enfoque clásico y al enfoque basado en máquinas con soporte vectorial y redes neuronales.

Existe una extensa literatura relacionada con este último enfoque. Ver, por ejemplo, los trabajos de Desai et al. (1996, 1997), West (2000), Yobas et al. (2000), Baesens et al. (2003), Van Gestel et al. (2005), West et al. (2005), Chen et al. (2006), Martens et al. (2007), Min y Lee (2008), Twala (2009) y Martens et al. (2011), entre otros.

Adicionalmente, en la literatura dedicada al análisis del riesgo de crédito existen obras en las que se recogen gran parte de las metodologías correspondientes a ambos enfoques. Ver, por ejemplo, los trabajos de Thomas et al. (1992), Altman y Saunders (1998), Hand (2001), Thomas et al. (2002), Crook et al. (2007), Baesens et al. (2009) y Baesens (2014).

En el presente capítulo se estudia la construcción de un modelo de puntuación crediticia utilizando técnicas de regresión logística univariante. El objetivo es elaborar una metodología estadística que permita obtener una herramienta de medición de la solvencia de quienes solicitan un crédito, o bien, de la conducta de pago de quienes poseen un crédito con una entidad financiera. El poder discriminante del modelo es evaluado en base a la aplicación del mismo a un estudio empírico con datos reales de tarjetas de crédito.

Además, cabe señalar que en los capítulos siguientes de esta memoria se estudian modelos de riesgo de crédito donde la covariable *puntuación crediticia* se puede construir siguiendo la metodología expuesta en este capítulo.

2.2. Modelo de puntuación crediticia vía regresión logística

Los modelos de puntuación crediticia forman parte de las herramientas cuantitativas utilizadas por las entidades financieras para evaluar el *perfil crediticio* de sus clientes. Según la literatura, el análisis de regresión logística es una de las técnicas más utilizadas en la construcción de esta clase de modelos. La popularidad de esta técnica de regresión se debe principalmente a su flexibilidad, ya que un modelo de regresión logística permite utilizar variables regresoras con distintos niveles de medición, y a la posibilidad de obtener un alto poder de discriminación entre clientes buenos y malos. Estas cualidades permiten aprovechar la máxima información disponible sobre el grado de solvencia y el comportamiento de pago de los clientes. El modelo logístico de puntuación crediticia ha sido estudiado, entre otros, por Wiginton (1980), Lo (1985), Srinivasan y Kim (1987), Steenackers y Goovaerts (1989), Thomas et al. (2002), Zellner et al. (2004), Bandyopadhyay (2006), Sun y Wang (2007) y Samreen y Zaidi (2012), entre otros.

El estudio de los modelos de regresión logística y sus aplicaciones en contextos más generales se pueden encontrar en obras de divulgación más extensas, como las debidas a Long (1997), Gourieroux (2000), Hosmer y Lemeshow (2000), Agresti (2002) y Green (2008), entre otros.

2.2.1. Formulación del modelo

Se define el vector aleatorio, $\boldsymbol{X}_i = (1, X_{i1}, ..., X_{ip-1}) \in \mathbb{R}^p$, que caracteriza el perfil crediticio del i-ésimo solicitante (o acreditado) con $1 \leq i \leq n$, siendo n el tamaño de la cartera de créditos. Además, la *función score* o *función de puntuación crediticia* se define como:

$$\psi : \mathbb{R}^p \to \mathbb{R} \quad \text{tal que} \quad \boldsymbol{x} \to \psi(\boldsymbol{x}) = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_j, \tag{2.1}$$

donde $\boldsymbol{\beta} = \left(\beta_0, ..., \beta_{p-1}\right)' \in \mathbb{R}^p$ es el vector de parámetros de la función score, $\psi(\boldsymbol{x})$, que es fijo y desconocido para toda la catera de créditos. El valor de $\psi(\boldsymbol{x})$ representa la importancia que tiene para el modelo la contribución conjunta de las variables crediticias.

Por simplicidad de notación, en ocasiones se escribirá $Z = \psi(\boldsymbol{X})$ para referirse a la función score y $z = \psi(\boldsymbol{x})$ para sus valores observados.

El modelo logístico de puntuación crediticia verifica las siguientes hipótesis:

H 2.1 *Sea Y_i la variable aleatoria indicadora de insolvencia (morosidad) del solicitante (acreditado) i-ésimo. Se verifica que las variables Y_i son condicionalmente independientes dada la covariable $\boldsymbol{X}_i$ y siguen una distribución de Bernoulli de parámetro π_i, que depende de la puntuación, Z_i, de forma que $\pi_i \equiv \pi(Z_i)$ con $Z_i \equiv Z_i(\boldsymbol{X}_i) = \psi(\boldsymbol{X}_i)$.*

La distribución de Y_i condicionada a $\boldsymbol{X}_i$ viene dada por:

$$Y_i|\boldsymbol{X}_i = \begin{cases} 1 & con \ \ probabilidad \quad \pi(Z_i) \\ 0 & con \ \ probabilidad \quad 1-\pi(Z_i) \end{cases} \tag{2.2}$$

donde $0 < E(Y_i|\boldsymbol{X}_i) = \pi(Z_i) < 1$ y $V(Y_i|\boldsymbol{X}_i) = \pi(Z_i)(1-\pi(Z_i))$ para todo i, con $1 \leq i \leq n$. Por simplicidad de notación, en ocasiones se utilizará la igualdad $\pi_i = \pi(Z_i) = \pi(\psi(\boldsymbol{X}_i))$ para hacer explícita la dependencia del parámetro π_i sobre la variable $\boldsymbol{X}_i$.

H 2.2 *La función de probabilidad condicional de mora o insolvencia, $\pi(z)$, sigue un modelo logístico y su fórmula viene dada por:*

$$\pi(z) = P(Y = 1|Z = z) = \frac{e^z}{1+e^z}. \tag{2.3}$$

De la fórmula (2.3) se obtiene que $\lim_{z \longrightarrow +\infty} \pi(z) = 1$ y $\lim_{z \longrightarrow -\infty} \pi(z) = 0$. Esto significa que cuando la puntuación crediticia tiende a $+\infty$, la probabilidad de insolvencia crece aproximándose a 1 mientras que cuando la puntuación tiende a $-\infty$, la probabilidad de insolvencia decrece aproximándose a 0.

2.2.2. Estimación del modelo

Sea $\{(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n)\}$ una muestra aleatoria de vectores mutuamente independientes donde $Y_i\,|\boldsymbol{X}_i \overset{d}{\equiv} Bernoulli(\pi_i)$ y $\boldsymbol{X}_i$ caracteriza el perfil del cliente i-ésimo como se definió previamente en el apartado 2.2.1. Denotando por $\boldsymbol{u}_i = (y_i, \boldsymbol{x}_i)$ al valor observado del vector $(Y_i, \boldsymbol{X}_i)$, con $1 \leq i \leq n$,

y bajo las hipótesis H2.1 y H2.2, el estimador máximo verosímil, $\hat{\boldsymbol{\beta}}$, del vector de parámetros del modelo de regresión, $\boldsymbol{\beta}$, se obtiene maximizando la siguiente función de verosimilitud condicionada:

$$\begin{aligned} L\left(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n,\boldsymbol{\beta}\right) &= \prod_{i=1}^{n} P\left(Y=y_i|\, Z=z_i\right) \\ &= \prod_{i=1}^{n} \left(\pi_i\right)^{y_i}\left(1-\pi_i\right)^{1-y_i} \\ &= \prod_{i=1}^{n} \left(\frac{e^{z_i}}{1+e^{z_i}}\right)^{y_i}\left(\frac{1}{1+e^{z_i}}\right)^{1-y_i} \\ &= \prod_{i=1}^{n} \frac{e^{y_i\boldsymbol{x}_i\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}}. \end{aligned} \tag{2.4}$$

Debido a la dificultad analítica de buscar el máximo de la función obtenida en (2.4), en su lugar, es habitual maximizar el logaritmo natural de dicha función, es decir, se busca maximizar la función de log-verosimilitud definida por:

$$\ell(\boldsymbol{\beta}) = \ln\left(L\left(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n,\boldsymbol{\beta}\right)\right) = \sum_{i=1}^{n}\left(y_i\boldsymbol{x}_i\boldsymbol{\beta} - \ln\left(1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}\right)\right). \tag{2.5}$$

Tomando derivadas parciales de $\ell(\boldsymbol{\beta})$ respecto de $\boldsymbol{\beta}$ e igualándolas a cero se obtiene el siguiente sistema de p ecuaciones no lineales:

$$\frac{\partial\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = 0 \Longrightarrow \left\{ \begin{array}{c} \sum_{i=1}^{n}\left(y_i-\pi\left(z_i\right)\right)=0 \quad \text{para } j=0 \\ \sum_{i=1}^{n} x_{ij}\left(y_i-\pi\left(z_i\right)\right)=0 \quad \text{para } j=1,...,p-1 \end{array} \right.$$

cuya solución es el estimador máximo verosímil de $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}\,\in\,\mathbb{R}^p}\ \ell(\boldsymbol{\beta}). \tag{2.6}$$

Finalmente, reemplazando (2.6) en la fórmula (2.3), se obtiene el estimador de la probabilidad condicional, π_i, que viene dado por:

$$\hat{\pi}_i = \pi\left(\hat{z}_i\right) = \frac{e^{\hat{z}_i}}{1+e^{\hat{z}_i}}, \tag{2.7}$$

donde $\hat{z}_i = \hat{\psi}\left(\boldsymbol{x}_i\right) = \hat{\beta}_0 + \sum_{j=1}^{p-1}\hat{\beta}_j x_{ij}$ para todo i, con $1 \leq i \leq n$.

En los siguientes capítulos de esta memoria se estudian modelos para el cálculo de la PD en créditos personales mediante técnicas de análisis de supervivencia. En dichos modelos se utiliza la covariable *puntuación crediticia* que, bajo hipótesis H2.1 y H2.2, se obtiene de la fórmula (2.7).

2.2.3. Tratamiento de las variables crediticias

Un importante problema a tomar en cuenta cuando se desarrolla un modelo de puntuación crediticia es el tratamiento de sus covariables. Esto conduce a estudiar criterios de inclusión para dichas variables en el modelo de regresión. En ese sentido, a continuación se presentan tres tipos de criterios ampliamente estudiados en análisis de regresión logística aplicada al riesgo de crédito.

Se redefine el vector de variables crediticias en la forma $\boldsymbol{X} = (1, \boldsymbol{X}_1, \boldsymbol{X}_2)$, donde, sin pérdida de generalidad, $\boldsymbol{X}_1 = (X_1, ..., X_m)$, con $m < p$, está compuesto por m características crediticias absolutamente continuas y $\boldsymbol{X}_2 = (X_{m+1}, ..., X_{p-1})$ es el vector compuesto por las $p - 1 - m$ características restantes que reciben el nombre de *factores* por tratarse de variables categóricas.

Tratamiento de variables continuas

El modelo de regresión logística permite incorporar variables continuas con dominio en todo $\mathbb{R}$ mediante la combinación lineal definida en (2.1). Sin embargo, a veces resulta conveniente transformar previamente las variables con el fin de evitar problemas difíciles de tratar con técnicas clásicas, como los efectos provocados por las diferentes escalas de cada una de las covariables originales, la falta de normalidad, la multicolinealidad o la ausencia de correlación entre las covariables y la puntuación crediticia, Z, entre otras anomalías que surgen en bases de datos extensas, como las pertenecientes a las entidades financieras. Aranda-Ordaz (1981), Guerrero y Johnson (1982) y Kay y Little (1987), entre otros, han estudiado este problema en modelos de regresión con variable respuesta binaria y sugieren como posible solución utilizar el método de las transformaciones de Box y Cox (1964). No obstante, debido a que dicho método tiene por finalidad conseguir la normalidad de la variable respuesta, en este capítulo se propone estudiar métodos alternativos

para el tratamiento de las variables crediticias continuas. Las técnicas que se estudian son dos:

Método de regresión segmentada. Este método consiste en ajustar una serie de segmentos rectilíneos sucesivos cuando, al representar la nube de puntos formada por la variable dependiente y la covariable, se observa la presencia de una tendencia lineal diferente en los distintos tramos del recorrido de la covariable.

Método de regresión polinómica. Este método se utiliza cuando, al representar la nube de puntos formada por la variable dependiente y la covariable, ésta última parece seguir una tendencia polinómica en distintos tramos de su recorrido. En tal caso, se ajusta un polinomio de grado mayor que 1 en cada uno de los segmentos de la regresión.

Método de regresión segmentada Sea X_{ij}, con $1 \leq j \leq m$, la j-ésima covariable continua asociada al perfil del i-ésimo solicitante (o acreditado). Supóngase además que, asociado a cada covariable X_j, existe un conjunto de r puntos de corte, c_{1j}, c_{2j}, ..., c_{rj}, tales que $c_{1j} < c_{2j} < \cdots < c_{rj}$. Los puntos c_{sj}, con $1 \leq s \leq r$, son constantes conocidas que determinan los límites de cada segmento del rango de X_j. Además, cada i-ésimo solicitante (o acreditado) posee una puntuación base inicial no nula, α_{0j}, según la covariable X_j.

El método consiste en ajustar sucesivos segmentos de recta, l_{ij}, definidos como:

$$l_{ij} = \begin{cases} \alpha_{0j} & \text{si} \quad c_{0j} \leq X_{ij} < c_{1j} \\ \alpha_{0j} + \alpha_{1j}(X_{ij} - c_{1j}) & \text{si} \quad c_{1j} \leq X_{ij} < c_{2j} \\ \alpha_{0j} + \sum_{u=1}^{s-2} \alpha_{uj}\delta_{uj} + \alpha_{s-1j}(X_{ij} - c_{s-1j}) & \text{si} \quad c_{s-1j} \leq X_{ij} < c_{sj} \\ & \text{para algún} \quad s : 3 \leq s \leq r \\ \alpha_{0j} + \sum_{u=1}^{r-1} \alpha_{uj}\delta_{uj} + \alpha_{rj}(X_{ij} - c_{rj}) & \text{si} \quad c_{rj} \leq X_{ij} < c_{r+1j} \end{cases}$$

donde se ha definido la variable auxiliar $\delta_{uj} = c_{u+1j} - c_{uj}$, α_{uj} es el coeficiente de la regresión lineal u-ésima en la j-ésima covariable y donde, por conveniencia, se definen los extremos $c_{0j} = -\infty$ y $c_{r+1j} = +\infty$ para todo j, con $1 \leq j \leq m$.

Equivalentemente, l_{ij} puede escribirse como:

$$l_{ij} = \sum_{s=1}^{r+1} I\left(c_{s-1j} \leq X_{ij} < c_{sj}\right)\left(\alpha_{0j} + \sum_{u=1}^{s-2} \alpha_{uj}(c_{u+1j} - c_{uj}) + \alpha_{s-1j}(X_{ij} - c_{s-1j})\right),$$

donde $I(A)$ es la indicadora del suceso A.

El valor de la puntuación, $Z_i = \psi(\boldsymbol{X}_i)$, correspondiente a la parte continua del perfil crediticio i-ésimo queda determinado por:

$$\begin{aligned} Z_i &= \sum_{j=1}^{m} l_{ij} \\ &= \sum_{j=1}^{m} \sum_{s=1}^{r+1} I\left(c_{s-1j} \leq X_{ij} < c_{sj}\right)\left(\beta_{0s-1j} + \beta_{1s-1j} X_{ij}\right), \end{aligned}$$

donde $\beta_{0j} = \alpha_{0j} + \sum_{u=1}^{s-2} \alpha_{uj}(c_{u+1j} - c_{uj}) - \alpha_{s-1j} c_{s-1j}$ y $\beta_{1j} = \alpha_{s-1j}$ son los coeficientes de la regresión lineal correspondiente a la j-ésima covariable, con $1 \leq j \leq m$.

En esta memoria se ha trabajado bajo el supuesto de que los puntos de corte, c_{sj}, son conocidos. Extensiones de este método donde los c_{sj} deben ser estimados han sido estudiados por Pastor y Guallar (1998), Pastor-Barriuso et al. (2003), Gurevich y Vexler (2005), Gill (2007) y Vexler y Gurevich (2009), entre otros.

Método de regresión polinómica Sea X_{ij}, con $1 \leq j \leq m$, la j-ésima covariable continua asociada al perfil del i-ésimo acreditado. Este método consiste en incluir en el modelo, como variables regresoras, transformaciones polinómicas de grado k_j de X_{ij}. Para ello, se definen las variables auxiliares $X_{ij}^{(s)} = \left(X_{ij} - \overline{X}_{\cdot j}\right)^s$, con $s = 1, 2, ..., k_j$, donde $\overline{X}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$, para algún entero positivo k_j.

El valor de la puntuación, $Z_i = \psi(\boldsymbol{X}_i)$, correspondiente a la parte continua del perfil crediticio i-ésimo queda determinado por:

$$\begin{aligned} Z_i &= \beta_0 + \sum_{j=1}^{m} \sum_{s=1}^{k_j} \beta_{sj} X_{ij}^{(s)} \\ &= \beta_0 + \sum_{j=1}^{m} \sum_{s=1}^{k_j} \beta_{sj} \left(X_{ij} - \overline{X_{\cdot j}}\right)^s, \end{aligned}$$

donde los β_{sj} son los coeficientes de la regresión que acompañan a los términos de grado k_j asociados a la j-ésima covariable.

En este punto, es importante señalar que las dos técnicas expuestas no son excluyentes, o incompatibles, y por tanto, en un modelo de puntuación crediticia unas variables continuas pueden ajustarse por el método de regresión segmentada y otras por el método de regresión polinómica.

Tratamiento de variables categóricas

Cuando las características crediticias analizadas por las entidades financieras son cualitativas y poseen más de dos categorías, se dice que dichas características corresponden a variables categóricas o *factores*, y por tanto, antes de ser incluidas en el modelo de regresión deben ser transformadas.

Supóngase que en el modelo de regresión se incluyen j factores, cada uno con r_j categorías nominales u ordinales. Entonces, para cada j-ésimo factor, se crean $r_j - 1$ variables de diseño dicotómicas llamadas *variables dummy*. Una vez creadas las variables dummy, para cada j-ésimo factor, se fija una de las categorías de la variable original, la que será la categoría de referencia, comparando el resto de categorías con ella.

En la literatura sobre modelos de regresión con covariables categóricas, existen numerosos trabajos relacionados con la utilización de variables dummy. Ver, por ejemplo, los trabajos de Long (1997), Gourieroux (2000), Agresti (2002) y Lawal (2003), entre otros.

Método de variables *dummy* Sea X_{ij}, con $m+1 \leq j \leq p-1$, el j-ésimo factor (covariable categórica) asociado al perfil del i-ésimo solicitante (o

acreditado) que posee r_j categorías distintas, nominales u ordinales, $\lambda_j^{(s)}$, con $1 \leq s \leq r_j$. Además, se definen las variables auxiliares $D_j^{(s)}$, con $1 \leq s \leq r_j$, a partir de la matriz de diseño que se muestra a continuación:

Tabla 2.1 Matriz de diseño para el j-ésimo factor

Categoría		Variable dummy asociada			
		$D_j^{(2)}$	$D_j^{(3)}$	$\cdots$	$D_j^{(r_j)}$
$\lambda_j^{(1)}$	(ref)	0	0	$\cdots$	0
$\lambda_j^{(2)}$		1	0	$\cdots$	0
$\lambda_j^{(3)}$		0	1	$\cdots$	0
$\vdots$		$\vdots$	$\vdots$		$\vdots$
$\lambda_j^{(r_j)}$		0	0	$\cdots$	1

En la Tabla 2.1 se observa que la primera categoría, $\lambda_j^{(1)}$, corresponde a la categoría de referencia del factor X_j. En este caso la codificación auxiliar es 0, y por tanto, si el solicitante i-ésimo posee un valor de $X_{ij} = \lambda_j^{(1)}$, entonces el valor de su puntuación, Z_i, será igual al término independiente, β_0, valor que coincide con la puntuación base para cada cliente de la cartera.

El valor de la puntuación, $Z_i = \psi(\boldsymbol{X}_i)$, correspondiente a la parte categórica del perfil crediticio i-ésimo queda determinado por:

$$\begin{aligned} Z_i &= \beta_0 + \sum_{j=m+1}^{p-1} \sum_{s=2}^{r_j} \beta_{sj} I\left(x_{ij} = \lambda_j^{(s)}\right) \\ &= \beta_0 + \sum_{j=m+1}^{p-1} \sum_{s=2}^{r_j} \beta_{sj} D_{ij}^{(s)}, \end{aligned}$$

donde los β_{sj} son los pesos correspondientes a cada s-ésima categoría del factor X_j.

2.2.4. Selección de las variables del modelo

Otro problema de gran trascendencia en la construcción de un modelo de puntuación crediticia estriba en poder determinar correctamente qué variables deben ser incluidas en el modelo de regresión. En el contexto del riesgo

de crédito, la literatura es más bien escasa en relación a este tema. Zellner et al. (2004) estudian la selección de variables crediticias comparando métodos del tipo *stepwise* o *paso a paso*, por medio de técnicas bootstrap a partir de un ejemplo con datos de créditos personales en Alemania.

En el contexto del riesgo actuarial y financiero, algunos autores han estudiado técnicas basadas en medidas de distancia y de similitud entre las variables. Ver, por ejemplo, los trabajos de Artís et al. (1994), Trías et al. (2005), Boj et al. (2007) y Boj et al. (2009) sobre este tema.

En contraste con lo anterior, en contextos distintos del ámbito financiero, el problema de la selección de variables en los modelos de regresión ha sido tratado extensamente en la literatura. Técnicas como el *contraste de Wald*, el *contraste de razón de verosimilitudes* y el *mejor subconjunto*, entre otras, han sido ampliamente estudiadas en modelos lineales generalizados y particularmente en regresión logística. Algunos trabajos en los que se tratan en profundidad dichas técnicas son, por ejemplo, las debidas a Hosmer y Lemeshow (2000), Agresti (2002), Greene (2008) y Hardin y Hilbe (2009), entre otros.

En este capítulo, se ha utilizado el método de inclusión de variables conocido como método stepwise (paso a paso) en combinación con los criterios de entrada (o salida) obtenidos con los contrastes de Wald y de razón de verosimilitudes.

Método de selección *stepwise* o *paso a paso* El método de selección de variables de tipo *stepwise* (o paso a paso) se debe a Efroymson (1960). Este método consiste en elegir el mejor modelo de manera secuencial incluyendo (o excluyendo) sólo una de las variables crediticias en cada paso de acuerdo con un criterio de entrada (o salida). Cada vez que una variable entra en el modelo, éste es cotejado con el modelo en el paso anterior, evaluando la ganancia de tal inclusión en base a los criterios mencionados anteriormente. El proceso de selección termina cuando una regla de parada previamente establecida es satisfecha y cuando se ha comprobado que ninguna de las variables no incluidas en el modelo es significativa.

En modelos de regresión logística, el método de selección *stepwise* ha sido estudiado, entre otros, por Nordberg (1981), Hosmer y Lemeshow (2000), Fahrmeir y Tutz (2001) y Shtatland et al. (2001).

Criterio basado en el contraste de *Wald* El contraste de *Wald* se utiliza para contrastar si una covariable debe ser incluida en el modelo de regresión logística en función de su significación estadística. La hipótesis nula establece que el coeficiente de regresión correspondiente a la j-ésima covariable es cero, es decir,

$$\begin{aligned} H_{0j} &: \beta_j = 0 \\ H_{1j} &: \beta_j \neq 0 \end{aligned} \tag{2.8}$$

El estadístico de *Wald*, denotado por $\hat{T}_{W_j}$, viene dado por:

$$\hat{T}_{W_j} = \frac{\hat{\beta}_j^{(k)}}{\sqrt{V\left(\hat{\beta}_j^{(k)}\right)}}, \tag{2.9}$$

donde $\hat{\beta}_j^{(k)}$ y $\sqrt{V\left(\hat{\beta}_j^{(k)}\right)}$ son los estimadores máximo verosímiles de β_j y de su correspondiente desviación estándar, obtenidos en el k-ésimo paso del ajuste del modelo. Se verifica que cuando la hipótesis nula en (2.8) es cierta, el cuadrado del estadístico definido en (2.9) converge en distribución a una $\chi^2_{(1)}$ para todo j, con $1 \leq j \leq p-1$.

Criterio basado en el contraste de *razón de verosimilitudes* El contraste de *razón de verosimilitudes* se define como el cociente entre el máximo valor que alcanza la función de log-verosimilitud del modelo sin la j-ésima covariable (*modelo reducido*) y el máximo valor alcanzado por dicha función con todas las covariables incluidas en el k-ésimo paso del ajuste del modelo (*modelo completo*). Este método es una alternativa al *contraste de Wald* para la selección de variables y también es utilizado como medida de bondad de ajuste global en el contexto de modelos generalizados de regresión.

El estadístico del contraste de razón de verosimilitudes, denotado por $\hat{T}_{RV_j}$, en términos de log-verosimilitudes, viene dado por:

$$\hat{T}_{RV_j} = -2\left(\ell(\hat{\boldsymbol{\beta}}_{-j}^{(k)}) - \ell(\hat{\boldsymbol{\beta}}^{(k)})\right), \tag{2.10}$$

donde $\ell(\hat{\boldsymbol{\beta}}_{-j}^{(k)})$ y $\ell(\hat{\boldsymbol{\beta}}^{(k)})$ se obtienen a partir de (2.5) siendo $\hat{\boldsymbol{\beta}}_{-j}^{(k)}$ y $\hat{\boldsymbol{\beta}}^{(k)}$ los vectores de coeficientes estimados en el paso (k), tanto del modelo reducido como del modelo completo, respectivamente. Se verifica que, cuando la

hipótesis nula en (2.8) es cierta, el estadístico definido en (2.10) converge en distribución a una $\chi^2_{(1)}$ para todo j, con $1 \leq j \leq p-1$.

2.3. Técnicas de validación del modelo

Como se ha explicado al inicio de este capítulo, los modelos de puntuación crediticia se utilizan como herramientas de clasificación y de predicción del grado de solvencia (o comportamiento de pago) de los clientes de las entidades financieras. Por ello, es fundamental que dichas entidades dispongan de técnicas cuantitativas que permitan garantizar que los modelos de riesgo de crédito que utilizan poseen un buen nivel predictivo.

Las técnicas de validación estadística se utilizan precisamente con ese propósito. Ellas permiten medir y comparar la bondad del ajuste, la exactitud de las predicciones y el poder discriminante del modelo.

En la literatura dedicada al análisis del riesgo de crédito, existen trabajos que se destacan por su influencia en las metodologías adoptadas por organismos supervisores, por agencias de calificación internacionales[2] y por las propias entidades financieras. Tal es el caso de los trabajos debidos a Sobehart et al. (2000), Hand (2001), Sobehart y Keenan (2001), Sobehart et al. (2001), Thomas et al. (2002), Engelman et al. (2003a, 2003b), Basel Committee on Banking Supervision (2005a), Stein (2005), Engelmann y Rauhmeier (2006), Rauhmeier (2006), Tasche (2006), Satchell y Xia (2007), Stein (2007) y Sun y Wang (2007), entre otros. En el presente capítulo se estudian las principales medidas de validación estadística propuestas por los autores mencionados:

- El contraste de *Hosmer-Lemeshow*
- Los criterios de información de *Akaike* y de *Schwarz*
- Los coeficientes pseudo-R^2 de *Cox-Snell* y de *Nagelkerke*
- El análisis de curvas *ROC* y de curvas *CAP*

[2] La agencia de calificación Moody's publica periódicamente una serie de documentos de trabajo llamada *Rating Methodology*. En esta publicación la agencia Moody's da a conocer los avances relacionados con las metodologías de validación que utiliza en sus modelos de rating. Uno de los trabajos más destacados sobre este tema es, por ejemplo, el debido a Sobehart et al. (2000).

- La distancia de *Kolmogorov-Smirnov*
- El análisis de los errores de clasificación de *tipo I* y de *tipo II*

2.3.1. Validación vía contrastes de bondad del ajuste

Una de las técnicas más utilizadas por las entidades financieras en la validación de sus modelos de puntuación crediticia es la que se realiza mediante contrastes de bondad del ajuste. Se verifica que el contraste Hosmer y Lemeshow (HL) es uno de los métodos de validación más aceptados por las entidades de crédito (Blochwitz et al. (2006), Rauhmeier (2006)).

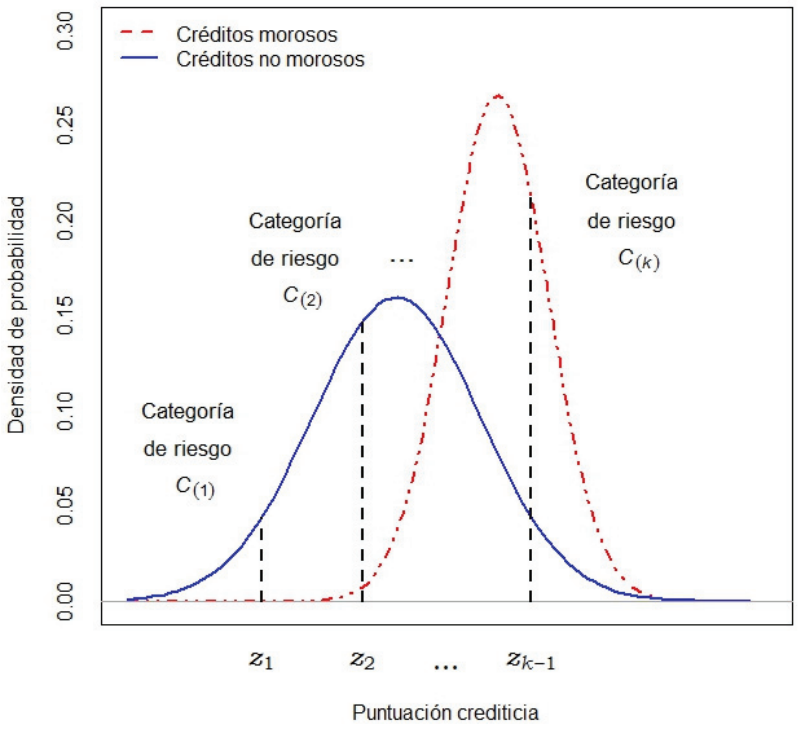

Figura 2.1 Categorías de riesgo para la puntuación de los créditos morosos y no morosos.

En la Figura 2.1 se ilustra el comportamiento de dos poblaciones de créditos, los *morosos* y los *no morosos*, a partir de las funciones de densidad de sus puntuaciones crediticias estimadas no paramétricamente. Tamibén se ilustra cómo se construyen los grupos de riesgo a partir de la categorización de las puntuaciones. A continuación, se verá cómo utilizar este mecanismo de categorización en el contraste de Hosmer-Lemeshow.

Contraste de *Hosmer-Lemeshow*

La prueba de Hosmer y Lemeshow (1980) contrasta la hipótesis nula de que el número observado de créditos en cada grupo (morosos y no morosos) es igual al número esperado de créditos (morosos y no morosos) estimado con el modelo. Este método agrupa a los acreditados en distintas *categorías o clases de riesgo* según la puntuación crediticia estimada con el modelo.

Dada una cartera de n créditos personales que verifica la situación representada en la Figura 2.1, la puntuación de la cartera se divide en k categorías de riesgo ordenadas, $C_{(1)}, C_{(2)}, ..., C_{(k)}$, con tamaños muestrales $n_1, n_2, \ldots, n_k$, donde $\sum_{j=1}^{k} n_j = n$. Para el cálculo del estadístico del contraste HL se construyen dos tablas de frecuencias en la cuales los créditos se agrupan según la situación de mora y la categoría de riesgo a la que pertenecen.

Tabla 2.2. Puntuaciones estimadas agrupadas por clases de riesgo

Clase de riesgo		$(-\infty, z_1]$	$(z_1, z_2]$	$\cdots$	$(z_{k-1}, +\infty)$
Tipo de crédito	Moroso	$\sum_{i=1}^{n_1} \hat{\pi}_{i1}$	$\sum_{i=1}^{n_2} \hat{\pi}_{i2}$	$\cdots$	$\sum_{i=1}^{n_k} \hat{\pi}_{ik}$
	No moroso	$n_1 - \sum_{i=1}^{n_1} \hat{\pi}_{i1}$	$n_2 - \sum_{i=1}^{n_2} \hat{\pi}_{i2}$	$\cdots$	$n_k - \sum_{i=1}^{n_k} \hat{\pi}_{ik}$

La Tabla 2.2 muestra, en cada celda, el número esperado de créditos morosos y no morosos en la categoría de riesgo $C_{(j)}$, con $1 \leq j \leq k$, a partir de las puntuaciones obtenidas con el modelo (tabla de valores esperados).

Tabla 2.3. Frecuencias observadas de morosidad agrupadas por clases de riesgo

Clase de riesgo		$(-\infty, z_1]$	$(z_1, z_2]$	$\cdots$	$(z_{k-1}, +\infty)$
Tipo de crédito	Moroso	$\sum_{i=1}^{n_1} y_{i1}$	$\sum_{i=1}^{n_2} y_{i2}$	$\cdots$	$\sum_{i=1}^{n_k} y_{ik}$
	No moroso	$n_1 - \sum_{i=1}^{n_1} y_{i1}$	$n_2 - \sum_{i=1}^{n_2} y_{i2}$	$\cdots$	$n_k - \sum_{i=1}^{n_k} y_{ik}$

La Tabla 2.3 muestra, en cada celda, las frecuencias observadas de créditos morosos y no morosos en la categoría de riesgo $C_{(j)}$, con $1 \leq j \leq k$ (tabla de valores observados).

El estadístico de *Hosmer-Lemeshow*, denotado por $\hat{T}_{HL}$, viene dado por:

$$\hat{T}_{HL} = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j \left(1 - \overline{\pi}_{\cdot j}\right)},$$

donde:

n_j es el tamaño de la categoría de riesgo $C_{(j)}$.

$O_j = \sum_{i=1}^{n_j} y_{ij}$ es el número observado de créditos morosos en la categoría de riesgo $C_{(j)}$.

$\overline{\pi}_{\cdot j} = \frac{1}{n_j} \sum_{i=1}^{n_j} \hat{\pi}_{ij}$ es la probabilidad de mora promedio estimada para los créditos pertenecientes a la categoría de riesgo $C_{(j)}$.

$E_j = n_j \overline{\pi}_{\cdot j}$ es el número esperado de créditos morosos en la categoría de riesgo $C_{(j)}$.

Hosmer y Lemeshow (1980) demostraron en un exhaustivo estudio de simulación que cuando la hipótesis nula es cierta y los tamaños muestrales n_j son suficientemente grandes para todo j, el estadístico $\hat{T}_{HL}$ converge en distribución a una $\chi^2_{(k-2)}$.

El *p-valor* obtenido con el contraste HL se utiliza como medida de bondad del ajuste del modelo de puntuación crediticia estimado. Si se obtiene un *p-valor* grande, dependiendo del nivel de significación escogido, entonces se considera que el modelo ajustado es adecuado.

Criterio de información de *Akaike*

El criterio de información de Akaike (Akaike (1974)) fue desarrollado como medida de discrepancia de propósito general en la modelización estadística. Este criterio sirve para cuantificar la distancia en información entre el modelo teórico y el modelo estimado y por ello es una medida apropiada para discriminar entre los distintos modelos de puntuación crediticia que se estudian en este capítulo.

El *estadístico de Akaike* (AIC) viene dado por:

$$AIC = -2\ell(\hat{\boldsymbol{\beta}}) + 2p,$$

donde $\ell(\hat{\boldsymbol{\beta}})$ se obtiene a partir de (2.5) y el término $2p$ es un valor de penalización debido a la parametrización del modelo que depende del número, p, de parámetros estimados.

La estrategia de selección basada en este criterio sostiene que se debe elegir aquel modelo de puntuación crediticia cuyo valor de AIC es el menor.

Criterio de información de *Schwarz*

El criterio de información de Schwarz (Schwarz (1978)) es una alternativa al criterio de Akaike y está construido sobre la base de argumentos de corte bayesiano, por este motivo este estadístico también recibe el nombre de *criterio bayesiano de información*. Debido a que las propiedades de consistencia del estadístico AIC han sido criticadas por no tomar en cuenta el efecto del tamaño muestral, n, en su construcción, el criterio BIC ha mostrado ser una alternativa apropiada cuando el tamaño muestral incide negativamente sobre la dimensión del modelo, es decir, cuando la complejidad de éste aumenta con n.

El *estadístico de Schwarz* (BIC) viene dado por:

$$BIC = -2\ell(\hat{\boldsymbol{\beta}}) + \ln(n)p,$$

donde el término $\ln(n)p$ se introduce para penalizar el efecto que provocan el tamaño muestral y el número de parámetros incluidos en el modelo. En general, si la complejidad del modelo estimado no aumenta de manera considerable con el tamaño muestral, basta con emplear el estadístico AIC para

elegir el mejor modelo, de lo contrario, es preferible emplear el estadístico BIC (Burnham y Anderson, 1998).

El criterio de selección de Schwarz es semejante al criterio de Akaike. Según este criterio se debe elegir aquel modelo cuyo valor de BIC es el menor.

Coeficientes de bondad del ajuste del tipo pseudo$-R^2$

El coeficiente de determinación lineal *R-cuadrado* (R^2) es la medida de bondad del ajuste por excelencia cuando se estudia el ajuste lineal en un modelo de regresión. En el contexto del *riesgo de crédito*, el coeficiente R^2 ha sido utilizado, por ejemplo, para evaluar la calidad del ajuste de modelos basados en el enfoque lineal multivariante de Altman (1968).

Por otra parte, es conocido en estadística que la aplicación del coeficiente R^2 puede no ser adecuado cuando se estudia el ajuste de modelos de regresión no lineales ya que, al no tener en cuenta el número de covariables en el modelo, en ocasiones, los valores que toma pueden ser de difícil interpretación. Debido a esto, existen medidas de bondad del ajuste que permiten corregir este problema, se conocen como *coeficientes pseudo*$-R^2$ y se utilizan comúnmente en el ajuste de modelos lineales generalizados. En adelante, por simplicidad de escritura, se utilizará la notación $\tilde{R}^2$ para hacer referencia a una medida o coeficiente pseudo$-R^2$.

En esta memoria se estudiarán dos de las medidas $\tilde{R}^2$ más utilizadas en la validación del ajuste de modelos de regresión logística, el coeficiente $\tilde{R}^2$ de *Cox-Snell* y el coeficiente $\tilde{R}^2$ de *Nagelkerke*. Las propiedades de estas medidas de bondad del ajuste han sido estudiadas, entre otros, por McFadden (1974), Amemiya (1981), Maddala (1983), Magee (1990), Laitila (1993), Veall y Zimmermann (1994), Windmeijer (1995), Veall y Zimmermann (1996) y Cameron y Windmeijer (1997).

Coeficiente $\tilde{R}^2$ de *Cox-Snell* Este coeficiente pertenece a la clase de medidas de bondad del ajuste construidas a partir del estadístico de razón de verosimilitudes. El coeficiente de *Cox-Snell*, denotado por $\tilde{R}^2_{CS}$, se define

como:

$$\tilde{R}^2_{CS} = 1 - \left(\frac{L(\hat{\beta}_0)}{L\left(\hat{\boldsymbol{\beta}}\right)} \right)^{\frac{2}{n}} = 1 - \exp\left(-\frac{\hat{T}_{RV_0}}{n} \right),$$

donde $\hat{T}_{RV_0} = -2\left(\ell(\hat{\beta}_0) - \ell(\hat{\boldsymbol{\beta}})\right)$ es el estadístico definido en (2.10) en el que se ha reemplazado la log-verosimilitud del modelo reducido, $\ell(\hat{\boldsymbol{\beta}}^{(k)}_{-j})$, por la log-verosimilitud del modelo sin covariables o modelo nulo, $\ell(\hat{\beta}_0)$. Debido a que este coeficiente verifica la desigualdad $0 < \tilde{R}^2_{CS} < 1$, el criterio de *Cox-Snell* establece que cuanto más próximo a 1 es el valor del estadístico $\tilde{R}^2_{CS}$, mejor es el ajuste del modelo. En tal sentido, al comparar entre dos, o más, modelos ajustados, es preferible aquel que exhibe el valor más alto de $\tilde{R}^2_{CS}$.

Coeficiente $\tilde{R}^2$ de *Nagelkerke* El coeficiente $\tilde{R}^2$ de *Nagelkerke* es una versión ajustada del coeficiente $\tilde{R}^2$ de *Cox-Snell*. Debido a que el coeficiente de *Cox-Snell* no alcanza el valor 1 para valores no nulos del cociente de verosimilitudes, $\frac{L(\hat{\beta}_0)}{L\left(\hat{\boldsymbol{\beta}}\right)}$, el coeficiente de *Nagelkerke* corrige este problema reescalando el valor de $\tilde{R}^2_{CS}$ por la cantidad $1 - \left(L(\hat{\beta}_0)\right)^{\frac{2}{n}}$, es decir, dividiendo por el máximo valor que alcanza el estadístico $\tilde{R}^2_{CS}$ cuando $L\left(\hat{\boldsymbol{\beta}}\right)$ toma el valor 1. Así, el coeficiente de *Nagelkerke*, denotado por $\tilde{R}^2_{Ng}$, viene dado por:

$$\tilde{R}^2_{Ng} = \frac{1 - \left(\frac{L(\hat{\beta}_0)}{L\left(\hat{\boldsymbol{\beta}}\right)} \right)^{\frac{2}{n}}}{1 - \left(L(\hat{\beta}_0)\right)^{\frac{2}{n}}} = \frac{\tilde{R}^2_{cs}}{1 - \left(L(\hat{\beta}_0)\right)^{\frac{2}{n}}}.$$

El criterio de bondad del ajuste de *Nagelkerke* funciona de forma análoga al criterio de *Cox-Snell*, es decir, cuanto más próximo a 1 es el valor de $\tilde{R}^2_{Ng}$, mejor es el ajuste del modelo. Así, entre dos, o más, modelos ajustados, es preferible aquel con el valor más alto de $\tilde{R}^2_{Ng}$.

2.3.2. Validación vía análisis de curvas *ROC*

Las curvas *ROC* (por su sigla en inglés de *receiver operating characteristic*) han sido tradicionalmente utilizadas como herramientas de análisis visual para comparar el poder discriminante de modelos de clasificación. Esta metodología, que surge en el contexto de la teoría del reconocimiento de señales (Egan (1975)), ha mostrado ser una técnica eficaz cuando se la utiliza para comparar diferentes mecanismos de clasificación. Una de las ventajas que ofrecen las curvas *ROC* como medidas de validación es la flexibilidad de su construcción, ya que no dependen de supuestos sobre el modelo probabilístico que subyace en la regla discriminante. La aplicación del análisis de curvas *ROC* a la validación de modelos discriminantes ha sido extensamente estudiado en diversas áreas del conocimiento. Ver, por ejemplo, los trabajos de Swets (1988), Hanley (1989), Pepe (2002), Zhou et al. (2002), Pepe (2003), Fawcett (2006) y Krzanowski y Hand (2009), entre otros.

En el contexto del riesgo de crédito, el análisis de curvas *ROC* fue dado a conocer en los trabajos de Sobehart et al. (2000) y de Sobehart y Keenan (2001), donde se explica la sencillez de la construcción de una curva *ROC* cuando se dispone de dos muestras de puntuaciones crediticias, una caracterizando los *créditos morosos* y la otra los *créditos no morosos*. Otros autores que han estudiado la aplicación de curvas *ROC* en el contexto de validación de modelos de riesgo de crédito son Engelmann et al. (2003a), Engelmann y Rauhmeier (2006), Tasche (2006), Satchell y Xia (2007), Stein (2007) y Hong (2009), entre otros.

Definición de la curva *ROC* y su capacidad predictiva

En una cartera de n clientes en la que sólo existen dos clases de créditos, los que provienen de la población de créditos morosos, denotada por Π_1 y los que provienen de la población de créditos no morosos, denotada por Π_0, se define la variable indicadora de la morosidad del crédito i-ésimo como:

$$Y_i = \begin{cases} 1 & \text{si} \quad \text{el crédito } i \in \Pi_1 \\ 0 & \text{si} \quad \text{el crédito } i \in \Pi_0 \end{cases}, \tag{2.11}$$

para todo i con $1 \leq i \leq n$. Así, dado un nuevo cliente (o solicitante) de un crédito, la entidad debe clasificarlo como bueno o malo comparando su

puntuación estimada, $\hat{Z}$, con un punto de corte llamado umbral de clasificación, denotado por z^*. Si $\hat{Z} > z^*$, entonces el crédito es asignado a Π_1, en caso contrario, es asignado a Π_0. Este mecanismo de clasificación, se formula definiendo la variable aleatoria

$$\hat{Y}|\boldsymbol{X} = \begin{cases} 1 & \text{si} \quad \hat{Z} > z^* \\ 0 & \text{si} \quad \hat{Z} \leq z^* \end{cases}, \tag{2.12}$$

que es un predictor de la variable definida en (2.2), donde $\hat{Z} = \hat{\psi}(\boldsymbol{X})$, como se definió previamente en el apartado 2.2.2. Además, este mecanismo de clasificación puede verse como el estadístico de contraste de un test de hipótesis sobre la condición de morosidad de un crédito, donde las hipótesis nula, H_0, y alternativa, H_1, se definen como:

$$\begin{array}{ll} H_0: & \text{el crédito es moroso, o bien, } Z > z^* \\ H_1: & \text{el crédito es no moroso, o bien, } Z \leq z^*. \end{array} \tag{2.13}$$

Como consecuencia de la regla definida en (2.12), se obtienen dos tipos de probabilidades de error de clasificación que dependen del punto de corte z^*. El primero corresponde a la probabilidad de clasificar como *no morosos* los créditos pertenecientes a Π_1. Esta probabilidad recibe el nombre de $P(\textit{error de tipo } I)$ y se define por:

$$P(\textit{error de tipo } I) = P(Z \leq z^*|\Pi_1) = P(Z \leq z^*|Y = 1). \tag{2.14}$$

El segundo tipo corresponde a la probabilidad de clasificar como *morosos* los créditos pertenecientes a Π_0. Esta probabilidad recibe el nombre de $P(\textit{error de tipo } II)$ y se define por:

$$P(\textit{error de tipo } II) = P(Z > z^*|\Pi_0) = P(Z > z^*|Y = 0). \tag{2.15}$$

Como la puntuación z^* toma valores en todo $\mathbb{R}$, el gráfico de la curva *ROC* representa todos los posibles resultados de la combinación de errores obtenidos a partir del mecanismo de clasificación definido en (2.12). Así, a partir de (2.14) y (2.15), la curva *ROC* teórica se define como el conjunto

$$\{(P(Z > z^*|\Pi_0), 1 - P(Z \leq z^*|\Pi_1)) \diagup z^* \in \mathbb{R}\},$$

o, equivalentemente, como el conjunto

$$\{(P(Z > z^*|Y = 0), P(Z > z^*|Y = 1)) \diagup z^* \in \mathbb{R}\}. \tag{2.16}$$

Por tanto, el gráfico de la curva *ROC* representa el balance entre la capacidad e incapacidad de clasificar correctamente que posee un modelo discriminante, como es el caso de los modelos de puntuación crediticia.

Finalmente, aunque la definición dada en (2.16) es la utilizada con más frecuencia en la literatura (ver Sobehart y Keenan (2001), Engelmann et al. (2003a), Pepe (2002, 2003), Fawcett (2006), Stein (2007) y Krzanowski y Hand (2009), entre otros), cabe mencionar que existen trabajos en los que la curva *ROC* es definida de forma diferente. Ver, por ejemplo, los trabajos de Greiner et al. (2000) y Metz (2006).

El espacio *ROC*

El dominio de representación de la curva *ROC* se denomina *espacio ROC* y está delimitado geométricamente por los puntos $(0,0)$, $(0,1)$, $(1,1)$ y $(1,0)$ del plano cartesiano.

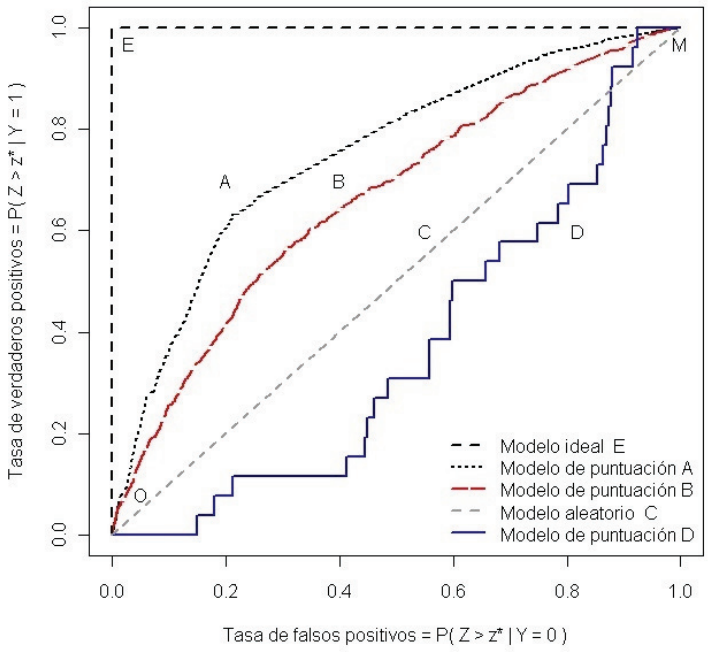

Figura 2.2 Curvas *ROC* asociadas a modelos con distinta capacidad discriminante.

En la Figura 2.2 se representa el espacio ROC asociado a cinco estrategias de clasificación. Cada una de ellas posee un poder de discriminación diferente y están representadas por las curvas ROC A, B, C, D y E.

La estrategia cuya curva ROC corresponde a la unión de las rectas OE y EM (línea segmentada en color negro) representa un *modelo de clasificación ideal*, es decir, un modelo que predice correctamente el grupo de pertenencia del 100 % de los créditos. Informalmente, se dice que un modelo de puntuación tiene más capacidad predictiva si los puntos pertencientes al conjunto definido en (2.16) forman una curva cóncava dibujada por encima de la diagonal OM (línea segmentada en color gris) y ésta pasa por algún punto próximo al punto $E = (0, 1)$. Las curvas ROC asociadas a los modelos A y B tienen formas cóncavas y están por encima de la diagonal OM. Esto permite suponer que ambos modelos representan estrategias de clasificación válidas para distinguir entre créditos morosos y no morosos, siendo el modelo A un clasificador más potente que el modelo B. A diferencia de lo anterior, el modelo C tiene asociada una curva ROC con todos sus puntos contenidos en la diagonal OM. Un gráfico ROC como éste representa una estrategia de clasificación en ausencia de información sobre los créditos, discriminando al azar si éstos son morosos o no. Debido a esto, el clasificador C recibe el nombre de *modelo de clasificación aleatorio*. Por último, el modelo de puntuación D tiene asociada una curva ROC cuyos puntos están contenidos en el interior del triángulo formado por los puntos $(0, 0)$, $(1, 0)$ y $(1, 1)$. Esta región del espacio ROC está compuesta por todas aquellas estrategias de clasificación que son incapaces de distinguir entre créditos morosos y no morosos, y por tanto, se dice que el modelo D carece de poder predictivo. Ante una situación como esta, se debe replantear la construcción del modelo de puntuación crediticia.

Medidas de precisión obtenidas de la curva ROC

Las medidas de precisión son índices que permiten evaluar la capacidad que posee un modelo de clasificación para predecir correctamente la población a la que pertenece un sujeto. Las situaciones que pueden darse en un problema de clasificación crediticia son cuatro:

- El crédito es *moroso* y el modelo lo clasifica como *moroso*. Esta clasificación del crédito corresponde a un acierto y se contabiliza como un *verdadero positivo* (VP).

- El crédito es *moroso* y el modelo lo clasifica como *no moroso*. Esta clasificación del crédito es errónea y se contabiliza como un *falso negativo* (FN). Este error recibe el nombre de *error de tipo* I, o bien, *error* α, nombre que toma la decisión de rechazar H_0 cuando ésta es verdadera en un contraste de hipótesis como el definido en (2.13).

- El crédito es *no moroso* y el modelo lo clasifica como *no moroso*. Esta clasificación del crédito corresponde a un acierto y se contabiliza como un *verdadero negativo* (VN).

- El crédito es *no moroso* y el modelo lo clasifica como *moroso*. Esta clasificación del crédito es errónea y se contabiliza como un *falso positivo* (FP). Este error recibe el nombre de *error de tipo* II, o bien, *error* β, nombre que toma la decisión de aceptar H_0 cuando ésta es falsa en un contraste de hipótesis como el definido en (2.13).

Tabla 2.4. Matriz de clasificación de los créditos según el punto de corte z^*

Modelo	Calidad crediticia	Moroso	No moroso
Puntuación	$Z > z^*$ (Mala calidad)	VP (Acierto)	FP (Error de tipo II)
	$Z \leq z^*$ (Buena calidad)	FN (Error de tipo I)	VN (Acierto)

La Tabla 2.4 muestra la matriz de clasificación de los créditos construida a partir de la regla definida en (2.12). Las cantidades VP, FN, VN y FP denotan las frecuencias conjuntas de la Tabla 2.4, donde VP es el número de verdaderos positivos, FN es el número de falsos negativos, VN es el número de verdaderos negativos y FP es el número de falsos positivos.

Dado el valor de corte, z^*, y una muestra aleatoria de n créditos con puntuaciones, $Z_1, ..., Z_n$, donde n_1 de ellos provienen de la población Π_1 y

$n_0 = n - n_1$ provienen de la población Π_0, se definen las siguientes medidas de precisión en la clasificación de los créditos:

Tasa de verdaderos positivos La tasa de verdaderos positivos (TVP) se define como la probabilidad condicional de clasificar como moroso un crédito proveniente de Π_1. Esta probabilidad mide la proporción de veces que el modelo clasifica correctamente los créditos morosos y viene dada por:

$$TVP\left(z^*\right) = P(Z > z^* | Y = 1).$$

El estimador empírico usual de $TVP\left(z^*\right)$ viene dado por:

$$\widehat{TVP}\left(z^*\right) = \frac{\sum_{i=1}^{n} I\left(Z_i > z^*, Y_i = 1\right)}{\sum_{i=1}^{n} I\left(Y_i = 1\right)} = \frac{VP}{n_1}.$$

Tasa de falsos negativos La tasa de falsos negativos (TFN) se define como la probabilidad condicional de clasificar como no moroso un crédito proveniente de Π_1. Esta probabilidad es equivalente a la definida en (2.14) y se utiliza como medida del *riesgo de crédito* asociado a los créditos que siendo morosos no son detectados por el modelo. La proporción $TFN\left(z^*\right)$ viene dada por:

$$TFN\left(z^*\right) = P\left(\textit{error de tipo } I\right) = P(Z \leq z^* | Y = 1),$$

y se calcula empíricamente mediante la fórmula:

$$\widehat{TFN}\left(z^*\right) = \frac{\sum_{i=1}^{n} I\left(Z_i \leq z^*, Y_i = 1\right)}{\sum_{i=1}^{n} I\left(Y_i = 1\right)} = 1 - \widehat{TVP}\left(z^*\right).$$

Tasa de verdaderos negativos La tasa de verdaderos negativos (TVN) se define como la probabilidad condicional de clasificar como no moroso un crédito proveniente de Π_0. Esta probabilidad mide la proporción de veces que el modelo clasifica correctamente los créditos no morosos y viene dada por:

$$TVN\left(z^*\right) = P(Z \leq z^* | Y = 0).$$

El estimador empírico usual de $TVN\left(z^*\right)$ viene dado por la fórmula:

$$\widehat{TVN}\left(z^*\right) = \frac{\sum_{i=1}^{n} I\left(Z_i \leq z^*, Y_i = 0\right)}{\sum_{i=1}^{n} I\left(Y_i = 0\right)} = \frac{VN}{n_0}.$$

Tasa de falsos positivos La tasa de falsos positivos (TFP) se define como la probabilidad condicional de clasificar como moroso un crédito proveniente de Π_0. Esta probabilidad es equivalente a la definida en (2.15) y corresponde a una medida del riesgo asociado al *coste de oportunidad* debido a los créditos buenos mal clasificados por el modelo. La proporción $TFP\,(z^*)$ viene dada por:

$$TFP\,(z^*) = P\,(\textit{error de tipo } II) = P(Z > z^*|Y = 0),$$

y se calcula empíricamente mediante la fórmula:

$$\widehat{TFP}\,(z^*) = \frac{\sum_{i=1}^{n} I\,(Z_i > z^*, Y_i = 0)}{\sum_{i=1}^{n} I\,(Y_i = 0)} = 1 - \widehat{TVN}\,(z^*)\,.$$

Tasa de precisión La tasa de precisión (TP) corresponde a la probabilidad de que un crédito sea correctamente clasificado por el modelo. La tasa de precisión, TP, se calcula empíricamente mediante la fórmula:

$$\widehat{TP} = \frac{\sum_{i=1}^{n} I\,(Z_i{>}z^*,Y_i{=}1) + \sum_{i=1}^{n} I\,(Z_i \leq z^*,Y_i{=}0)}{n_1 + n_0} = \frac{VP + VN}{n}.$$

Área bajo la curva *ROC* El área bajo la curva *ROC* (AUC) se define como el área delimitada por el gráfico de la curva *ROC* y por las rectas $TFP = 1$ y $TVP = 0$.

Denotando por $\bar{G}_j(z) = 1 - G_j(z) = P(Z > z|Y = j) = P(Z_j > z)$, la función de supervivencia condicional de la variable puntuación crediticia, Z, dada la variable indicadora de morosidad, $Y = j$, con $j \in \{0, 1\}$ (definida en (2.11)), se verifica que el área bajo la curva *ROC*, AUC, viene dada por:

$$AUC = \int_0^1 R\,(p)\,dp = P(Z_1 > Z_0), \tag{2.17}$$

donde se ha definido la función $R(p) = 1 - G_1\left(G_0^{-1}\,(1-p)\right)$ con $p \equiv p\,(z^*) = 1 - G_1(z^*)$ variando entre 0 y 1 a medida que z^* varía entre $-\infty$ e $+\infty$.

Es conocido en la literatura que el estimador empírico usual de AUC se obtiene a partir del estadístico del contraste no paramétrico de *Mann-Whitney* para dos muestras independientes. La demostración de este resultado puede verse en Hanley y McNeil (1982).

Se define la variable indicadora, ψ_{ij}, por medio de la regla

$$\psi_{ij} = \begin{cases} 1 & \text{si} \quad Z_{1i} > Z_{0j} \\ 0 & \text{si} \quad Z_{1i} \leq Z_{0j} \end{cases} ,$$

donde Z_{1i}, con $1 \leq i \leq n_1$, es la variable puntuación del i-ésimo crédito moroso y Z_{0j}, con $1 \leq j \leq n_0$, es la variable puntuación del j-ésimo crédito no moroso, entonces, el estimador no paramétrico de AUC viene dado por:

$$\widehat{AUC} = \frac{1}{n_1 n_0} \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} \psi_{ij}. \tag{2.18}$$

Bajo los supuestos de independencia y continuidad de las puntuaciones Z_{1i} y Z_{0j}, se demuestra que el estimador definido en (2.18) es insesgado y que su esperanza coincide con el valor de la probabilidad definida en (2.17).

En relación con la varianza del estadístico $\widehat{AUC}$, Hanley y McNeil (1982) proponen su cálculo a partir de la fórmula de aproximación dada por:

$$\begin{aligned} Var\left(\widehat{AUC}\right) &= \frac{1}{n_1 n_0} \left(AUC\left(1 - AUC\right) + \left(n_1 - 1\right)\left(P_1 - AUC^2\right)\right. \\ &\qquad \left. + \left(n_0 - 1\right)\left(P_2 - AUC^2\right)\right) , \end{aligned}$$

donde $P_1 = P(Z_{11} > Z_{01}, Z_{12} > Z_{01})$ y $P_2 = P(Z_{11} > Z_{01}, Z_{11} > Z_{02})$ siendo (Z_{11}, Z_{12}) y (Z_{01}, Z_{02}) las puntuaciones correspondientes a dos parejas de acreditados tomados al azar de las poblaciones Π_1 y Π_0, respectivamente.

Debido a que el cálculo de $Var\left(\widehat{AUC}\right)$ sigue siendo un problema abierto en estadística, otros autores proponen fórmulas alternativas a la obtenida por Hanley y McNeil (1982). Ver, por ejemplo, los resultados obtenidos por Bamber (1975), DeLong et al. (1988) y Hanley y Hajian-Tilaki (1997), entre otros.

Contrastes de hipótesis para el AUC Las curvas ROC permiten comparar distintos mecanismos de clasificación de forma simultánea y sencilla, vía inspección visual. Sin embargo, cuando se analizan dos o más estrategias en las que no es posible distinguir visualmente (y tampoco a partir de sus valores de AUC) qué modelo domina en capacidad de clasificación sobre los demás, es necesario contrastar estadísticamente si ellos poseen o no la misma

capacidad predictiva. Para resolver este problema, Hanley y McNeil (1982) proponen el siguiente contraste de hipótesis:

$$\begin{aligned} H_0 &: AUC_1 = AUC_2 \\ H_1 &: AUC_1 \neq AUC_2. \end{aligned} \tag{2.19}$$

El estadístico del contraste viene dado por:

$$\hat{T}_{AUC} = \frac{\widehat{AUC_1} - \widehat{AUC_2}}{\sqrt{Var\left(\widehat{AUC_1} - \widehat{AUC_2}\right)}},$$

donde

$$\begin{aligned} Var\left(\widehat{AUC_1} - \widehat{AUC_2}\right) &= Var\left(\widehat{AUC_1}\right) + Var\left(\widehat{AUC_2}\right) \\ &\quad - 2\rho_{12}\sqrt{Var\left(\widehat{AUC_1}\right) Var\left(\widehat{AUC_2}\right)} \end{aligned}$$

y ρ_{12} es la correlación entre los estimadores $\widehat{AUC_1}$ y $\widehat{AUC_2}$. Uno de los métodos que se utilizan habitualmente para el cálculo del coeficiente ρ_{12} es el debido a Hanley y McNeil (1983).

Bajo la hipótesis nula definida en (2.19), se demuestra que la distribución del estadístico $\hat{T}_{AUC}$ converge en distribución a una normal estándar. Así, si el valor absoluto observado de $\hat{T}_{AUC}$ es mayor que el percentil $1 - \frac{\alpha}{2}$ de la distribución normal estándar, se rechaza la hipótesis de igualdad entre AUC_1 y AUC_2, o en otras palabras, se rechaza la hipótesis de que ambos modelos de puntuación poseen la misma capacidad predictiva. Este contraste de hipótesis también ha sido estudiado, entre otros, por McNeil y Hanley (1984) y por Hajian-Tilaki (1997).

2.3.3. Validación vía análisis de curvas CAP

La curva del perfil de precisión acumulativa, o curva CAP (por sus siglas en inglés de *cumulative accuracy profile*), es una técnica de análisis visual que permite evaluar la capacidad de un modelo para clasificar correctamente los créditos morosos en una muestra compuesta por distintas clases de créditos. El análisis de curvas CAP fue introducido en el contexto de los modelos de

puntuación crediticia por la agencia de calificacion Moody's (Sobehart et al. 2000). Posteriormente, esta técnica ha sido extensamente estudiada en trabajos sobre validación de modelos de riesgo de crédito. Ver, por ejemplo, Sobehart y Keenan (2001), Sobehart et al. (2001), Engelman et al. (2003a, 2003b), Basel Committee on Banking Supervision (2005a) y Engelmann (2006), entre otros.

Definición de la curva CAP y su capacidad predictiva

El gráfico de la curva CAP teórica se define como el conjunto de puntos de la forma

$$\{(P(Z > z^*), P(Z > z^*|Y = 1))/z^* \in \mathbb{R}\}, \tag{2.20}$$

donde la puntuación z^* es el valor de corte a partir del cual se determina la condición de morosidad del crédito.

En la Figura 2.3 se ilustra el concepto de curva CAP a partir de la representación de tres estrategias de clasificación de los créditos.

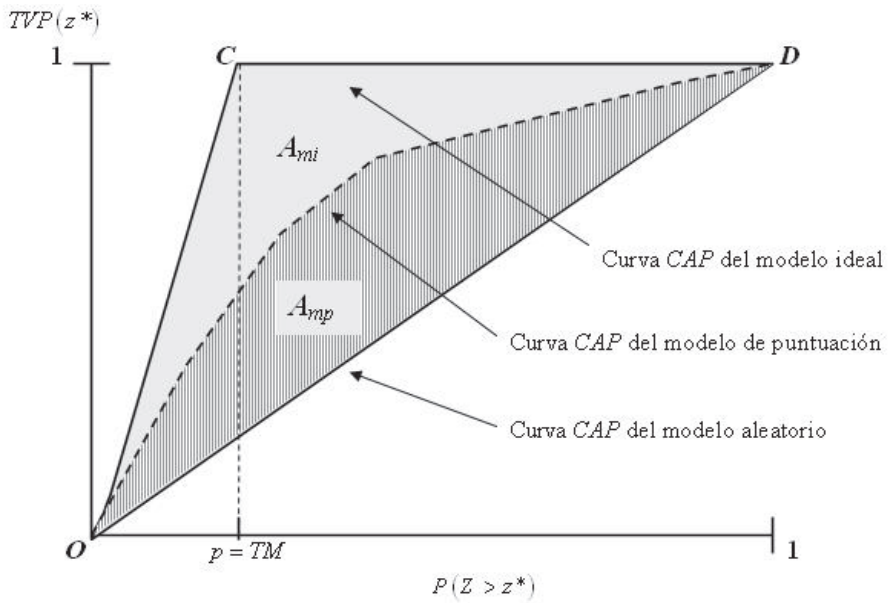

Figura 2.3 Curvas CAP asociadas a modelos con distinta capacidad discriminante.

La primera curva CAP, representada gráficamente por la línea que une los segmentos OC y CD, corresponde a un modelo de clasificación ideal, es decir, un modelo que permite identificar correctamente el 100 % de los créditos morosos a partir de sus puntuaciones. Como consecuencia, en una muestra de créditos con tasa de morosidad fija, denotada por TM, la totalidad de los créditos morosos se concentran a la izquierda del gráfico CAP, en este caso, a la izquierda de la abscisa dada por la ecuación $P(Z > z^*) = TM$.

Por otro lado, dado un valor de puntuación z^*, la curva CAP contenida en el espacio delimitado por el triángulo OCD (dibujada en línea segmentada) representa la variación del porcentaje de créditos morosos correctamente identificados por el modelo ($TVP(z^*)$) en función de la proporción de créditos cuya puntuación supera el valor $z^*(P(Z > z^*))$.

El tercer modelo de clasificación está representado por una curva CAP coincidente con la diagonal OD. A esta curva se la denomina curva CAP aleatoria. Análogamente a la curva ROC aleatoria, esta curva representa una estrategia de clasificación en ausencia de información sobre los créditos, y por tanto, su poder predictivo es equivalente al obtenido cuando dicha clasificación se realiza al azar.

En general, si el modelo de puntuación está bien calibrado, se espera que la proporción de créditos morosos concentrados a la izquierda del gráfico CAP sea próximo al $100TM$ % de las puntuaciones más altas de la muestra.

Coeficiente de precisión de la curva CAP En la Figura 2.3 se ilustra el poder predictivo de un modelo de puntuación obtenido a partir de la curva CAP. El coeficiente de precisión de dicho modelo, denotado por AR (por sus siglas en inglés de *accuracy ratio*), se define como el cociente entre su capacidad de clasificación, medida como el área encerrada entre su curva CAP y la curva CAP aleatoria, denotada por A_{mp}, y la capacidad de clasificación del modelo ideal, medida como el área encerrada entre la curva CAP ideal y la curva CAP aleatoria, denotada por A_{mi}. Así, el valor del coeficiente de precisión, AR, viene dado por:

$$AR = \frac{\text{área bajo la curva } CAP \text{ del modelo} - \frac{1}{2}}{\text{área bajo la curva } CAP \text{ ideal} - \frac{1}{2}} = \frac{A_{mp}}{A_{mi}}.$$

Al observar la Figura 2.3, se deduce que el gráfico de la curva CAP asociada a un modelo de clasificación está acotado por las curvas CAP ideal y aleatoria, y por tanto, la capacidad predictiva del modelo, medida en términos del área A_{mp}, varía entre 0 y A_{mi}. Como consecuencia, se obtiene que el coeficiente AR toma valores entre 0 y 1.

Debido a que las curvas CAP y ROC se construyen de forma similar, se sabe que existe una relación matemática entre sus medidas de precisión, AR y AUC, respectivamente. Esta propiedad puede formalizarse a partir de la siguiente proposición.

