Trabajo Fin de Máster

Estimación no paramétrica de la probabilidad de mora en riesgo de crédito

Un estudio comparativo

Rebeca Peláez Suárez

Máster en Técnicas Estadísticas

Curso 2018-2019

Propuesta de Trabajo Fin de Máster

Título en galego: Estimación non paramétrica da probabilidade de incumplimiento no risco de crédito. Un estudo comparativo.
Título en español: Estimación no paramétrica de la probabilidad de mora en riesgo de crédito. Un estudio comparativo.
English title: Nonparametric estimation of the probability of default in credit risk. A comparative study.
Modalidad: Modalidad A
Autor/a: Rebeca Peláez Suárez, Universidade da Coruña
Director/a: Ricardo Cao Abad, Universidade da Coruña; Juan Manuel Vilar Fernández, Universidade da Coruña
Breve resumen del trabajo: En el presente trabajo se estudia el problema de la estimación de la probabilidad de mora en riesgo de crédito mediante técnicas de estimación de la distribución condicional con datos censurados. Se consideran los estimadores de Beran, Cai y Van Keilegom y Akritas, así como versiones suavizadas de los mismos. Los distintos métodos se ilustran aplicándolos a datos simulados y a datos reales.
Recomendaciones:
Otras observaciones:

Don Ricardo Cao Abad, catedrático de la Universidade da Coruña y Don Juan Manuel Vilar Fernández, catedrático de la Universidade da Coruña informan que el Trabajo Fin de Máster titulado

Estimación no paramétrica de la probabilidad de mora en riesgo de crédito. Un estudio comparativo

fue realizado bajo su dirección por doña Rebeca Peláez Suárez para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En A Coruña, a 30 de Enero de 2019.

El director: El director:

Don Ricardo Cao Abad Don Juan Manuel Vilar Fernández

La autora:

Doña Rebeca Peláez Suárez

Este trabajo ha sido financiado por la Xunta de Galicia (Grupo de Referencia Competitiva ED431C-2016-015), ayuda cofinanciada por el FEDER.

Este traballo foi financiado pola Xunta de Galicia (Grupo de Referencia Competitiva ED431C-2016-015), axuda cofinanciada polo FEDER.

This work has been supported by the Xunta de Galicia (Grupo de Referencia Competitiva ED431C-2016-015), through the ERDF.

Índice general

Resumen

En el presente trabajo se estudia el problema de la estimación de la probabilidad de mora en riesgo de crédito mediante técnicas de estimación de la distribución condicional con datos censurados. Se consideran los estimadores de Beran, Cai y Van Keilegom y Akritas, así como versiones suavizadas de los mismos. Los distintos métodos se ilustran aplicándolos a datos simulados y a datos reales.

Abstract

This work studies the problem of estimating the probability of default in credit risk using techniques for estimating conditional distribution with censored data. The estimators of Beran, Cai and Van Keilegom and Akritas are considered, as well as smoothed versions of them. These methods are illustrated by applying them to simulated and real data.

Introducción

En 1992, Naraim publicó su artículo "*Survival analysis and the credit granting decision*" en el que defendía el uso del análisis de supervivencia en el contexto del riesgo de crédito. Concretamente, argumentó la posibilidad de analizar todas las operaciones crediticias que involucran variables predictoras y en las que el tiempo hasta la ocurrencia de un evento es la variable de interés mediante el análisis de supervivencia.

Si el riesgo de crédito se define como la posible pérdida que asume un agente económico en caso de que la contraparte incumpla sus obligaciones contractuales, es evidente que la posibilidad de que un cliente que ha recibido un préstamo de una entidad bancaria por medio de una tarjeta de crédito, una hipoteca, un préstamo personal, etc, se declare incapaz de pagar la deuda contraída compromete el capital de dicha entidad. Es por ello que para la entidad bancaria resulta importante determinar la probabilidad de que un crédito en riesgo de incumplimiento caiga en impago o mora, convirtiéndose en un crédito moroso. En este escenario, la variable de interés a la que Naraim hacía referencia sería, precisamente, el tiempo hasta la caída en mora. Esta variable no es completamente observable: sólo es posible conocer el tiempo de vida de un crédito hasta que el cliente deja de pagarlo cuando la caída en impago tiene lugar durante el tiempo de observación de los créditos; en otro caso, el dato es censurado y el tiempo observado es el tiempo hasta la censura.

Por otro lado, la puntuación crediticia, *credit scoring* en inglés, es una calificación que el banco asigna a clientes o futuros clientes con la intención de evaluar su capacidad para hacer frente a una posible deuda que contraiga con el banco a través de un préstamo. La puntuación crediticia juega, por tanto, el papel de la variable predictora. Entonces, salta

a la vista la analogía existente, y de la que Naraim se percató (véase Naraim (1992)), entre el *tiempo hasta la caída en mora* y el *tiempo hasta el suceso de interés*, habitual en modelos biométricos y en esta analogía reside la motivación para aplicar técnicas de análisis de supervivencia en problemas de riesgo de crédito.

En esta memoria, y siguiendo esta idea, se proponen modelos de supervivencia que permitan estimar la *probabilidad de mora* (denotada como PD por sus siglas en inglés *probability of default*) como función de la puntuación crediticia en créditos personales, pues es claro el interés que tienen las entidades financieras en conocer la probabilidad de que un acreditado se declare incapaz de pagar su deuda al cabo de cierto tiempo (en general un año) de su formalización.

En el capítulo 1 de este trabajo se muestran las condiciones bajo las cuales el análisis de supervivencia tiene cabida en el estudio de la probabilidad de mora. En los capítulos 2, 3 y 4 se presentan tres estimadores para dicha probabilidad. El primero de ellos, expuesto en el capítulo 2, se construye a partir del estimador límite-producto generalizado de Beran para la supervivencia; las otras dos propuestas, capítulos 3 y 4, están basadas en modelos de regresión no paramétricos. En cada uno de estos tres capítulos se analiza el comportamiento de estos estimadores en su uso sobre muestras de datos simulados. Finalmente, se comparan los resultados obtenidos con cada uno de ellos, discutiendo cuál de los tres estimadores propuestos arroja mejores estimaciones de la probabilidad de mora y se aplican a un conjunto de datos reales procedentes de tarjetas de crédito.

Capítulo 1

Preliminares

El uso del análisis de supervivencia condicional en el contexto del riesgo de crédito permite construir estimadores para la probabilidad de mora utilizando como variable de interés el *tiempo hasta que se produce el impago*, denotada por T. El suceso que se pretende observar es la caída en mora que está, en parte, determinada por la covariable unidimensional X, que denota la puntuación crediticia o scoring del acreditado. Sin embargo, la variable T no es completamente observable y en numerosas ocasiones lo que se conoce es el tiempo hasta la cancelación o vencimiento del crédito o el tiempo hasta la censura, es decir, se observa la variable unidimensional denotada por C, *tiempo de potencial censura*.

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. Suponiendo que el tiempo de observación sea el intervalo $[0, \tau]$, existen tres situaciones posibles.

(a) **El crédito cae en mora.**

El instante de tiempo en que se produce la caída en mora se encuentra en el intervalo $[0, \tau]$ y, por tanto, puede ser observado; se tiene $T \leq C$. En este escenario el tiempo de vida del crédito es *no censurado*.

(b) **El crédito está activo y se está pagando.**

El crédito aún no ha caído en impago cuando el tiempo de observación finaliza. La mora, de producirse, no puede ser observada. En este caso se tiene $T > C$ y el tiempo

de vida del crédito es *censurado por la derecha.*

(c) **El crédito es cancelado anticipadamente.**

El crédito ha terminado de pagarse o bien ha sido cancelado de forma anticipada, siendo su tiempo de vida menor que τ. En cualquier caso, no se observa el impago, se tiene $T > C$. Por tanto, se considera un dato *censurado por la derecha.*

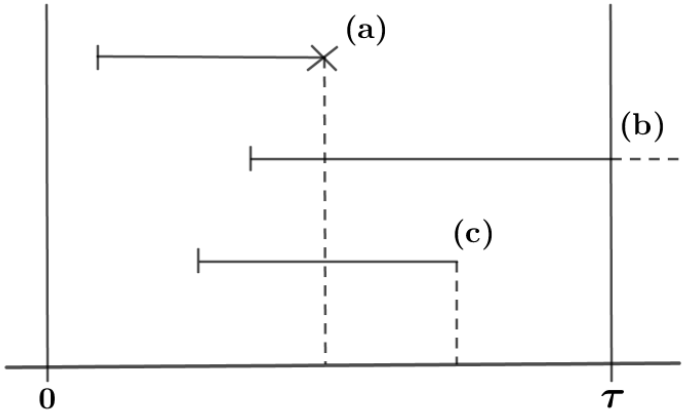


Figura 1.1: Estructura de censura

En este escenario, la información disponible será una muestra aleatoria simple de la terna (X, Z, δ), es decir, n ternas de variables aleatorias independientes e idénticamente distribuidas $(X_1, Z_1, \delta_1), ..., (X_n, Z_n, \delta_n)$, donde X es la covariable que representa la puntuación crediticia, Z es el tiempo de vida observado dado por $Z = \text{mín}\{T, C\}$ y $\delta = I_{\{T \leq C\}}$ es el indicador de no censura. De este modo, X_i es el grado de solvencia del i-ésimo acreditado, δ_i indica si se ha producido o no la caída en mora, tomando los valores 1 o 0, respectivamente, y Z_i será igual a T_i, el tiempo hasta la mora para i-ésimo acreditado, en el primer caso, o igual a C_i, el tiempo hasta la censura, en el segundo.

La puntuación crediticia, X, es una variable aleatoria real que tiene como soporte el intervalo $[0, 1]$ de manera que valores altos indican mayor solvencia del cliente. Existen diferentes métodos que permiten resumir la información de variables tales como género, estado civil, profesión, lugar de residencia, tipo de vivienda, edad, antigüedad laboral o cantidad de dinero disponible en una única variable, denominada scoring o puntuación crediticia, que mide la capacidad de un cliente para hacer frente al pago de la deuda

contraída con el banco. Algunas de las técnicas clásicas utilizadas para este fin son la regresión lineal multivariante, los modelos lineales generalizados y el análisis discriminante no paramétrico. Recientemente, también han cogido fuerza métodos basados en técnicas de automatización y minería de datos. En esta memoria no se analizará la bondad de estos métodos y la puntuación crediticia se supondrá conocida.

Como consecuencia del problema expuesto, existe una relación de dependencia entre T y X. Sin embargo, las variables T y C, dada X, se consideran condicionalmente independientes.

A lo largo de esta memoria se utilizará la siguiente notación:

- $F(t|x) = P(T \leq t|X = x)$ denota la función de distribución del tiempo hasta que se produce la mora condicionada a la puntuación crediticia.
- $G(t|x) = P(C \leq t|X = x)$ denota la función de distribución del tiempo hasta que se produce la censura condicionada a la puntuación crediticia.
- $H(t|x) = P(Z \leq t|X = x)$ denota la función de distribución del tiempo de vida observado condicionada a la puntuación crediticia.
- f_X denota la función de densidad de la variable X, puntuación crediticia.

Es fácil comprobar que se verifica la relación $1 - H(t|x) = \big(1 - F(t|x)\big)\big(1 - G(t|x)\big)$.

Además, se fijarán los siguientes supuestos:

H1. Las variables T, C y X son variables aleatorias no negativas absolutamente continuas.

H2. La variable aleatoria X no depende del tiempo. Se asume que la puntuación crediticia de un acreditado es constante en el tiempo.

H3. Las funciones de distribución condicionadas $F(t|x)$ y $G(t|x)$ son absolutamente continuas.

La función de probabilidad de mora a horizonte b, entendida como la probabilidad de que un crédito que ha sido pagado hasta el instante t y cuyo titular tiene asociada una puntuación crediticia x, caiga en impago en un periodo de tiempo siguiente de duración b, es decir, en el intervalo $(t, t+b]$, se puede escribir en términos de la función de supervivencia

condicional, tal y como se muestra a continuación:

$$\begin{aligned} PD(t|x) &= P(T \leq t+b|T > t, X = x) \\ &= \frac{P(T \leq t+b, T > t|X = x)}{P(T > t|X = x)} = \frac{P(t < T \leq t+b|X = x)}{1 - P(T \leq t|X = x)} \\ &= \frac{P(T \leq t+b|X = x) - P(T \leq t|X = x)}{1 - P(T \leq t|X = x)} = \frac{F(t+b|x) - F(t|x)}{1 - F(t|x)} \\ &= \frac{1 - F(t|x) - (1 - F(t+b|x))}{1 - F(t|x)} = 1 - \frac{S(t+b|x)}{S(t|x)}, \end{aligned} \tag{1.1}$$

siendo $S(t|x) = 1 - F(t|x)$ la función de supervivencia condicional.

La idea principal de este trabajo es encontrar estimadores adecuados de la función de supervivencia $S(t|x)$, que permitan estimar la función de probabilidad de mora y analizar su comportamiento. Dado $\widehat{S}(t|x)$, un estimador de la supervivencia condicionada a la puntuación crediticia x, es posible obtener un estimador de la probabilidad de mora a horizonte b sin más que sustituirlo en (1.1) como sigue:

$$\widehat{PD}(t|x) = 1 - \frac{\widehat{S}(t+b|x)}{\widehat{S}(t|x)}$$

Supóngase por un instante que se dispone de una muestra aleatoria simple no censurada $T_1, ..., T_n$ de una variable T y para la que no se tiene en cuenta ninguna covariable; por ejemplo, $T \sim Exp(\lambda)$, siendo $Exp(\lambda)$ la distribución exponencial de parámetro λ. En tal caso, si $t > 0$, la función de supervivencia de T viene dada por

$$S(t) = P(T > t) = 1 - F_{Exp(\lambda)}(t) = e^{-\lambda t}$$

y la función de probabilidad de mora a horizonte b puede hallarse como sigue

$$PD(t) = P(T \leq t+b|T > t) = 1 - \frac{S(t+b)}{S(t)} = 1 - e^{-\lambda b}$$

Como se explicó anteriormente, a lo largo de este trabajo se buscarán formas de estimar $S(t)$ que deriven en un buen estimador de $PD(t)$. En un primer acercamiento parece razonable estimar la función de supervivencia de T a partir de la función de distribución empírica, F_n, que se obtiene de la muestra $T_1, ..., T_n$ y, como resultado, conseguir una estimación de la probabilidad de mora. Esto es, obtener la supervivencia estimada como

$$S_n(t) = 1 - F_n(t) = 1 - \frac{1}{n}\sum_{i=1}^{n} I_{\{T_i \leq t\}} = \frac{1}{n}\sum_{i=1}^{n} I_{\{T_i > t\}}$$

entonces, la probabilidad de mora puede estimarse mediante

$$PD_n(t) = 1 - \frac{S_n(t+b)}{S_n(t)}.$$

En la figura 1.2 se muestra la gráfica de la función de supervivencia y la PD estimadas y teóricas para una muestra de tamaño $n = 400$ de una $Exp(2)$.

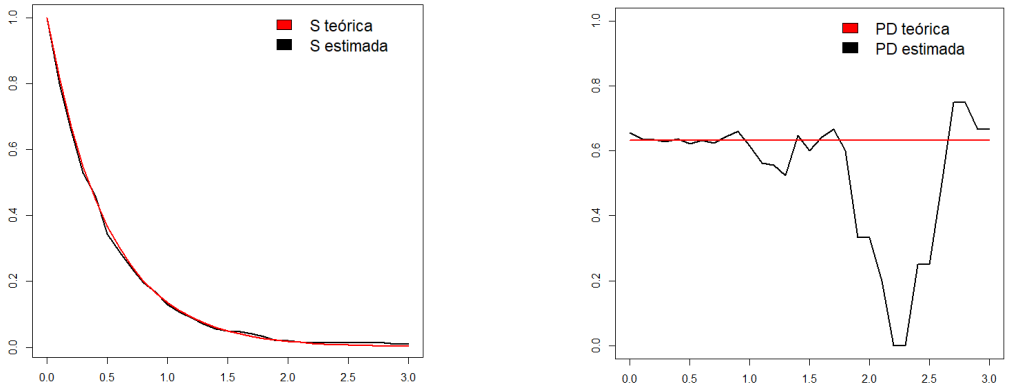


Figura 1.2: Estimación mediante la distribución empírica para una muestra de tamaño $n = 400$ de una $Exp(2)$. Izquierda: Función de supervivencia (línea roja) y su estimación (línea negra). Derecha: Función de probabilidad de mora (línea roja) y su estimación (línea negra).

Pese a estar bajo las hipótesis más sencillas (ausencia de censura y covariable) y conseguir mediante F_n una muy buena estimación de la supervivencia, el cociente de supervivencias en tiempos t y $t + b$ provoca saltos en la estimación de la PD como los que se ven en la gráfica de la derecha de la figura 1.2. Para intentar solucionar esto, parece adecuado proponer un estimador suavizado de la supervivencia $S(t)$.

Considérese brevemente el estimador tipo núcleo de la densidad dado por

$$\widehat{f}_g(t) = \frac{1}{ng} \sum_{i=1}^{n} K\left(\frac{t - T_i}{g}\right).$$

que fue propuesto por Parzen (1962) y Rosenblatt (1956). Se trata de un estimador no paramétrico donde K es una función núcleo (típicamente una densidad simétrica en torno al cero) y $g > 0$ es un parámetro de suavizado denominado ventana. La función de distribución asociada al estimador tipo núcleo de la función de densidad viene dada por

$$\begin{aligned} \widehat{F}_g(t) &= \int_{-\infty}^{t} \widehat{f}_g(u)du = \int_{-\infty}^{t} \frac{1}{ng} \sum_{i=1}^{n} K\left(\frac{u-T_i}{g}\right) du \\ &= \frac{1}{ng} \sum_{i=1}^{n} \int_{-\infty}^{t} K\left(\frac{u-T_i}{g}\right) du = \frac{1}{n} \sum_{i=1}^{n} \mathbb{K}\left(\frac{t-T_i}{g}\right) \end{aligned} \tag{1.2}$$

donde $\mathbb{K}(t) = \int_{-\infty}^{t} K(u)\, du$ es la función de distribución asociada al núcleo K. Entonces, $\widehat{F}_g(t)$ es el estimador tipo núcleo de la función de distribución y proporciona un buen estimador suavizado para $S(t)$ sin más que definir

$$\widehat{S}_g(t) = 1 - \widehat{F}_g(t)$$

En la figura 1.3 se muestran la supervivencia y la PD estimadas de forma suavizada y teóricas para la misma muestra de tamaño $n = 400$ de una $Exp(2)$ donde se observa la gran mejoría con respecto al estimador empírico mostrado en la figura 1.2.

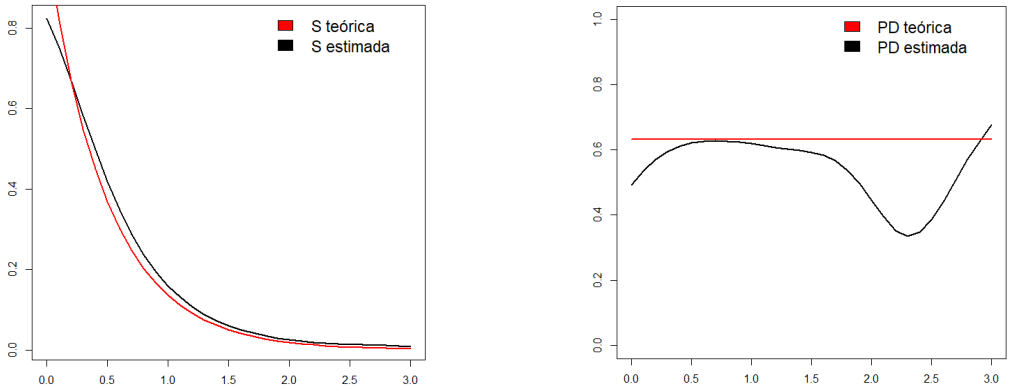


Figura 1.3: Estimación mediante la distribución empírica suavizada para una muestra de tamaño $n = 400$ de una $Exp(2)$. Izquierda: Función de supervivencia (línea roja) y su estimación (línea negra). Derecha: Función de probabilidad de mora (línea roja) y su estimación (línea negra).