Proposición 2.1 *En un modelo de clasificación con variable respuesta continua, se verifica que el coeficiente de precisión, AR, y el área bajo la curva ROC, AUC, satisfacen la igualdad:*

$$AR = 2AUC - 1.$$

Demostración 2.1 Sean Z_0, Z_1 y Z_1' variables aleatorias independientes con distribuciones condicionadas $Z_0 \overset{d}{\equiv} Z|Y = 0$, $Z_1 \overset{d}{\equiv} Z|Y = 1$ y $Z_1' \overset{d}{\equiv} Z|Y = 1$. Denotando por S_{mp} al área bajo la curva CAP, por construcción se obtiene:

$$S_{mp} = A_{mp} + \frac{1}{2}.$$

Análogamente a lo visto para las curvas ROC se obtiene:

$$\begin{aligned}
S_{mp} &= P(Z < Z_1) = E(P(Z < Z_1|Z_1)) \\
&= E\left(P(Z < Z_1|Z_1, Y = 0)\right)(Y = 0) \\
&\qquad + E\left(P(Z < Z_1|Z_1, Y = 1)\right)P\left(Y = 1\right) \\
&= E\left(P(Z_0 < Z_1|Z_1, Y = 0)\right)(1 - TM) \\
&\qquad + E\left(P(Z_1' < Z_1|Z_1, Y = 1)\right)TM \\
&= (1 - TM)\, E\left(P(Z_0 < Z_1|Z_1)\right) + TM\, E\left(P(Z_1' < Z_1|Z_1)\right) \\
&= (1 - TM)\, P(Z_0 < Z_1) + TM\, P(Z_1' < Z_1) \\
&= (1 - TM)\, AUC + \frac{1}{2}\, TM.
\end{aligned}$$

Por tanto,

$$S_{mp} = A_{mp} + \frac{1}{2} = (1 - TM)\, AUC + \frac{1}{2}\, TM,$$

de donde se deduce que

$$A_{mp} = (1 - TM)\, AUC + \frac{1}{2} TM - \frac{1}{2}.$$

Por otro lado, por construcción se deduce que el área bajo la curva CAP del modelo ideal, denotada por S_{mi}, viene dada por:

$$S_{mi} = A_{mi} + \frac{1}{2} = 1 - \frac{TM}{2},$$

de donde se obtiene que

$$A_{mi} = \frac{1}{2}(1 - TM).$$

Por tanto,

$$AR = \frac{A_{mp}}{A_{mi}} = \frac{(1 - TM)\, AUC - \frac{1}{2}\,(1 - TM)}{\frac{1}{2}(1 - TM)} = 2\, AUC - 1.$$

□

2.3.4. Elección del punto de corte para la clasificación de los créditos

La elección del punto de corte para la clasificación de los créditos es un problema crucial tanto en la validación del modelo de scoring como en el proceso de concesión de los créditos. Por este motivo, es de gran importancia contar con una metodología que permita obtener el valor de corte óptimo para distinguir entre créditos de buena y mala calidad.

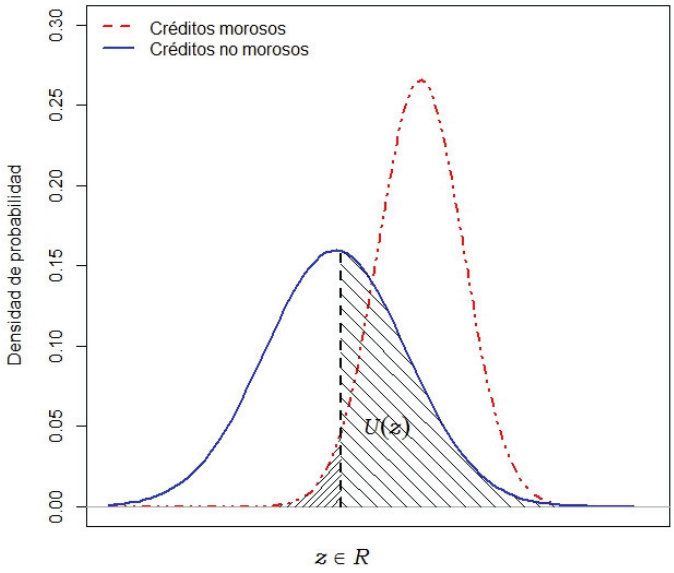

Figura 2.4 Región de solapamiento de las densidades de los créditos morosos y no morosos.

En la Figura 2.4, la superficie sombreada representa la región de solapamiento generada por la intersección de las funciones de densidad asociadas a las distribuciones G_1 (créditos morosos) y G_0 (créditos no morosos). El área de la región sombreada, definida por la función $U\,(z)$, con $z \in \mathbb{R}$, se obtiene a partir de (2.14) y (2.15) como:

$$\begin{aligned} U\,(z) &= P\,(Z \leq z|Y=1) + P\,(Z > z|Y=0) \\ &= G_1(z) + 1 - G_0(z). \end{aligned}$$

Es conocido en la literatura que el valor de corte óptimo para la clasificación de los créditos se obtiene a partir de:

$$z_U^* = \arg\min_{z\in\mathbb{R}} U(z). \tag{2.21}$$

Además, se verifica que si las distribuciones G_0 y G_1 poseen densidades cuyos soportes no se intersectan, entonces la función $U\,(z)$ toma el valor 0

para los valores z mayores o iguales que el extremo superior del soporte de G_0 y menores o iguales que el extremo inferior del soporte de G_1. Si ambas distribuciones poseen densidades idénticas en todo el soporte, entonces la función $U(z)$ toma el valor 1. En cualquier otro caso, la función $U(z)$ toma valores en el intervalo abierto $(0, 1)$.

En los trabajos de Kraft et al. (2003, 2004) y Clavero (2006) se estudia la solución del problema (2.21) como una medida del poder discriminante del modelo de puntuación.

En lugar del enfoque anterior, en este capítulo se utiliza, como valor de corte óptimo para la clasificación de los créditos, el valor obtenido con el estadístico del contraste de *Kolmogorov-Smirnov* para la igualdad de las distribuciones G_1 y G_0. Este método proporciona un valor óptimo equivalente a la solución obtenida en (2.21).

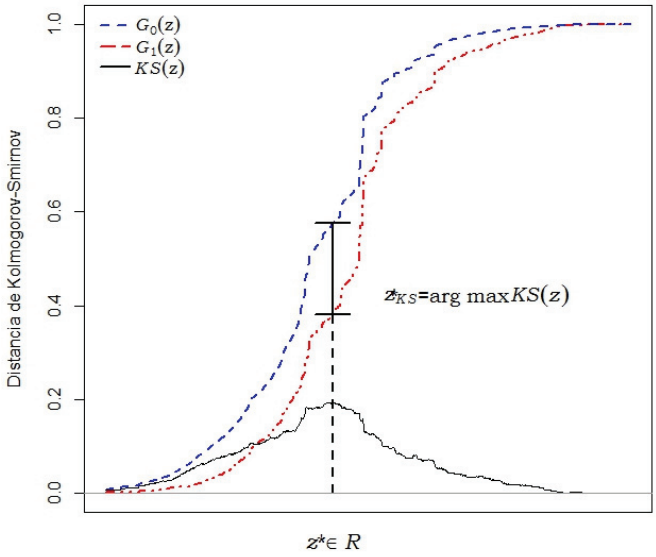

Figura 2.5 Distancia de Kolmogorov-Smirnov para un modelo de puntuación crediticia

La Figura 2.5 ilustra el método de *Kolmogorov-Smirnov*, $KS(z)$, para determinar el valor óptimo de puntuación, z^*_{KS}, que maximiza la distancia

entre las distribuciones $G_0(z)$ y $G_1(z)$. Además, el valor z^*_{KS} coincide con el valor que produce la máxima separabilidad entre las densidades ilustradas en la Figura 2.4.

El estadístico, o distancia, de *Kolmogorov-Smirnov*, $KS(z)$, se define como:

$$KS(z) = \max_{z \in \mathbb{R}} \left(G_0(z) - G_1(z) \right),$$

donde el valor de corte óptimo viene dado por:

$$z^*_{KS} = \arg \max_{z \in \mathbb{R}} KS(z).$$

La aplicación de este método en el contexto del riesgo de crédito ha sido estudiado, entre otros, por Appel (2002), Thomas et al. (2002), Basel Committee on Banking Supervision (2005a) y Hong (2009).

2.4. Aplicación a una cartera de tarjetas de crédito

En esta sección se estudia la construcción de un modelo de puntuación para los titulares de tarjetas de crédito de una entidad financiera española.

En la primera parte de esta sección se realiza un análisis descriptivo de los datos utilizados. Allí se describen las variables de las que se dispone y a partir de las cuales se ajustaron los modelos de puntuación crediticia. Posteriormente se realizó la validación de los mismos según las técnicas expuestas en la Sección 2.3.

Es importante explicar que, por motivos de confidencialidad, en esta memoria no se utilizaron directamente los datos entregados por la entidad colaboradora. Por tal motivo, se hicieron transformaciones y cambios de escala en algunas de las variables, además de alterar arbitrariamente la proporción de créditos morosos en la muestra.

2.4.1. Análisis de la base de datos

La base de datos utilizada en este capítulo se compone de dos muestras aleatorias tomadas de forma independiente. La primera muestra corresponde

a 25 000 registros de tarjetas de crédito que fueron formalizadas entre 2004 y 2007. Esta muestra contiene 970 créditos morosos y 24 030 créditos no morosos, lo que equivale a una tasa de mora global del 3.88 %. De los 25 000 registros, 20 000 se utilizaron como muestra de entrenamiento para ajustar los modelos y los 5 000 restantes se utilizaron en la validación de los mismos. La segunda muestra contiene 327 solicitudes de tarjetas de crédito que nunca fueron formalizadas, por lo que no se pudo observar la mora en estas solicitudes. De esta muestra sólo se conocen las puntuaciones calculadas por la entidad y el resultado del dictamen de concesión, que fue uno de los tres posibles: *concedido*, *dudoso* y *denegado*.

2.4.2. Las variables del modelo

Con el objetivo de estudiar cómo se distribuye la proporción de los créditos morosos en función de las variables consideradas, en este capítulo se han construido tablas con datos agrupados en categorías de riesgo ordenadas según la tasa de mora observada. En el caso de las variables categóricas (factores), se fijaron como grupos de referencia aquellas categorías donde se obtuvieron las tasas de mora más altas. Para elegir las variables a tomar en cuenta en el ajuste del modelo de puntuación se consideraron, por una parte, los criterios aportados por la entidad colaboradora basados en la experiencia empírica, y por otra, la literatura sobre el tratamiento de variables explicativas en modelos de riesgo de crédito. Algunos trabajos en los que se han estudiado modelos de riesgo de crédito con variables similares a las tratadas aquí son Hand y Henley (1997), Fahrmeir y Tutz (2001) y Samreen y Zaidi (2012), entre otros.

Las variables consideradas en este capítulo son:

(a) *Género* La variable género ($GENE$) se incluyó como factor explicativo en el ajuste de los modelos debido a los resultados obtenidos con los contrastes de hipótesis sobre la igualdad de proporciones de morosidad entre clientes del sexo *masculino* y *femenino*.

Los contrastes empleados (test exacto de *Fisher*, test Chi-cuadrado y el test basado en la *aproximación normal*) arrojaron p-valores $\ll 0.001$, por lo que la variable *género* es un factor predictivo de la morosidad.

Las categorías de riesgo de la variable $GENE$ son:

1. Femenino
2. Masculino

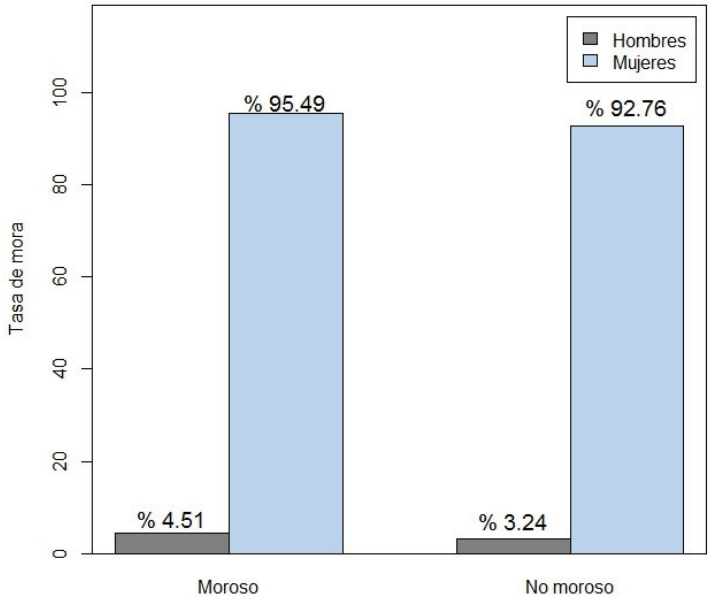

Figura 2.6 Distribución de la morosidad por género de los clientes.

La Figura 2.6 ilustra la diferencia entre las proporciones de morosidad de los clientes de género masculino y femenino. Como se ha comentado antes, dichas proporciones resultaron estadísticamente distintas con un nivel de significación del 95 %, y por tanto, es razonable pensar que el género es un factor predictivo de la situación de morosidad de los acreditados [3].

[3]La inclusión de variables como el género, el estado civil, la edad, u otras similares a las tratadas en este capítulo, en modelos de puntuación crediticia está sujeta a la normativa de regulación y supervisión financiera vigente en cada país. En España, por ejemplo, rige la normativa europea que permite el tratamiento estadístico de este tipo de datos.

Tabla 2.5. Tasa de morosidad por grupo de riesgo para la variable *género*.

	Situación de morosidad			
Grupo de riesgo	Moroso	No moroso	Total	TM (%)
Femenino	401	11 987	12 388	3.24
Masculino	569	12 043	12 612	4.51
Total	970	24 030	25 000	3.88

En la Tabla 2.5 se ve que la tasa de mora de los clientes de género masculino es mayor que la correspondiente a los de género femenino, por lo que la categoría de referencia para la variable $GENE$ es la categoría *masculino*. Como resultado, el peso de la variable $GENE$ en el modelo es cero si el cliente pertenece al sexo masculino y distinto de cero si no.

(b) *Estado civil* Es conocido, por experiencia empírica, que el *estado civil* (EST_CVL) de los acreditados es un factor socio-demográfico altamente predictivo del comportamiento crediticio de los mismos. Por este motivo, la variable *estado civil* fue incluida como factor explicativo en todos los modelos estudiados en este capítulo. Los grupos de riesgo de la variable EST_CVL son los siguientes:

1. Casado y pareja de hecho.
2. Viudo
3. Soltero
4. Divorciado, separado judicial y separado de hecho.

Tabla 2.6. Tasa de morosidad por grupo de riesgo para la variable *estado civil*.

	Situación de morosidad			
Grupo de riesgo	Moroso	No moroso	Total	TM (%)
Grupo 1	17	598	615	2.76
Grupo 2	408	12491	12899	3.16
Grupo 3	61	1443	1504	4.06
Grupo 4	484	9498	9982	4.85
Total	970	24030	25000	3.88

En la Tabla 2.6 se observa que la mayor tasa de mora de la cartera de créditos, un 4.85 %, se concentra entre los clientes divorciados, separados judicialmente y separados de hecho. Por este motivo se fijó este grupo como categoría de referencia de la variable EST_CVL en el modelo ajustado. En contraste, los clientes que están casados o que tienen parejas de hecho, representan el menor riesgo de morosidad para la entidad, alcanzando una tasa de mora del 2.76 %.

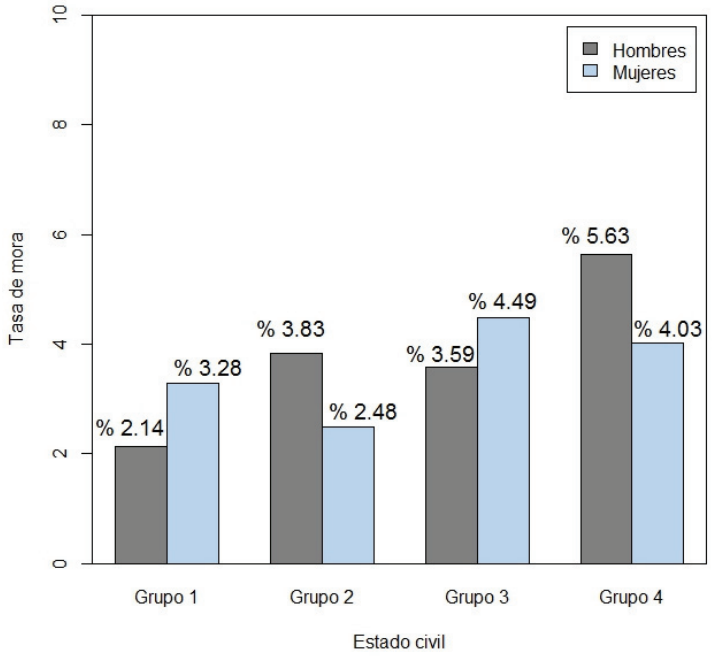

Figura 2.7 Distribución de la morosidad según el género y el estado civil de los clientes.

La Figura 2.7 muestra la distribución de la morosidad por género en cada grupo de riesgo construido a partir del estado civil de los acreditados. Se observa que la morosidad entre los clientes del género masculino es mayor en el grupo 4, de los divorciados y separados, mientras que entre las mujeres, la morosidad es mayor en el grupo 3, de las solteras.

(c) *Profesión* La profesión ($PROF$) de los clientes es uno de los factores socio-demográficos con mayor capacidad discriminante en el estudio de la

morosidad en tarjetas de crédito. Los grupos de riesgo formados a partir de la *profesión* de los acreditados son los siguientes:

1. Técnico superior, mando intermedio, administrativo, encargado de tienda, profesor, notario, procurador, militar, policía, arquitecto, ingeniero y rentista.

2. Gerente – alto cargo ejecutivo, médico, dentista, veterinario, farmacéutico y jubilado – pensionista.

3. Obrero especializado y abogado.

4. Vendedor comisionista, obrero, periodista, escritor, traductor, artista, deportista, religioso, ama de casa, estudiante, parado, otra profesión liberal y otra profesión no liberal.

Tabla 2.7. Tasa de morosidad por grupo de riesgo para la variable *profesión*.

	Situación de morosidad			
Grupo de riesgo	Moroso	No moroso	Total	TM (%)
Grupo 1	71	2637	2708	2.62
Grupo 2	222	6175	6397	3.47
Grupo 3	160	3730	3890	4.11
Grupo 4	517	11488	12005	4.31
Total	970	24030	25000	3.88

En la Tabla 2.7 se observa que la tasa de mora más alta se concentra entre los clientes con profesiones menos tradicionales, o liberales, como por ejemplo: los artistas, los músicos, algunos deportistas, estudiantes, amas de casa y parados, entre otros. La morosidad en este grupo alcanza el 4.31 %, y por tanto, el perfil de referencia para la variable profesión está formado por acreditados con profesiones pertenecientes al grupo de riesgo 4.

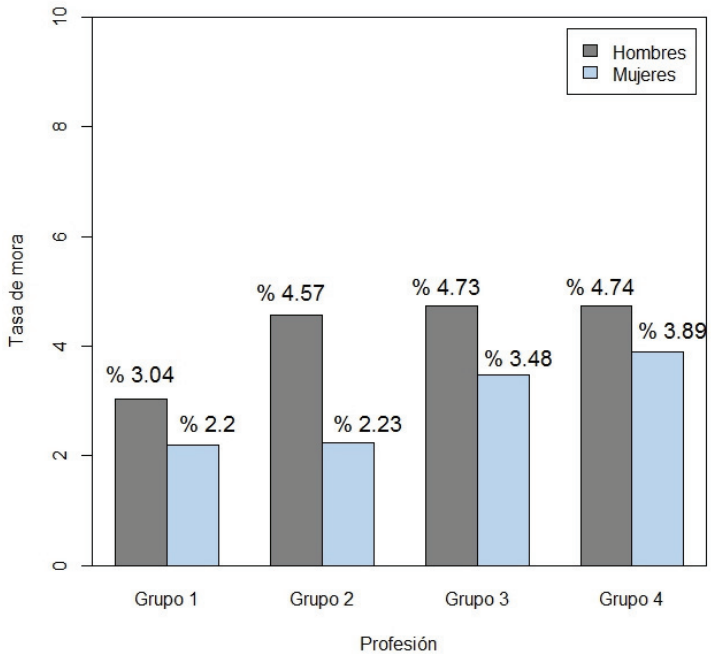

Figura 2.8 Distribución de la morosidad por género y profesión de los clientes.

La Figura 2.8 muestra la distribución de la morosidad por género en cada grupo de riesgo construido a partir de las profesiones de los acreditados. Se observa además que la morosidad es más alta en los clientes del género masculino que en los del género femenino, sea cual sea el perfil profesional de éstos.

(d) *Lugar de residencia* En esta memoria, la variable *lugar de residencia* ($LRESID$) corresponde a la provincia en la que reside el titular de la tarjeta. Se conoce, por experiencia empírica, que la morosidad de los créditos personales está relacionada con el lugar geográfico donde residen los clientes. Por este motivo, el *lugar de residencia* se ha incluido como factor explicativo en los modelos estudiados en este capítulo. Los grupos de provincias fueron conformados siguiendo las recomendaciones hechas por la entidad colaboradora, quedando denifidos de la siguiente manera:

1. Álava, Ávila, Baleares, Burgos, Cádiz, Cuenca, Guadalajara, Huelva, Málaga, Murcia, Navarra, Pontevedra, Cantabria, Santa Cruz de Tenerife, Soria, Teruel, Toledo, Valencia, Valladolid, Vizcaya, Zamora y Zaragoza.

2. Badajoz, Castellón, Córdoba, Granada, Guipúzcoa, León, A Coruña, Segovia, Sevilla y Tarragona.

3. Alicante, Girona, Ourense, Huesca, Jaén, Lleida, Lugo, Madrid y Asturias.

4. Albacete, Almería, Barcelona, Cáceres, Ciudad Real, La Rioja, Palencia, Las Palmas de Gran Canaria, Salamanca, Ceuta y Melilla y Extranjero.

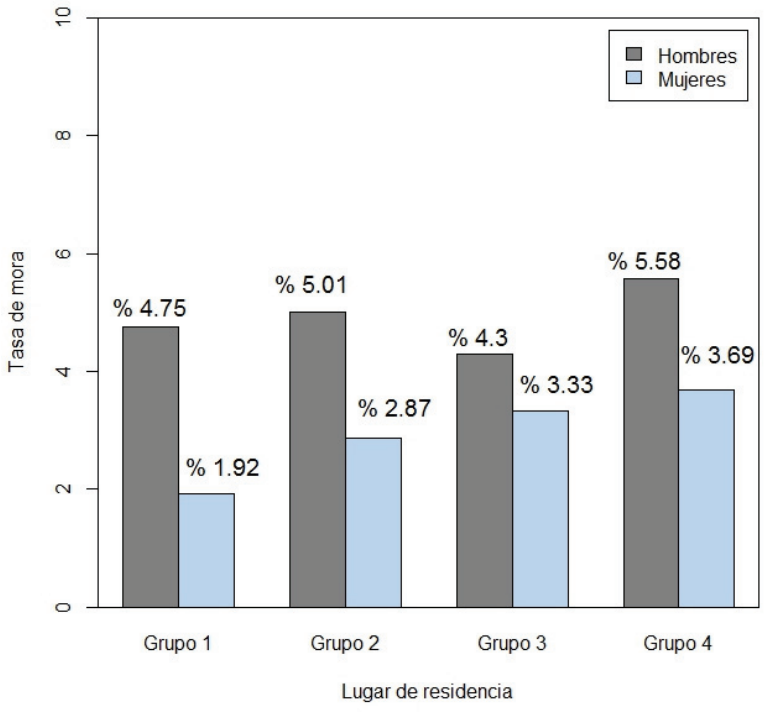

Figura 2.9 Distribución de la morosidad por género y lugar de residencia.

En la Figura 2.9 se ilustra la distribución de la morosidad por género en cada grupo de riesgo construido a partir de las provincias españolas (y extranjero) donde residen los acreditados. La figura muestra que, según esta conformación de grupos de riesgo por provincias, la proporción de créditos morosos es más alta en los hombres que en las mujeres. Además, la evolución de la tasa de morosidad es distinta entre hombres y mujeres, como se aprecia en los clientes pertenecientes al grupo 3.

Tabla 2.8. Tasa de morosidad por grupo de riesgo para la variable *lugar de residencia.*

	Situación de morosidad			
Grupo de riesgo	Moroso	No moroso	Total	TM (%)
Grupo 1	58	1752	1810	3.20
Grupo 2	60	1464	1524	3.94
Grupo 3	715	17936	18651	3.98
Grupo 4	137	2878	3015	4.54
Total	970	24030	25000	3.88

La Tabla 2.8 muestra que la tasa de morosidad no parece ser muy distinta entre los grupos 2 y 3. Los grupos 1 y 4, en cambio, muestran diferencias de más de un 0.5 % con respecto a las demás. Las tasa de mora más alta, un 4.54 %, se concentra entre los clientes que residen en las provincias que conforman el grupo 4, razón por la que este grupo se ha fijado como la categoría de referencia de la variable $LRESID$ en los modelos estudiados.

(e) ***Tipo de vivienda*** Se conoce, por experiencia empírica, que uno de los factores determinantes del nivel de solvencia de los clientes bancarios corresponde al *tipo de vivienda* en la que habitan. Por este motivo, la variable *tipo de vivienda* ($VVNDA$) ha sido incluida en los modelos que se estudian en este capítulo. Los grupos de riesgo de la variable $VVNDA$ son:

1. Propiedad libre de cargas
2. Propiedad hipotecada
3. Domicilio con familiares y otros.
4. Alquiler

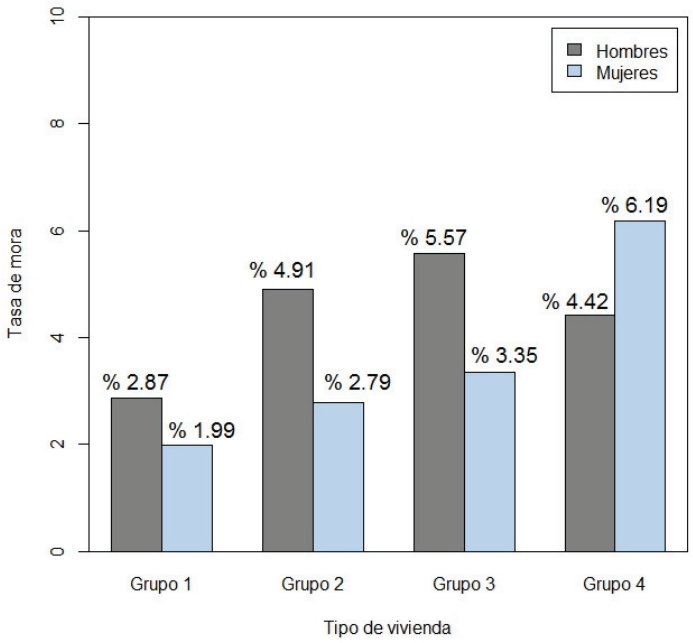

Figura 2.10 Distribución de la morosidad por género y tipo de vivienda de los clientes.

En la Figura 2.10 se muestra la distribución de la morosidad según el género y el tipo de vivienda de los acreditados. Se observa que la tasa de morosidad es más alta en los hombres que en las mujeres en todos los tipos de vivienda con excepción de los clientes que viven de alquiler, donde la morosidad es mayor en las mujeres (6.19 %) que en los hombres (4.42 %).

Tabla 2.9. Tasa de morosidad por grupo de riesgo para la variable *tipo de vivienda*.

	Situación de morosidad			
Grupo de riesgo	Moroso	No moroso	Total	TM (%)
Grupo 1	166	6614	6780	2.44
Grupo 2	260	6357	6617	3.93
Grupo 3	385	8286	8671	4.44
Grupo 4	159	2773	2932	5.42
Total	970	24030	25000	3.88

La Tabla 2.9 muestra la distribución de la tasa de morosidad según el *tipo de vivienda* de los clientes. Allí se observa que la tasa de morosidad es más alta en los clientes que alquilan su vivienda, razón por la que el grupo de referencia de la variable $VVNDA$ es la propiedad en alquiler.

(f) *Edad* Se conoce, de la experiencia empírica, que la variable *edad* es un buen predictor de la solvencia de los clientes bancarios, razón por la que fue incluida en todos los modelos ajustados. La variable *edad* ($EDAD$) fue segmentada en cuatro grupos de riesgo:

1. De 15 hasta menos de 25 años
2. De 25 hasta menos de 45 años
3. De 45 hasta menos de 65 años
4. De 65 años o más

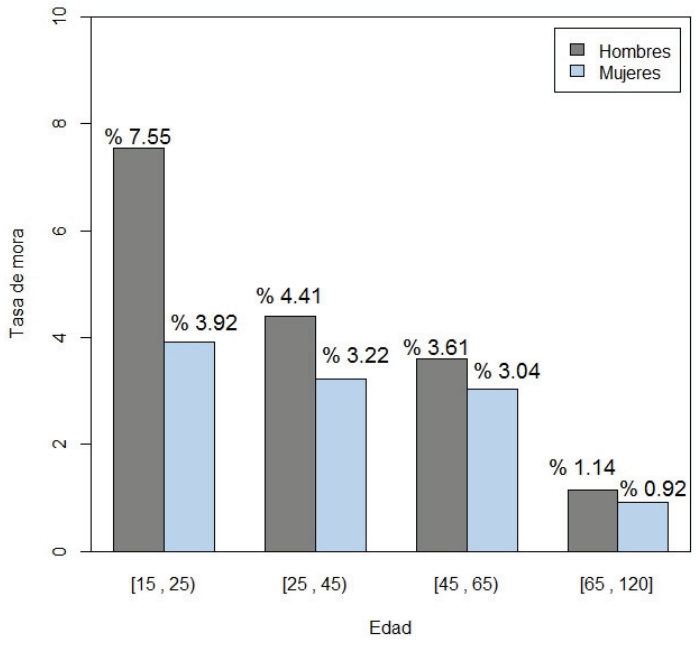

Figura 2.11 Distribución de la morosidad por edad y género de los clientes.

En la Figura 2.11 se observa que la tasa de morosidad de los hombres supera a la de las mujeres en los cuatro segmentos de edad definidos. Llama la atención la gran diferencia de proporción de morosos entre hombres y mujeres en el rango de edad comprendido entre los 15 hasta menos de 25 años, donde los hombres exhiben una tasa de mora del 7.55 % frente al 3.92 % de la mujeres. En las categorías restantes, las diferencias de tasa de morosidad entre hombres y mujeres decrecen hasta hacerse muy parecidas, como ocurre entre los clientes de 65 o más años (diferencia de sólo un 0.22 %).

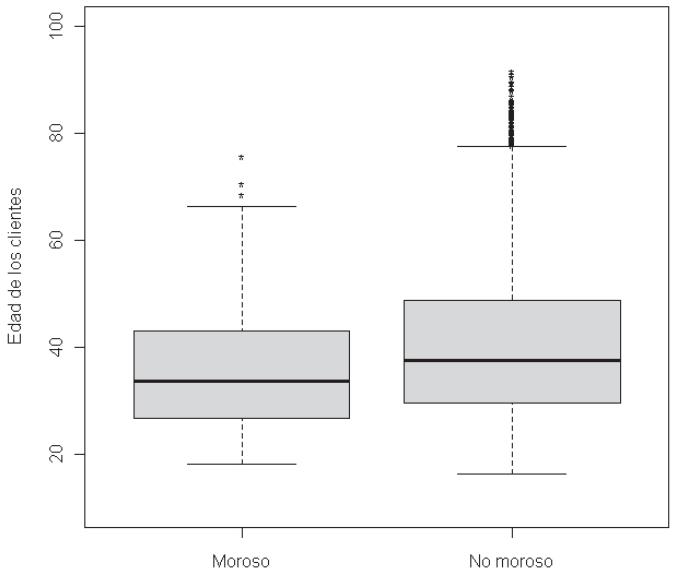

Figura 2.12 Gráfico de cajas de la distribución de la edad según la situación de morosidad.

En la Figura 2.12 se presenta un gráfico de cajas para comparar el grado de dispersión de la distribución de la *edad* de los acreditados según su situación de morosidad. En la figura se observa que, en el caso de los clientes no morosos, existe una alta concentración de valores atípicos en los clientes con más de 80 años. De este hecho se deduce que la distribución de la edad de los clientes no morosos posee una alta asimetría positiva, dando evidencia de no normalidad.

Tabla 2.10. Tasa de morosidad por grupo de riesgo para la variable *edad*.

	Situación de morosidad			
Grupo de riesgo	Moroso	No moroso	Total	TM (%)
Grupo 1	189	2471	2660	7.10
Grupo 2	567	13679	14246	3.98
Grupo 3	203	6815	7018	2.89
Grupo 4	11	1065	1076	1.02
Total	970	24030	23 000	3.88

La Tabla 2.10 muestra que la distribución de la morosidad de los acreditados decrece con la edad. Los datos muestran que el grupo de mayor riesgo para la entidad lo conforman los clientes con edades comprendidas entre los 15 hasta menos de 25 años, alcanzando un 7.10 % de morosidad.

(g) *Antigüedad laboral* La variable antigüedad laboral (ANT_LAB) se define como el tiempo que el acreditado ha permanecido trabajando en el empleo actual hasta el momento de solicitar el crédito. Para aquellos acreditados que son jubilados o pensionistas, la atigüedad laboral se refiere al tiempo que duró el acreditado en su último empleo antes de jubilarse.

Los grupos de riesgo para la variable *Antigüedad laboral* son:

1. De 0 hasta menos de 5 años
2. De 5 hasta menos de 10 años
3. De 10 hasta menos de 20 años
4. De 20 años o más

Tabla 2.11. Tasa de morosidad por grupo de riesgo para la variable *antigüedad laboral*.

	Situación de morosidad			
Grupo de riesgo	Moroso	No moroso	Total	TM (%)
Grupo 1	218	4090	4308	5.06
Grupo 2	426	9707	10133	4.20
Grupo 3	321	9757	10078	3.19
Grupo 4	5	476	481	1.04
Total	970	24030	25000	3.88

La Tabla 2.11 muestra que la distribución de la morosidad de los acreditados decrece a medida que aumenta la antigüedad en el empleo. El grupo de mayor riesgo para la entidad lo conforman aquellos clientes cuya antigüedad laboral es inferior a 5 años, alcanzando una tasa de mora del 5.06 %.

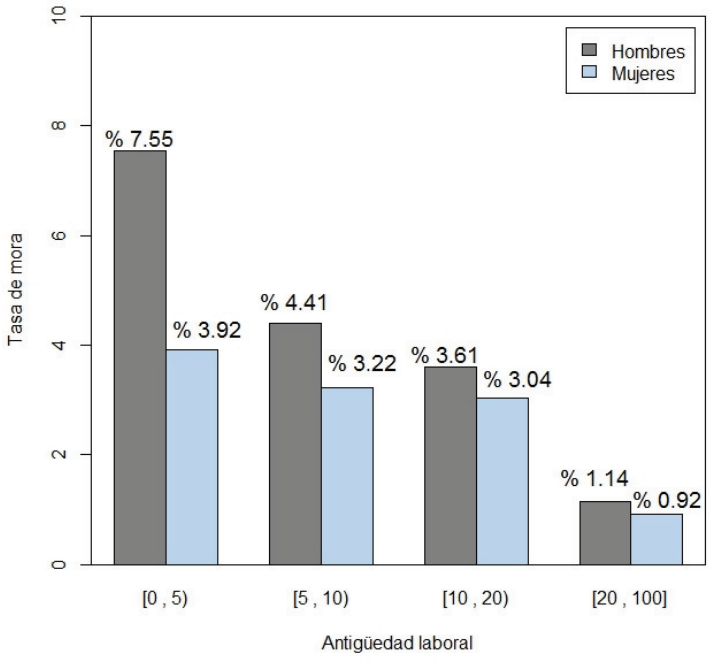

Figura 2.13 Distribución de la morosidad por género y antigüedad laboral.

La Figura 2.13 muestra que en el grupo de acreditados con la menor antigüedad laboral (menos de 5 años) se concentra la mayor tasa de morosidad, tanto en hombres como en mujeres. No obstante, es llamativa la diferencia que existe entre los hombres y las mujeres de este grupo, aproximadamente de un 3.63 % de morosidad. Una posible explicación de la alta morosidad en este grupo de clientes es que un porcentaje considerable de ellos (más del 38 %) resultó tener antigüedad laboral igual a 0, donde parte de estos clientes concuerdan con el perfil profesional conformado por el grupo 4 de la variable $PROF$.

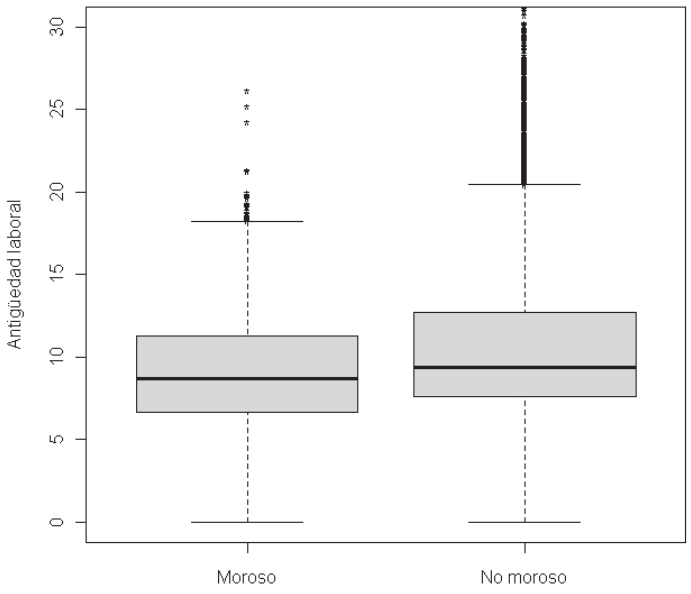

Figura 2.14 Gráfico de cajas de la distribución de la antigüedad laboral según la situación de morosidad.

En la Figura 2.14 se presenta un gráfico de cajas para comparar el grado de dispersión de la distribución de la *antigüedad laboral* de los acreditados según su situación de morosidad. En ambos tipos de acreditados se aprecia una alta concentración de valores atípicos en la cola superior de la distribución (a partir de una antigüedad de más de 18 años en el caso de los clientes

morosos y a partir de más de 20 años en los clientes no morosos). Este hecho indica que la distribución de la antigüedad laboral posee una alta asimetría positiva, dando evidencia de no normalidad.

(h) *Saldo a la vista* La variable saldo a la vista ($SALDO$) se define como la cantidad de dinero disponible que está a la vista en la cuenta de ahorros del titular en la fecha en la que el banco hace efectivo el cobro de la tarjeta.

Los grupos de riesgo resultantes para la variable *saldo a la vista* son los siguientes:

1. De 0 hasta menos de 1 000 €
2. De 1 000 hasta menos de 5 000 €
3. De 5 000 hasta menos de 10 000 €
4. De 10 000 hasta 1 000 000 €

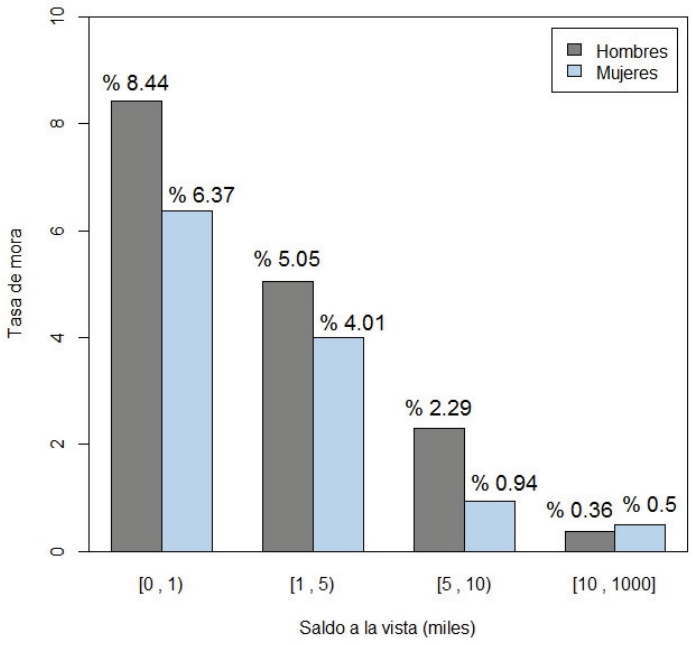

Figura 2.15 Distribución de la morosidad por género y saldo a la vista de los clientes.

En la Figura 2.15 se ilustra la distribución de la morosidad según el género y el saldo a la vista de los clientes. Se observa que la morosidad es más alta en los hombres que en la mujeres, con excepción de los clientes pertenecientes al grupo 4 (saldo a la vista entre 10 000 € y 1 000 000 €). Además, se aprecia que la diferencia en la tasa de mora entre hombres y mujeres es más alta en el grupo 1 (8.44 % en los hombres versus 6.37 % de las mujeres).

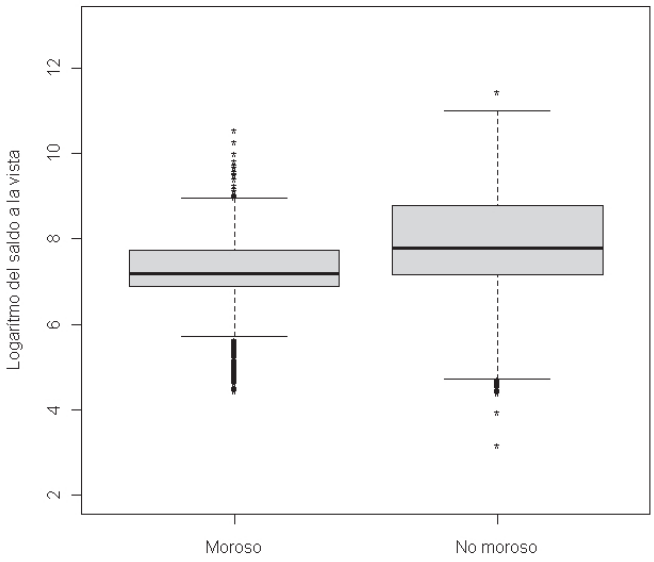

Figura 2.16 Gráfico de cajas de la distribución de la variable saldo a la vista en escala logarítmica.

En la Figura 2.16 se presenta un gráfico de cajas para examinar el grado de dispersión de la variable *saldo a la vista* según la situación de morosidad de los clientes. Para facilitar su inspección visual, esta variable ha sido convenientemente reescalada tomando su logarítmo natural. En la figura se puede ver que en los créditos morosos existe una alta concentración de valores atípicos en ambos extremos de la distribución. En contraste, los créditos no morosos presentan una distribución con más asimetría negativa. En ambos casos, la distribución del logaritmo de la variable *saldo a la vista* muestra evidencia de no normalidad.

Tabla 2.12. Tasa de morosidad por grupo de riesgo para la variable *saldo a la vista*.

	Situación de morosidad			
Grupo de riesgo	Moroso	No moroso	Total	TM (%)
Grupo 1	248	3219	3467	7.15
Grupo 2	647	13368	14015	4.62
Grupo 3	58	3676	3734	1.55
Grupo 4	17	3767	3784	0.45
Total	970	24030	25000	3.88

La Tabla 2.12 muestra que la morosidad de los acreditados disminuye a medida que aumenta el saldo a la vista en la cuenta corriente de los clientes. Además, en esta tabla se observa que, tal y como se podía esperar, la tasa de mora más alta, un 7.15 %, se concentra en aquellos clientes cuya cuenta corriente verifica un valor de saldo a la vista de menos de 1 000 €.

2.4.3. Modelos de puntuación crediticia ajustados

En esta sección se analizan los resultados de la estimación y validación de modelos estadísticos de puntuación crediticia. El objetivo que se persigue es mostrar la funcionalidad de los *modelos de regresión logística* como herramientas para la toma de decisiones sobre la concesión de tarjetas de crédito. En el ajuste de los modelos de regresión se utilizó como variable dependiente la variable indicadora de morosidad del crédito y como variables regresoras las características socio-demográficas descritas en la Sección 2.4.2 (variables (a) – (h)), que fueron tratadas previamente con las técnicas descritas en la Sección 2.2.3.

A continuación, se exponen los resultados obtenidos con los dos modelos que mejor se ajustaron a los datos. Las estimaciones obtenidas con estos modelos fueron comparadas con las puntuaciones del modelo proporcionado por la entidad colaboradora que, por simplicidad de escritura, ha sido etiquetado como modelo $ENTIDAD$.

El primer modelo fue ajustado con variables continuas tratadas con el método de *regresión polinómica* que, por simplicidad de escritura, fue etiquetado como $MPOL$. El segundo modelo fue ajustado con variables continuas tratadas con el método de *regresión segmentada* y etiquetado como $MSEG$.

Es importante mencionar que, por motivos de confidencialidad, los coeficientes de la regresión del modelo $ENTIDAD$ no se conocieron durante el período de colaboración con la entidad financiera que proporcionó los datos utilizados en esta memoria.

Modelo de puntuación con variables polinómicas

La Tabla 2.13 muestra el resultado del ajuste del modelo logístico con covariables continuas tratadas con el método de *regresión polinómica*. La expresión POL_j, con $1 \leq j \leq 3$, se refiere al término de grado j-ésimo de la variable continua incluida en el modelo de regresión. El subíndice que acompaña a los factores $GENE$, $VVNDA$, $LRESID$ y $PROF$ corresponde a la j-ésima variable auxiliar (dummy) incluida en el modelo de regresión, con $1 \leq j \leq 3$.

Tabla 2.13. Resumen del ajuste del modelo logístico $MPOL$

Covariable	Coeficiente estimado ($\hat{\beta}$)	Error estándar	Estadístico de contraste	p-valor
$Constante$	-3.16400	0.1633	-19.379	< 0.00001
$GENE_1$	-0.16120	0.0761	-2.118	0.034174
$VVNDA_1$	-0.61480	0.1380	-4.455	< 0.00001
$VVNDA_2$	-0.32060	0.1190	-2.693	0.007077
$VVNDA_3$	-0.37740	0.1125	-3.354	0.000796
$LRESID_1$	-0.45820	0.1827	-2.507	0.012161
$LRESID_2$	-0.14710	0.1761	-0.835	0.403593
$LRESID_3$	-0.25990	0.1092	-2.380	0.017334
$PROF_1$	0.06230	0.1530	0.408	0.683536
$PROF_2$	-0.27670	0.0946	-2.925	0.003446
$PROF_3$	-0.11450	0.1047	-1.094	0.274023
$EDAD\_POL_1$	0.01970	0.0068	2.901	0.003718
$EDAD\_POL_3$	-0.00010	0.0002	-4.370	< 0.00001
$SALDO\_POL_1$	-0.22360	0.0174	-12.840	< 0.00001
$SALDO\_POL_2$	0.01080	0.0026	4.083	< 0.00001
$SALDO\_POL_3$	-0.00017	0.0001	-2.104	0.035412

Modelo de puntuación con variables segmentadas

En la Tabla 2.14 se muestra el resultado de la estimación del modelo logístico con covariables continuas tratadas con el método de regresión segmentada. La expresión SEG_j, con $1 \leq j \leq 4$, se refiere al j-ésimo segmento de la variable continua incluida en modelo de regresión. Las variables categóricas inicialmente seleccionadas para este modelo fueron las mismas que las empleadas en el ajuste del modelo $MPOL$, aunque en el paso final del ajuste del modelo $MSEG$, el factor explicativo $GENE$ no fue significativo.

Tabla 2.14. Resumen del ajuste del modelo logístico $MSEG$

Covariable	Coeficiente estimado ($\hat{\beta}$)	Error estándar	Estadístico de contraste	p-valor
$Constante$	-0.995012	0.332422	-2.993	0.002761
$VVNDA_1$	-0.637742	0.137152	-4.650	< 0.00001
$VVNDA_2$	-0.300360	0.119039	-2.523	0.011629
$VVNDA_3$	-0.396792	0.113660	-3.491	0.000481
$LRESID_1$	-0.457927	0.182725	-2.506	0.012207
$LRESID_2$	-0.139239	0.176096	-0.791	0.429120
$LRESID_3$	-0.258262	0.109100	-2.367	0.017923
$PROF_1$	-0.023517	0.151928	-0.155	0.876985
$PROF_2$	-0.253108	0.095431	-2.652	0.007996
$PROF_3$	-0.103021	0.104980	-0.981	0.326423
$EDAD\_SEG_1$	-0.106345	0.036992	-2.875	0.004043
$EDAD\_SEG_2$	0.014774	0.007121	2.075	0.038006
$EDAD\_SEG_4$	-0.228563	0.082677	-2.765	0.005700
$ANT\_LAB\_SEG_1$	0.110416	0.044143	2.501	0.012373
$ANT\_LAB\_SEG_2$	-0.070249	0.034054	-2.063	0.039127
$SALDO\_SEG_1$	-0.487620	0.173810	-2.805	0.005024
$SALDO\_SEG_2$	-0.255052	0.041148	-6.198	< 0.00001
$SALDO\_SEG_3$	-0.275041	0.053566	-5.135	< 0.00001

Análisis descriptivo de las puntuaciones estimadas

En este apartado se analizan los resultados de las puntuaciones obtenidas con los modelos ajustados, $MPOL$ y $MSEG$, frente a las puntuaciones utilizadas por la entidad colaboradora en su política de concesión de tarjetas de crédito. Para ello, en la Tabla 2.15 se expone la comparación de las puntuaciones obtenidas mediante el cálculo de estadísticas descriptivas usuales. El análisis descriptivo se complementa comparando gráficamente las distribuciones de las tres muestras de puntuaciones a partir de estimadores no paramétricos de sus funciones de densidad y de distribución (Figuras 2.17 a la 2.20).

Tabla 2.15. Estadísticos descriptivos de las puntuaciones estimadas.

Estadístico	Modelo de puntuación		
	$ENTIDAD$	$MPOL$	$MSEG$
min	-8.517	-36.044	-11.508
Q_1	-4.984	-4.284	-4.321
Q_2	-3.985	-3.261	-3.256
$media$	-3.973	-3.651	-3.611
$moda$	-4.462	-2.857	-2.975
Q_3	-2.899	-2.827	-2.863
max	-0.062	-1.358	-1.280
$d.e.$	1.442	1.210	1.065
$asimetría$	-0.109	-2.333	-1.008
$curtosis$	2.591	29.302	4.254

Los resultados expuestos en la Tabla 2.15 muestran que prácticamente no hay diferencias entre las medidas de tendencia central de las puntuaciones obtenidas con los modelos $MPOL$ y $MSEG$, mientras que sí las hay en términos de sus valores extremos. La diferencia más importante se observa en el valor mínimo de las puntuaciones obtenidas con el método de variables polinómicas ($min = -36.044$).

En las figuras siguientes, se representa la estimación no paramétrica de la función de densidad de la variable puntuación crediticia en la muestra completa (Figura 2.17), en la muestra de créditos morosos (Figura 2.18) y en la de créditos no morosos (Figura 2.19). Para esto se utilizó el estimador

núcleo de *Rosenblatt-Parzen* (Rosenblatt (1956) y Parzen (1962)) con función núcleo *gaussiana* y parámetro de suavizado, h, obtenido con el método *plug-in* de Sheather y Jones (1991), cuya implementación está disponible en la librería *KernSmooth* (Wand y Ripley (2008) del paquete estadístico *R* (*R* Core Team (2015)).

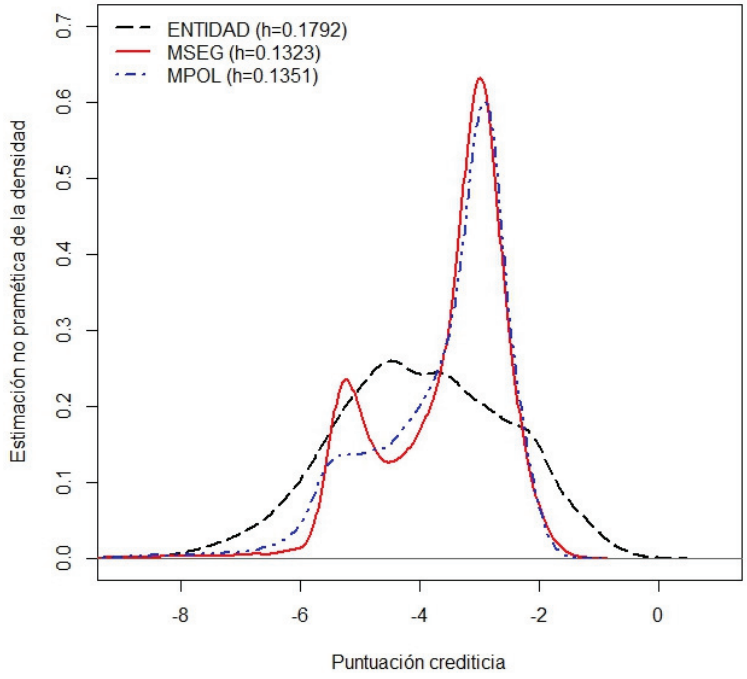

Figura 2.17 Estimación no paramétrica de las curvas de densidad de las puntuaciones crediticias.

En la Figura 2.17 se ilustran las curvas de densidad de las puntuaciones estimadas no paramétricamente. Debido al efecto de suavizamiento obtenido con las estimaciones no paramétricas, estas curvas permiten detectar rasgos importantes en la forma de las distribuciones de las puntuaciones. En primer lugar, se observa que existe una bimodalidad muy acusada de las puntuaciones obtenidas con el modelo $MSEG$. También se aprecia gran similitud entre las curvas obtenidas con los modelos ajustados ($MPOL$ y

$MSEG$) en comparación con la curva obtenida con el modelo $ENTIDAD$. Los coeficientes de curtosis y asimetría empíricos indican que las puntuaciones obtenidas con los modelos ajustados poseen distribuciones de tipo leptocúrticas con asimetría negativa, es decir, que son distribuciones muy concentradas en torno a sus valores modales, con más peso en la cola inferior. En particular, es llamativo el grado de curtosis de las puntuaciones obtenidas con el modelo $MPOL$ ($curtosis = 29.302$), que es aproximadamente 10 veces el de una distribución normal estándar ($curtosis = 3$). En contraste, las puntuaciones obtenidas con el modelo $ENTIDAD$ presentan más variabilidad, menos asimetría y menos concentración en torno a su valor modal ($curtosis = 2.5913 < 3$).

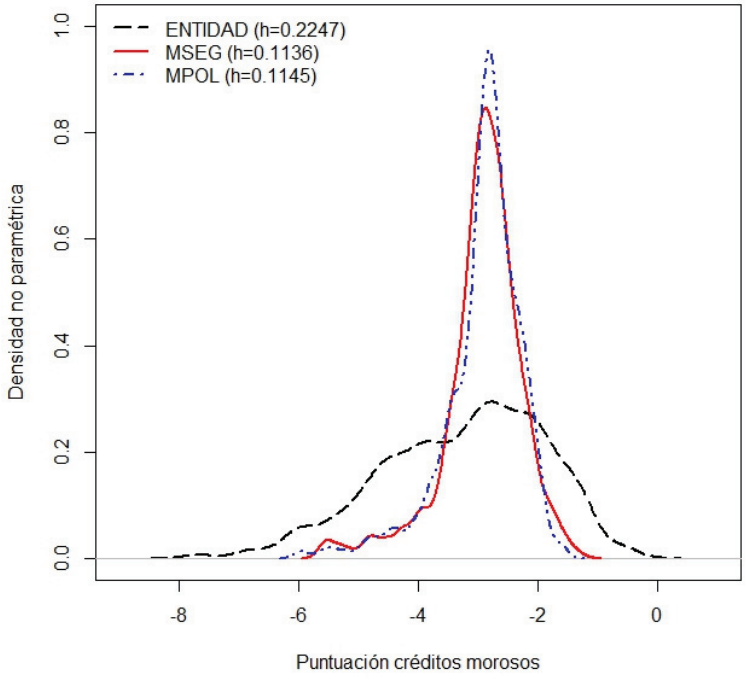

Figura 2.18 Estimación no paramétrica de las curvas de densidad de las puntuaciones de los créditos morosos.

Las curvas representadas en la Figura 2.18 corresponden a las funciones de densidad estimadas no paramétricamente para las puntuaciones asociadas

a los créditos morosos. A diferencia de las curvas representadas en la Figura 2.17, las puntuaciones obtenidas con el método de variables segmentadas no exhiben bimodalidad. Además, en el caso de los créditos morosos, existe una gran semejanza entre las funciones de densidad no paramétricas de las puntuaciones ajustadas con los modelos $MPOL$ y $MSEG$.

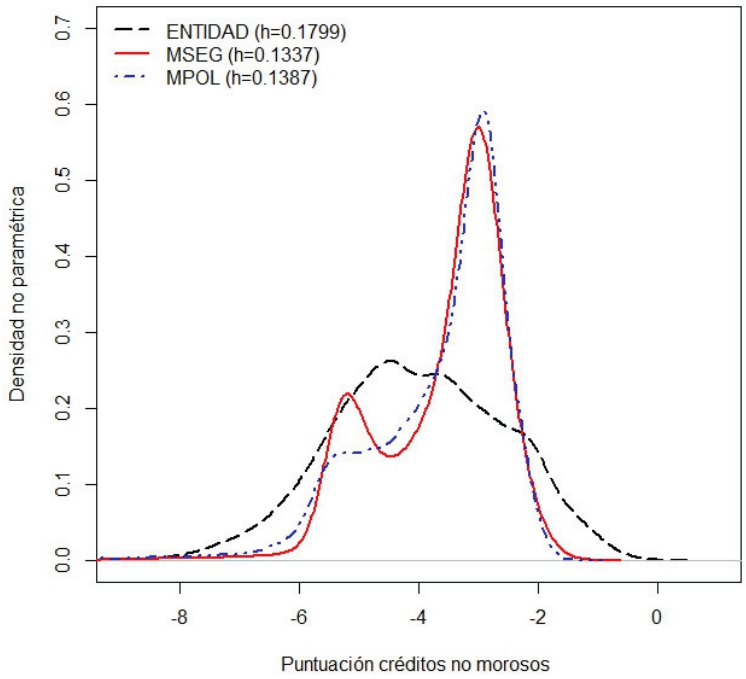

Figura 2.19 Estimación no paramétrica de las curvas de densidad de las puntuaciones de los créditos no morosos.

En la Figura 2.19 se ilustran las funciones de densidad estimadas no paramétricamente para las puntuaciones de los créditos no morosos. A diferencia de lo que ocurrre con los créditos morosos (Figura 2.18), en la Figura 2.19 se observa que las estimaciones obtenidas para los créditos no morosos son muy parecidas a las representadas en la Figura 2.17, obtenidas con toda la muestra. Este resultado se debe a la influencia que ejerce en las estimaciones la gran cantidad de clientes no morosos presentes en la muestra ($n_0 = 24030$

), quienes aportan el 96.12 % de la información del perfil crediticio global de la muestra.

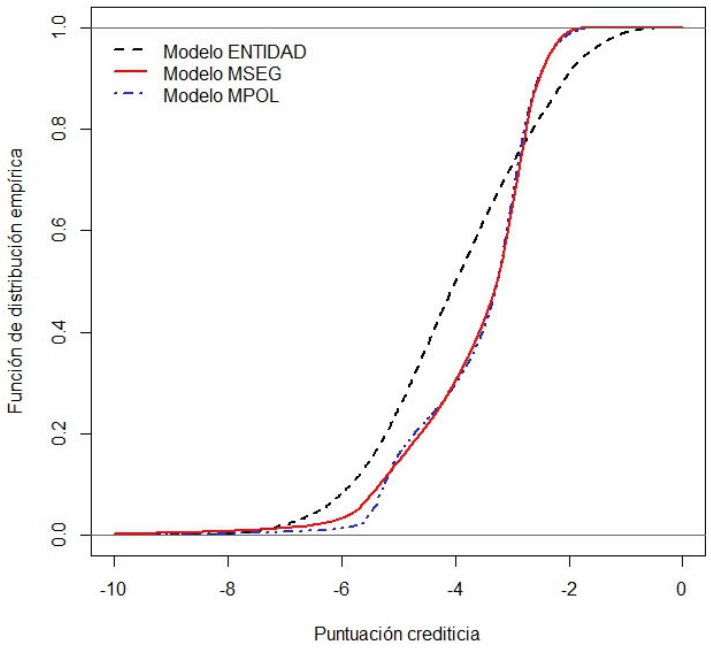

Figura 2.20 Funciones de distribución empíricas de las puntuaciones crediticias obtenidas con los tres modelos estudiados.

En la Figura 2.20 se representan las funciones de distribución empíricas de los tres modelos de puntuación. Se observa que, similarmente a los resultados obtenidos con las curvas de densidad no paramétricas, las funciones de distribución empíricas asociadas a los modelos de puntuación ajustados, $MPOL$ y $MSEG$, resultaron ser muy parecidas entre sí, a la vez que llamativamente distintas de la curva de distribución empírica obtenida con el modelo $ENTIDAD$. Para verificar estadísticamente este resultado, se compararon ambas distribuciones mediante un contraste de hipótesis sobre la igualdad de las mismas. Para esto se utilizó la prueba no paramétrica de *Wilcoxon* para dos muestras relacionadas (debido a la alta correlación lineal existente entre ambas muestras de puntuaciones, $\hat{\rho} = 0.9463$). Como resultado, se aceptó la hipótesis nula de igualdad de las distribuciones con un p-valor $= 0.4867$. Al

aplicar el mismo contraste entre las puntuaciones obtenidas con los modelos $MSEG$ y $ENTIDAD$, se obtuvo un p-valor < 0.00001, rechazando la hipótesis nula de igualdad de las distribuciones.

2.4.4. Análisis de validación y poder predictivo de los modelos ajustados

A continuación se expone una comparación de los modelos estudiados en este capítulo a partir de los resultados del análisis de bondad del ajuste aplicado a los datos.

Tabla 2.16. Medidas de bondad del ajuste y contraste de Hosmer-Lemeshow. Por motivos de confidencialidad, no se tuvo acceso a los valores sustituidos por $(*)$.

Modelo de puntuación	AIC	BIC	$\tilde{R}^2_{CS}$	$\tilde{R}^2_{Ng}$	χ^2_{HL}	p-valor
$ENTIDAD$	$(*)$	$(*)$	$(*)$	$(*)$	720.34	< 0.0001
$MPOL$	6092.46	6211.01	0.0217	0.0784	13.028	0.1109
$MSEG$	6097.65	6232.01	0.0217	0.0782	4.109	0.8471

En la Tabla 2.16 se muestran los resultados de los estadísticos de bondad del ajuste de *Akaike* (AIC) y de *Schwarz* (BIC), las medidas $\tilde{R}^2$ de *Cox-Snell* y de *Nagelkerke*, y el resultado del contraste de bondad del ajuste de *Hosmer-Lemeshow* (HL).

El resultado del contraste de HL indica que se rechaza la hipótesis nula de un ajuste logístico para el modelo $ENTIDAD$. Esto ocurre debido a que existen diferencias significativas entre las tasas de mora estimadas con el modelo $ENTIDAD$ y las proporciones de morosos reales observados en cada una de las clases de riesgo construidas para este contraste. A diferencia de esto, la hipótesis nula de que el modelo ajustado es adecuado no se rechaza para los modelos $MPOL$ y $MSEG$, aunque las medidas $\tilde{R}^2_{CS}$ y $\tilde{R}^2_{Ng}$ indican que el grado de información que proporcionan estos modelos es bajo.

A partir de los resultados obtenidos con las distintas medidas de bondad del ajuste aplicadas en ambas muestras se concluye que no existen diferencias significativas en la calidad del ajuste de los modelos $MPOL$ y $MSEG$.

Resultado del análisis de curvas *ROC*

A continuación, en la Figura 2.21 se representan las curvas *ROC* asociadas a los modelos de puntuación ajustados y se comparan con la curva *ROC* obtenida con las puntuaciones del modelo propuesto por la entidad.

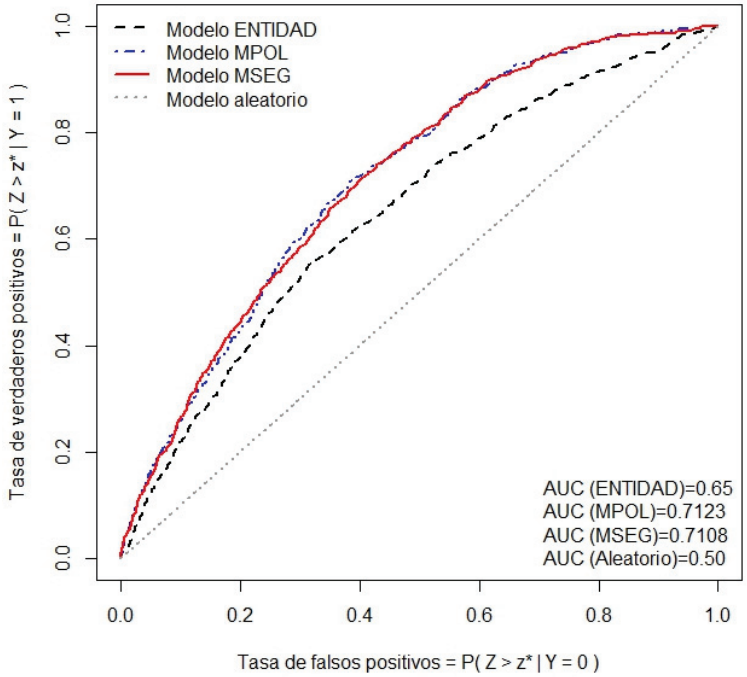

Figura 2.21 Curvas *ROC* resultantes para los tres modelos de puntuación crediticia.

En la Figura 2.21 se observa que las curvas *ROC* obtenidas con los modelos $MPOL$ y $MSEG$ son muy similares entre sí, estando prácticamente la primera (con trazo de color azul) superpuesta sobre la segunda (con línea de color rojo). También se observa que el área bajo la curva *ROC*, *AUC*, del modelo $MPOL$ es ligeramente mayor que el valor *AUC* obtenido con el modelo $MSEG$. Como las curvas *ROC* de los modelos ajustados están completamente por encima de la curva *ROC* obtenida a partir del modelo $ENTIDAD$, se concluye que los modelos ajustados ofrecen mejor capacidad de discriminación de los créditos.

Resultado del análisis de curvas CAP

En la Figura 2.22 se representan las curvas CAP obtenidas con los modelos ajustados y se comparan con la curva CAP asociada al modelo propuesto por la entidad.

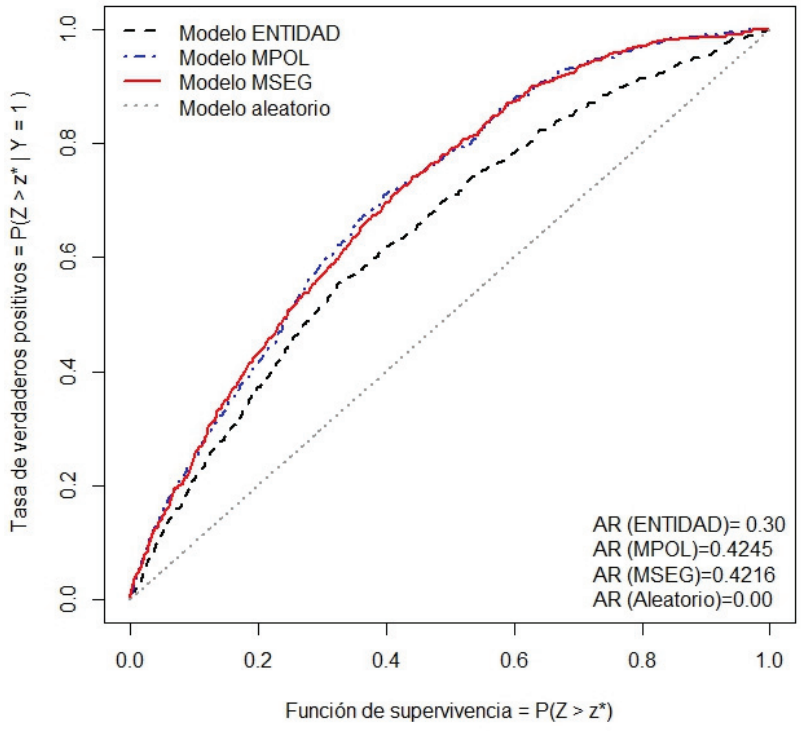

Figura 2.22 Curvas CAP resultantes para los tres modelos de puntuación crediticia.

Como se puede apreciar en la Figura 2.22, el análisis de curvas CAP ofrece resultados análogos a los obtenidos con las curvas ROC, es decir, la capacidad de discriminación entre créditos morosos y no morosos ofrecida por los modelos ajustados, $MPOL$ y $MSEG$, es equivalente entre sí y superior a la capacidad de discriminación que exhiben las puntuaciones obtenidas con el modelo $ENTIDAD$.