Nótese que se está produciendo el denominado efecto frontera en la estimación en un entorno de $t = 0$. La función de distribución estimada está tomando valores positivos para tiempos negativos, lo que se traduce en una sobrestimación de $F(t)$ (y en consecuencia una infraestimación de $S(t)$) en valores de $t > 0$ cercanos a cero.

Es habitual que el dominio de definición de una función de densidad sea un intervalo real acotado por uno o ambos extremos y no la recta real completa. Por ejemplo, una función

de densidad $f(t)$ puede no estar definida para valores de t negativos y el objetivo sería obtener un estimador $\widehat{f}(t)$ que imite este comportamiento. Para ello, en Silverman (1986) se propone el *método de reflexión*, consistente en obtener $\widehat{f}(t)$ para t positivo y fijar el valor de $\widehat{f}(t)$ a cero para todo t negativo. Una vez hecho esto, ha de corregirse el estimador de manera que la densidad estimada integre la unidad. Esta idea se traslada al estimador dado en (1.2) para obtener un estimador de la función de distribución que corrige el efecto frontera en cero como sigue:

$$\widehat{F}_g^R(t) = \begin{cases} 0 & \text{si } t < 0 \\ \widehat{F}_g(t) - \widehat{F}_g(-t) & \text{si } t \geq 0 \end{cases}$$

De este modo se obtiene un estimador de la función de supervivencia que evita la infra-estimación de $S(t)$ en torno a $t = 0$, para $t > 0$:

$$\widehat{S}_g^R(t) = \begin{cases} 1 & \text{si } t < 0 \\ 1 + \widehat{S}_g(t) - \widehat{S}_g(-t) & \text{si } t \geq 0 \end{cases}$$

En la figura 1.4 se muestra la supervivencia y la probabilidad de mora estimadas teniendo en cuenta esta corrección del efecto frontera. En este trabajo todos los estimadores no paramétricos serán corregidos de esta forma.

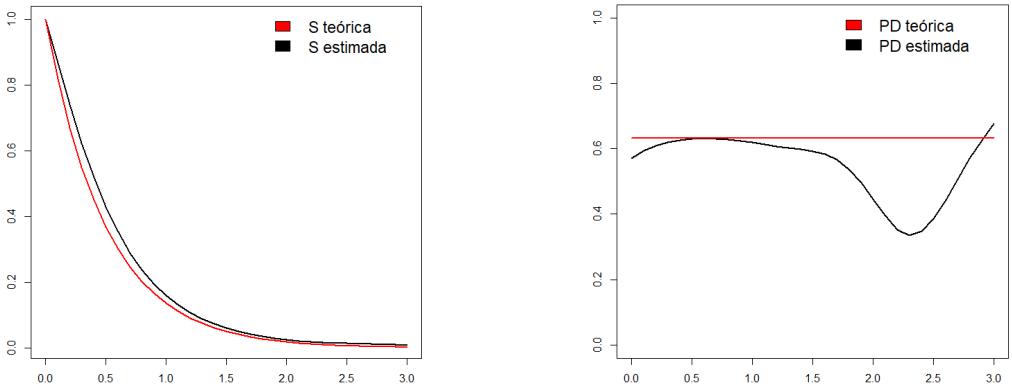


Figura 1.4: Estimación mediante la distribución empírica suavizada con corrección del efecto frontera para una muestra de tamaño $n = 400$ de una $Exp(2)$. Izquierda: Función de supervivencia (línea roja) y su estimación (línea negra). Derecha: Función de probabilidad de mora (línea roja) y su estimación (línea negra).

Supóngase ahora que la variable observada está censurada. Sean $T \sim Exp(\lambda_1)$ y $C \sim$

$Exp(\lambda_2)$ el tiempo de supervivencia y el tiempo de censura con distribuciones exponenciales de parámetros λ_1 y λ_2, respectivamente, y sea $(Z_1, \delta_1), ..., (Z_n, \delta_n)$ una muestra aleatoria simple censurada, donde $Z_i = \text{mín}\{T_i, C_i\}$ y $\delta_i = I_{\{I_i \leq C_i\}}$.

Dado $t > 0$, la función de supervivencia teórica es $S(t) = e^{-\lambda_1 t}$, de modo que la función de probabilidad de mora a horizonte b tiene la expresión

$$PD(t) = 1 - e^{-\lambda_1 b}.$$

En este caso, el estimador límite-producto propuesto por Kaplan and Meier (1958) proporciona un buen estimador de la función de supervivencia y viene dado por

$$\widehat{S}(t) = \prod_{Z_{(i)} \leq t} \left(\frac{n-i}{n-i+1} \right)^{\delta_{[i]}},$$

donde $Z_{(i)}$ es el i-ésimo elemento de la muestra $Z_1, ..., Z_n$ ordenada y $\delta_{[i]}$ su concomitante. El estimador $\widehat{S}$ es el estimador máximo verosímil no paramétrico de S y se basa en otorgar masa de probabilidad únicamente a datos no censurados de la variable Z, pero esa masa de probabilidad se ve afectada por cómo se distribuyen los datos censurados entre los no censurados. Nótese que en ausencia de censura, coincidiría con S_n.

En la figura 1.5 se representan gráficamente la supervivencia y la probabilidad de mora teóricas y estimadas para una muestra de tamaño $n = 400$ de tiempo de supervivencia $Exp(3)$ y tiempo de censura $Exp(2)$. Razonablemente, bajo censura se observa el mismo problema que en el caso sin censura: debido al cociente entre supervivencias que aparece en la expresión de la PD, la estimación es muy errática y tiene excesiva variabilidad. Se plantea la misma solución que en ausencia de censura, una suavización del estimador.

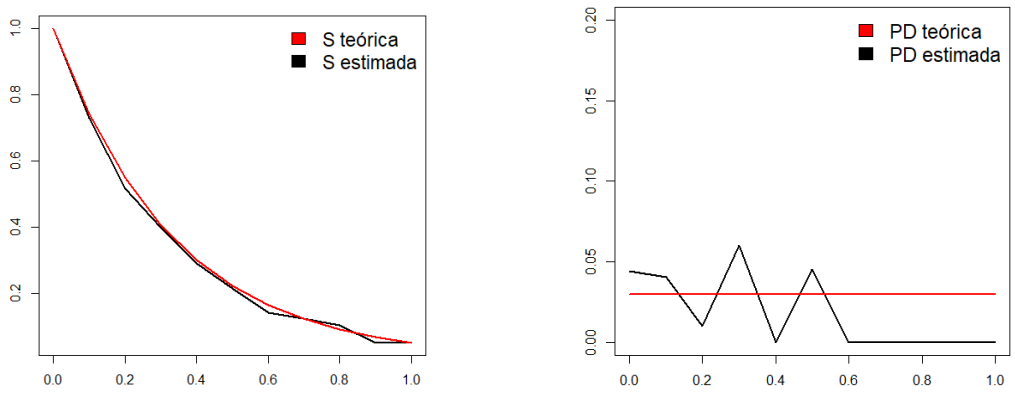


Figura 1.5: Estimación mediante Kaplan-Meier para una muestra censurada de tamaño $n = 400$ de una $Exp(3)$ con tiempo de censura $Exp(2)$. Izquierda: Función de supervivencia (línea roja) y su estimación (línea negra). Derecha: Función de probabilidad de mora (línea roja) y su estimación (línea negra).

La suavización propuesta se basa en estimar $S(t)$ mediante el promedio de los valores que toma $S_n(\cdot)$ en puntos de la muestra cercanos a t. Sea g un parámetro ventana que mide esa cercanía al punto t y sea $\mathbb{K}$ de nuevo la función de distribución de algún núcleo K. El estimador suavizado de la supervivencia construido a partir del estimador de Kaplan-Meier es el siguiente:

$$\widehat{S}_g(t) = 1 - \sum_{i=1}^{n} s_i \, \mathbb{K}\left(\frac{t - Z_{(i)}}{g}\right)$$

donde s_i es el salto que da el estimador de Kaplan-Meier en el punto $Z_{(i)}$, es decir,

$$s_i = \widehat{S}\big(Z_{(i-1)}\big) - \widehat{S}\big(Z_{(i)}\big),$$

y la estimación de la probabilidad de mora a horizonte b es

$$PD(t) = 1 - \frac{\widehat{S}_g(t+b)}{\widehat{S}_g(t)}$$

En la figura 1.6 puede verse cómo mejora la estimación suavizada y con corrección del efecto frontera de la PD con respecto a la estimación obtenida directamente a partir del estimador de Kaplan-Meier para esta muestra.

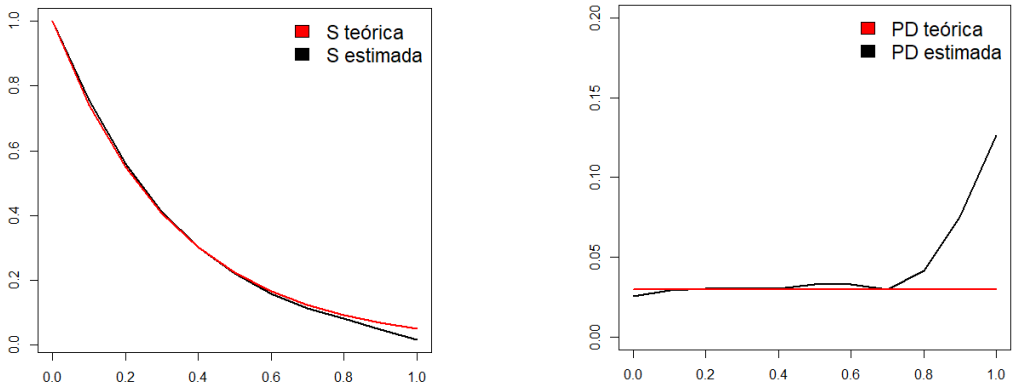


Figura 1.6: Estimación mediante Kaplan-Meier suavizado con corrección del efecto frontera para una muestra censurada de tamaño $n = 400$ de una $Exp(3)$ con tiempo de censura $Exp(2)$. Izquierda: Función de supervivencia (línea roja) y su estimación (línea negra). Derecha: Función de probabilidad de mora (línea roja) y su estimación empírica (línea negra).

En todo lo anterior se utilizó un parámetro ventana global denotado por g y que determina el grado de suavización en t. En posteriores páginas se discutirá el interés que pueda tener utilizar una ventana local, de manera que el parámetro de suavizado g varíe en función de la densidad de datos en torno al valor de t donde se pretende realizar la estimación.

El motivo de mostrar en primer lugar estos casos más sencillos, ambos sin tener en cuenta covariables, es que los inconvenientes que surgen en estos estimadores se mantienen al considerar el caso en el que la variable T, tiempo hasta el evento de interés, está censurada y depende de cierta covariable X. Por ello, los razonamientos que se seguirán al estudiar el comportamiento de los estimadores presentados en ese contexto seguirán el mismo hilo conductor; así como las propuestas de mejora, que se basarán en construir versiones suavizadas de los estimadores originales. De hecho, el estimador de Beran para la supervivencia condicional que se tratará en el capítulo 2 es la generalización del estimador límite-producto de Kaplan-Meier y sigue la misma idea para su construcción.

Desde la publicación en 1992 del artículo de Naraim (Naraim (1992)), se ha desarrollado abundante literatura donde el análisis de supervivencia es usado para enfrentarse a problemas en riesgo de crédito. Por citar algunos, en Hanson and Schuermann (2004) el análisis de supervivencia permite obtener intervalos de confianza para la probabilidad de mora; en Glennon and Nigro (2005) se estima la función de distribución del tiempo hasta la mora mediante un modelo *hazard* y en Allen and Rose (2006) se propone el estimador

de Kaplan-Meier para estimar la función de supervivencia del tiempo hasta el impago.

En Naraim (1992), la propuesta es un modelo de riesgos proporcionales de Cox para estimar la función de supervivencia condicional $S(t|x)$ y en Cao et al. (2009) obtienen a partir de él, y escribiendo la probabilidad de mora en términos de la supervivencia condicional como en (1.1), un estimador para la PD. Una segunda alternativa dada en Cao et al. (2009) consiste en asumir un modelo lineal generalizado para la distribución del tiempo de vida del crédito bajo censura: $P(T \leq t|X = x) = F_\theta(t|x) = g(\theta_0 + \theta_1 t + \theta_2 x)$, donde g es una función de enlace desconocida y $\theta = (\theta_0, \theta_1, \theta_2)$. La tercera alternativa de estos autores para estimar la probabilidad de mora es la obtenida a partir del estimador de Beran de la supervivencia condicional, precisamente el que se estudiará en esta memoria a lo largo del capítulo siguiente.

Capítulo 2

Estimador de la PD basado en el estimador de Beran

A lo largo de este capítulo se estudia el estimador de la probabilidad de mora construido a partir del estimador de Beran para la función de supervivencia condicional, la generalización del estimador límite-producto de la supervivencia al caso de una covariable continua. Se enuncian algunas de las propiedades asintóticas de las que goza el estimador obtenido y se observa su comportamiento al aplicarlo sobre muestras simuladas. Además, se propone una versión suavizada del estimador de Beran para la supervivencia y se analizan las estimaciones de la probabilidad de mora obtenidas a partir del mismo.

2.1. Estimador no paramétrico de la función de supervivencia

Sea una muestra aleatoria simple $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$ de la terna (X, Z, δ) donde X es la puntuación crediticia $Z = \text{mín}\{T, C\}$ y δ el indicador de no censura, $\delta = I_{\{T \leq C\}}$. Fijado el valor $X = x_0$ de la covariable, el estimador límite-producto generalizado propuesto por Beran (1981) para estimar la función de supervivencia condicional es el siguiente:

$$\widehat{S}_h^B(t|x_0) \;=\; \prod_{i=1}^{n} \left(1 - \frac{I_{\{Z_i \leq t,\, \delta_i = 1\}} w_{i,n}(x_0)}{1 - \sum_{j=1}^{n} I_{\{Z_j < Z_i\}} w_{n,j}(x_0)}\right) \tag{2.1}$$

donde

$$w_{i,n}(x_0) = \frac{K\big((x_0 - X_i)/h\big)}{\sum_{j=1}^{n} K\big((x_0 - X_j)/h\big)}$$

con $i = 1, ..., n$ y $h = h_n$, parámetro de suavizado para la covariable.

En Dabrowska (1989) e Iglesias-Pérez and González-Manteiga (1999) se prueba, bajo ciertas hipótesis, la consistencia fuerte uniforme y la normalidad asintótica del estimador $\widehat{S}_h^B(t|x_0)$. Además, se obtienen expresiones asintóticas del sesgo y la varianza de dicho estimador.

2.2. Estimador de la probabilidad de mora

Siendo $\widehat{S}_h^B(t|x_0)$, el estimador de Beran de la supervivencia, el estimador de la probabilidad de mora a horizonte b condicionado al valor x_0 de la covariable se denota por $\widehat{PD}_h^B(t|x_0)$ y se obtiene como sigue

$$\widehat{PD}_h^B(t|x_0) \;=\; 1 - \frac{\widehat{S}_h^B(t+b|x_0)}{\widehat{S}_h^B(t|x_0)} \tag{2.2}$$

A partir de las buenas propiedades del estimador $\widehat{S}_h^B(t|x_0)$, se puede demostrar la consistencia fuerte uniforme y la normalidad asintótica para el estimador $\widehat{PD}_h^B(t|x_0)$. Además, es posible obtener el sesgo y varianza asintóticos del mismo. Estos resultados fueron enunciados y demostrados en Cao et al. (2009) y Devia (2016). A continuación se muestran algunos de ellos y las hipótesis bajo las cuales tienen validez.

Sea la covariable X una variable aleatoria unidimensional con función de distribución absolutamente continua y densidad $f_X(x)$. Considérense las funciones de subdistribución:

$$\begin{aligned} H_1(t|x) &= P(Z \le t, \delta = 1|X = x) = \int_0^t \big(1 - G(u|x)\big) dF(u|x) \\ H_0(t|x) &= P(Z \le t, \delta = 0|X = x) = \int_0^t \big(1 - F(u|x)\big) dG(u|x) \end{aligned}$$

y sean

$$\begin{aligned} \underline{\tau}_H(x) &= \text{ínf}\{t \,:\, H(t|x) > 0\} \\ \overline{\tau}_H(x) &= \sup\{t \,:\, H(t|x) < 1\} \end{aligned}$$

Análogamente, se definen $\underline{\tau}_F(x)$, $\underline{\tau}_G(x)$, $\overline{\tau}_F(x)$ y $\overline{\tau}_G(x)$, verificando

$$\underline{\tau}_H(x) = \text{mín}\{\underline{\tau}_F(x), \underline{\tau}_G(x)\}, \quad \overline{\tau}_H(x) = \text{mín}\{\overline{\tau}_F(x), \overline{\tau}_G(x)\}.$$

Las siguientes hipótesis son necesarias para probar resultados asintóticos del estimador de la PD.

B1. El núcleo K es una función de densidad simétrica absolutamente continua y diferenciable con soporte compacto $\Omega_K \subset \mathbb{R}$.

B2. Sea Ω_X el soporte de la densidad f_X y sea $I = [x_1, x_2] \subseteq \Omega_X$ de manera que existen $\alpha, \beta, \delta > 0$ verificando $\alpha\delta \leq \beta\delta < 1$ y

$$\alpha \leq \text{ínf}\{f_X(x) : \ x \in I_\delta\} \leq \sup\{f_X(x) : \ x \in I_\delta\} \leq \beta$$

donde $I_\delta = [x_1 - \delta, x_2 + \delta]$. Además, f_X es dos veces diferenciable y las funciones $f'_X(x)$ y $f''_X(x)$ son continuas y acotadas en I_δ.

B3. Existe $\theta \in \mathbb{R}^+$ tal que

$$\text{ínf}\{1 - H(t|x) : \ t \in [0, \overline{\tau}_H], \ x \in I_\delta\} > \theta.$$

B4. Las funciones $H'(t|x) = \dfrac{\partial H(t|x)}{\partial x}$, $H''(t|x) = \dfrac{\partial^2 H(t|x)}{\partial x^2}$, $H'_1(t|x) = \dfrac{\partial H_1(t|x)}{\partial x}$ y $H''_1(t|x) = \dfrac{\partial^2 H_1(t|x)}{\partial x^2}$ existen, son continuas y acotadas en $(t, x) \in [0, \overline{\tau}_H] \times I_\delta$.

B5. Las funciones $\dot{H}(t|x) = \dfrac{\partial H(t|x)}{\partial t}$, $\ddot{H}(t|x) = \dfrac{\partial^2 H(t|x)}{\partial t^2}$, $\dot{H}_1(t|x) = \dfrac{\partial H_1(t|x)}{\partial t}$ y $\ddot{H}_1(t|x) = \dfrac{\partial^2 H_1(t|x)}{\partial t^2}$ existen, son continuas y acotadas en $(t, x) \in [0, \overline{\tau}_H] \times I_\delta$.

B6. Las funciones $\dot{H}'(t|x) = \dfrac{\partial^2 H(t|x)}{\partial t \partial x} = \dfrac{\partial^2 H(t|x)}{\partial x \partial t}$, $\dot{H}'_1(t|x) = \dfrac{\partial^2 H_1(t|x)}{\partial t \partial x} = \dfrac{\partial^2 H_1(t|x)}{\partial x \partial t}$ existen, son continuas y acotadas en $(t, x) \in [0, \overline{\tau}_H] \times I_\delta$.

B7. El parámetro de suavizado $h = h(n)$ verifica que

$$(\ln n)^3/nh \longrightarrow 0, \quad nh^5/\ln n = O(1)$$

cuando $n \to \infty$.

Las hipótesis sobre el núcleo K dadas en B1 garantizan que $c_K = \int K(t)^2 dt$ y $d_k = \int t^2 K(t) dt$ sean cantidades finitas, ya que, como K es continua y Ω_K compacto, se tiene que $K(\Omega_K)$ también es compacto; con lo cual, K está acotada en Ω_K. Además, por definición se tiene que $K(u) = 0 \; \forall \, u \notin \Omega_K$, por tanto, K está acotada en todo $\mathbb{R}$. Entonces,

$$c_K = \int K(t)^2 dt \leq \int_{\Omega_K} \|K\|_\infty^2 dt \leq \|K\|_\infty^2 \int_{\Omega_K} dt = \|K\|_\infty^2 \, \mu(\Omega_K) < \infty$$

$$d_K = \int t^2 K(t) dt \leq \int_{\Omega_K} t^2 \|K\|_\infty dt \leq \|K\|_\infty \int_{\Omega_K} t^2 dt < \infty$$

En Dabrowska (1989) se exige la condición B2 para obtener cotas exponenciales para las colas de la distribución del estimador $\widehat{S}_h^B(t|x)$ y, a partir de ellas, obtener la convergencia débil y fuerte del estimador. La hipótesis dada en B3 es necesaria para estimar las colas de las distribuciones $F(t|x)$, $G(t|x)$, $H(t|x)$ y $H_1(t|x)$. Las hipótesis B4, B5 y B6 junto con las condiciones impuestas a la función núcleo aseguran la insesgadez asintótica del estimador $\widehat{S}_h^B(t|x)$. Las condiciones sobre el parámetro ventana dadas en B7 permiten estimar la velocidad de convergencia del estimador $\widehat{S}_h^B(t|x)$ a la función de supervivencia condicional (para más detalles véase Dabrowska (1989))

Teorema 2.2.1. *Si se verifican las hipótesis B1-B7, entonces el error cuadrático integrado del estimador $\widehat{PD}_h^B(t|x)$ viene dado por la siguiente expresión*

$$ECM\big(\widehat{PD}_h^B(t|x)\big) \;=\; h^4\big(b(t|x)\big)^2 + \frac{1}{nh}\nu(t|x) + o\left(h^4 + \frac{1}{nh}\right)$$

donde

$$b(t|x) \;=\; -\frac{1}{2} d_k \big(1 - PD(t|x)\big) B_H(t, t+b|x)$$

$$\nu(t|x) \;=\; \frac{1}{f_X(x)} c_k \big(1 - PD(t|x)\big)^2 D_H(t, t+b|x)$$

$$\begin{aligned} B_H(t, t+b|x) \;=\; & \int_t^{t+b} \left(\ddot{H}(s|x) + 2\frac{f_X'}{f_X(x)} \dot{H}(s|x) \right) \frac{H_1(ds|x)}{\big(1 - H(t|x)\big)^2} \\ & + \int_t^{t+b} \frac{1}{1 - H(t|x)} \left(\dot{H}_1(ds|x) + 2\frac{f_X'(x)}{f_X(x)} H(ds|x) \right) \end{aligned}$$

$$D_H(t, t+b|x) \;=\; \int_t^{t+b} \frac{H_1(ds|x)}{\big(1 - H(t|x)\big)^2}$$

Corolario 2.2.1. *Sea el par* (t, x) *tal que* $0 < PD(t|x) < 1$*, bajo las condiciones B1-B7 con* $nh^5 \longrightarrow C \in (0, +\infty)$*, se verifica*

$$\sqrt{nh}\big(\widehat{PD}_h^B(t|x) - PD(t|x)\big) \xrightarrow{d} N\big(C^{1/2}b(t|x), \nu(t|x)\big)$$

donde $b(t|x)$ *y* $\nu(t|x)$ *siguen las expresiones dadas en el Teorema 2.2.1.*

Las demostraciones de estos resultados pueden consultarse en Cao et al. (2009).

2.3. Aplicación del estimador a datos simulados

En esta sección se estudiará el comportamiento del estimador de la probabilidad de mora construido a partir del estimador de Beran para la supervivencia condicionada. Se estimará por este medio la supervivencia y la PD de dos muestras simuladas para las que se conocerá la expresión teórica de ambas funciones. De este modo, se valorará la bondad del estimador en términos del error cuadrático integrado.

Este estimador fue utilizado previamente por Cao et al. (2009) y Devia (2016). No obstante, estos autores no corrigieron el efecto frontera, corrección que sí se aplica en esta memoria.

2.3.1. Generación de las muestras

A continuación se explican los supuestos bajo los cuales se simulan dos muestras censuradas de tiempos de vida de créditos personales con distribuciones exponencial y Weibull, que resultan distribuciones habituales para modelar el tiempo hasta la ocurrencia de un evento de interés.

Se obtienen muestras con un porcentaje de censura entre el 25 % y el 35 %. Es una censura considerablemente inferior a la que se esperaría en una muestra real de tiempos de vida de un crédito personal, pero esto facilita el análisis de los estimadores al ser más eficientes desde el punto de vista estadístico y reducirse el tiempo de computación.

Muestra 1

La primera muestra es simulada considerando tiempos de vida y censura exponenciales. En primer lugar, se considera una distribución beta de parámetros α y β para la covariable, $X \sim \mathcal{B}(\alpha, \beta)$ cuyo soporte sea $[0, 1]$. Se buscan valores de α y β tales que una parte importante de la densidad de X se encuentre en el subintervalo $[0.5, 1]$, pues esta variable refleja la solvencia de un cliente y es de esperar que la mayor parte de los clientes lo sean.

En la figura 2.1 se puede ver cómo cambia la función de densidad de una distribución beta al variar sus parámetros. En vista de esto se escogen los valores $\alpha = 7$ y $\beta = 3$ como parámetros de la distribución de la covariable.

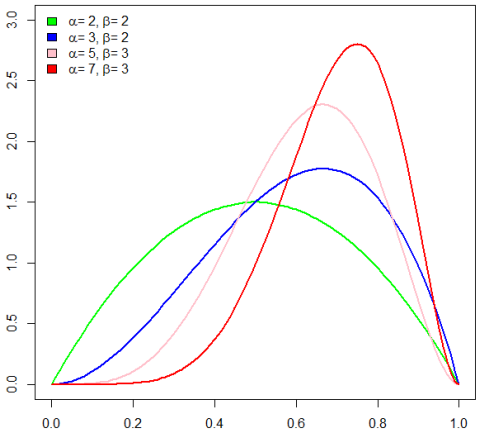


Figura 2.1: Densidades de $\mathcal{B}(\alpha, \beta)$

Más adelante se mostrarán las curvas de supervivencia condicional y de probabilidad de mora condicionadas a ciertos valores de la covariable; en concreto los cuantiles 0.25, 0.5, 0.75 de la distribución $\mathcal{B}(7, 3)$. Su valor se muestra en la tabla 2.1.

$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
0.609	0.713	0.804

Tabla 2.1: Cuartiles de la distribución $\mathcal{B}(7, 3)$

Siguiendo la idea de Van Keilegom et al. (2001), se considera una distribución exponencial

de parámetro $1/P(x)$ para el tiempo hasta la mora condicionado a X,

$$T|_{X=x} \sim Exp\big(1/P(x)\big),$$

con $P(x) = a_0 + a_1 x + a_2 x^2$ y una distribución exponencial con parámetro $1/Q(x)$ para el tiempo hasta la censura condicionado a la covariable,

$$C|_{X=x} \sim Exp\big(1/Q(x)\big),$$

con $Q(x) = b_0 + b_1 x + b_2 x^2$.

En estas condiciones, puede demostrarse que la probabilidad de censura condicionada a la covariable X viene dada por

$$P(\delta = 0|X = x) = H_0(\infty|X = x) = \int_0^{\infty} \big(1 - F(u|x)\big) dG(u|x) = \frac{Q(x)}{P(x) + Q(x)}$$

y la probabilidad de censura incondicional por

$$P(\delta = 0) = \int_{-\infty}^{+\infty} P(\delta = 0|X = x) f_X(x) dx.$$

Entonces, la probabilidad de censura de la muestra viene determinada por la elección de los coeficientes de los polinomios P y Q. Por otro lado, cuanto mayor es la solvencia de un cliente (valores más altos de la covariable), menos probable es que caiga en impago y, por tanto, la probabilidad de ser un dato censurado ha de ser mayor. Entonces, han de escogerse valores para los coeficientes de los polinomios de manera que valores altos de x arrojen valores de la probabilidad $P(\delta = 0|X = x)$ comprendidos en el rango de censura que se busca para la muestra final; en este caso, entorno al 30 %.

Los polinomios escogidos para generar la muestra 1 son los siguientes:

$$\begin{aligned} P(x) &= 0.5 + 0.5x + 5x^2 \\ Q(x) &= 6 + x + 2x^2 \end{aligned} \tag{2.3}$$

En la figura 2.2 se muestra la curva de probabilidad de censura condicional obtenida para los mismos.

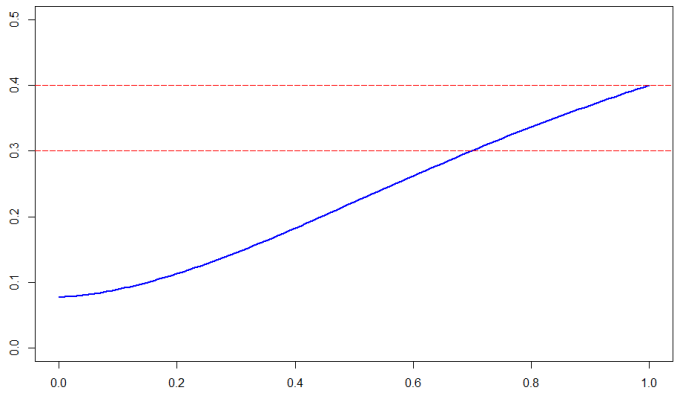


Figura 2.2: Probabilidad de censura condicional para $P(x) = 0.5 + 0.5x + 5x^2$ y $Q(x) = 6 + x + 2x^2$.

En efecto, para valores de la covariable superiores a 0.7, la probabilidad de censura condicional para los polinomios $P(x)$ y $Q(x)$ dados en (2.3) se encuentra entre 0.3 y 0.4. Además, la probabilidad de censura incondicional, $P(\delta = 0)$, aproximada para estos polinomios es 0.29.

Una vez elegidos los coeficientes adecuados de los polinomios $P(x)$ y $Q(x)$, se obtiene una muestra aleatoria simple de tamaño n del tiempo hasta la mora condicionado a $X = x$, $T_1, ..., T_n$, y del tiempo hasta la censura condicionado a $X = x$, $C_1, ..., C_n$. Finalmente, teniendo en cuenta que Z_i es el mínimo T_i y C_i, y δ_i es el indicador $I_{\{T_i \leq C_i\}}$ con $i = 1, ..., n$, se obtiene la terna $(X_i, Z_i, \delta_i)_{i=1}^n$. El tamaño muestral considerado es $n = 400$.

La proporción de censura de la muestra obtenida es de 0.34; 34 de cada 100 tiempos observados no se corresponden con un tiempo hasta la mora, si no con un tiempo hasta la censura. En la figura 2.3 se presenta el histograma de la muestra $Z_1, ..., Z_n$. Nótese que la mayor densidad de datos se encuentra en el intervalo de valores $[0, 6]$.

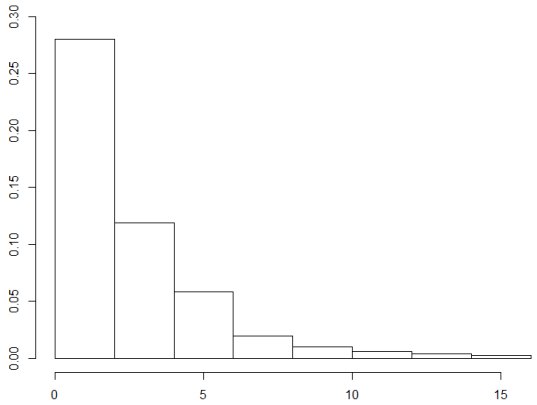


Figura 2.3: Histograma de la m.a.s. de Z en la muestra 1.

En este escenario es posible conocer la función de supervivencia condicional teórica y, en consecuencia, la probabilidad de mora. A continuación se muestran sus expresiones.

$$S_{T|X}(t|x) = e^{-t/P(x)}$$

$$PD(t|x) = 1 - \frac{S_{T|X}(t+b|x)}{S_{T|X}(t|x)} = 1 - e^{-b/P(x)}$$

En la figura 2.4 se muestran las gráficas de la función de supervivencia condicional y la probabilidad de mora teóricas condicionadas al cuantil $Q_{0.5}$ de la covariable, $x = 0.713$, obtenidas en una rejilla de tiempos en el intervalo $[0, 6]$ y a horizonte $b = 0.5$.

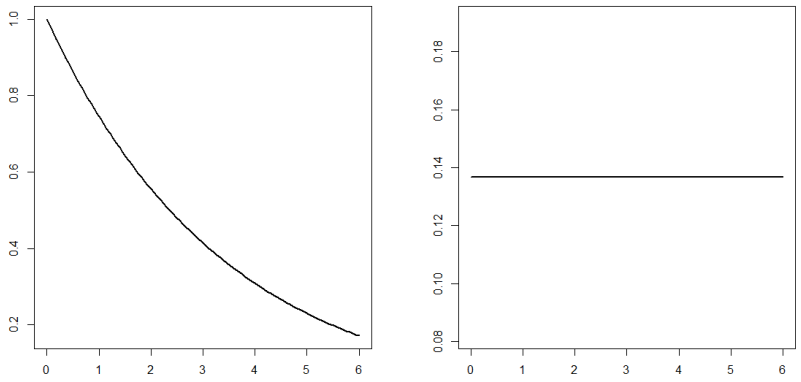


Figura 2.4: Izquierda: función de supervivencia condicional teórica para la muestra 1. Derecha: Probabilidad de mora teórica para la muestra 1.

Obsérvese que la función de probabilidad de mora para un tiempo de vida de crédito con distribución exponencial no depende del tiempo, $PD(t|x)$, es constante como función de t.

Muestra 2

La segunda muestra se simula considerando tiempos de vida y censura con distribuciones Weibull. Tal y como se hizo para la muestra 1, se considera una distribución beta de parámetros α y β para la covariable, $X \sim \mathcal{B}(\alpha, \beta)$, cuyo soporte es el intervalo $[0, 1]$. De nuevo atendiendo a la figura 2.1, se buscaron valores de α y β tales que una parte importante de la densidad de X se encontrase en el subintervalo $[0.5, 1]$. Se escogieron los valores $\alpha = 3$ y $\beta = 2$ para la distribución de X.

Se obtendrán las curvas de supervivencia condicional y de probabilidad de mora condicionadas a los cuartiles 0.25, 0.5 y 0.75 de la covariable, que en el caso de una distribución $\mathcal{B}(3, 2)$ toman los valores mostrados en la tabla 2.2.

$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
0.456	0.614	0.757

Tabla 2.2: Cuartiles de la distribución $\mathcal{B}(3, 2)$

Se considera una distribución Weibull de parámetros d y $C(x)^{-1/d}$ para el tiempo hasta la mora condicionado a X,

$$T|_{X=x} \sim \mathcal{W}(d, C(x)^{-1/d}),$$

con $d \in \mathbb{R}$ y $C(x) = c_0 + c_1 x + c_2 x^2$ y una distribución Weibull de parámetros d y $D(x)^{-1/d}$ para el tiempo hasta la censura condicionado a la covariable,

$$C|_{X=x} \sim \mathcal{W}(d, (D(x))^{-1/d}),$$

con $d \in \mathbb{R}$ y $D(x) = d_0 + d_1 x + d_2 x^2$.

De nuevo, se puede probar que la probabilidad de censura condicionada a la covariable X

que viene dada por

$$P(\delta = 0|X = x) = H_0(\infty|X = x) = \int_0^{\infty} \big(1 - F(u|x)\big) dG(u|x) = \frac{D(x)}{D(x) + C(x)}$$

y la probabilidad de censura incondicional

$$P(\delta = 0) = \int_{-\infty}^{+\infty} P(\delta = 0|X = x) f_X(x).$$

Se fija $d = 2$ y, siguiendo el mismo razonamiento que para la muestra 1, se escogen valores de los coeficientes de C y D de manera que la probabilidad de censura condicional esté aproximadamente entre 0.3 y 0.4 para valores altos de la covariable. Los polinomios elegidos con este fin son los siguientes:

$$\begin{aligned} C(x) &= 6 + x + 2x^2 \\ D(x) &= 0.5x + 5x^2 \end{aligned} \tag{2.4}$$

En la figura 2.5 se muestra la curva de probabilidad condicional para dichos polinomios.

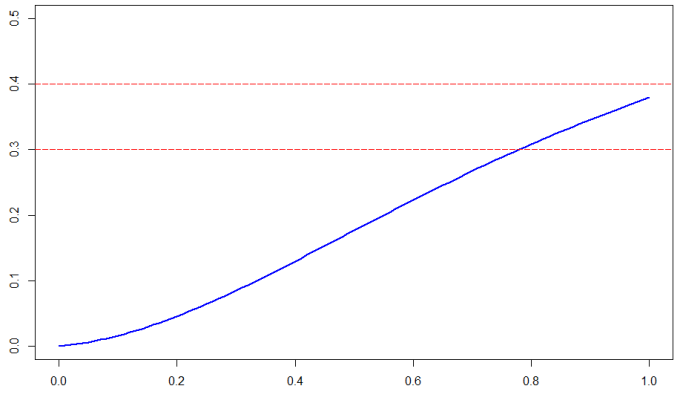


Figura 2.5: Probabilidad de censura condicional para $C(x) = 6 + x + 2x^2$ y $D(x) = 0.5x + 5x^2$.

En efecto, para los polinomios elegidos, valores de x superiores a 0.8 arrojan probabilidades de censura condicionales, $P(\delta = 0|X = x)$, entre 0.3 y 0.4 y la probabilidad de censura incondicional estimada para ellos es 0.38.

Una vez fijados estos coeficientes, las distribuciones para los tiempos de vida y censura quedan completamente determinadas y es posible obtener una muestra aleatoria simple

de tamaño n del tiempo hasta la mora condicionado a X, $T_1, ..., T_n$, y del tiempo hasta la censura condicionado a X, $C_1, ..., C_n$, a partir de dichas distribuciones. Para obtener la terna $(X_i, Z_i, \delta_i)_{i=1}^n$, basta tener en cuenta que $Z_i = \text{mín}\{T_i, C_i\}$ y $\delta_i = I_{\{T_{\leq} C_i\}}$. El tamaño muestral considerado es $n = 400$.

La proporción de censura de la muestra obtenida es de 0.25 y en la figura 2.3 se presenta el histograma de la muestra $Z_1, ..., Z_n$. Nótese que la mayor densidad de datos se encuentra en el intervalo de valores $[0, 0.5]$.

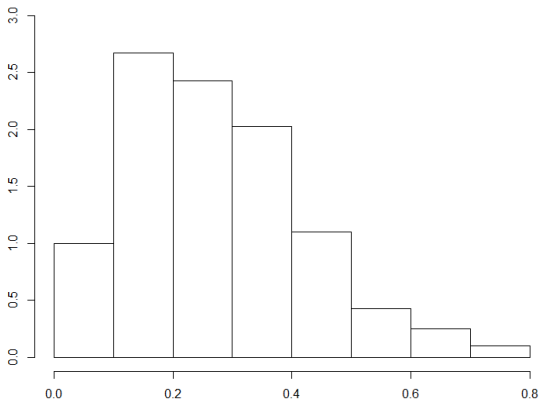


Figura 2.6: Histograma de la m.a.s. de Z en la muestra 2.