Elección del punto de corte y errores de clasificación asociados

En este apartado se analizan los resultados del procedimiento de validación estadística de los modelos. Para esto, se ha calculado el error de clasificación de los créditos en función del punto de corte obtenido para cada modelo además de otras técnicas de validación descritas en las Secciones 2.3.2, 2.3.3 y 2.3.4.

Tabla 2.17. Resultados acerca del poder discriminante de los modelos

Modelo de puntuación	z^*_{KS}	$KS\,(z^*_{KS})$	AUC	AR	$TP\ (\%)$
$ENTIDAD$	-3.25633	0.2379	0.650	0.300	67.94
$MPOL$	-3.06187	0.3249	0.712	0.426	61.84
$MSEG$	-3.10485	0.3115	0.711	0.422	60.43

En la Tabla 2.17 se muestra el punto de corte óptimo, z^*_{KS}, obtenido con el método de la distancia de *Kolmogorov-Smirnov*, $KS\,(z)$, y el resultado de la tasa TP asociada al punto z^*_{KS}. Además, se ofrecen las medidas de validación AUC y AR que no dependen de z^*_{KS}, sino que son medidas globales obtenidas de las curvas ROC y CAP, respectivamente. Los resultados obtenidos con los índices AUC y AR permiten concluir que la capacidad de discriminación de los modelos $MPOL$ y $MSEG$ son equivalentes. Además, ambos modelos superan en capacidad predictiva al modelo $ENTIDAD$. Este resultado fue probado por medio del contraste de hipótesis definido en (2.19). Como resultado, se aceptó la hipótesis nula sobre la igualdad de áreas bajo las curvas ROC obtenidas con los modelos $MPOL$ y $MSEG$ (p-valor > 0.9335). Al realizar el mismo contraste para comparar los modelos $MSEG$ y $ENTIDAD$, se rechazó la hipótesis nula con un p-valor < 0.00115. Estos resultados están en concordancia con lo obtenido en la Tabla 2.16

En cada una de las siguientes tablas se ha calculado el riesgo de crédito, medido como el error de clasificación de tipo I, y el coste de oportunidad, medido como el error de clasificación de tipo II. Además, cada tabla está acompañada de un gráfico con las distribuciones empíricas condicionadas de las puntuaciones dada la situación de morosidad versus la curva del estadístico $KS\,(z)$.

Tabla 2.18. Matriz de clasificación correspondiente al modelo $ENTIDAD$

	Clasificación según el modelo			
Situación	$Z >$ -3.25633	$Z \leq$ -3.25633	Total	Error (%)
Moroso	424	342	766	Tipo I: 44.65
No moroso	6 070	13 164	19 234	Tipo II: 31.56

La Tabla 2.18 muestra que, utilizando el modelo de la entidad, el riesgo de crédito debido a la concesión de la tarjeta a un cliente moroso cuya calificación crediticia resultó ser buena (puntuación inferior al punto de corte) es de un 44.65 %, mientras que el coste de oportunidad debido a la denegación de la misma a un cliente no moroso cuya calificación crediticia resultó ser mala (puntuación superior al punto de corte) es de un 31.56 %.

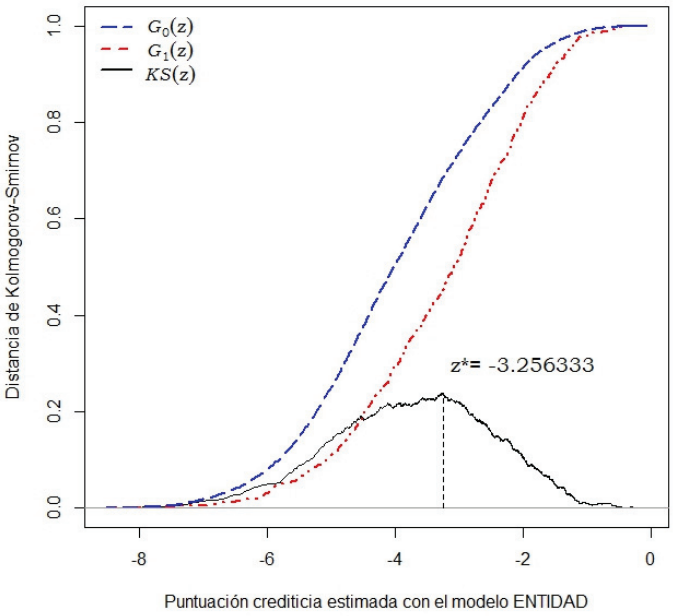

Figura 2.23 Funciones de distribución empíricas de las puntuaciones de los créditos no morosos (G_0), de los créditos morosos (G_1) y estadístico KS para el modelo *ENTIDAD*.

La Figura 2.23 muestra el gráfico del estadístico $KS(z)$ y el valor de corte, $z^*_{KS} = -3.25633$, calculado a partir de las puntuaciones obtenidas con el modelo $ENTIDAD$.

Tabla 2.19. Matriz de clasificación correspondiente al modelo $MPOL$

	Clasificación según el modelo			
Situación	$Z >$ -3.06187	$Z \leq$ -3.06187	Total	Error (%)
Moroso	544	222	766	Tipo I: 28.98
No moroso	7 410	11 824	19 234	Tipo II: 38.53

La Tabla 2.19 muestra que, utilizando el modelo $MPOL$, el riesgo de crédito debido a la concesión de la tarjeta a un cliente cuya calificación crediticia es buena y que resultó ser moroso es de un 28.98 %, mientras que el coste de oportunidad debido a la denegación de la misma a un cliente con mala calificación crediticia que resultó ser no moroso es de un 38.53 %.

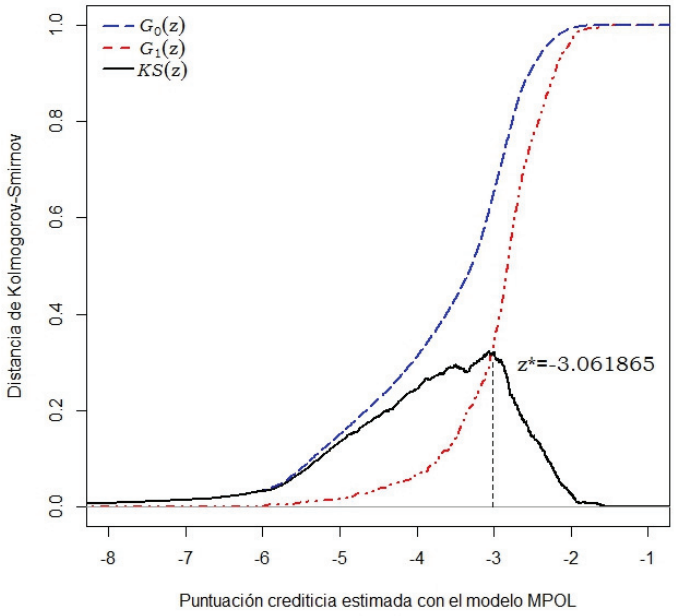

Figura 2.24 Funciones de distribución empíricas de las puntuaciones de los créditos no morosos (G_0), de los créditos morosos (G_1) y estadístico KS para el modelo $MPOL$.

La Figura 2.24 muestra el gráfico del estadístico $KS\,(z)$ y el valor de corte, $z^*_{KS} = -3.06187$, calculado a partir de las puntuaciones estimadas con el modelo $MPOL$.

Tabla 2.20. Matriz de clasificación correspondiente al modelo $MSEG$

	Clasificación según el modelo			
Situación	$Z >$ -3.10485	$Z \leq$ -3.10485	Total	Error (%)
Moroso	544	222	766	Tipo I: 28.98
No moroso	7 693	11 541	19 234	Tipo II: 40.00

La Tabla 2.20 muestra que el modelo $MSEG$ es equivalente al modelo $MPOL$ en términos de riesgo de crédito, obteniéndose en ambos casos un error de clasificación de tipo I de un 28.98 %. En cambio, el modelo $MSEG$ provoca un leve aumento del coste de oportunidad debido a una disminución del poder predictivo con respecto a lo obtenido con modelo $MPOL$. Como consecuencia, el error de tipo II asociado al modelo $MSEG$ es de un 40.00 %.

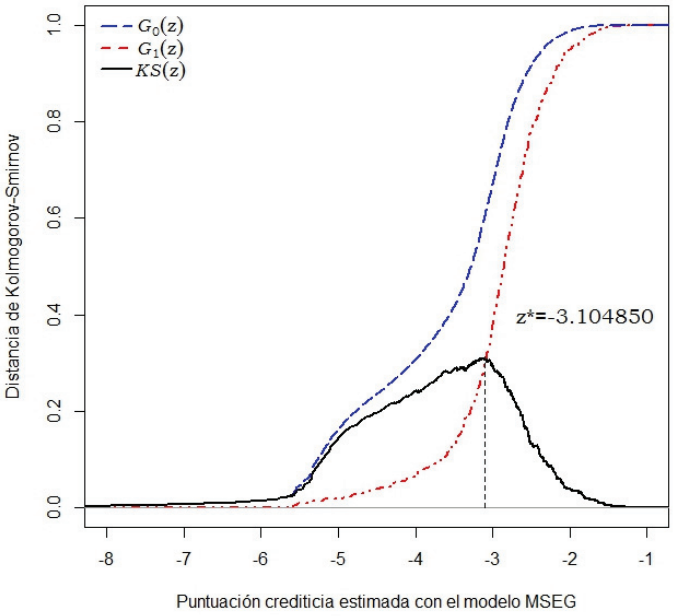

Figura 2.25 Funciones de distribución empíricas de las puntuaciones de los créditos no morosos (G_0), de los créditos morosos (G_1) y estadístico KS para el modelo *MSEG*.

La Figura 2.25 muestra el gráfico del estadístico $KS(z)$ y el valor de corte, $z^*_{KS} = -3.10485$, calculado a partir de las puntuaciones estimadas con el modelo $MSEG$.

A partir de los resultados contenidos en las Tablas 2.16 a la 2.20, y del resultado del análisis de curvas ROC y curvas CAP, se obtiene que el modelo $MPOL$ es preferido a los otros dos modelos debido a que su utilización en la clasificación de los créditos permite minimizar el riesgo de crédito (*error de tipo I*) y obtener mejores resultados con medidas de validación como los índices AUC y AR. Sin embargo, en términos del riesgo por coste de oportunidad (*error de tipo II*) y de la tasa de precisión (TP), el modelo que mostró mejor rendimiento fue el modelo de la entidad.

2.4.5. Resultados obtenidos con la muestra de validación

A continuación, se exponen los resultados obtenidos con la muestra de validación. Esta muestra se compone de 5 000 registros de tarjetas de crédito, de las cuales 204 son morosas, lo que equivale a una tasa de mora observada de un 4.08 %. Se ha analizado el poder predictivo de los modelos a partir de los índices AUC, AR y TP. También se calcularon los errores de clasificación de *tipo I* y de *tipo II*, respectivamente. El análisis concluye con la ilustración de las diferencias en la capacidad discriminante de los modelos a partir de sus gráficos ROC y CAP.

Tabla 2.21. Medidas de validación de los modelos de puntuación crediticia

Modelo de puntuación	z^*_{KS}	AUC	AR	TP (%)
$ENTIDAD$	-3.25633	0.612	0.223	66.44
$MPOL$	-3.06187	0.688	0.376	60.50
$MSEG$	-3.10485	0.679	0.359	58.98

En la Tabla 2.21 se ofrecen los resultados de las medidas de validación obtenidas con los tres modelos de puntuación. Análogamente a lo obtenido con la muestra de entrenamiento, de la Tabla 2.21 se desprende que el modelo con mayor poder de discriminación entre créditos buenos y malos es el modelo $MPOL$. Este modelo produce los valores más altos de área bajo la curva ROC ($AUC = 0.688$) e índice de precisión ($AR = 0.376$). Sin embargo, el valor más alto de tasa de precisión ($TP = 66.44\,\%$) se obtiene con el modelo $ENTIDAD$.

Tabla 2.22. Poder discriminante de los modelos obtenido con la muestra de validación

Modelo de puntuación	Riesgo de crédito (%)	Coste de oportunidad (%)
$ENTIDAD$	48.04	32.94
$MPOL$	32.84	39.78
$MSEG$	35.29	41.26

En la Tabla 2.22 se ofrecen los resultados correspondientes a los errores de clasificación de tipo I y de tipo II obtenidos con la muestra de validación. Los resultados muestran que el modelo $MPOL$ supera al modelo $MSEG$ en la reducción de ambos tipos de errores pero pierde capacidad de reducir el *error de tipo II* frente al modelo $ENTIDAD$. Además, el modelo $MPOL$ arroja el menor valor de *error de tipo I*, por lo que representa la herramienta de clasificación de los créditos que minimiza el *riesgo de crédito*.

Resultado del análisis de curvas ROC

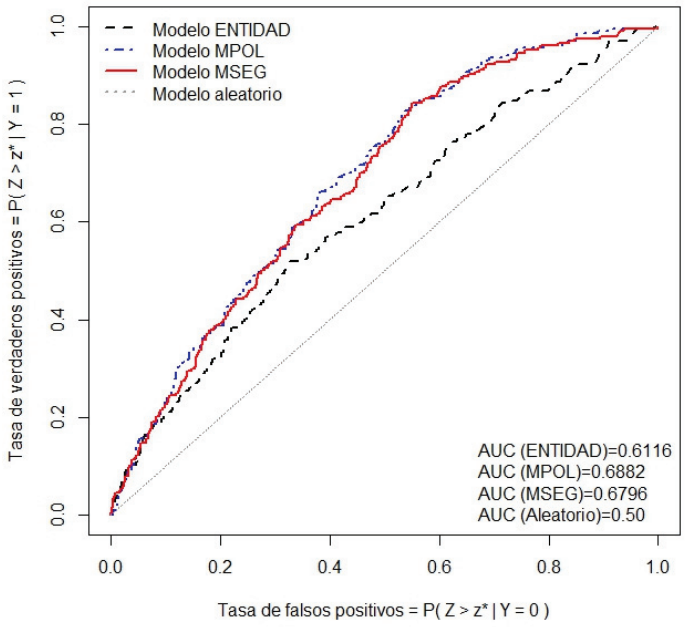

Figura 2.26 Resultado de las curvas ROC obtenidas con la muestra de validación.

La Figura 2.26 muestra las curvas ROC obtenidas con las puntuaciones de los tres modelos considerados. Visualmente, se confirman los resultados ofrecidos en las Tablas 2.21 y 2.22, en los que parece existir evidencia de una ligera ganancia en capacidad discriminante ofrecida por el modelo $MPOL$ en relación al modelo $MSEG$. Por otra parte, ambos modelos presentan un comportamiento sustancialmente mejor que el del modelo $ENTIDAD$.

Resultado del análisis de curvas CAP

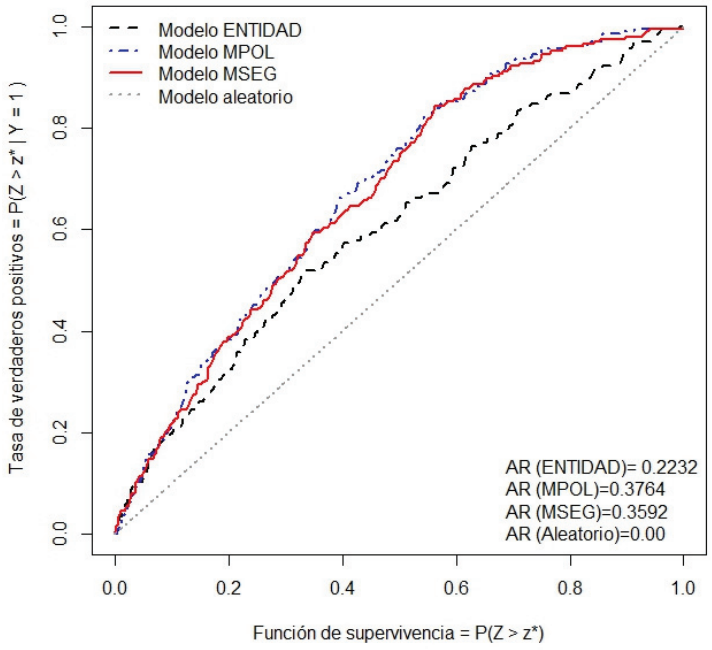

Figura 2.27 Resultado de las curvas CAP obtenidas con la muestra de validación.

En la Figura 2.27 se representan las curvas CAP de los tres modelos de puntuación obtenidas con la muestra de validación. Análogamente a lo obtenido con las curvas ROC (Figura 2.26), las curvas CAP resultantes son coherentes con los resultados de las medidas de validación expuestas en las Tablas 2.21 y 2.22.

Como consecuencia de los resultados obtenidos con las técnicas de validación aplicadas a los datos reales, se concluye que el modelo de puntuación $MPOL$ supera levemente al modelo $MSEG$ en capacidad discriminante, y que ambos modelos superaron significativamente la capacidad de reducción del riesgo de crédito del modelo $ENTIDAD$. Como contrapartida, el modelo de la entidad ofreció los mejores resultados en términos de reducción del riesgo por coste de oportunidad así como los valores más altos de la tasa de precisión de la clasificación de los créditos (TP).

2.4.6. Determinación de la frontera de concesión de la tarjeta

La estrategia de concesión de una tarjeta de crédito puede caracterizarse por el valor esperado de su función de costes y beneficios. Esta función, denotada por $V(z^*) \equiv E(U(Z, z^*))$, depende de la puntuación del solicitante, Z, y del valor de corte elegido, z^*. Para obtener la estrategia de concesión óptima de la tarjeta, se requiere determinar los puntos de corte que establecen las fronteras de decisión a partir de las cuales el dictamen de concesión (DC) es uno de los siguientes: *conceder*, *consultar* o *denegar*.

Los puntos de corte de la estrategia de concesión se obtienen a partir de la puntuación estimada por el modelo estadístico. La estrategia de concesión se describe representando los costes y los beneficios del dictamen de concesión en una matriz, como se verá más adelante. El procedimiento para elaborar la estrategia de concesión se resume a continuación.

Se denotará por F_c el punto de corte, o frontera, del rango de puntuaciones (o equivalentemente de la PD) a partir del cual el dictamen es *consultar* (créditos considerados dudosos por el modelo). Así, para todo valor de puntuación a la izquierda de F_c el dictamen es *conceder*. Además, se denotará por F_d al punto de corte o frontera a partir del cual el dictamen es *denegar*. La Figura 2.28 ilustra este criterio en función de la PD de las tarjetas.

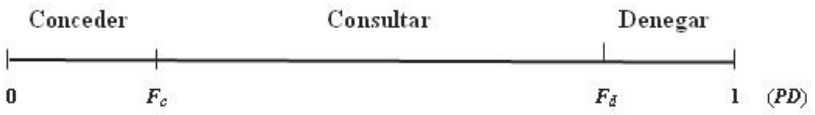

Figura 2.28 Puntos de corte del dictamen de concesión de la tarjeta de crédito.

El objetivo es determinar una estrategia óptima que permita obtener los puntos F_c y F_d. Para ello, se construye la función de costes, $V(z^*)$, donde la estrategia óptima se alcanza en el valor de z^* para el cual $V(z^*)$ toma su valor mínimo. Ambos valores se obtienen resolviendo ecuaciones análogas bajo dos escenarios económicos distintos, como se verá más adelante.

Tabla 2.23. Matriz de costes y beneficios asociada a la estrategia de concesión

Dictamen de concesión	Coste debido a los créditos morosos	Beneficio debido a los créditos no morosos
Denegar si $Z > z^*$	0	$-C_3$
Conceder si $Z \leq z^*$	C_1	$-C_2$

En la Tabla 2.23 se han definido las cantidades C_1, C_2 y C_3. La cantidad C_1 representa el coste debido a la concesión de la tarjeta a un solicitante que será moroso, C_2 representa el beneficio obtenido por la concesión de la tarjeta de crédito a un solicitante que no será moroso y C_3 representa el coste de oportunidad debido a la no concesión de la tarjeta a un cliente que no resultará moroso.

Supuestos del modelo

En la literatura es posible encontrar modelos para la función de costes y beneficios en los que se tienen en cuenta tanto el error de tipo I como el de tipo II, razón por la que incorporan la cantidad C_3 en la definición de la función objetivo $V(z^*)$. Ver, por ejemplo, los trabajos de Koh (1992) y Darayseh y Waples (2010) hechos en el contexto de la predicción de la quiebra corporativa. En esta memoria, en cambio, se ha trabajado bajo supuestos que permiten definir la función de costes, $V(z^*)$, de manera diferente:

H 2.3 *El coste debido a la no concesión de la tarjeta a un cliente mal clasificado como moroso es insignificante, y por tanto,* $C_3 = 0$.

Si se tienen en cuenta las pérdidas debidas a la morosidad de las tarjetas de crédito y los gastos de recobro en los que deben incurrir las entidades financieras para recuperar el capital adeudado por los clientes morosos, la cantidad C_3 resultará despreciable con respecto a C_1 (coste asociado al riesgo de crédito de la tarjeta).

H 2.4 *Las cantidades C_1 y C_2 son constantes positivas que no dependen de las covariables del modelo de puntuación.*

En esta memoria se ha tomado como referencia el modelo utilizado por la entidad colaboradora, donde las cantidades C_1 y C_2 fueron constantes dadas. Sin embargo, es importante comentar que el modelo estudiado puede ser generalizado permitiendo que ambas cantidades dependan, por ejemplo, de la puntuación, Z, o de otras variables.

Al formular la estrategia óptima de concesión surge un conjunto de probabilidades que dependen del valor z^* fijado para cada dictamen. Dichas probabilidades pueden representarse en términos de la distribución conjunta del vector (Z, Y). La distribución conjunta de (Z, Y) representada en forma de matriz viene dada en la Tabla 2.24

Tabla 2.24. Matriz de probabilidades de clasificación

Puntuación	créditos morosos	créditos no morosos
$Z > z^*$	P_{11}	P_{10}
$Z \leq z^*$	P_{01}	P_{00}

De la Tabla 2.24 se obtiene que el coste esperado en función del punto de corte, z^*, viene dado por:

$$V(z^*) = E\left(U(Z, z^*)\right) = C_1 P_{01}(z^*) - C_2 P_{00}(z^*),$$

donde se han definido las siguientes probabilidades:

$$P_{0j}(z^*) = P\left(Z \leq z^*, Y = j\right) \tag{2.22}$$

y

$$P_{1j}(z^*) = P\left(Z > z^*, Y = j\right). \tag{2.23}$$

Reemplazando las funciones $P_{0j}(z^*)$, para $j = 0$ y $j = 1$, por sus estimadores muestrales, $\hat{P}_{0j}(z^*)$, se obtiene que la función empírica del coste esperado viene dada por:

$$\hat{V}(z^*) = C_1 \hat{P}_{01}\left(z^*\right) - C_2 \hat{P}_{00}\left(z^*\right). \tag{2.24}$$

Por conveniencia, en lo sucesivo se considerará que las n primeras realizaciones del vector aleatorio (Z, Y), dadas por la muestra $(z_1, y_1), \ldots, (z_n, y_n)$, corresponden a las solicitudes concedidas y que las $m - n$ realizaciones restantes, $(z_{n+1}, y_{n+1}), \ldots, (z_m, y_m)$, corresponden a las solicitudes denegadas. Así, los estimadores empíricos de las probabilidades definidas en (2.22) y en (2.23) vienen dados por:

$$\hat{P}_{00}(z^*) = \frac{1}{m}\left\{\sum_{i=1}^{n}(1-y_i)\,I\,(z_i \leq z^*) + \sum_{i=n+1}^{m}(1-\widehat{y}_i)\,I\,(z_i \leq z^*)\right\}$$

y

$$\hat{P}_{01}(z^*) = \frac{1}{m}\left\{\sum_{i=1}^{n} y_i I\,(z_i \leq z^*) + \sum_{i=n+1}^{m}\widehat{y}_i I\,(z_i \leq z^*)\right\},$$

donde $I_{(A)}$ es la indicadora del suceso A y los valores de la variable indicadora de morosidad, Y, se estiman en los créditos denegados usando la fórmula (2.7), esto es, utlizando las probabilidades de mora proporcionadas por el modelo ajustado, se obtiene:

$$\widehat{y}_i = \pi\,(\hat{z}_i) = \frac{e^{\hat{z}_i}}{1+e^{\hat{z}_i}}, \quad i = n+1, ..., m.$$

Sean C_1^A y C_2^A los valores que toman las constantes C_1 y C_2 de la Tabla 2.23 bajo el escenario económico A, y C_1^B y C_2^B los valores que toman bajo el escenario económico B. Supóngase además, que A es un escenario económico más favorable para la entidad financiera que el escenario B. Así, la estrategia óptima para la concesión de tarjetas se obtiene resolviendo el siguiente par de ecuaciones:

(a) Bajo el escenario económico A, calcular:

$$z_A^* = \arg\min_{z^* \in \mathbb{R}} \hat{V}_A(z^*), \tag{2.25}$$

siendo $\hat{V}_A(z^*) = C_1^A \hat{P}_{01}(z^*) - C_2^A \hat{P}_{00}(z^*)$.

La solución de la ecuación (2.25) se utiliza para decidir si *conceder* o *consultar* la solicitud de la tarjeta tomando como valor de frontera el punto $F_c = z_A^*$ (o, en escala de probabilidad, $F_c = \pi\,(z_A^*)$).

(b) Bajo el escenario económico B, calcular:

$$z_B^* = \arg\min_{z^* \in \mathbb{R}} \hat{V}_B(z^*), \tag{2.26}$$

siendo $\hat{V}_B(z^*) = C_1^B \hat{P}_{01}(z^*) - C_2^B \hat{P}_{00}(z^*)$.

La solución de la ecuación (2.26) se utiliza para decidir si *consultar* o *denegar* la solicitud de la tarjeta tomando como valor de frontera el punto $F_d = z_B^*$ (o, en escala de probabilidad, $F_d = \pi\left(z_B^*\right)$).

Los resultados obtenidos a partir de los datos proporcionados por la entidad colaboradora se resumen en la siguiente tabla:

Tabla 2.25. Límites de la frontera de concesión óptima para tarjetas de crédito

	Frontera	F_c	F_d
Escala	Puntuación Z	-3.46087	-2.71139
	$\widehat{PD}$	0.0304463	0.062305

La Tabla 2.25 muestra los puntos de corte obtenidos resolviendo las ecuaciones (2.25) y (2.26). Los costes utilizados en dichas ecuaciones fueron predeterminados por la entidad colaboradora y están medidos en escala porcentual:

- $C_1^A = 1.37$ y $C_1^B = 2.42$
- $C_2^A = 49$ y $C_2^B = 60$

Así, el dictamen de concesión en función de la puntuación Z viene dada por:

$$DC(z) = \begin{cases} \text{conceder} & \text{si} & z \leq \text{-3.46087} \\ \text{consultar} & \text{si} & \text{-3.46087} < z \leq \text{-2.71139} \\ \text{denegar} & \text{si} & z > \text{-2.71139} \end{cases}$$

o, equivalentemente, en función de la PD estimada:

$$DC\left(\widehat{PD}\right) = \begin{cases} \text{conceder} & \text{si} & \widehat{PD} \leq 0.0304463 \\ \text{consultar} & \text{si} & 0.0304463 < \widehat{PD} \leq 0.062305 \\ \text{denegar} & \text{si} & \widehat{PD} > 0.062305 \end{cases} \tag{2.27}$$

Esta metodología de cálculo del DC otorga a la entidades financieras una estrategia óptima para la concesión de las tarjetas de crédito determinada empíricamente a partir de las soluciones de las ecuaciones (2.25) y (2.26).

Así, si el modelo de puntuación del que se obtienen los valores de Z posee una capacidad de discriminación suficientemente alta, el riesgo de conceder la tarjeta a un cliente propenso a la morosidad será mínimo mientras que el beneficio que se obtendrá por conceder la tarjeta a un cliente solvente será máximo.

2.4.7. Cálculo del límite de la tarjeta de crédito

En este apartado se trata el problema del cálculo del límite de la tarjeta de crédito. Este procedimiento se realiza una vez que se conoce el dictamen de concesión del crédito (fórmula (2.27)) y constituye un requisito indispensable para la formalización del contrato comercial entre la entidad financiera y el solicitante.

Para calcular el límite comercial de la tarjeta, las entidades de crédito utilizan indicadores de solvencia que se construyen a partir de características propias de los clientes, como sus ingresos y sus gastos mensuales, su grado de endeudamiento, su patrimonio y su propensión a la morosidad, medida por su probabilidad de insolvencia (PD), entre otras. También utilizan características que afectan a la población en su conjunto, como las condiciones del mercado de créditos (reflejadas, por ejemplo, en los tipos de interés) y los índices macroeconómicos, entre otras fuentes de información. Siguiendo el modelo utilizado por la entidad colaboradora, existen dos clases de indicadores de solvencia que se utilizan para calcular el límite de crédito asignable a la tarjeta, los que determinan el balance de caja mensual del cliente y los que determinan sus ingresos mensuales. En este capítulo se exponen dos de estos indicadores, el *factor de balance de caja*, denotado por F^{BC}, y el *factor de ingreso mensual*, denotado por F^{IM}.

Durante la investigación que dió lugar a esta memoria, después de conocer la metodología empleada por la entidad colaboradora, se pudo verificar que los índices F^{BC} y F^{IM} se construyen a partir de otros indicadores financieros que, en algunos casos, requieren necesariamente de la utilización de tablas con datos externos (estadísticas oficiales) y de una serie de cálculos que harían demasiado extenso el presente capítulo, por lo que no serán tratados aquí. Por este motivo, a continuación se exponen sólo las fórmulas de cálculo de los índices F^{BC} y F^{IM} sin profundizar en el estudio de sus componentes.

El límite de la tarjeta de crédito, denotado por λ, se define como:

$$\lambda = comp^{BC} + comp^{IM}, \tag{2.28}$$

donde $comp^{BC}$ es la parte del crédito debida al *balance de caja* del solicitante mientras que $comp^{IM}$ es la parte del crédito debida a sus *ingresos mensuales*.

(a) *Componente de balance de caja* La componente de balance de caja, $comp^{BC}$, corresponde a la parte del límite de crédito que la entidad financiera otorga al cliente tomando como base de solvencia exclusivamente su balance de caja mensual. La $comp^{BC}$ se obtiene a partir del coeficiente de balance de caja, BC, del índice F^{BC}, de la PD estimada ($\widehat{PD}$) y del punto de corte F_d obtenido de la ecuación (2.26). Por otro lado, el índice F^{BC} se obtiene por medio de una recta de interpolación lineal entre dos valores predeterminados por la entidad: el coeficiente BC mínimo ($BC_{min} \geq 0$) y el coeficiente BC máximo ($BC_{max} > 0$). Si el dictamen es *conceder* o *consultar*, el valor del índice F^{BC} viene dado por la fórmula:

$$F^{BC}\left(\widehat{PD}\right) = BC_{max} - \frac{\widehat{PD}}{F_d}\left(BC_{max} - BC_{min}\right).$$

Llamando IM al coeficiente de ingresos mensuales, GM al coeficiente de gastos mensuales y FM al factor monetario correspondiente al lugar de residencia del cliente[4], la cantidad $comp^{BC}$ del límite de crédito λ viene dada por:

$$comp^{BC} = F^{BC}\left(\widehat{PD}\right) mod^{BC},$$

donde $mod^{BC} \equiv g_1(IM, GM, FM)$, siendo g_1, por ejemplo, una función lineal segmentada según el FM.

[4]El factor monetario, FM, es un coeficiente numérico construido a partir de estadísticas oficiales del ingreso per cápita por provincia y comunidad autónoma. Se utiliza para estimar el poder adquisitivo de los clientes según la localidad en la que residen.

(b) ***Componente de ingreso mensual*** La componente de ingreso mensual, $comp^{IM}$, corresponde a la parte del límite de crédito que la entidad financiera otorga al cliente tomando como base de solvencia exclusivamente sus ingresos mensuales. Para obtener la $comp^{IM}$ se requiere conocer el coeficiente IM, el valor del índice F^{IM}, el valor de la $\widehat{PD}$ y el punto de corte F_d. El valor del índice F^{IM} se obtiene de forma análoga al índice F^{BC}, es decir, por medio de una recta de interpolación lineal entre dos valores predeterminados por la entidad financiera: el coeficiente IM mínimo (IM_{min}) y el coeficiente IM máximo (IM_{max}). Si el dictamen de concesión es *conceder* o *consultar*, el valor del índice F^{IM} viene dado por la fórmula:

$$F^{IM}\left(\widehat{PD}\right) = IM_{max} - \frac{\widehat{PD}}{F_d}\left(IM_{max} - IM_{min}\right).$$

Así, la cantidad $comp^{IM}$ del límite de crédito λ viene dada por:

$$comp^{IM} = F^{IM}\left(\widehat{PD}\right) mod^{IM},$$

donde $mod^{IM} \equiv g_2(IM, GM)$, siendo g_2, por ejemplo, una función de interpolación lineal o cuadrática.

(c) ***Límite comercial de la tarjeta*** Una vez aprobada la tarjeta de crédito, la entidad debe determinar el límite de crédito que asignará al contrato comercial con el cliente. Por cómo está definido el límite de crédito λ (ver fórmula (2.28)), éste puede tomar valores que superan el máximo recomendado para este tipo de operaciones provocando un aumento indebido del riesgo de exposición asociado a la concesión del crédito. Por este motivo, la entidad colaboradora utiliza un mecanismo que le permite mitigar dicho riesgo acotando el límite de crédito por el que finalmente se formalizará el contrato comercial con el cliente. El mecanismo consiste en definir una función que permita modular o acotar la asignación de crédito que contratará el cliente tomando como argumento el valor de λ, lo que se conoce con el nombre de *límite comercial de la tarjeta crédito*, denotado por LCT. Tomando como ejemplo a la entidad colaboradora, el $LCT \equiv LCT(\lambda)$ y se define como una función positiva, continua y creciente en λ, acotada inferior y superiormente por dos valores predefinidos por la entidad: el límite comercial mínimo, m_c, y el límite comercial máximo, M_c.

Un ejemplo de la función $LCT(\lambda)$ es el que se define a continuación. Dados los valores de $m_c = 1\,000$€ y $M_c = 20\,000$€ con saltos de 250€, se define el límite comercial de la tarjeta como:

$$LCT(\lambda) \begin{cases} 0 & \text{si} \quad \lambda < 1\,000 \\ 1\,000 + 250 \left\lceil \frac{\lambda - 1\,000}{250} \right\rceil & \text{si} \quad 1\,000 \leq \lambda \leq 20\,000 \\ 20\,000 & \text{si} \quad \lambda > 20\,000 \end{cases} \tag{2.29}$$

donde $\lceil x \rceil$ es la función parte entera de x.

De la fórmula (2.29) se obtiene que el límite comercial de una tarjeta crédito irá desde los 1 000€ hasta los 20 000€ como máximo, aumentando en 250€ cada vez que el cociente $\frac{\lambda - 1\,000}{250}$ alcanza o supera un número entero.

2.5. Cálculo de la tasa de mora a partir de la puntuación crediticia

Uno de los principales problemas asociados a la concesión de tarjetas de crédito es poder determinar de manera anticipada la tasa de mora esperada de la cartera. La tasa de mora, TM, se define como la proporción de créditos morosos en una cartera de tamaño n observada en un período de tiempo determinado y es el principal parámetro a tener en cuenta en el análisis de la solvencia de la misma, por lo que es fundamental contar con mecanismos que permitan obtener estimaciones precisas de esta cantidad. La técnica que se propone a continuación permite estimar la TM a partir de la media de las probabilidades condicionales de mora (PD) asociadas a las tarjetas de crédito.

A continuación, se exponen dos mecanismos de estimación de la TM de la cartera a partir de las puntuaciones obtenidas con los modelos $ENTIDAD$, $MSEG$ y $MPOL$. El primero corresponde al modelo de probabilidad condicional definido en (2.3), válido bajo la hipótesis de un modelo logístico (H2.2). Alternativamente, se utilizan dos estimadores tipo núcleo de la regresión para obtener estimaciones no paramétricas de las curvas de PD, el estimador *local constante* o de *Nadaraya-Watson* (Nadaraya (1964) y Watson (1964)) y el *estimador local lineal* (Fan (1992, 1993), Hastie y Loader (1993)). Los resultados obtenidos son comparados en términos de la suavidad de las curvas de

PD estimadas, del valor estimado de la TM de la cartera y de las medidas de validación expuestas anteriomente (Sección 2.4.4). Algunos trabajos en los que se estudian modelos de regresión no paramétrica con variable dependiente binaria son, entre otros, los debidos a Chu y Cheng (1995), Aragaki y Altman (1997), Signorini (1998), Altman y McGibbon (1998), Signorini y Jones (2004), Frölich (2006), Hazelton (2007) y Okumura (2011).

Hasta ahora, se han utilizado técnicas de estimación no paramétrica de curvas como una herramienta de apoyo visual en el análisis exploratorio de los datos, sin embargo, la utilización de este tipo de técnicas cobrará más relevancia a partir del Capítulo 3, donde se utilizan técnicas de análisis de supervivencia para el estudio de la morosidad con datos censurados siguiendo las ideas presentadas en la Sección 1.2.2, y donde el estudio del suavizado no paramétrico de las curvas de PD con estimadores tipo núcleo se convierte en la parte central de la investigación llevada a cabo en esta memoria.

2.5.1. Tasa de mora esperada de la cartera

De acuerdo con el modelo de puntuación crediticia definido en la Sección 2.2.1, las variables indicadoras de morosidad, Y_i, son condicionalmente independientes dada la puntuación crediticia $Z_i = \psi(\mathbf{X}_i)$ (H2.1). Se denota por $M_n = \sum_{i=1}^{n} Y_i$ la variable número de créditos morosos en una cartera de tamaño n y las variables $Y_i|Z_i \overset{d}{\equiv} Bernoulli(\pi_i)$, donde el parámetro $\pi_i \equiv \pi(z_i) = E(Y_i|Z_i = z_i) = PD(z_i)$ corresponde a la probabilidad condicional de mora del crédito i-ésimo dada la puntuación $Z_i = z_i$, para todo i, con $1 \leq i \leq n$. Entonces, la tasa de mora esperada de la cartera, TM, se define como:

$$\begin{aligned} TM &= \frac{1}{n} E(M_n) \\ &= \frac{1}{n} \sum_{i=1}^{n} E(E(Y_i|Z_i)) \\ &= \frac{1}{n} \sum_{i=1}^{n} E(\pi(Z_i)|Z_i = z_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} PD(z_i), \end{aligned} \tag{2.30}$$

donde, por simplicidad de notación, se ha omitido expresar z_i como función del vector $\boldsymbol{x}_i$, cuya relación se define en la fórmula (2.1) de la Sección 2.2.1.

Aunque el estudio de la variable aleatoria M_n va más allá del alcance de este capítulo, es conocido que su tratamiento es de gran importancia en modelos de riesgo de crédito en los que interesa conocer la distribución exacta del número de créditos morosos en una cartera (Duffie et al. (2007), Hong (2012)) y su relación con la distribución de la pérdida dada la mora (LGD), uno de los parámetros fundamentales del riesgo de crédito (ver Sección 1.1.3).

En este capítulo, la TM se utiliza como medida de validación del modelo de puntuación crediticia, siendo complementaria a las medidas de validación estadística expuestas en la Sección 2.4.4.

La obtención de la TM a partir de las PD condicionadas asociadas a las tarjetas de crédito tiene su justificación en el siguiente planteamiento.

Supóngase que las variables Y, indicadora de morosidad, y $Z \equiv Z(\boldsymbol{X}) = \psi(\boldsymbol{X})$, puntuación crediticia o scoring, están relacionadas de forma que:

$$Y = m(Z) + \varepsilon,$$

donde $m(z) = E(Y \mid Z = z) = PD(z)$ es una función de regresión desconocida y la variable ε es el término residual del modelo. Supóngase además que m y ε verifican que:

H 2.5 *La función $m(\cdot)$ es continua y posee p+1 derivadas continuas definidas en todo el soporte de Z para algún entero $p > 1$.*

H 2.6 *El término residual ε es independiente de la covariable Z y verifica las propiedades:*

i) $E(\varepsilon \mid Z = z) = E(\varepsilon) = 0$

ii) $E(\varepsilon^2 \mid Z = z) = \sigma_\varepsilon^2(z) < \infty$.

Entonces, dada la muestra $(z_1, y_1), \ldots, (z_n, y_n)$ de n realizaciones independientes del vector (Z, Y), la función $m(\cdot)$ puede estimarse a partir de los datos utilizando los mecanismos que se describen a a continuación.

Estimación de la TM vía modelo *logit*

Bajo la hipótesis de un modelo logístico (H2.2), dada la puntuación crediticia z_i, la probabilidad de mora condicional asociada al crédito i-ésimo viene dada por:

$$\widehat{PD}_{logit}\left(z_i\right) = \hat{m}_{logit}(z_i) = \frac{e^{z_i}}{1+e^{z_i}} = \frac{e^{\boldsymbol{x}_i\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}}, \tag{2.31}$$

donde $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip-1}) \in \mathbb{R}^p$ y $\boldsymbol{\beta} = \left(\beta_0, ..., \beta_{p-1}\right)' \in \mathbb{R}^p$ para todo i, con $1 \leq i \leq n$.

Así, a partir de la fórmula (2.31) se define el estimador *logit* de la TM como:

$$\widehat{TM}_{logit} = \frac{1}{n}\sum_{i=1}^{n} \hat{m}_{logit}\left(z_i\right). \tag{2.32}$$

Estimación de la TM con técnicas de regresión locales

Alternativamente a la obtención de la TM a partir de la fórmula (2.32), se considera la estimación de la función de regresión $m(\cdot)$ con técnicas de suavizado no paramétrico tipo núcleo. Para ello se utiliza una de las técnicas más estudiadas en regresión no paramétrica de curvas, la *estimación polinómica local* (Fan y Gijbels (1996)). La idea de la estimación local de la regresión es la siguiente. De acuerdo con H2.5, $m(\cdot)$ posee $p+1$ derivadas continuas en el soporte de Z, en este caso, todo el conjunto $\mathbb{R}$. Entonces, para z_i en una vecindad del punto z, el desarrollo de Taylor de orden p de $m(z_i)$ viene dado por

$$m(z_i) \approx m(z) + m'(z)(z_i - z) + \ldots + \frac{1}{p!}m^{(p)}(z)(z_i - z)^p,$$

para todo i, con $1 \leq i \leq n$. Denotando por $\boldsymbol{B} = (b_0, ..., b_p)' \in \mathbb{R}^{p+1}$ el vector de coeficientes de la regresión local cuyo término $b_j \equiv b_j(z) = \frac{1}{j!}m^{(j)}(z)$, y donde $m^{(j)}(z)$ es la derivada j-ésima de la función $m(\cdot)$, entonces, el estimador polinómico local de $m(z)$, para $z \in \mathbb{R}$, se obtiene resolviendo el siguiente problema de mínimos cuadrados locales

$$Q_{p,h}\left(\boldsymbol{B}\right) = \min_{\boldsymbol{B}\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} \left(y_i - \sum\nolimits_{j=0}^{p} b_j(z_i - z)^j\right)^2 K\left(\frac{z_i - z}{h}\right), \tag{2.33}$$

donde $K(\cdot)$ es una función de densidad de probabilidad simétrica y acotada que recibe el nombre de función núcleo y $h \equiv h(n)$ es el parámetro de suavizado encargado de regular el ancho de la vecindad donde se calcula la estimación no paramétrica local, y por tanto, el grado de suavizamiento de esta.

Estimador *local constante* o de *Nadaraya-Watson* Cuando se ajusta un polinomio local de grado $p = 0$, es decir, una constante local, el problema de mínimos cuadrados locales definido en la ecuación (2.33) se reduce a:

$$Q_{0,h}(b_0) = \min_{b_0 \in \mathbb{R}} \sum_{i=1}^{n} (y_i - b_0)^2 K\left(\frac{z_i - z}{h}\right), \ z \in \mathbb{R},$$

cuya solución $\hat{b}_0 = \hat{m}_{0,h}(z) = \widehat{PD}_{NW}(z)$ viene dada por

$$\begin{aligned} \hat{m}_{0,h}(z) &= \sum_{i=1}^{n} \frac{K_h(z_i - z) y_i}{\sum_{j=1}^{n} K_h(z_j - z)}, \\ &= \sum_{i=1}^{n} w_{ih}(z)\, y_i, \end{aligned} \tag{2.34}$$

donde los términos $w_{ih}(z) = K_h(z_i - z) \left(\sum_{j=1}^{n} K_h(z_j - z)\right)^{-1}$, con $1 \leq i \leq n$, son los pesos no paramétricos de la regresión local y $K_h(u) = (1/h)\, K(u/h)$ es la función núcleo reescalada por h.

Así, a partir de la fórmula (2.34), se define el estimador *local constante* o de *Nadaraya-Watson* (NW) de la TM como:

$$\widehat{TM}_{NW} = \frac{1}{n} \sum_{i=1}^{n} \hat{m}_{0,h}(z_i). \tag{2.35}$$

Estimador por *regresión local lineal* Si el polinomio local ajustado es de grado $p = 1$, es decir, una recta local, el problema de mínimos cuadrados locales definido en la ecuación (2.33) toma la forma:

$$Q_{1,h}(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 (z_i - z))^2 K\left(\frac{z_i - z}{h}\right), z \in \mathbb{R}. \tag{2.36}$$

Resolviendo el problema (2.36), se obtiene como solución $\hat{b}_0 = \hat{m}_{1,h}(z) = \widehat{PD}_{RLL}(z)$, cuya fórmula viene dada por

$$\hat{m}_{1,h}(z) = \frac{\sum_{i=1}^{n} \omega_{ih}(z)\, y_i}{\sum_{i=1}^{n} \omega_{ih}(z)}, \tag{2.37}$$

donde los términos $\omega_{ih}(z) = K_h(z_i - z)\,(s_{n,2} - (z_i - z)s_{n,1})$ son los pesos no paramétricos de la regresión local lineal y los términos, $s_{n,j}$, vienen dados por

$$s_{n,j} = \sum_{i=1}^{n} (z_i - z)^j K_h(z_i - z),\ z \in \mathbb{R},$$

con $j = 1, 2$.

Así, a partir de la fórmula (2.37), se define el estimador de la TM por *regresión local lineal* como:

$$\widehat{TM}_{RLL} = \frac{1}{n} \sum_{i=1}^{n} \hat{m}_{1,h}(z_i). \tag{2.38}$$

Las técnicas utilizadas para determinar el parámetro h en (2.34) y en (2.37) se describen en el apartado siguiente.

2.5.2. Resultados obtenidos con la muestra de validación

En este apartado se presentan los resultados de las estimaciones no paramé-tricas de las curvas de probabilidad de mora, $PD(z) = m(z)$, obtenidas con los tres estimadores descritos en el apartado anterior, $\hat{m}_{logit}$, $\hat{m}_{0,h}$ y $\hat{m}_{1,h}$. Los datos utilizados corresponden a la muestra de validación utilizada con anterioridad en la Sección 2.4.5. La muestra la componen $n_1 = 204$ tarjetas de crédito morosas y $n_0 = 4\,796$ tarjetas no morosas, lo que equivale a una cartera de $n = 5\,000$ créditos con una tasa de morosidad observada de un $4.08\,\%$. Las puntuaciones crediticias obtenidas con los modelos estudiados en este capítulo son utilizados como covariables en los modelos de regresión ajustados. Además, por tratarse de un estudio de validación, las estimaciones no paramétricas se calculan utilizando los datos de la muestra de entrenamiento, es decir, con $n = 20\,000$ créditos, mientras que los valores del parámetro de suavizado, h, fueron calculados con la muestra de validación, es decir con $n = 5\,000$ créditos.

Curvas de PD estimadas con métodos de regresión local

Los resultados que se exponen a continuación se obtuvieron empleando diferentes valores del parámetro de suavizado, h, calculados con distintos métodos de selección. La elección del parámetro de suavizado, h, es crucial en el resultado final de la estimación no paramétrica debido a que este se encarga de controlar el grado de suavizamiento de la curva estimada. Si el valor de h es demasiado grande, se obtiene como resultado un aumento del sesgo de la estimación, provocando un efecto de exceso de suavizamiento (sobresuavizado) de la curva resultante. Por el contrario, si el valor de h es demasiado pequeño, se obtiene como resultado un aumento no deseado de la variabilidad de la estimación, lo que se traduce en una curva infrasuavisada (poco suave). El estudio de este problema ha motivado que la búsqueda de técnicas de selección automática del parámetro de suavizado óptimo se mantenga vigente desde los años 80 del siglo XX hasta nuestros días.

Aunque existe abundante literatura dedicada a los métodos de selección de la ventana de suavizado, h, se verifica que la mayoría de las técnicas existentes pueden agruparse en tres grandes bloques: los métodos de validación cruzada, los métodos basados en sustitución directa o *plug-in* y los métodos basados en remuestreo o *bootstrap*. En esta sección se utilizan cuatro selectores automáticos del parámetro h. El primer selector automático corresponde a la regla del pulgar o *rule-of-thumb* (rot) de Fan y Gijbels (1996), implementado en la función *thumbBw* de la librería *locpol* de R (Ojeda Cabrera (2012)). El segundo selector corresponde al método de validación cruzada *leave-one-out* ($loo\text{-}cv$) de Härdle y Marron (1985). El tercer selector es el método de validación cruzada basado en el *criterio de información de Akaike correjido* ($AICc\text{-}cv$) de Hurvich et al. (1998), implementado en la función *npregbw* de la librería *np* de R (Hayfield y Racine (2008)). Finalmente, el cuarto selector automático corresponde al método *plug-in directo* (dpi) de Ruppert et al. (1995), implementado en la función *dpill* de la librería *KernSmooth* de R (Wand y Ripley (2008)).

Estimador de la curva de PD de *Nadaraya-Watson* (NW)

En las figuras siguientes se muestran las estimaciones no paramétricas de las curvas de PD de las tarjetas de crédito obtenidas con el estimador de *Nadaraya-Watson,* $\hat{m}_{0,h}(z)$. Las curvas allí representadas se obtienen con

distintos valores de h para cada uno de los modelos de puntuación estudiados. Los valores del parámetro de suavizado, h, obtenidos para los tres modelos de puntuaciones se muestran en la Tabla 2.26.

Tabla 2.26. Selectores del parámetro de suavizado del estimador de NW de la PD

Selector de h	$ENTIDAD$	$MSEG$	$MPOL$
rot	0.1623	0.0959	0.0962
$loo\text{-}cv$	0.4925	0.3960	0.4826
$AICc\text{-}cv$	0.4116	0.3082	0.3652
dpi	0.5437	0.4930	0.2645

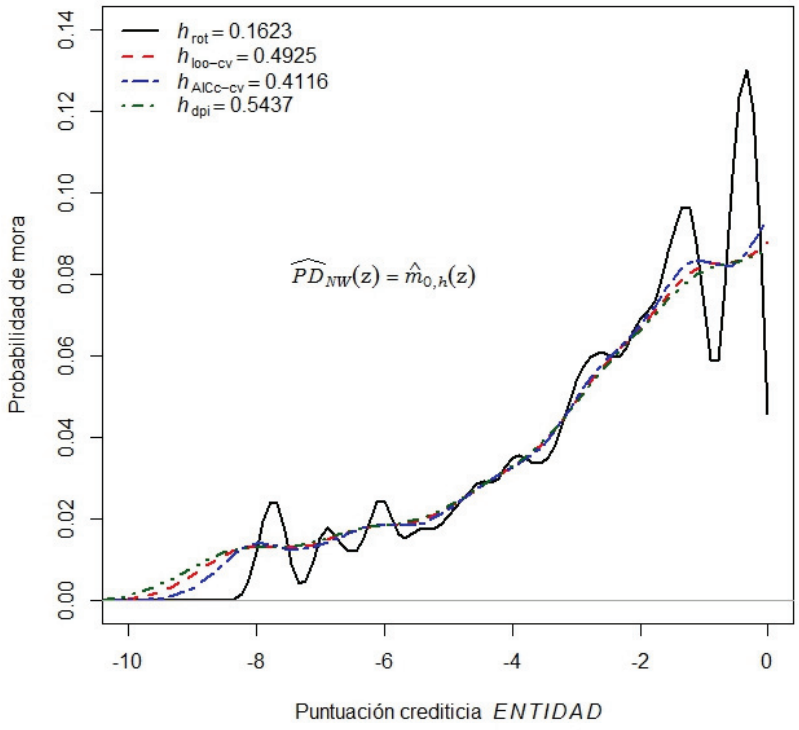

Figura 2.29 Curvas de $PD_{ENTIDAD}$ obtenidas con el estimador de *Nadaraya-Watson*.

La Figura 2.29 muestra las estimaciones de la curva de PD obtenidas con los valores de h ofrecidos en la Tabla 2.26 para las puntuaciones del modelo $ENTIDAD$. Se observa que las curvas obtenidas con los valores de h calculados con los selectores automáticos $h_{loo\text{-}cv} = 0.4925$ (trazo de color rojo), $h_{AICc\text{-}cv} = 0.4116$ (trazo de color azul) y $h_{dpi} = 0.5437$ (trazo de color verde), exhiben un buen nivel de suavizamiento y de ajuste en los extremos. En cambio, al utilizar la ventana $h_{rot} = 0.1623$ (línea de color negro), se obtienen estimaciones demasiado variables dando como resultado una curva infrasuavizada y con problemas de falta de ajuste en los extremos del rango de estimación. En base a estos resultados, es posible concluir que los valores adecuados que puede tomar la ventana de suavizado del modelo $ENTIDAD$, denotada por $h^0_{ENTIDAD}$, se encuentran en el rango $0.40 \leq h \leq 0.55$.

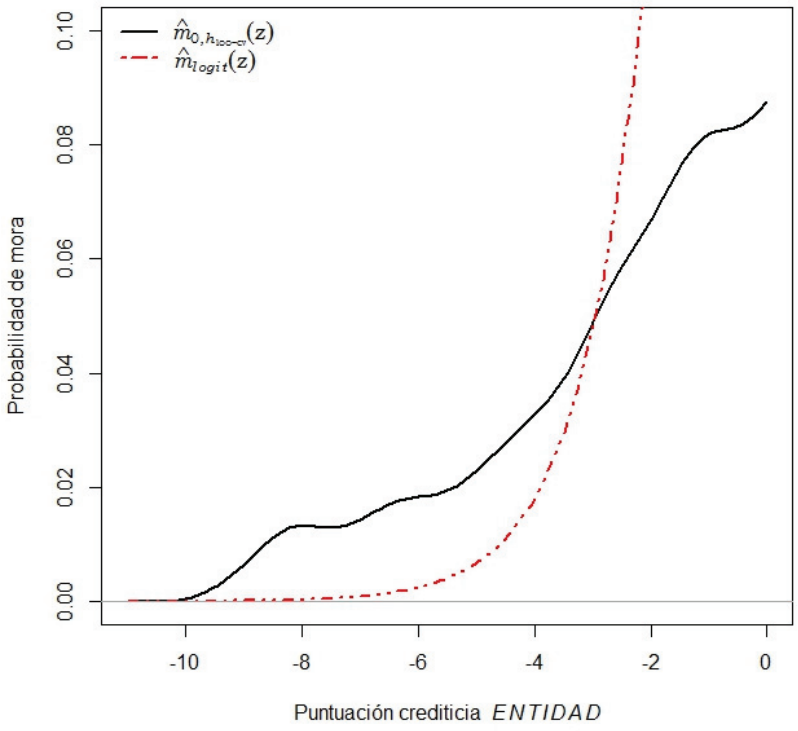

Figura 2.30 Curvas estimadas de $PD_{ENTIDAD}$. Estimador de NW vs. estimador $logit$.

En la Figura 2.30, se representan las curvas de PD para el modelo $ENTIDAD$ obtenidas con los estimadores $\hat{m}_{logit}$ y $\hat{m}_{0,h^0_{ENTIDAD}}$, donde la ventana seleccionada es $h^0_{ENTIDAD} = h_{loo\text{-}cv} = 0.4925$. Se observa que las diferencias entre ambas estimaciones es muy notable, donde la curva obtenida con el estimador no paramétrico, $\hat{m}_{0,h_{loo\text{-}cv}}$, deja de poder calcularse cuando los puntos de evaluación alcanzan el máximo muestral de las puntuaciones, en torno al punto $z = -0.062$. Este problema es consecuencia del conocido efecto frontera (en inglés *boundary bias*), el cual se sabe, por la literatura, que afecta a estimadores tipo núcleo como el de NW. Ver, por ejemplo, el estudio de este problema en los trabajos de Chu y Marron (1991) y de Fan y Gijbels (1996, págs. 17-18), entre otros. En el extremo inferior, en cambio, se ve que la curva decrece a cero más suavemente y es prácticamente cero para valores de puntuación $z < -10$.

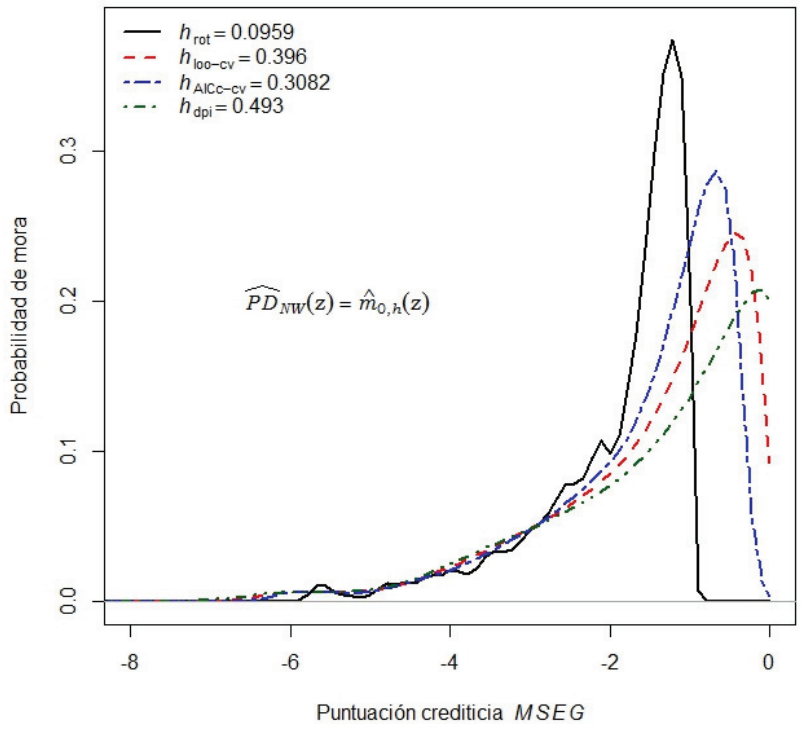

Figura 2.31 Curvas de PD_{MSEG} obtenidas con el estimador de *Nadaraya-Watson*.

En la Figura 2.31 se muestran las estimaciones de la curva de PD obtenidas con las puntuaciones del modelo $MSEG$. Allí se observa que las curvas obtenidas con las ventanas $h_{AICc\text{-}cv} = 0.3082$ (trazo de color azul), $h_{loo\text{-}cv} = 0.396$ (trazo de color rojo) y $h_{dpi} = 0.4930$ (trazo de color verde) exhiben un buen grado de suavizamiento aunque muestran signos de verse afectadas por el efecto frontera en puntos cercanos al máximo muestral de las puntuaciones. Sin embargo, cuando la curva no paramétrica $\hat{m}_{0,h}(z)$ es calculada utilizando la ventana $h_{rot} = 0.0959$ (trazo de color negro), el nivel de suavizamiento obtenido empeora considerablemente y el efecto frontera se vuelve más acusado en los extremos del rango de estimación, tal y como se aprecia en el trazado de la curva de color negro. Como consecuencia, se obtiene que los valores adecuados que puede tomar la ventana de suavizado del modelo $MSEG$, denotada por h^0_{MSEG}, se encuentran en el rango $0.30 < h \leq 0.5$.

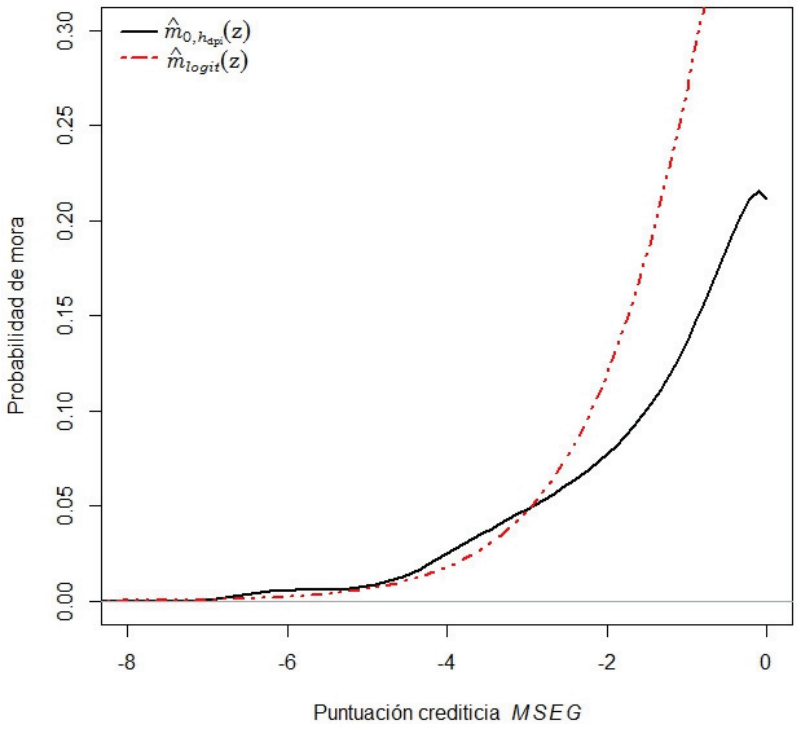

Figura 2.32 Curvas estimadas de PD_{MSEG}. Estimador de NW vs. estimador $logit$.

En la Figura 2.32 se comparan las curvas de PD obtenidas con los estimadores $\hat{m}_{logit}$ y $\hat{m}_{0,h^0_{MSEG}}$, donde la ventana seleccionada es $h^0_{MSEG} = h_{dpi} = 0.493$. Allí se observa que, a partir de la intersección entre ambas curvas (aproximadamente en el punto $z_0 = -3$), se obtiene que $\hat{m}_{0,h_{dpi}}(z) < \hat{m}_{logit}(z)$ para todo $z > z_0$. También se observa que el efecto frontera es más acusado en zonas próximas al extremo superior del rango muestral, lo que impide que la curva $\hat{m}_{0,h_{loo\text{-}cv}}(z)$ pueda calcularse más allá del máximo muestral de la puntuación $MSEG$.

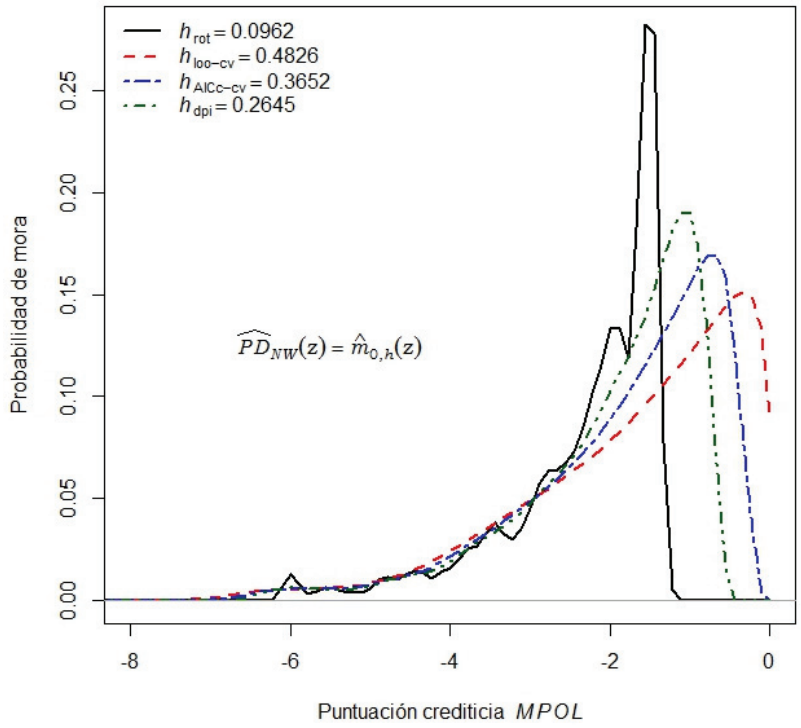

Figura 2.33 Curvas de PD_{MPOL} obtenidas con el estimador de *Nadaraya-Watson*.

En la Figura 2.33 se muestran las estimaciones de la curva de PD obtenidas con las puntuaciones del modelo $MPOL$. En ella se observan curvas con un buen grado de suavizamiento obtenidas con las ventanas $h_{loo\text{-}cv} = 0.4826$ (trazo de color rojo), $h_{AICc\text{-}cv} = 0.3652$ (trazo de color azul) y $h_{dpi} = 0.2645$ (trazo

de color verde). Ademas, se observa que al utilizar la ventana $h_{rot} = 0.0962$, se obtiene una curva con alta variabilidad y con falta de ajuste en los extremos, tal y como se aprecia en la curva dibujada con trazo de color negro. Como consecuencia, se obtiene que los valores adecuados que puede tomar la ventana de suavizado del modelo $MPOL$, denotada como h^0_{MPOL}, se encuentran en el rango $0.26 < h \leq 0.50$.

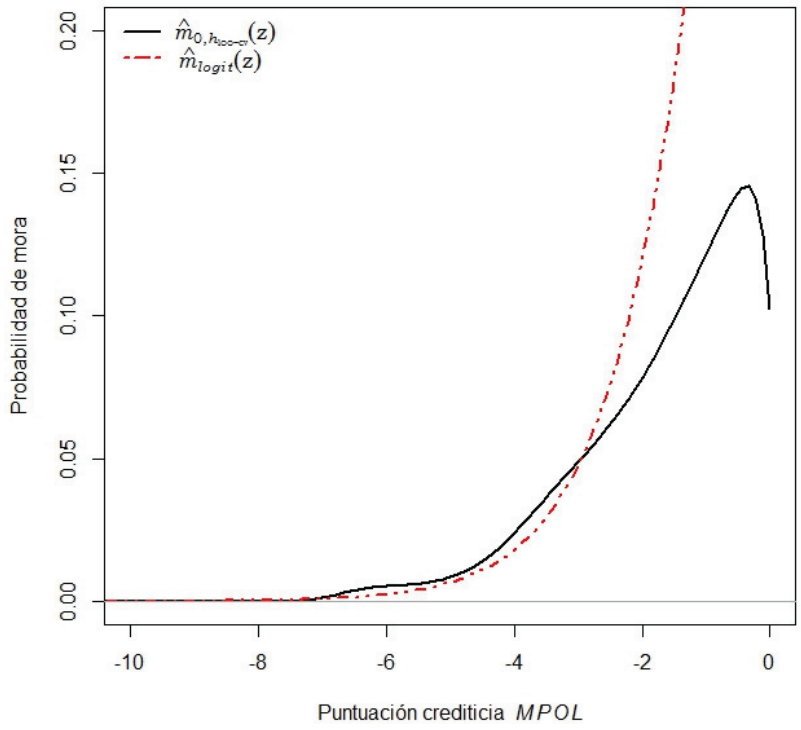

Figura 2.34 Curvas estimadas de PD_{MPOL}. Estimador de NW vs. estimador $logit$.

En la Figura 2.34 se comparan las curvas de PD obtenidas con los estimadores $\hat{m}_{logit}$ y $\hat{m}_{0,h^0_{MPOL}}$, donde la ventana seleccionada es $h^0_{MPOL} = h_{loo\text{-}cv} = 0.4826$. Se observa que el efecto frontera afecta a la estimación con las puntuaciones $MPOL$ de forma similar a lo que sucede con las puntuaciones del modelo $MSEG$, es decir, que la curva $\hat{m}_{0,h_{loo\text{-}cv}}(z)$ decrece a cero en zonas en que la puntuación está próxima a los extremos muestrales.

Estimador de la curva de PD vía regresión local lineal (RLL)

A continuación, en las figuras siguientes, se muestran los resutados de las estimaciones no paramétricas de las curvas de PD obtenidas con el estimador de la *regresión local lineal* (RLL), $\hat{m}_{1,h}(z)$. Los valores de las ventanas de suavizado del estimador $\hat{m}_{1,h}(z)$ obtenidos para los modelos de puntuaciones estudiados se muestran a continuación, en la Tabla 2.27.

Tabla 2.27. Selectores del parámetro de suavizado del estimador de RLL de la PD

Selector de h	*ENTIDAD*	*MSEG*	*MPOL*
rot	0.4595	0.3505	0.2645
loo-cv	0.3750	1.1596	0.8132
AICc-cv	1.5300	0.6657	0.7860
dpi	0.6279	0.3909	0.2858

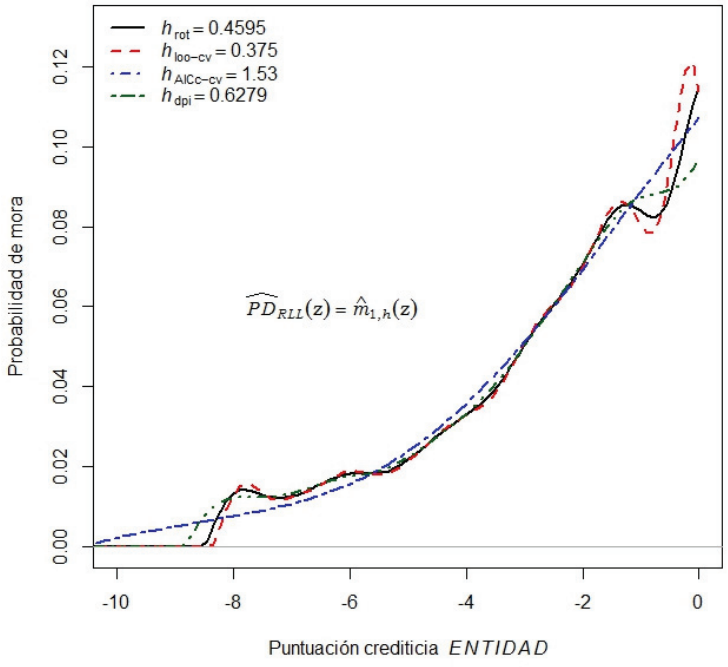

Figura 2.35 Curvas de $PD_{ENTIDAD}$ obtenidas con el estimador de RLL.

En la Figura 2.35 se muestran las estimaciones de la curva de PD obtenidas con los valores de h determinados para las puntuaciones del modelo $ENTIDAD$. Allí se observa que la curva obtenida con la ventana $h_{dpi} = 0.6279$ (trazo de color verde) ofrece un buen grado de suavizamiento. También se observa que las curvas calculadas con las ventanas $h_{rot} = 0.4595$ y $h_{loo\text{-}cv} = 0.375$ (trazos de color negro y rojo, respectivamente) muestran un grado de suavizamiento aceptable, aunque también se observa la presencia del efecto frontera en los extremos del rango de la estimación. Por último, la curva calculada con la ventana $h_{AICc} = 1.53$ (trazo de color azul) ofrece una curva sobresuavizada, y por tanto, poco adecuada para estimar la curva de $PD_{ENTIDAD}$. En consecuencia, los valores adecuados que puede tomar la ventana de suavizado del modelo $ENTIDAD$, denotada por $h^1_{ENTIDAD}$, se encuentran en el rango $0.46 \leq h \leq 0.63$.

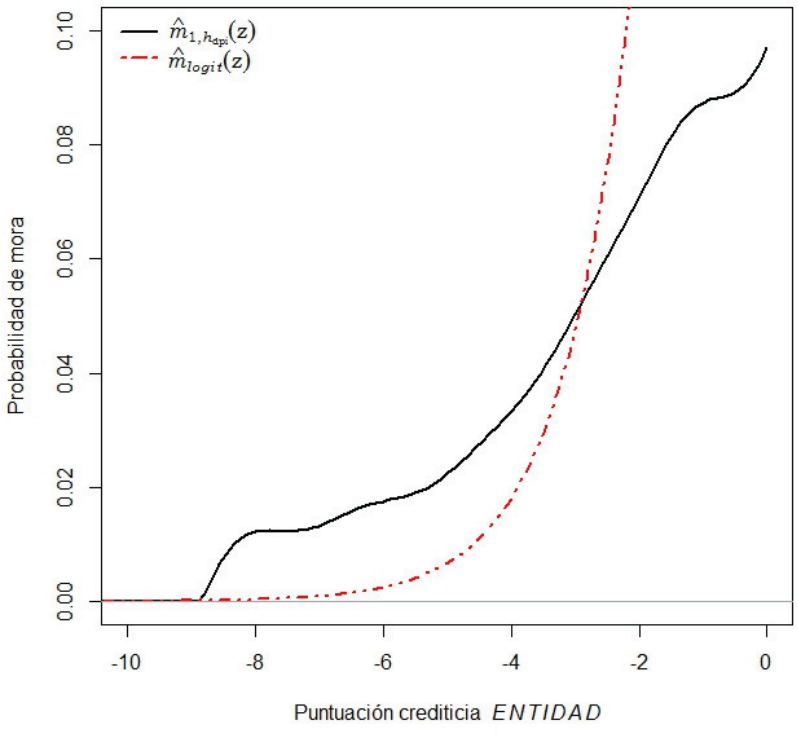

Figura 2.36 Curvas estimadas de $PD_{ENTIDAD}$. Estimador de RLL vs. estimador $logit$.

En la Figura 2.36, se representan las curvas de probabilidad de mora para el modelo $ENTIDAD$ obtenidas con los estimadores $\hat{m}_{logit}$ y $\hat{m}_{1,h^1_{ENTIDAD}}$, donde la ventana seleccionada es $h^1_{ENTIDAD} = h_{dpi} = 0.6279$. Se observa que el resultado obtenido con el estimador de RLL no difiere sustancialmente de lo obtenido con el estimador de NW (ver Figura 2.30). Visualmente, ambas curvas ofrecen un grado de suavizamiento similar, aunque el estimador $\hat{m}_{1,h^1_{ENTIDAD}}$ requiere de un ancho de ventana mayor para conseguirlo. Además, cuando la curva es evaluada en puntos próximos al mínimo muestral, en torno a $z = -8.517$, esta decrece a cero más abruptamente que la curva análoga calculada con el estimador de NW, es decir, con la curva $\hat{m}_{0,h^0_{ENTIDAD}}(z)$.

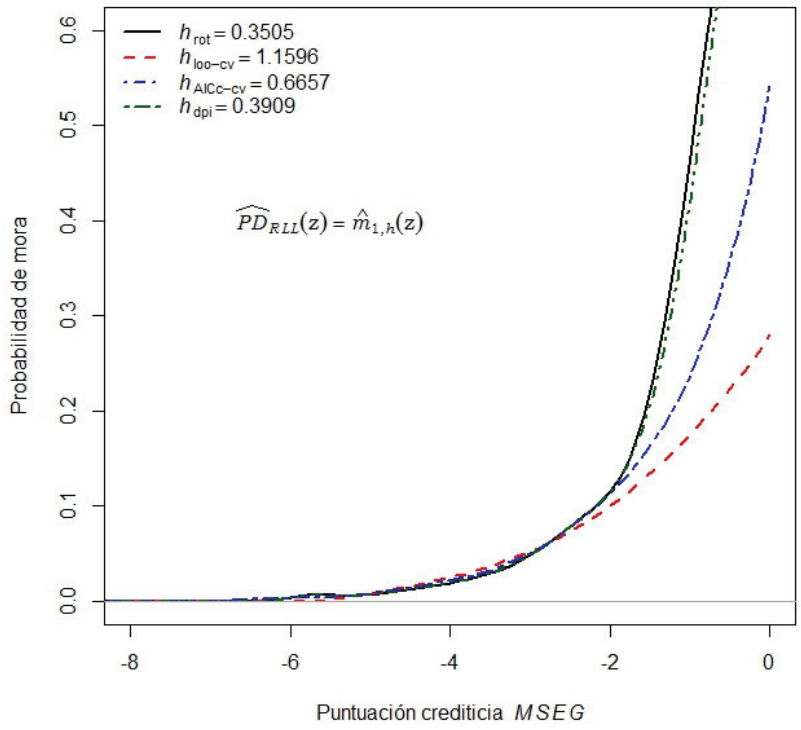

Figura 2.37 Curvas de PD_{MSEG} obtenidas con el estimador de RLL.

La Figura 2.37 muestra las estimaciones de la curva de PD obtenidas con las puntuaciones del modelo $MSEG$. Se observa que las estimaciones

obtenidas con tres de las cuatro ventanas de suavizado calculadas, $h_{rot} = 0.3505$ (línea de color negro), $h_{dpi} = 0.3909$ (trazo de color verde) y $h_{AICc\text{-}cv} = 0.6657$ (trazo de color azul), muestran un buen grado de suavizamiento de las curvas de PD, alcanzando niveles de PD cada vez menores a medida que crece el ancho de la ventana de suavizado. Sin embargo, la curva obtenida con la ventana $h_{loo\text{-}cv} = 1.1596$ presenta un exceso de suavizado, lo que lleva a pensar que $\hat{m}_{1,h_{loo\text{-}cv}}(z)$ no es un estimador adecuado de la curva de PD_{MSEG}.

Por otra parte, se observa que las curvas obtenidas con el estimador de RLL no parecen dar muestras del efecto frontera, como sí ocurre con el estimador de NW de la PD_{MSEG}. Como consecuencia, los valores adecuados que puede tomar la ventana de suavizado del modelo $MSEG$, denotada como h^1_{MSEG}, se encuentran en el rango $0.35 < h \leq 0.67$.

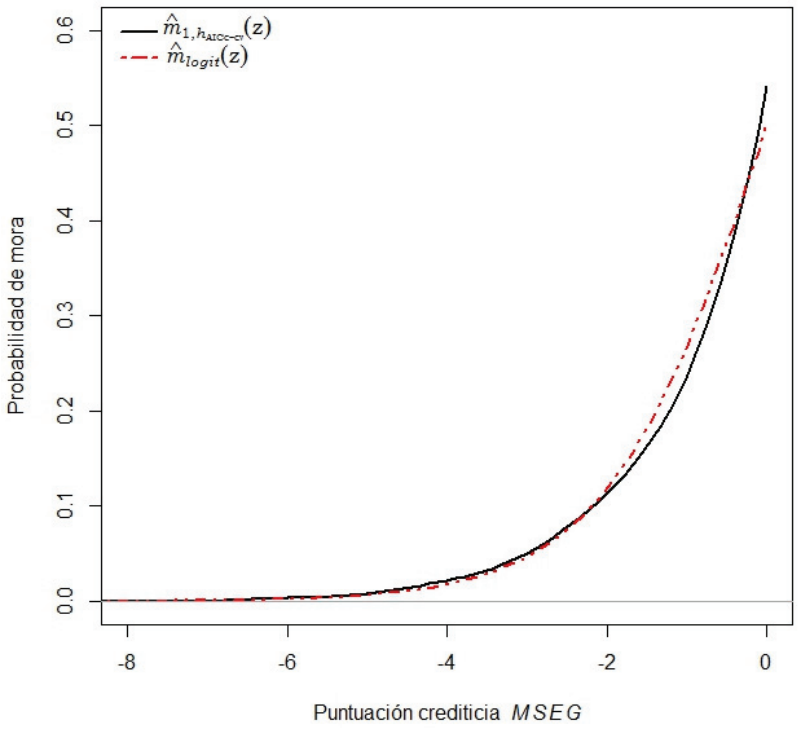

Figura 2.38 Curvas estimadas de PD_{MSEG}. Estimador de RLL vs. estimador $logit$.

En la Figura 2.38 se comparan las curvas de PD obtenidas con los estimadores $\hat{m}_{logit}$ y $\hat{m}_{1,h^1_{MSEG}}$, donde la ventana seleccionada es $h^1_{MSEG} = h_{AICc\text{-}cv} = 0.6657$. Allí se observa que ambas curvas son prácticamente coincidentes, es decir, que el estimador RLL coincide con el estimador *logit* de la PD. También se observa que el estimador $\hat{m}_{1,h_{loo\text{-}cv}}$ prácticamente no sufre el efecto frontera, como se aprecia en la curva dibujada en color negro. Por tanto, ambos estimadores, $\hat{m}_{1,h_{loo\text{-}cv}}$ y $\hat{m}_{logit}$, ofrecen resultados comparables en términos de suavidad de las curvas y de capacidad predictiva en los extremos del rango muestral, lo que mejora sustancialmente los resultados obtenidos hasta ahora con ambos estimadores no paramétricos, RLL y NW.

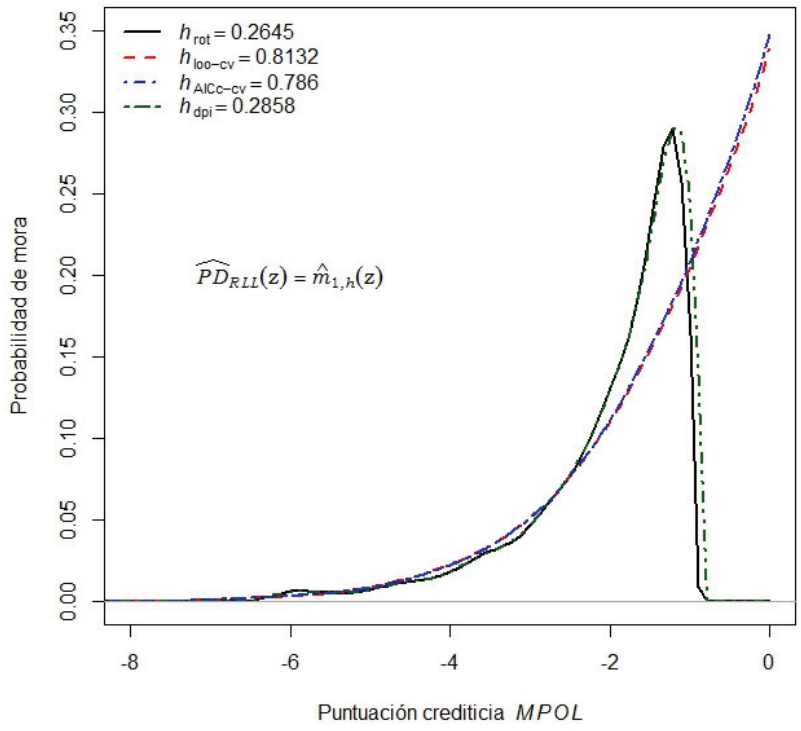

Figura 2.39 Curvas de PD_{MPOL} obtenidas con el estimador de RLL.

En la Figura 2.39 se muestran las estimaciones no paramétricas de la curva de PD para el modelo $MPOL$. Allí se observa que las curvas obtenidas con las

ventanas $h_{rot} = 0.2645$ y $h_{dpi} = 0.2858$ presentan un marcado efecto frontera en torno al máximo muestral, provocando que las curvas de PD decrezcan hasta cero de manera abrupta, tal y como se aprecia en las curvas dibujadas con trazos de color negro y verde, respectivamente. En cambio, cuando el estimador $\hat{m}_{1,h}(z)$ es calculado utilizando las ventanas $h_{loo\text{-}cv} = 0.8132$ (trazo de color rojo) y $h_{AICc\text{-}cv} = 0.786$ (trazos de color azul) los resultados obtenidos son mucho más satisfactorios, logrando un buen grado de suavizamiento y ausencia del efecto frontera en ambas curvas. Como consecuencia, se obtiene que los valores adecuados que puede tomar la ventana de suavizado del modelo $MPOL$, denotada como h^1_{MPOL}, se encuentran en el rango $0.78 \leq h \leq 0.82$.

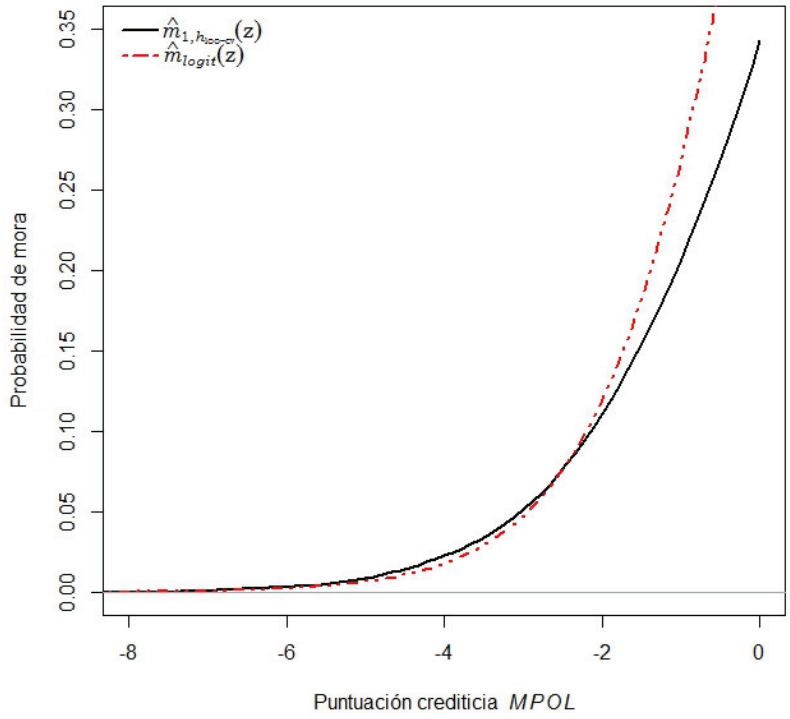

Figura 2.40 Curvas estimadas de PD_{MPOL}. Estimador de RLL vs. estimador $logit$.

En la Figura 2.40 se comparan las curvas de probabilidad de mora obtenidas con los estimadores $\hat{m}_{logit}$ y $\hat{m}_{1,h^1_{MPOL}}$, donde la ventana óptima selecciona-

da es $h^1_{MPOL} = h_{loo\text{-}cv} = 0.8132$. Allí se observa que ambas curvas son crecientes de forma monótona y que arrojan probabilidades de mora muy similares para valores de $z \leq -2.56$. A partir de la intersección entre ambas curvas, aproximadamente en $z_0 = -2.56$, se verifica que $\hat{m}_{logit}(z) \geq \hat{m}_{1,h_{loo\text{-}cv}}(z)$ para $z \geq z_0$. Además, se observa que, aproximadamente para valores de $z < -7$, la curva no paramétrica decrece hasta cero mucho antes de alcanzar el mínimo muestral, que se encuentra en el punto $z = -10.636$.

A continuación, en las figuras siguientes se comparan las curvas de PD obtenidas con los tres estimadores estudiados en esta sección, $\hat{m}_{logit}$, $\hat{m}_{0,h^0_k}$ y $\hat{m}_{1,h^1_k}$, donde h^0_k y h^1_k son ventanas adecuadas para los estimadores de NW y de RLL, respectivamente, y donde k denota uno de los tres modelos de puntuaciones posibles, $ENTIDAD$, $MSEG$ y $MPOL$.

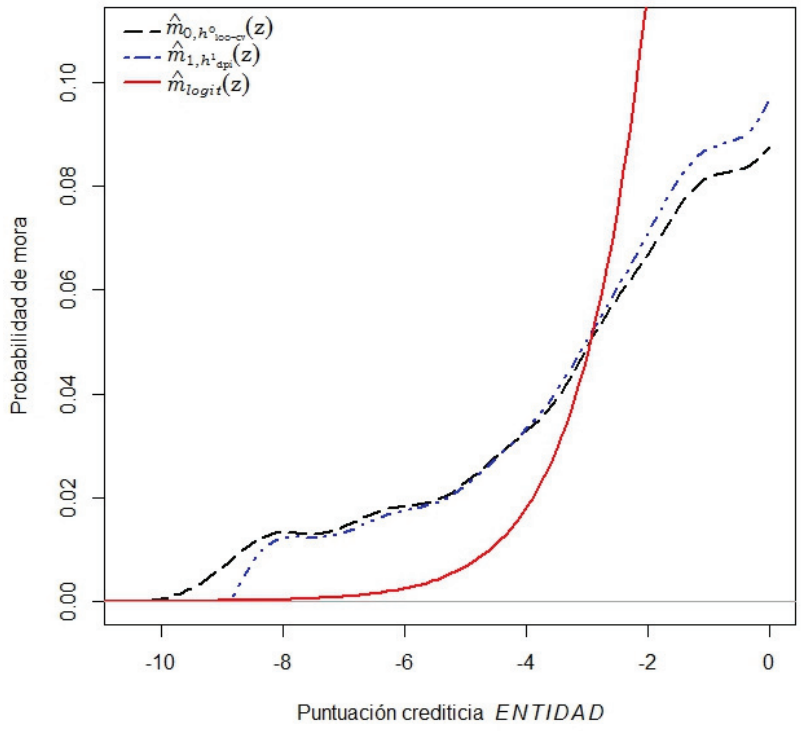

Figura 2.41 Curvas estimadas de $PD_{ENTIDAD}$. Estimadores de NW, RLL y $logit$.

En la Figura 2.41 se representan las curvas de PD para el modelo de puntuación $ENTIDAD$ obtenidas con los estimadores $\hat{m}_{0,h^0_{loo\text{-}cv}}$, $\hat{m}_{1,h^1_{dpi}}$ y $\hat{m}_{logit}$, donde la ventana seleccionada para el estimador de NW es $h^0_{loo\text{-}cv} = 0.4925$ (trazo de color negro) y la ventana seleccionada para el estimador de RLL es $h^1_{dpi} = 0.6279$ (trazo de color azul). Los resultados muestran que ambas curvas no paramétricas ofrecen grados de ajuste y de suavizamiento similares, aunque el estimador de NW ofrece mejores resultados en zonas próximas al mínimo muestral, tal y como se aprecia en la parte inferior izquierda del trazado de las curvas (en el rango $-10 \leq z \leq -8$). Como resultado, se obtiene que es preferible utilizar el estimador de NW para calcular la PD con puntuaciones próximas al mínimo muestral mientras que para el resto de puntuaciones (por ejemplo, para $z > -8$), ambos estimadores ofrecen similar calidad del ajuste no paramétrico de la curva de $PD_{ENTIDAD}$.

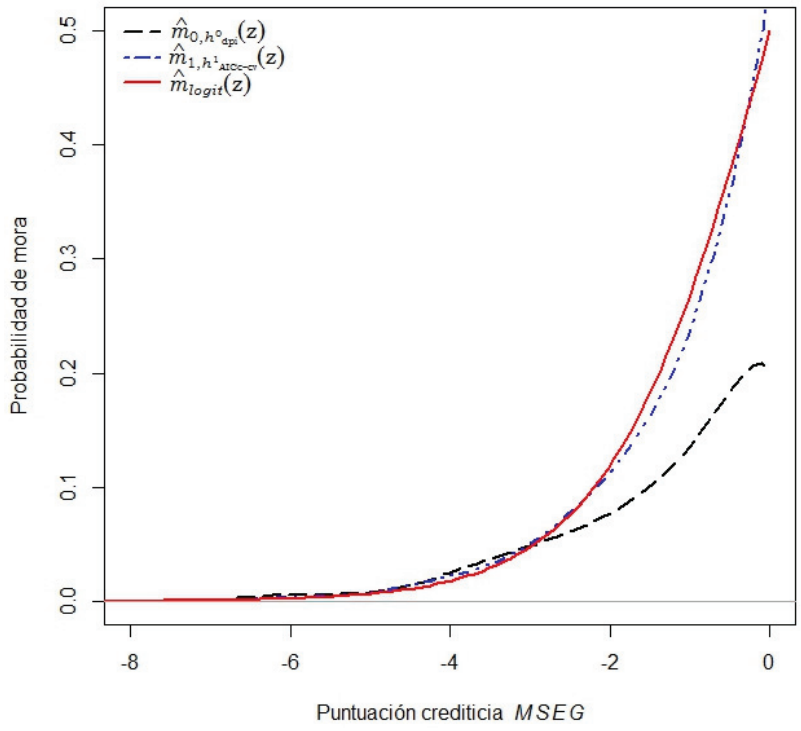

Figura 2.42 Curvas estimadas de PD_{MSEG}. Estimadores de NW, RLL y $logit$.

En la Figura 2.42 se representan las curvas de PD para el modelo de puntuaciones $MSEG$ obtenidas con los estimadores $\hat{m}_{0,h^0_{loo\text{-}cv}}$, $\hat{m}_{1,h^1_{loo\text{-}cv}}$ y $\hat{m}_{logit}$, donde la ventana seleccionada para el estimador de NW es $h^0_{loo\text{-}cv} = 0.3484$ (trazo de color negro) y la ventana seleccionada para el estimador de RLL es $h^1_{loo\text{-}cv} = 0.3615$ (trazo de color azul). Los resultados muestran que ambos estimadores no paramétricos producen curvas con buen grado de suavizamiento, aunque el estimador de NW parece verse más afectado por el efecto frontera. Esto hace preferible al estimador de RLL para el ajuste de las curvas de PD_{MSEG} en zonas próximas al máximo muestral (para $z \geq -1.28$). Además, como se vio antes (Figura 2.38), las probabilidades de mora obtenidas con el estimador de RLL prácticamente coinciden con las obtenidas con el modelo *logit* (línea de color rojo).

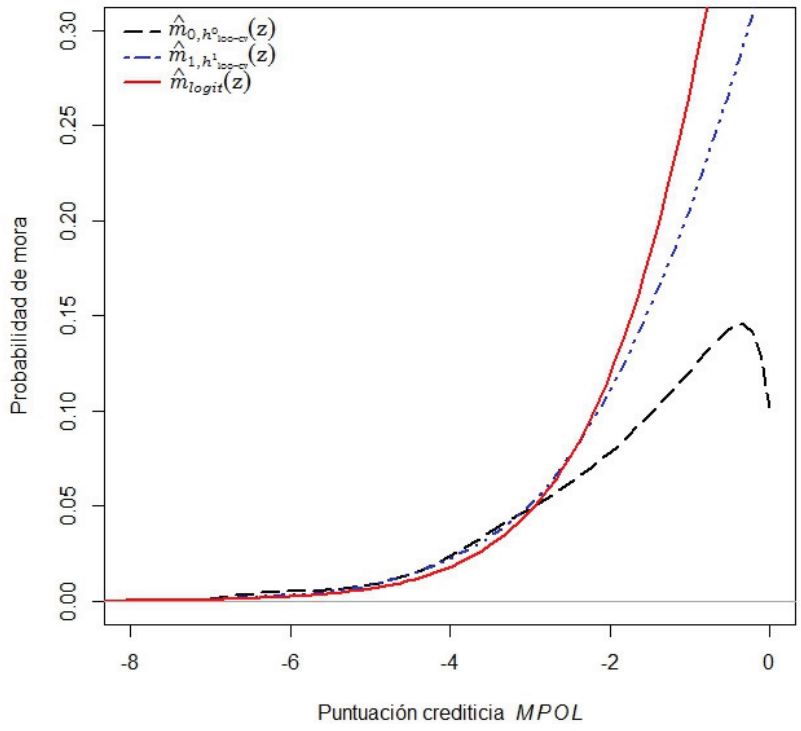

Figura 2.43 Curvas estimadas de PD_{MPOL}. Estimadores de NW, RLL y $logit$.

En la Figura 2.43 se representan las curvas de PD calculadas con las puntuaciones $MPOL$, obtenidas con los estimadores $\hat{m}_{0,h^0_{loo\text{-}cv}}$, $\hat{m}_{1,h^1_{loo\text{-}cv}}$ y $\hat{m}_{logit}$, donde la ventana escogida para el estimador de NW es $h^0_{loo\text{-}cv} = 0.4826$ y la ventana escogida para el estimador de RLL es $h^1_{loo\text{-}cv} = 0.8132$. Se observa que los resultados obtenidos con ambos estimadores, NW y RLL, son similares a los obtenidos con las puntuaciones $MSEG$. Aunque ambas curvas son suaves, el estimador de la RLL es preferible al de NW ya que no parece verse afectado por el efecto frontera. También se ve mejor ajuste del estimador RLL en puntos cercanos al máximo muestral. Sin embargo, a diferencia de lo obtenido para las puntuaciones $MSEG$, la curva de PD_{MPOL} estimada por RLL no coincide con la curva obtenida con el estimador *logit*, aunque ambas son relativamente cercanas.

Validación de los estimadores de la PD a partir de la TM

Hasta ahora, se ha visto un análisis descriptivo de las curvas de PD obtenidas con los tres estimadores estudiados, *logit*, NW y RLL (ver Figuras 2.29 a la 2.43). El estudio de los estimadores tipo núcleo, NW y RLL, se ha centrado especialmente en el grado de suavizamiento de las curvas de PD estimadas, obtenido con distintos selectores del parámetro de suavizado, h, y su comparación con la curva de PD obtenida con el estimador *logit*.

Para complementar la parte descriptiva, se ha realizado un estudio de validación de los tres estimadores con el fin de evaluar el poder predictivo de los mismos. El estudio de validación está compuesto por cuatro medidas empíricas. La tasa de mora estimada, $\widehat{TM}$, que ofrece una medida global del poder de clasificación del estimador de la PD, los errores de clasificación de tipo I y de tipo II ($err\ I$ y $err\ II$), y la tasa de precisión ($\widehat{TP}$), estos dos últimos utilizados previamente en la Sección 2.4.4.

El mecanismo de validación es el siguiente. Se tomaron 100 submuestras aleatorias con reemplazamiento, de tres tamaños distintos, de la muestra original de validación ($n = 5\,000$). Los tamaños de submuestra utilizados son: $m = 300, 500$ y $1\,000$. Para cada submuestra se calcularon las cuatro medidas de validación mencionadas, tomando como resultado la media de los 100 valores calculados. Para obtener las estimaciones no paramétricas de las curvas de PD se emplearon los mismos valores de h utilizados en las Figuras 2.41 a la 2.43. Los resultados obtenidos se muestran en la Tabla 2.28.

Tabla 2.28. Medidas de validación de la capacidad predictiva de los estimadores *logit*, *NW* y *RLL* de las curvas de *PD*

	Estimador	*logit*			*NW*			*RLL*		
$TM\,\%$	Medida de Validación	*ENTIDAD*	*MSEG*	*MPOL*	*ENTIDAD*	*MSEG*	*MPOL*	*ENTIDAD*	*MSEG*	*MPOL*
4.260 m =300	$\widehat{TM}\,\%$	4.32	3.91	3.90	3.86	3.79	3.81	3.93	4.14	4.16
	err I %	46.77	35.86	32.07	41.16	25.50	27.25	40.13	29.37	28.25
	err II %	33.28	41.32	39.79	42.08	49.34	45.53	44.39	46.60	45.33
	$\widehat{TP}\,\%$	66.12	58.91	60.55	57.95	51.68	55.21	55.79	54.13	55.40
3.912 m =500	$\widehat{TM}\,\%$	4.36	3.92	3.91	3.86	3.79	3.81	3.96	4.15	4.17
	err I %	47.83	34.25	32.50	42.00	24.95	28.04	41.14	28.73	28.04
	err II %	33.03	41.52	40.21	41.99	49.43	45.65	44.44	46.58	45.45
	$\widehat{TP}\,\%$	66.39	58.77	60.10	58.01	51.54	55.05	55.68	54.12	55.23
4.021 m =1 000	$\widehat{TM}\,\%$	4.33	3.90	3.89	3.86	3.78	3.80	3.95	4.13	4.15
	err I %	48.84	34.68	32.32	41.88	24.75	28.49	41.04	29.05	28.49
	err II %	32.85	41.17	39.75	41.67	49.30	45.50	44.10	46.41	45.29
	$\widehat{TP}\,\%$	66.55	59.09	60.55	58.33	51.69	55.19	56.02	54.29	55.39

De los valores contenidos en la Tabla 2.28 se extraen tres resultados principales. En primer lugar, se observa que la tasa de mora media ($\widehat{TM}$) varía entre 3.912 % y 4.26 %, rango que contiene el valor 4.08 % de tasa de morosidad observada en la muestra de validación. También se observa que para tamaños de submuestra pequeños ($m = 300$ y $m = 500$), el estimador *logit* arroja los valores de $\widehat{TM}$ más ajustados a la TM de la muestra de validación ($n = 5\,000$). Con estos tamaños de submuestra, también se obtienen buenos resultados utilizando el estimador de RLL. Sin embargo, cuando se toman submuestras de tamaño $m = 1\,000$, el mejor resultado se obtiene utilizando el estimador de RLL. En este sentido, se deduce que a medida que aumenta el tamaño muestral, el estimador de RLL tiende a minimizar la distancia en valor absoluto entre la TM observada (muestra de validación) y la TM estimada con el modelo, $\widehat{TM}$, utilizando como covariables las puntuaciones obtenidas con los modelos $ENTIDAD$ y $MPOL$.

Con respecto a los errores de clasificación, se observa que, independientemente del tamaño de submuestra utilizado, el estimador de NW siempre logra minimizar el error de tipo I (representativo del riesgo de crédito) cuando se utiliza como covariable del modelo de regresión la puntuación crediticia $MSEG$. En cambio, con respecto el error de clasificación de tipo II (representativo del riesgo asociado al coste de oportunidad), se observa que los mejores resultados (errores mínimos) se logran con el estimador *logit* utilizando como covariable la puntuación crediticia $ENTIDAD$. El resultado es el mismo cuando se observan los valores de la tasa media de precisión, $\widehat{TP}$. Así, independientemente del tamaño de muestra, el modelo *logit* de la PD logra minimizar el error de clasificación de tipo II y al mismo tiempo logra maximizar la precisión ($\widehat{TP}$) en la clasificación de los créditos cuando se utiliza como covariable del modelo la puntuación crediticia dada por la entidad. Finalmente, es importante señalar que los resultados relativos a los errores de clasificación y a la $\widehat{TP}$ se obtuvieron a partir de los mismos valores (puntos de corte) utilizados en el mecanismo de validación de los modelos logísticos de la Sección 2.4.4 (Tabla 2.17).

2.6. Comentarios y conclusiones

En este capítulo se han estudiado técnicas estadísticas para la construcción de un modelo de puntuación crediticia (o credit scoring) cuya implementación es una parte fundamental de los procesos de control y administración del riesgo de crédito de las entidades financieras. En este estudio se han tenido en cuenta las principales etapas que componen el proceso de modelización del perfil crediticio de los acreditados, siendo las más importantes el tratamiento estadístico de los datos, las técnicas de modelización utilizadas y la validación de los modelos obtenidos. La metodología expuesta en este capítulo se fundamenta en las técnicas utilizadas por la entidad financiera colaboradora en el desarrollo de esta memoria y en trabajos de reconocida influencia en la literatura dedicada al análisis del riesgo de crédito, de entre los cuales más adelante se citan algunos ejemplos.

En la Sección 2.1 se ofrece una breve revisión del estado del arte sobre las distintas estrategias y metodologías utilizadas en el estudio y modelización de la calidad crediticia de los clientes de las entidades financieras. Sin adentrarse exhaustivamente en las fuentes consultadas, en esta sección se pretende explicar cómo han ido evolucionando las técnicas utilizadas por los expertos de las entidades financieras y por quienes se han dedicado al estudio de la modelización del riesgo de crédito desde mediados del siglo XX hasta la actualidad. Trabajos como los debidos a Altman y Saunders (1998), Sobehart et al. (2001), Basel Committee on Banking Supervision (2005a), Engelmann y Rauhmeier (2006), Crook et al. (2007), Baesens et al. (2009) y Baesens (2014), son considerados de lectura recomendada sobre este tema, debido a la variedad de sus contenidos, la profundidad con que estos son tratados, y a su influencia en la literatura especializada.

En la Sección 2.2 se expone brevemente la teoría que sustenta la aplicación de un modelo de regresión logística al estudio y modelización del perfil de solvencia de los aceditados. Se establecen los supuestos del modelo y se explica el mecanismo de estimación de los parámetros del mismo. También se describen los métodos utilizados en el tratamiento previo de las variables regresoras y se establecen los criterios estadísticos de selección e inclusión de las mismas en el modelo.

En la Sección 2.3 se describen las técnicas utilizadas en el análisis de validación de modelos ajustados. Métodos de validación como el contraste de

Hosmer-Lemeshow (Figura 2.1 y Tablas 2.2 y 2.3), el análisis de curvas ROC (Figura 2.2) y de curvas CAP (Figura 2.3), la distancia de Kolmogorov-Smirnov (Figura 2.5) y el análisis de los errores de clasificación de tipo I y II (Tabla 2.4 y Figura 2.4) son algunas de las técnicas tratadas en este capítulo, siguiendo la metodología utilizada por Sobehart et al. (2000), Sobehart et al. (2001), Thomas et al. (2002), Basel Committee on Banking Supervision (2005a), Engelmann (2006), Engelmann y Rauhmeier (2006) y Crook et al. (2007), entre otros.

La Sección 2.4 está dedicada a la aplicación de las técnicas descritas anteriormente a una base de datos de tarjetas de crédito. Como se ha explicado antes, se trabajó en la construcción de un modelo de puntuación crediticia bajo la hipótesis de que la probabilidad condicional de mora o insolvencia de un cliente (o solicitante) de tarjetas de crédito sigue una distribución logística. Las covariables del modelo (ver Sección 2.4.2) fueron tratadas conforme a las técnicas descritas en la Sección 2.2, esto es, las variables cuantitativas fueron tratadas con los métodos de regresión polinómica y regresión segmentada, mientras que las variables categóricas fueron tratadas con el método de variables auxiliares o dummies. Como resultado del ajuste de los modelos, fueron seleccionados los dos que ofrecieron mejor poder predictivo, el modelo denominado $MPOL$ y el modelo denominado $MSEG$. Las puntuaciones obtenidas con ambos modelos fueron comparadas con las puntuaciones del modelo proporcionado por la entidad colaboradora, denominado como $ENTIDAD$. Los resultados obtenidos muestran que los dos modelos ajustados, $MPOL$ y $MSEG$, ofrecen similar capacidad predictiva en términos del análisis de curvas ROC y CAP, y de sus medidas de precisión, AUC y AR, respectivamente, siendo ambos superiores en este sentido al modelo $ENTIDAD$. También se vio que la utilización del modelo $MPOL$ como herramienta de clasificación de los créditos permitió obtener el menor error de tipo I (representativo del *riesgo de crédito*) mientras que al utilizar el modelo $ENTIDAD$ se obtuvo el menor error de tipo II (representativo del riesgo por *coste de oportunidad*). El mismo resultado se obtuvo cuando se compararon los tres modelos en términos de la tasa de precisión (TP), donde el modelo de la entidad fue significativamente superior a los otros modelos.

Como aplicación práctica de los modelos estudiados se obtuvo un método para determinar la frontera de concesión de la tarjeta a partir de un dictamen basado en las puntuaciones obtenidas con los modelos (ver Sección 2.4.6). La técnica consitió en la construcción de una función de utilidad a partir de una

matriz de costes y beneficios, y de la distribución conjunta del vector (Z, Y), que fue estimada empíricamente a partir de la muestra. Los límites del dictamen se obtuvieron a partir de las soluciones del problema de maximización de la función de utilidad bajo dos escenarios económicos diferentes, uno más favorable y otro menos favorable para la entidad de crédito.

La parte final de este capítulo, la Sección 2.5, está dedicada al estudio de técnicas que permiten obtener la tasa de mora esperada (TM) de una cartera de tarjetas de crédito a partir de las probabilidades de mora individuales (PD) asociadas a cada acreditado o tarjeta. Para obtener los valores de las PD se utilizaron tres mecanismos de cálculo, una fórmula directa dada por el estimador *logit* (fórmula (2.31)) y dos estimadores no paramétricos tipo núcleo, el estimador de *Nadaraya-Watson* (NW) y el estimador de la *regresión local lineal* (RLL), cuyos resultados gráficos se comparan en las Figuras 2.29 a la 2.40. Los resultados muestran que uno de los métodos de selección más adecuados para determinar la ventana de suavizado en modelos de regresión con variable dependiente binaria es el de validación cruzada de tipo *loo-cv*, aunque la validación cruzada de tipo *AICc-cv* y el método de *plug-in* directo (*dpi*) funcionaron bien en ocasiones. Además, tal y como se ve en las Figuras 2.42 y 2.43, se verifica que el estimador de RLL funciona mejor, en general, que el estimador de NW cuando se quiere estimar las curvas de PD de los tres modelos de puntuaciones. Sólo es recomendable utilizar el estimador de NW cuando se trata de puntuaciones provenientes del modelo $ENTIDAD$ que están próximas al mínimo muestral (ver Figura 2.41). También se vio que el estimador de RLL resultó ser menos vulnerable al efecto frontera cuando se calculan las curvas de PD utilizando las puntuaciones obtenidas con los modelos $MSEG$ y $MPOL$. Por último, del estudio de validación cuyos resultados están contenidos en la Tabla 2.28, se obtuvo que el estimador de RLL constituye el mejor mecanismo para estimar la TM de la cartera mientras que el estimador de NW permite minimizar el error de tipo I y el estimador *logit* permite minimizar el error de tipo II a la vez que logra maximizar la tasa de precisión en la clasificación de los créditos.

Finalmente, es importante hacer una breve reflexión sobre posibles extensiones de las técnicas estudiadas en este capítulo que, por limitaciones de tiempo, se han dejado como objetivo de investigación para futuros trabajos. Como se ha dicho al inicio de este capítulo, el modelo de regresión logística es una de las técnicas estadísticas más populares en la construcción de modelos de riesgo de crédito, en particular de modelos de puntuación crediticia.

Sin embargo, como suele ocurrir con los modelos de regresión paramétricos, este modelo está sujeto a supuestos fuertemente restrictivos, por ejemplo, que la variable de puntuación crediticia, Z, sigue una distribución logística, y que el modelo de regresión está limitado a una forma lineal y aditiva en las covariables. En este sentido, una extensión natural de este trabajo es la utilización de otros *modelos lineales generalizados* para estudiar el perfil crediticio y la morosidad de los créditos contenidos en la base de datos utilizada en esta memoria. Así, sería interesante comparar la capacidad predictiva de los modelos *logit* estudiados con modelos de regresión de tipo *probit* y *log-log* complementarios (*clog-log*), donde la teoría estudiada aquí es fácilmente adaptable utilizando las funciones de enlace (o *link*) asociadas a las distribuciones normal estándar (*probit*) y del valor extremo[5] (*clog-log*), respectivamente. Ver, por ejemplo, los trabajos de Neagu et al. (2009) y Gurný y Gurný (2013), en este contexto. Otra posible extensión, menos conocida en el contexto de la modelización del riesgo de crédito, es la utilización de modelos semiparamétricos que permiten relajar las limitaciones anteriores permitiendo que la función de enlace se suponga desconocida debiendo ser estimada no paramétricamente a partir de los datos. En otros modelos semiparamétricos, en cambio, se relaja el supuesto de que todas las covariables están relacionadas linealmente, permitiendo que entre algunas de ellas exista una relación desconocida que puede estimarse no paramétricamente, esta clase de modelos se conoce como *modelos parcialmente lineales*. Ver, por ejemplo, el trabajo de Müller y Rönz (2000) en este mismo contexto.

Por otra parte, con respecto a las técnicas de validación vistas en la Sección 2.4.4, en este capítulo se utilizó el método de Kolmogorov-Smirnov para determinar el punto de corte óptimo de discriminación entre clientes buenos y malos. En ese sentido, es conocido que éste y otros métodos existentes en la literatura presentan difultades cuando se trabaja con muestras altamente desbalanceadas, tal y como ocurre con los datos utilizados en este capítulo (sólo 766 créditos cumplen la condición de ser morosos versus 19 234 que no la cumplen). Una de las razones de esto es que las técnicas de clasificación convencionales están construidas principalmente bajo el supuesto de que cada

[5] La distribución del valor extremo, o distribución de *Gumbel*, es conocida en el contexto del análisis de fiabilidad y se utiliza frecuentemente para obtener la función de distribución del mímimo de una muestra de variables aleatorias *i.i.d.* con distribución exponencial. La función de enlace *clog-log* está relacionada con la distribución de *Gumbel* a partir de su función de distribución acumulativa definida por $F(x) = 1-\exp[-\exp(x)]$, para todo $x \in \mathbb{R}$ (ver, por ejemplo, Aitkin et. al (2005)).

una de las clases de sujetos están suficientemente representadas en el conjunto de datos. Debido a que este tema aún sigue abierto, una posible línea de trabajo futura es la búsqueda de métodos más robustos que pemitan obtener el punto de corte óptimo evitando (o reduciendo) el efecto del desequilibrio entre el número de créditos morosos y el de no morosos en la muestra. Ver, por ejemplo, las ideas propuestas por Calabrese (2014).

Por último, otra línea de trabajo que se propone para más adelante, es la generalización de la estrategia óptima de concesión de las tarjetas de crédito estudiada en la Sección 2.4.6. La función de utilidad definida en (2.24) está construida bajo la hipótesis de que los costes de mala clasificación de los créditos son constantes conocidas. Sin embargo, en una situación más general, donde los supuestos son menos restrictivos, es posible pensar que los costes de mala clasificación son variables que dependen, por ejemplo, del tamaño de la cartera, de la capacidad de solvencia del acreditado (la puntuación Z) o de otras variables inherentes al mercado crediticio y al estado de la economía.

Capítulo 3

Estimación de la probabilidad de mora vía análisis de supervivencia

3.1. Introducción

En el contexto de Basilea II, es conocido que uno de los objetivos fundamentales de los modelos de riesgo de crédito es la estimación de la *probabilidad de mora* (PD) del crédito, es decir, la probabilidad de que un acreditado sea incapaz de cumplir sus obligaciones con la entidad financiera en el período de tiempo pactado para ello. En el caso de los créditos a particulares, Basilea II establece este período en 90 días (ver Sección 1.1.3). El estudio de este problema constituye el eje central de la investigación llevada a cabo en esta memoria, y en este capítulo se proporciona una metodología para su modelización a partir del estudio de los tiempos de vida de los créditos.

Debido a que los modelos de riesgo financiero tradicionales no son aptos para predecir la probabilidad de mora en créditos personales, como ocurre, por ejemplo, con los modelos construidos bajo el enfoque de Merton (1974), desde finales del siglo XX han surgido diversas perspectivas de análisis dedicadas al estudio de la morosidad en créditos a particulares, siendo el análisis de supervivencia uno de los enfoques que ha captado más interés en la literatura.

La idea de aplicar técnicas de análisis de supervivencia en la construcción de modelos de riesgo de crédito se atribuye originalmente a Narain (1992), cuyo enfoque fue desarrollado más tarde por Banasik et al. (1999). Aprovechando la analogía existente entre el estudio del *tiempo hasta que se produce la mora* de un crédito y el *tiempo hasta el suceso de interés* en problemas de naturaleza biomédica, estos autores demostraron que la teoría del análisis de supervivencia es útil para modelizar estadísticamente el suceso *entrada en mora del crédito* a partir del estudio de la función de supervivencia del *tiempo de vida del crédito*. Así, bajo el enfoque de análisis de supervivencia para el riesgo de crédito, no sólo es importante ser capaces de predecir si un crédito será o no moroso, sino también, poder estimar el tiempo que tardará en producirse la mora. En este sentido, Narain (1992) estudió el tiempo hasta que se produce la mora en una base de datos de préstamos personales con vencimiento de 24 meses utilizando un modelo de regresión de vida exponencial acelerada. Narain (1992) encontró evidencia de que su modelo estima con un buen nivel de precisión el número de créditos fallidos en diferentes instantes del tiempo. Además, mostró que las decisiones sobre la concesión de créditos pueden ser mejoradas mediante el uso de técnicas de análisis de supervivencia en comparación con la utilización de regresión lineal múltiple. Finalmente, este autor argumenta que el análisis de supervivencia puede utilizarse para estudiar la morosidad en todo tipo de operaciones crediticias en las que existen variables predictoras y la variable dependiente es el tiempo hasta que se produce la mora del crédito.

Por otra parte, Banasik et al. (1999) compararon el desempeño de *modelos de regresión exponencial*, de *Weibull* y de *Cox* versus la *regresión logística*, encontrando que los modelos de supervivencia son competitivos, y en ocasiones, superiores al enfoque de regresión logística tradicional. De acuerdo con los resultados obtenidos, estos autores concluyen que el análisis de supervivencia es una herramienta útil para obtener estimaciones precisas de la PD con horizonte de predicción fijo de 12 meses para diversos tipos de préstamos, validando así la utilidad de este enfoque según el *Acuerdo de Basilea II* (Tong et al. (2012)).

Otros autores que han estudiado técnicas de modelización de la PD a partir del tiempo de vida de los créditos son, entre otros, Carling et al. (1998), Stepanova y Thomas (2002), Roszbach (2004), Glennon y Nigro (2005), Allen y Rose (2006), Malik y Thomas (2006) y Beran y Djaïdja (2007).

En el presente capítulo se desarrollan las ideas sobre modelización del riesgo de crédito vía análisis de supervivencia propuestas por Cao et al. (2009). En dicho trabajo, los autores plantean un modelo de regresión que permite obtener estimaciones de la PD condicionada al perfil de solvencia del acreditado suponiendo que existe una relación de dependencia entre el tiempo de vida del crédito, T, y el vector de covariables $\boldsymbol{X}$. Para estudiar el ajuste de los modelos propuestos se utilizan tres enfoques diferentes: técnicas de regresión paramétricas (vía máxima verosimilitud), técnicas de regresión semiparamétricas (vía regresión de Cox) y técnicas de regresión no paramétricas (vía suavizado tipo núcleo).

3.1.1. Distribución del tiempo hasta la mora

El enfoque de análisis de supervivencia para el riesgo de crédito permite construir modelos predictivos de la morosidad crediticia utilizando como variable de interés el tiempo hasta que se produce la mora o impago del mismo, denotada por T. Si por el contrario, lo que se observa es el tiempo hasta la cancelación (en vencimiento o anticipada) del crédito, entonces se está en presencia de la variable aleatoria censurante, C. Sin pérdida de generalidad, se suponen conocidas p características que permiten conocer el grado de solvencia individual de cada cliente, caracterizadas por el vector $\boldsymbol{X} \in \mathbb{R}^p$, cuya información puede resumirse en una variable unidimensional, denotada por X. Para transferir la información del espacio de covariables, $\mathbb{R}^p$, al conjunto $\mathbb{R}$, es habitual utilizar técnicas de reducción de la dimensionalidad, por ejemplo, como los modelos de regresión logística estudiados en el Capítulo 2. Más adelante se verá que en gran parte de esta memoria se utiliza como covariable de los modelos de regresión una variable unidimensional, X, ya que matemáticamente su tratamiento es más sencillo y las curvas de regresión resultantes pueden ser representadas gráficamente, haciendo más simple su interpretación.

Hipótesis sobre las variables del modelo

En presencia de datos completos, existe una variedad de técnicas de regresión clásicas que permiten estudiar la relación de dependencia entre la variable T y el vector de covariables $\boldsymbol{X}$. Sin embargo, cuando los datos no

son completamente observables, como ocurre cuando se está en presencia de datos censurados, es necesario emplear técnicas más sofisticadas para modelizar dicha relación. Además, debido a que en el ámbito de las entidades de crédito tradicionales el perfil de los clientes se conoce durante todo el tiempo de vida del crédito, en este capítulo se estudiarán modelos de regresión para estimar la probabilidad de mora (PD) utilizando un enfoque de supervivencia condicional en el que se consideran las siguientes hipótesis sobre las variables del modelo:

H 3.1 *T, C y $\boldsymbol{X}$ son variables aleatorias no negativas absolutamente continuas.*

H 3.2 *La variable tiempo hasta que se produce la mora, T, no es completamente observable debido a la censura aleatoria por la derecha que la afecta.*

H 3.3 *La variable C, es la variable censurante que, en ocasiones, impide observar el suceso de interés, es decir, la mora del crédito.*

H 3.4 *El vector de covariables $\boldsymbol{X}$ no está afectado por la censura y su comportamiento probabilístico no depende del tiempo.*

H 3.5 *Existe una relación de dependencia entre las variables T y $\boldsymbol{X}$.*

H 3.6 *Las variables T y C son condicionalmente independientes dado el vector $\boldsymbol{X} = \boldsymbol{x}$.*

H 3.7 *Las variables T y C poseen funciones de distribución condicionales absolutamente continuas dado el vector $\boldsymbol{X} = \boldsymbol{x}$.*

Los datos observables provienen de una muestra aleatoria de n ternas de variables independientes e idénticamente distribuidas ($i.i.d$) representadas por el conjunto $\{(\xi_1, \boldsymbol{X}_1, \delta_1), \ldots, (\xi_n, \boldsymbol{X}_n, \delta_n)\}$, con idéntica distribución que el vector $(\xi, \boldsymbol{X}, \delta)$, donde $\xi_i = \min\{T_i, C_i\}$ corresponde al tiempo de vida observado del crédito i-ésimo, T_i es el tiempo hasta que se produce la mora del crédito i-ésimo, C_i es el valor de la variable censurante (tiempo de censura) del crédito i-ésimo, $\delta_i = I(T_i \leq C_i)$ es la indicadora de no censura, o lo que

es lo mismo, la indicadora de que se ha producido la mora del crédito i-ésimo y $\boldsymbol{X}_i$ es un vector p-dimensional de covariables que contienen información sobre el grado de solvencia del i-ésimo acreditado, por ejemplo, como las variables estudiadas en el Capítulo 2 (Sección 2.4.2), con $1 \leq i \leq n$.

Además, para garantizar la correcta especificación de los modelos de probabilidad que se estudian en esta memoria, en adelante, para cada i-ésimo crédito se utilizará la siguiente notación:

(i) $F(t|\boldsymbol{x}) = P\left(T_i \leq t|\boldsymbol{X}_i = \boldsymbol{x}\right)$ es la función de distribución condicional del tiempo hasta que se produce la mora.

(ii) $G(t|\boldsymbol{x}) = P\left(C_i \leq t|\boldsymbol{X}_i = \boldsymbol{x}\right)$ es la función de distribución condicional del tiempo de censura.

(iii) $H(t|\boldsymbol{x}) = P\left(\xi_i \leq t|\boldsymbol{X}_i = \boldsymbol{x}\right)$ es la función de distribución condicional del tiempo de vida observado.

(iv) $M(\boldsymbol{x}) = P\left(\boldsymbol{X}_i \leq \boldsymbol{x}\right)$ y $m\left(\boldsymbol{x}\right) = M'(\boldsymbol{x})$ son las funciones de distribución y de densidad de la covariable $\boldsymbol{X}_i$, respectivamente.

Además, como consecuencia de H3.5 se obtiene la igualdad:

$$1 - H(t|\boldsymbol{x}) = \left(1 - F(t|\boldsymbol{x})\right)\left(1 - G(t|\boldsymbol{x})\right).$$

La función de distribución condicional del tiempo hasta que se produce la mora, T, dado el vector de covariables $\boldsymbol{X} = \boldsymbol{x}$ se define por

$$F(t|\boldsymbol{x}) = P\left(T \leq t|\boldsymbol{X} = \boldsymbol{x}\right) = \int_0^t f(u|\boldsymbol{x})du,$$

donde $f(t|\boldsymbol{x})$ denota la función de densidad condicional de T dada $\boldsymbol{X} = \boldsymbol{x}$.

Para obtener un estimador de la función de distribución condicional, $F(t|\boldsymbol{x})$, o equivalentemente, de la función de supervivencia condicional definida por $S(t|\boldsymbol{x}) = 1 - F(t|\boldsymbol{x})$, en este capítulo se estudian las ideas propuestas por Cao et al. (2009). Estos autores emplean tres metodologías que permiten estimar localmente la función $F(t|\boldsymbol{x})$ desde tres perspectivas diferentes.

Modelo de ajuste paramétrico Consiste en suponer que existe una relación de dependencia entre la variable T y la covariable $\boldsymbol{X}$ que es gobernada por un

modelo de probabilidad conocido. Por simplicidad del estudio, en este capítulo se han estudiado sólo modelos pertenecientes a la familia de distribuciones exponenciales, sin embargo, este es un problema abierto que puede ser extentido a otra clase de distribuciones paramétricas. La estimación de los parámetros se realiza vía máxima verosimilitud, obteniéndose como resultado el estimador máximo verosimímil de la PD para cada crédito.

Modelo de ajuste semiparamétrico Consiste en estimar la función de supervivencia condicional del tiempo hasta la mora, $S(t|\boldsymbol{x}) = 1 - F(t|\boldsymbol{x})$, bajo el supuesto de riesgos proporcionales de Cox (1972). Para ello se utiliza un modelo de regresión semiparamétrico compuesto por una parte no paramétrica, dada por la función de fallo acumulativa basal que se obtiene vía el estimador de *Nelson-Aalen*, y una parte paramétrica que se estima por máxima verosimilitud.

Modelo de ajuste no paramétrico Consiste en modelizar la relación de dependencia entre T y $\boldsymbol{X}$ empleando un modelo de regresión no paramétrica con datos censurados en diseño aleatorio. Para estimar la función de distribución condicional del tiempo hasta la mora, $F\left(t|\boldsymbol{x}\right)$, se utiliza el estimador no paramétrico tipo núcleo de Beran (1981).

3.2. Estimación de la probabilidad de mora condicional

A continuación, se desarrollan las ideas expuestas de forma introductoria en el Capítulo 1 de esta memoria (Sección 1.2.3), donde se estudian mecanismos de estimación de la probabilidad de mora en créditos personales siguiendo la metodología propuesta por Cao et al. (2009). Según estos autores, la probabilidad de que un crédito se convierta en moroso no más tarde del instante $t+b$ sabiendo que aún está vivo en el tiempo t, o lo que es lo mismo, que no se ha producido la mora para $T \leq t$ y que el perfil crediticio $\boldsymbol{X}$ toma

el valor $\boldsymbol{x}$, se obtiene a partir de la fórmula:

$$\begin{aligned}\varphi\left(t|\boldsymbol{x}\right) &= P(t \leq T \leq t+b|T \geq t, \mathbf{X} = \boldsymbol{x}),\\ &= \frac{F(t+b|\boldsymbol{x}) - F(t|\boldsymbol{x})}{1 - F(t|\boldsymbol{x})},\\ &= 1 - \frac{S\left(t+b|\boldsymbol{x}\right)}{S\left(t|\boldsymbol{x}\right)}. \end{aligned} \tag{3.1}$$

Así definida, $PD\left(t|\boldsymbol{x}\right) = \varphi\left(t|\boldsymbol{x}\right)$ es la función de probabilidad de mora condicional evaluada en el punto $(t, t+b, \boldsymbol{x})$, donde b es el horizonte de tiempo para el que se quiere predecir la mora del crédito. Si bien, el horizonte de predicción, b, puede tomar cualquier valor no negativo, en la práctica las entidades trabajan con valores de b iguales o mayores a un año según lo recomendado por Basilea II.

En lo que resta de este capítulo, se proponen tres estimadores de la función de probabilidad de mora condicional, $\varphi\left(t|\boldsymbol{x}\right)$, y se estudia su aplicación a una base de datos reales de créditos personales. Los resultados obtenidos por Cao et al. (2009) se reproducen íntegramente al final del capítulo.

3.2.1. Estimador de la PD basado en un modelo lineal generalizado

Para obtener un estimador de la función de probabilidad de mora condicional, $\varphi\left(t|\boldsymbol{x}\right)$, definida en (3.1), se requiere previamente de un estimador de la función de distribución condicional, $F(t|\boldsymbol{x})$, para la que se define el siguiente modelo lineal generalizado. Se supone que la función de distribución condicional de T dada $\mathbf{X} = \boldsymbol{x}$, verifica la siguiente hipótesis:

H 3.8 *La función de distribución condicional del tiempo hasta la mora del crédito, $F\left(t|\boldsymbol{x}\right)$, se obtiene a partir de un modelo lineal generalizado definido por*

$$\begin{aligned}F\left(t|\boldsymbol{x}\right) &= P\left(T \leq t|\mathbf{X} = \boldsymbol{x}\right),\\ &= g\left(\theta_0 + \theta_1 t + \boldsymbol{\theta}'\boldsymbol{x}\right), \end{aligned} \tag{3.2}$$

donde la función de distribución g pertenece a la familia exponencial, θ_0 y θ_1 son parámetros de posición y de escala, respectivamente, y donde $\boldsymbol{\theta} = (\theta_2, \theta_3, \ldots, \theta_p)' \in \mathbb{R}^{p-1}$ es el vector de pesos de las covariables del modelo.

Típicamente, en el estudio de modelos lineales generalizados, las funciones de enlace son funciones de distribución invertibles pertenecientes a la familia exponencial (Nelder y Wedderburn (1972)). Tal es el caso, por ejemplo, de los modelos *probit* y *logit*, cuyas funciones de enlace corresponden a las funciones de distribución inversas de la normal estándar y la logística estándar, respectivamente. En general, cualquier función de distribución de probabilidad invertible puede ser utilizada como función de enlace en modelos de regresión paramétricos o semiparamétricos.

Estimación del modelo

Como se ha explicado anteriormente, para simplificar la metodología de estimación del modelo lineal generalizado (MLG) definido en (3.2), en adelante se considerará un modelo de probabilidad condicional con covariable univariante, X. Así, el vector de parámetros del modelo es $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)'$ y la función de distribución condicional de T dada $X = x$ viene dada por:

$$F\left(t|x\right) = g\left(\theta_0 + \theta_1 t + \theta_2 x\right), \tag{3.3}$$

cuya función de densidad se define por:

$$f(t|x) = \frac{\partial F\left(t|x\right)}{\partial t} = \theta_1 g'\left(\theta_0 + \theta_1 t + \theta_2 x\right).$$

Como se está trabajando sobre una muestra aleatoria $i.i.d.$ de datos censurados, la función de verosimilitud condicional se obtiene como el producto de los términos que implican la densidad condicional, en el caso de los datos no censurados (los tiempos de mora observados), y la función de supervivencia condicional, para aquellos tiempos correspondientes a datos censurados. Así, definiendo el vector $\boldsymbol{u}_i = (\xi_i, x_i, \delta_i)$ con $1 \leq i \leq n$, la función de verosimilitud condicional de la muestra se define como:

$$L(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n, \boldsymbol{\theta}) = \prod_{i=1}^{n} \left(f(\xi_i|x_i)\right)^{\delta_i} \left(1 - F(\xi_i|x_i)\right)^{1-\delta_i}, \tag{3.4}$$

donde ξ_i es el tiempo de vida observado del i-ésimo crédito, x_i es el valor observado de la puntuación crediticia del i-ésimo crédito y δ_i es la indicadora de que el i-ésimo crédito ha resultado moroso. El problema de estimación se resuelve buscando el valor de $\boldsymbol{\theta}$ que maximize el logaritmo natural de la

función definida en (3.4), es decir, la función de log-verosimilitud condicional definida como:

$$\begin{aligned}\ell(\boldsymbol{\theta}) = \ln\left(L(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n,\boldsymbol{\theta})\right) &= \sum_{i=1}^{n}\delta_i \ln\left(f(\xi_i|x_i)\right) \\ &\quad + \sum_{i=1}^{n}(1-\delta_i)\ln\left(1-F(\xi_i|x_i)\right) \\ = \sum_{i=1}^{n}\delta_i \ln\left(\theta_1 g'\left(\theta_0+\theta_1\xi_i+\theta_2 x_i\right)\right) &\\ &\quad + \sum_{i=1}^{n}(1-\delta_i)\ln\left(1-g\left(\theta_0+\theta_1\xi_i+\theta_2 x_i\right)\right) \\ = \sum_{i=1}^{n}\delta_i \left[\ln(\theta_1)+\ln\left(g'\left(\theta_0+\theta_1\xi_i+\theta_2 x_i\right)\right)\right] &\\ &\quad + \sum_{i=1}^{n}(1-\delta_i)\ln\left(1-g(\theta_0+\theta_1\xi_i+\theta_2 x_i)\right).\end{aligned}$$

El estimador del vector de parámetros del modelo se obtiene como el valor de $\boldsymbol{\theta}$ que maximiza la función de log-verosimilitud, $\ell(\boldsymbol{\theta})$, es decir, la solución del sistema de ecuaciones:

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_0} = \sum_{i=1}^{n}\frac{\delta_i g''\left(\theta_0+\theta_1\xi_i+\theta_2 x_i\right)}{g'\left(\theta_0+\theta_1\xi_i+\theta_2 x_i\right)} - \sum_{i=1}^{n}\frac{(1-\delta_i)\,g'\left(\theta_0+\theta_1\xi_i+\theta_2 x_i\right)}{1-g\left(\theta_0+\theta_1\xi_i+\theta_2 x_i\right)} = 0,$$

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1} &= \sum_{i=1}^{n}\frac{\delta_i}{\theta_1} + \sum_{i=1}^{n}\frac{\delta_i\xi_i g''\left(\theta_0+\theta_1\xi_i+\theta_2 x_i\right)}{g'\left(\theta_0+\theta_1\xi_i+\theta_2 x_i\right)} \\ &\quad - \sum_{i=1}^{n}\frac{(1-\delta_i)\,\xi_i g'\left(\theta_0+\theta_1\xi_i+\theta_2 x_i\right)}{1-g\left(\theta_0+\theta_1\xi_i+\theta_2 x_i\right)} = 0,\end{aligned}$$

y

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_2} &= \sum_{i=1}^{n}\frac{\delta_i x_i g''\left(\theta_0+\theta_1\xi_i+\theta_2 x_i\right)}{g'\left(\theta_0+\theta_1\xi_i+\theta_2 x_i\right)} \\ &\quad - \sum_{i=1}^{n}\frac{(1-\delta_i)\,x_i g'\left(\theta_0+\theta_1\xi_i+\theta_2 x_i\right)}{1-g\left(\theta_0+\theta_1\xi_i+\theta_2 x_i\right)} = 0.\end{aligned}$$

Calculado el valor del estimador $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ y sustituyéndolo en (3.3) se obtiene el estimador $\hat{F}_g(t|x)$ de la distribución condicionada $F(t|x)$ dado por:

$$\hat{F}_g(t|x) = g\left(\hat{\theta}_0 + \hat{\theta}_1 t + \hat{\theta}_2 x\right),$$

donde $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2\right)'$ es el estimador máximo verosímil de $\boldsymbol{\theta}$. Así, el estimador $\hat{S}_g(t|x)$ de la función de supervivencia condicionada, $S(t|x)$, se obtiene como

$$\hat{S}_g(t|x) = 1 - \hat{F}_g(t|x). \tag{3.5}$$

Reemplazando la expresión (3.5) en la fórmula (3.1) se obtiene el estimador máximo verosímil de la PD condicional, $\hat{\varphi}^{MLG}$, cuya fórmula viene dada por:

$$\hat{\varphi}^{MLG}(t|x) = 1 - \frac{\hat{S}_g(t+b|x)}{\hat{S}_g(t|x)}. \tag{3.6}$$

Modelo lineal generalizado con censura extrema

Alternativamente al modelo definido en (3.3), en un contexto en el que las muestras son altamente desbalanceadas (entre créditos morosos y no morosos), el tiempo hasta que se produce la mora del crédito puede llegar a ser infinito, es decir, que la mora del crédito no será observada. En la práctica, si se analizan los registros históricos de morosidad de las entidades financieras, se observa que en períodos de tiempo sin crisis económicas o finacieras, el suceso de entrada en mora de un crédito es poco frecuente, resultando en una proporción de créditos morosos muy pequeña. Esto puede interpretarse como que el tiempo de supervivencia del crédito es muy grande, o incluso infinito, comparado con la duración del estudio. Como consecuencia de esto, la mayoría de los tiempos hasta que se produce la mora resultan ser datos censurados. Por esta razón, es necesario que el modelo que se quiere ajustar para estimar la función de distribución condicionada, $F(t|x)$, tenga en cuenta esa alta proporción de censura (ver discusión hecha por el profesor Jan Beran en Cao et al. (2009), págs. 39-40).

Siguiendo las ideas propuestas por Beran (2009), en este apartado se propone utilizar una variante del modelo (3.3) denominado modelo lineal

generalizado modificado ($MLGM$). Para ello se adopta el enfoque de supervivencia condicional con censura extrema estudiado por Maller y Zhou (1996) y Beran y Djaïdja (2007), entre otros.

Bajo este nuevo enfoque, se supondrá que las observaciones de los tiempos hasta la entrada en mora provienen de una mezcla de distribuciones compuesta por una alta proporción, q, de *créditos inmunes al riesgo de morosidad* y por una proporción más pequeña, $1-q \ll q$, de créditos propensos a convertirse en morosos. Este enfoque entra de lleno en los llamados modelos de curación dentro del análisis de supervivencia (ver Farewell (1982)).

Las siguientes hipótesis dan validez a este modelo:

H 3.9 *El tiempo hasta la entrada en mora, T, puede representarse como una mezcla entre un punto con masa de probabilidad en $T=\infty$ y una variable aleatoria continua no negativa, W, con función de distribución, g_W, conocida. Así, la variable aleatoria T puede ser escrita como*

$$T = \zeta \cdot \infty + (1-\zeta)W, \tag{3.7}$$

donde ζ es la variable aleatoria indicadora de inmunidad que es independiente de W. La función de masa de probabilidad de ζ viene dada por $P(\zeta=1) = 1-P(\zeta=0) = q$.