En este caso también se conocen las funciones de supervivencia condicional y probabilidad de mora teóricas:

$$S_{T|X}(t|x) = e^{-C(X)t^d}$$

$$PD(t|x) = 1 - \frac{S_{T|X}(t+b|x)}{S_{T|X}(t|x)} = 1 - \frac{e^{-C(X)(t+b)^d}}{e^{-C(X)t^d}}$$

En la figura 2.7 se muestran las gráficas de las funciones de supervivencia condicional y probabilidad de mora teóricas condicionadas al cuantil $Q_{0.5}$ de la covariable, $x = 0.614$, obtenidas en una rejilla de tiempos en el intervalo $[0, 0.5]$ y a horizonte $b = 0.05$.

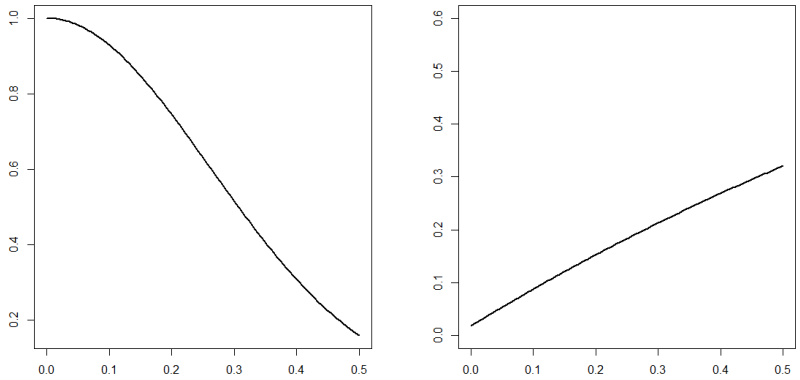


Figura 2.7: Izquierda: Función de supervivencia condicional teórica para la muestra 2. Derecha: Función de probabilidad de mora teórica para la muestra 2.

2.3.2. Resultados

En esta sección se utiliza el estimador de Beran de la supervivencia condicional para estimar tanto la supervivencia como la probabilidad de mora para las muestras obtenidas bajo los supuestos anteriores, las denominadas muestra 1 y muestra 2.

Para la suavización en la covariable se utiliza el núcleo de Epanechnikov que tiene soporte compacto y viene dado por

$$K(u) = \frac{3}{4}(1-u^2)I_{\{|u|<1\}},$$

Por otro lado, las gráficas mostradas en esta sección han sido obtenidas utilizando un valor del parámetro ventana h que, de ser posible, minimiza el error cuadrático integrado cometido en la estimación de la curva de probabilidad de mora. El error cuadrático integrado cometido en la estimación de la probabilidad de mora mediante el estimador $\widehat{PD}_h^B$ se define como sigue:

$$ECI = \int \left(\widehat{PD}_h^B(t|x) - PD(t|x)\right)dt$$

y el procedimiento seguido para seleccionar el parámetro de suavizado en la covariable es el siguiente: se obtiene, para diferentes valores de h, la curva estimada en una rejilla de tiempos en el intervalo correspondiente ($[0,6]$ para la muestra 1 y $[0,0.5]$ para la muestra 2) y se calcula para cada una de las estimaciones la raíz del error cuadrático integrado

(en adelante $RECI$). El valor de h elegido para obtener las curvas estimadas es el que arroja un menor error.

En la figura 2.8 se muestran las gráficas del $RECI$ en la estimación de la probabilidad de mora condicionada a los cuartiles $Q_{0.25}$, $Q_{0.5}$, $Q_{0.75}$ de la covariable X para la muestra 1. Nótese que, en este caso, el valor del parámetro de suavizado en la covariable mediante el cual se obtiene un menor error de estimación es también en el que la función $RECI(h)$ alcanza un mínimo.

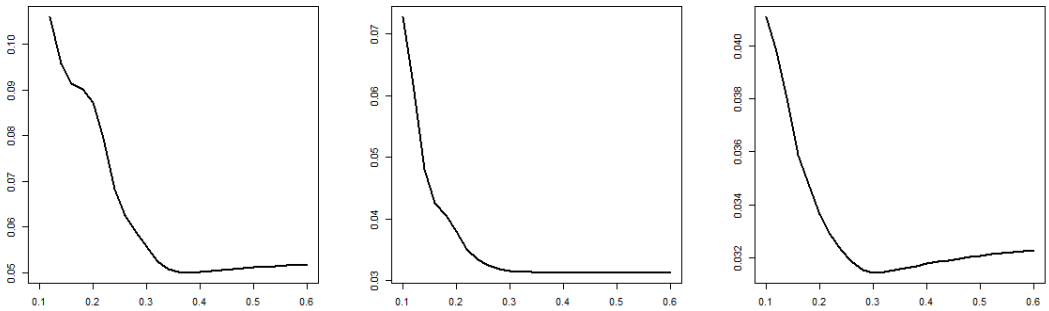


Figura 2.8: $RECI(h)$ para el estimador $\widehat{PD}_h^B(t|x)$ con $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$ en la muestra 1.

En la tabla 2.3 se muestran los valores del parámetro ventana que, para cada cuartil, arrojan un menor error cuadrático integrado y la raíz de dicho error. Cabe destacar que el error cometido en la estimación es menor para los cuantiles más grandes de la covariable.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
h	0.38	0.44	0.30
$RECI$	0.050	0.031	0.031

Tabla 2.3: Ventana óptima y $RECI$ obtenido mediante $\widehat{PD}_h^B(t|x)$ en la muestra 1.

En la figura 2.9 se muestran las estimaciones de la supervivencia condicional y la probabilidad de mora de la muestra 1 obtenidas para los valores del parámetro h dados en la tabla 2.3. Al igual que en los valores del $RECI$, en las gráficas se observa que el ajuste es mejor en el cuartil $Q_{0.75}$, pero es en general bueno para los tres cuartiles, especialmente para

la supervivencia. La estimación de la probabilidad de mora presenta mucha variabilidad, aunque las oscilaciones tienen lugar en torno a la verdadera curva de probabilidad.

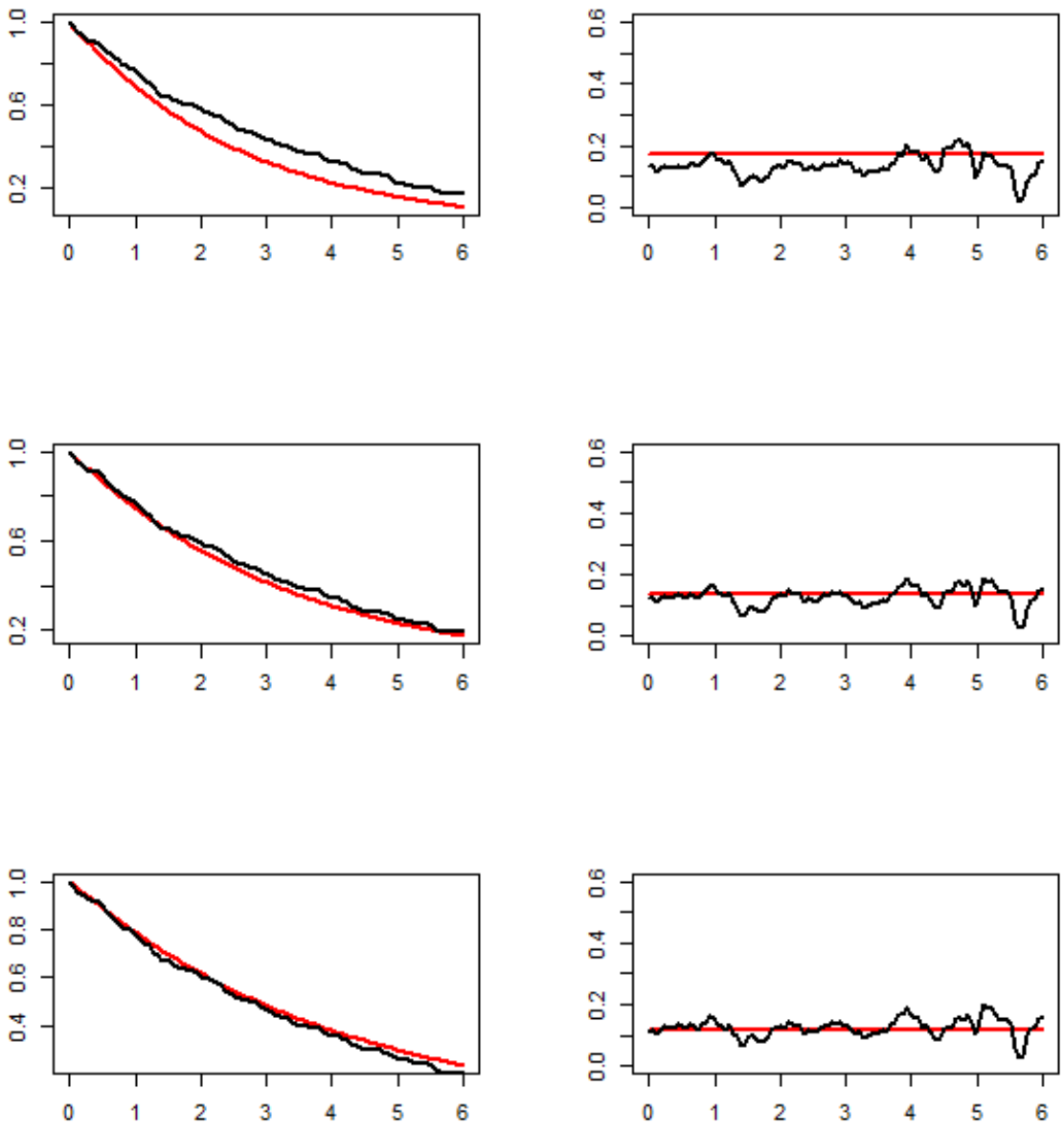


Figura 2.9: Izquierda: Supervivencia condicional (línea roja) y su estimación $\widehat{S}_h^B(t|x)$ (línea negra). Derecha: Probabilidad de mora (línea roja) y su estimación $\widehat{PD}_h^B(t|x)$ (línea negra). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 1.

Considerando ahora la muestra 2, en la figura 2.10 se muestran las gráficas de la raíz del

error cuadrático integrado cometido en la estimación de la probabilidad de mora obtenida mediante el estimador de la PD basado en el estimador de Beran para la supervivencia. En este caso, la función $RECI(h)$ es decreciente, sugiriendo que se tome como parámetro de suavizado el valor más grande del rango elegido.

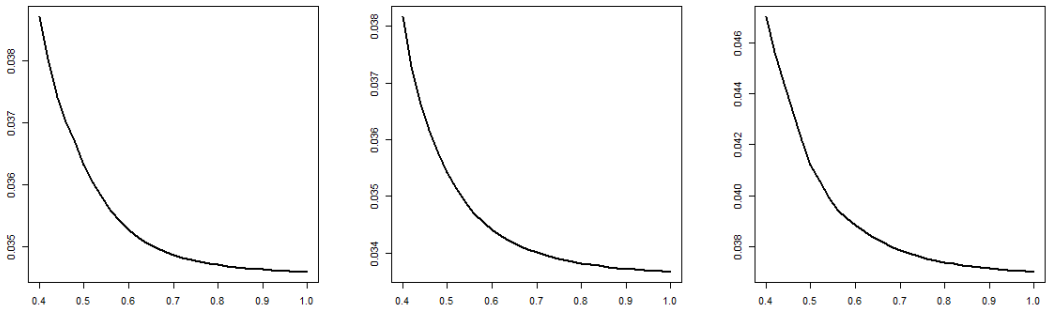


Figura 2.10: $RECI(h)$ para el estimador $\widehat{PD}_h^B(t|x)$ con $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$ en la muestra 2.

Tomando en la estimación de la PD para cada cuartil un valor grande del parámetro ventana, $h = 1$, se obtienen los valores de $RECI$ mostrados en la tabla 2.4. El error cometido es similar para los tres cuartiles de la covariable y puede decirse que no es notablemente grande.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
h	1	1	1
$RECI$	0.035	0.034	0.037

Tabla 2.4: Ventana óptima y $RECI$ obtenido mediante $\widehat{PD}_h^B(t|x)$ en la muestra 2.

En la figura 2.11 se muestran las gráficas de las curvas de supervivencia condicional y probabilidad de mora estimadas para dicho valor del parámetro ventana. A diferencia de lo que ocurría para la muestra 1, en la muestra 2 no se observan diferencias en la estimación para los diferentes cuartiles; en los tres casos, la supervivencia condicional estimada ajusta razonablemente bien la verdadera supervivencia y la curva de probabilidad de mora estimada oscila en torno a la verdadera.

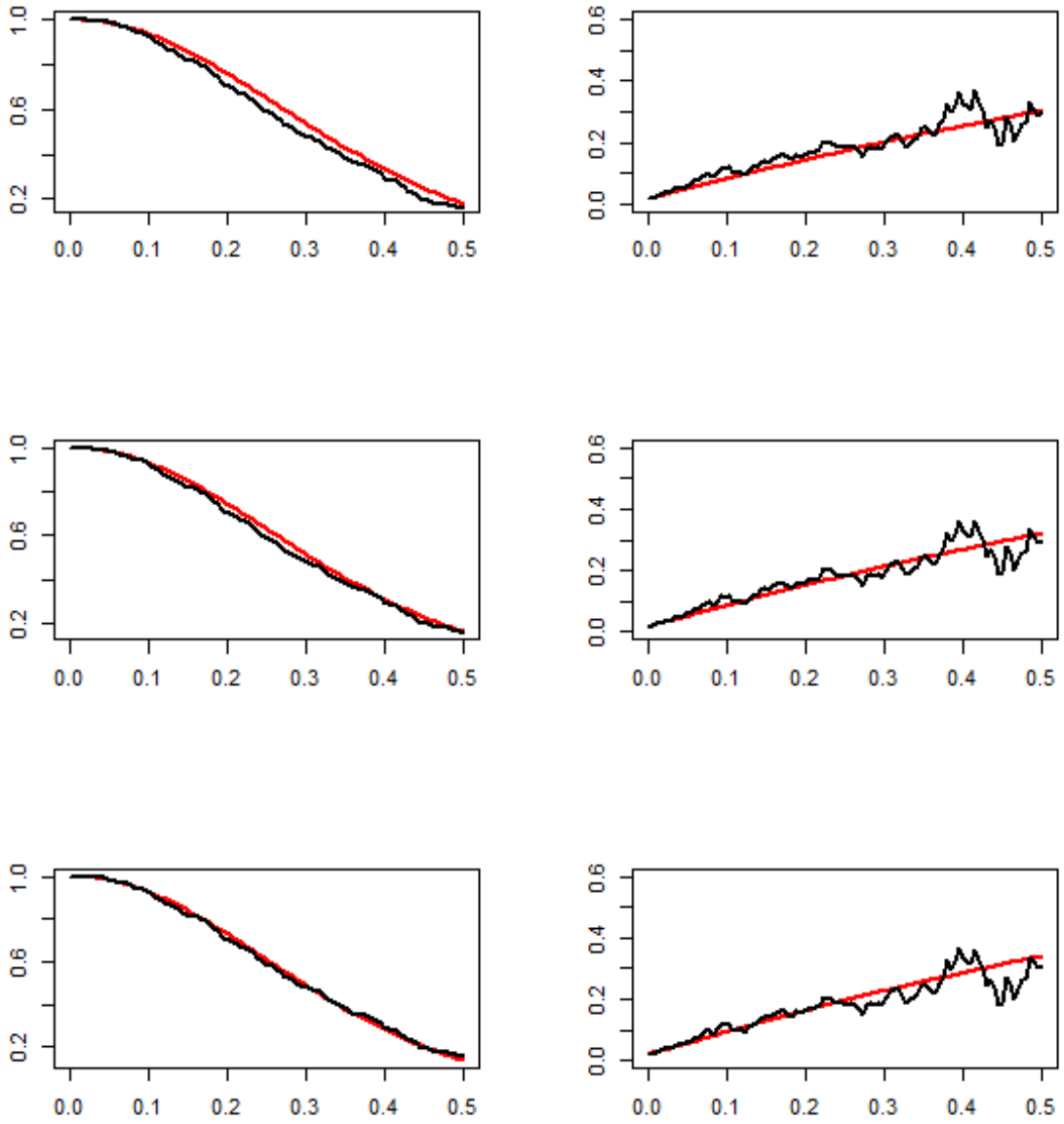


Figura 2.11: Izquierda: Supervivencia condicional (línea roja) y su estimación $\widehat{S}_h^B(t|x)$ (línea negra). Derecha: Probabilidad de mora (línea roja) y su estimación $\widehat{PD}_h^B(t|x)$ (línea negra). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 2.

Si bien es cierto que para ambas muestras, las estimaciones tanto de la supervivencia condicional como de la PD son razonables, se observa que la estimación de la probabilidad de mora presenta mucha variabilidad, especialmente en la cola derecha, donde la proporción de datos censurados es mayor. En el siguiente apartado se propone una suavización del estimador en la variable tiempo, que podría solucionarlo.

2.4. Estimador de Beran suavizado

El estimador de la probabilidad de mora construido a partir del estimador de Beran para la función de supervivencia arroja estimaciones razonables de las verdaderas curvas de probabilidad, pero éstas presentan excesiva variabilidad. La causa de esto podría ser el cociente entre supervivencias estimadas que ha de hacerse para obtener una estimación de la PD (véase (2.2)): aunque el error de predicción relativo de la supervivencia sea pequeño, se puede ver incrementado notablemente en la estimación de la PD, provocando la variabilidad observada en la sección anterior. La propuesta para solucionar esto es una versión suavizada del estimador dado en (2.2).

La idea intuitiva es estimar la supervivencia en un punto t condicionado al valor x_0 no mediante el valor que toma el estimador $\widehat{S}_h^B(t|x_0)$, si no mediante una ponderación de los valores que toma el estimador $\widehat{S}_h^B(\cdot|x_0)$ en puntos cercanos a t. De este modo, la estimación será suave. La expresión formal de esta idea es la siguiente:

$$\widehat{S}_{h,g}^B(t|x_0) \quad = \quad 1 - \sum_{i=1}^{n} s_i \mathbb{K}\left(\frac{t - Z_{(i)}}{g}\right)$$

donde $s_i = \widehat{S}_h^B(Z_{(i-1)}|x_0) - \widehat{S}_h^B(Z_{(i)}|x_0)$ siendo $\widehat{S}_h^B(\cdot|x_0)$ el estimador de Beran de la supervivencia y $\mathbb{K}(t)$ la función de distribución de un núcleo K, $\mathbb{K}(t) = \int_{-\infty}^{t} K(u)du$.

En efecto, la estimación $\widehat{S}_{h,g}^B(t|x_0)$ en el tiempo t es un promedio de los saltos que da la estimación con $\widehat{S}_h^B(t|x_0)$ en los puntos de $\{Z_i\}_{i=1}^n$ más cercanos a t. Esta noción de cercanía se concreta en el parámetro de suavizado g, que será, por el momento, un parámetro global independiente del punto t.

Suavizar en la variable tiempo el estimador de la supervivencia permite obtener estimaciones suaves también para la probabilidad de mora mediante la expresión:

$$\widehat{PD}_{h,g}^B(t|x_0) \quad = \quad 1 - \frac{\widehat{S}_{h,g}^B(t+b|x_0)}{\widehat{S}_{h,g}^B(t|x_0)}$$

En la siguiente sección se muestran los resultados obtenidos al aplicar este estimador sobre las muestras 1 y 2.