Observación 3.1 (i) *La expresión $T=\infty$ en H3.9 puede interpretarse como que el tiempo en el que se produce la mora es mayor que el tiempo de duración del estudio, y por tanto, la expresión $T=\zeta\cdot\infty$ en (3.7) puede reemplazarse por $T=\zeta\cdot\tau_\infty$, para algún $\tau_\infty > \max\limits_{\delta_i=0}\{\xi_i\}_{i=1}^n \in \mathbb{R}^+$.* (ii) *La cantidad, q, puede interpretarse como la probabilidad de que un crédito no entre en mora en el intervalo $[0,\tau_\infty]$, es decir, créditos cuya probabilidad conjunta $P(\delta=1,\zeta=1)=0$. Como consecuencia, se obtiene que $P(T\leq C|\zeta=1) = P(\delta=1|\zeta=1) = P(\delta=1,\zeta=1)/P(\zeta=1)=0$. Además, q es distinta, en general, de la probabilidad de que una observación sea censurada, ya que, T, puede ser censurada en ausencia de inmunidad $(\zeta=0)$ y por tanto, $P(\delta=0,\zeta=0)>0$, en general. Si bien, la proporción de datos censurados en la muestra es aleatoria, esta es siempre conocida debido a que el conjunto $\{(\xi_i,x_i,\delta_i)\}_{i=1}^n$ es siempre observable.* (iii) *En la hipótesis H3.9, la función de distribución, g_W, de la variable W se supondrá perteneciente a la familia exponencial manteniendo el enfoque clásico de los modelos lineales generalizados (Nelder y Wedderburn (1972)).*

H 3.10 *Suponiendo que la probabilidad condicional* $q(x) = P(\zeta = 1|X = x) = q$*, no depende de* x*, la función de distribución condicional del tiempo hasta la mora depende de la probabilidad,* q*, y de la de la función de distribución,* g_W*, de la variable,* W*. Entonces, usando la ley de probabilidad total se obtiene*

$$\begin{aligned} P(T \leq t|X = x) &= P(T \leq t|\zeta = 0, X = x)\, P(\zeta = 0|X = x) \\ &\quad + P(T \leq t|\zeta = 1, X = x)\, P(\zeta = 1|X = x) \\ &= P(W \leq t|\zeta = 0, X = x)\,(1 - q(x)) \\ &\quad + P(\infty \leq t|\zeta = 1, X = x)\, q(x) \\ &= (1 - q)\, g_W(t|x). \end{aligned} \tag{3.8}$$

Así, bajo las hipótesis H3.9-H3.10, tomando $X \in \mathbb{R}$, existen parámetros $(\theta_0, \theta_1, \theta_2) \in \mathbb{R}^3$ tales que la función de distribución condicional, $F(t|x)$, viene dada por:

$$F(t|x) = (1 - q)\, g_W(\theta_0 + \theta_1 t + \theta_2 x), \tag{3.9}$$

cuya función de densidad está dada por:

$$f(t|x) = \frac{\partial F(t|x)}{\partial t} = (1 - q)\,\theta_1 g'_W(\theta_0 + \theta_1 t + \theta_2 x)$$

y la función de log-verosimilitud se define como:

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n} [\delta_i \ln(f(\xi_i|x_i)) + (1 - \delta_i) \ln(1 - F(\xi_i|x_i))] \\ &= \sum_{i=1}^{n} \delta_i [\ln(1 - q) + \ln(\theta_1)] + \sum_{i=1}^{n} \delta_i \ln[g'_W(\theta_0 + \theta_1 \xi_i + \theta_2 x_i)] \\ &\quad + \sum_{i=1}^{n} (1 - \delta_i) \ln[1 - (1 - q)\, g_W(\theta_0 + \theta_1 \xi_i + \theta_2 x_i)], \end{aligned} \tag{3.10}$$

donde $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, q)'$ es el vector de parámetros a estimar.

El estimador máximo verosímil, $\hat{\boldsymbol{\theta}}$, de $\boldsymbol{\theta}$ se obtiene resolviendo el siguiente sistema de ecuaciones:

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_0} &= \sum_{i=1}^{n} \frac{\delta_i g''_W(\theta_0 + \theta_1 \xi_i + \theta_2 x_i)}{g'_W(\theta_0 + \theta_1 \xi_i + \theta_2 x_i)} \\ &\quad - \sum_{i=1}^{n} \frac{(1 - \delta_i)(1 - q)\, g'_W(\theta_0 + \theta_1 \xi_i + \theta_2 x_i)}{1 - (1 - q)\, g_W(\theta_0 + \theta_1 \xi_i + \theta_2 x_i)} = 0, \end{aligned}$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1} = \sum_{i=1}^{n} \frac{\delta_i}{\theta_1} + \sum_{i=1}^{n} \frac{\delta_i \xi_i g''_W(\theta_0 + \theta_1 \xi_i + \theta_2 x_i)}{g'_W(\theta_0 + \theta_1 \xi_i + \theta_2 x_i)} - \sum_{i=1}^{n} \frac{(1-\delta_i)\xi_i (1-q) g'_W(\theta_0 + \theta_1 \xi_i + \theta_2 x_i)}{1-(1-q) g_W(\theta_0 + \theta_1 \xi_i + \theta_2 x_i)} = 0,$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_2} = \sum_{i=1}^{n} \frac{\delta_i x_i g''_W(\theta_0 + \theta_1 \xi_i + \theta_2 x_i)}{g'_W(\theta_0 + \theta_1 \xi_i + \theta_2 x_i)} - \sum_{i=1}^{n} \frac{(1-\delta_i) x_i (1-q) g'_W(\theta_0 + \theta_1 \xi_i + \theta_2 x_i)}{1-(1-q) g_W(\theta_0 + \theta_1 \xi_i + \theta_2 x_i)} = 0,$$

y

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial q} = -\sum_{i=1}^{n} \frac{\delta_i}{(1-q)} + \sum_{i=1}^{n} \frac{(1-\delta_i) g_W(\theta_0 + \theta_1 \xi_i + \theta_2 x_i)}{1-(1-q) g_W(\theta_0 + \theta_1 \xi_i + \theta_2 x_i)} = 0.$$

Bajo el modelo $MLGM$, definido en (3.9), el estimador máximo verosímil de $F(t|x)$, denotado por $\hat{F}_W(t|x)$, viene dado por:

$$\hat{F}_W(t|x) = (1-\hat{q})\, g_W\left(\hat{\theta}_0 + \hat{\theta}_1 t + \hat{\theta}_2 x\right), \tag{3.11}$$

donde $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2, p\right)'$ es la solución del sistema de ecuaciones descrito anteriormente.

Análogamente a la fórmula obtenida en (3.5), el estimador de la función de supervivencia condicionada, $\hat{S}_W(t|x)$, de $S(t|x)$ se obtiene como:

$$\hat{S}_W(t|x) = 1 - \hat{F}_W(t|x). \tag{3.12}$$

Cabe recordar que la cantidad $(1-\hat{q})$ es una estimación de la proporción de créditos carentes de inmunidad al riesgo de morosidad que, en general, será muy pequeña en comparación con $\hat{q}$, y por tanto, $0 < \hat{F}_W(t|x) < 1$ cuando el máximo tiempo de vida observado en la muestra es censurado.

Reemplazando la expresión (3.12) en la fórmula (3.1) se obtiene el estimador máximo verosímil de la PD condicional, $\hat{\varphi}^{MLGM}$, bajo el modelo (3.9), es decir,

$$\hat{\varphi}^{MLGM}(t|x) = 1 - \frac{\hat{S}_W(t+b|x)}{\hat{S}_W(t|x)}. \tag{3.13}$$

3.2.2. Estimador de la PD basado en un modelo de riesgos proporcionales de Cox

En este apartado se propone un método semiparamétrico para el estudio de la función de probabilidad de mora condicional, $\varphi(t|x)$. Este método consiste en utilizar el modelo de riesgos proporcionales de Cox (PHM) para estimar la función de supervivencia condicional del tiempo de vida de los créditos, $S(t|x)$. Recordando la fórmula definida en (1.2) (Sección 1.2.2 del Capítulo 1), las funciones de supervivencia condicional, $S(t|x)$, y de fallo acumulativo condicional, $\Lambda(t|x)$, verifican la siguiente igualdad:

$$S(t|x) = \exp(-\Lambda(t|x)), \tag{3.14}$$

donde $\Lambda(t|x) = \int_0^t \lambda(s|x)\, ds$, siendo $\lambda(t|x)$ la función razón de fallo condicional. En el contexto de modelos de regresión de Cox para el riesgo de crédito, la función $\lambda(t|x)$ ofrece una medida de la tasa de mora instantánea del crédito en el instante $T = t$ condicionada al perfil crediticio $X = x$.

En un modelo de regresión de Cox, bajo a las hipótesis H3.4-H3.6, se define el estimador de la función de fallo acumulativo condicional, $\Lambda(t|x)$, como:

$$\hat{\Lambda}(t|x) = \hat{\Lambda}_0(t) \exp\left(\hat{\beta}x\right), \tag{3.15}$$

donde $\hat{\Lambda}_0(t)$ es el estimador no paramétrico de la función de fallo acumulativo basal, $\Lambda_0(t)$, y $\hat{\beta}$ es el estimador de máxima verosimilitud del coeficiente de la regresión de Cox, β.

El estimador usual de $\Lambda_0(t)$ es el de Nelson-Aalen definido por:

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \frac{I(\xi_i \leq t, \delta_i = 1)}{\sum_{j=1}^{n} I\left(\xi_j \geq \xi_i\right)}. \tag{3.16}$$

Por otro lado, el estimador, $\hat{\beta}$, se obtiene como el valor que maximiza la función de verosimilitud parcial de Cox, $L(\beta)$, definida por:

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta x_i)}{\sum_{j=1}^{n} I\left(\xi_j \geq \xi_i\right) \exp(\beta x_j)} \right)^{\delta_i}. \tag{3.17}$$

La ecuación de verosimilitud parcial obtenida a partir de (3.17) puede escribirse como:

$$\frac{d\ln\left(L(\beta)\right)}{d\beta} = \sum_{i=1}^{n} \delta_i \left[x_i - \left(\sum_{j=1}^{n} x_j \varpi_{nj}\left(\xi_i;\beta\right) \right) \right] = 0, \tag{3.18}$$

donde se definen los pesos $\varpi_{nj}\left(t;\beta\right)$ como:

$$\varpi_{nj}\left(t;\beta\right) = \frac{I\left(\xi_j \geq t\right)\exp\left(\beta x_j\right)}{\sum_{k=1}^{n} I\left(\xi_k \geq t\right)\exp\left(\beta x_k\right)}.$$

Resolviendo la ecuación definida en (3.18) se obtiene el estimador máximo verosímil, $\hat{\beta}$, de β. Sustituyendo $\hat{\beta}$ y la expresión (3.16) en la fórmula (3.15) se obtiene el estimador de la función de fallo acumulativo, $\hat{\Lambda}(t|x)$. Así, sustituyendo la expresión (3.15) en la ecuación (3.14), se obtiene el estimador de la función de supervivencia condicional, $\hat{S}^{PHM}\left(t|x\right)$, definido por:

$$\hat{S}^{PHM}\left(t|x\right) = \exp\left(-\hat{\Lambda}(t|x)\right). \tag{3.19}$$

Finalmente, insertando la expresión (3.19) en la fórmula (3.1) se obtiene el estimador de la PD condicional bajo un modelo de riesgos proporcionales de Cox, $\hat{\varphi}^{PHM}(t|x)$, cuya fórmula viene dada por:

$$\hat{\varphi}^{PHM}(t|x) = 1 - \frac{\hat{S}^{PHM}\left(t+b|x\right)}{\hat{S}^{PHM}(t|x)}. \tag{3.20}$$

3.2.3. Estimación de la PD basada en el estimador producto límite generalizado de Beran

En este aparado se propone un enfoque de modelización no paramétrico para la función de probabilidad de mora condicional, $\varphi(t|x)$. Una ventaja importante que presenta el enfoque no paramétrico en comparación con los dos modelos anteriores es que el modelo no paramétrico no requiere de supuestos acerca de las distribuciones de T, C o X.

En este capítulo se utiliza el estimador no paramétrico tipo núcleo conocido como estimador producto límite generalizado (PLG) de Beran (1981).

Según la literatura, Beran (1981) fue el primero en estudiar el problema de la regresión con datos censurados bajo un enfoque completamente no paramétrico. El estimador PLG propuesto por Beran (1981) es una generalización del estimador producto límite de Kaplan y Meier (1958), razón por la que algunos autores también lo denominan estimador de Kaplan-Meier condicional o estimador de Kaplan-Meier generalizado. Muchos investigadores han contribuido al estudio de este estimador en el contexto de la estimación no paramétrica de la función de distribución condicional con datos censurados, tanto en diseño fijo como aleatorio. El estudio de este estimador en diseño fijo se encuentra, por ejemplo, en los trabajos de González Manteiga y Cadarso Suárez (1994), Van Keilegom y Veraverbeke (1997b) y Van Keilegom (1998). Además, las propiedades asintóticas del estimador de Beran (1981) en diseño aleatorio han sido estudiadas por Dabrowska (1987, 1989), McKeague y Utikal (1990), Dabrowska (1992a, 1992b), Akritas (1994), Li y Doss (1995), Van Keilegom y Veraverbeke (1996), Iglesias Pérez y González Manteiga (1999), Van Keilegom et al. (2001), Du y Akritas (2002), Iglesias Pérez (2001) y Strzalkowska-Kominiak y Cao (2014), entre otros. En todos ellos se aborda el problema de la estimación no paramétrica de la distribución condiconal, $F(t|x)$, con covariable unidimensional. El estudio de una extensión del estimador PLG de Beran al caso en que la covariable es multidimensional se encuentra en el trabajo de Lopez (2011), quien propone un estimador que permite enfrentar el problema de *la maldición de la dimensionalidad* añadiendo condiciones adicionales a las hipótesis tradicionales (H3.5 y H3.6) sobre la variable censurante, C.

El estimador PLG de Beran de la función de supervivencia del tiempo hasta la entrada en mora, denotado por $\hat{S}_h^{PLG}$, se define por:

$$\hat{S}_h^{PLG}(t|x) = \prod_{i=1}^{n} \left(1 - \frac{I\left(\xi_i \leq t, \delta_i = 1\right) B_{ih}(x)}{1 - \sum_{j=1}^{n} I\left(\xi_j < \xi_i\right) B_{jh}(x)} \right), \tag{3.21}$$

para todo $t \leq \xi_{(n)} = \max_{\delta_i=1} \{\xi_i\}$, donde ξ_i es el tiempo de vida observado del i-ésimo crédito, δ_i es la variable indicadora de morosidad de i-ésimo crédito y los términos $B_{ih}(x)$ son pesos no paramétricos de tipo Nadaraya-Watson definidos como:

$$B_{ih}(x) = \frac{K((x - x_i)/h)}{\sum_{j=1}^{n} K((x - x_j)/h)}, \ 1 \leq i \leq n, \tag{3.22}$$

donde $K(\cdot)$ es una función de densidad de probabilidad llamada *función núcleo*, x_i es el valor de la covariable X observada en el crédito i-ésimo y h es el parámetro de suavizado del estimador núcleo. Tanto la función núcleo, K, como el parámetro de suavizado, h, son elementos cruciales para la estimación no pramétrica de las funciones $S(t|x)$ y $\varphi(t|x)$, por lo que han de cumplir ciertas condiciones de regularidad que garantizarán la correcta definición de los estimadores $\hat{S}_h^{PLG}(t|x)$ y $\hat{\varphi}_n^{PLG}(t|x)$.

La ventana de suavizado, $h \equiv h_n$, es una sucesión de valores reales positivos que depende del tamaño muestral, n, y su función es controlar el grado de suavidad de la estimación no paramétrica actuando como un modulador entre el sesgo (sobresuavizado) y la varianza (infrasuavizado) de la misma. El parámetro h debe verificar las condiciones:

$$h \longrightarrow 0 \text{ y } nh \longrightarrow \infty \text{ cuando } n \longrightarrow \infty.$$

En general, un valor adecuado de h es aquel que permite obtener un equilibrio óptimo entre el sesgo y la varianza de la estimación no paramétrica.

Insertando la expresión (3.21) en la fórmula (3.1), se obtiene como resultado un estimador tipo núcleo de la PD condicional basado en el estimador PLG de Beran. Llamando a este estimador, $\hat{\varphi}_n^{PLG}$, su fórmula viene dada por:

$$\hat{\varphi}_n^{PLG}(t|x) = 1 - \frac{\hat{S}_h^{PLG}(t+b|x)}{\hat{S}_h^{PLG}(t|x)}. \tag{3.23}$$

Hipótesis para la función núcleo

En la estimación no paramétrica de curvas, la función núcleo es una función no negativa, típicamente una función de densidad de probabilidad, que se utiliza para asignar pesos a las observaciones muestrales en la vecindad local considerada para la estimación no paramétrica, cuya amplitud es controlada por el parámetro h. En lo sucesivo, se utilizan funciones núcleo, K, que verifican las siguientes hipótesis:

H 3.11 *K es absolutamente continua y diferenciable con soporte compacto denotado por $\Omega_K \subseteq \mathbb{R}$.*

H 3.12 *K es una densidad de probabilidad simétrica respecto del origen, y por tanto, satisface las siguientes condiciones:*

(i) $\int_{\Omega_K} K(u)\,du = 1$

(ii) $K(u) = K(-u)$ *para todo* $u \in \Omega_K$

(iii) El momento de orden j de K se denota por $\mu_j(K) = \int u^j K(u)\,du$. *Entonces* $\mu_j(K) = 0$ *para todo j impar*

Observación 3.2 *Como K es continua y Ω_K es compacto (H3.11), entonces la imagen por K de Ω_K es compacta, y en particular, está acotada. Además, si $u \notin \Omega_K$ entonces $K(u) = 0$, por lo que K está acotada en todo $\mathbb{R}$. Como consecuencia, se obtiene que (i)* $d_K = \int u^2 K(u)\,du \leq \int_{\Omega_K} u^2 \|K\|_\infty du \leq \|K\|_\infty \int_{\Omega_K} u^2 du < \infty$ *y que (ii) el coeficiente de rugosidad de K denotado por* $c_K = \int K^2(u)\,du \leq \int_{\Omega_K} \|K\|_\infty^2 du \leq \|K\|_\infty^2 \int_{\Omega_K} du = \|K\|_\infty^2 \mu(\Omega_K) < \infty$.

La literatura sobre estimación no paramétrica de curvas recoge varios ejemplos de funciones tipo núcleo de segundo orden (Silverman (1986), Scott (1992), Wand y Jones (1995)). Los siguientes son algunos ejemplos de funciones núcleo que pueden utilizarse en la fórmula definida en (3.22).

Tabla 3.1. Funciones de densidad núcleo de segundo orden más comunes

Función núcleo	$K(u)$	d_K^2	c_K
Gaussiano	$1/\sqrt{2\pi}\exp(-u^2/2)I(\lvert u\rvert < \infty)$	1	$1/\sqrt{2\pi}$
Uniforme	$(1/2)\,I(\lvert u\rvert \leq 1)$	1/3	1/2
Triangular	$(1-\lvert u\rvert)\,I(\lvert u\rvert \leq 1)$	1/6	2/3
Epanechnikov	$(3/4)\,(1-u^2)\,I(\lvert u\rvert \leq 1)$	1/5	3/5
Biweight	$(15/16)\,(1-u^2)^2\,I(\lvert u\rvert \leq 1)$	1/7	5/7
Triweight	$(35/32)\,(1-u^2)^3\,I(\lvert u\rvert \leq 1)$	1/9	350/429
Tricubo	$(70/81)\left(1-\lvert u\rvert^3\right)^3 I(\lvert u\rvert \leq 1)$	35/243	175/247
Coseno	$(\pi/4)\cos(\pi u/2)\,I(\lvert u\rvert \leq 1)$	$(\pi^2-8)/\pi^2$	$\pi^2/16$

Es conocido en la literatura sobre estimación no paramétrica de curvas que la elección de la función núcleo no influye de manera sustancial en el grado de

suvizamiento obtenido, aunque se demuestra que el núcleo de *Epanechnikov* verifica ciertas propiedades de optimalidad que lo hacen preferible frente a otras funciones núcleo (Hodges y Lehman (1956)). En la práctica, se verifica que las dos funciones núcleo que más se utilizan en estudios empíricos y de simulación son las de *Epanechnikov* y *Gaussiano*.

Elección del parámetro de suavizado

La elección del *parámetro de suavizado* (o ancho de la ventana) es un problema de gran relevancia en estimación no paramétrica de curvas con técnicas tipo núcleo, y por tanto, es un problema crucial en la estimación de la función de probabilidad de mora condicional, $\varphi(t|x)$. Como el estimador no paramétrico $\hat{\varphi}_n^{PLG}$ depende funcionalmente del estimador $\hat{S}_h^{PLG}$, es importante mencionar que la elección óptima del parámetro de suavizado de $\hat{S}_h^{PLG}$ no garantiza que dicho valor también sea óptimo para el estimador $\hat{\varphi}_n^{PLG}$. Más adelante, en el Capítulo 4 se volverá a este punto, donde se demuestra la diferencia (en términos del error cuadrático medio asintótico) entre ambas cantidades. En virtud de lo anterior, en esta sección se propone estudiar la implementación de un método de selección automática del parámetro de suavizado, h, tanto para $\hat{S}_h^{PLG}$ como para $\hat{\varphi}_n^{PLG}$.

En la actualidad, existen varios métodos automáticos para elegir el valor del parámetro de suavizado del estimador núcleo de la función de densidad, de la función de regresión y de la función de distribución, entre otras funciones de interés en el análisis de supervivencia. Entre los más utilizados se encuentran el *método plug-in*, los métodos de *validación cruzada* y los *métodos de remuestreo*, de los cuales el más conocido es el método *bootstrap*. Algunos trabajos en los que se ha estudiado el problema de la estimación no paramétrica de la función de distribución son los debidos a Reiss (1981), Jones (1990), Lejeune y Sarda (1992), Sarda (1993), Altman y Léger (1995) y Bowman et al. (1998), entre otros.

En el contexto de la función de distribución condicional con datos censurados, la literatura es menos extensa. Algunos de los trabajos en los que se estudia este problema son los debidos a Dabrowska (1992a), Leconte et al. (2002), Cai (2003), Gannoun et al. (2005), Gannoun et al. (2007) y Lopez (2011), entre otros.

Método bootstrap con datos censurados

Antes se han mencionado algunos de los métodos que tradicionalmente se utilizan para elegir el parámetro, h, de suavizado en la estimación no paramétrica tipo núcleo. Sin embargo, algunos de ellos no siempre son aplicables cuando se analizan problemas con datos reales. Por ejemplo, la utilización de métodos de *validación cruzada*, en general, no es recomendable cuando el tamaño de la muestra utilizada es grande, como ocurre normalmente con las bases de datos de las entidades financieras.

Las técnicas de sustitución, o *plug-in*, tampoco parecen ser recomendables cuando se trabaja con bases de datos extensas. La utilización de este método requiere de la estimación de cantidades poblacionales que dependen de funciones desconocidas que deben ser estimadas a partir de los datos, lo que lleva a procedimientos complejos que hacen que el proceso de estimación sea laborioso y excesivamente costoso en tiempos de computación.

Alternativamente, las técnicas de remuestreo de tipo *bootstrap* se han hecho cada vez más populares entre los métodos de selección automática del parámetro de suavizado, h, sin ofrecer grandes limitaciones debido al tamaño de la muestra utilizada. Esta técnica consiste en fijar una muestra de datos, definida por $\mathcal{M}_n = \{(\xi_i, \delta_i, X_i)\}_{i=1}^n$, y aprovechar la información allí contenida tomando sucesivas remuestras de tamaño n de $\mathcal{M}_n$ escogidas aleatoriamente, denotadas por $\mathcal{M}_n^{*b} = \left\{\left(\xi_i^{*b}, \delta_i^{*b}, X_i^{*b}\right)\right\}_{i=1}^n$, con $b = 1, .., B$, donde la terna $\left(\xi_i^{*b}, \delta_i^{*b}, X_i^{*b}\right)$ es la información de el i-ésimo acreditado escogido en la b-ésima réplica de $\mathcal{M}_n$. Una vez obtenidos los estimadores $\hat{F}_h(t|x)$ y $\hat{\varphi}_n^{PLG}(t|x)$ con la muestra original, $\mathcal{M}_n$, se emplea el método bootstrap para calcular un número, B, suficientemente grande de estimaciones bootstrap de $F(t|x)$ y $\varphi(t|x)$ a partir de las remuestras $\mathcal{M}_{nb}^*$, de donde se obtendrá el valor (óptimo) bootstrap del parámetro de suavizado h. Existe una extensa literatura sobre aplicaciones de este método en la estimación no paramétrica de las funciones de densidad, de distribución y de regresión, entre otras. Ver, por ejemplo, las contribuciones sobre este tema debidas a Freedman (1981), Härdle y Bowman (1988), Taylor (1989), Faraway y Jhun (1990), Falk (1992), Marron (1992), Cao y González Manteiga (1993) y Delaigle y Gijbels (2004), por citar sólo algunos trabajos. Por otra parte, la extensión de este problema al caso de datos censurados ha sido estudiada, entre otros, por Van Keilegom y Veraverbeke (1997a, 1998), Cao et al. (2001), Li y Datta (2001) e Iglesias Pérez y González Manteiga (2003).

El objetivo es elegir un valor para el parámetro de suavizado que minimice el error cuadrático medio (ECM) del estimador $\hat{\varphi}_n^{PLG}(t|x)$. Esto no siempre es fácil y algunos métodos tratan de aproximar el valor del parámetro de suavizado que minimiza la expresión asintótica del error cuadrático medio asintótico ($ECMA$). El método de remuestreo se utiliza para aproximar el ECM del estimador $\hat{\varphi}_n^{PLG}(t|x)$, por su análogo bootstrap, $ECM_{t,x}^{*}(h)$, y obtener así una estimación del valor óptimo de la ventana de suavizado de $\hat{\varphi}_n^{PLG}(t|x)$. Ver, por ejemplo, los trabajos de Hall (1990) y Cao (1993) quienes han estudiado este método en el contexto de la estimación no paramétrica de la función de densidad.

Las ideas de Cao (1993) pueden adaptarse para obtener el parámetro óptimo de suavizado bootstrap del estimador $\hat{F}_h^{PLG}(t|x)$ y posteriormente estimar no paramétricamente la función $\varphi(t|x)$. Este método requiere del uso de dos ventanas piloto, g_1 y g_2, para estimar $F(t|x)$ y $G(t|x)$, respectivamente, y una tercera ventana piloto, g_X, para estimar la función de densidad de la covariable, $m(x)$.

La elección de una ventana piloto óptima es un problema de investigación abierto sobre el que no se profundizará ya que se escapa del alcance de esta memoria, dejándose propuesto como una posible extensión para futuros trabajos.

En lo sucesivo, por simplicidad de notación, no se hará distinción entre $\hat{F}_h^{PLG}(t|x)$ y $\hat{F}_h(t|x)$ para referise al estimador PLG de Beran de $F(t|x)$. Asimismo, no habrá distinción entre $\hat{G}_h^{PLG}(t|x)$ y $\hat{G}_h(t|x)$, y entre $\hat{\varphi}_n^{PLG}(t|x)$ y $\hat{\varphi}_n(t|x)$, para referirse a los estimadores PLG de $G(t|x)$ y de $\varphi(t|x)$, respectivamente.

Ventana de suvizado bootstrap del estimador PLG de la PD

El algoritmo de remuestreo propuesto por Cao et al. (2009) para obtener la ventana de suavizado óptima bootstrap se compone de los siguientes pasos:

1. Calcular $\hat{F}_{g_1}(t|x)$, el estimador de Beran de $F(t|x)$ y $\hat{G}_{g_2}(t|x)$, el estimador de Beran de $G(t|x)$.

2. Estimar $m(x)$ por medio de $\hat{m}_{g_X}(x)$, estimador de Parzen-Rosenblatt.

3. Generar una muestra aleatoria con reemplazo, $(X_1^*, X_2^*, \ldots, X_n^*)$, de la función de densidad estimada, $\hat{m}_{g_X}(x)$.

4. Para cada $i = 1, 2, \ldots, n$, generar T_i^* a partir de $\hat{F}_{g_1}(t|x_i^*)$ y C_i^* a partir de $\hat{G}_{g_2}(t|x_i^*)$.

5. Para cada $i = 1, 2, \ldots, n$, calcular $\xi_i^* = \min\{T_i^*, C_i^*\}$ y la indicadora $\delta_i^* = I\left(T_i^* \leq C_i^*\right)$.

6. Usar la muestra $\{(\xi_1^*, \delta_1^*, X_1^*), (\xi_2^*, \delta_2^*, X_2^*), \ldots, (\xi_n^*, \delta_n^*, X_n^*)\}$ para calcular $\hat{\varphi}_{g_1}^*(t|x)$, el análogo bootstrap de $\hat{\varphi}_{n,h}^*(t|x)$.

7. Aproximar el error cuadrático medio asintótico de $\hat{\varphi}_n^{PLG}(t|x)$ por su versión bootstrap:

$$ECM_{t,x}^*(h) = E^*\left[\left(\hat{\varphi}_{n,h}^*(t|x) - \hat{\varphi}_{g_1}(t|x)\right)^2\right]. \tag{3.24}$$

8. El error cuadrático medio bootstrap definido en (3.24) puede ser aproximado tomando un número grande, digamos B, de réplicas bootstrap de la muestra original siguiendo los pasos 4-6 y calculando:

$$\widehat{ECM}_{t,x}^*(h) = \frac{1}{B}\sum_{j=1}^{B}\left(\hat{\varphi}_{n,h}^{*j}(t|x) - \hat{\varphi}_{g_1}(t|x)\right)^2. \tag{3.25}$$

9. Finalmente, la ventana óptima bootstrap, $h_{ECM,t,x}^*$, se obtiene minimizando la expresión (3.25) con respecto a h, esto es

$$h_{ECM,t,x}^* = \arg\min_{h>0} \widehat{ECM}_{t,x}^*(h).$$

Como este algoritmo de remuestreo puede llegar a ser demasiado costoso en tiempo de computación cuando el tamaño de la muestra es muy grande, una forma de remediar este problema, por ejemplo tomando $n = 25\,000$, es la que se describe a continuación:

Se toma una submuestra de tamaño $m < n$, por ejemplo, con $m = 2\,500$, es decir, que $n = \lambda m$ con un valor de λ típicamente grande, en este ejemplo $\lambda = 10$. Aplicando este algoritmo de remuestreo se obtiene la ventana bootstrap, $h^*_{ECM,m,t,x}$, para $m = 2\,500$. Finalmente, utilizando la fórmula (4.9) en el Capítulo 4, y motivado por la fórmula asintótica para la ventana bootstrap óptima local:

$$h^*_{ECM,n,t,x} = \lambda^{-1/5} h^*_{ECM,m,t,x} \ , \tag{3.26}$$

para tamaños de muestra m y n tales que $\lambda = n/m > c$ para algún entero $c > 1$.

El estudio de las propiedades asintóticas del estimador $\widehat{ECM}^*_{t,x}(h)$ y de la ventana $h^*_{ECM,t,x}$, es un problema de gran relevancia estadística, cuyo tratamiento puede llegar a ser analíticamente muy complejo. Por este motivo, su estudio se ha dejado como parte de un trabajo futuro en el que pueden adoptarse, por ejemplo, las ideas de Li y Datta (2001) e Iglesias Pérez y González Manteiga (2003), quienes han estudiado este problema en el contexto de la estimación bootstrap de la función de distribución condicional con datos censurados.

3.3. Aplicación a una cartera de créditos personales

En este capítulo se estudian tres métodos de estimación de la función de probabilidad de mora condicional, $\varphi(t|x)$, donde la muestra analizada exhibe una alta proporción de datos censurados por la derecha. Con el fin de evaluar el comportamiento empírico de dichos estimadores, en esta sección se presenta un análisis comparativo de éstos por medio de estadísticas descriptivas y curvas de PD estimadas. En el análisis se han utilizado distintos valores de tiempo hasta la entrada en mora, T, de puntuación crediticia, X, de funciones de enlace (para los estimadores MLG y $MLGM$) y del parámetro de suavizado h (para el estimador PLG). Los resultados obtenidos se presentan a continuación.

3.3.1. Análisis de la base de datos

Los datos utilizados corresponden a una muestra de tamaño $n = 25\,000$ de préstamos personales de una entidad financiera española formalizados entre julio de 2004 y noviembre de 2006. Por motivos de confidencialidad con la entidad colaboradora, los datos de la muestra fueron tomados aleatoriamente cambiando la probabilidad de seleccionar un crédito moroso. De esta forma, la tasa de morosidad de la cartera de créditos original fue distorsionada arbitrariamente y no refleja su estado real de solvencia. La tasa de mora observada en la muestra de desarrollo fue de un 7.2 %.

Se considera que la muestra utilizada es representativa de dos poblaciones de acreditados:

(*i*) Π_1: Es el conjunto de todos los préstamos personales emitidos por la entidad entre julio de 2004 y noviembre de 2006 cuya fecha de entrada en mora fue registrada dentro de ese período de tiempo. Así, Π_1 representa la población de datos no censurados.

(*ii*) Π_0: Es el conjunto conformado por dos tipos de préstamos. Los emitidos por la entidad entre julio de 2004 y noviembre de 2006 cuya fecha de vencimiento, o plazo, fue anterior a noviembre de 2006 y en los que no se registró la mora, y los préstamos con fecha de vencimiento posterior a noviembre de 2006 cuya fecha de entrada en mora nunca se observó, o bien, fue registrada en una fecha posterior a noviembre de 2006. Así, Π_0 representa la población de datos censurados. Por último, $\Pi = \Pi_0 \bigcup \Pi_1$ representa el conjunto de todos los créditos formalizados entre julio de 2004 y noviembre de 2006.

Las variables analizadas en este estudio son:

(*iii*) ξ_i es la madurez o tiempo de vida observado del i-ésimo crédito. Si el crédito $i \in \Pi_1$, lo que se observa es $\xi_i = T_i$, es decir, el tiempo hasta que se produce la mora del crédito i-ésimo. Por el contrario, si el crédito $i \in \Pi_0$, entoces se observa el tiempo de censura, $\xi_i = C_i$, es decir, el tiempo de vida del crédito i-esimo hasta la cancelación de este, ya sea en el vencimiento del plazo o de forma anticipada. Esta variable se ha medido en meses y la profundidad de la muestra es aproximadamente de 30 meses.

(*iv*) X_i es la variable puntuación crediticia y mide el grado de propensión a cometer impago del crédito i-ésimo. El rango de X_i es el intervalo $[0, 100]$,

donde $X_i = 0$ representa la ausencia de riesgo de mora del acreditado i-ésimo y $X_i = 100$ representa el 100 % de certeza de que el crédito i-ésimo será moroso. Así, cuanto más cercano a 100 es el valor de X_i, más propenso a convertirse en moroso es el crédito i-ésimo.

(v) δ_i es la variable indicadora de morosidad (no censura) asociada al crédito i-ésimo. Si el crédito $i \in \Pi_1$ entonces se observa que la indicadora $\delta_i = 1$, en caso contrario, se observa que la indicadora $\delta_i = 0$.

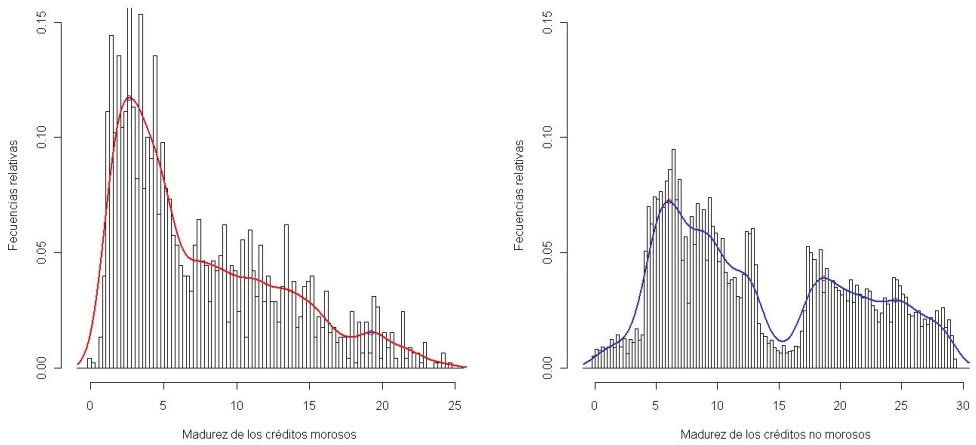

Figura 3.1. Histogramas de la madurez de los créditos morosos (izquierda) y de los créditos no mororos (derecha).

En la Figura 3.1 se representan los histogramas de frecuencias de los tiempos de vida de los créditos morosos y no morosos. En ambos casos, se observa ausencia de simetría lo que da indicios de no normalidad de los datos. En el caso de los créditos morosos (parte izquierdo) se observa un grado de asimetría positiva moderado (0.89) y su coeficiente de curtosis es inferior a 3 (2.89), valores que difieren de los parámetros correspondientes a una distribución normal estándar (0 y 3, respectivamente). En el caso de los créditos no morosos (parte derecha) también se obtienen coeficientes distintos a los de la distribución normal, asimetría de 0.41 y curtosis de 1.89, respectivamente. Para verificar la falta de normalidad de los datos, se aplicaron dos contrastes de bondad del ajuste apropiados para utilizarse con muestras grandes como son, el test de *Lilliefors* (versión del test de

Kolmogorov-Smirnov para el caso particular de la distribución normal) y el test de *Jarque-Bera* (Jarque y Bera (1987)), que se basa precisamente en comparar los coeficientes de asimetría y curtosis muestrales con los de la distribución normal. Como resultado, se obtuvo que los dos contrastes rechazaron la hipótesis de normalidad de ambas muestras con p-valores $<$ 0,001.

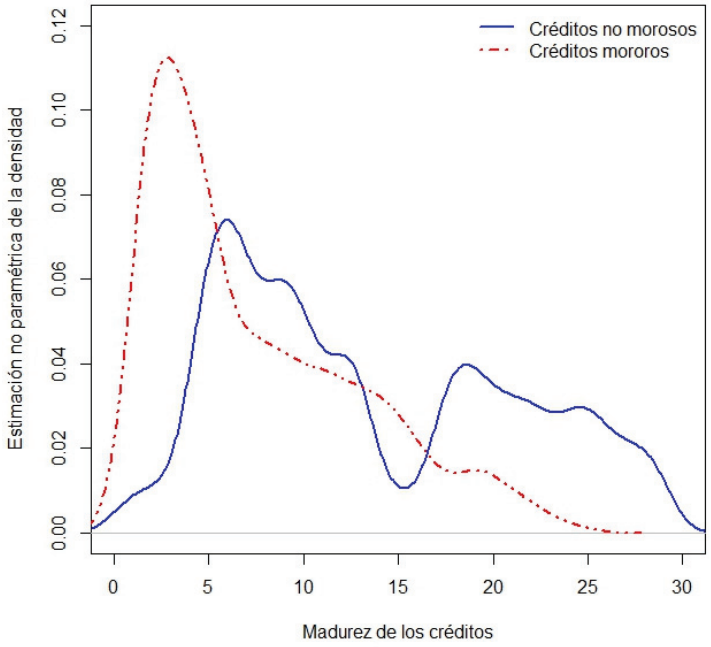

Figura 3.2. Estimación no paramétrica de la función de densidad de la madurez de los créditos morosos y no mororos.

En la Figura 3.2 se muestra la estimación no paramétrica de las funciones de densidad del tiempo de vida observado, o madurez, de los créditos. Para esto se utilizó el estimador núcleo de *Rosenblatt-Parzen* con función núcleo *gaussiana* y parámetro de suavizado, h, obtenido con el método *plug-in* de Sheather y Jones (1991). Se observa, a partir de las curvas de densidad estimadas no paramétricamente en ambas muestras, que existen claras diferencias en el comportamiento de pago de los acreditados, aunque comparten el hecho de mostrar un alto grado de asimetría positiva. Además, la forma

bimodal (con gran peso en la cola derecha) de la curva de densidad estimada de los créditos no morosos (dibujada en color azul) revela la presencia en la muestra de las dos clases de créditos no morosos descritos la población Π_0, es decir, aquellos créditos con fecha de vencimiento anterior a la finalización del estudio que pagaron dentro del plazo establecido, y aquellos créditos con fecha de vencimiento posterior a la finalización del estudio, cuya fecha de mora nunca llegó a ser registrada.

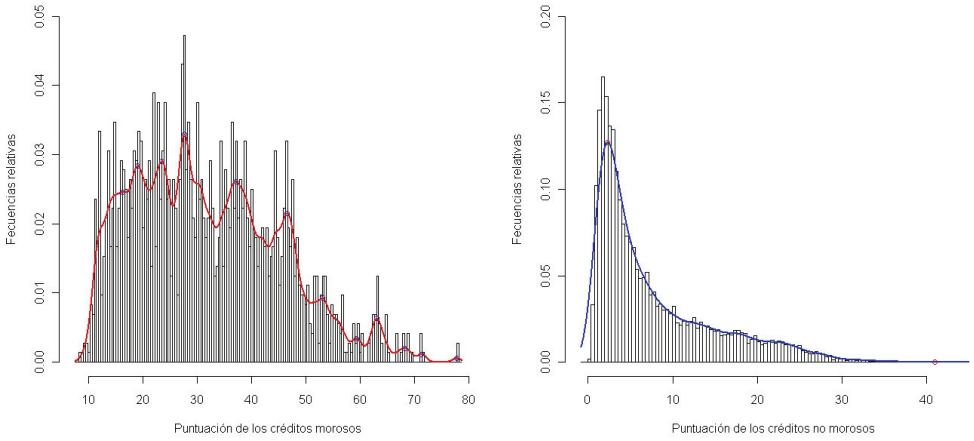

Figura 3.3. Histogramas de la puntuación crediticia de los créditos morosos (izquierda) y de los créditos no mororos (derecha).

En la Figura 3.3 se muestran los histogramas de frecuencia de las puntuaciones de los créditos morosos (parte izquierda) y los no morosos (parte derecha). La forma asimétrica de las distribuciones, medidas a partir de sus coeficientes de curtosis (2.68 y 4.29) y de asimetría (0.51 y 1.37), reflejan la falta de normalidad de las puntuaciones crediticias en ambas submuestras. Esto se verificó utilizando los contrastes para la normalidad mencionados anteriormente (*Lilliefors* y *Jarque-Bera*), obteniéndose un resultado similar al de los tiempos de vida de los créditos, es decir, que se rechazó la hipótesis de normalidad de los datos.

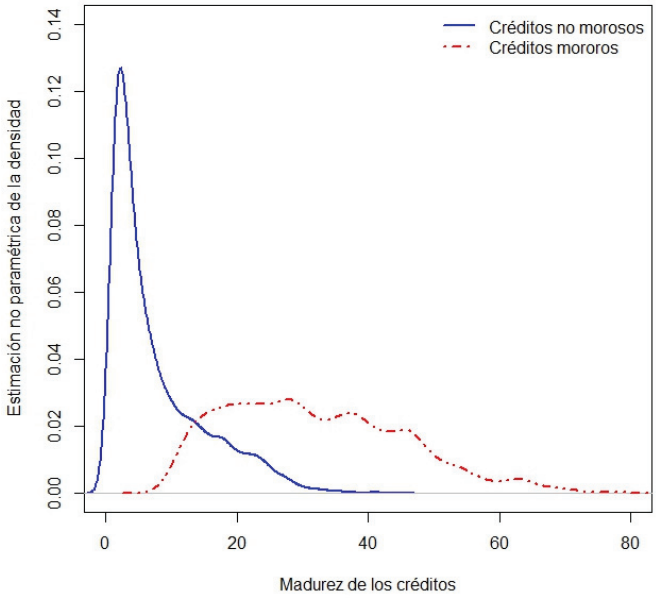

Figura 3.4. Densidad estimada no paramétricamente de la puntuación crediticia para los créditos morosos y los créditos no mororos.

La Figura 3.4 muestra la estimación no paramétrica de la densidad de las puntuaciones de los créditos. Como se puede apreciar en esta figura, las características de la puntuación crediticia observada en ambas muestras de acreditados son claramente diferentes. La curva de densidad estimada de los créditos morosos (trazo de color rojo) muestra una forma platicúrtica con asimetría positiva y gran peso en la cola derecha. En contraste, la curva de densidad estimada en los créditos no morosos (línea de color azul) muestra una distribución con un alto grado de leptocurtosis, también con asimetría positiva pero con menos peso en la cola superior. Además, se observa que el punto en el que se intersectan ambas curvas de densidad estimadas, aproximadamente en $X = 15.33$, permite discriminar, en una gran proporción, entre créditos morosos y no morosos. En efecto, en la muestra existe una alta concentración (aproximadamente el 84.20 %) de clientes no morosos cuya puntuación $X \leq 15.33$. Al mismo tiempo, existe una alta concentración (aproximadamente el 89.56 %) de clientes morosos cuya puntuación $X \geq 15.33$.

Tabla 3.2. Análisis descriptivo de la madurez y la puntuación crediticia con las muestras de créditos morosos (*CM*), créditos no morosos (*CNM*) y la muestra agregada (*MA*).

	Muestra	*min*	Q_1	Q_2	*media*	Q_3	*max*
CM	Madurez ξ	0.033	2.933	5.500	7.458	11.150	24.767
	Puntuación X	8.398	20.295	30.066	31.817	41.167	77.819
CNM	Madurez ξ	0.000	6.767	11.367	13.455	20.033	29.500
	Puntuación X	0.150	2.412	4.857	7.688	11.070	43.920
MA	Madurez ξ	0.000	6.500	10.870	13.020	19.570	29.500
	Puntuación X	0.150	2.540	5.440	9.425	13.405	77.819

En la Tabla 3.2 se ofrece un resumen descriptivo complementario a la información representada en las Figuras 3.1 a la 3.4. Allí se observa que la mayoría de los créditos no morosos (aproximadamente el 75 %) tiene puntuaciones inferiores a la puntuación observada en el 25 % inferior de los créditos morosos, es decir que, $Q_1 = 20.295$ observado en la submuestra *CM* es mayor que $Q_3 = 11.07$ observado en la submuestra *CNM*. También se observa que el percentil 75 % de la puntuación de los créditos morosos prácticamente cuadruplica la puntuación del 75 % de los créditos no morosos ($Q_3 = 41.17$ observado en la submuestra *CM* frente a $Q_3 = 11.07$ observado en la submuestra *CNM*, respectivamente).

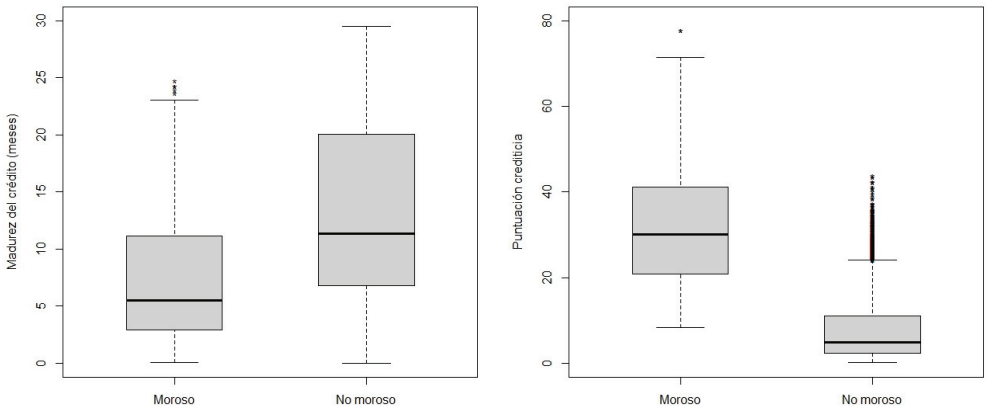

Figura 3.5. Gráficos de cajas para la dispersión de las variables madurez observada (izquierda) y puntuación crediticia (derecha) por situación de morosidad.

En la Figura 3.5 se representan gráficos de cajas para examinar el grado de dispersión de la variable *madurez* o *tiempo de vida observado* del crédito (a la izquierda) y de la variable *puntuación crediticia* (a la derecha). En cuanto a la madurez de los créditos, se observa que, en el caso de los morosos, existe un alto grado de asimetría positiva donde más de la mitad de ellos registraron la mora antes de los 6 meses de vida del crédito (52.83 %), lo que es llamativo. Además, se observa que el peso en la cola superior de la distribución está siendo afectado por la existencia de datos atípicos, que en el caso de los créditos no morosos no existen, y donde la asimetría es menor que en el caso de los créditos morosos.

Con respecto a la puntuación crediticia (figura de la derecha), la situación es distinta que la observada para el tiempo de vida del crédito, es decir, existe mayor concentración de créditos no morosos en torno a valores pequeños de X ($Q_2 < 5$) y la asimetría positiva es mayor que en los créditos no morosos. En esta subuestra de créditos la presencia de valores atípicos es más alta que en el caso de los tiempos de vida (aproximadamente un 3.8 %). Finalmente, en el caso de los créditos morosos, como es lógico, la mayor parte de la distribución se desplaza hacia valores de puntuación más altos (peores) que en el caso de los créditos no morosos donde, por ejemplo, la mediana de X supera en más de seis veces a su percentil equivalente medido en los créditos no morosos (30.066 versus 4.857, respectivamente).

Estimador empírico incondicional de la TM

A continuación, en la Figura 3.6, se ilustra la estimación empírica de la tasa de mora incondicional, denotada por $TM\,(t)$. Las estimaciones de la función $TM\,(t)$ se utilizan para obtener curvas empíricas de la tasa de mora incondicional dado un horizonte fijo de predicción de la entrada en mora, b.

El estimador empírico de la tasa de mora incondicional con datos censurados por la derecha se define por:

$$\widehat{TM}_{n,b}(t) = 1 - \frac{\hat{S}_n^{KM}\,(t+b)}{\hat{S}_n^{KM}\,(t)}, \tag{3.27}$$

donde b es el horizonte de tiempo en el que se desea predecir la entrada en mora del crédito y $\hat{S}_n^{KM}\,(t)$ es el estimador de la función de distribución con datos censurados de Kaplan-Meier (Kaplan y Meier (1958)).

La fórmula utilizada para el cálculo del estimador de Kaplan-Meier viene dada por

$$\hat{S}_n^{KM}(t) = \prod_{\xi_{(i)} \leq t} \left(1 - \frac{1}{n-i+1}\right)^{\delta_{(i)}},$$

donde $\delta_{(i)}$ es el concomitante del i-ésimo estadístico ordenado del tiempo hasta la mora, $\xi_{(i)}$.

El estimador empírico definido en (3.27), se ha utilizado exclusivamente como herramienta de referencia de la morosidad (no censura) presente la muestra utilizada, y por tanto, el estudio de sus propiedades no forma parte de los objetivos de esta memoria dejándose como un problema propuesto para un trabajo futuro.

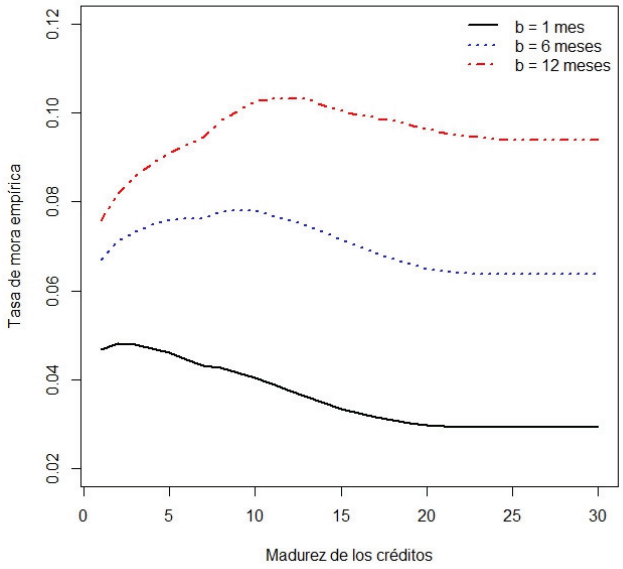

Figura 3.6. Curvas de la tasa de mora empírica incondicional obtenidas con horizontes de predición, b=1, 6 y 12 meses.

En la Figura 3.6 se observa que, ante falta de información adicional, como la aportada por una o más covariables, las curvas empíricas de la tasa de mora (TM) alcanzan proporciones más altas cuando aumenta el horizonte de predicción de la entrada en mora, b, tal y como se puede esperar. Se

observa también que las curvas decrecen hasta volverse constantes cuando el tiempo de vida observado se acerca al máximo muestral (aproximadamente a partir del mes $t = 25$). Esto provoca que las estimaciones en el extremo de la cola derecha resulten deficientes.

Por otro lado, en la Figura 3.6, no es posible observar si los cambios en las curvas obtenidas con el estimador $\widehat{TM}_{n,b}$ se deben sólo al cambio del horizonte de predicción, b, o si además éstas están siendo influenciadas por cambios en otras variables, como puede ser el efecto de la puntuación crediticia de los clientes. Estos resultados permiten pensar que el enfoque incondicional representado por el estimador $\widehat{TM}_{n,b}$ ofrece un mecanismo útil para estudiar la variación en la tasa de mora dado un horizonte de predición fijo, b, en el que no interesa la interacción entre el tiempo de vida del crédito y otras variables. Sin emabrgo, este método no es recomendable para obtener estimaciones de la PD asociadas a un determinado perfil de solvencia de los clientes, como sí ocurre cuando se utiliza el enfoque de supervivencia condicional propuesto en esta memoria. Estos aspectos de estimación de la PD en función de la solvencia crediticia de los clientes resultan muy importantes para las entidades de crédito.

3.3.2. Resultados de la estimación de la PD

En esta sección se analizan los resultados obtenidos con los cuatro estimadores propuestos para el modelo definido en (3.1) bajo el enfoque de análisis de supervivencia. Estos resultados muestran que, en general, cuando se incorpora la información de la covariable puntuación crediticia, X, se obtienen estimaciones más realistas de la PD individual, en particular cuando éstas se comparan con las obtenidas bajo el enfoque incondicional.

Resultados obtenidos con el modelo lineal generalizado MLG

A continuación, en las Figuras 3.7 y 3.8, se muestran los resultados de la estimación de las funciones de distribución condicionada, $F(t|x)$, y de la PD condicionada, $\varphi(t|x)$, obtenidas con el modelo lineal generalizado (MLG). Se tomaron como funciones de enlace dos familias de distribuciones pertenecientes a la familia exponencial, la distribución de Pareto de parámetro α y la distribución F de Snedecor de parámetros ν_1 y ν_2 (los grados

de libertad). Ambas distribuciones fueron elegidas empíricamente después de varias pruebas con otros modelos de distribuciones exponenciales.

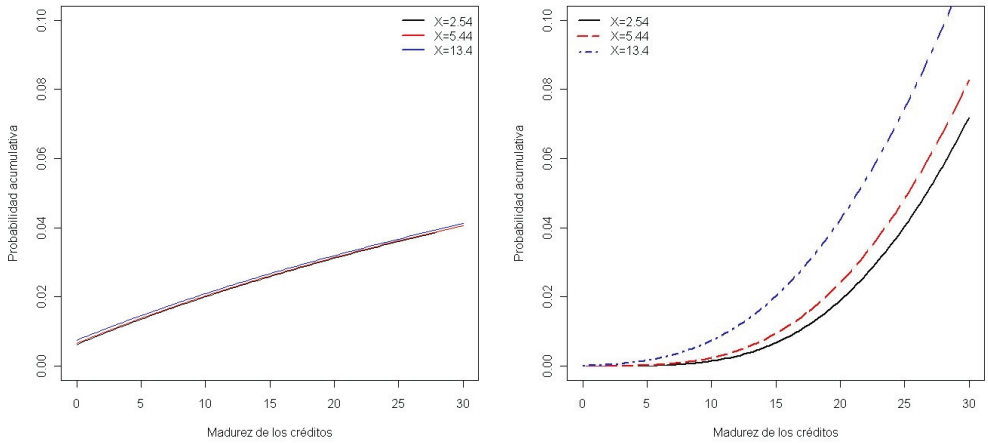

Figura 3.7. Funciones de distribución condicional del tiempo hasta la mora con funciones de enlace de Pareto (izquierda) y F de Snedecor (derecha).

En la Figura 3.7 se muestran las estimaciones de la función de distribución de T condicionada a la puntuación crediticia X, $F(t|x)$, donde X toma los valores 2.54, 5.44 y 13.4, los tres cuartiles de la variable X.

En el recuadro de la izquierda se muestran los resultados obtenidos con la función de enlace de Pareto ($\alpha = 0.6$). En este caso, se observa una aparente falta de influencia de la puntuación crediticia X en las distribuciones condicionadas, siendo difícil distinguir entre cada una de las curvas de distribución. Además, se observa que las tres curvas de distribución de probabilidad condicional estimadas son casi lineales en el rango muestral $t \in [0, 30]$ meses.

En contaste con lo anterior, en el recuadro de la derecha se muestran los resultados obtenidos tomando como función de enlace la distribución $F_{10,50}$. En este caso, se observa claramente la influencia de la variable X en las estimaciones de $F(t|x)$. Así, por ejemplo, se ve que para $t > 5$ meses, la curva $\hat{F}_g(t|13.4)$ (dibujada en azul) crece más rápidamente que las otras dos curvas, $\hat{F}_g(t|5.44)$ (en color rojo), y $\hat{F}_g(t|2.54)$ (en color negro), obteniéndose la relación de orden $\hat{F}_g(t|2.54) < \hat{F}_g(t|5.44) < \hat{F}_g(t|13.4)$, para valores de $t \geq 8$ meses.

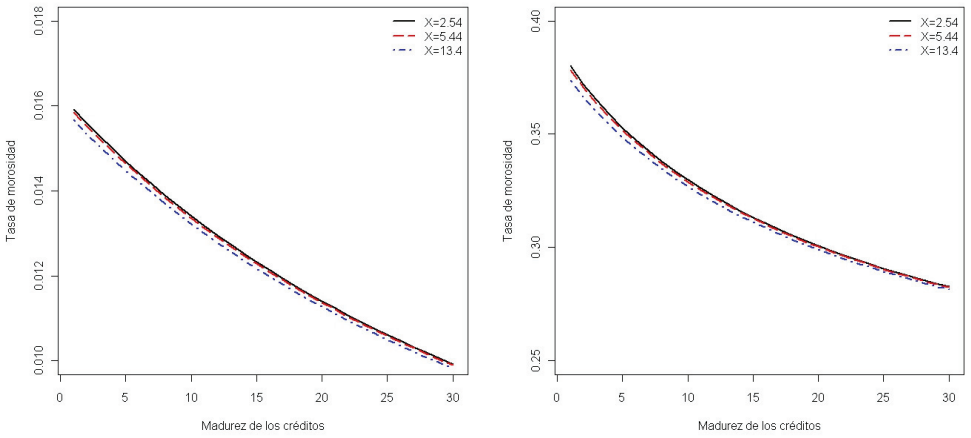

Figura 3.8. Funciones de probabilidad de mora condicional estimada utilizando funciones de enlace de *Pareto* (izquierda) y F de *Snedecor* (derecha).

En la Figura 3.8 se representan las curvas de PD obtenidas con el estimador $\hat{\varphi}^{MLG}$. Teniendo en cuenta que la tasa de mora observada en la muestra de créditos es de un 7.2 %, los resultados obtenidos con ambas funciones enlace parecen no ajustarse lo suficiente a los datos empíricos. En el caso de la distribución de Pareto (recuadro de la izquierda), las curvas de PD resultantes sufren el mismo efecto que el observado en la estimación de la función de distribución condicionada (Figura 3.7, izquierda), es decir, que aparentemente el estimador $\hat{\varphi}^{MLG}$ no tiene en cuenta la contribución de la variable X, provocando que las curvas estén prácticamente superpuestas.

Por otro lado, los valores de la PD estimada con la función de enlace $F_{10,50}$ (recuadro de la derecha), resultan ser demasiado altos en comparación con las curvas empíricas ilustradas en la Figura 3.6. También se observa un efecto de falta de contribución de la variable X, similar al caso de la distribución de Pareto, haciendo difícil distinguir entre la curvas de PD.

Resultados obtenidos con el modelo de regresión de Cox PHM

Estimando la PD con el modelo de riesgos proporcionales de Cox (PHM), se observan claras diferencias con los resultados obtenidos por el método empírico incondicional (ver Figura 3.6).

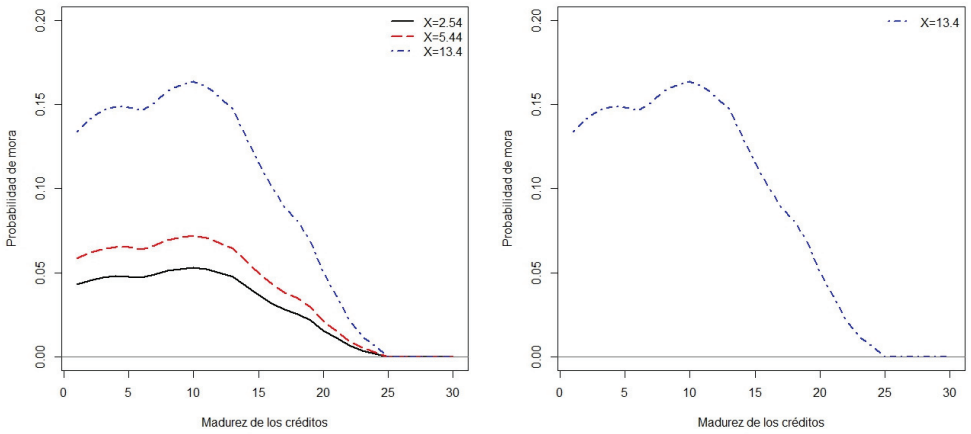

Figura 3.9. Probabilidad de mora condicional estimada con horizonte de predicción b=1 año para X=2.54, 5.44, 13.4 (izquierda) y para la media de X (derecha).

En la Figura 3.9, se observa que las curvas obtenidas con el estimador $\hat{\varphi}^{PHM}(t|x)$ decrecen a cero rápidamente cuando el tiempo hasta la mora alcanza el máximo valor de madurez muestral (aproximadamente en el mes $t = 25$). Este efecto en la curva de la PD estimada se debe fundamentalmente a la alta censura de los datos y la falta de profundidad de la muestra. Como consecuencia de esto, el grado de ajuste de las estimaciones en el extremo de la cola derecha de la curva de PD es deficiente. Recordando que la variable X mide la propensión que tiene el acreditado a convertirse en moroso, se observa además que las estimaciones de la PD crecen a medida que crece el grado de insolvencia de los acreditados.

Resultados obtenidos con el estimador no paramétrico PLG

En esta sección se analizan los resultados de la estimación de la función de PD condicional de los créditos utilizando el estimador no paramétrico, $\hat{\varphi}_n^{PLG}(t|x)$, que por simplicidad de notación se escribirá simplemente como $\hat{\varphi}_n^{PLG}(t|x)$, como se ha explicado anteriormente en la Sección 3.2.3.

Determinación del parámetro de suavizado

Con el objeto de utilizar un selector local, para elegir el parámetro de suavizado, h, se utilizó la técnica de los k vecinos más próximos. Este método consiste en fijar el valor entero positivo, k, de puntos en la vencindad considerada para la estimación local y determinar la ventana h definida por

$$h := h_{[k]}(x),$$

donde $h_{[k]}(x)$ es el valor de la distancia entre x y el k-ésimo valor más próximo a este punto tomado de la muestra ordenada de la covariable X para los tiempos no censurados, es decir, $h_{[k]}(x) = d\left(x, X_{[k]}\right)$ tales que $\delta_{[k]} = 1$, donde $\delta_{[k]}$ es la indicadora de no censura del k-ésimo tiempo hasta la mora.

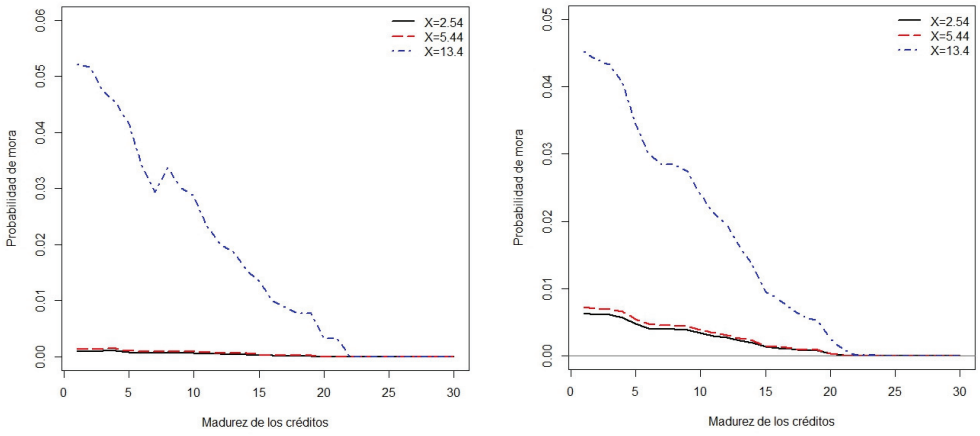

Figura 3.10. Probabilidad de mora condicional estimada con horizonte de predicción b=1 año y parámetro de suavizado fijo k=100 (izquierda) y k=400 (derecha) con distintos valores de puntuación.

En la Figura 3.10, se representan estimaciones no paramétricas de la PD donde el paramétro de vecinos más próximos, k, se ha mantenido fijo en $k = 100$ y el estimador $\hat{\varphi}_n^{PLG}(t|x)$ se ha calculado con tres valores distintos de $X = 2.54, 5.44, 13.4$. Las dos primeras curvas muestran valores pequeños de la PD cuando el valor de X es menor que el primer cuartil de la distribución. Para $k = 100$ (a la izquierda de la figura) hay un aparente efecto de

infrasuavizado de las curvas estimadas. La situación mejora en las curvas que se ilustran en el lado derecho de la figura. Allí, tomando $k = 400$, la curva estimada muestra mayor grado de suavizamiento. También se observa que las estimaciones parecen ser sensibles ante pequeños cambios en la puntuación crediticia de los clientes. Como resultado, la PD puede ser sobreestimada al comienzo de la vida del préstamo para algunos valores de X.

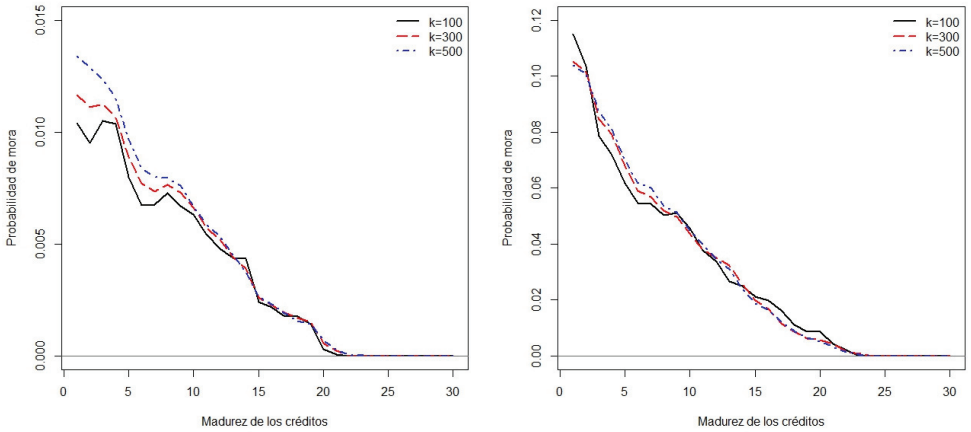

Figura 3.11. Probabilidad de mora condicional estimada con horizonte de predicción b=1 año y con tres diferentes valores del parámetro de suavizado dado X=9.43 (izquierda) y dado X=20 (derecha).

En la Figura 3.11 se ilustran las curvas de PD obtenidas con el estimador PLG, $\hat{\varphi}_n^{PLG}(t|x)$, para dos valores fijos de X (9.43, 20) y variando el número de vecinos más próximos, $k = 100, 300, 500$. Observando ambas imágenes (izquierda y derecha), se aprecia que para un valor fijo del parámetro de suavizado, k, el nivel de probabilidad de mora obtenido con la función curva suavizada aumenta a si la puntación crediticia crece.

3.3.3. Análisis de los resultados obtenidos

A continuación, se ilustran los resultados obtenidos con el estimador de la función de PD condicional con censura extrema, $\hat{\varphi}^{MLGM}$, cuyas curvas se comparan con las obtenidas con los estimadores $\hat{\varphi}^{MLG}$, $\hat{\varphi}^{PHM}$ y $\hat{\varphi}_n^{PLG}$.

Todas las curvas estimadas fueron obtenidas condicionando sobre $X = 5.44$ (el valor mediano de la covariable X).

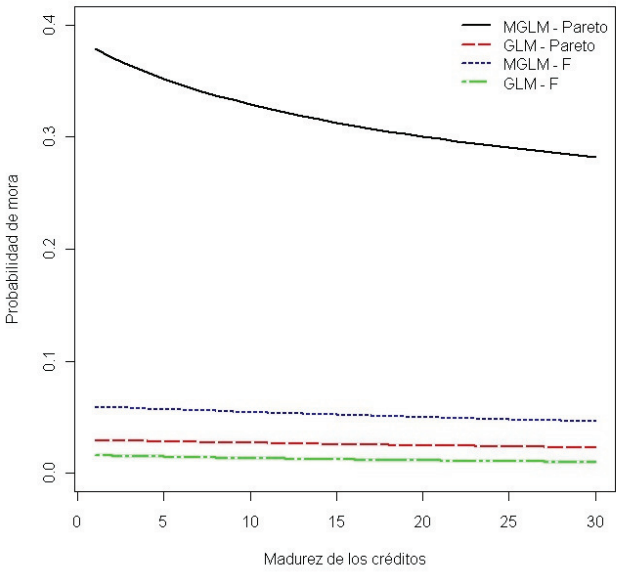

Figura 3.12. Curvas de probabilidad de mora condicional estimadas con los modelos *MLG* y *MLGM*, ambos con funciones de enlace de Pareto y $F_{10,50}$.

En la Figura 3.12 se comparan las curvas estimadas de la PD condicional obtenidas con el modelo MLG y con la variante para el caso de censura extrema, $MLGM$. Las funciones de enlace utilizadas fueron las mismas en ambos casos, la distribución de Pareto de parámetro $\alpha = 0.6$ y la distribución F de Snedecor con $v_1 = 10$ y $v_2 = 50$ grados de libertad. En todas las curvas de PD estimadas vía $MLGM$ se utilizó la proporción estimada de créditos inmunes al riesgo de morosidad (censura) $q = 0.99$. Se observa que la curva $MLGM$ con distribución de Pareto (dibujada en color negro) ofrece estimaciones de la PD muy por encima de las otras tres curvas, llegando a una escala de probabilidad en la que es difícil hacer la compración. En contraste, teniendo en cuenta que la tasa de mora observada en la muestra es de un 7.2 %, se observa que al condicionar sobre el valor mediano de la variable X, las curvas de PD obtenidas vía MLG y $MLGM$ con distribución $F_{10,50}$, ofrecen estimaciones por debajo del 10 %, ajustándose más a la tasa

de mora muestral.

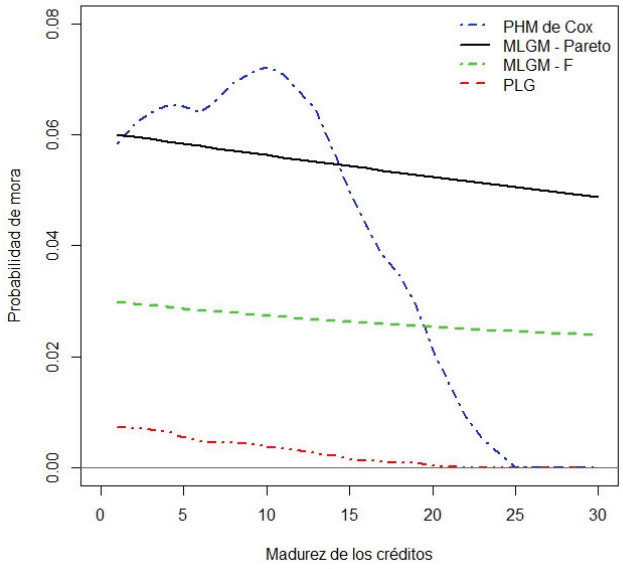

Figura 3.13. Curvas de probabilidad de mora condicional estimadas con los modelos *MLGM (Pareto y* $F_{10,50}$*), PHM* y *PLG*.

En la Figura 3.13 se comparan las curvas de la PD condicional obtenidas con los modelos $MLGM$, PHM de Cox y no paramétrico, PLG. Se observa que, dada $X = 5.44$ (valor mediano de X), la escala de las PD obtenidas con los tres modelo se ajusta bien a la tasa de mora muestral, fluctuando todas ellas por debajo del 7 %. Sin embargo, se observa que los resultados obtenidos con el enfoque basado en modelos lineales generalizados, en general, no son buenos, tanto con la versión MLG como con la versión modificada, $MLGM$.

Con relación a las funciones de enlace consideradas en este estudio, llama la atención que, en general, en las distintas pruebas realizadas con éstas y otras distribuciones de la familia exponencial (como la distribución normal), no se observó una mejora de las estimaciones obtenidas con el modelo $MLGM$ respecto de lo obtenido con el modelo MLG.

A continuación, en la Tabla 3.3, se ofrece un análisis descriptivo de las PD estimadas con los modelos estudiados en este capítulo.

Tabla 3.3. Estadisticos descriptivos de la tasa de mora empírica ($\widehat{TM}_{n,b}$) y de la PD obtenida con los modelos deCox (PHM), lineal generalizado (MLG) y no paramétrico (PLG) tomando un horizonte de predicción de b=12 meses

Modelo de PD			min	Q_1	Q_2	$media$	Q_3	max
		b						
$\widehat{TM}_{n,b}$		1	0.076	0.094	0.095	0.095	0.099	0.103
		6	0.064	0.064	0.069	0.070	0.076	0.078
		12	0.029	0.029	0.033	0.036	0.042	0.048
		X						
PHM (b=12)		2.54	0.000	0.005	0.034	0.029	0.014	0.016
		5.44	0.000	0.006	0.047	0.039	0.014	0.016
		13.4	0.000	0.015	0.108	0.089	0.014	0.016
	Función de enlace	X						
MLG (b=12)	$Pareto_{\alpha=0.6}$	2.54	0.010	0.011	0.012	0.013	0.014	0.016
		5.44	0.001	0.011	0.012	0.012	0.014	0.016
		13.4	0.010	0.011	0.012	0.012	0.014	0.016
	$F_{10,50}$	2.54	0.283	0.295	0.312	0.318	0.337	0.447
		5.44	0.282	0.295	0.311	0.318	0.336	0.378
		13.4	0.282	0.294	0.310	0.316	0.334	0.374
	k	X						
PLG (b=12)	100	2.54	0.0000	0.0000	0.0002	0.0004	0.0007	0.0012
	100	5.44	0.0000	0.0000	0.0003	0.0005	0.0010	0.0015
	100	13.4	0.0000	0.0000	0.0118	0.0175	0.0300	0.0520
	400	2.54	0.0000	0.0000	0.0012	0.0021	0.0039	0.0064
	400	5.44	0.0000	0.0001	0.0014	0.0024	0.0045	0.0073
	400	13.4	0.0000	0.0001	0.0089	0.0152	0.0282	0.0452
	100	9.43	0.0000	0.0000	0.0023	0.0037	0.0067	0.0105
	300	9.43	0.0000	0.0000	0.0024	0.0040	0.0073	0.0117
	500	9.43	0.0000	0.0001	0.0025	0.0042	0.0079	0.0134
	100	20	0.0000	0.0005	0.0205	0.0301	0.0509	0.1149
	300	20	0.0000	0.0009	0.0183	0.0302	0.0514	0.1054
	500	20	0.0000	0.0006	0.0177	0.0306	0.0531	0.1040

Los resultados contenidos en la Tabla 3.3 se obtuvieron utilizando como funciones de enlace del modelo MLG la distribución de Pareto (0.6) y la distribución $F_{10,50}$. Para obtener los valores del estimador no paramétrico, PLG, se utilizaron distintos valores del parámetro k (vecinos más próximos) para un valor fijo de X y después se fijó el valor del parámetro k para distintos valores de X. Para las distribuciones condicionadas se utilizaron valores representativos de la distribución de la variable X como son los cuartiles Q_1, Q_2 (mediana) y Q_3.

Se observa que la tasa de mora empírica, $\widehat{TM}_n$, disminuye con el tiempo de vida del crédito a medida que aumenta el horizonte de predicción de entrada en mora, b. Los valores máximo y mínimo de $\widehat{TM}_n$, 0.103 y 0.029, se obtienen para $b = 1$ mes y $b = 12$ meses, respectivamente. En contraste con lo anterior, los tres modelos de PD condicional, PHM de Cox, MLG y PLG, tienen en cuenta la influencia de la covariable X, por lo que se han obtenido las curvas de PD condicionadas a X, considerando para esto tres valores representativos de la distribución de X, como son los tres cuartiles muestrales, Q_1, Q_2 y Q_3. Las curvas estimadas con el modelo PHM ofrece estimaciones bien acotadas, ya que no se escapan (significativamente) del rango de PD entre 0.0 y 0.10 que puede considerarase como un rango de valores realista, o bien, ajustado a la experiencia empírica, teniendo en cuenta que la TM muestral es de un 7.2 %. Además, las PD crecen cuando X aumenta de valor hasta un valor máximo y después vuelven a decrecer a medida que el horizonte de predicción aumenta.

Por otra parte, el modelo MLG ofrece estimaciones poco realistas, muy por debajo del 7.2 % y prácticamente invariantes con a medida que cambian los valores de X. Utilizando la función de enlace de Pareto, las curvas de probabilidad de mora estimadas varían levemente entre 0.01 y 0.016, resultando en curvas prácticamente superpuestas para los tres valores de X. En cambio, cuando se utiliza la función de enlace F, se las curvas de PD obtenidas arrojan probabilidades de mora entre 0.282 y 0.447, valores de PD muy por encima del rango considerado realista, $0.01 < PD \leq 0.1$. Similarmente al caso de la distribución de Pareto, la curvas obtenidas con el modelo MLG con función de enlace $F_{10,50}$ resultaron prácticamente invariantes ante cambios en la variable X.

Por último, con respecto a las PD obtenidas con el estimador no paramétrico, PLG, se observa que fijando el parámetro de vecinos más cercanos en

$k = 100$ y $k = 400$, al ir cambiando los valores de X, las curvas estimadas arrojan probabilidades de mora mejor justadas que el modelo MLG. Se observa también, que las PD estimadas no paramétricamente varían cuando aumenta el valor de la variable X. Además, cuando se fija el valor de la covariable en $X = 9.43$ y $X = 20$, se obtiene que las curvas de PD aumentan tanto con el valor de la variable X como con el valor del parámetro $h = h_{[k]}$.

3.3.4. Estudio de validación de los modelos

En la literatura existen interesantes contrastes para evaluar la idoneidad de los modelos que se utilizan para evaluar la solvencia de las entidades financieras basados en criterios de predicción (Beran (2009)). La probabilidad de mora estimada es uno de ellos y se puede utilizar para discriminar entre créditos morosos y no morosos. Usando los métodos propuestos en Cao et al. (2009) y fijando un valor de madurez de $t = 5$ meses y un horizonte de predicción de $b = 12$ meses, se ha estimado la PD para cada uno de los créditos de una cartera de préstamos personales.

Partiendo de una muestra de créditos vivos en el instante t, se han tomado dos submuestras, una de créditos morosos y otra de créditos no morosos, en el tiempo $t + b$. Para evaluar el poder de discriminación de los modelos considerados se utilizaron curvas ROC. También se calcularon los coeficientes del área bajo la curva ROC (AUC) que, como se ha visto en Capítulo 2, es una medida de validación global del poder de discriminación de los modelos de riesgo de crédito.

El estudio se realizó dividiendo la muestra original, de $N = 25\,000$ créditos, en dos submuestras, una de entrenamiento de tamaño $n = 20\,000$ y otra más pequeña, llamada de validación, de tamaño $m = 5\,000$.

Los parámetros de los modelos de PD condicional se obtuvieron con los datos de la muestra de entrenamiento mientras que las PD fueron estimadas para ada uno de los créditos de la muestra de validación.

El resultado del análisis de curvas ROC obtenidas con la muestra de validación se ilustra a continuación, en la Figura 3.14.

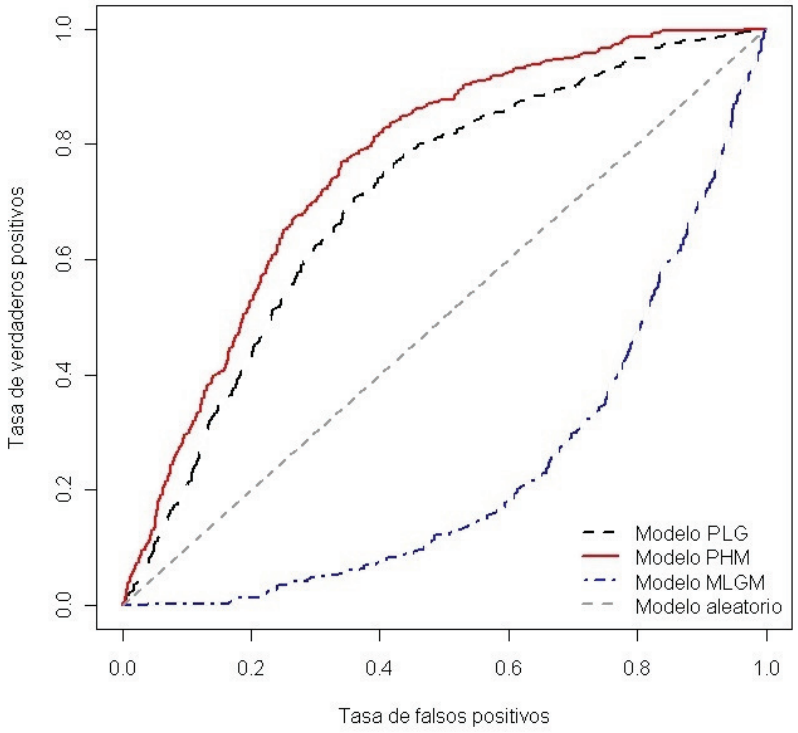

Figura 3.14. Curvas *ROC* para los modelos *PLG*, *PHM* y *MLGM.*

En la Figura 3.14 se muestra el resultado del mecanismo de validación del poder predictivo de los modelos considerados. Allí se ovserva una pobre capacidad de discriminación del modelo $MLGM$ en comparación con los modelos de regresión de Cox, PHM, y no paramétrico, PLG. Esto se verifica a partir de los valores del coeficiente AUC que se muestran en la Tabla 3.4. En contraste con este resultado, el desempeño de los modelos de regresión de Cox, PHM, y no paramétrico basado en el estimador de Beran (1981), PLG, resultaron más satisfactorios y acorde con lo esperado. Ambos modelos son competitivos y ofrecen buena capacidad predictiva, observándose un ligera ventaja del modelo PHM, cuyo índice $AUC = 76{,}2\,\%$, con respecto al modelo PLG, cuyo índice $AUC = 71{,}5\,\%$.

Tabla 3.4. Comparación del poder predictivo de los modelos *MGLM*, *PHM* y *PLG* vía el cálculo del Área bajo la curva *ROC*.

Modelo	*AUC*	Intervalo de confianza asintótico 95 %	
$MLGM$	0.238	0.217	0.260
PHM	0.762	0.740	0.783
PLG	0.715	0.691	0.739

De acuerdo con los resultados de la Tabla 3.4 junto con las curvas ilustradas en las Figuras 3.12 a la 3.14, se puede concluir que la variante del modelo MLG para el caso de censura extrema, $MLGM$, no arrojó los resultados esperados en el sentido de mejorar la capacidad predictiva del modelo original, MLG. En efecto, un valor del coeficiente $AUC < 0.5$ es equivalente a decir que el modelo utilizado carece de capacidad predictiva, e incluso, en el caso de esta muestra, que tiene tendencia a confundir los clientes no morosos con los morosos. Estos resultados conducen a proponer un estudio más profundo sobre las hipótesis utilizadas, especialmente en lo referente al mecanismo de selección de la función de enlace para lo cual se pueden implementar contrastes de bondad del ajuste sobre la elección de la función de enlace, como los estudiados por Pregibon (1980) y Czado (1992, 1997), por citar sólo algunos autores.

Por otro lado, las estimaciones de la PD obtenidas con el modelo de regresión de Cox, PHM, y con el modelo no paramétrico basado en el estimado de Beran, PLG, proporcionaron los mecanismos más potentes para discriminar entre créditos morosos y no morosos. Además, si se observa con detenimiento la Tabla 3.4, debido a que sus respectivos intevalos de confianza para el índice AUC no se intersectan, es posible afirmar que el estimador de la PD condicional vía el modelo de regresión de Cox, $\hat{\varphi}^{PHM}$, mostró mejor capacidad predictiva de la morosidad de los créditos que componen la muestra de validación.

Capítulo 4

Estudio asintótico del estimador de la probabilidad de mora condicional basado en el estimador de Beran

En este capítulo se estudian las propiedades asintóticas del estimador no paramétrico de la función de probabilidad de mora condicional propuesto por Cao et al. (2009), $\hat{\varphi}_n^{PLG}$, cuya fórmula viene dada por

$$\hat{\varphi}_n^{PLG}(t|x) = 1 - \frac{\hat{S}_h^{PLG}(t+b|x)}{\hat{S}_h^{PLG}(t|x)},$$

donde $\hat{S}_h^{PLG}(\cdot|x)$ es el estimador producto límite generalizado (PLG) de Beran (1981) para la función de supervivencia condicional definido en (3.21) para todo $t > 0$ y $x \in \mathbb{R}$, y donde $b > 0$ es el horizonte de tiempo en el que se quiere predecir la futura mora del crédito.

Las propiedades asintóticas del estimador de la PD, $\hat{\varphi}_n^{PLG}$, se obtienen a partir de resultados análogos obtenidos previamente por Dabrowska (1989), Van Keilegom y Veraverbeke (1996) e Iglesias Pérez y González Manteiga (1999), entre otros, para el estimador PLG de Beran (1981) de la función de distribución condicional con datos censurados. Dependiendo de si se trata de un modelo de regresión en diseño fijo o aleatorio, y del tipo de pesos empleados en la estimación no paramétrica, se han estudiado distintas variantes y

propiedades asintóticas del estimador de Beran, entre las que cabe señalar la consistencia uniforme débil y fuerte probada por Dabrowska (1989) y la representación casi segura y normalidad asintótica obtenidas por Iglesias Pérez y González Manteiga (1999), todas ellas en un contexto de diseño aleatorio con pesos de tipo Nadaraya-Watson.

4.1. Definiciones e hipótesis

En esta sección se establecen las hipótesis necesarias para la obtención de los resultados asintóticos del estimador no paramétrico de la PD condicionada, $\hat{\varphi}_n^{PLG}$, que en este capítulo también se denotará simplemente por $\hat{\varphi}_n$.

En adelante se supondrá un contexto de diseño aleatorio en el que la covariable, X, es unidimensional con función de distribución absolutamente continua, $M(x)$, y función de densidad $m(x) = M'(x)$, cuyo soporte se denota por $\Omega_X = \{x \in \mathbb{R}^+ : m(x) > 0\}$. Además, es conveniente recordar algunas de las notaciones y definiciones dadas en el Capítulo 3:

(i) $H_1(t|x) = P(\xi_i \leq t, \delta_i = 1|X_i = x) = \int_0^t (1 - G(u|x))\, dF(u|x)$ es la subdistribución condicional del tiempo de vida de un crédito moroso.

(ii) $H_0(t|x) = P(\xi_i \leq t, \delta_i = 0|X_i = x) = \int_0^t (1 - F(u|x))\, dG(u|x)$ es el análogo de $H_1(t|x)$ para un crédito no moroso.

(iii) $H(t|x) = P(\xi_i \leq t|X_i = x) = 1 - (1 - F(t|x))(1 - G(t|x))$ es la función de distribución del tiempo de vida observado del crédito.

(iv) La función núcleo, K, verifica las hipótesis H3.11 y H3.12 donde

$$d_K = \int u^2 K(u)\, du,\ c_K = \int K^2(u)\, du.$$

(v) Denotando por L_x la función de distribución condicionada, $L(\cdot|x)$, con $x \in \Omega_X$, la función complementaria de L_x se denota por $\overline{L}_x = 1 - L_x$ y los extremos superior e inferior del soporte de L_x se denotan, respectivamente, por $\underline{\tau}_{L_x} = \inf\{t : L(t|x) > 0\}$ y $\bar{\tau}_{L_x} = \sup\{t : L(t|x) < 1\}$.

Como consecuencia de H3.5, a partir de (iii) se verifica que:

$$\begin{aligned} \underline{\tau}_{H_x} &= \min\left\{\underline{\tau}_{F_x}, \underline{\tau}_{G_x}\right\}, \\ \overline{\tau}_{H_x} &= \min\left\{\overline{\tau}_{F_x}, \overline{\tau}_{G_x}\right\}. \end{aligned}$$

(vi) Existen las derivadas parciales primera y segunda de L_x respecto de x y se denotan por $\dot{L}_x = \frac{\partial}{\partial x} L_x$ y $\ddot{L}_x = \frac{\partial^2}{\partial x^2} L_x$, respectivamente.

(vii) Existen las derivadas parciales primera y segunda de L_x respecto de t y se denotan por $L'_x = \frac{\partial}{\partial t} L_x$ y $L''_x = \frac{\partial^2}{\partial t^2} L_x$, respectivamente.

$(viii)$ Existe la derivada parcial segunda de L_x respecto de x y de t, y se denota por $\dot{L}'_x = \frac{\partial^2}{\partial x \partial t} L_x = \frac{\partial^2}{\partial t \partial x} L_x$.

4.1.1. Hipótesis para el estimador $\hat{\varphi}_n^{PLG}$

H 4.1 *Sea $I = [x_1, x_2]$ un intervalo contenido en Ω_X, entonces se verifica que:*

$$0 < \gamma = inf\ \{m(x) : x \in I_\delta\} < sup\ \{m(x) : x \in I_\delta\} = \Gamma < \infty,$$

para alguna vecindad $I_\delta = [x_1 - \delta, x_2 + \delta]$ de Ω_X con $\delta > 0$ y $0 < \delta\Gamma < 1$.

H 4.2 *Existe $\theta > 0$ tal que:*

$$\inf_{t \in [0, \overline{\tau}_{H_x}]} \{1 - H(t|x) : x \in I_\delta\} \geq \theta.$$

H 4.3 *La función $m(x)$ tiene derivadas de primer y segundo orden continuas y acotadas en el intervalo I_δ denotadas por $m'(x)$ y $m''(x)$, respectivamente.*

H 4.4 *Las funciones $\dot{H}(t|x)$, $\ddot{H}(t|x)$, $\dot{H}_1(t|x)$ y $\ddot{H}_1(t|x)$ existen y son absolutamente continuas y acotadas para todo $(t, x) \in [0, \overline{\tau}_{H_x}] \times I_\delta$.*

H 4.5 *Las funciones $H'(t|x)$, $H''(t|x)$, $H'_1(t|x)$ y $H''_1(t|x)$ existen y son absolutamente continuas y acotadas para todo $(t, x) \in [0, \overline{\tau}_{H_x}] \times I_\delta$.*

H 4.6 *Las funciones $\dot{H}'(t|x)$ y $\dot{H}'_1(t|x)$ existen y son absolutamente continuas y acotadas en $(t, x) \in [0, \overline{\tau}_{H_x}] \times I_\delta$.*

H 4.7 *El parámetro de suavizado, $h \equiv h_n$, satisface las condiciones:*

$$(\ln n)^3/nh \longrightarrow 0 \text{ y } nh^5/\ln n = O(1) \text{ cuando } n \longrightarrow \infty.$$

Observación 4.1 (i) *La hipótesis H3.12 permite asegurar que los estimadores tipo núcleo de las funciones $F(t|x)$, $G(t|x)$, $H(t|x)$ y $H_1(t|x)$, son, en efecto, funciones de distribución de probabilidad bien definidas, es decir, funciones no negativas y crecientes en t con recorrido en $[0,1]$ para cualquier $x \in \Omega_X$.* (ii) *La hipótesis H4.1 es utilizada por Dabrowska (1989) para obtener cotas exponenciales para las colas de la distribución del estimador $\hat{S}_h(t|x)$. Este resultado es utilizado después para obtener la convergencia débil y fuerte del mismo.* (iii) *La hipótesis H4.2 es necesaria para evitar problemas de estimación en las colas de las distribuciones mencionadas en* (i). (iv) *Las condiciones sobre la función núcleo, K, explicadas en la observación 3.2 junto con las hipótesis H4.3-H4.6 permiten establecer condiciones necesarias para asegurar la insesgadez asintótica del estimador de Beran, $\hat{S}_h(t|x)$.* (v) *Por último, las condiciones impuestas sobre el parámetro de ventana, h, en H4.7 son necesarias para determinar la velocidad de convergencia del estimador, $\hat{S}_h(t|x)$, a la función de supervivencia condicional, $S(t|x)$. Una discusión más profunda acerca de estas hipótesis puede verse en Dabrowska (1989).*

4.2. Propiedades asintóticas del estimador PLG de la PD

Cao et. (2009) demostraron que el estimador no paramétrico de la PD basado en el estimador Beran, $\hat{\varphi}_n^{PLG}$, verifica las propiedades de consistencia fuerte uniforme y normalidad asintótica. Ambas propiedades se formalizan a continuación en dos teoremas y un corolario. Las demostraciones de estos resultados asintóticos se desarrollan en la Sección 4.3 mientras que en la Sección 4.4 se exponen algunos resultados auxiliares relativos al estimador de Beran imprescindibles para la obtención de los resultados de la Sección 4.3.

Teorema 4.1 (*Consistencia fuerte uniforme*) *Fijados t y x tales que $0 < \varphi(t|x) < 1$, bajo las hipótesis H4.1-H4.7, $\hat{\varphi}_n(t|x)$ es un estimador fuertemente consistente de la probabilidad de mora condicionada, $\varphi(t|x)$. Además, si $b < \overline{\tau}_{H_x}$ e $\inf_{x\in I} \{S(\overline{\tau}_{H_x}|x)\} > 0$, entonces, definiendo $\tau^*_H = \overline{\tau}_{H_x} - b$, la consistencia es uniforme en $(t, x) \in [0, \tau^*_H] \times I$, es decir:*

$$\sup_{t\in\left[0,\tau^*_H\right]} \sup_{x\in I} |\hat{\varphi}_n(t|x) - \varphi(t|x)| \longrightarrow 0 \ \ c.s.$$

Teorema 4.2 (*Sesgo y varianza asintóticos*) *Si se verifican las hipótesis H4.1-H4.7, se obtiene que el error cuadrático medio del estimador PLG de la probabilidad de mora condicional, $\hat{\varphi}_h(t|x)$, viene dado por:*

$$ECM\left(\hat{\varphi}_h(t|x)\right) = h^4 \left(b(t|x)\right)^2 + \frac{1}{nh} v(t|x) + o\left(h^4 + \frac{1}{nh}\right), \tag{4.1}$$

donde

$$b(t|x) = -\frac{1}{2} d_K \left(1 - \varphi(t|x)\right) B_H\left(t, t+b|x\right), \tag{4.2}$$

y

$$v(t|x) = \frac{1}{m\left(x\right)} c_K \left(1 - \varphi(t|x)\right)^2 D_H\left(t, t+b|x\right), \tag{4.3}$$

y donde se definen las expresiones

$$\begin{aligned} B_H\left(t, t+b|x\right) &= \int_t^{t+b} \left[\ddot{H}(s|x) + 2\frac{m'\left(x\right)}{m\left(x\right)} \dot{H}(s|x)\right] \frac{dH_1(s|x)}{\left(1 - H(t|x)\right)^2} \\ &\quad + \int_t^{t+b} \frac{d}{1 - H(t|x)} \left[\dot{H}_1(s|x) + 2\frac{m'\left(x\right)}{m\left(x\right)} H(s|x)\right] \end{aligned} \tag{4.4}$$

y

$$D_H\left(t, t+b|x\right) = \int_t^{t+b} \frac{dH_1\left(s|x\right)}{\left(1 - H\left(s|x\right)\right)^2}. \tag{4.5}$$

Observación 4.2 *Como consecuencia del Teorema 4.2, se obtiene que el valor del parámetro de suavizado que minimiza los términos dominantes del ECM en (4.1) viene dado por la fórmula asintótica:*

$$h_0 = \left(\frac{v(t|x)}{4\left(b(t|x)\right)^2} \right)^{1/5} n^{-1/5}. \tag{4.6}$$

Corolario 4.1 (*Normalidad asintótica*) *Bajo las mismas hipótesis del Teorema 4.1, si además* $nh^5 \longrightarrow C \in (0, +\infty)$*, entonces la distribución límite del estadístico* $\sqrt{nh}\left(\hat{\varphi}_n(t|x) - \varphi(t|x)\right)$ *viene dada por*

$$\sqrt{nh}\left(\hat{\varphi}_n(t|x) - \varphi(t|x)\right) \xrightarrow{d} N\left(C^{1/2} b(t|x), v(t|x)\right). \tag{4.7}$$

4.3. Demostraciones

Demostración del Teorema 4.1 (a) Recordando las ecuaciones (3.1) y (3.23), se definen las igualdades

$$\begin{aligned} \varphi(t|x) &= 1 - \frac{P}{Q}, \\ \hat{\varphi}_n(t|x) &= 1 - \frac{\hat{P}}{\hat{Q}}, \end{aligned}$$

con $P = S\left(t + b|x\right)$, $Q = S\left(t|x\right)$, $\hat{P} = \hat{S}_h\left(t + b|x\right)$ y $\hat{Q} = \hat{S}_h\left(t|x\right)$. Utilizando la Propiedad 4.1 de la Sección 4.4 se obtiene la convergencia fuerte del vector $\left(\hat{P}, \hat{Q}\right)$, esto es

$$\left(\hat{P}, \hat{Q}\right) \longrightarrow (P, Q) \text{ c. s.} \tag{4.8}$$

Aplicando el teorema de la función continua al resultado obtenido en (4.8), como $g\left(x, y\right) = \dfrac{x}{y}$ es continua en (P, Q), ya que $Q > 0$, se obtiene que

$$\hat{\varphi}_n(t|x) \longrightarrow \varphi(t|x) \text{ c.s.},$$

lo que concluye la primera parte de la demostración.

(b) Para la segunda parte de la demostración se utiliza la Propiedad 4.2 de la Sección 4.4 que permite obtener de manera inmediata los siguientes resultados de consistencia fuerte uniforme del estimador de Beran:

$$\sup_{t\in\left[0,\tau_H^*\right]}\sup_{x\in I}\left|\hat{S}_h\left(t+b|x\right)-S\left(t+b|x\right)\right| \longrightarrow 0 \text{ c.s.} \tag{4.9}$$

$$\sup_{t\in\left[0,\tau_H^*\right]}\sup_{x\in I}\left|\hat{S}_h\left(t|x\right)-S\left(t|x\right)\right| \longrightarrow 0 \text{ c.s.} \tag{4.10}$$

con $\tau_H^* = \overline{\tau}_{H_x} - b$.

Considérese ahora la identidad:

$$\frac{1}{z} = 1-(z-1)+\cdots+(-1)^p(z-1)^p+(-1)^{p+1}\frac{(z-1)^{p+1}}{z}, \tag{4.11}$$

válida para $z \neq 0$ y para todo $p \in \mathbb{N}$. Utilizando (4.11) con $p = 1$ y tomando $\frac{1}{z} = \frac{Q}{\hat{Q}}$ se obtiene que:

$$\begin{aligned} 1-\hat{\varphi}_n(t|x) &= \frac{\hat{P}}{\hat{Q}} = \frac{\hat{P}}{Q}\frac{Q}{\hat{Q}} \\ &= \frac{\hat{P}}{Q}\left[1-\left(\frac{\hat{Q}}{Q}-1\right)+\frac{Q}{\hat{Q}}\left(\frac{\hat{Q}}{Q}-1\right)^2\right] \\ &= \frac{\hat{P}}{Q}-\frac{\hat{P}\left(\hat{Q}-Q\right)}{Q^2}+\frac{\hat{P}}{\hat{Q}}\frac{\left(\hat{Q}-Q\right)^2}{Q^2}, \end{aligned}$$

así

$$|(1-\hat{\varphi}_n(t|x))-(1-\varphi(t|x))| \leq A_1+A_2+A_3 \tag{4.12}$$

donde

$$\begin{aligned} A_1 &= \frac{\left|\hat{P}-P\right|}{Q}, \\ A_2 &= \frac{\hat{P}\left|\hat{Q}-Q\right|}{Q^2}, \\ A_3 &= \frac{\hat{P}}{\hat{Q}}\frac{\left(\hat{Q}-Q\right)^2}{Q^2}. \end{aligned}$$

Por otra parte, si $x \in I$ y $t \leq \tau_H^*$, se tiene que

$$A_1 \leq \frac{\sup\limits_{t\in\left[0,\tau_H^*\right]}\sup\limits_{x\in I}\left|\hat{S}_h\left(t+b|x\right)-S\left(t+b|x\right)\right|}{\inf\limits_{x\in I}S(\tau_H^*|x)}, \tag{4.13}$$

$$A_2 \leq \frac{\sup\limits_{t\in\left[0,\tau_H^*\right]}\sup\limits_{x\in I}\left|\hat{S}_h\left(t|x\right)-S\left(t|x\right)\right|}{\inf\limits_{x\in I}\left(S(\tau_H^*|x)\right)^2}, \tag{4.14}$$

$$A_3 \leq \frac{\sup\limits_{t\in\left[0,\tau_H^*\right]}\sup\limits_{x\in I}\left|\hat{S}_h\left(t|x\right)-S\left(t|x\right)\right|^2}{\inf\limits_{x\in I}\left(S(\tau_H^*|x)\right)^2}. \tag{4.15}$$

Finalmente, utilizando los resultados (4.9) y (4.10) en las expresiones (4.13), (4.14) y (4.15), la ecuación (4.12) da como resultado que

$$\sup_{t\in\left[0,\tau_H^*\right]}\sup_{x\in I}|\hat{\varphi}_n(t|x)-\varphi(t|x)| \to 0 \text{ c.s.}$$

lo que concluye la demostración del Teorema 4.1 ⊡

Demostración del Teorema 4.2 Para estudiar el sesgo de $\hat{\varphi}_n(t|x)$, se utiliza (4.11) con $p=1$ y $\frac{1}{z}=\frac{E\left(\hat{Q}\right)}{\hat{Q}}$, lo que conduce a la igualdad:

$$\begin{aligned}
1-\hat{\varphi}_n(t|x) &= \frac{\hat{P}}{\hat{Q}}=\frac{\hat{P}}{E\left(\hat{Q}\right)}\frac{E\left(\hat{Q}\right)}{\hat{Q}} \\
&= \frac{\hat{P}}{E\left(\hat{Q}\right)}\left[1-\left(\frac{\hat{Q}}{E(\hat{Q})}-1\right)\right. \\
&\qquad \left.+\frac{E\left(\hat{Q}\right)}{\hat{Q}}\left(\frac{\hat{Q}}{E\left(\hat{Q}\right)}-1\right)^2\right] \\
&= \frac{\hat{P}}{E\left(\hat{Q}\right)}-\frac{\hat{P}\left(\hat{Q}-E\left(\hat{Q}\right)\right)}{\left(E\left(\hat{Q}\right)\right)^2}+\frac{\hat{P}}{\hat{Q}}\frac{\left(\hat{Q}-E\left(\hat{Q}\right)\right)^2}{\left(E\left(\hat{Q}\right)\right)^2}
\end{aligned} \tag{4.16}$$

Como consecuencia, se tiene que

$$E\left(1-\hat{\varphi}_n(t|x)\right) = A_1 + A_2 + A_3, \tag{4.17}$$

con

$$A_1 = \frac{E\left(\hat{P}\right)}{E\left(\hat{Q}\right)}, \tag{4.18}$$

$$A_2 = -\frac{Cov\left(\hat{P},\hat{Q}\right)}{\left(E\left(\hat{Q}\right)\right)^2}, \tag{4.19}$$

$$A_3 = \frac{E\left[\frac{\hat{P}}{\hat{Q}}\left(\hat{Q}-E\left(\hat{Q}\right)\right)^2\right]}{\left(E\left(\hat{Q}\right)\right)^2}. \tag{4.20}$$

Utilizando la Propiedad 4.3 de la Sección 4.4 se obtiene que:

$$E\left(\hat{P}\right) = P\left(1-\frac{1}{2}d_K A_H\left(t+b|x\right)h^2+o\left(h^2\right)\right), \tag{4.21}$$

$$E\left(\hat{Q}\right) = Q\left(1-\frac{1}{2}d_K A_H\left(t|x\right)h^2+o\left(h^2\right)\right), \tag{4.22}$$

donde

$$\begin{aligned} A_H\left(t|x\right) &= \int_0^t \left[\ddot{H}(s|x)+2\frac{m'\left(x\right)}{m\left(x\right)}\dot{H}(s|x)\right]\frac{dH_1(s|x)}{\left(1-H(t|x)\right)^2} \\ &\quad + \int_0^t \frac{1}{1-H(t|x)}\left[d\dot{H}_1(s|x)+2\frac{m'\left(x\right)}{m\left(x\right)}dH(s|x)\right]. \end{aligned} \tag{4.23}$$

Las expresiones definidas en (4.4) y en (4.23) junto con las fórmulas (4.21) y (4.22) se utilizan para determinar expresiones asintóticas para (4.18) y (4.19):

$$\begin{aligned} A_1 &= \frac{P\left(1-\frac{1}{2}d_K A_H\left(t+b|x\right)h^2+o\left(h^2\right)\right)}{Q\left(1-\frac{1}{2}d_K A_H\left(t|x\right)h^2+o\left(h^2\right)\right)} \\ &= \left(1-\varphi(t|x)\right)\frac{1-\frac{1}{2}d_K A_H\left(t+b|x\right)h^2+o\left(h^2\right)}{1-\frac{1}{2}d_K A_H\left(t|x\right)h^2+o\left(h^2\right)} \\ &= \left(1-\varphi(t|x)\right)\left[1-\frac{1}{2}d_K\left(A_H\left(t+b|x\right)-A_H\left(t|x\right)\right)h^2\right]+o\left(h^2\right) \\ &= \left(1-\varphi(t|x)\right)\left[1-\frac{1}{2}d_K B_H\left(t,t+b|x\right)h^2\right]+o\left(h^2\right), \end{aligned} \tag{4.24}$$

$$A_2 = -\frac{Cov\left(\hat{P},\hat{Q}\right)}{\left(E\left(\hat{Q}\right)\right)^2} = O\left(\frac{1}{nh}\right), \tag{4.25}$$

donde para obtener (4.24) se ha utilizado la identidad definida en (4.11) con $z = 1-\frac{1}{2}d_K A_H\left(t+b|x\right)h^2+o\left(h^2\right)$.

Finalmente, a partir de que $1-\hat{\varphi}_n(t|x) = \dfrac{\hat{P}}{\hat{Q}} \in [0,1]$, entonces el término (4.20) puede ser acotado por:

$$0 \leq A_3 \leq \frac{Var\left[\hat{Q}\right]}{\left(E\left(\hat{Q}\right)\right)^2} = O\left(\frac{1}{nh}\right). \tag{4.26}$$

Utilizando (4.24), (4.25), (4.26) y (4.2) en (4.17) se obtiene una expresión para el sesgo de $\hat{\varphi}_n(t|x)$:

$$E\left(\hat{\varphi}_n(t|x)-\varphi(t|x)\right) = h^2 b(t|x)+o\left(h^2\right)+O\left(\frac{1}{nh}\right). \tag{4.27}$$

Para obtener la varianza de $\hat{\varphi}_n(t|x)$ se utiliza (4.11) con $p=3$ y tomado $\frac{1}{z}=\frac{\left(E\left(\hat{Q}\right)\right)^2}{\hat{Q}^2}$ para obtener:

$$\begin{aligned}\frac{\left(E\left(\hat{Q}\right)\right)^2}{\hat{Q}^2} &= 1+\sum_{i=1}^{3}(-1)^i\left(\frac{\hat{Q}^2-E\left(\hat{Q}\right)^2}{\left(E\left(\hat{Q}\right)\right)^2}\right)^i \\ &\quad+\left(\frac{\hat{Q}^2-E\left(\hat{Q}\right)^2}{\left(E\left(\hat{Q}\right)\right)^2}\right)^4\frac{\left(E\left(\hat{Q}\right)\right)^2}{\hat{Q}^2}. \qquad (4.28)\end{aligned}$$

Por otra parte, la igualdad

$$\hat{Q}^2-\left(E\left(\hat{Q}\right)\right)^2=\left[\hat{Q}-E\left(\hat{Q}\right)\right]^2+2E\left(\hat{Q}\right)\left[\hat{Q}-E\left(\hat{Q}\right)\right]$$

da como resultado

$$\begin{aligned}\left(\frac{\hat{Q}^2-\left(E\left(\hat{Q}\right)\right)^2}{\left(E\left(\hat{Q}\right)\right)^2}\right)^i &= \sum_{j=0}^{i}\binom{i}{j}\left[\frac{\left(\hat{Q}-E\left(\hat{Q}\right)\right)^2}{\left(E\left(\hat{Q}\right)\right)^2}\right]^j \\ &\quad\times\left[\frac{2E\left(\hat{Q}\right)\left[\hat{Q}-E\left(\hat{Q}\right)\right]}{\left(E\left(\hat{Q}\right)\right)^2}\right]^{i-j} \\ &= \sum_{j=0}^{i}\binom{i}{j}\frac{2^{i-j}\left(\hat{Q}-E\left(\hat{Q}\right)\right)^{j+i}}{\left(E\left(\hat{Q}\right)\right)^{j+i}}. \qquad (4.29)\end{aligned}$$

Sustituyendo (4.29) en (4.28) se obtiene:

$$\begin{aligned}\frac{\left(E\left(\hat{Q}\right)\right)^2}{\hat{Q}^2} &= 1+\sum_{i=1}^{3}(-1)^i\sum_{j=0}^{i}\binom{i}{j}\frac{2^{i-j}\left(\hat{Q}-E\left(\hat{Q}\right)\right)^{j+i}}{\left(E\left(\hat{Q}\right)\right)^{j+i}} \\ &\quad+\sum_{j=0}^{4}\binom{4}{j}\frac{2^{4-j}\left(\hat{Q}-E\left(\hat{Q}\right)\right)^{j+4}}{\left(E\left(\hat{Q}\right)\right)^{j+4}}\frac{\left(E\left(\hat{Q}\right)\right)^2}{\hat{Q}^2}. \qquad (4.30)\end{aligned}$$

Utilizando la fórmula (4.30) se obtiene un desarrollo para el momento de segundo orden:

$$\begin{aligned} E\left[(1-\hat{\varphi}_n(t|x))^2\right] &= E\left(\frac{\hat{P}^2}{\hat{Q}^2}\right) = E\left(\frac{\hat{P}^2}{\left(E\left(\hat{Q}\right)\right)^2}\frac{\left(E\left(\hat{Q}\right)\right)^2}{\hat{Q}^2}\right) \\ &= \frac{E\left[\left(\hat{P}-E\left(\hat{P}\right)\right)^2\right]}{\left(E\left(\hat{Q}\right)\right)^2} + \frac{E\left(\hat{P}\right)^2}{\left(E\left(\hat{Q}\right)\right)^2} \\ &\quad + \sum_{i=1}^{3}(-1)^i \sum_{j=0}^{i}\binom{i}{j}\frac{2^{i-j}E\left[\hat{P}^2\left(\hat{Q}-E\left(\hat{Q}\right)\right)^{j+i}\right]}{\left(E\left(\hat{Q}\right)\right)^{j+i+2}} \\ &\quad + \sum_{j=0}^{4}\binom{4}{j}\frac{2^{4-j}E\left[\frac{\hat{P}^2}{\hat{Q}^2}\left(\hat{Q}-E\left(\hat{Q}\right)\right)^{j+4}\right]}{\left(E\left(\hat{Q}\right)\right)^{j+6}}. \end{aligned} \tag{4.31}$$

Definiendo, para $i, j = 0, 1, \ldots$, las notaciones

$$A_{ij} = E\left[\left(\hat{P}-E\left(\hat{P}\right)\right)^i\left(\hat{Q}-E\left(\hat{Q}\right)\right)^j\right], \tag{4.32}$$

$$B_{ij} = E\left[\hat{P}^i\left(\hat{Q}-E\left(\hat{Q}\right)\right)^j\right], \tag{4.33}$$

$$C_i = \left(E\left(\hat{Q}\right)\right)^i, \tag{4.34}$$

$$D_{ij} = E\left[(1-\hat{\varphi}_n(t|x))^i\left(\hat{Q}-E\left(\hat{Q}\right)\right)^j\right], \tag{4.35}$$

entonces, la expresión (4.31) puede escribirse como

$$\begin{aligned} E\left[(1-\hat{\varphi}_n(t|x))^2\right] &= \frac{A_{20}}{C_2} + \frac{B_{10}^2}{C_2} + \sum_{i=1}^{3}(-1)^i\sum_{j=0}^{i}\binom{i}{j}2^{i-j}\frac{B_{2\ i+j}}{C_{i+j+2}} \\ &\quad + \sum_{j=0}^{4}\binom{4}{j}2^{4-j}\frac{D_{2\ j+4}}{C_{j+6}}, \end{aligned} \tag{4.36}$$

donde se puede demostrar fácilmente la relación:

$$B_{2\ i+j} = A_{2\ i+j} + 2B_{10}A_{1\ i+j} - B_{10}^2 A_{0\ i+j},$$

con $i = 1, 2, 3$ y $j = 0, \ldots, i$.

A partir de cálculos similares a los utilizados en la obtención de (4.31) y (4.36), que debido a lo extensos que resultan no serán desarrollados aquí, se puede demostrar que, para momentos de orden igual o superior a 3, la tasa de convergencia es

$$\begin{aligned} E\left[\left(\hat{P} - E\left(\hat{P}\right)\right)^i\right] &= o\left(\frac{1}{nh}\right), \text{ para } i \geq 3, \\ E\left[\left(\hat{Q} - E\left(\hat{Q}\right)\right)^i\right] &= o\left(\frac{1}{nh}\right), \text{ para } i \geq 3. \end{aligned}$$

Ahora, retomando las expresiones (4.32), (4.33), (4.34) y (4.35), y utilizando la desigualdad de Cauchy-Schwartz junto con otras propiedades de acotamiento, se puede demostrar que:

$$A_{01} = A_{10} = 0, \tag{4.37}$$

$$A_{ij} = o\left(\frac{1}{nh}\right), \text{ para } i + j \geq 3, \tag{4.38}$$

$$B_{ij} = o\left(\frac{1}{nh}\right), \text{ para } j \geq 3, \tag{4.39}$$

$$D_{ij} = o\left(\frac{1}{nh}\right), \text{ para } j \geq 3. \tag{4.40}$$

Utilizando las expresiones (4.37), (4.38), (4.39) y (4.40) en la fórmula (4.36), se obtiene que:

$$\begin{aligned} E\left[(1 - \hat{\varphi}_n(t|x))^2\right] &= \frac{A_{20}}{C_2} + \frac{B_{10}^2}{C_2} - \frac{4B_{10}A_{11}}{C_3} - \frac{3B_{10}^2 A_{02}}{C_4} + o\left(\frac{1}{nh}\right) \\ &= \frac{Var\left(\hat{P}\right)}{\left(E\left(\hat{Q}\right)\right)^2} + \frac{E\left(\hat{P}\right)^2}{\left(E\left(\hat{Q}\right)\right)^2} - \frac{4E\left(\hat{P}\right)Cov\left(\hat{P}, \hat{Q}\right)}{\left(E\left(\hat{Q}\right)\right)^3} \\ &\quad + \frac{3E\left(\hat{P}\right)^2 Var\left(\hat{Q}\right)}{\left(E\left(\hat{Q}\right)\right)^4} + o\left(\frac{1}{nh}\right) \end{aligned} \tag{4.41}$$

Por otra parte, sustituyendo (4.16) en el término A_3 de la expresión (4.17) y utilizando las expresiones (4.37), (4.38), (4.39) y (4.40) se obtiene que:

$$\begin{aligned} E\left(1-\hat{\varphi}_n(t|x)\right) &= \frac{B_{10}}{C_1}-\frac{A_{11}}{C_2}+\frac{A_{12}+B_{10}A_{02}}{C_3}-\frac{A_{13}+B_{10}A_{03}}{C_4}+\frac{D_{14}}{C_4} \\ &= \frac{E\left(\hat{P}\right)}{E\left(\hat{Q}\right)}-\frac{Cov\left(\hat{P},\hat{Q}\right)}{\left(E\left(\hat{Q}\right)\right)^2} \\ &\quad +\frac{E\left(\hat{P}\right)Var\left(\hat{Q}\right)}{\left(E\left(\hat{Q}\right)\right)^3}+o\left(\frac{1}{nh}\right) \end{aligned} \tag{4.42}$$

Iglesias Pérez y González Manteiga (1999) demostraron en el Teorema 2(c) que el término residual, $R'_n(t|x)$, obtenido en la representación casi segura de $\hat{S}_h(t|x)$ (expresión definida en (4.51)), es uniformemente despreciable. De forma similar, también pueden obtenerse tasas de convergencia para los momentos de $R'_n(t|x)$. Como consecuencia, aplicando la Propiedad 4.3 junto con el Corolario 4 en Iglesias Pérez y González Manteiga (1999), pueden obtenerse expresiones asintóticas para la estructura de covarianzas del proceso $\left\{\hat{S}_h(\cdot|x)\right\}_{x\in\Omega_X}$. Estos resultados pueden utilizarse para obtener expresiones asintóticas para las varianzas de $\hat{P}$ y $\hat{Q}$, esto es:

$$Var\left(\hat{P}\right) = \frac{1}{nh}v_1(t+b|x)+o\left(\frac{1}{nh}\right), \tag{4.43}$$

$$Var\left(\hat{Q}\right) = \frac{1}{nh}v_1(t|x)+o\left(\frac{1}{nh}\right), \tag{4.44}$$

$$Cov\left(\hat{P},\hat{Q}\right) = \frac{1}{nh}v_2(t,t+b|x)+o\left(\frac{1}{nh}\right), \tag{4.45}$$

donde

$$v_1(t|x) = \frac{\left(S(t|x)\right)^2}{m(x)}c_K C_H(t|x), \tag{4.46}$$

$$v_2(t,y|x) = \frac{S(t|x)\,S(y|x)}{m(x)}c_K C_H(t\wedge y|x), \tag{4.47}$$

$$C_H(t|x) = \int_0^t \frac{dH_1(s|x)}{\left(1-H(s|x)\right)^2}. \tag{4.48}$$

Ahora, utilizando los órdenes de convergencia determinados en (4.43), (4.44) y (4.45) en las expresiones (4.41) y (4.42) se obtiene:

$$\begin{aligned} Var\left(\hat{\varphi}_n(t|x)\right) &= Var\left(1-\hat{\varphi}_n(t|x)\right) = \frac{Var\left(\hat{P}\right)}{\left(E\left(\hat{Q}\right)\right)^2} - \frac{2E\left(\hat{P}\right)Cov\left(\hat{P},\hat{Q}\right)}{\left(E\left(\hat{Q}\right)\right)^3} \\ &\quad + \frac{\left(E\left(\hat{P}\right)\right)^2 Var\left(\hat{Q}\right)}{\left(E\left(\hat{Q}\right)\right)^4} + o\left(\frac{1}{nh}\right). \end{aligned}$$

Finalmente, recopilando las expresiones asintóticas (4.21), (4.22), (4.43), (4.44) y (4.45) junto con las definiciones (4.46), (4.47), (4.48), (4.5) y (4.3), se obtiene:

$$\begin{aligned} Var\left(\hat{\varphi}_n(t|x)\right) &= \frac{1}{nh}\frac{v_1\left(t+b|x\right)}{\left(S(t|x)\right)^2} - \frac{2}{nh}\frac{v_2\left(t,t+b|x\right)S(t+b|x)}{\left(S(t|x)\right)^3} + \\ &\qquad \frac{1}{nh}\frac{v_1\left(t|x\right)\left(S(t+b|x)\right)^2}{\left(S(t|x)\right)^4} + o\left(\frac{1}{nh}\right) \\ &= \frac{1}{nh}\frac{c_K}{m\left(x\right)}\left[\frac{C_H\left(t+b|x\right)\left(S(t+b|x)\right)^2}{\left(S(t|x)\right)^2} - \right. \\ &\qquad \frac{2C_H\left(t|x\right)\left(S(t+b|x)\right)^2}{\left(S(t|x)\right)^2} + \\ &\qquad \left.\frac{C_H\left(t|x\right)\left(S(t+b|x)\right)^2}{\left(S(t|x)\right)^2}\right] + o\left(\frac{1}{nh}\right) \\ &= \frac{1}{nh}\frac{c_K\left[C_H\left(t+b|x\right)-C_H\left(t|x\right)\right]}{m\left(x\right)}\left(\frac{S\left(t+b|x\right)}{S(t|x)}\right)^2 + o\left(\frac{1}{nh}\right) \\ &= \frac{1}{nh}\frac{c_K D_H\left(t,t+b|x\right)}{m\left(x\right)}\left(1-\varphi(t|x)\right)^2 + o\left(\frac{1}{nh}\right) \\ &= \frac{1}{nh}v\left(t|x\right) + o\left(\frac{1}{nh}\right). \end{aligned} \tag{4.49}$$

Por último, a partir de las expresiones (4.27) y (4.49) se obtiene la expresión (4.1) para el error cuadrático medio de $\hat{\varphi}_n(t|x)$, lo que concluye la demostración del Teorema 4.2 $\boxdot$

Demostración de Corolario 4.1 Para probar la normalidad asintótica del estimador $\hat{\varphi}_n(t|x)$ se utiliza la Propiedad 4.3 de la Sección 4.4 junto con el Corolario 4 en Iglesias Pérez y González Manteiga (1999). Estas propiedades conducen a

$$\sqrt{nh}\left[\left(\hat{P},\hat{Q}\right)^t-(P,Q)^t\right]\xrightarrow{\mathrm{d}}N_2\left(\mathbf{b},\mathbf{V}\right),$$

donde

$$\begin{aligned}\mathbf{b} &= (b_1,b_2)^t=-C^{1/2}\frac{1}{2}d_K\left(A_H\left(t+b|x\right)P,A_H\left(t|x\right)Q\right)^t,\\ \mathbf{V} &= \begin{pmatrix} v_{11} & v_{12}\\ v_{21} & v_{22}\end{pmatrix}=\begin{pmatrix} v_1\left(t+b|x\right) & v_2\left(t,t+b|x\right)\\ v_2\left(t,t+b|x\right) & v_1\left(t|x\right)\end{pmatrix}.\end{aligned}$$

Aplicando el teorema de la función continua con $g\left(u,v\right)=\frac{u}{v}$ a la variable aleatoria bivariante $\left(\hat{P},\hat{Q}\right)$ y utilizando el método delta, se obtiene que

$$\sqrt{nh}\left(\frac{\hat{P}}{\hat{Q}}-\frac{P}{Q}\right)\xrightarrow{\mathrm{d}}N\left(\mu,\sigma^2\right),\tag{4.50}$$

donde

$$\begin{aligned}\mu &= \left(\frac{\partial g\left(u,v\right)}{\partial u},\frac{\partial g\left(u,v\right)}{\partial v}\right)\Bigg|_{(u,v)=(P,Q)}\mathbf{b}\\ &= \frac{1}{Q}b_1-\frac{P}{Q^2}b_2=-C^{1/2}\frac{1}{2}d_K\frac{P}{Q}\left(A_H\left(t+b|x\right)-A_H\left(t|x\right)\right)\\ &= -\frac{C^{1/2}}{2}d_K\left(1-\varphi(t|x)\right)B_H\left(t,t+b|x\right)\\ &= C^{1/2}b\left(t|x\right),\\ \sigma^2 &= \left(\frac{\partial g\left(u,v\right)}{\partial u},\frac{\partial g\left(u,v\right)}{\partial v}\right)\Bigg|_{(u,v)=(P,Q)}\mathbf{V}\left(\frac{\partial g\left(u,v\right)}{\partial u},\frac{\partial g\left(u,v\right)}{\partial v}\right)^t\Bigg|_{(u,v)=(P,Q)}\\ &= \frac{1}{Q^2}v_1\left(t+b|x\right)-\frac{2P}{Q^3}v_2\left(t,t+b|x\right)+\frac{P^2}{Q^4}v_1\left(t|x\right)\\ &= \frac{1}{\left(S\left(t|x\right)\right)^2}\left[\frac{\left(S\left(t+b|x\right)\right)^2}{m\left(x\right)}c_KC_H\left(t+b|x\right)-\frac{\left(S\left(t+b|x\right)\right)^2}{m\left(x\right)}c_KC_H\left(t|x\right)\right]\\ &= \frac{\left(1-\varphi(t|x)\right)^2}{m\left(x\right)}c_KD_H\left(t,t+b|x\right)\\ &= v\left(t|x\right).\end{aligned}$$

Finalmente, la demostración del Corolario 4.1 concluye sustituyendo las cantidades $\dfrac{\hat{P}}{\hat{Q}} = 1 - \hat{\varphi}_n(t|x)$ y $\dfrac{P}{Q} = 1 - \varphi(t|x)$ en la fórmula (4.50) $\boxdot$

4.4. Apéndice: resultados auxiliares relativos al estimador de Beran

En este apéndice se enuncian algunos de los resultados asintóticos relativos al estimador de Beran (1981) que se han utilizado en las demostraciones de los Teoremas 4.1 y 4.2, y del Corolario 4.1 relativos al estimador PLG de la PD condicional, $\hat{\varphi}_n(t|x)$. Estos resultados se recopilan y formalizan a continuación, en tres propiedades fundamentales del estimador de la función de supervivencia condicional con datos censurados de Beran (1981). Las demostraciones de estos resultados astintóticos pueden verse en los trabajos citados al final de cada enunciado y en otros, como por ejemplo, en la Tesis de Iglesias Pérez (2001).

Propiedad 4.1 (*Convergencia fuerte uniforme*) *Bajo las hipótesis H3.1 y H4.1-H4.6, junto con* $h \longrightarrow 0$, $\ln n/nh \longrightarrow 0$ *y* $nh/\ln n = O(1)$, *si* $a, b \in \mathbb{R}_+$ *tales que* $a \leq b \leq \tau_H^*$, *para* $x \in I$ *y para* $t \in [a, b]$, *entonces se verifica que*:

$$\left(\hat{S}_h(t|x) - S(t|x)\right) = S(t|x) \sum_{i=1}^{n} B_{ih}(x) \eta\left(\xi_i, \delta_i, t, x\right) + R'_n\left(t|x\right), \qquad (4.51)$$

donde se define la variable aleatoria η como

$$\eta\left(\xi_i, \delta_i, t, x\right) = -\frac{I\left(\xi_i \leq t, \delta_i = 1\right)}{1 - H\left(\xi_i|x\right)} + \int_0^t \frac{I\left(\xi > u\right) dH_1\left(u|x\right)}{\left(1 - H\left(u|x\right)\right)^2},$$

y donde el término residual, $R'_n\left(t|x\right)$, satisface que

$$\sup_{t\in[a,b]} \sup_{x\in I} \left|R'_n\left(t|x\right)\right| = O\left(\frac{\ln n}{nh}\right)^{3/4} \text{ c.s.}$$

Este resultado de convergencia fuerte uniforme se obtiene a partir del Teorema 2(c) en Iglesias Pérez y González Manteiga (1999).

Propiedad 4.2 (*Consistencia fuerte uniforme*) *Bajo las hipótesis H3.1 y H4.1-H4.6 junto con* $h \longrightarrow 0$ *y* $nh \longrightarrow \infty$*, si* $a, b \in \mathbb{R}_+$ *tales que* $a \leq b \leq \tau_H^*$*, para* $x \in I$ *y para* $t \in [a, b]$*, entonces se verifica que:*

$$\sup_{t\in[a,b]} \sup_{x\in I} \left|\hat{S}_h(t|x) - S(t|x)\right| \longrightarrow 0 \text{ c.s.}$$

Además, si se impone la condición $nh^5 \longrightarrow 0$, entonces

$$\sup_{t\in[a,b]} \sup_{x\in I} \left|\hat{S}_h(t|x) - S(t|x)\right| = O\left(\frac{\ln n}{nh}\right)^{1/2} \text{ c.s.}$$

Estos resultados de consistencia del estimador de Beran (1981) se deben a Dabrowska (1989), Corolarios 2.1(iii) y 2.2 (iii), respectivamente.

Propiedad 4.3 (*Normalidad asintótica*) *En un contexto de diseño aleatorio con pesos de tipo Nadaraya-Watson, bajo las mismas hipótesis de la Propiedad 4.1, para* $x \in I$ *y para* $t \in [a, b]$*, se verifica que:*

(a) Si $nh^5 \longrightarrow 0$ y $\ln n/nh \longrightarrow 0$ entonces:

$$\sqrt{nh}\left(\hat{S}_h(t|x) - S(t|x)\right) \xrightarrow{\text{d}} N\left(0, \sigma_A^2(x)\right),$$

donde la varianza asintótica, $\sigma_A^2(x)$, viene dada por la expresión

$$\begin{aligned}\sigma_A^2(x) &= \frac{(S(t|x))^2}{m(x)}\left(\int K^2(u)\,du\right)\int_0^t \frac{dH_1(s|x)}{(1-H(s|x))^2} \\ &= \frac{(S(t|x))^2}{m(x)} c_K C_H(t|x).\end{aligned}$$

(b) Si además se exige que $h = Cn^{-1/5}$, entonces:

$$\sqrt{nh}\left(\hat{S}_h(t|x) - S(t|x)\right) \xrightarrow{\text{d}} N\left(b_A(x), \sigma_A^2(x)\right),$$

donde la varianza asintótica, $\sigma_A^2(x)$, es la misma que en (a) y el sesgo asin-

tótico, $b_A(x)$, viene dado por la expresión

$$\begin{aligned} b_A(x) &= \frac{-C^{5/2}}{2m(x)} S(t|x) \left(\int uK(u)\, du \right) \left(\Phi''(x)\, m(x) + 2\Phi'(x)\, m'(x) \right) \\ &= \frac{-C^{5/2}}{2} S(t|x) d_K \left(\Phi''(x) + \frac{2\Phi'(x)\, m'(x)}{m(x)} \right) \\ &= -\frac{C^{5/2}}{2} d_K S(t|x) A_H(t|x), \end{aligned}$$

y donde se define convenientemente la función $\Phi(u) = E(\eta(\xi, \delta, t, x)\,|X = u)$.

Estos resultados relativos al sesgo y la varianza asintóticos, así como la normalidad asintótica del estimador de Beran (1981), se obtienen a partir del Teorema 2(c) y del Corolario 3(b) en Iglesias Pérez y González Manteiga (1999).

Las expresiones resultantes para las derivadas parciales de la función $\Phi(u)$ en la defínción de $b_A(x)$ se han omitido debido a que son complejas y extensas, y no se utilizan explícitamente en ninguno de los resultados obtenidos en este capítulo. Sin embargo, el lector interesado en estos resultados puede ver las fórmulas explícitas para $\Phi'(u)$ y para $\Phi''(u)$ en la demostración del Corolario 3 (parte (b)) en Iglesias Pérez y González Manteiga (1999).

Capítulo 5

Estudio de la reincidencia de la morosidad crediticia vía análisis de supervivencia

5.1. Introducción

Hasta ahora, en esta memoria se han estudiado modelos de probabilidad, bajo distintos enfoques, como herramientas de predicción de la morosidad en créditos bancarios minoristas como son, por ejemplo, los modelos de puntuación crediticia vía regresión logística para tarjetas de crédito (en el Capítulo 2) y los modelos de probabilidad de mora (PD) en créditos de consumo, ajustados con técnicas de regresión paramétrica, semiparamétrica y no paramétrica con datos censurados (en el Capítulo 3).

En este capítulo, se aborda el problema de la morosidad en créditos personales desde una perspectiva diferente. Ahora, el interés se centra en estudiar el comportamiento de pago irregular de los clientes que, habiendo cometido uno o más impagos en el pasado, al cabo de un tiempo vuelven a incurrir en alguno de los incumplimientos del crédito que se describen más adelante. Este problema, denominado en esta memoria como *reincidencia de la morosidad*, está relacionado con la capacidad de solvencia (más bien limitada) de clientes con un determinado perfil socio-económico que, por lo general, dependen de una única fuente de ingresos mensuales.

En adelante, se entenderá por *reincidencia de la morosidad* el mal comportamiento de pago de los clientes de una entidad financiera, en el que situaciones como los retrasos en los pagos, el descubierto en la cuenta bancaria u otros tipos de incumplimientos, se suceden de manera repetitiva o recurrente en el tiempo.

A juicio de los expertos que colaboraron en este estudio, el aumento de los clientes que incumplen sus obligaciones de manera recurrente en el tiempo provoca un importante deterioro de la calidad crediticia de las carteras, afectando directamente a las provisiones de capital, y por tanto, la solvencia de la entidad. Sin embargo, hasta el momento de finalizar este estudio, no fue posible encontrar trabajos en los que se tratase este problema de forma particular en la literatura sobre riesgo de crédito.

Con el fin de contribuir al estudio de este problema, y a la búsqueda de posibles soluciones que ayuden a la prevención del mismo, la empresa de consultoría financiera AFI-*Analistas financieros internacionales* junto con el grupo de investigación *MODES* de la *Universidade da Coruña*, se propusieron colaborar en el estudio de un modelo de estimación de la *probabilidad de reincidencia de la mora* en carteras de créditos personales.

Para abordar este problema, se adoptó el enfoque no paramétrico de análisis de supervivencia visto en el Capítulo 3, donde la idea central es la formulación de un modelo de probabilidad de mora condicional análogo al estudiado en los Capítulos 3 y 4.

El estudio se realizó a partir de los datos empíricos de una entidad financiera española observados entre enero de 2003 y diciembre de 2008, a los que se tuvo acceso gracias a la colaboración con AFI. En este sentido, es importante explicar que por tratarse de datos reales, la entidad finaciera los proporcionó exclusivamente para su tratamiento estadístico bajo un estricto compromiso de confidencialidad, por el cual queda prohibida la utilización de estos datos para otros fines distintos de aquel para el que fueron recopilados, así como la cesión total o parcial de estos datos a tereceros sin la correspondiente autorización de la entidad.

El resto del capítulo está organizado de la siguiente manera. En la Sección 5.2 se exponen las notaciones, las definiciones y las hipótesis que se utilizarán en este capítulo. En la Sección 5.3 se define la función de distribución condicional de los tiempos de reincidencia, la función de probabilidad condicional de reincidencia de la mora y los estimadores no paramétricos de estas dos

funciones. En la Sección 5.4 se ofrece una aplicación del modelo de reincidencia de la morosidad a una base de datos reales de tarjetas de crédito. Los resultados son evaluados a partir de la inspección y comparación de las curvas, y de los resultados numéricos obtenidos. Por último, en la Sección 5.5 se entrega una breve discusión de los resultados obtenidos y de posibles extensiones del modelo estudiado para trabajos futuros.

5.2. Modelización de la reincidencia de la morosidad crediticia

Como se ha dicho anteriormente, el objetivo de este capítulo es determinar un modelo de probabilidad que dote a la entidad de una herramienta predictiva de la reincidencia de los impagos (u otros tipos de incumplimientos) en una cartera de créditos personales. Por tanto, lo que se busca es un modelo que permita calcular la probabilidad de que un acreditado incurra en un nuevo impago, por ejemplo, el j-ésimo de ellos, en un horizonte futuro de predicción, b, sabiendo que existen $j-1$ impagos anteriores asociados al mismo crédito.

Para tratar este problema, se propone visualizar la existencia de impagos o incumplimientos sucesivos de los créditos como un problema de *análisis de sucesos recurrentes*, donde las variables que se estudian son los tiempos hasta que se producen los sucesivos impagos, el tiempo de interocurrencia entre dos impagos consecutivos y el número total impagos en los que incurre el cliente.

La literatura sobre análisis de sucesos recurrentes es abundante en contextos distintos del riesgo de crédito, como son la epidemiología, las ciencias actuariales, el análisis de fiabilidad y la ingeniería de sistemas, entre otras. Dos de los enfoques más conocidos en el análisis estadístico de sucesos recurrentes son el basado en la teoría de procesos de conteo (Aalen y Husebye (1991), Lin et al. (1998) y Jin et al. (2006)) y el enfoque basado en modelos de regresión de Cox (1972) (Wei et al. (1989), Lin (1994), Lawless y Nadeau (1995) y Gjessing et al. (2010)). Como alternativa a estos enfoques, recientemente han aparecido gran cantidad de trabajos en los que se estudian modelos paramétricos y semiparamétricos para describir los procesos de conteo y las tasas de intensidad de los sucesos recurrentes, como son el enfoque

basado en modelos de tiempos de fallo acelerado (Lu (2005) y Zeng y Lin (2007)) y los modelos de tasas de intensidad aditiva (Schaubel et al. (2006), Liu y Wu (2011) y Liu et al. (2014)). Uno de los trabajos más relevantes sobre técnicas estadísticas para el análisis de sucesos recurrentes es el libro de Cook y Lawless (2007). Allí se tratan la mayoría de las ténicas que se conocen en la actualidad, siendo muy importante el tratamiento de los temas desde la perspectiva del análisis de supervivencia.