2.4.1. Aplicación del estimador suavizado a datos simulados

Para estudiar el comportamiento de $\widehat{PD}^{B}_{h,g}(t|x)$ como estimador de la probabilidad de mora, se aplica en esta sección sobre las muestras 1 y 2. El núcleo de suavizado para la covariable utilizado es, de nuevo, el de Epanechnikov. La ventana de suavizado h es fijada a los valores óptimos obtenidos en la sección 2.3.1 para cada caso. Para la suavización en la variable tiempo se considera una distribución de núcleo gaussiano y la ventana de suavizado, g, se escoge siguiendo el mismo criterio que se siguió anteriormente para h: fijado el valor óptimo de h, se obtiene la probabilidad de mora estimada por $\widehat{PD}^{B}_{h,g}(t|x)$ en una rejilla de valores del parámetro ventana g y se escoge el valor de g que arroja un menor error cuadrático integrado.

Se considera en primer lugar la muestra 1, en la figura 2.12 se muestra la gráfica de la raíz del error cuadrático integrado, $RECI$, como función de g en cada cuantil de la covariable. Se utiliza como ventana de suavizado g el valor en el que $RECI(h, g)$ alcanza el mínimo.

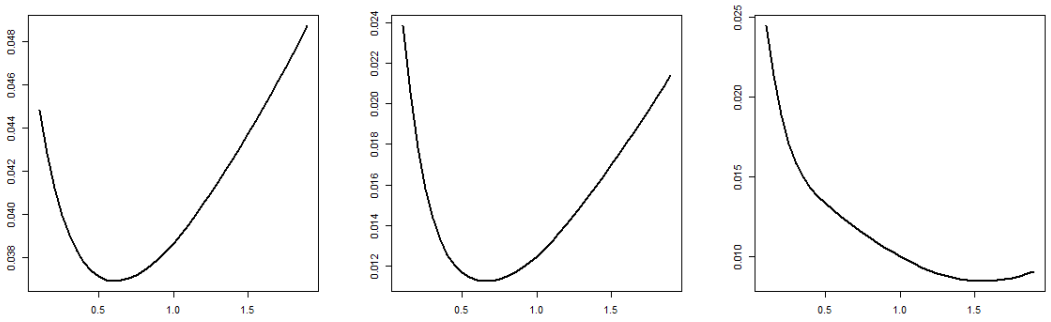


Figura 2.12: $RECI(h, g)$ para el estimador $\widehat{PD}^{B}_{h,g}(t|x)$ con $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$ en la muestra 1.

En la tabla 2.5 se muestran estos valores óptimos de la ventana g, el valor de la ventana h que se consideró y el valor de $RECI$ obtenido. Se observa que el error en la estimación se reduce considerablemente con respecto al error cometido en la estimación mediante el estimador basado en el de Beran sin suavización (ver tabla 2.3), especialmente para los cuantiles más grandes de la covariable.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
h	0.38	0.44	0.30
g	0.60	0.65	1.55
$RECI$	0.037	0.011	0.008

Tabla 2.5: Ventana óptima y $RECI$ obtenido mediante $\widehat{PD}^{B}_{h,g}(t|x)$ en la muestra 1.

En la figura 2.14 se muestran las estimaciones de la supervivencia condicional y la probabilidad de mora obtenidas para los valores de h, g y x dados en la tabla anterior. Aunque perdiendo rugosidad, la estimación para la supervivencia condicional es muy similar a la obtenida en la figura 2.9. Sin embargo, la mejora en las estimaciones de la probabilidad de mora con respecto a las mostradas en dicha figura son importantes.

En la tabla 2.6 se muestra la raíz del error cuadrático integrado cometido en la estimación de la probabilidad de mora en la muestra 1 mediante el estimador basado en el de Beran y mediante su versión suavizada, quedando patente que el segundo reduce el error cometido, mejorando, por tanto, la estimación de la PD.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
$RECI(h)$	0.050	0.031	0.031
$RECI(h,g)$	0.037	0.011	0.008

Tabla 2.6: Valores $RECI$ para $\widehat{PD}^{B}_{h}(t|x)$ y para $\widehat{PD}^{B}_{h,g}(t|x)$ en la muestra 1.

En la figura 2.15 se pueden ver las diferencias entre la estimación de la supervivencia y la PD mediante el estimador de Beran y mediante su versión suavizada para la muestra 1. Las diferencias en la estimación de la supervivencia condicional no son notables. Sin embargo, en la estimación de la probabilidad de mora se observa una mejora considerable, que se podía intuir por los valores del error cuadrático integrado dados en la tabla anterior.

Considerando ahora la muestra 2, en la figura 2.13 se pueden ver las gráficas del error cuadrático integrado como función de g para la estimación de la probabilidad de mora. En la tabla 2.7 se muestra el valor de la ventana óptima, en la que el $RECI(h,g)$ alcanza un mínimo, y la raíz del error cuadrático integrado cometido, para cada cuantil de la covariable. La reducción del error con respecto al cometido mediante el estimador sin suavización en la variable tiempo es notable.

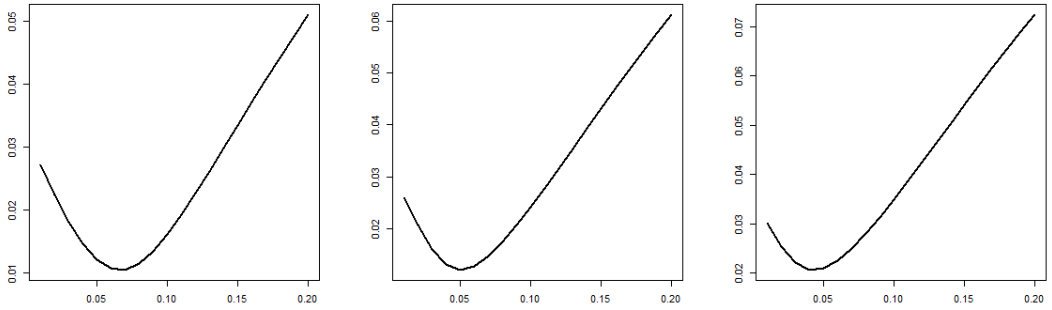


Figura 2.13: $RECI(h,g)$ para el estimador $\widehat{PD}^{B}_{h,g}(t|x)$ con $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$ en la muestra 2.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
h	1	1	1
g	0.07	0.05	0.04
$RECI$	0.010	0.012	0.021

Tabla 2.7: Ventana óptima y $RECI$ obtenido mediante $\widehat{PD}^{B}_{h,g}(t|x)$ en la muestra 2.

En la figura 2.16 se muestran las gráficas de la supervivencia condicional y la probabilidad de mora estimadas incorporando la suavización en la variable tiempo al estimador de Beran para cada cuantil de la covariable en la muestra 2. Debe destacarse el buen ajuste del estimador a la verdadera curva, tanto para la supervivencia condicional como para la probabilidad de mora.

En la tabla 2.8 se resumen los valores del $RECI$ cometido en la estimación de la probabilidad de mora mediante el estimador basado en Beran y mediante su versión suavizada

y la figura 2.17 muestra las diferencias entre uno y otro estimador para la muestra 2. Es claro que la estimación suavizada en la variable tiempo ajusta mejor la verdadera curva de probabilidad de mora, aunque las diferencias en la estimación de la supervivencia no son notables.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
$RECI(h)$	0.035	0.034	0.037
$RECI(h,g)$	0.010	0.012	0.021

Tabla 2.8: Valores $RECI$ para $\widehat{PD}_h^B(t|x)$ y para $\widehat{PD}_{h,g}^B(t|x)$ en la muestra 2.

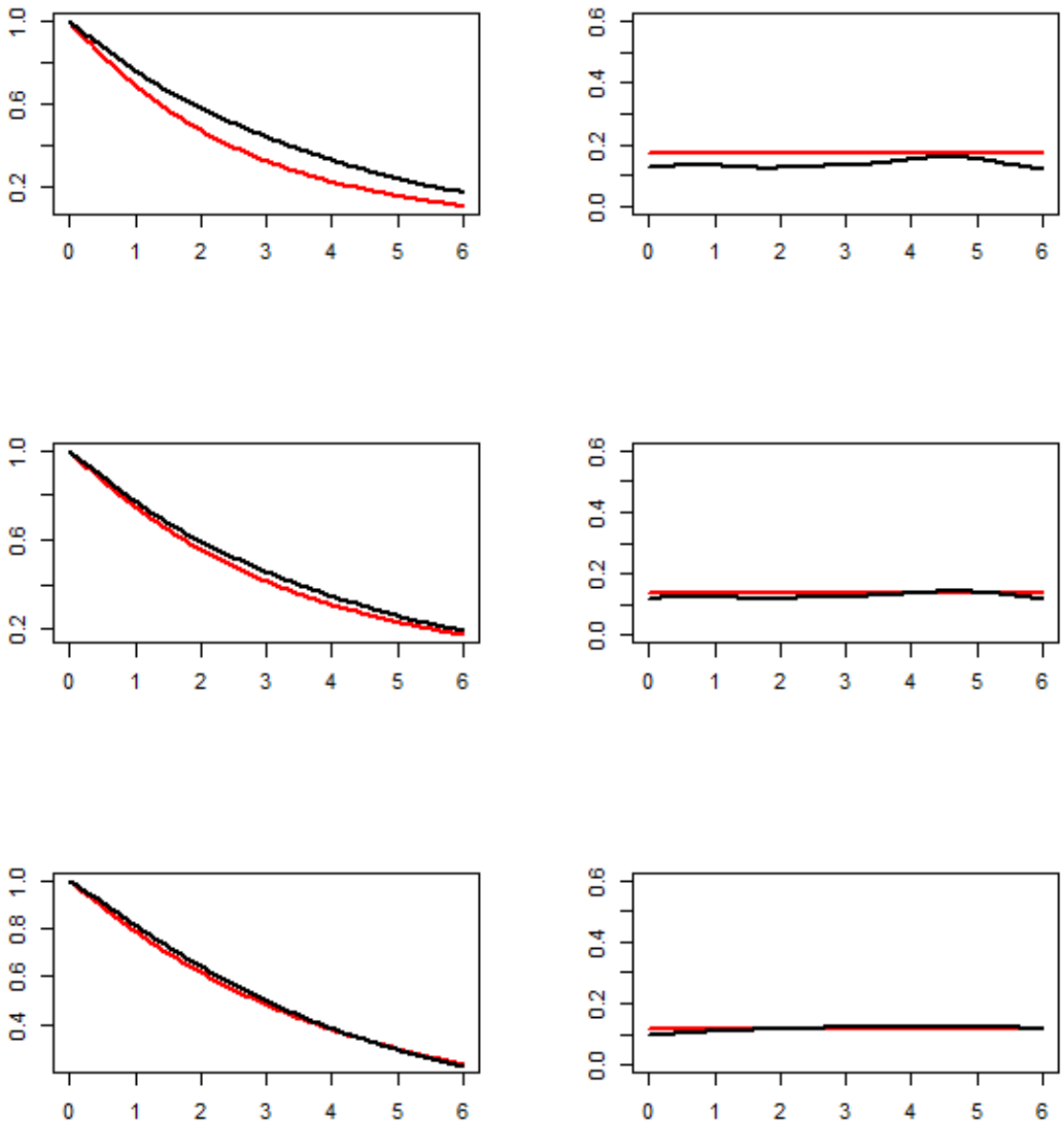


Figura 2.14: Izquierda: Supervivencia condicional (línea roja) y su estimación $\widehat{S}^B_{h,g}(t|x)$ (línea negra). Derecha: Probabilidad de mora (línea roja) y su estimación $\widehat{PD}^B_{h,g}(t|x)$ (línea negra). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 1.

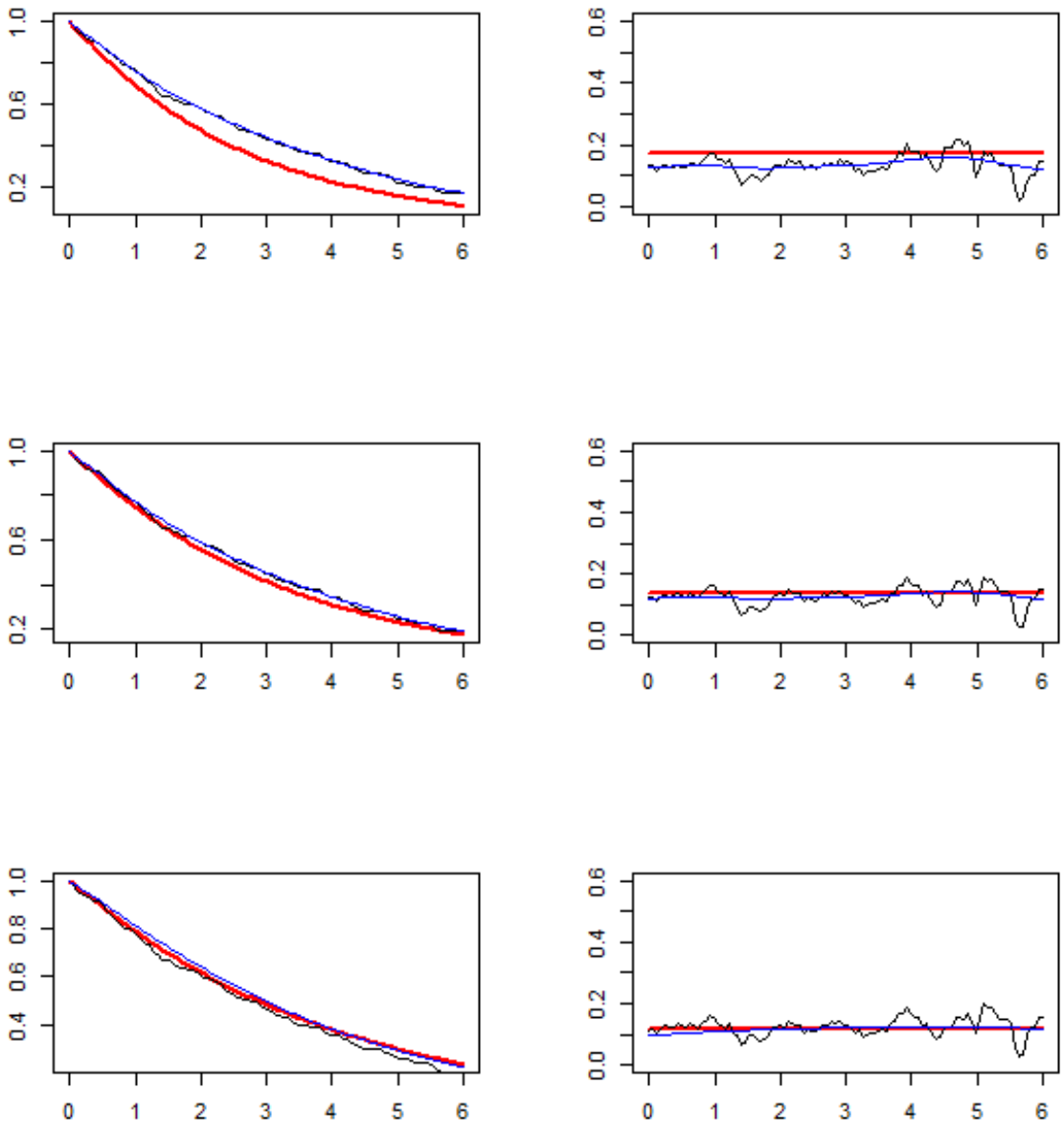


Figura 2.15: Izquierda: Supervivencia condicional (línea roja), su estimación por $\widehat{S}_h^B(t|x)$ (línea negra) y por $\widehat{S}_{h,g}^B(t|x)$ (línea azul). Derecha: Probabilidad de mora (línea roja), su estimación por $\widehat{PD}_h^B(t|x)$ (línea negra) y por $\widehat{PD}_{h,g}^B(t|x)$ (línea azul). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 1.

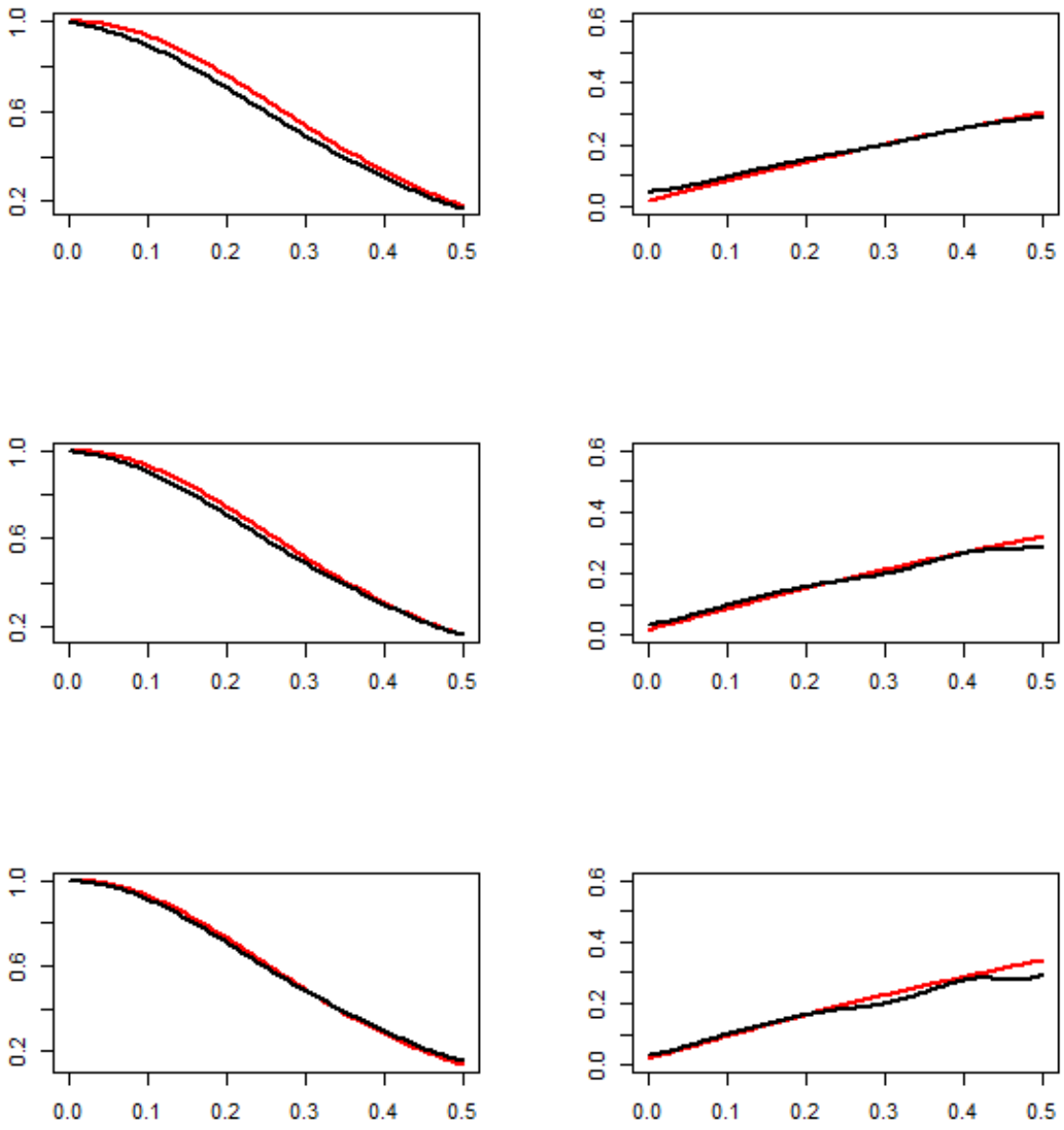


Figura 2.16: Izquierda: Supervivencia condicional (línea roja) y su estimación $\widehat{S}^B_{h,g}(t|x)$ (línea negra). Derecha: Probabilidad de mora (línea roja) y su estimación $\widehat{PD}^B_{h,g}(t|x)$ (línea negra). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 2.