Con respecto al enfoque no paramétrico, la estimación de la función de distribución de los tiempos hasta (o entre) los sucesos recurrentes en un contexto de datos dependientes o correlacionados también ha recibido gran atención en la literatura. Algunos de los trabajos más relevantes en este contexto son los debidos a Visser (1996), Wang y Wells (1998), Lin et al. (1999), Wang y Chang (1999), Peña et al. (2001), Schaubel y Cai (2004) y más recientemente Lakhal-Chaieb et al. (2013) y Zhu (2014). Sin embargo, el tipo de dependencia que se aborda en estos trabajos se limita sólo a los tiempos de ocurrencia de los sucesos, sin adentrarse en el caso en que se incorporan covariables al modelo. Otra característica importante del enfoque no paramétrico representado por los trabajos citados anteriormente es, por ejemplo, que los estimadores propuestos por Wang y Wells (1998), Lin et al. (1999) y Peña et al. (2001), todos ellos pueden ser vistos como extensiones de los estimadores de Kaplan-Meier (Kaplan y Meier (1958)) y de Nelson-Aalen (Nelson (1969, 1972) y Aalen (1978)) al caso de datos multivariantes, por lo que estos estimadores no siempre ofrecen estimaciones suavizadas de la curvas de supervivencia o de la razón de fallo acumulativas.

En contraste con lo anterior, ha sido difícil encontrar trabajos en los que se aborde el problema del análisis de sucesos recurrentes en el contexto del riesgo de crédito. Por ello, es destacable el trabajo de Chen et al. (2012), en el que los autores aplican el análisis de sucesos recurrentes para examinar los determinantes de los cambios en las calificaciones de crédito de las empresas. Para ello, adoptan el enfoque basado en modelos de riesgo proporcional de Cox para examinar los cambios en las puntuaciones crediticias de las empresas empleando como variables explicativas los componentes (o ratios) financieros utilizados en el modelo Z-score de Altman (1968). Recientemente, en un contexto de riesgo de crédito similar (para pequeñas y medianas empresas o *pymes*), Gupta et al. (2015) apuntan como una de las causas de la falta de atención que ha recibido el análisis de sucesos recurrentes en la literatura sobre riesgo de crédito, el hecho de no tener en cuenta en la especificación de

los modelos, la posible dependencia existente entre los sucesos de incumplimiento observados en un mismo sujeto durante el tiempo de vida del crédito.

En este capítulo se propone tratar este problema adaptando el enfoque no paramétrico de supervivencia estudiado en el Capítulo 3 al estudio de las distribuciones condicionadas de los tiempos asociados a los impagos recurrentes de un grupo de clientes en una cartera de créditos personales.

Si bien, bajo el modelo de supervivencia del Capítulo 3 se acepta la hipótesis de independencia entre los tiempos de entrada en mora entre distintos acreditados, en este nuevo enfoque, denominado *modelo de reincidencia para la morosidad crediticia*, los tiempos hasta el j-ésimo impago o incumplimiento observados en un mismo cliente ya no son independientes. La razón de la no independencia de sucesos en este caso puede explicarse con un sencillo ejemplo. Supóngase que un cliente de condición socio-económica media, que posee un único ingreso familiar al mes, repentinamente pierde su fuente de ingresos. En un caso como este, lo más habitual es que el cliente reaccione reorganizando las prioridades de sus compromisos económicos debido a la pérdida parcial, o total, de su capacidad de solvencia. Como consecuencia, es razonable creer que en algún momento del tiempo el cliente dejará de pagar sus deudas, o al menos, que comenzará a retrasarse de forma recurrente en el pago de las mismas. De este ejemplo se deduce fácilmente que los impagos o incumplimientos recurrentes observados en el tiempo de vida de un crédito pueden estar provocados por las mismas causas, y por tanto, la hipótesis de independiencia entre dichos tiempos no es realista.

Teniendo en cuenta lo anterior, en este capítulo se propone estudiar un modelo de estimación de la probabilidad de reincidenicia de los impagos o incumplimientos a partir del estimador de la función de supervivencia condicional tipo Beran (1981) propuesto por Akritas y Van Keilegom (2003), en el que el estimador PLG de Beran (1981) es adaptado al caso de tiempos de vida múltiples con censura. En este capítulo, se añade un paso más al cálculo del estimador PLG de Akritas y Van Keilegom (2003) modificándolo para incorporar también la información aportada por la covariable unidimensional puntuación crediticia, X.

Por último, debido a las limitaciones de tiempo y a la extensión de esta memoria, se ha dejado para un futuro trabajo el estudio asintótico del estimador PLG de Akritas y Van Keilegom (2003) en su versión con covariables y del estimador de la probabilidad condicional de reincidencia con tiempos

de impagos múltiples que se obtiene a partir del anterior. Las propiedades asintóticas de ambos estimadores pueden obtenerse siguiendo argumentos y técnicas análogas a las utilizadas en el Capítulo 4 de esta memoria, así como los resultados asintóticos obtenidos por Akritas y Van Keilegom (2003).

5.2.1. Definiciones y notación

A continuación, se utilizarán las siguientes definiciones y notación:

(i) Las variables $T_{i1}, T_{i2}, \ldots, T_{iJ_i}$, donde $J_i \geq 1$ para todo i, con $1 \leq i \leq n$ son los tiempos hasta que se producen los primeros J_i impagos del crédito i-ésimo. Además, se denota por T_{i0} el tiempo de formalización del crédito i.

(ii) Las variables $C_{i1}, C_{i2}, \ldots, C_{iJ_i}$, con $1 \leq i \leq n$ son los tiempos de censura que impiden observar los primeros J_i impagos del crédito i-ésimo.

(iii) La variable ξ_{ij}= $\min\{T_{ij}, C_{ij}\}$, para todo i, j, con $1 \leq i \leq n$ y $1 \leq j \leq J_i$, es la variable que registra el tiempo de vida observado asociado al j-ésimo impago del crédito i-ésimo. Por simplicidad de notación, en adelante, para el caso de $j = 1$, se utilizará la convención ξ_{i1}=$\min\{T_{i1}, C_{i1}\}$=$\min\{T_i, C_i\}$=ξ_i, para todo i, con $1 \leq i \leq n$.

(iv) La variable, X_i, representa la puntuación crediticia observada en el crédito i-ésimo. La variable X posee función de distribución absolutamente continua, denotada por $M(x)$, y función de densidad, $m(x) = M'(x)$, cuyo soporte se denota por $\Omega_X = \{x \in \mathbb{R}^+ : m(x) > 0\}$.

Prescindiendo del subíndice i para referirse al crédito i-ésimo, las siguientes definiciones serán válidas para los n créditos de la cartera:

(v) La variable tiempo hasta el j-ésimo impago, T_j, posee función de distribución condicional dados $T_{j-1} = s$ y $X = x$ definida por

$$\begin{aligned} F_{j|j-1}(t|s,x) &= P\left(T_j \leq t | T_{j-1} = s, X = x\right) \\ &= 1 - S_{j|j-1}(t|s,x), \end{aligned}$$

con $0 \leq s \leq t$ y donde $S_{j|j-1}(\cdot|\cdot, x)$ es la función de supervivencia condicional de T_j dados T_{j-1} y X.

(*vii*) La variable censurante del tiempo hasta el j-ésimo impago, C_j, posee función de distribución condicional dados $C_{j-1} = s$ y $X = x$ definida por

$$\begin{aligned} G_{j|j-1}(t|s,x) &= P\left(C_j \leq t | C_{j-1} = s, X = x\right) \\ &= 1 - \bar{G}_{j|j-1}(t|s,x), \end{aligned}$$

con $0 \leq s \leq t$ y donde $\bar{G}_{j|j-1}(\cdot|\cdot,x)$ es la función de supervivencia condicional de la variable C_j dados C_{j-1} y X.

(*viii*) La variable tiempo de vida observado hasta el j-ésimo impago, ξ_j, posee función de distribución condicional dados $\xi_{j-1} = s$ y $X = x$ definida por

$$\begin{aligned} H_{j|j-1}(t|s,x) &= P\left(\xi_j \leq t | \xi_{j-1} = s, X = x\right) \\ &= 1 - \bar{H}_{j|j-1}(t|s,x), \end{aligned}$$

con $0 \leq s \leq t$ y donde $\bar{H}_{j|j-1}(\cdot|\cdot,x)$ es la función de supervivencia condicional de la variable ξ_j dados ξ_{j-1} y X.

(*ix*) Se denota por $\mathcal{F}_{j-1} = \{T_1, T_2, \ldots, T_{j-1}\}$ el conjunto de tiempos hasta el impago $j-1$. $\mathcal{F}_{j-1}$ representa la historia o pasado inmediatamente anterior al tiempo hasta el j-ésimo impago, T_j. Así, la función de distribución condicional de T_j dada una realización de $\mathcal{F}_{j-1}$ y $X = x$ se define por

$$\begin{aligned} F_{j|\mathcal{F}_{j-1}}(t|x) &= P\left(T_j \leq t | T_1 = t_1, T_2 = t_2, \ldots, T_{j-1} = t_{j-1}, X = x\right) \\ &= P\left(T_j \leq t | \mathcal{F}_{j-1}, X = x\right) \\ &= 1 - S_{j|\mathcal{F}_{j-1}}(t|x). \end{aligned} \tag{5.1}$$

5.2.2. Hipótesis del modelo de reincidencia

El modelo de reincidencia de la morosidad verifica las siguientes hipótesis:

H 5.1 *El número máximo de impagos o incumplimientos en los que puede incurrir el i-ésimo cliente es J_i, con $1 \leq i \leq n$, donde T_{J_i} es el tiempo hasta que el crédito es considerado como moroso según la regulación de Basilea II, es decir, cuando el retraso del pago se extiende por un periodo igual o superior a 90 días.*

5.2.3. Función de probabilidad condicional de reincidencia

Se define la función de probabilidad condicional de reincidencia de la mora, denotada por $PD_{j|j-1}$, como la probabilidad condicional de que ocurra el j-ésimo nuevo impago (o incumplimiento) en el tiempo T_j, sabiendo que el crédito ha incurrido en impago en un tiempo anterior, T_{j-1}, y dada la puntuación crediticia $X = x$. Por analogía con la notación utilizada en el Capítulo 3 para la función de probabilidad de mora condicional, $\varphi(\cdot|x)$, en este capítulo se utilizará la equivalencia $PD_{j|j-1} \equiv \varphi_{j|j-1}(\cdot|\cdot, x)$ para referirse a la probabilidad condicional de reincidencia, para todo j, con $1 \leq j \leq J$.

La probabilidad condicional de reincididencia del j-ésimo impago en algún instante anterior a $t+b$ sabiendo que es posterior al instante t, que el impago $j-1$ se ha producido en el instante s y que la puntuación crediticia X toma el valor x, se obtiene a parir de la siguiente expresión:

$$\begin{aligned}
\varphi_{j|j-1}(t+b|s,t,x) &= P(T_j \leq t+b|T_j > t, T_{j-1} = s, X = x) \\
&= \frac{P(t < T_j \leq t+b|T_{j-1} = s, X = x)}{P(T_j > t|T_{j-1} = s, X = x)} \\
&= \frac{F_{j|j-1}(t+b|s,x) - F_{j|j-1}(t|s,x)}{1 - F_{j|j-1}(t|s,x)} \\
&= 1 - \frac{S_{j|j-1}(t+b|s,x)}{S_{j|j-1}(t|s,x)}, \qquad (5.2)
\end{aligned}$$

donde $b > 0$ es el horizonte de predicción del j-ésimo nuevo impago y el

soporte de $\varphi_{j|j-1}(t+b|t,x)$ se define como el conjunto:

$$\boldsymbol{D}_{j-1} = \left\{(s,t,x) \in [0,\infty)^2 \times \mathbb{R} : S_{j|j-1}(t|s,x) > 0\right\}.$$

De la fórmula (5.2) se deduce que en el caso particular de $j = 1$, se obtiene que T_1 es el tiempo hasta el primer impago, es decir, que aún no existe la primera reincidencia del impago, por lo que T_0 representa el momento de la formalización del crédito. Así, tomando $T_1 = t+b$ y $T_0 = 0$ se obtiene el caso particular en que la probabilidad de reincidencia, $PD_{1|0} = \varphi_1(t+b|t,x)$,

coincide con la probabilidad de mora, $\varphi(t|x)$, definida en (3.1), esto es

$$\begin{aligned}\varphi_1(t+b|t,x) &= P(T_1 \leq t+b|T_1 > t, X = x) \\ &= \frac{P(t < T_1 \leq t+b|X=x)}{P(T_1 > t|X=x)} \\ &= 1 - \frac{S_1(t+b|x)}{S_1(t|x)}. \end{aligned} \tag{5.3}$$

Fijados tres valores para s, t y x, y el horizonte de predicción, b, la probabilidad condicional de mora definida en (5.3) puede estimarse con alguna de las fórmulas de la PD propuestas en el Capítulo 3, por ejemplo, utilizando el estimador PLG, $\hat{\varphi}_n^{PLG}(t|x)$, definido en (3.23).

En general, la probabilidad condicional de reincididencia del j-ésimo impago en el instante $T_j = t + b$ dados los $j-1$ impagos anteriores en $T_1 = t_1, T_2 = t_2, \ldots, T_{j-1} = t_{j-1}$ y dada la puntuación crediticia $X = x$, se obtiene reemplazando la función de supervivencia condicional definida en (5.1) en la fórmula (5.2), esto es:

$$\begin{aligned}\varphi_{j|\mathcal{F}_{j-1}}(t+b|t_1,\ldots,t_{j-1},t,x) &= P(T_j \leq t+b|\, T_1 = t_1, \ldots, \\ &\qquad T_{j-1} = t_{j-1}, T_j > t, X = x) \\ &= \frac{P(t \leq T_j \leq t+b|\mathcal{F}_{j-1} = \boldsymbol{t}_{j-1}, X = x)}{P(T_j > t|\mathcal{F}_{j-1} = \boldsymbol{t}_{j-1}, X = x)} \\ &= \frac{S_{j|\mathcal{F}_{j-1}}(t|x) - S_{j|\mathcal{F}_{j-1}}(t+b|x)}{S_{j|\mathcal{F}_{j-1}}(t|x)} \\ &= 1 - \frac{S_{j|\mathcal{F}_{j-1}}(t+b|x)}{S_{j|\mathcal{F}_{j-1}}(t|x)}, \end{aligned} \tag{5.4}$$

donde se denota por $\boldsymbol{t}_{j-1} = (t_1, \ldots, t_{j-1})$ al vector de tiempos correspondientes a la realización del conjunto $\mathcal{F}_{j-1}$ sobre la que se condiciona el cálculo de la probabilidad del j-ésimo impago.

5.3. Estimación de la probabilidad condicional de reincidencia

En la fórmula (5.2) se ve que la probabilidad condicional de reincidencia de la mora se define a partir del cociente entre dos funciones de supervivencia.

Por tanto, para estimar la función $\varphi_{j|j-1}(t+b|t,x)$ se requiere únicamente de un estimador de la función de supervivencia condicional (o equivalentemente de la función de distribución condicional) del tiempo hasta el j-ésimo impago, $S_{j|j-1}(t+b|t,x)$, dados el anterior impago en el tiempo $T_{j-1}=t$ y la puntuación crediticia $X=x$.

Como en este capítulo se considera que los tiempos entre sucesivos impagos son dependientes y algunos de ellos no son observables debido a la censura aleatoria que los afecta, en este capítulo se trabajará en un contexto de regresión no paramétria en diseño aleatorio. En un contexto como este, lo que observa la entidad es una muestra de n realizaciones independientes de la forma, $\left\{\left(\boldsymbol{\xi}_{ij},\boldsymbol{\delta}_{ij},X_{ij}\right)\right\}_{i=1}^{n}$, con idéntica distribución que el vector $\left(\boldsymbol{\xi}_j,\boldsymbol{\delta}_j,X_j\right)$, donde $\boldsymbol{\xi}_j=\left(\left(T_1\wedge C_1\right),\left(T_2\wedge C_2\right),\ldots,\left(T_j\wedge C_j\right)\right)$, el vector $\boldsymbol{\delta}_j=\left(I\left(T_1\leq C_1\right),\left(T_2\leq C_2\right),\ldots,\left(T_j\leq C_j\right)\right)$ y $T_j\wedge C_j=\min\{T_j,C_j\}$, para todo $j\geq 1$.

Por este motivo, se propone utilizar como estimador de la función de supervivencia condicional, $S_{j|\mathcal{F}_{j-1}}(\cdot|x)$, una versión modificada del estimador PLG de Akritas y Van Keilegom (2003), denotado por $\hat{S}^{AVK}_{j|\mathcal{F}_{j-1}}(\cdot|x)$, en el que además de conocer los tiempos de ocurrencia de los $j-1$ impagos anteriores, $\mathcal{F}_{j-1}$, también se tiene en cuenta la información aportada por la puntuación crediticia, X. Así, la fórmula del estimador PLG de Akritas-Van Keilegom, $\hat{S}^{AVK}_{j|\mathcal{F}_{j-1}}(\cdot|x)$, con covariable univariante viene dada por:

$$\hat{S}^{AVK}_{j|\mathcal{F}_{j-1}}(t|x)=\prod_{\delta_{ij}=1}\left(1-\frac{I\left(\xi_{ij}\leq t\right)w_{ih_{j-1}}(\boldsymbol{t}_{j-1})B_{ih_X}(x)}{\sum_{l=1}^{n}I\left(\xi_{lj}\geq\xi_{ij}\right)w_{lh_{j-1}}(\boldsymbol{t}_{j-1})B_{lh_X}(x)}\right), \tag{5.5}$$

donde $\boldsymbol{t}_{j-1}=(t_1,\ldots,t_{j-1})$ es el vector de tiempos asociados a los $j-1$ impagos ocurridos antes del j-ésimo nuevo impago que ocurrirá algún $T_{ij}=t$, tal que $t\geq t_{j-1}\geq t_{j-2}\geq\ldots\geq t_1$ y donde los términos, $w_{ih_{j-1}}(\cdot)$ y $B_{ih_X}(\cdot)$, son los pesos no paramétricos de Nadaraya-Watson asociados a los tiempos hasta los primeros $j-1$ impagos, $T_1,\ldots,T_{j-1}$, y a la covariable, X, respectivamente.

Los pesos no paramétricos, $w_{ih_{j-1}}(\boldsymbol{t}_{j-1})$, asociados a la distribución conjunta de los $j-1$ impagos, $T_{i1},\ldots,T_{ij-1}$, del crédito i-ésimo vienen dados por:

$$w_{ih_{j-1}}(\boldsymbol{t})=\begin{cases}\frac{K(\boldsymbol{H}_n^{-1/2}(\boldsymbol{\xi}_{ij-1}-\boldsymbol{t}_{j-1}))}{\sum_{l=1}^{n}K(\boldsymbol{H}_n^{-1/2}(\boldsymbol{\xi}_{lj-1}-\boldsymbol{t}_{j-1}))} & \text{si}\quad \boldsymbol{\delta}_{ij-1}=\mathbf{1}\\ 0 & \text{si}\quad \boldsymbol{\delta}_{ij-1}=\mathbf{0}\end{cases},$$

donde el vector $\boldsymbol{\delta}_{ij-1}$ es el análogo de la variable indicadora, δ_{ij}, para el vector de tiempos de los primeros $j-1$ impagos, $\boldsymbol{T}_{ij-1} = (T_{i1}, \ldots, T_{ij-1})$, para todo i, con $1 \leq i \leq n$. La función núcleo, $K(\cdot)$, es una función de densidad multivariante esféricamente simétrica con soporte compacto y momentos de segundo orden finitos que satisface las siguientes condiciones:

$$\begin{aligned} \int_{\mathbb{R}^d} K(\boldsymbol{u})\, d\boldsymbol{u} &= \mathbf{1} \\ \int_{\mathbb{R}^d} \boldsymbol{u} K(\boldsymbol{u})\, d\boldsymbol{u} &= \mathbf{0} \\ \int_{\mathbb{R}^d} \boldsymbol{u}\boldsymbol{u}^t K(\boldsymbol{u})\, d\boldsymbol{u} &= c_K \boldsymbol{I}_d \quad , \end{aligned}$$

es decir,

$$c_K \boldsymbol{I}_d = \begin{cases} \int u_i u_j K(\boldsymbol{u})\, d\boldsymbol{u} = 0 & \text{si} \quad i \neq j \\ \int u_j^2 K(\boldsymbol{u})\, d\boldsymbol{u} = c_K & \text{si} \quad i = j \end{cases}$$

para $i = 1, \ldots, d$, donde $\boldsymbol{I}_d$ es la matriz identidad en $\mathbb{R}^d$.

El parámetro de suavizado asociado a la distribución conjunta del vector de tiempos hasta los primeros j impagos, $\boldsymbol{T}_j = (T_1, T_2, \ldots, T_j)$, viene dado por la matriz simétrica definida positiva,

$$\boldsymbol{H}_n = \begin{bmatrix} h_1 & h_{12} & \cdots & h_{1j} \\ h_{21} & h_2 & \cdots & h_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ h_{j1} & h_{j2} & \cdots & h_j \end{bmatrix},$$

donde cada componente, $h_{kj} \equiv h_{kj}(n)$, corresponde al parámetro de suavizado asociado al vector de tiempos, (T_k, T_j) y es una sucesión real decreciente a medida que aumenta el tamaño de la muestra, n. La matriz de paramétros de suavizado, $\boldsymbol{H}_n$, verifica la condición:

$$\boldsymbol{H}_n \longrightarrow \mathbf{0} \text{ y } n^{-1} |\boldsymbol{H}_n| \longrightarrow 0 \text{ cuando } n \longrightarrow \infty,$$

donde $|A|$ es el determinante de la matriz A.

Además, los pesos de Nadaraya-Watson, $B_{ih_X}(\cdot)$, asociados a la covariable, X, vienen dados por:

$$B_{ih_X}(x) = \frac{K((x - x_i)/h_X)}{\sum_{l=1}^{n} K((x - x_l)/h_X)}, 1 \leq i \leq n,$$

y satisfacen las condiciones impuestas en el Capítulo 3 (fórmula (3.22)) en el caso del estimador de Beran (1981), $\hat{S}_h^{PLG}(t|x)$.

Finalmente, sustituyendo las funciones de supervivencia condicional teóricas, $S_{j|\mathcal{F}_{j-1}}(t+b|x)$ y $S_{j|\mathcal{F}_{j-1}}(t|x)$, en la fórmula (5.4) por sus estimadores no paramétricos definidos en (5.5), se obtiene como resultado un estimador PLG de la probabilidad condicional de reincidencia del j-ésimo impago basado en el estimador PLG de Akritas y Van Keilegom (2003). Llamando a este estimador, $\hat{\varphi}^{AVK}_{j|\mathcal{F}_{j-1}}$, su fórmula viene dada por la expresión:

$$\hat{\varphi}^{AVK}_{j|\mathcal{F}_{j-1}}(t+b|x) = 1 - \frac{\hat{S}^{AVK}_{j|\mathcal{F}_{j-1}}(t+b|x)}{\hat{S}^{AVK}_{j|\mathcal{F}_{j-1}}(t|x)}. \tag{5.6}$$

5.4. Aplicación a una cartera de tarjetas de crédito

En esta sección se exponen los resultados de un estudio empírico sobre la estimación no paramétrica de la probabilidad de reincidencia de los impagos que puede observar una entidad financiera durante la trayectoria de cada uno de los clientes de una cartera de créditos.

Para tratar estadísticamente el problema de la existencia de impagos o incumplimientos que se reiteran en el tiempo para un mismo acreditado, en este capítulo se han combinado las ideas del análisis de sucesos recurrentes con el modelo no paramétrico de estimación de la función de supervivencia condicional con datos censurados visto en el Capítulo 3, representado por el estimador PLG de Beran (1981). Bajo este enfoque, se ha utilizado el estimador tipo Beran (1981) propuesto por Akritas y Van Keilegom (2003) para la función de distribución bivariante con datos censurados. En este caso, el estimador PLG de Akritas y Van Keilegom (2003) se utiliza para estimar la función de distribución condicional del tiempo hasta el j-ésimo impago, T_j, dados los tiempos hasta el impago $j-1$, $T_1, T_2, \ldots, T_{j-1}$, y dada la covariable $X = x$, particularizando la definición del estimador de Akritas-Van Keilegom al caso en el que existe una covariable unidimensional. Así, aplicando un planteamiento similar al utilizado en el Capítulo 3 (basado en las ideas de Cao et al. (2009)) para el estimador PLG de la PD definido en (3.23), en este capítulo se obtiene un estimador núcleo análogo al de la PD para la

probabilidad condicional de reincidencia de los impagos de los créditos.

El estimador PLG de Akritas-Van Keilegom, $\hat{\varphi}^{AVK}_{J|\mathcal{F}_{J-1}}$, se ha utilizado para obtener las probabilidades condicionales de las primeras $J-1$ reincidencias de impagos simulados para cada uno de los clientes de una base de datos reales de tarjetas de crédito. Las características descriptivas de los datos y los resultados obtenidos con este nuevo mecanismo predictivo de la morosidad se analizan a continuación.

5.4.1. Análisis de la base de datos

La base de datos utilizada en este estudio se compone de 89 646 registros de tarjetas de crédito observadas entre enero de 2003 y diciembre de 2008, donde 15 193 de ellas corresponden a clientes morosos y los restantes 74 453 son clientes no morosos, es decir, una tasa de morosidad global de la cartera de un 16.95 %.

Si bien, en este estudio se dispuso de un buen tamaño de muestra, $n = 89\,646$, en la que los datos no están excesivamente desbalanceados (aproximadamente, 1 de cada 5 créditos resultaron morosos), desafortunadamente, debido a dificultades en la extracción de los datos orginales, no fue posible acceder a las fechas en las que se produjeron los sucesivos incumplimientos de los clientes (los impagos de las cuotas de las tarjetas), por lo que sólo se conoció la fecha de formalización de la tarjeta, la fecha de fin del estudio y, en el caso de los clientes morosos, la fecha de entrada en situación de mora de acuerdo con la normativa de Basilea II, es decir, los clientes cuyo retraso en el pago de la tarjeta se prolongó por un período igual o superior a los 90 días (ver Sección 1.1.3 del Capítulo 1).

A continuación, en la Figura 5.1, se representa un gráfico con la evolución mensual de la tasa de mora de la cartera registrada entre enero de 2003 y diciembre de 2008. Allí, se observan picos de morosidad en distintas épocas del año, detectándose un posible patrón estacional entre los meses de abril y mayo en el período 2003-2007, que es justamente cuando se alcanza el valor más alto de la serie ($TM = 23.82\,\%$).

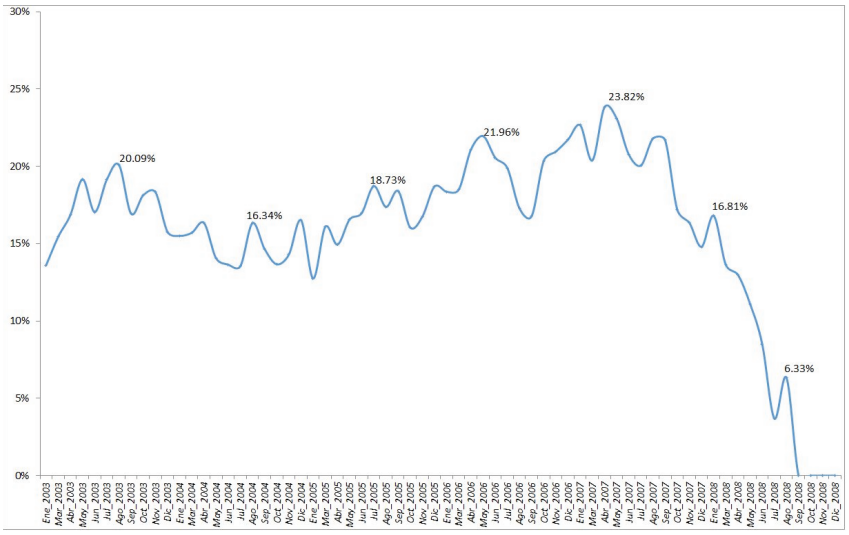

Figura 5.1. Tasa de morosidad mensual de las tarjetas de crédito entre 2003 y 2008.

La variable tiempo hasta la mora o madurez

La variable tiempo hasta la mora, o madurez, se obtiene a partir de las fechas de formalización y de mora contenidas en la base de datos de la entidad colaboradora.

Tabla 5.1. Estadísticos descriptivos del tiempo de vida de los créditos.

Muestra	*min*	Q_2	*media*	*max*	*d.e.*	*asimetría*	*curtosis*
Morosos	2.00	21.0	24.23	70.00	14.83	0.814	1.647
No morosos	1.00	29.0	30.86	73.00	18.18	0.326	2.126
Todos	1.00	29.0	29.74	73.00	17.83	0.416	2.208

En la Tabla 5.1 se observan valores moderados de asimetría positiva y de curtosis en las tres muestras. Sin embargo, estas se diferencian significativamente de la distribución normal estándar. Al realizar las pruebas de normalidad correspondientes (vía los contrastes de bondad del ajuste de *Lilliefors* y de *Jarque-Bera*) se obtuvo como resultado el rechazo de la hipótesis de normalidad con p-valores $<< 0.0001$.

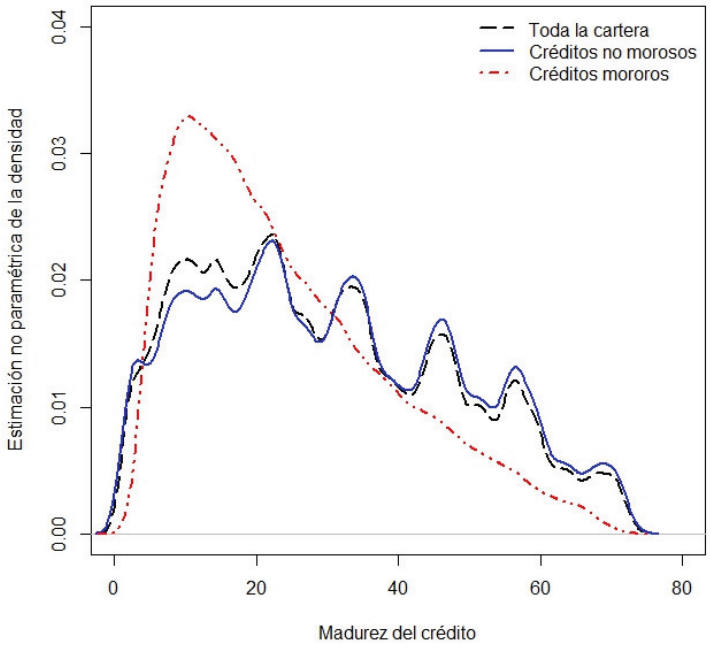

Figura 5.2. Estimación de la función de densidad de la madurez de los créditos.

En la Figura 5.2 se representa la estimación no paramétrica de la función de densidad del tiempo de vida observado de los créditos en la muestra completa (trazo segmentado de color negro), en la muestra de créditos morosos (trazo segmentado de color rojo) y en la muestra de créditos no morosos (línea de color azul). Para ello, se utilizó el estimador núcleo de *Rosenblatt-Parzen* (Rosenblatt (1956) y Parzen (1962)) con función núcleo gaussiana y parámetro de suavizado, h, obtenido con el método *plug-in* de Sheather y Jones (1991), implementado en la librería *KernSmooth* (Wand y Ripley (2008)) del paquete estadístico *R* (*R* Core Team (2015)). Allí se observan con claridad las características asimétricas del tiempo de vida de los créditos, tanto de los clientes morosos como de los no mororos. También se aprecia que, debido a la alta proporción de censura presente en la muestra, la curva de densidad de toda la muestra es muy parecida a la curva de densidad de los créditos no morosos.

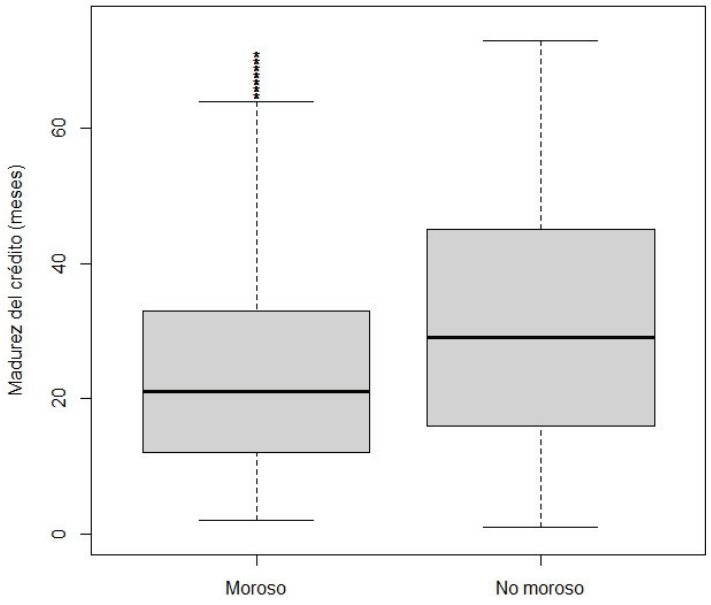

Figura 5.3. Gráfico de cajas para la madurez observada de los créditos.

En la Figura 5.3 se representan dos gráficos de cajas para comparar el grado de dispersión de los datos entre los créditos morosos y los no morosos. En el caso de los créditos morosos, se aprecia claramente tanto la asimetría positiva como la existencia de valores atípicos, mientras que en la muestra de créditos no morosos la asimetría es menos acusada y no se detectan valores atípicos.

La variable puntuación crediticia

La variable puntuación crediticia es un valor numérico calculado por la entidad a partir de un modelo de regresión logística similar al estudiado en el Capítulo 2 de esta memoria. Su valor varía entre 0 y 100, y se utiliza para asociar a cada acreditado el grado de la propensión del cliente a cometer impagos o incumplimientos del crédito.

Tabla 5.2. Estadísticos descriptivos de la puntuación crediticia de los créditos.

Muestra	*min*	Q_2	*media*	*max*	*d.e.*	*asimetría*	*curtosis*
Morosos	0.00	4.84	7.12	48.40	6.707	0.814	1.647
No morosos	0.00	1.16	2.55	44.10	3.497	3.166	17.426
Todos	0.00	1.51	3.32	48.40	4.55	2.849	13.779

En la Tabla 5.2 se observa que los coeficientes de asimetría y curtosis obtenidos para las tres muestras son muy distintos de los valores asociados a la distribución normal estándar ($asimetría = 0$ y $curtosis = 3$). Después de realizar los contrastes de bondad del ajuste, se rechaza la hipótesis de normalidad de la puntuación crediticia con p-valores $<< 0.0001$.

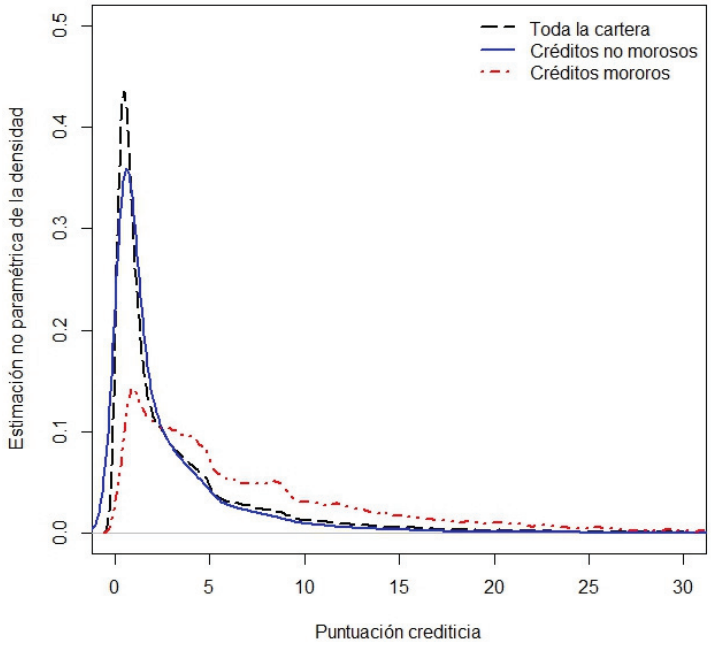

Figura 5.4. Estimación de la función de densidad de la puntuación crediticia.

En la Figura 5.4 se ilustran las estimaciones no paramétricas de las funciones de densidad de los tiempos de vida observados de los créditos en la

muestra completa (trazo segmentado de color negro), en la muestra de créditos morosos (trazo segmentado de color rojo) y en la muestra de créditos no morosos (curva de color azul). Allí, se observa claramente tanto el alto grado de asimetría positiva de las curvas estimadas de la densidad, como el marcado grado de apuntamiento (curtosis), con valores altos de la función de densidad en torno a los valores pequeños de las puntuaciones, siendo esto último más acusado en los créditos no morosos que en los morosos, como se ve en la Tabla 5.2.

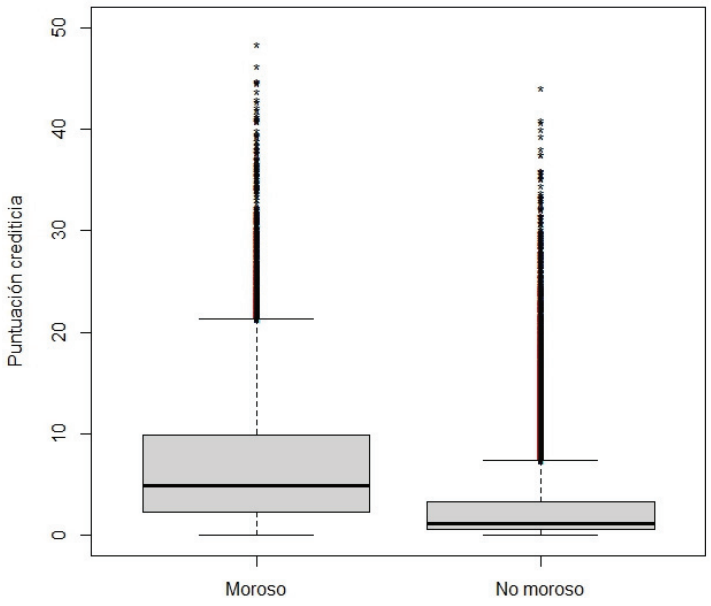

Figura 5.5. Gráfico de cajas para la puntuación crediticia.

En la Figura 5.5 se representan dos gráficos de cajas para comparar el grado de dispersión de la puntuación crediticia entre los créditos morosos y los no morosos. Ambos gráficos son útiles ya que permiten verificar lo observado para ambas muestras en las curvas ilustradas en la Figura 5.4, es decir, que con respecto a la puntuación crediticia, ambos tipos de créditos poseen distribuciones con fuerte asimetría positiva y colas derechas pesadas debido a la alta concentración de datos atípicos.

5.4.2. Resultados de la estimación no paramétrica de la probabilidad condicional de reincidencia

En este apartado se analizan los resultados de las estimaciones no paramétricas de la función de supervivencia condicional y de la función de probabilidad condicional de reincidencia de los impagos en una cartera de créditos con datos reales. Sin embargo, tal y como se explicó al comienzo de esta sección, al no contar con los registros de las fechas en las que se produjeron los sucesivos impagos de las tarjetas, no fue posible calcular los tiempos de supervivencia de los mismos. Por este motivo, para poder examinar empíricamente los resultados de las estimaciones no paramétricas, fue necesario completar la información de la muestra con tiempos de impagos ficticios, generados aleatoriamente para cada uno de los créditos de la cartera.

Para obtener los tiempos de los impagos recurrentes asociados a las tarjetas, se adopta el siguiente razonamiento:

Denotando por Π el conjunto de todos los créditos de la cartera, en virtud de la hipótesis H5.1, existe un número máximo de reincidencias de los impagos, J, que pueden observarse en cada crédito. En este caso, por simplicidad computacional y de interpretación de los resultados, este valor se fijó en $J = 3$, para todo $i \in \Pi$. Los datos muestrales se completaron generando aleatoriamente vectores de la forma $(\boldsymbol{T}_i, \boldsymbol{C}_i, \boldsymbol{\delta}_i)$ con $\boldsymbol{T}_i = (T_{i1}, T_{i2}, T_{i3})$, $\boldsymbol{C}_i = (C_{i1}, C_{i2}, C_{i3})$ y $\boldsymbol{\delta}_i = (\delta_{i1}, \delta_{i2}, \delta_{i3})$, donde los tiempos de los impagos, T_{ij}, cumplen la condición de orden cronológico:

$$T_{i0} + 1 \leq T_{i1} < T_{i2} < T_{i3} \leq \tau_i - 2, \tag{5.7}$$

y donde T_{i0} y τ_i son las fechas o tiempos de formalización y de entrada en mora del credito i, respectivamente.

Observación 5.1 *La condición (5.7) permite establecer el orden cronológico en el que se producen los sucesivos impagos de los créditos. La condición inferior, $T_{i1} \geq T_{i0} + 1$, sirve para asegurar que ningún crédito pueda incurrir en un impago antes de que transcurra el primer mes desde su formalización. Además, la condición superior, $T_{ij} \leq \tau_i - 2$, permite asegurar que la j-ésima reincidencia, T_{ij}, ocurrirá, como máximo, 2 meses antes del tiempo en el que el crédito i-ésimo se convierte en moroso según la normativa de Basilea II, denotado por τ_i. Así, el i-ésimo crédito sólo puede reincidir en los impagos entre los tiempos $T_{i0} + 1$ y $\tau_i - 2$, para todo i, con $1 \leq i \leq n$.*

Generación de los tiempos hasta los impagos

El siguiente algoritmo se utiliza para generar aleatoriamente los tiempos en los que ocurren los sucesivos impagos de los créditos. En él se ha procurado que los tiempos, T_{ij} y C_{ij}, estén correlacionados con la puntuación crediticia, X. Para todo $i \in \Pi$, con $1 \leq i \leq n$, ejecutar los siguientes pasos:

Paso 1: Sin pérdida de generalidad, considerar que $T_{i0} = 0$, para todo i, y fijar el límite máximo de tiempo hasta la mora, τ_i, y el máximo valor de puntuación crediticia muestral, x_n. De la Tabla 5.1 se extraen los valores $\tau_i = 73$ y $x_n = 44.8$.

Paso 2: Generar el tiempo hasta el primer impago como $T_{i1}|X = x_i \overset{d}{\equiv} U(1, a + bx_i)$, donde $U(\alpha, \beta)$ es una variable aleatoria uniforme en el intervalo (α, β) y x_i es el valor de la puntuación crediticia, X, observado en el crédito i-ésimo. Simultáneamente, generar la variable censurante asociada, C_{i1}, como $C_{i1}|X = x_i \overset{d}{\equiv} U(1, c + dx_i)$.

Paso 3: Para $j = 1$ obtener la indicadora $\delta_{ij} = I(T_{ij} \leq C_{ij})$ y el tiempo observado hasta el primer impago, $\xi_{ij} = \min\{T_{ij}, C_{ij}\}$.

Paso 4: Generar el tiempo condicionado hasta el segundo impago como $T_{i2}|T_{i1} = t_1, X = x_i \overset{d}{\equiv} U(t_1 + 1, a + bx_i)$ y la variable censurante, C_{i2}, como $C_{i2}|C_{i1} = t_1, X = x_i \overset{d}{\equiv} U(t_1 + 1, c + dx_i)$.

Paso 5: Volver a realizar el paso 3 para $j = 2$.

Paso 6: Generar el tiempo condicionado hasta el tercer impago como $T_{i3}|T_{i2} = t_2, X = x_i \overset{d}{\equiv} U(t_2 + 1, a + bx_i)$ y la variable censurante asociada, C_{i3}, como $C_{i3}|C_{i2} = t_2, X = x_i \overset{d}{\equiv} U(t_2 + 1, c + dx_i)$.

Paso 7: Volver a realizar el paso 3 para $j = 3$.

Observación 5.2 *En este estudio se utilizaron los valores de los parámetros* $a = c = 2$, $b = (\tau_i - 4)/x_n$ *y* $d = (\tau_i - 2)/100$, *donde* x_n *es el máximo valor de la muestra de los* x_i *y* τ_i *es el tiempo de vida real obervado para cada crédito* i, *con* $1 \leq i \leq n$. *Se observó que con esta forma de simular los* T_{ij} *y los* C_{ij} *se logran proporciones de censura (clientes no morosos) más altas a medida que aumenta el número de impagos en los clientes, es decir, que se logra que la proporción de clientes reincidentes con tres impagos sea menor que la*

proporción de clientes con dos impagos, y que a su vez ésta sea menor que la proporcion de clientes con sólo un impago. Para simular los tiempos entre reincidencias, se utilizan variables aleatorias con distribuciones uniformes, suponiendo así que estas pueden ocurrir en cualquier momento dentro del intervalo de tiempo disponible antes de la entrada en mora.

Análisis descriptivo de los tiempos generados

En la Figura 5.6 se ilustran los histogramas de los tiempos hasta los impagos, T_1, T_2 y T_3, y sus tiempos de censura asociados, C_1, C_2 y C_3. De la observación de los histogramas se deduce que, en general, todos los tiempos simulados poseen distribuciones muy asimétricas con colas pesadas a la derecha, aunque en el caso de T_3 y C_3 esta característica parece más pronunciada que en las demás variables, tal y como se ve más adelante en la Figura 5.7.

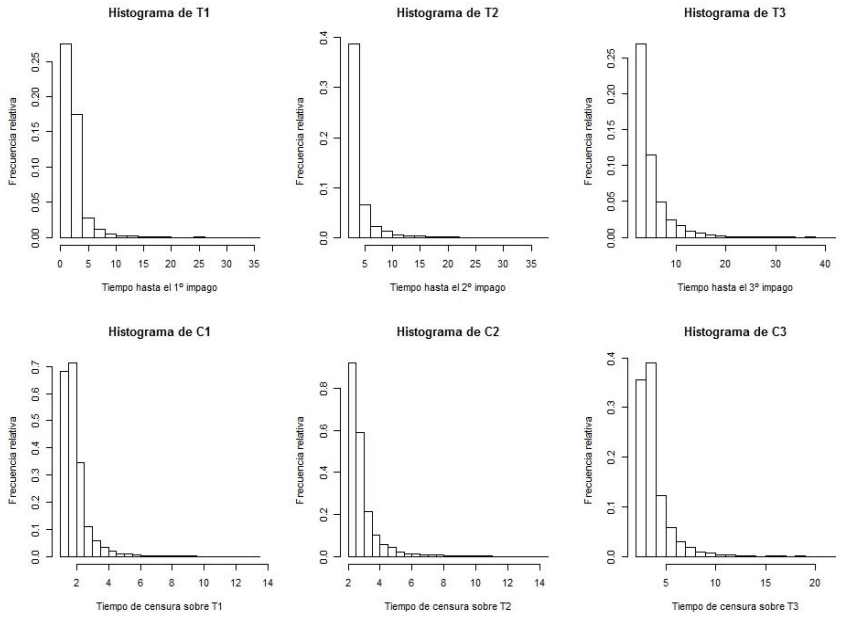

Figura 5.6. Histogramas de los tiempos de reincidencia de los impagos y de sus tiempos de censura (j=3).

Tabla 5.3. Estadísticos descriptivos de los tiempos hasta el j-ésimo.

Tiempo condicionado	min	Q_1	Q_2	$media$	Q_3	max
T_1	1.001	1.449	1.904	2.417	2.599	34.748
T_2	2.003	2.414	2.834	3.613	3.804	37.688
T_3	3.013	3.191	3.832	5.135	5.758	41.716
C_1	1.000	1.369	1.717	1.891	2.093	13.425
C_2	2.003	2.129	2.366	2.728	2.872	14.407
C_3	2.029	2.843	3.205	3.706	4.013	21.295

En la Tabla 5.3 se muestran los estadísticos descriptivos usuales para los tiempos hasta el j-ésimo impago y para los tiempos de censura asociados. Se observa que los tiempos C_1, C_2 y C_3, tienen llegan a valores máximos meores que $T1$, $T2$ y $T3$, respectivamente a concentrarse en valores inferiores a los 5 meses de vida a partir de su formalización. Además, bajo este mecanismo de simulación, todos los j-ésimos impagos hasta $j = 3$ reincidencias ocurren antes de llegar a los 20 meses desde la formalización de los créditos.

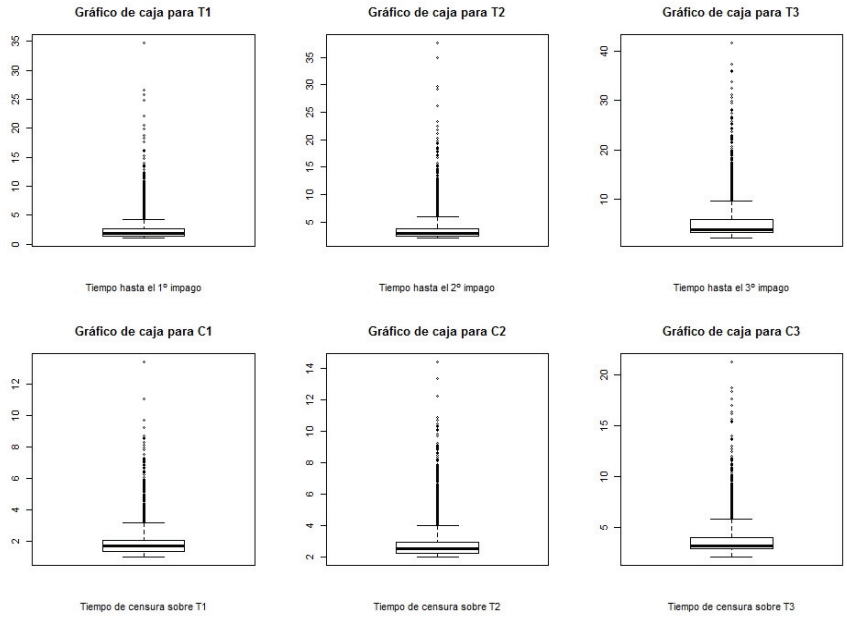

Figura 5.7. Gráficos de cajas de los tiempos de reincidencia de los impagos y de sus tiempos de censura (j=3).

En la Figura 5.7 se muestran seis gráficos de cajas en los que se ofrece una visión simultánea del grado de variabilidad y de asimetría positiva observada en los tiempos hasta la reincidencia, T_1, T_2, T_3 y en sus tiempos de censura asociados. Se confirma lo observado en la Tabla 5.3 y en la Figura 5.6. También se obseva una alta presencia de datos atípicos, lo que provoca el significativo peso de la masa de probabilidad concentrada a la derecha de la distribución de las variables.

Estimación de la función de supervivencia condicional del tiempo hasta la reincidencia

En este apartado se exponen los resultados de la estimación no paramétrica de la función de supervivencia condicional del tiempo hasta el j-ésimo impago, dados los $j-1$ impagos anteriores y dada la covariable $X = x$, obtenidos con el estimador PLG de Akritas y Van Keilegom (2003). Como se ha expuesto anteriormente, las estimaciones se obtienen a partir de una mezcla de datos reales, como es el tiempo hasta la entrada en mora, τ, y de datos simulados, como son los tiempos de reincidencia de los impagos ocurridos antes de τ.

Como las curvas de probabilidades de supervivencia que pueden obtenerse en función de los valores que pueden tomar las variables condicionantes son múltiples, en este caso se optó por calcular distintos escenarios, representados por $\mathcal{F}_{J-j}$, para $j = 1, 2, 3$, considerando como número máximo de impagos, $J = 4$, donde el tiempo hasta el cuarto impago, T_{i4}, coincide precisamente con el tiempo hasta la entrada en mora según Basilea II, τ_i, es decir, la mora real observada en el crédito i-esimo. Así, los valores que se muestran a continuación en las tablas siguientes corresponden a la media aritmética de las estimaciones puntuales obtenidas con el estimador de la función de supervivencia condicional, $\hat{S}^{AVK}_{j|\mathcal{F}_{j-1}}(t|x)$, sobre una secuencia creciente ordenada de $T_j = t$, con $j = 1, 2, 3$, dados los valores de la covariable, $X = 2.24$, 4.48, 7.12 y 9.85. Los resultados se obtuvieron a partir de una muestra aleatoria de $m = 5\,000$ créditos tomada de la base de datos original ($n = 89\,646$).

Por último, la curvas estimadas de supervivencia condicional y de probabilidad de reincidencia condicional dados $\mathcal{F}_2 = \{T_1, T_2\}$ y $\mathcal{F}_3 = \{T_1, T_2, T_3\}$, que se exponen más adelante, respectivamente, se obtuvieron con cuatro selectores automáticos multivariantes, tres de los cuales están implementados

en la librería *ks* del paquete estadítico *R* (Duong (2007, 2015)). Los selectores considerados en este capítulo vienen adaptados en las funciones, Hpi, basada en el método *plug-in* de Wand y Jones (1994), $Hscv$, basada en el método de *validación cruzada suavizada* (scv) de Duong y Hazelton (2005) y $Hpi.kcde$, basada en el método propuesto por Duong (2014). También se utilizó el selector basado en la regla de Scott (1992), que como se verá más adelante, arrojó algunos de los mejores resultados. En el caso univariante, para las variables tiempo hasta la entrada en mora (Basilea II), $T_1 = \tau$, y puntuación crediticia, X, se utilizaron selectores análogos a los anteriores (regla de Scott, *plug-in* de Wand y Jones (1995), dpi, y *validación cruzada* de Rudemo (1982) y Bowman (1984), $lscv$).

A continuación, en la Tabla 5.4, se muestran los valores de los parámetros de suavizado utilizados en las estimaciones no paramétricas con $j = 1$ impago previo a la entrada en mora, $T = \tau$.

Tabla 5.4. Selectores del parámetro de suavizado en el caso univariante.

Selector de h	T_1	X
Regla de $Scott$	0.3142	0.8114
$lscv$	0.8043	0.1313
dpi	0.05107	0.0759

Los paramámetros de suavizado obtenidos para el caso de $j = 2$ impagos:

$$\boldsymbol{H}_{2,n}^{Scott} = \begin{bmatrix} 0.46086 & 0.48381 \\ 0.48381 & 0.56326 \end{bmatrix},$$

$$\boldsymbol{H}_{2,n}^{scv} = \begin{bmatrix} 0.025877 & 0.02258 \\ 0.02258 & 0.02569 \end{bmatrix},$$

$$\boldsymbol{H}_{2,n}^{dpi} = \begin{bmatrix} 0.02881 & 0.02766 \\ 0.02766 & 0.03215 \end{bmatrix}.$$

Los paramámetros de suavizado obtenidos para el caso de $j = 3$ impagos son:

$$\begin{aligned}
\boldsymbol{H}_{3,n}^{Scott} &= \begin{bmatrix} 0.56447 & 0.59258 & 1.65148 \\ 0.59258 & 0.68989 & 0.77117 \\ 1.65148 & 0.77117 & 1.03225 \end{bmatrix}, \\
\boldsymbol{H}_{3,n}^{scv} &= \begin{bmatrix} 6.08011 & 7.72332 & 10.4484 \\ 7.72332 & 9.82616 & 13.30568 \\ 10.4484 & 13.30568 & 18.05227 \end{bmatrix}, \\
\boldsymbol{H}_{3,n}^{dpi} &= \begin{bmatrix} 0.01623 & 0.01443 & 0.01577 \\ 0.01443 & 0.02278 & 0.02399 \\ 0.01577 & 0.02399 & 0.05641 \end{bmatrix}.
\end{aligned}$$

Después de realizar múltiples pruebas con los parámetros de suavizado obtenidos con los tres métodos de selección, se llegó a la conclusión de que las estimaciones de las probabilidades de supervivencia resultaron mejores cuando se utilizaron valores del parámetro de suavizado más grandes que los obtenidos con los tres selectores automáticos propuestos. En particular, los resultados fueron mejores (visualmente) cuando se utilizaron las ventanas de suavizado proporcionales a las obtenidas con el método de Scott (1992). Por ello, el procedimiento utilizado para generar las curvas de probabilidad rvas que se presentan a continuación se basa en un criterio de corte más más empírico, en el que se fue ajustando el parámetro de ventana hasta conseguir resultados acordes con la teoría, es decir, que se varió el valor del parámetro de suavizado hasta conseguir que las curvas de probabilidad de superviviencia estimadas tuviesen un aspecto similar al de las curvas de superviviencia teóricas.

A continuación, en las Figuras 5.8 a la 5.13, se ilustran algunos ejemplos de las funciones de supervivencia condicionales obtenidas para algunos valores de la covaribale X.

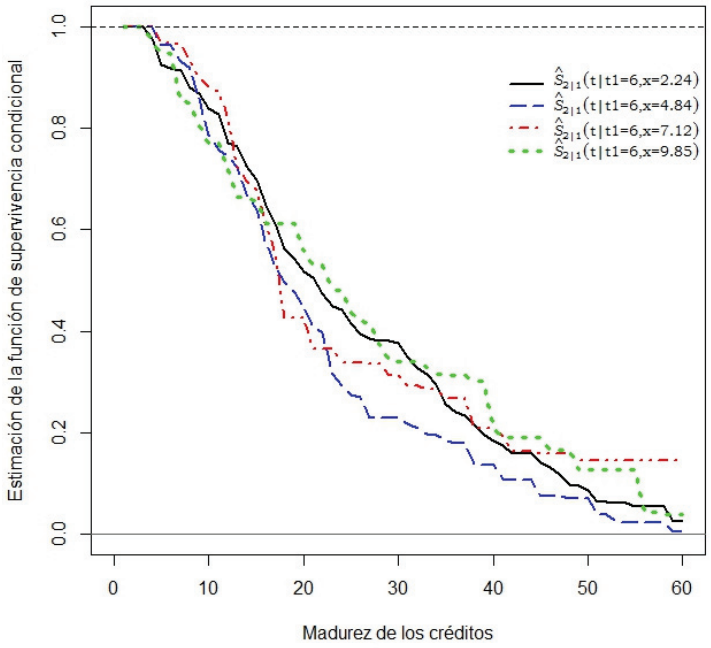

Figura 5.8. Estimaciones de las funciones de supervivencia condicional de T_2 dada T_1=6 y dada la X=2.24, 4.84, 7.12 y 9.85.

En la Figura 5.8 se ilustran las estimaciones de las funciones de supervivencia condicional para el tiempo hasta el segundo impago (primera reincidencia), T_2, dado que el primero ocurre en $T_1 = 6$, y dados los valores de puntuación crediticia, $X = 2.24$, 4.48, 7.12 y 9.85. En el cálculo de esta curvas se utlizaron parámetros de suavizado, $h_X^{Scott} = 0.8114$, para la variable, X, y $h = 10h_n^{Scott} = 3.47$ para la variable T_1.

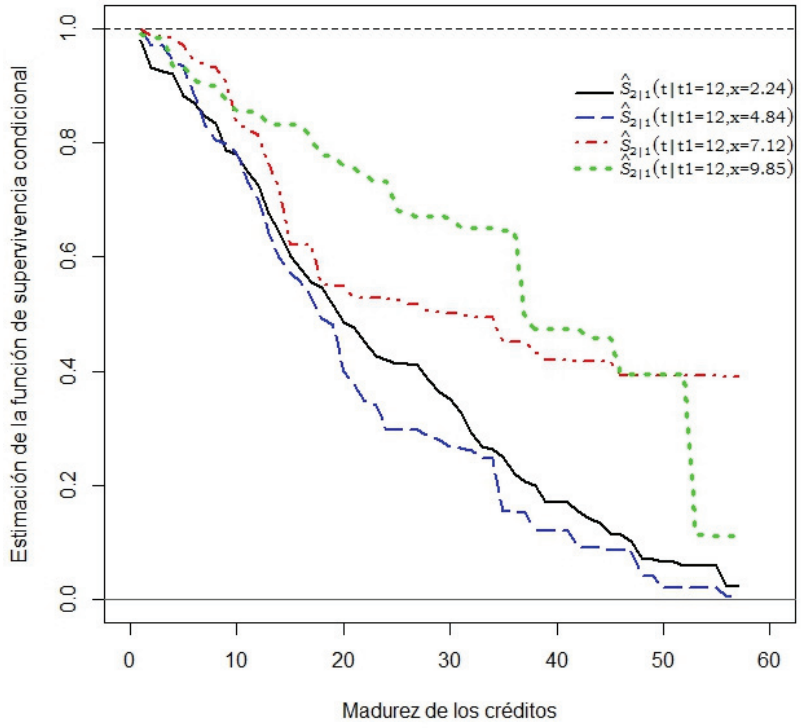

Figura 5.9. Estimaciones de las funciones de supervivencia condicional de T_2 dada T_1=12 y dada la X=2.24, 4.84, 7.12 y 9.85.

En la Figura 5.9 ilustran las estimaciones de las funciones de supervivencia condicionales para el tiempo hasta el segundo impago, T_2, dado que el primero ocurre en $T_1 = 12$, y dados los valores de puntuación crediticia, $X = 2.24$, 4.48, 7.12 y 9.85.

Se observa que las estimaciones obtenidas con parámetros de suavizado $h_X^{Scott} = 0.8114$, para la variable, X, y $h = 8h_n^{Scott} = 2{,}174$ para la variable T_1. En ambos casos se observa un grado de suavizado aceptable y que las nivel de probabilidad alcanzado es coherente con la curva de supervivencia teórica, puesto que las cuatro curvas deberían valer uno (aproximadamente) hasta $T_2 = 12$ y luego comenzar a decreder hasta valores cercanos a cero, tal y como está ocurre en ambas gráficas (Figrura 5.8 y Figura 5.9).

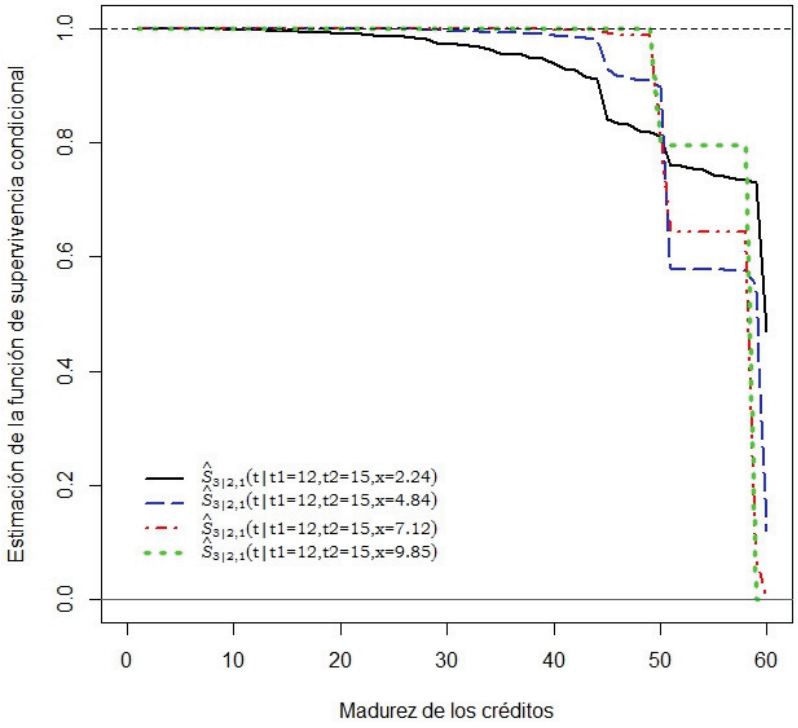

Figura 5.10. Estimaciones de las funciones de supervivencia condicional de T_3 dada T_2=15 y T_1=12 y dada la X=2.24, 4.84, 7.12 y 9.85.

En la Figura 5.10 se ilustran cuatro estimaciones de las funciones de supervivencia condicionadas para el tiempo hasta el tercer impago (segunda reincidencia), T_3, dado que el primer impago ocurre en $T_1 = 12$, el segundo impago ocurre en $T_2 = 15$, y dados los valores de puntuación crediticia, $X = 2.24$, 4.48, 7.12 y 9.85. Las curvas allí representadas se obtuvieron con parámetros de suavizado calculados con la regla de Scott (1992), $h_X^{Scott} = 0.8114$, para la variable, X, y $H = 4.5\boldsymbol{H}_{2,n}^{Scott}$ para el vector (T_1, T_2).

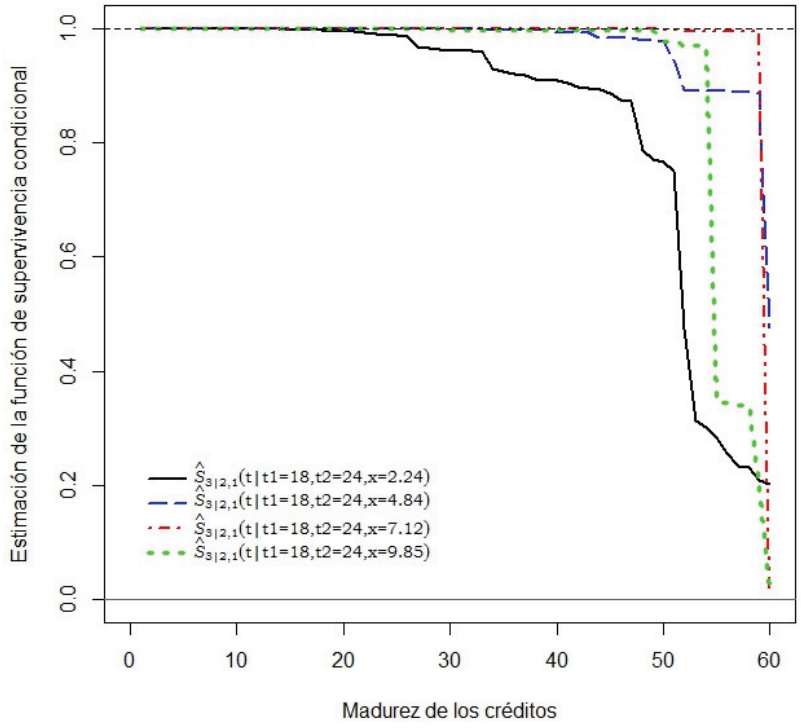

Figura 5.11. Estimaciones de las funciones de supervivencia condicional de T_3 dada T_2=24 y T_1=18 y dada la X=2.24, 4.84, 7.12 y 9.85.

En la Figura 5.11 se ilustran las estimaciones de las funciones de supervivencia condicionadas para el tiempo hasta el tercer impago (o segunda reincidencia), T_3, dado que el segundo impago ocurre en en $T_2 = 24$ y el primero en $T_1 = 18$, y dados los valores de puntuación crediticia, $X = 2.24$, 4.48, 7.12 y 9.85.

Tal y como se vio para las curvas estimadas representadas en la Figuras 8 y 9, en las Figuras 10 y 11 también se observan resultados razonables, que van de acuerdo con la teoría. Las curvas que aparecen allí representadas se obtuvieron con parámetros de suavizado calculados con la regla de Scott (1992), $h_X^{Scott} = 0.8114$, para la variable, X, y $H = 8\boldsymbol{H}_{2,n}^{Scott}$ para el vector (T_1, T_2).

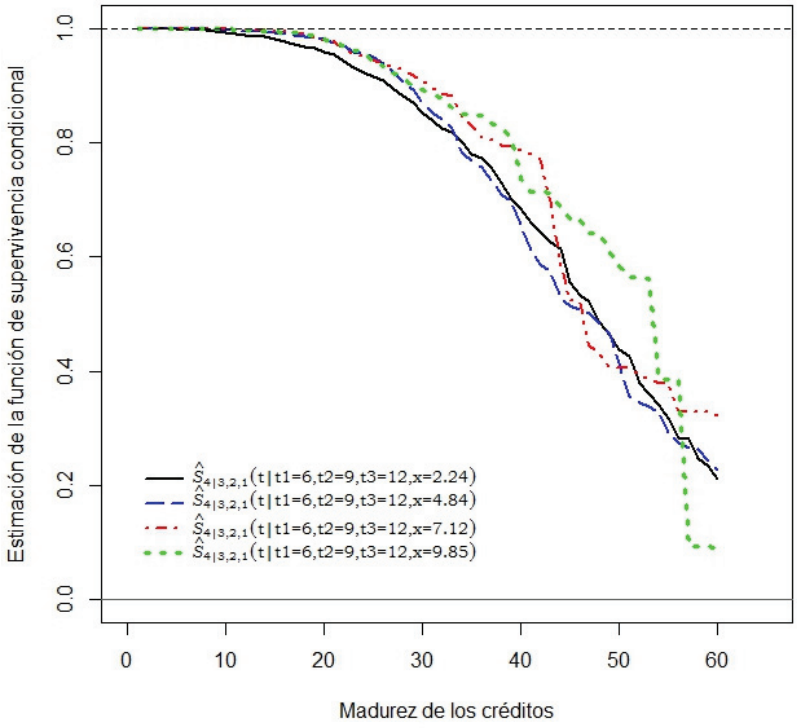

Figura 5.12. Estimaciones de las funciones de supervivencia condicional de T_4 dados T_3=12, T_2=9, T_1=6 y dada X=2.24, 4.84, 7.12 y 9.85.

En la Figura 5.12 se ilustran las estimaciones de las funciones de supervivencia condicionadas para el tiempo hasta el cuarto impago (o tercera reincidencia), T_4, dado que el tercer impago ocurre en $T_3 = 12$, el segundo en $T_2 = 9$ y el primero en $T_1 = 6$, y dados los valores de puntuación crediticia, $X = 2.24$, 4.48, 7.12 y 9.85. Las curvas representadas se obtuvieron con parámetros de suavizado calculados con la regla de Scott (1992), $h_X^{Scott} = 0.8114$, para la variable, X, y $H = 4{,}5\boldsymbol{H}_{3,n}^{Scott}$ para el vector (T_1, T_2, T_3).

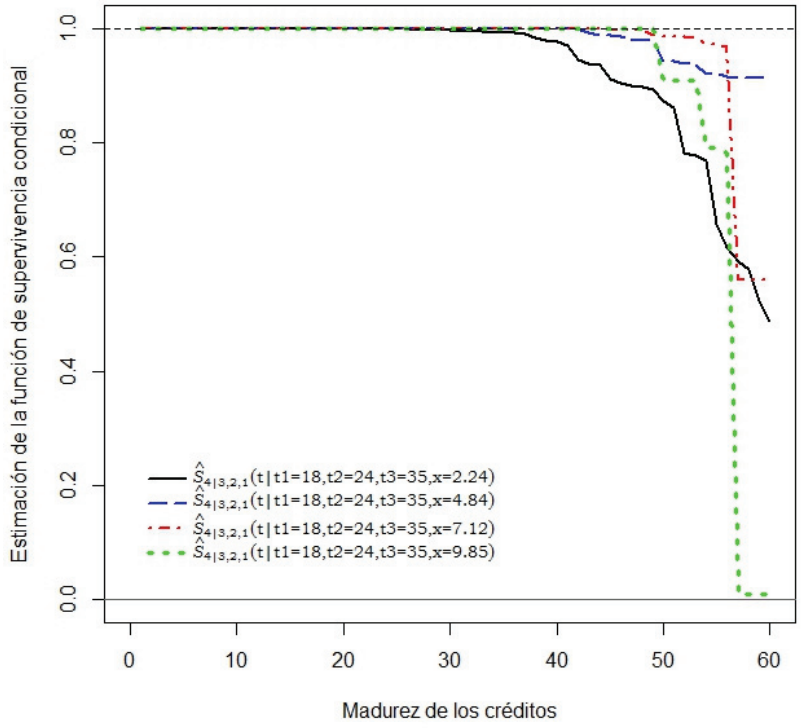

Figura 5.13. Estimaciones de las funciones de supervivencia condicional de T_4 dados T_3=35, T_2=24, T_1=18 y dada X=2.24, 4.84, 7.12 y 9.85.

En la Figura 5.13 se ilustran las estimaciones de las funciones de supervivencia condicionadas para el tiempo hasta el cuarto impago (o tercera reincidencia), T_4, dado que el tercer impago ocurre en $T_3 = 35$, el segundo en $T_2 = 24$ y el primero en $T_1 = 18$, y dados los valores de puntuación crediticia, $X = 2.24$, 4.48, 7.12 y 9.85. Las curvas representadas se obtuvieron con parámetros de suavizado calculados con la regla de Scott (1992), $h_X^{Scott} = 0.8114$, para la variable, X, y $H_n = 8\boldsymbol{H}_{3,n}^{Scott}$ para el vector (T_1, T_2, T_3).

Resultados de la estimación no paramétrica de las curvas de probabilidad condicional de reincidencia

En las siguientes figuras se representan gráficos con las curvas de probabilidades de reincidencia condicional obtenidas con el estimador PLG, $\hat{\varphi}^{AVK}_{J|\mathcal{F}_{J-1}}$. Análogamente a los resultados obtenidos para la funciones de supervivencia condicional estimadas, las curvas de probabilidad condicional de reincidencia de los impagos se ilustran condicionando sobre cuatro valores representativos de la covariable, X, como son los valores $Q_1 = 2.24$, $Q_2 = 4.48$, la $\bar{X} = 7.12$, y $Q_3 = 9.85$, además de condicionar sobre los valores de los $j-1$ tiempos anteriores al j-ésimo impago, con $j = 1, 2, 3$.

Estimación de la probabilidad condicional de reincidencia

En este apartado se exponen los resultados de las estimaciones de las probabilidades condicionales de reincidencia de los impagos. Análogamente a lo visto para la función de supervivencia condicional, a continuación se muestran gráficos en los que se ilustran algunos de los resultados obtenidos en una serie de pruebas realizadas bajo distintos escenarios (condiciones similares a las utilizadas para el cálculo del estimador de la supervivencia condicional, $\hat{S}^{AVK}_{J|\mathcal{F}_{J-1}}$). Prácticamente en el 100 % de la pruebas realizadas se observó que las estimaciones de las probabilidades condicionales de reincidencia obtenidas arrojaron mejores resultados utilizando valores grandes del parámetro de suavizado. Así, análogamente al caso de la función de supervivencia estimada, en este apartado se fueron probando distintos valores del parámetro de suavizado partiendo por la ventana (o matriz de ventanas) calculada con el método de Scott (1992). Como resultado se obtuvieron curvas de probabilidad de reincidencia de los impagos menos variables y más acordes con la teoría cuando se utilizaron parámetros de ventana que fueron desde $4\boldsymbol{H}_n^{Scott}$ hasta $10\boldsymbol{H}_n^{Scott}$.

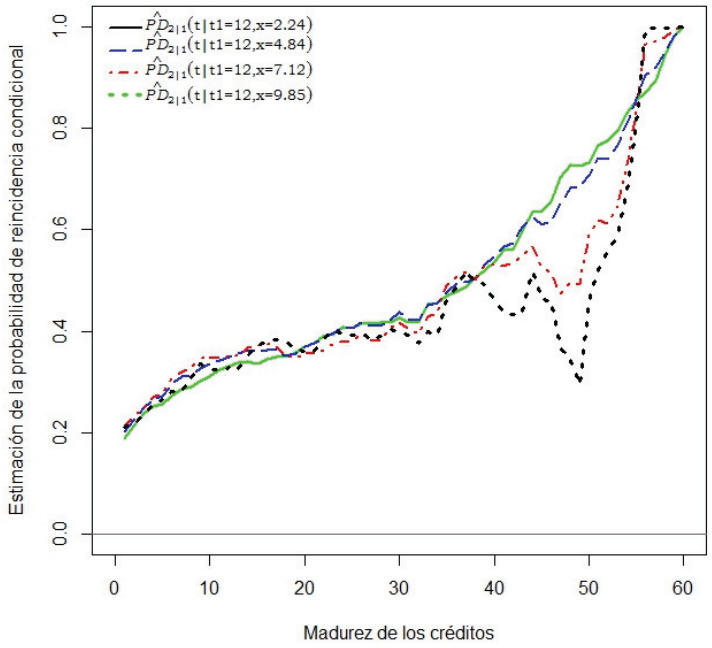

Figura 5.14. Estimaciones de las funciones de supervivencia condicional de T_2 dada T_1=12 y dada la X=2.24, 4.84, 7.12 y 9.85.

En la Figura 5.14 se representan las estimaciones de las funciones de probabilidad condicional de reincidencia del segundo impago, T_2, dado el primer impago en $T_1 = 12$, y dados los valores de puntuación crediticia, $X = 2.24$, 4.48, 7.12 y 9.85. La estimación de las curvas de probabilidad de reincidencia se obtuvo fijando el horizonte de predicción en $b = 12$ meses con parámetros de suavizado calculados con la regla de Scott (1992), $h_X^{Scott} = 0.8114$, para la variable X, y $h = 8h_n^{Scott} = 2.7756$ para la variable T_1. En el caso de los vectores de variables condicionantes (T_1, T_2) y (T_1, T_2, T_3), los mejores resultados se obtuvieron con el parámetro de suavizado vía método de Scott (1992), $\boldsymbol{H}_{3,n}^{Scott}$. Como resultado, se observa que las curvas obtenidas con el estimador de Akritas-Van Keilegom, $\hat{\varphi}_{2|1}^{AVK}(t+12|x)$, exhiben un grado de suavizamiento aceptable, comparable con los resultados obtenidos con el estimador de la PD condicional, $\hat{\varphi}_n^{PLG}(t|x)$, calculado a un año vista ($b = 12$) en el Capítulo 3 de esta memoria (ver Sección 3.3.2, Figuras 3.9 a la 3.11).

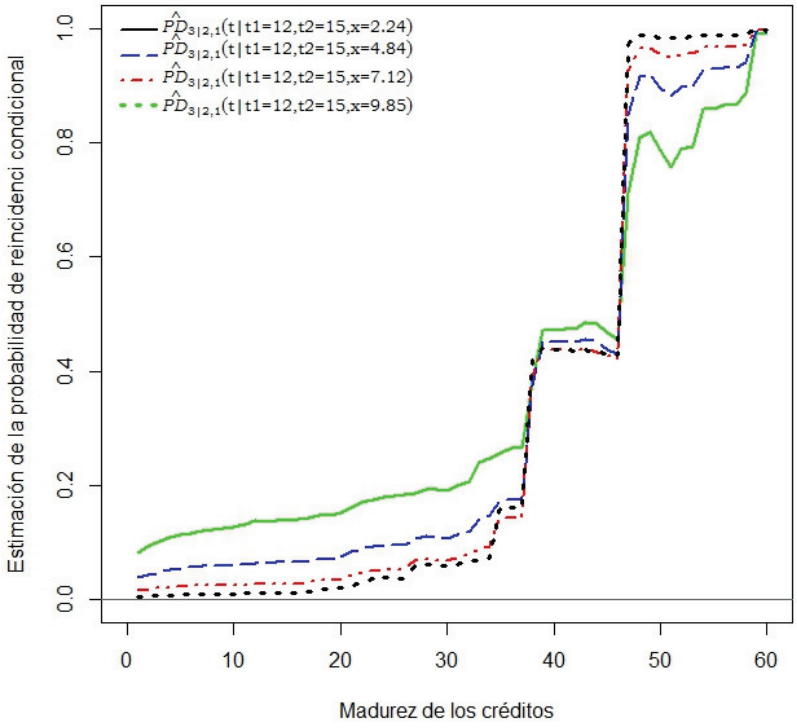

Figura 5.15. Estimaciones de las funciones de supervivencia condicional de T_3 dada T_2=15 y T_1=12 y dada la X=2.24, 4.84, 7.12 y 9.85.

En la Figura 5.15 se muestran las estimaciones de la probabilidad de reincidencia asociadas al impago, T_2, dado que el primer impago ocurre en $T_1 = 12$, el segundo impago ocurre en $T_2 = 15$, y dados los valores de puntuación crediticia, $X = 2.24$, 4.48, 7.12 y 9.85. Se observa que con una ventana 10 veces más grande que la matriz de ventanas de suavizado, $\boldsymbol{H}_{2,n}^{Scott}$, los resultados obtenidos parecen ser satisfactorios ya que al aumentar el tamaño de la ventana, $\boldsymbol{H}_{2,n}^{Scott}$, en un factor de entre 4 y 10 veces más grandes, se obtienen como resultado curvas menos variables y más acordes con las curvas teóricas, es decir, alcanzando el valor 1 para aquellos clientes con varios episodios de impagos ocurridos en el pasado.

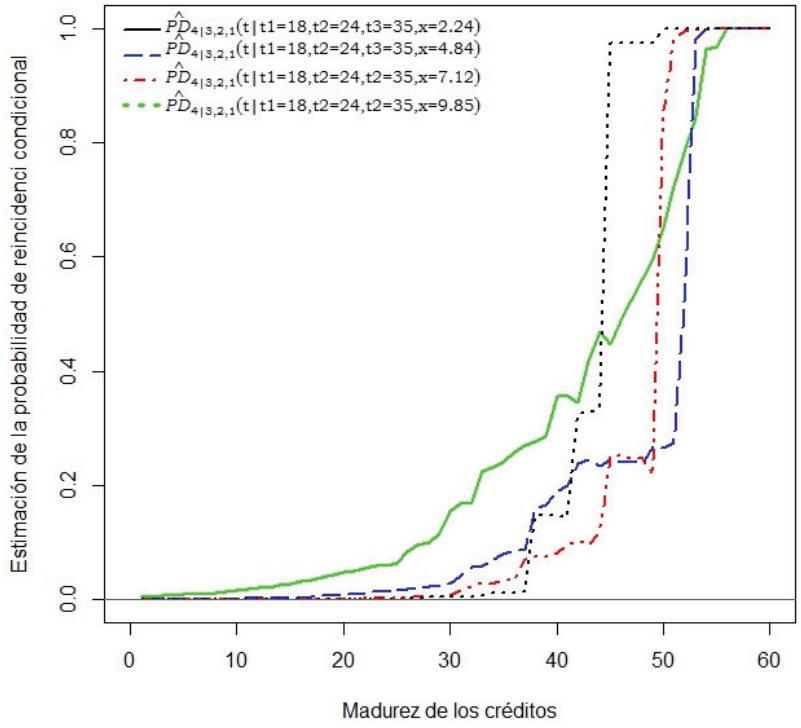

Figura 5.16. Estimaciones de las probabilidades de reincidencia condicional de T_4 dados T_3=35, T_2=24, T_1=18 y dada X=2.24, 4.84, 7.12 y 9.85.

En la Figura 5.16 se observan resultados similares a los obtenidos en las Figuras 5.14. y 5.15, es decir, mejores resultados con ventanas más grandes que las obtenidas con el método de Scott (1992). En este caso, se ha multiplicando por 2 el valor del parámetro de suavizado asociado a la covariable, X, y se ha multiplicado por 10 el parámetro de suavizado, $\boldsymbol{H}_{3,n}^{Scott}$, asociado a los tiempos (T_1, T_2, T_3). Estos resultados sugieren que un valor adecuado del parámetro de suavizado asociado a la covariable X, h_X, para el cálculo de las curvas de probabilidad condicional del segundo impago (primera reincidencia), T_2, dado el primer impago en $T_1 = 12$ (doce meses antes) y dados los valores de puntuación crediticia, $X = 2.24$, 4.48, 7.12 y 9.85, se encuentra en en el rango $2h_X^{Scott} \leq h_X \leq 3h_X^{Scott}$, con $h_X^{Scott} = 0.8114$.

5.5. Comentarios y conclusiones

En este capítulo se aborda el problema de la *reincidencia de la morosidad crediticia* en una cartera de tarjeas de crédito, donde el concepto de *reincidencia* hace referencia al hecho de que un acreditado incurra en sucesos de incumplimiento de su deuda de forma recurrente en el tiempo. En el caso de las tarjetas de crédito, es usual obsevar situaciones de incumplimiento provocadas, por ejemplo, por el decubierto en la cuenta bancaria vinculada al pago de la misma.

Debido a la posible correlación existente entre los impagos sucesivos, y al hecho de que algunos de estos sucesos no son completamente observables debido a la censura presente en los datos, se adoptó un enfoque de modelización basado en el análisis de sucesos recurrentes con datos censurados bajo dependencia. Como herramienta predictiva de los impagos recurrentes, se utilizó una fórmula para la probabilidad condicionada de la j-ésima reincidencia obtenida a partir de la función de supervivencia condicional del tiempo hasta el j-ésimo impago, $S_{j|\mathcal{F}_{j-1}}(t|x)$. Esta fórmula se obtiene de forma análoga a la fórmula definida en (3.1) para la PD condicionada del Capítulo 3.

Debido a que la reincidencia de la morosidad crediticia ha recibido escaso interés en la literatura sobre riesgo de crédito, se considera que el tratamiento estadístico que ha dado a este problema en este capítulo contribuye a su divulgación y puede aplicarse al estudio de modelos de scoring crediticio de conducta (Thomas (2000), Thomas et al. (2002), Chen et al. (2009)), donde el análisis de riesgo se centra en la detección y prevención de aquellos acreditados con más propensión a convertirse en malos pagadores, por ejemplo, como los clientes reincidentes durante el tiempo de vida del acreditado.

Por otra parte, con respecto a los resultados de las estimaciones obtenidas, se observó que las curvas de supervivencia y de probabilidad de reincidencia, resultaron mejores, o más razonables en el sentido de ser más parecidas a las curvas esperadas o teóricas, cuando se utilizaron valores del parámetro de suavizado significativamente más grandes que los obtenidos con algunos de los métodos implementados en la librería *ks* de *R* (Duong (2007, 2015)). En efecto, en numerosas pruebas realizadas con submuestras de distintos tamaños, se observó que, tanto el grado de suavidad de las curvas estimadas como el grado de sensibilidad del estimador frente a los distintos valores que toman las variables del espacio condicionante, comenzaban a mejorar a par-

tir de valores significativamente más grandes de las ventanas de suavizado obtenidas, por ejemplo, con los métodos *plug-in* de Wand y Jones (1995) y Duong (2014). Estos resultados pueden deberse a varios motivos, por ejemplo, al conocido efecto de la maldición de la dimensionalidad, padeciendo los efectos de utilizar hasta cuatro variables condicionantes en el cálculo de las probabilidades de supervivencia. También, como se ha mencionado antes, la falta de suavidad de las curvas puede deberse a la elección de un selector del parámetro de suavizado inadecuado para trabajar con funciones de distribución con más de dos variables condicionantes. En este sentido, resulta natural proponer como objetivo de investigación futura la búsqueda de un mecanismo de selección automática del (los) parámetro(s) de suavizado óptimos para los estimadores condicionales basados en el estimador de Akritas y Van Keilegom (2003) estudiados en este capítulo. En ese sentido, sería interesante adaptar el algoritmo bootstrap propuesto en el Capítulo 3 para el estimador de la PD condicional al caso de la probabilidad condicional de reincidencia, y comparar esos resultados con otros selectores automáticos, por ejemplo, como los propuestos por Duong y Hazelton (2005) y por Chacón y Duong (2010), entre otros. Con respecto a lo anterior, es importante volver a explicar que, debido a la imposibilidad de contar con datos reales completos sobre los incumplimientos anteriores a la entrada en mora (los sucesos recurrentes), se vio que los resultados obtenidos dependen fuertemente tanto del mecanismo utilizado para simular los datos faltantes, como de las hipótesis consideradas en su implementación. Por este motivo, es importante hacer hincapié en que este trabajo responde principalmente a un estudio de carácter exploratorio, donde las conclusiones obtenidas tienen una validez comparable a las extraídas en un estudio de simulación. Por tanto, sería recomendable elaborar un segundo estudio en el que éstas puedan ser contrastadas frente a los resultados observados en una base de datos con información completa sobre los sucesos de incumplimientos.

Finalmente, es importante explicar aquí que, a partir de las técnicas utilizadas en este capítulo para el tratamiento estadístico del problema de *la reincidencia de la morosidad crediticia*, en el camino surgieron problemas derivados del propio enfoque metodológico seleccionado y que, principalmente por motivos de tiempo y de extensión de la propia memoria, se dejaron para un trabajo posterior. En efecto, uno de los problemas que se propone estudiar a partir de los resultados obtenidos en este capítulo es un estudio comparativo del estimador de la probabilidad condicional de reincidencia

basado en el estimador de Akritas-Van Keilegom (2003) versus estimadores análogos construidos a partir de otros estimadores de la función de supervivencia condicional bajo el enfoque de análisis de sucesos recurrentes con datos censurados bajo dependencia, por ejemplo, como los debidos a Visser (1996), Lin et al. (1999) y Peña et al. (2001). También sería interesante comparar los resultados obtenidos bajo el enfoque no paramétrico versus modelos paramétricos y semiparamétricos como los basados en el enfoque de regresión de Cox y en los modelos de vida acelerada.

Otra extensión natural derivada del problema de la estimación no paramétrica de la función de supervivecnia condicional, y de la función de probabilidad condicional de reincidencia obtenida a partir de la primera, es la relacionada con el estudio asintótico de estos dos estimadores. Propiedades como la convergencia del estimador, la consistencia uniforme débil y fuerte, la normalidad asintótica y la determinación del parámetro de suavizado óptimo vía la minimización del error cuadrático medio asintótico ($ECMA$), o del error cuadrático medio integrado asintótico ($ECMIA$), son algunos de los resultados que tradicionalmente se utlizan para medir y comparar la eficiencia de los estimadores en estadística no paramétrica. Para obtener algunas de estas propiedades, pueden adoptarse, por ejemplo, las ideas de Cao et al. (2009) desarrolladas íntegramente en el Capítulo 4 de esta memoria para el estimador PLG de la PD basado en el estimador de Beran (1981), $\hat{\varphi}_n^{PLG}$, así como las técnicas estudiadas, por ejemplo, por Dabrowska (1989), Iglesias Pérez y González Manteiga (1999), Du y Akritas (2002) y Akritas y Van Keilegom (2003), entre otros, quienes han obtenido propiedades análogas a las mencionadas y que pueden servir, sin duda, de fuente de inspiración, tanto para el autor de esta memoria, como para otros investigadores interesados en el estudio de técnicas de estimación no paramétrica de curvas como una poderosa herramienta de análisis estadístico del riesgo de crédito desde la perspectiva del análisis de supervivencia.

Bibliografía

[1] Aalen, O. O. (1978). Nonparametric inference for a family of counting processes, *The Annals of Statistics*, **6**, 701–726.

[2] Aalen, O. O. y E. Husebye (1991). Statistical analysis of repeated events forming renewal processes. *Statistics in Medicine*, **10**, 1227–1240.

[3] Agresti, A. (2002). *Categorical Data Analysis*, 2nd Ed. John Wiley & Sons, New York.

[4] Aitkin, M., B. Francis y J. Hinde (2005). *Statistical Modelling in GLIM 4*, 2nd Ed. Oxford University Press, New York.

[5] Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**, 716–723.

[6] Akritas, M. G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring, *The Annals of Statistics*, **22**, 1299–1327.

[7] Akritas, M. G. e I. Van Keilegom (2003). Estimation of bivariate and marginal distributions with censored data, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**, 457–471.

[8] Alexakis, P. (2008). Altman Z-score model and prediction of bussines failures, *International Journal of Monetary Economics and Finance*, **1,** 329–337.

[9] Allen, L. N. y L. C. Rose (2006). Financial survival analysis of defaulted debtors, *Journal of the Operational Research Society*, **57**, 630–636.

[10] Altman, E. I. (1968). Finantial ratios, discriminant analysis, and the prediction of corporate bankruptcy, *Journal of Finance*, **23**, 589–611.

[11] Altman, E. I., R. G. Haldeman y P. Narayanan (1977). $ZETA$ analysis: a new model to identify bankruptcy risk of corporations, *Journal of Banking & Finance*, **1**, 29–54.

[12] Altman, E. I. y A. Saunders (1998). Credit risk measurement: developments over the last 20 years, *Journal of Banking & Finance*, **25**, 1721–1742.

[13] Altman, E. I. (2000). Revisiting credit scoring models in a Basel 2 environment, Stern School of Business Working Paper No. S-CDM-02-06, New York University.

[14] Altman, E. I. (2002). Predicting financial distress of companies: revisiting the Z-score and $ZETA^{®}$ models, Stern School of Business Working Paper, New York University.

[15] Altman, E. I., M. De la Fuente, A. Elizondo, C. C. Finger, J. Gutiérrez, R. Gutiérrez, J. Márquez, J. Mina y M. Segoviano (2003). *Medición Integral del Riesgo de Crédito*. Limusa, México D. F.

[16] Altman, N. y C. Léger (1995). Bandwidth selection for kernel distribution function estimation, *Journal of Statistical Planning and Inference*, **46**, 195–214.

[17] Altman, N. y B. McGibbon (1998). Consistent bandwidth selection for kernel binary regression, *Journal of Statistical Planning and Inference*, **70**, 121–137.

[18] Amemiya, T. (1981). Qualitative response models: a survey, *Journal of Economic Literature*, **19**, 1483–1536.

[19] Apilado, V. P., D. C. Warner y J. J. Dauten (1974). Evaluative techniques in consumer finance, *Journal of Financial and Quantitative Analysis*, **March**, 275–283.

[20] Appel, M. J. (2002). Best estimation of a binary outcome in the sense of KS and $AROC$, *Communications in Statistics – Theory and Methods*, **31**, 1251–1257.

[21] Aragaki, A. y N. Altman. (1997). Local polynomial regression for binary response, Biometrics Unit, Cornell University, Technical Report No. BU-1397-M.

[22] Aranda-Ordaz, F. J. (1981). On two families of transformations to additivity for binary response data, *Biometrika*, **68**, 357–363.

[23] Artís, M, M. Guillén y J. M. Martínez (1994). A model for credit scoring: an application of discriminant analysis, *QUESTIIÓ*, **18**, 385–395.

[24] Baesens, B. (2014). *Analytics in a Big Data World: The esssential guide to Data Science and its applications,* John Wiley & Sons, New Jersey.

[25] Baesens, B., R. Setiono, C. Mues y J. Vanthienen (2003). Using neural network rule extraction and decision tables for credit-risk evaluation, *Management Science*, **49**, 312–329.

[26] Baesens, B., C. Mues, D. Martens y J. Vanthienen (2009). 50 years of data mining and OR: upcoming trends and challenges, *Journal of the Operational Research Society*, **60**, 16–23.

[27] Bamber, D. (1975). The area above the ordinal dominance graph and the area under the receiver operating characteristic graph, *Journal of Mathematical Psychology*, **12**, 387–415.

[28] Banasik, J., J. N. Crook y L. C. Thomas (1999). Not if but when will borrowers default, *The Journal of the Operational Research Society*, **50**, 1185–1190.

[29] Bandyopadhyay, A. (2006). Predicting probability of default of Indian corporate bonds: logistic and Z-score model approaches, *The Journal of Risk Finance*, **7**, 255–272.

[30] Bank for International Settlements (2009). *79th Annual Report: 1 April 2008 - 31 March 2009*. Bank for International Settlements, Basel.

[31] Basel Committee on Banking Supervision (1999). Credit risk modelling: current practices and applications. Bank for International Settlements, Basel.

[32] Basel Committee on Banking Supervision (2001a). The New Basel Capital Accord. Bank for International Settlements, Basel.

[33] Basel Committee on Banking Supervision (2001b). The internal ratings-based approach. Bank for International Settlements, Basel.

[34] Basel Committee on Banking Supervision (2004). International convergence of capital measurement and capital standards: a revised framework, Basel.

[35] Basel Committee on Banking Supervision (2005a). Studies on the validation of internal rating systems. Revised version. Bank for International Settlements, Working Paper No. 14, Basel.

[36] Basel Committee on Banking Supervision (2005b). An explanatory note on the Basel II IRB risk weight functions. Bank for International Settlements, Basel.

[37] Beaver, W. (1967). Financial ratios as predictors of failure, *Journal of Accounting Research*, **5**, 71–111.

[38] Begley, J., J. Ming y S. Watts (1996). Bankruptcy classification errors in the 1980s: an empirical analysis of Altman's and Ohlson's models, *Review of Accounting Studies*, **1**, 267–284.

[39] Beran, J. (2009). Discussion of modelling consumer credit risk via survival analysis, *Statistics and Operations Research Transactions*, **33**, 39–40.

[40] Beran, J. y A. K. Djaïdja (2007). Credit risk modeling based on survival analysis with immunes, *Statistical Methodology*, **4**, 251–276.

[41] Beran, R. (1981). Nonparametric regression with randomly censored survival data, Unpublished technical report, University of California, Berkeley.

[42] Bessis, J. (2002). *Risk Management in Banking*, John Wiley & Sons, London.

[43] Black, F. y M. Scholes (1973). The pricing of options and corporate liabilities, *Journal of Political Economy*, **81**, 637–654.

[44] Blochwitz, S., M. R. W. Martin y C. S. Wehn (2006). Statistical approaches to *PD* validation. En: Engelmann, B. y R. Rauhmeier (Editores). *The Basel II Risk Parameters: Estimation, Validation and Stress Testing*, Springer, 289–306.

[45] Bluhm, C., L. Overbeck y C. Wagner (2003). *An Introduction to Credit Risk Modeling*, Chapman & Hall/CRC, New York.

[46] Boj, E., M. M. Claramunt, A. Esteve y J. Fortiana (2009). Criterios de selección de modelo en el credit scoring. Aplicación del análisis discriminante basado en distancias, *Anales del Instituto de Actuarios Españoles*, **3**^{a} **época**, 209–230.

[47] Boj, E., M. M. Claramunt, y J. Fortiana (2007). Selection of predictors in distance-based regression. *Communications in Statistics–Simulation and Computation*, **36**, 87–98.

[48] Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates, *Biometrika*, **71**, 353–360.

[49] Bowman, A. W., P. Hall y T. Prvan (1998). Bandwidth selection for the smoothing of distribution functions, *Biometrika*, **85**, 799–808.

[50] Box, G.E.P. y D. R. Cox (1964). An analysis of transformations, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **26**, 211–252.

[51] Burnham, K. P. y D. R. Anderson (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*. Springer, New York.

[52] Cai, Z. (2003). Weighted local linear approach to censored nonparametric regression. En: Akritas, M. G. y D. N. Politis (Editores). *Recent Advances and Trends in Nonparametric Statistics*, Elsevier, 217–231.

[53] Calabrese, R. (2014). Optimal cut-off for rare events and unbalanced misclassification costs, *Journal of Applied Statistics*, **41**, 1678–1693.

[54] Cameron, A. C. y F. A. G. Windmeijer (1997). An *R*-squared measure of goodness of fit for some common nonlinear regression models, *Journal of Econometrics*, **77**, 329–342.

[55] Cao, R. (1993). Bootstrapping the mean integrated squared error, *Journal of Multivariate Analysis*, **45**, 137–160.

[56] Cao, R. y W. González Manteiga (1993). Bootstrap methods in regression smoothing, *Journal of Nonparametric Statistics*, **2**, 379–388.

[57] Cao, R., P. Janssen y N. Veraverbeke (2001). Relative denstity estimation and local bandwidth selection with censored data, *Computational Statistics & Data Analysis*, **36**, 497–510.

[58] Cao, R., J. M. Vilar y A. Devia (2009). Modelling consumer credit risk via survival analysis (with discussion), *Statistics and Operations Research Transactions*, **33**, 3–30.

[59] Carling, K., T. Jacobson y K. Roszbach (1998). Duration of consumer loans and bank lending policy: dormancy versus default risk, Working Paper Series in Economics and Finance No. 280, Stockholm School of Economics.

[60] Cecchetti, S. G., M. Kohler y C. Upper (2009). Financial crises and economic activity. Bank for International Settlements, Basel.

[61] Chacón, J. E. y T. Duong (2010). Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices, *Test*, **19**, 375–398.

[62] Chan-Lau, J. (2006). Fundamentals-based estimation of default probabilities: a survey, International Monetary Fund, Working Paper No. WP/06/149, Washington D. C.

[63] Chen, S., W. K. Härdle y R. A. Moro (2006). Estimation of default probabilities with support vector machines, Center of Applied Statistics and Economics, Discussion Paper No. 2006-077, Humboldt University of Berlin.

[64] Chen, Y., R. J. Guo y R. L. Huang (2009). Two stages credit evaluation in bank loan appraisal, *Economic Modelling*, **26**, 63–70.

[65] Chen, Y. S., P. H. Ho, C. Y. Lin y W. C. Tsai (2012). Applying recurrent event analysis to understand the causes of changes in firm credit ratings, *Applied Financial Economics*, **22**, 977–988.

[66] Chu, C. K. y K. F. Cheng (1995). Nonparametric regression estimates using misclassified binary responses, *Biometrika*, **82**, 315–325.

[67] Chu, C. K. y J. S. Marron (1991). Choosing a kernel regression estimator (with discussion), *Statistical Science*, **6**, 404–436.

[68] Clavero, B. (2006). *Statistical aspects of setting up a credit rating syste,* Doctoral Thesis of the Technical University of Kaiserslautern, Germany.

[69] Cook, R. J. y J. F. Lawless (2007). *The statistical analysis of recurrent events*, Springer, New York.

[70] Cox, D. R. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **34**, 187–220.

[71] Credit Suisse Financial Products (1997). *CreditRisk$^+$: A Credit Risk Management Framework*, Credit Suisse Financial Products, London.

[72] Crook, J. N., D. B. Edelmann y L. C. Thomas (2007). Recent developments in consumer credit risk assessment, *European Journal of Operational Research*, **183**, 1447–1465.

[73] Crouhy, M., D. Galai y R. Mark (2000). A comparative analysis of current credit risk models, *Journal of Banking & Finance*, **24**, 59–117.

[74] Czado, C. (1992). On link selection in generalized linear models. En: L. Fahrmeir, B., R. Francis, R. Gilchrist y G. Tutz (Editores). *Advances in GLIM and Statistical Modelling: Lecture Notes in Statistics,* **78**. Springer Verlag, New York, 60–65.

[75] Czado, C. (1997). On selecting parametric link transformation families in generalized linear models, *Journal of Statistical Planning and Inference*, **61**, 125–139.

[76] Dabrowska, D. (1987). Non-parametric regression with censored survival time data, *Scandinavian Journal of Statistics*, **14**, 181–197.

[77] Dabrowska, D. (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate, *The Annals of Statistics*, **17**, 1157–1167.

[78] Dabrowska, D. (1992a). Variable bandwidth conditional Kaplan-Meier estimate, *Scandinavian Journal of Statistics*, **19**, 351–361.

[79] Dabrowska, D. (1992b). Nonparametric quantile regression with censored data, *Sankhyä – The Indian Journal Of Statistics: Series A*, **19**, 351–361.

[80] Darayseh, M. y E. Waples (2010). Measuring type I and type II errors in an estimation model: an empirical analysis; the case of bankruptcy, *International Journal of Behavioural Accounting and Finance*, **1**, 268–275.

[81] Delaigle, A. e I. Gijbels (2004). Bootstrap bandwidth selection in kernel density estimation from a contaminated sample, *Annals of the Institute of Statistical Mathematics*, **56**, 19–47.

[82] DeLong, E. R., D. M. DeLong y D. L. Clarke-Pearson (1988). Comparing the areas under two or more correlated receiver opoerating characteristic curves: a nonparametric approach, *Biometrics*, **44**, 837–845.

[83] Desai, V.S., D. G. Convay, J. N. Crook y G. A. Overstreet (1997). Credit scoring models in the credit union environment using neural networks and genetic algorithms, *IMA Journal of Mathematics Applied in Business and Industry*, **8**, 323–346.

[84] Desai, V.S., J. N. Crook y G. A. Overstreet (1996). A comparison of neural networks and linear scoring models in the credit environment, *European Journal of Operational Research*, **95**, 24–37.

[85] Du, Y. y M. G. Akritas (2002). Uniform strong representation of the conditional Kaplan-Meier process, *Mathematical Methods of Statistics*, **11**, 152–182.

[86] Duffie, D. y K. J. Singleton (2003). *Credit Risk: Pricing, Measurement and Management*, Princeton University Press, Princeton.

[87] Duffie, D., L. Saita y K. Wang (2007). Multi-period corporate default prediction with stochastic covariates, *Journal of Financial Economics*, **83**, 635–665.

[88] Duong, T. (2007). Kernel density estimation and kernel discriminant analysis for multivariate data in *R*, *Journal of Statistical Software*, **21**, 1–16.

[89] Duong, T. (2014). Non-parametric kernel estimation of multivariate cumulative distribution and survival functions, and receiver operating characteristic curves. Enviado para su publicación.

[90] Duong, T. (2015). *ks:* Kernel Smoothing. *R* package version 1.9.4. URL: https://cran.r-project.org/web/packages/ks/ks.pdf

[91] Duong, T. y M. L. Hazelton (2005). Cross-validation bandwidth matrices for multivariate kernel density estimation, *Scandinavian Journal of Statistics*, **32**, 485–506.

[92] Durand, D. (1941). Risk elements in consumer instalment financing, National Bureau of Economics Research, New York.

[93] Efroymson, M. A. (1960). Multiple regression analysis. En: Ralston, A. y H. S. Wilf (Editores). *Mathematical Methods for Digital Computers*, John Wiley & Sons, New York, 191–203.

[94] Egan, J. P. (1975). *Signal detection theory and ROC analysis*. Academic Press, New York.

[95] Eisenbeis, R. A. (1977). Pitfalls in the applications of discriminant analysis in business, finance and economics, *Journal of Finance*, **32**, 875–900.

[96] Eisenbeis, R. A. (1978). Problems in applying discriminant anlaysis in credit scoring models, *Journal of Banking & Finance*, **2**, 205–219.

[97] Engelmann, B. (2006). Measures of a rating's discriminative power-applications and limitations. En: Engelmann, B. y R. Rauhmeier (Editores). *The Basel II Risk Parameters: Estimation, Validation and Stress Testing*, Springer, Berlin, 263–287.

[98] Engelmann, B., E. Hayden y D. Tasche (2003a). Testing rating accuracy, *Risk*, **16**, 82–86.

[99] Engelmann, B., E. Hayden y D. Tasche (2003b). Measuring the discriminative power of rating systems, Discussion paper Series 2: Banking and Financial Supervision, Document No. 01/2003, Deutsche Bank, Frankfurt.

[100] Engelmann, B. y R. Rauhmeier (2006). *The Basel II Risk Parameters: Estimation, Validation and Stress Testing*, Springer, Berlin.

[101] Fahrmeir, L. y G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer, New York.

[102] Falk, M. (1992). Bootstrap optimal bandwidth selection for kernel density estimates, *Journal of Statistical Planning and Inference*, **30**, 13–22.

[103] Fan, J. (1992). Design-adaptive nonparametric regression, *Journal of the American Statistical Association*, **87**, 998–1004.

[104] Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies, *The Annals of Statistics*, **21**, 196–216.

[105] Fan, J. e I. Gijbels (1996). *Local polynomial modelling and its applications.* Chapman & Hall/CRC, London.

[106] Faraway, J. J. y M. Jhun (1990). Bootstrap choice of bandwidth for density estimation, *Journal of The American Statistical Association*, **85**, 1119–1122.

[107] Farewell, W. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors, *Biometrics*, **38**, 1041–1046.

[108] Fawcett, T. (2006). An introduction to *ROC* analysis, *Pattern Recognition Letters*, **27**, 861–874.

[109] Finger, C. (1999). Conditional approaches for *CreditMetrics* portfolio distributions, *CreditMetrics Monitor*, **1**, 14–33.

[110] Finger, C. (2001). The one-factor *CreditMetrics* model in the New Basel Capital Accord, *RiskMetrics Journal*, **2**, 9–18.

[111] Fisher, R. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, **7**, 179–188.

[112] Freedman, D. A. (1981). Bootstrapping regression models, *The Annals of Statistics*, **6**, 1218–1228.

[113] Frölich, M. (2006). Non-parametric regression for binary dependent variables, *Econometrics Journal*, **9**, 511–540.

[114] Gannoun, A., J. Saracco y K. Yu (2007). Comparison of kernel estimators of conditional distribution function and quantile regression under censoring, *Statistical Modelling*, **7**, 329–344.

[115] Gannoun, A., J. Saracco, A. Yuan y G. E. Bonney (2005). Nonparametric quantile regression with censored data, *Scandinavian Journal of Statistics*, **32**, 527–550.

[116] Gjessing, H. K., K. Røysland, E. A. Peña y O. O. Aalen (2010). Recurrent events and the exploding Cox model, *Lifetime Data Analysis*, **16**, 525–546.

[117] Gill, R., K. Lee y S. Song (2007). Computation of estimates in segmented regression and a liquidity effect model, *Computational Statistics & Data Analysis*, **51**, 6459–6475.

[118] Girón, A. G. (1998). Crisis financieras y crisis financieras, Universidad de Marne-La-Vallée, Paris.

[119] Glennon, D. y P. Nigro (2005). Measuring the default risk of small business loans: a survival analysis approach, *Journal of Money, Credit, and Banking*, **37**, 923–947.

[120] González Manteiga, W. y M. C. Cadarso Suárez (1994). Asymptotic properties of a generalized Kaplan-Meier estimator with some applications, *Journal of Nonparametric Statistics*, **4**, 65–78.

[121] Gordy, M. (2000). A comparative anatomy of credit risk models, *Journal of Banking & Finance*, **24**, 119–149.

[122] Gordy, M. (2003). A risk-factor model foundation for ratings-based bank capital rules, *Journal of Financial Intermediation*, **12**, 199–232.

[123] Gourieroux, C. (2000). *Econometrics of Qualitative Dependent Variables*. Cambridge University Press, Cambridge.

[124] Greene, W. H. (1992). A statistical model for credit scoring, Stern School of Business Working Paper No. EC-92-29, New York University.

[125] Greene, W. H. (2008). *Econometric Analysis*, 6th Ed. Prentice Hall, New Jersey.

[126] Greiner, M., D. Pfeiffer y R. D. Smith (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests, *Preventive Veterinary Medicine*, **45**, 23–41.

[127] Grice, S. J. y R. W. Ingram (2001). Tests of generalizability of Altman's bankruptcy prediction model, *Journal of Business Research*, **54**, 53–61.

[128] Guerrero, V. M. y R. A. Johnson (1982). Use of the Box-Cox transformation with binary response models, *Biometrika*, **69**, 309–314.

[129] Gupta, J., A. Gregoriou y T. Ebrahimi (2015). Using hazard models correctly: a comparison employing different definitions of SMES finantial distress, Paper No. 422, Annual Meetings 2015 of the European Financial Management Association, Amsterdam. URL: http://www.efmaefm.org/EFMA2015_0422_fullpaper.pdf

[130] Gurevich, G. y A. Vexler (2005). Change point problems in the model of logistic regression, *Journal of Statistical Planning and Inference*, **131**, 313–331.

[131] Gurný, P. y M. Gurný (2013). Comparison of credit scoring models on probability of default estimation for US banks, *Prague Economic Papers*, **2**, 163–181.

[132] Hall, P. (1990). Using the bootstrap to estimate the mean squared error and select smoothing parameters in nonparametric problems, *Journal of Multivariate Analysis*, **32**, 177–203.

[133] Hamerle, A., T. Liebig y D. Rösch (2003). Credit risk factor modeling and the Basel II IRB approach, Discussion paper Series 2: Banking and Financial Supervision, Document No. 02/2003, Deutsche Bank, Frankfurt.

[134] Hand, D. J. (2001). Modelling consumer credit risk, *IMA Journal of Management Mathematics*, **12**, 139–155.

[135] Hanley, J. A. y B. J. McNeil (1982). The meaning and the use of the area under a receiver operating characteristic (*ROC*) curve, *Radiology*, **143**, 29–36.

[136] Hand, D. J. y W. E. Henley (1997). Statistical classification in consumer credit scoring: a review, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **160**, 523–541.

[137] Hanley, J. A. (1989). Receiver operating characteristic (*ROC*) methodology: the state of the art, *Critical Reviews in Diagnostic Imaging*, **29**, 307–335.

[138] Hanley, J. A. y K. O. Hajian-Tilaki (1997). Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update, *Academic Radiology*, **4**, 49–58.

[139] Hanley, J. A. y B. J. McNeil (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases, *Radiology*, **148**, 839–843.

[140] Hanson, S. y T. Schuerman (2004). Estimating probabilities of default, Federal Reserve Bank of New York, Staff Report no. 190.

[141] Hardin, J. W. y J. M. Hilbe (2007). *Generalized Linear Models and Extensions*, 2nd Ed., Stata Press, New York.

[142] Härdle, W. y A. W. Bowman (1988). Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands, *Journal of the American Statistical Association*, **83,** 102–110.

[143] Härdle, W. y J. S. Marron (1985). Optimal bandwidth selection in nonparametric regression function estimation, *The Annals of Statistics*, **13**, 1465–1481.

[144] Hastie, T. y C. Loader (1993). Local regression: automatic kernel carpentry, *Statistical Science*, **8**, 120–143.

[145] Hayfield, T. y J. S. Racine (2008). Nonparametric Econometrics: The np Package, *Journal of Statistical Software*, **27**, 1–34.

[146] Hazelton, M. L. (2007). Bias reduction in kernel binary regression, *Computational Statistics and Data Analysis*, **51**, 4393–4402.

[147] Henley, W. E. y D. J. Hand (1996). A k-nearest-neighbour classifier for assessing consumer credit risk, *Statistician*, **45**, 77–95.

[148] Hodges, J. L. y E. L. Lehmann (1956). The e ciency of some nonparametriccompetitors of the t-test, *The Annals of Mathematical Statistics*, **27**, 324–335.

[149] Hong, C. S. (2009). Optimal threshold from ROC and CAP curves, *Communications in Statistics – Simulation and Computation*, **38**, 2060–2072.

[150] Hong, Y. (2012). On computing the distribution function for the Poisson binomial distribution. *Computational Statistics & Data Analysis*, **59**, 41–51.

[151] Hosmer, D. W. y S. Lemeshow (1980). A goodness-of-fit test for the multiple logistic regression model, *Communications in Statistics – Theory and Methods*, **10**, 1043–1069.

[152] Hosmer, D. W. y S. Lemeshow (2000). *Applied Logistic Regression*, 2nd Ed. John Wiley & Sons, New York.

[153] Hurvich, C. M., J. S. Simonoff y C. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **60**, 271–293.

[154] Iglesias Pérez, M. C. (2001). *Estimación de la función de distribución condicional con censura y truncamiento*. Tesis Doctoral de la Universidad de Santiago de Compostela, España.

[155] Iglesias Pérez, M. C. y W. González Manteiga (1999). Strong representation of a generalized product-limit estimator for truncated and censored data with some applications, *Journal of Nonparametric Statistics*, **10**, 213–244.

[156] Iglesias Pérez, M. C. y W. González Manteiga (2003). Bootstrap for the conditional distribution function with truncated and censored data, *Annals of the Institute of Statistical Mathematics*, **55**, 331–357.

[157] Jarque, C. y A. Bera (1987). A test for normality of observations and regression residuals, *International Statistical Review*, **55**, 163–172.

[158] Jin, Z., D. Y. Lin y Z. Ying (2006). Rank regression analysis of multivariate failure time data based on marginal linear models *Scandinavian Journal of Statistics*, **33**, 1–23.

[159] Jones (1990). The performance of kernel density functions in kernel distribution function estimation, *Statistics and Probability Letters*, **41**, 163–168.

[160] Kaplan, E. L. y P. Meier (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, **53**, 457–481.

[161] Kay, R. y S. Little (1987). Transformations of the explanatory variables in the logistic regression model for binary data, *Biomatrika*, **74**, 495–501.

[162] Koh, H. C. (1992). The sensitivity of optimal cutoff points to misclassification costs of type I and type II errors in the going-concern prediction context, *Journal of Business Finance & Accounting*, **19**, 187–197.

[163] Kraft, H., G. Kroisandt y M. Müller (2003). Assessing the discriminatory power of credit scores under censoring. Working Paper from Fraunhofer Institute for Technical and Industrial Mathematics (ITWM).

[164] Kraft, H., G. Kroisandt y M. Müller (2004). Redisigning ratings: assesing the discriminatory power of credit scores under censoring. Retrieved January 10, 2006, Working Paper from Fraunhofer Institute for Technical and Industrial Mathematics (ITWM).

[165] Krzanowski, W. J. y D. J. Hand (2009). *ROC Curves for Continuous Data*, Chapman & Hall/CRC, London.

[166] Laitila, T. (1993). A pseudo-R^2 measure for limited and qualitative dependent variable models, *Journal of Econometrics*, **56**, 341–356.

[167] Lakhal-Chaieb, L., B. Abdous y T. Duchesne (2013). Nonparametric estimation of the conditional survival function for bivariate failure times, *The Canadian Journal of Statistics*, 41, 439–452.

[168] Lane, S. (1972). Submarginal credit risk classification, *Journal of Financial and Quantitative Analysis*, **7**, 1379–1385.

[169] Lawal, B. (2003). *Categorical Data Analysis with SAS® and SPSS Applications*, Laurence Erlbaum Associates, London.

[170] Lawless, J. F. y Nadeau, C. (1995). Some simple robust methods for the analysis of recurrent events, *Technometrics*, **37**, 158–168.

[171] Lawrence, E. C., S. L. Douglas y M. Rhoades (1992). An analysis of default risk in mobile home credit, *Journal of Banking & Finance*, **2**, 299–312.

[172] Leconte, E., S. Poiraud-Casanova y C. Thomas-Agnan (2002). Smooth conditional distribution function and quantiles under random censorship, *Lifetime Data Analysis*, **8**, 229–246.

[173] Lejeune, M. y P. Sarda (1992). Smooth estimators of distribution and density functions, *Computational Statistics & Data Analysis*, **14**, 457–471.

[174] Li, G. y S. Datta (2001). A bootstrap approach to nonparametric regression for right censored data, *Annals of the Institute of Statistical Mathematics*, **53**, 708–729.

[175] Li, G. y H. Doss (1995). An approach to nonparametric regression life history data using local linear fitting, *The Annals of Statistics*, **23**, 787–823.

[176] Lin, D. Y. (1994). Cox regression analysis of multivariate failure time data: the marginal approach, *Statistics in Medicine*, **13**, 2233–2247.

[177] Lin, D. Y., L. J. Wei y Z. Ying (1998). Accelerated failure time models for counting processes, *Bimetrika*, **85**, 605–618.

[178] Lin, D. Y., W. Sun y Z. Ying (1999). Nonparametric estimation of the gap time distributions for serial events with censored data, *Biometrika*, **86**, 59–70.

[179] Liu, B. y W. Lu (2011). Semiparametric additive intensity model with frailty for recurrent events, *Acta Mathematica Sinica, English Series*, **27**, 1831–1842.

[180] Liu, B., W. Lu y J. Zhang (2014). Accelerated intensity frailty model for recurrent events data, *Biometrics*, **70**, 579–587.

[181] Lo, A. W. (1985). Logit versus discriminant analysis - a specification test and application to corporate bankruptcies, *Journal of Econometrics*, **31**, 151–178.

[182] Long, S. (1997). *Regression Models for Categorical and Limited Dependent Variables*, SAGE Publications, California.

[183] Lopez, O. (2011). Nonparametric estimation of the multivariate distribution function in a censored regression model with applications, *Communications in Statistics – Theory and Methods*, **40**, 2639–2660.

[184] Lu, W. (2005). Marginal regression of multivariate event times based on linear transformation models, *Lifetime Data Analysis*, **11**, 389–404.

[185] Lucas, A., P. Klaasen, P. Spreij y S. Straetmans (2001). An analytic approach to credit risk of large corporate bond and loan portfolios, *Journal of Banking & Finance*, **25**, 1635–1664.

[186] Maddala, G. S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, Cambridge.

[187] Magge, L. (1990). R^2 measures based on Wald and likelihood ratio joint significance tests, *The American Statistician*, **44**, 250–253.

[188] Malik, M. y L. C. Thomas (2006). Modelling credit risk of portfolio of consumer loans, School of Management Working Paper Series No. CORMSIS-07-12, University of Southampton.

[189] Maller, R. A. y X. H. Zhou (1996). *Survival Analysis with Long-Term Survivors*. John Wiley & Sons, New York.

[190] Marron, J. S. (1992). Bootstrap bandwidth selection. En: LePage, R. y L. Billards (Editores). *Exploring the Limits of the Bootstrap*, 249–262.

[191] Martens D., B. Baesens, T. Van Gestel y J. Vanthienen (2007). Comprehensible credit scoring models using rule extraction from support vector machines, *European Journal of Operational Research*, **183**, 1466–1476.

[192] Martens D., J. Vanthienen, W. Verbeke y B. Baesens (2011). Performance of classification models from a user perspective, *Decision Support Systems*, **51,** 782–793.

[193] McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. En: Zarembka, P. (Editor). *Frontiers in Econometrics*. Academic Press, New York, 105–142.

[194] McKeague, I. W. y K. J. Utikal (1990). Inference for a nonlinear counting process regression model, *The Annals of Statistics*, **18**, 1172–1187.

[195] McNeil, B. J. y J. A. Hanley (1984). Statistical approaches to the analysis of receiver operating characteristic (*ROC*) curves, *Medical Decision Making*, **4**, 137–150.

[196] Merton, R. C. (1974). On the pricing of corporate debt: the risk structure of interest rates, *Journal of Finance*, **29**, 449–470.

[197] Metz, C. (2006). Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems, *Journal of the American College of Radiology*, **3**, 413–422.

[198] Min, J. H., y Y. C. Lee (2008). A practical approach to credit scoring, *Expert Systems with Applications*, **35**, 1762–1770.

[199] Morgan, J. P. (1997). *CreditMetrics^{TM}– Technical Document: The benchmark for understanding credit risk*, J. P. Morgan & Co. Inc., New York.

[200] Moses, D. y S. S. Liao (1987). On developing models for failure prediction, *Journal of Commercial Bank Lending*, **69**, 27–38.

[201] Müller, M. y B. Rönz (2000). Credit ccoring using semiparametric methods. En: Franke, J., W. Härdle y W. Stahl (Editores). *Measuring Risk in Complex Stochastic Systems*. Springer Verlag, New york, 83–97.

[202] Myers, J. H. y E. W. Forgy (1963). The development of numerical credit evaluation systems, *Journal of the American Statistical Association*, **58**, 799–806.

[203] Nadaraya, E. A. (1964). On estimating regression, *Theory of Probability and its Applications*, **9**, 141–142.

[204] Narain, B. (1992). Survival analysis and the credit granting decision. En: Thomas, L., J. N. Crook y D. B. Edelman (Editores). *Credit Scoring and Credit Control*, Oxford University Press, Oxford, 109–121.

[205] Neagu, R., S. Keenan y K. Chalermkraivuth (2009). Internal credit rating systems: methodology and economic value, *The Journal of Risk Model Validation*, **3**, 11–34.

[206] Nelder, J. A. y R. W. M., Wedderburn (1972). Generalized linear models, *Journal of the Royal Statistical Society: Series A (General)*, **135**, 370–384.

[207] Nelson, W. (1969). Hazard plotting for incomplete failure data, *Journal of Quality Technology*, **1**, 27–52.

[208] Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data, *Technometrics*, **14**, 945–965.

[209] Nordberg, L. (1981). Stepwise selection of explanatory variables in the binary logit model, *Scandinavian Journal of Statistics*, **8**, 17–26.

[210] Ojeda Cabrera, J. L. (2012). *locpol*: Kernel local polynomial regression, *R* package version 0.6-0. Url: http://cran.r-project.org/web/packages/locpol/

[211] Okumura, H. (2011). Kernel regression for binary response data, *Memoirs of the Faculty of Science, Shimane University. Series B: Mathematical Science*, **44**, 33–53.

[212] Parzen, E. (1962). On estimation of a probability density function and mode, *The Annals of Mathematical Statistics*, **33**, 1065–1076.

[213] Pastor, R. y E. Guallar (1998). Use of two-segmented logistic regression to estimate change-points in epidemiologic studies, *American Journal of Epidemiology*, **148**, 631–642.

[214] Pastor-Barriuso, R., E. Guallar y J. Coresh (2003). Transition models for change-point estimation in logistic regression, *Statistics in Medicine*, **15**, 1141–1162.

[215] Peña, E. A., R. L. Strawderman y M. Hollander (2001). Nonparametric estimation with recurrent event data, *Journal of the American Statistical Association*, **96**, 1299–1315.

[216] Pepe, M. S. (2002). Receiver operating characteristic methodology. En: Raftery, E. A., M. A. Tanner y M. T. Wells (Editores). *Statistics in the 21st Century*, Chapman & Hall/CRC, London.

[217] Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, Oxford.

[218] Platt, H. D. y M. B. Platt (1991). A note on the industry-relative rations in bankruptcy-prediction, *Journal of Banking & Finance*, **15**, 1183–1194.

[219] Pregibon, D. (1980). Goodness of link tests for generalized linear models, *Journal of the Royal Statistical Society: Series C ((Applied Statistics)*, **29**, 15–23.

[220] Pyle, D. H. (1997). Bank risk management, Research Program in Finance Working Papers, University of California, Berkeley.

[221] *R* Core Team (2015). *R*: A Language and Environment for Statistical Computing. *R* Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Url: http://www.R-project.org/

[222] Rauhmeier, R. (2006). *PD*-validation - experience from banking practice. En: Engelmann, B. y R. Rauhmeier (Editores). *The Basel II Risk Parameters: Estimation, Validation and Stress Testing*, Springer, 307–346.

[223] Reichert, A. K., C. C. Cho y G. M. Wagner (1983). An examination of the conceptual issues involved in developing credit-scoring models, *Journal of Business & Economic Statistics*, **1,** 101–114.

[224] Reiss, R. D. (1981). Nonparametric estimation of smooth distribution functions, *Scandinavian Journal of Statistics*, **8**, 116–119.

[225] Rosenberg, E. y A. Gleit (1994). Quantitative methods in credit management: a survey, *Operations Research*, **42**, 589–613.

[226] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function, *The Annals of Mathematical Statistics*, **27**, 832–837.

[227] Rosenblatt, M. (1971). Curve estimates, *The Annals of Statistics*, **42**, 1815–1842.

[228] Roszbach, K. (2004). Bank lending policy, credit scoring and the survival of loans, Working Paper Series No. 154, Sveriges Riksbank, Stockholm.

[229] Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators, *Scandinavian Journal of Statistics*, **9**, 65–78.

[230] Ruppert, D., S. J. Sheather y M. P. Wand (1995). An effective bandwidth selector for local least squares regression, *Journal of the American Statistical Association*, **90**, 1257–1270.

[231] Samreen, A. y F. B. Zaidi (2012). Design and development of credit scoring model for the commercial banks of Pakistan: forecasting creditworthiness of individual borrowers, *International Journal of Business and Social Science*, **17**, 155–166.

[232] Sarda, P. (1993). Smoothing parameter selection for smooth distribution functions, *Journal of Statistical Planning and Inference*, **35**, 65–75.

[233] Satchell, S. y W. Xia (2007). Analytic models of the *ROC* curve: applcations to credit rating model validation, Quantitative Finance Research Centre, Research Paper No. 181, University of Technology, Sydney.

[234] Saunders, A. y L. Allen (2002). *Credit Risk Measurement: New Approaches to Value at Risk and Other Paradigms*, 2nd Ed., John Wiley & Sons, New York.

[235] Saunders, A. y M. M. Cornett (2008). *Financial Institutions Management: A Risk Manangement Approach*, McGraw Hill, Singapore.

[236] Schaubel, D. E. y J. Cai (2004). Non-parametric estimation of gap time survival functions for ordered multivariate failure time data, *Statistics in Medicine*, **23**, 1885–1900.

[237] Schaubel, D. E., D. Zeng y J. Cai (2006). A semiparametric additive rates model for recurrent event data, *Lifetime Data Analysis*, **12**, 389–406.

[238] Schönbucher, P. J. (2000). Factor models for portfolio credit risk, Working Paper, University of Bonn.

[239] Schönbucher, P. J. (2003). *Credit Derivatives Pricing Models*, Springer, New York.

[240] Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**, 461–464.

[241] Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*, John Wiley & Sons, Inc., New York.

[242] Scott, J. (1981). The probability of bankruptcy: a comparison of empirical predictions and theoretical models, *Journal of Banking & Finance*, **5**, 317–344.

[243] Sheather, S. J. y M. C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **53**, 683–690.

[244] Shtatland, E. S., E. Cain y M. B. Barton (2001). The perils of stepwise logistic regression and how to escape them using information criteria and the output delivery system, *Proceedings of the Twenty-Sixth Annual SAS® Users Group International Conference*, Paper No. p222-26, SAS Institute Inc, North Carolina.

[245] Signorini, D. F. (1998). *Practical aspects of kernel smoothing for binary regression and density estimation.* Doctoral Thesis of The Open University, United Kingdom.

[246] Signorini, D. F. y M. C. Jones (2004). Kernel estimators for univariate binary regression, *Journal of The American Statistical Association*, **99**, 119–126.

[247] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall/CRC, London.

[248] Sobehart, J. R. y S. C. Keenan (2001). Measuring deafult accurately, *Risk*, **1** , 31–33.

[249] Sobehart, J. R., S. C. Keenan y R. M. Stein (2000). Validation methodologies for default risk models, *Rating Methodology*, March, Moody's KMV Investors Services Working Paper Series.

[250] Sobehart, J. R., S. C. Keenan y R. M. Stein (2001). Benchmarking quantitative default risk models: a validation methodology, *Algo Research Quarterly*, March-June, **4**, 57–72.

[251] Srinivasan, V. y Y. H. Kim (1987). Credit granting: a comparative analysis of clasification procedures, *Journal of Finance*, **42**, 665–681.

[252] Steenackers, A. y M. J. Goovaerts (1989). A credit scoring model for personal loans, *Insurance: Mathematics and Economics*, **8**, 31–34.

[253] Stein, R. M. (2005). The relationship between default prediction and lending profits: integrating the *ROC* analysis and loan pricing, *Journal of Banking and Finance*, **29**, 1213–1236.

[254] Stein, R. M. (2007). Benchmarking default prediction models: pitfalls and remedies in model validation, *Journal of Risk Model Validation*, **1**, 77–133.

[255] Stepanova, M. y L. C. Thomas (2002). Survival analysis methods for personal loan data, *Operations Research*, **50**, 277–289.

[256] Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators, *The Annals of Statistics*, **8,** 1348–1360.

[257] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression (Special Invited Paper), *The Annals of Statistics*, **18,** 907–924.

[258] Strzalkowska-Kominiak, E. y R. Cao (2014). Beran-based approach for single-index models under censoring, *Computational Statistics*, **29**, 1243–1261.

[259] Sun, M.Y y S. F. Wang (2007). Validation of credit rating models: a preliminary look at methodology and literature review, *Review of Financial Risk Management*, Joint Credit Information Center Risk Research Team, Taipei.

[260] Swets, J. A. (1988). Measuring the accuracy of diagnostic systems, *Science*, **240**, 1285–1293.

[261] Tasche, D. (2006). Validation of internal rating systems and PD estimates, *Physics and Society*, Working Paper No. physics/0606071.

[262] Taylor, C. C. (1989). Bootstrap choice of the smoothing parameter in kernel density estimation, *Biometrika*, **76**, 705–712.

[263] Thomas, L. C. (2000). A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers, *International Journal of Forecasting*, **16**, 149–172.

[264] Thomas, L. C., J. N. Crook y D. B. Edelman (1992). *Credit Scoring and Credit Control*. Oxford University Press, Oxford.

[265] Thomas, L. C., D. B. Edelman y J. N. Crook (2002). *Credit Scoring and Its Applications*. Society for Industrial and Applied Mathematics, Philadelphia.

[266] Tong, E. N. C., C. Mues y L. C. Thomas (2012). Mixture cure models in credit scoring: If and when borrowers default, *European Journal of Operational Research*, **218**, 132–139.

[267] Trías, R., F. Carrascosa, D. Fernández, Ll. Parés y G. Negot (2005). *Riesgo de créditos: Conceptos para su medición, Basilea II, Herramientas de Apoyo a la Gestión*, AIS Group-Financial Decisions, Barcelona.

[268] Trucharte, C. y A. M. Antuña (2001). Modelos factoriales de riesgo de crédito: el modelo de Basilea II y sus implicaciones, *Estabilidad Financiera*, **1**, 205–218.

[269] Turnbull, S. M. (2003). Practical issues in modeling default dependence, Unpublished manuscript, University of Houston.

[270] Twala, B. (2009). Multiple classifier application to credit risk assessment, *Expert Systems with Applications*, **37**, 3326–3336.

[271] Van Gestel, T., B. Baesens, P. Van Dijcke, J. Suykens, J. García y T. Alderweireld (2005). Linear and non-linear credit scoring by combining logistic regression and support vector machines, *Journal of Credit Risk*, **1**, 31–60.

[272] Van Keilegom, I. (1998). *Nonparametric Estimation of the Conditional Distribution in Regression with Censored Data*. Doctoral Thesis of The Hasselt University, Belgium.

[273] Van Keilegom, I. y N. Veraverbeke (1996). Uniform strong convergence results for the conditional Kaplan-Meier estimator and its quantiles, *Communications in Statistics - Theory and Methods*, **25**, 2251–2265.

[274] Van Keilegom, I. y N. Veraverbeke (1997a). Weak convergence of the bootstrapped conditional Kaplan-Meier process and its quantile process, *Communications in Statistics - Theory and Methods*, **26**, 853–869.

[275] Van Keilegom, I. y N. Veraverbeke (1997b). Estimation and bootstrap with censored data in fixed design nonparametric regression, *Annals of the Institute of Statistical Mathematics*, **49**, 467–491.

[276] Van Keilegom, I. y N. Veraverbeke (1998). Bootstrapping quantiles in a fixed design regression model with censored data, *Journal of Statistical Planning and Inference*, **69**, 115–131.

[277] Van Keilegom, I., M. G. Akritas y N. Veraverbeke (2001). Estimation of the conditional distribution in regression with censored data: a comparative study, *Computational Statistics & Data Analysis*, **35**, 487–500.

[278] Vasicek, O. (1987). Probability of loss on loan portfolio, Working Paper, *Moody's KMV Corporation*.

[279] Vasicek, O. (1997). The loan loss distribution, Technical Report, *Moody's KMV Corporation*.

[280] Veall, M. R. y K. F. Zimmermann (1994). Evaluating pseudo-R^2's for binary probit models, *Quality and Quantity*, **28**, 151–164.

[281] Veall, M. R. y K. F. Zimmermann (1996). Pseudo-R^2 measures for some common limited dependent variable models, Institute for Statistics, University of Munich.

[282] Vexler, A. y G. Gurevich (2009). Average most powerful tests for a segmented regression, *Communications in Statistics - Theory and Methods*, **38**, 2214–2231.

[283] Visser, M. (1996). Nonparametric estimation of the bivariate survival function with an application to vertically transmitted AIDS, *Biometrika*, **83**, 507–518.

[284] Wand, M. P. y M. C. Jones (1994). Multivariate plug-in bandwidth selection, *Computational Statistics*, **9**, 97–116.

[285] Wand, M. P. y M. C. Jones (1995). *Kernel Smoothing*, Chapman & Hall/CRC, London.

[286] Wand, M. P. y B. Ripley (2008). *KernSmooth:* Functions for kernel smoothing. *R* package version 2.22-22. Url: http://cran.rproject.org/web/packages/KernSmooth/

[287] Wang, M. C. y S. H. Chang (1999). Nonparametric estimation of a recurrent survival function, *Journal of the American Statistical Association*, **94**, 146–153.

[288] Wang, W. y M. T. Wells (1998). Nonparametric estimation of successive duration times under dependent censoring, *Biometrika*, **85**, 561–572.

[289] Watson, G. S. (1964). Smooth regression analysis, *Sankhya – The Indian Journal Of Statistics: Series A*, **26**, 359–372.

[290] Wei, L. J., D. Y. Lin y L. Weissfeld (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions, *Journal of the American Statistical Association*, **84**, 1065–1073.

[291] West, D. (2000). Neural network credit scoring models, *Computers & Operations Research*, **27**, 1131–1152.

[292] West, D., S. Dellana y J. Qian (2005). Neural network esemble strategies for financial decision applications, *Computers & Operations Research*, **32**, 2543–2559.

[293] Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior, *Journal of Financial and Quantitative Analysis*, **15**, 757–770.

[294] Wilson, T. (1997a). Portfolio credit risk I, *Risk*, September.

[295] Wilson, T. (1997b). Portfolio credit risk II, *Risk*, October.

[296] Windmeijer, F. A. G. (1995). Goodness of fit measures in binary choice models, *Econometric Reviews*, **14**, 101–116.

[297] Yobas, M. B., J. N. Crook y P. Ross (2000). Credit scoring using neural and evolutionary techniques, *IMA Journal of Mathematics Applied in Business and Industry*, **11**, 111–125.

[298] Zellner, D., F. Keller y G. E. Zellner (2004). Variable Selection in Logistic Regression Models, *Communications in Statistics - Simulation and Computation*, **33**, 787–805.

[299] Zeng, D. y D. Y. Lin (2007). Semiparametric transformation models with random effects for recurrent events, *Journal of the American Statistical Association*, **102**, 167–180.

[300] Zhou, X. H., N. A. Obuchowski y D. K. McClish (2002). *Statistical Methods in Diagnostic Medicine*, John Wiley & Sons, New York.

[301] Zhu, H. (2014). Non-parametric analysis of gap times for multiple event data: an overview, *International Statistical Review*, **82**, 106–122.