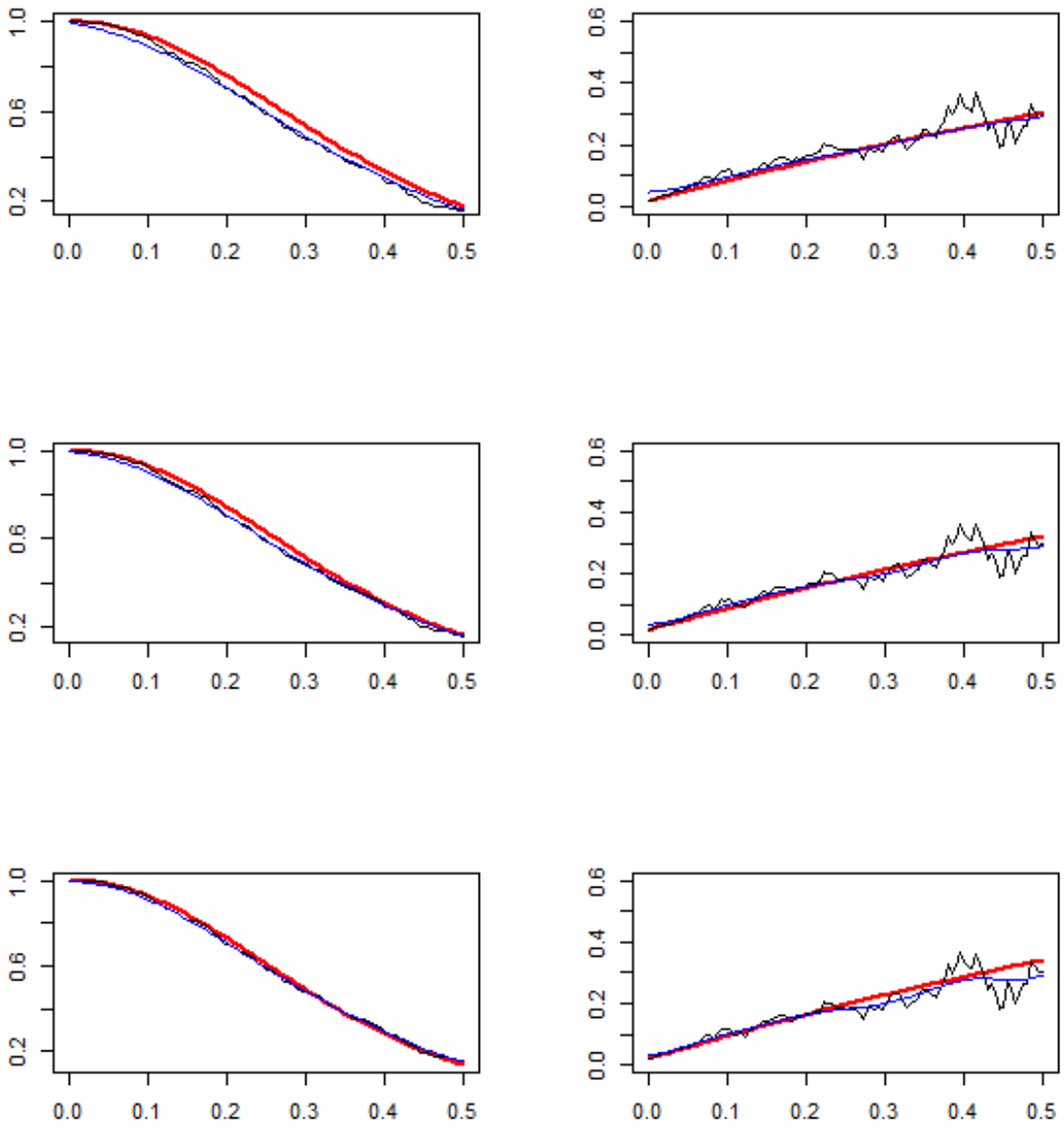


Figura 2.17: Izquierda: Supervivencia condicional (línea roja), su estimación por $\widehat{S}_h^B(t|x)$ (línea negra) y por $\widehat{S}_{h,g}^B(t|x)$ (línea azul). Derecha: Probabilidad de mora (línea roja), su estimación por $\widehat{PD}_h^B(t|x)$ (línea negra) y por $\widehat{PD}_{h,g}^B(t|x)$ (línea azul). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 2.

2.4.2. Discusión sobre la ventana

La ventana de suavización en la variable tiempo, g, se eligió minimizando el error cuadrático integrado y es una ventana global. Se plantea ahora el uso de un parámetro ventana local que atienda a la densidad de datos que hay en la muestra en torno al punto donde se pretende realizar la estimación. En esta sección se utiliza el método del k-vecino más próximo (o k-NN) con este propósito.

Este método, dado un valor de $k \in \{1, ..., n\}$, propone como parámetro ventana la distancia al k-ésimo vecino más próximo a t, o un múltiplo de ella. La elección del valor del entero k se hace siguiendo el mismo criterio que para la ventana g y el estimador resultante se denota por $\widehat{PD}^{B}_{h,k}(t|x)$.

En la tabla 2.9 se muestra la ventana h fijada, el valor óptimo de k y el valor del $RECI$ en la estimación para cada cuartil de la covariable de la muestra 1. En la tabla 2.10 se pueden ver los mismos datos para la muestra 2. En ambos casos es notable el aumento de la raíz de error cuadrático integrado con respecto al suavizado con ventana global.

Intuitivamente, se esperaba que el uso de una ventana local mejorase la estimación en la cola derecha, pues en dicha zona la proporción de datos no censurados es menor; sin embargo, la ventana local obtenida mediante el método k-NN no es más adecuada que la ventana global. Si bien es cierto que el error cometido en la estimación mediante Beran suavizado con ventana k-NN se reduce con respecto al estimador de Beran sin suavización en la variable tiempo, en las figuras 2.18 y 2.19, se constata que el comportamiento de este estimador no mejora el del estimador de Beran suavizado con ventana global.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
h	0.38	0.44	0.30
k	15	20	40
$RECI$	0.043	0.020	0.013

Tabla 2.9: Ventana óptima, k óptimo y $RECI$ obtenido mediante $\widehat{PD}^{B}_{h,k}(t|x)$ en la muestra 1.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
h	1	1	1
k	20	20	20
$RECI$	0.024	0.024	0.029

Tabla 2.10: Ventana óptima, k óptimo y $RECI$ obtenido mediante $\widehat{PD}^{B}_{h,k}(t|x)$ en la muestra 2.

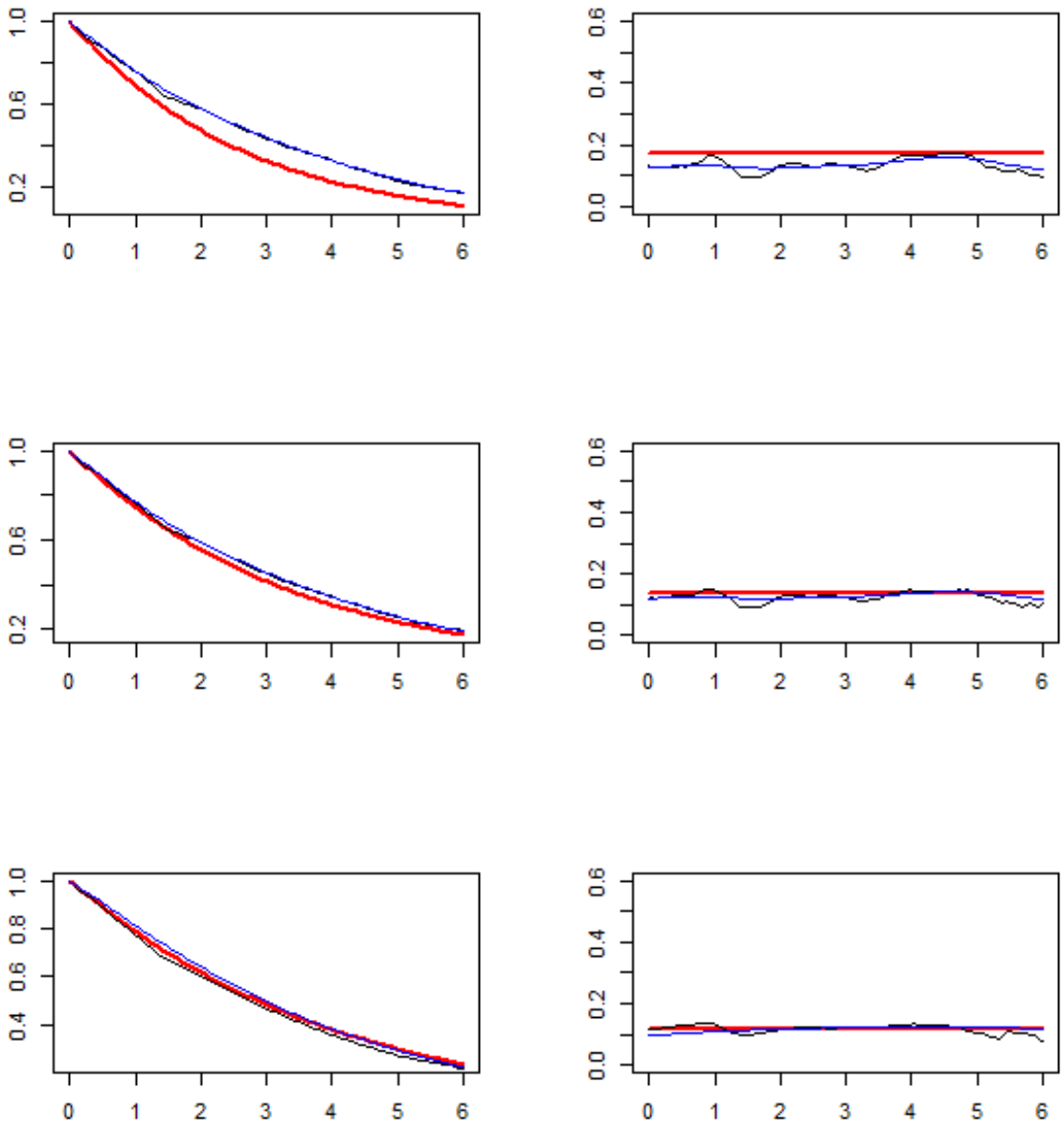


Figura 2.18: Izquierda: Supervivencia condicional (línea roja), su estimación por $\widehat{S}^B_{h,g}(t|x)$ (línea azul) y por $\widehat{S}^B_{h,k}(t|x)$ (línea negra). Derecha: Probabilidad de mora (línea roja), su estimación por $\widehat{PD}^B_{h,g}(t|x)$ (línea azul) y por $\widehat{PD}^B_{h,k}(t|x)$ (línea negra). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 1.

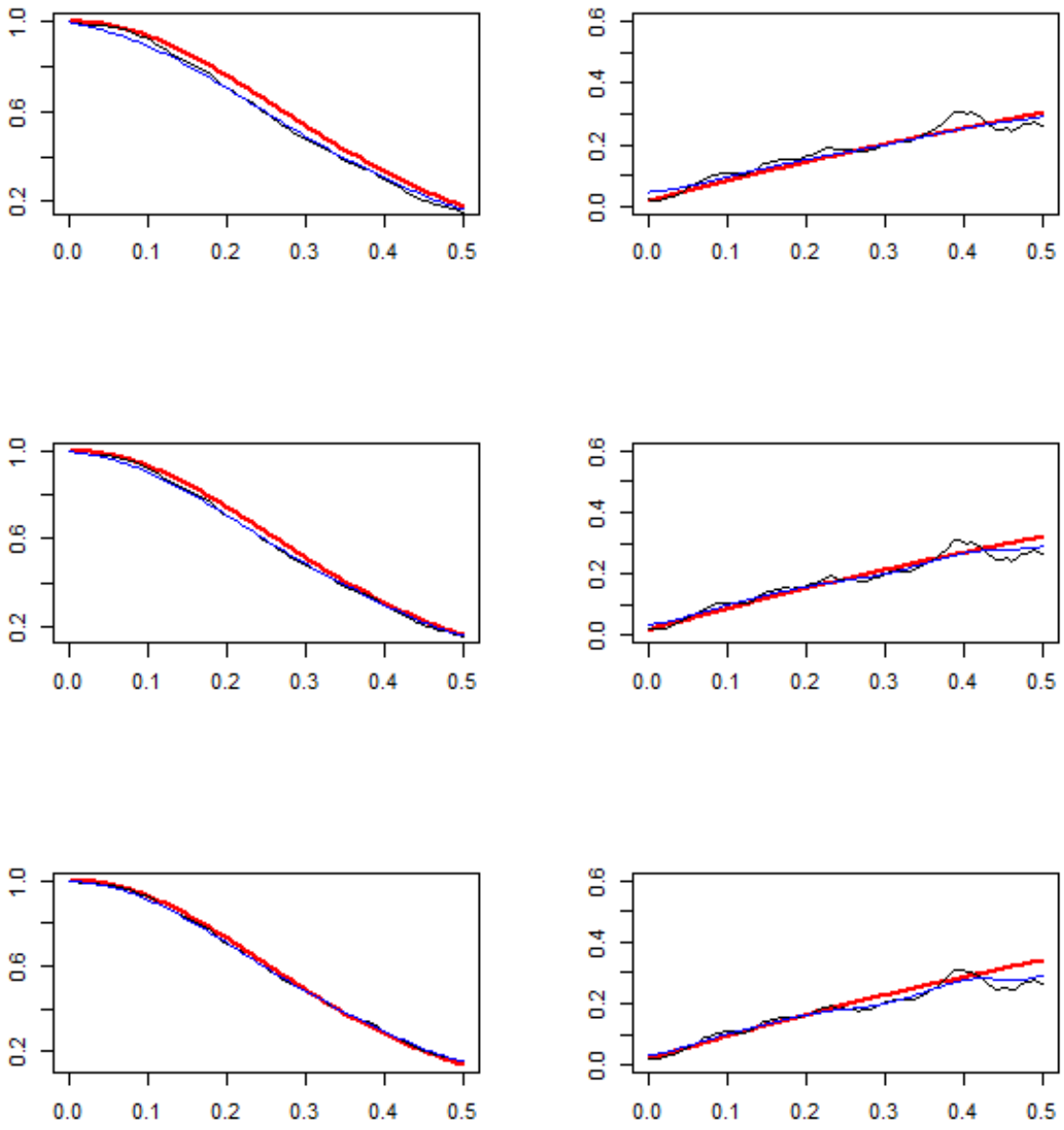


Figura 2.19: Izquierda: Supervivencia condicional (línea roja), su estimación por $\widehat{S}^B_{h,g}(t|x)$ (línea azul) y por $\widehat{S}^B_{h,k}(t|x)$ (línea negra). Derecha: Probabilidad de mora (línea roja), su estimación por $\widehat{PD}^B_{h,g}(t|x)$ (línea azul) y por $\widehat{PD}^B_{h,k}(t|x)$ (línea negra). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 2.

Capítulo 3

Estimador de la PD basado en el estimador de Cai

En este capítulo se construye un estimador de la probabilidad de mora a partir de un estimador de la función de supervivencia condicional obtenido mediante técnicas de regresión. Se trata del estimador propuesto en Cai (2003) que asume un modelo de regresión no paramétrico entre cierta transformación de la variable T, tiempo hasta la caída en mora, y la puntuación crediticia. De este modo, al estimar la función de regresión correspondiente se obtiene un estimador de la función de supervivencia condicional y a partir del mismo es posible obtener un estimador de la probabilidad de mora. Se demuestran propiedades de sesgo y varianza asintóticos para dicho estimador de la PD; se analizan las estimaciones obtenidas sobre muestras simuladas y finalmente, se propone una suavización en la variable tiempo del estimador de la función de supervivencia que disminuye el error de estimación.

3.1. Estimador no paramétrico de la regresión en presencia de censura

En Cai (2003) se propone un método para estimar la función de regresión de un modelo de regresión no paramétrico donde la variable respuesta está censurada. Siguiendo esta

idea se plantea el siguiente escenario.

Dada la variable T, tiempo hasta la entrada en mora, y X, puntuación crediticia, se considera una función arbitraria, Ψ, y la variable $V = \Psi(T)$ y se establece la siguiente relación no paramétrica de regresión:

$$V = \Psi(T) = m(X) + \varepsilon \tag{3.1}$$

donde $m(x) = E(V|X = x)$ es la función de regresión y ε es la variable error que verifica $E(\varepsilon|X) = 0$ y $Var(\varepsilon|X) = \sigma^2(X)$.

Según la elección de Ψ, en Masry (1996) se presentan tres casos de especial interés. En el primero $\Psi(t) = t$, tratándose entonces de una regresión ordinaria; en el segundo $\Psi(t) = t^2$ que corresponde con la estimación del momento de orden dos y el tercero, consiste en considerar, para un tiempo fijo t, la función indicadora $\Psi_t(T) = I_{\{T \leq t\}}$. En este último caso, la función de regresión es la función de distribución de $T_{|X=x}$. Sin embargo, en este trabajo interesa obtener estimadores de la función de supervivencia condicional $S(t|X = x)$ y para ello, se puede considerar $\Psi_t(T) = I_{\{T>t\}} = V_t$. Así,

$$m(x) = E(V_t|X = x) = E(I_{\{T>t\}}|X = x) = S(t|X = x)$$

y, por tanto, un estimador de la función de regresión $m(x)$ en estas condiciones será un estimador de la función de supervivencia, que permitirá estimar la PD.

Sea $\{(X_{[i]}, Z_{(i)}, \delta_{[i]})\}_{i=1}^n$ una muestra ordenada en función de los valores Z_i de la población (X, Z, δ), siendo $X_{[i]}$, $\delta_{[i]}$ sus concomitantes. A continuación se presenta el razonamiento seguido en Cai (2003) para proponer un estimador polinómico local de la función de regresión del modelo (3.1). Si se quisiese aproximar la función de regresión en un punto x_0, $m(x_0)$, mediante un polinomio de grado q en ausencia de censura, es decir, cuando $Z_i = Y_i$ para todo $i = 1, ..., n$, se buscarían valores de los coeficientes de dicho polinomio, $a_0, a_1, ..., a_q \in \mathbb{R}$, tales que minimicen la suma de cuadrados ponderados

$$MSE(a_0, a_1, ..., a_q) = \frac{1}{n} \sum_{i=1}^{n} \left(\Psi(Z_{(i)}) - \sum_{j=0}^{q} a_j (X_{[i]} - x_0)\right)^2 w_{i,h} \tag{3.2}$$

donde $w_{i,h}$ son los pesos de suavización de tipo Nadaraya-Watson definidos por $w_{i,h} = K_h(X_{[i]} - x_0) = K\left(\dfrac{X_{[i]} - x_0}{h}\right)$ siendo K una función núcleo y h un parámetro ventana que controla ese grado de suavización.

Sin embargo, en presencia de censura, deben incorporarse unos pesos que la tengan en cuenta. Estos pesos son los saltos del estimador de Kaplan-Meier y se definen como sigue:

$$W_{i,n} = \frac{\delta_{[i]}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{[j]}}$$

de manera que la función de regresión $m(x)$ sea aproximada por un polinomio de grado q, $m(x_0) \cong a_0 + a_1(X - x_0) + \cdots + a_q(X - x_0)^q$, cuyos coeficientes minimicen la suma de cuadrados ponderados dada por:

$$MSE(a_0, a_1, ..., a_q) = \frac{1}{n} \sum_{i=1}^{n} \left(\Psi(Z_{(i)}) - \sum_{j=0}^{q} a_j (X_{[i]} - x_0) \right)^2 w_{i,h} W_{i,n}$$

Puesto que se pretende estimar $m(x_0)$ y no sus derivadas, será suficiente considerar $q = 1$ (estimador local lineal) y minimizar:

$$MSE(a_0, a_1) = \sum_{i=1}^{n} \left(\Psi(Z_{(i)}) - a_0 - a_1 (X_{[i]} - x_0) \right)^2 w_{i,h} W_{i,n} \tag{3.3}$$

Nótese que en ausencia de censura, se tiene $W_{i,n} = 1/n$ y $Z_{(i)} = T_{(i)}$, para todo $i = 1, ..., n$, con lo que la suma de cuadrados en (3.3) es proporcional a la descrita en (3.2).

Entonces, el estimador lineal local ponderado de la regresión se puede expresar como

$$\widehat{m}(x_0) = \frac{S_{n,2}(x_0) T_{n,0}(x_0) - S_{n,1}(x_0) T_{n,1}(x_0)}{S_{n,2}(x_0) S_{n,0}(x_0) - S_{n,1}^2(x_0)} \tag{3.4}$$

donde

$$\begin{aligned} S_{n,l}(x_0) &= \sum_{i=1}^{n} (X_{[i]} - x_0)^l \, w_{i,h} W_{i,n} \\ T_{n,l}(x_0) &= \sum_{i=1}^{n} \Psi(Z_{(i)}) (X_{[i]} - x_0)^l \, w_{i,h} W_{i,n} \end{aligned}$$

para $l = 0, 1, 2$.

A continuación, se enumeran las condiciones que han de imponerse para que el estimador construido en las líneas anteriores goce de buenas propiedades.

C1. $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$ es una muestra aleatoria simple de (X, Z, δ).

C2. La densidad $f_X(x)$ es continua en $X = x_0$ y $f_X(x_0) > 0$.

C3. La segunda derivada de $m(x)$ existe y es continua en un entorno de x_0.

C4. La probabilidad de no censura es independiente de la covariable X dada la variable T, es decir, $P(T \leq C|T, X) = P(T \leq C|T)$.

C5. La función núcleo K es simétrica y tiene soporte compacto.

C6. Para cada $t \in \mathbb{R}$ y $k = 0, 1$, las funciones

$$a_k(x,t) = E\big[\Psi(T)^k\, I_{\{T>t\}}|X = x\big]$$

y

$$b_{2k}(x) = E\big[\Psi(T)^{2k}\{1 - G(T)\}^{-1}|X = x\big]$$

son continuas en un entorno de $x = x_0$.

C7. La función $E\big[|\Psi(T)|^{2+\alpha}/(1 - G(T))^{1+\alpha}|X = x\big]$ es continua en un entorno de $x = x_0$ para $\alpha > 0$.

Las hipótesis anteriores son estándar en la literatura, exceptuando C4. La premisa de independencia entre la probabilidad de no censura y la covariable, dada T quiere decir que, dado el tiempo hasta la mora, la covariable X no proporciona información sobre si tendrá lugar o no la censura. Esta hipótesis tan restrictiva también es impuesta en algunos modelos paramétricos y aunque se conjetura que no es necesaria, los resultados no han podido ser probados sin ella. Para más detalles véase Stute (1999).

Bajo las las condiciones anteriores, se pueden encontrar expresiones asintóticas para el sesgo y la varianza del estimador de la regresión propuesto por Cai. Para ello, en primer lugar, se enuncia el siguiente teorema sobre la distribución asintótica del mismo.

Teorema 3.1.1. *Si la función de distribución de la variable Z condicionada a $X = x_0$, $H(\cdot|X = x_0)$ es continua, $\overline{\tau}_H = \overline{\tau}_F$ y se verifican las hipótesis C1-C7, entonces,*

$$\sqrt{nh}\Big(\widehat{m}(x_0) - m(x_0) - \frac{h^2}{d_K} m''(x_0) + o_p(h^2)\Big) \longrightarrow \mathcal{N}\big(0,\, \Sigma(x_0)\big)$$

donde $\Sigma(x_0) = \dfrac{c_K b_2(x_0)}{f_X(x_0)}$.

El siguiente corolario, consecuencia directa de este teorema, proporciona expresiones del sesgo y la varianza asintóticos del estimador lineal local no paramétrico de la función de regresión.

Corolario 3.1.1. *En las condiciones del Teorema 3.1.1, el estimador dado en* (3.4) *es un estimador consistente de* $m(x_0)$ *y, además,*

$$\begin{aligned} Sesgo\big(\widehat{m}(x_0)\big) &= \frac{h^2}{2} d_K m''(x_0) + o(h^2) \\ Var\big(\widehat{m}(x_0)\big) &= \frac{1}{nh} \frac{c_K b_2(x_0)}{f_X(x_0)} \end{aligned}$$

El enunciado del Corolario 3.1.1 está en términos del caso que se desarrolla en Cai (2003) y se corresponde con la elección de $\Psi(T) = T$. Para el caso tratado aquí, $\Psi(T) = I_{\{T>t\}} = V_t$ será de interés disponer de las siguientes expresiones asintóticas para el sesgo y la varianza del estimador de la regresión.

Sean t, b y x_0 cualesquiera pero fijos y sea $\widehat{m}_t(x_0)$ el estimador de la regresión en x_0 para la variable respuesta V_t, entonces:

$$\begin{aligned} Sesgo\left(\widehat{m}_t(x_0)\right) &= Sesgo\left(\widehat{S}(t|X=x_0)\right) = A_t(x_0)h^2 + o(h^2) \\ Var\left(\widehat{m}_t(x_0)\right) &= Var\left(\widehat{S}(t|X=x_0)\right) = B_t(x_0)\frac{1}{nh} + o\left(\frac{1}{nh}\right) \\ Cov\left(\widehat{m}_t(x_0),\, \widehat{m}_{t+b}(x_0)\right) &= C_{t,t+b}(x_0)\frac{1}{nh} + o\left(\frac{1}{nh}\right) \end{aligned} \tag{3.5}$$

donde

$$A_t(x_0) = \frac{1}{2} d_K S''(t|x_0)$$

$$B_t(x_0) = \frac{c_K \beta_t(x_0)}{f_X(x_0)}$$

$$\beta_t(x_0) = E\Big(V_t^2\{1 - G(T)\}^{-1}|X = x_0\Big)$$

$$C_{t,t+b}(x_0) = \frac{c_K \gamma_{t,t+b}(x_0)}{f_X(x_0)}$$

$$\gamma_{t,t+b}(x_0) = E\Big(V_t V_{t+b}\{1 - G(T)\}^{-1}|X = x_0\Big)$$

3.2. Estimador de la probabilidad de mora

Se utilizará este estimador no paramétrico de la función de regresión (o de la función de supervivencia) en el punto x_0 para estimar la probabilidad de mora a horizonte b

condicionada al valor de la covariable $X = x_0$ del siguiente modo:

$$\widehat{PD}_h^C(t|x_0) = 1 - \frac{\widehat{S}_h^C(t+b|x_0)}{\widehat{S}_h^C(t|x_0)} = 1 - \frac{\widehat{m}_{t+b}(x_0)}{\widehat{m}_t(x_0)} \tag{3.6}$$

Es de interés disponer de expresiones asintóticas del sesgo y la varianza de este estimador de la PD y éstas se obtendrán a partir del sesgo y la varianza asintóticos de $\widehat{m}_t(x_0)$ dados en (3.5). A continuación se enuncia este resultado.

Teorema 3.2.1. *En las condiciones del teorema 3.1.1, para el estimador de la probabilidad de mora dado en* (3.6) *se tiene*

$$\begin{aligned} Sesgo\left(\widehat{PD}(t|x)\right) &= \frac{1}{S(t|x_0)}\Big(A_{t+b}(x_0) - \frac{S(t+b|x_0)}{S(t|x_0)}A_t(x_0)\Big)h^2 + \left(o(h^2) + O\left(\frac{1}{nh}\right)\right) \\ Var\left(\widehat{PD}(t|x)\right) &= \frac{1}{\left(S(t|x_0)\right)^2}\Bigg(B_{t+b}(x_0) - 2\frac{S(t+b|x_0)}{S(t|x_0)}C_{t,t+b}(x_0) \\ &\quad + \left(\frac{S(t+b|x_0)}{S(t|x_0)}\right)^2 B_t(x_0)\Bigg)\frac{1}{nh} + o\left(\frac{1}{nh}\right) \end{aligned}$$

Demostración.
Se denota $PD(t|x_0) = 1 - \dfrac{S(t+b|x_0)}{S(t|x_0)} = 1 - \varphi$, $P = S(t+b|x_0)$ y $Q = S(t|x_0)$, con lo que $\varphi = P/Q$. Del mismo modo, se denota $\widehat{P} = \widehat{S}(t+b|x_0)$, $\widehat{Q} = \widehat{S}(t|x_0)$ y $\widehat{\varphi} = \widehat{P}/\widehat{Q}$.

Considérese la igualdad:

$$\frac{1}{z} = 1 - (z-1) + ... + (-1)^p(z-1)^p + (-1)^{(p+1)}\frac{(z-1)^{(p+1)}}{z} \tag{3.7}$$

que será utilizada en varios puntos de la demostración.

Se comenzará buscando una expresión asintótica para el sesgo del estimador $\widehat{PD}(t|x_0)$.

Para $p = 1$ en (3.7) se tiene

$$\frac{1}{z} = 1 - (z-1) + \frac{(z-1)^2}{z}$$

y haciendo $z = \dfrac{\widehat{Q}}{E(\widehat{Q})}$ se obtiene

$$\begin{aligned} \widehat{\varphi} &= \frac{\widehat{P}}{\widehat{Q}} = \frac{\widehat{P}}{E(\widehat{Q})}\frac{E(\widehat{Q})}{\widehat{Q}} = \frac{\widehat{P}}{E(\widehat{Q})}\left(1 - \left(\frac{\widehat{Q}}{E(\widehat{Q})} - 1\right) + \frac{E(\widehat{Q})}{\widehat{Q}}\left(\frac{\widehat{Q}}{E(\widehat{Q})} - 1\right)^2\right) = \\ &= \frac{\widehat{P}}{E(\widehat{Q})} - \frac{\widehat{P}\big(\widehat{Q} - E(\widehat{Q})\big)}{E(\widehat{Q})^2} + \frac{\widehat{P}}{\widehat{Q}}\frac{\big(\widehat{Q} - E(\widehat{Q})\big)^2}{E(\widehat{Q})^2} \end{aligned}$$

Tomando esperanzas,

$$\begin{aligned} E(\widehat{\varphi}) &= \frac{E(\widehat{P})}{E(\widehat{Q})} - \frac{E\big[\widehat{P}\big(\widehat{Q} - E(\widehat{Q})\big)\big]}{E(\widehat{Q})^2} + \frac{E\Big[\frac{\widehat{P}}{\widehat{Q}}\big(\widehat{Q} - E(\widehat{Q})\big)^2\Big]}{E(\widehat{Q})^2} = \\ &= \frac{E(\widehat{P})}{E(\widehat{Q})} - \frac{Cov(\widehat{P},\widehat{Q})}{E(\widehat{Q})^2} + \frac{E\Big[\frac{\widehat{P}}{\widehat{Q}}\big(\widehat{Q} - E(\widehat{Q})\big)^2\Big]}{E(\widehat{Q})^2} \end{aligned}$$

Entonces,

$$E(\widehat{\varphi}) = A_1 + A_2 + A_3 \tag{3.8}$$

donde

$$\begin{aligned} A_1 &= \frac{E(\widehat{P})}{E(\widehat{Q})} \\ A_2 &= -\frac{Cov(\widehat{P},\widehat{Q})}{E(\widehat{Q})^2} \\ A_3 &= \frac{E\Big[\frac{\widehat{P}}{\widehat{Q}}\big(\widehat{Q} - E(\widehat{Q})\big)^2\Big]}{E(\widehat{Q})^2} \end{aligned}$$

Utilizando las expresiones del sesgo y varianza de $\widehat{m}_t(x_0)$ dadas en (3.5), se tiene lo siguiente:

$$E(\widehat{P}) = E\big(\widehat{S}(t+b|x_0)\big) = P + A_{t+b}(x_0)h^2 + o(h^2)$$

$$E(\widehat{Q}) = E\big(\widehat{S}(t|x_0)\big) = Q + A_t(x_0)h^2 + o(h^2)$$

$$Var(\widehat{Q}) = Var\big(\widehat{S}(t|x_0)\big) = B_t(x_0)\frac{1}{nh} + o\left(\frac{1}{nh}\right)$$

$$Cov(\widehat{P},\widehat{Q}) = Cov\big(\widehat{S}(t+b|x_0),\ \widehat{S}(t|x_0)\big) = C_{t,t+b}(x_0)\frac{1}{nh} + o\left(\frac{1}{nh}\right)$$

lo cual se utiliza a continuación para estudiar cada uno de los sumandos en (3.8):

$$
\begin{aligned}
A_1 &= \frac{E(\widehat{P})}{E(\widehat{Q})} = \frac{P + A_{t+b}(x_0)h^2 + o(h^2)}{Q + A_t(x_0)h^2 + o(h^2)} \simeq \frac{P + A_{t+b}(x_0)h^2}{Q + A_t(x_0)h^2} \cdot \frac{Q - A_t(x_0)h^2}{Q - A_t(x_0)h^2} \\
&= \frac{PQ + A_{t+b}(x_0)Qh^2 - PA_t(x_0)h^2 - A_{t+b}(x_0)A_t(x_0)h^4}{Q^2 - (A_t(x_0))^2h^4} \\
&\simeq \frac{P}{Q} + \left(\frac{A_{t+b}(x_0)}{Q} - \frac{P}{Q}\frac{A_t(x_0)}{Q}\right)h^2 + o(h^2) \\
&= \varphi + \left(\frac{A_{t+b}(x_0)}{Q} - \varphi\frac{A_t(x_0)}{Q}\right)h^2 + o(h^2) \\
A_2 &= -\frac{Cov(\widehat{P},\widehat{Q})}{E(\widehat{Q})^2} = -\frac{C_{t,t+b}(x_0)\frac{1}{nh} + o\left(\frac{1}{nh}\right)}{\left(Q + A_t(x_0)h^2 + o(h^2)\right)^2} \simeq -\frac{C_{t,t+b}(x_0)\frac{1}{nh}}{(Q + A_t(x_0)h^2)^2} \\
&\simeq -C_{t,t+b}(x_0)\frac{1}{Q^2}\frac{1}{nh} = O\left(\frac{1}{nh}\right) \\
A_3 &= \frac{E\left[\frac{\widehat{P}}{\widehat{Q}}\left(\widehat{Q} - E(\widehat{Q})\right)^2\right]}{E(\widehat{Q})^2} \leq \frac{Var(\widehat{Q})}{E(\widehat{Q})^2} = \frac{B_t(x_0)\frac{1}{nh} + o\left(\frac{1}{nh}\right)}{Q + A_t(x_0)h^2 + o(h^2)} = O\left(\frac{1}{nh}\right)
\end{aligned}
$$

Así, se obtiene finalmente el sesgo del estimador $\widehat{PD}(t|x_0)$:

$$
\begin{aligned}
Sesgo\left(\widehat{PD}(t|x_0)\right) &= E\left(\widehat{PD}(t|x_0) - PD(t|x_0)\right) = E(1-\widehat{\varphi}) - (1-\varphi) \\
&= \varphi - (A_1 + A_2 + A_3) \\
&= \frac{1}{Q}\left(\frac{P}{Q}A_t(x_0) - A_{t+b}(x_0)\right)h^2 + o(h^4) + O\left(\frac{1}{nh}\right)
\end{aligned}
$$

A continuación, se tratará de encontrar una expresión de la varianza. Para ello, en primer lugar se recurrirá a la igualdad dada en (3.7) con $p = 3$ y $z = \dfrac{\widehat{Q}^2}{E(\widehat{Q})^2}$:

$$
\begin{aligned}
\frac{E(\widehat{Q})^2}{\widehat{Q}^2} &= 1 + \sum_{i=1}^{3}(-1)^i\left(\frac{\widehat{Q}^2}{E(\widehat{Q})^2} - 1\right)^i + \frac{(\widehat{Q}^2/E(\widehat{Q})^2 - 1)^4}{\widehat{Q}^2/E(\widehat{Q})^2} \\
&= 1 + \sum_{i=1}^{3}(-1)^i\left(\frac{\widehat{Q}^2 - E(\widehat{Q})^2}{E(\widehat{Q})^2}\right)^i + \left(\frac{\widehat{Q}^2 - E(\widehat{Q})^2}{E(\widehat{Q})^2}\right)^4\frac{E(\widehat{Q})^2}{\widehat{Q}^2}
\end{aligned} \tag{3.9}
$$

Nótese que:

$$
\begin{aligned}
(\widehat{Q} - E(\widehat{Q}))^2 &= \widehat{Q}^2 - 2\widehat{Q}E(\widehat{Q}) + E(\widehat{Q})^2 + E(\widehat{Q})^2 - E(\widehat{Q})^2 \\
&= \widehat{Q}^2 - 2\widehat{Q}E(\widehat{Q}) - E(\widehat{Q})^2 + 2E(\widehat{Q})^2 \\
&= \widehat{Q}^2 - E(\widehat{Q})^2 - 2E(\widehat{Q})(\widehat{Q} - E(\widehat{Q})),
\end{aligned}
$$

entonces,

$$\widehat{Q}^2 - E(\widehat{Q})^2 = (\widehat{Q} - E(\widehat{Q}))^2 + 2E(\widehat{Q})(\widehat{Q} - E(\widehat{Q}))$$

Esta igualdad junto con el binomio de Newton permite obtener lo siguiente:

$$\begin{aligned}
\left(\frac{\widehat{Q}^2 - E(\widehat{Q})^2}{E(\widehat{Q})^2}\right)^i &= \left(\frac{(\widehat{Q} - E(\widehat{Q}))^2 + 2E(\widehat{Q})(\widehat{Q} - E(\widehat{Q}))}{E(\widehat{Q})^2}\right)^i \\
&= \left(\frac{(\widehat{Q} - E(\widehat{Q}))^2}{E(\widehat{Q})^2} + \frac{2E(\widehat{Q})(\widehat{Q} - E(\widehat{Q}))}{E(\widehat{Q})^2}\right)^i \\
&= \sum_{j=0}^{i} \binom{i}{j} \left(\frac{(\widehat{Q} - E(\widehat{Q}))^2}{E(\widehat{Q})^2}\right)^j \left(\frac{2E(\widehat{Q})(\widehat{Q} - E(\widehat{Q}))}{E(\widehat{Q})^2}\right)^{i-j} \\
&= \sum_{j=0}^{i} \binom{i}{j} \frac{(\widehat{Q} - E(\widehat{Q}))^{2j}}{E(\widehat{Q})^{2j}} \cdot \frac{2^{i-j}(\widehat{Q} - E(\widehat{Q}))^{i-j}}{E(\widehat{Q})^{i-j}} \\
&= \sum_{j=0}^{i} \binom{i}{j} \frac{2^{i-j}(\widehat{Q} - E(\widehat{Q}))^{i+j}}{E(\widehat{Q})^{i+j}}
\end{aligned}$$

que se sustituye en la expresión (3.9):

$$\begin{aligned}
\frac{E(\widehat{Q})^2}{\widehat{Q}^2} &= 1 + \sum_{i=1}^{3}(-1)^i \left(\sum_{j=0}^{i} \binom{i}{j} \frac{2^{i-j}(\widehat{Q} - E(\widehat{Q}))^{i+j}}{E(\widehat{Q})^{i+j}}\right) \\
&\quad + \left(\sum_{j=0}^{4} \binom{4}{j} \frac{2^{4-j}(\widehat{Q} - E(\widehat{Q}))^{4+j}}{E(\widehat{Q})^{4+j}}\right) \frac{E(\widehat{Q})^2}{\widehat{Q}^2}
\end{aligned}$$

y se utiliza lo obtenido para calcular el momento de orden dos de $\widehat{\varphi}$:

$$
\begin{aligned}
E(\widehat{\varphi}^2) &= E\left(\frac{\widehat{P}^2}{\widehat{Q}^2}\right) = E\left(\frac{\widehat{P}^2}{E(\widehat{Q})^2}\frac{E(\widehat{Q})^2}{\widehat{Q}^2}\right) \\
&= E\left[\frac{\widehat{P}^2}{E(\widehat{Q})^2} + \sum_{i=1}^{3}(-1)^i\left(\sum_{j=0}^{i}\binom{i}{j}\frac{\widehat{P}^2}{E(\widehat{Q})^2}\frac{2^{i-j}(\widehat{Q}-E(\widehat{Q}))^{i+j}}{E(\widehat{Q})^{i+j}}\right)\right. \\
&\quad \left. + \sum_{j=0}^{4}\binom{4}{j}\frac{2^{4-j}(\widehat{Q}-E(\widehat{Q}))^{4+j}}{E(\widehat{Q})^{4+j}}\frac{E(\widehat{Q})^2}{\widehat{Q}^2}\frac{\widehat{P}^2}{E(\widehat{Q})^2}\right] \\
&= \frac{E(\widehat{P}^2)}{E(\widehat{Q})^2} + \sum_{i=1}^{3}(-1)^i\sum_{j=0}^{i}\binom{i}{j}\frac{2^{i-j}E(\widehat{P}^2(\widehat{Q}-E(\widehat{Q})^{i+j}))}{E(\widehat{Q})^{i+j+2}} \\
&\quad + \sum_{j=0}^{4}\binom{4}{j}\frac{2^{4-j}E\left(\frac{\widehat{P}^2}{\widehat{Q}^2}\left(\widehat{Q}-E(\widehat{Q})\right)^{4+j}\right)}{E(\widehat{Q})^{4+j}} \\
&= \frac{E(\widehat{P}^2)-E(\widehat{P})^2+E(\widehat{P})^2}{E(\widehat{Q})^2} + \sum_{i=1}^{3}(-1)^i\sum_{j=0}^{i}\binom{i}{j}\frac{2^{i-j}E(\widehat{P}^2(\widehat{Q}-E(\widehat{Q})^{i+j}))}{E(\widehat{Q})^{i+j+2}} \\
&\quad + \sum_{j=0}^{4}\binom{4}{j}\frac{2^{4-j}E\left(\frac{\widehat{P}^2}{\widehat{Q}^2}\left(\widehat{Q}-E(\widehat{Q})\right)^{4+j}\right)}{E(\widehat{Q})^{4+j}} \\
&= \frac{E\left(\widehat{P}^2-E(\widehat{P})^2\right)}{E(\widehat{Q})^2} + \frac{E(\widehat{P})^2}{E(\widehat{Q})^2} + \sum_{i=1}^{3}(-1)^i\sum_{j=0}^{i}\binom{i}{j}\frac{2^{i-j}E(\widehat{P}^2(\widehat{Q}-E(\widehat{Q})^{i+j}))}{E(\widehat{Q})^{i+j+2}} \\
&\quad + \sum_{j=0}^{4}\binom{4}{j}\frac{2^{4-j}E\left(\frac{\widehat{P}^2}{\widehat{Q}^2}\left(\widehat{Q}-E(\widehat{Q})\right)^{4+j}\right)}{E(\widehat{Q})^{4+j}}
\end{aligned}
$$

Se define

$$
\begin{aligned}
A_{ij} &= E\left[\left(\widehat{P}-E(\widehat{P})\right)^i\left(\widehat{Q}-E(\widehat{Q})\right)^j\right] \\
B_{ij} &= E\left[\widehat{P}^i\left(\widehat{Q}-E(\widehat{Q})\right)^j\right] \\
C_i &= E(\widehat{Q})^i \\
D_{ij} &= E\left[(1-\widehat{\varphi})^i\left(\widehat{Q}-E(\widehat{Q})\right)^j\right]
\end{aligned}
$$

para $i, j = 0, 1, ...$ Teniendo en cuenta que

$$A_{0j} = B_{0j}, \quad \forall\, j = 0, 1, ...$$

y

$$A_{2j} = B_{2j} - 2B_{10}A_{1j} + B_{10}^2A_{0j}$$

se sustituye en la expresión de $E(\varphi^2)$, obteniendo:

$$\begin{aligned} E(\widehat{\varphi}^2) &= \frac{A_{20}}{C_2} + \frac{B_{10}^2}{C_2} + \sum_{i=1}^{3}(-1)^i \sum_{j=0}^{i} \binom{i}{j} 2^{i-j} \frac{B_{2,i+j}}{C_{i+j+2}} + \sum_{j=0}^{4} \binom{4}{j} 2^{4-j} \frac{D_{2,4+j}}{C_{4+j}} \\ &= \frac{A_{20}}{C_2} + \frac{B_{10}^2}{C_2} + \sum_{i=1}^{3}(-1)^i \sum_{j=0}^{i} \binom{i}{j} 2^{i-j} \frac{A_{2,i+j} + 2B_{10}A_{1,i+j} - B_{10}^2 A_{0,i+j}}{C_{i+j+2}} \\ &\quad + \sum_{j=0}^{4} \binom{4}{j} 2^{4-j} \frac{D_{2,4+j}}{C_{4+j}} \end{aligned}$$

Se puede probar que para $i \geq 3$ se cumple que:

$$A_{i0} = E\big[\big(\widehat{P} - E(\widehat{P})\big)^i\big] = o\left(\frac{1}{nh}\right), \quad A_{0i} = B_{0i} = E\big[\big(\widehat{Q} - E(\widehat{Q})\big)^i\big] = o\left(\frac{1}{nh}\right)$$

con lo que

$$\begin{aligned} A_{ij} &= o\left(\frac{1}{nh}\right), \quad i + j \geq 3 \\ B_{ij} &= o\left(\frac{1}{nh}\right), \quad j \geq 3 \\ D_{ij} &= o\left(\frac{1}{nh}\right), \quad j \geq 3 \end{aligned}$$

y, además, $A_{01} = 0 = A_{10}$. Entonces,

$$\begin{aligned} E(\widehat{\varphi}^2) &= \frac{A_{20}}{C_2} + \frac{B_{10}^2}{C_2} - \frac{4B_{10}A_{11}}{C_3} - \frac{3B_{10}^2 A_{02}}{C_4} + o\left(\frac{1}{nh}\right) \\ &= \frac{Var(\widehat{P})}{E(\widehat{Q})^2} + \frac{E(\widehat{P})^2}{E(\widehat{Q})^2} - \frac{4E(\widehat{P})Cov(\widehat{P},\widehat{Q})}{E(\widehat{Q})^3} - \frac{3E(\widehat{P})^2 Var(\widehat{Q})}{E(\widehat{Q})^4} + o\left(\frac{1}{nh}\right) \end{aligned}$$

Por otro lado, en el cálculo del sesgo del estimador $\widehat{PD}(t|x_0)$ se halló una expresión para $E(\widehat{\varphi})$ que puede escribirse como sigue:

$$\begin{aligned} E(\widehat{\varphi}) &= \frac{B_{10}}{C_1} - \frac{A_{11}}{C_2} + \frac{A_{12} + B_{10}A_{02}}{C_3} - \frac{A_{13} + B_{10}A_{03}}{C_4} + \frac{D_{14}}{C_4} \\ &= \frac{E(\widehat{P})}{E(\widehat{Q})} - \frac{Cov(\widehat{P},\widehat{Q})}{E(\widehat{Q})^2} + \frac{E(\widehat{P})Var(\widehat{Q})}{E(\widehat{Q})^3} + o\left(\frac{1}{nh}\right) \end{aligned}$$

Entonces, usando que $Cov(\widehat{P},\widehat{Q}) = O\left(\frac{1}{nh}\right)$, $Var(\widehat{Q}) = O\left(\frac{1}{nh}\right)$, $E(\widehat{P}) = O(1)$ y $E(\widehat{Q}) = Q + o(1)$, se tiene,

$$
\begin{aligned}
Var\big(\widehat{PD}(t|x_0)\big) &= Var(1-\widehat{\varphi}) = Var(\widehat{\varphi}) = E\big(\widehat{\varphi}^2\big) - E(\widehat{\varphi})^2 \\
&= \frac{Var(\widehat{P})}{E(\widehat{Q})^2} + \frac{E(\widehat{P})^2}{E(\widehat{Q})^2} - \frac{4E(\widehat{P})Cov(\widehat{P},\widehat{Q})}{E(\widehat{Q})^3} - \frac{3E(\widehat{P})^2Var(\widehat{Q})}{E(\widehat{Q})^4} + o\left(\frac{1}{nh}\right) \\
&\quad - \left[\frac{E(\widehat{P})}{E(\widehat{Q})} - \left(\frac{Cov(\widehat{P},\widehat{Q})}{E(\widehat{Q})^2} - \frac{E(\widehat{P})Var(\widehat{Q})}{E(\widehat{Q})^3} + o\left(\frac{1}{nh}\right)\right)\right]^2 \\
&= \frac{Var(\widehat{P})}{E(\widehat{Q})^2} + \frac{E(\widehat{P})^2}{E(\widehat{Q})^2} - 4\frac{E(\widehat{P})Cov(\widehat{P},\widehat{Q})}{E(\widehat{Q})^3} - 3\frac{E(\widehat{P})^2Var(\widehat{Q})}{E(\widehat{Q})^4} + o\left(\frac{1}{nh}\right) \\
&\quad - \left[\frac{E(\widehat{P})^2}{E(\widehat{Q})^2} - 2\frac{E(\widehat{P})}{E(\widehat{Q})}\left(\frac{Cov(\widehat{P},\widehat{Q})}{E(\widehat{Q})^2} - \frac{E(\widehat{P})Var(\widehat{Q})}{E(\widehat{Q})^3} + o\left(\frac{1}{nh}\right)\right)\right. \\
&\quad \left. + \left(\frac{Cov(\widehat{P},\widehat{Q})}{E(\widehat{Q})^2} - \frac{E(\widehat{P})Var(\widehat{Q})}{E(\widehat{Q})^3} + o\left(\frac{1}{nh}\right)\right)^2\right] \\
&= \frac{Var(\widehat{P})}{E(\widehat{Q})^2} - 2\frac{E(\widehat{P})Cov(\widehat{P},\widehat{Q})}{E(\widehat{Q})^3} - 5\frac{E(\widehat{P})^2Var(\widehat{Q})}{E(\widehat{Q})^4} + o\left(\frac{1}{nh}\right) \\
&\quad + \left(\frac{Cov(\widehat{P},\widehat{Q})}{E(\widehat{Q})^2} - \frac{E(\widehat{P})Var(\widehat{Q})}{E(\widehat{Q})^3} + o\left(\frac{1}{nh}\right)\right)^2 + o\left(\frac{1}{nh}\right) \\
&= \frac{Var(\widehat{P})}{E(\widehat{Q})^2} - 2\frac{E(\widehat{P})Cov(\widehat{P},\widehat{Q})}{E(\widehat{Q})^3} - 5\frac{E(\widehat{P})^2Var(\widehat{Q})}{E(\widehat{Q})^4} + o\left(\frac{1}{nh}\right)
\end{aligned}
$$

Finalmente, basta tener en cuenta que

$$E(\widehat{P}) = E(\widehat{S}(t+b|x)) = P + A_{t+b}(x_0)h^2 + o(h^2)$$

$$E(\widehat{Q}) = E(\widehat{S}(t|x)) = Q + A_t(x_0)h^2 + o(h^2)$$

$$Var(\widehat{P}) = Var(\widehat{S}(t+b|x)) = B_{t+b}(x_0)\frac{1}{nh} + o\left(\frac{1}{nh}\right)$$

$$Var(\widehat{Q}) = Var(\widehat{S}(t|x)) = B_t(x_0)\frac{1}{nh} + o\left(\frac{1}{nh}\right)$$

$$Cov(\widehat{P},\widehat{Q}) = Cov(\widehat{S}(t+b|x),\ \widehat{S}(t|x)) = C_{t,t+b}(x_0)\frac{1}{nh} + o\left(\frac{1}{nh}\right)$$

y sustituyendo se obtiene la varianza asintótica de $\widehat{PD}(t|x_0)$:

$$
\begin{aligned}
Var\big(\widehat{PD}(t|x)\big) &= \frac{Var(\widehat{P})}{E(\widehat{Q})^2} - 2\frac{E(\widehat{P})Cov(\widehat{P},\widehat{Q})}{E(\widehat{Q})^3} - 5\frac{E(\widehat{P})^2Var(\widehat{Q})}{E(\widehat{Q})^4} + o\left(\frac{1}{nh}\right) \\
&= \frac{1}{Q^2}\left(B_{t+b}(x_0) - 2\frac{P}{Q}C_{t,t+b}(x_0) - 5\left(\frac{P}{Q}\right)^2 B_t(x_0)\right)\frac{1}{nh} + o\left(\frac{1}{nh}\right)
\end{aligned}
$$

□

3.3. Aplicación del estimador a datos simulados

Los datos utilizados son los de las muestras simuladas según los supuestos explicados en la sección 2.3.1. En esta sección se utiliza el estimador de la PD construido a partir del estimador de Cai de la supervivencia condicional para estimar tanto la supervivencia como la probabilidad de mora en ambas muestras.

Para la suavización en la covariable se elige el núcleo de Epanechnikov, de soporte compacto. Tal y cómo se hizo anteriormente, las gráficas mostradas han sido obtenidas utilizando un valor del parámetro ventana h que, de ser posible, minimiza el error cuadrático integrado cometido en la estimación de la curva de probabilidad de mora.

En la figura 3.1 se muestran las gráficas del $RECI$ en la estimación de la probabilidad de mora condicionada a los cuantiles $Q_{0.25}$, $Q_{0.5}$, $Q_{0.75}$ de la covariable X para la muestra 1. Cabe destacar que la raíz del error cuadrático integrado como función de la ventana h es decreciente para los cuantiles $Q_{0.25}$ y $Q_{0.5}$, sugiriendo que se tome una ventana lo más grande posible y, por tanto, suavizando al máximo en la covariable. Por el contrario, la raíz del error cuadrático integrado es creciente para el cuantil $Q_{0.75}$ en el rango analizado, sugiriendo tomar un valor de h muy pequeño, lo que implica una suavización muy leve en la covariable.

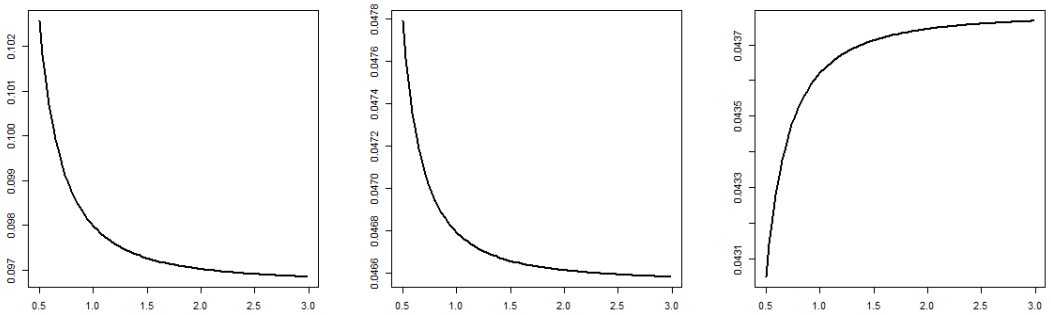


Figura 3.1: $RECI(h)$ para el estimador $\widehat{PD}_h^C(t|x)$ con $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$ en la muestra 1.

Aunque en este caso no se alcanza un valor óptimo del parámetro de suavizado, se escoge aquel que arroja un menor error cuadrático integrado, dentro del rango de valores de h considerados. En la tabla 3.1 se muestra dicho valor para cada cuantil de la covariable

y la raíz del error cuadrático integrado que se obtiene. Se observa que el error cometido en la estimación se reduce al aumentar el valor de la covariable al que se condiciona la función de supervivencia.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
h	3.0	3.0	0.5
$RECI$	0.097	0.047	0.043

Tabla 3.1: Ventana óptima y $RECI$ obtenido mediante $\widehat{PD}_h^C(t|x)$ en la muestra 1.

En la figura 3.3 se muestran las estimaciones de la supervivencia condicional y la probabilidad de mora de la muestra 1 obtenidas para los valores del parámetro h dados en la tabla anterior. Nótese que la estimación de la supervivencia ajusta realmente bien la verdadera curva; sin embargo la estimación de la PD, pese a ser razonable, presenta excesiva variabilidad, problema que ya se detectó en el estimador basado en el de Beran para la supervivencia.

Considerando ahora la muestra 2, en la figura 3.2 se muestran las gráficas de la raíz del error cuadrático integrado cometido en la estimación de la probabilidad de mora obtenida del estimador de Cai para la supervivencia. En este caso, la función $RECI(h)$ es decreciente para los dos primeros cuartiles, sugiriendo que se tome un parámetro de suavizado lo más grande posible. Para el cuantil $Q_{0.75}$ la función $RECI$ presenta un mínimo, proporcionando por tanto un parámetro de ventana óptimo.

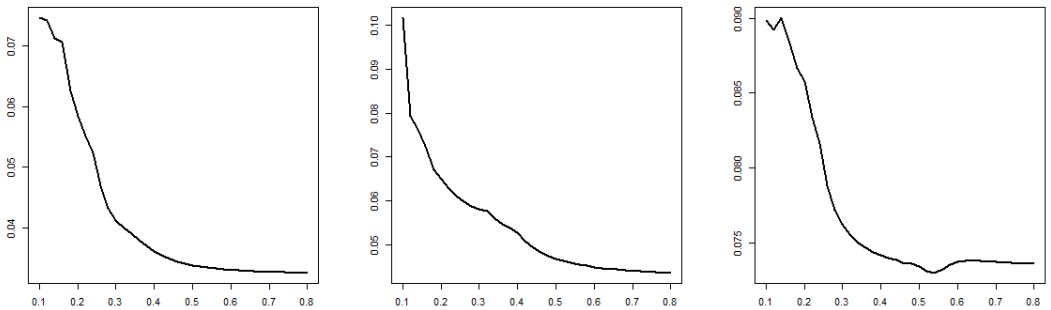


Figura 3.2: $RECI(h)$ para el estimador $\widehat{PD}_h^C(t|x)$ con $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$ en la muestra 2.

Tomando en la estimación de la PD para cada cuartil la ventana de suavizado apropiada, se obtienen los valores de $RECI$ mostrados en la tabla 3.2.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
h	0.80	0.80	0.54
$RECI$	0.033	0.044	0.073

Tabla 3.2: Ventana óptima y $RECI$ obtenido mediante $\widehat{PD}_h^C(t|x)$ en la muestra 2.

En la figura 3.4 se muestran las gráficas de las curvas de supervivencia condicional y probabilidad de mora estimadas para los valores óptimos de la ventana obtenidos. De nuevo, la estimación de la supervivencia condicional es buena, pero la estimación de la probabilidad de mora presenta excesiva variabilidad, especialmente en la cola derecha, aunque las oscilaciones se producen siguiendo la tendencia de la verdadera curva de probabilidad.

Tal y como ocurrió con el estimador de la PD obtenido a partir del estimador de Beran para la supervivencia, las estimaciones halladas a partir del estimador de Cai son razonables, pero presentan mucha variabilidad. Para tratar de solucionarlo, en la siguiente sección se propone una suavización del estimador en la variable tiempo.

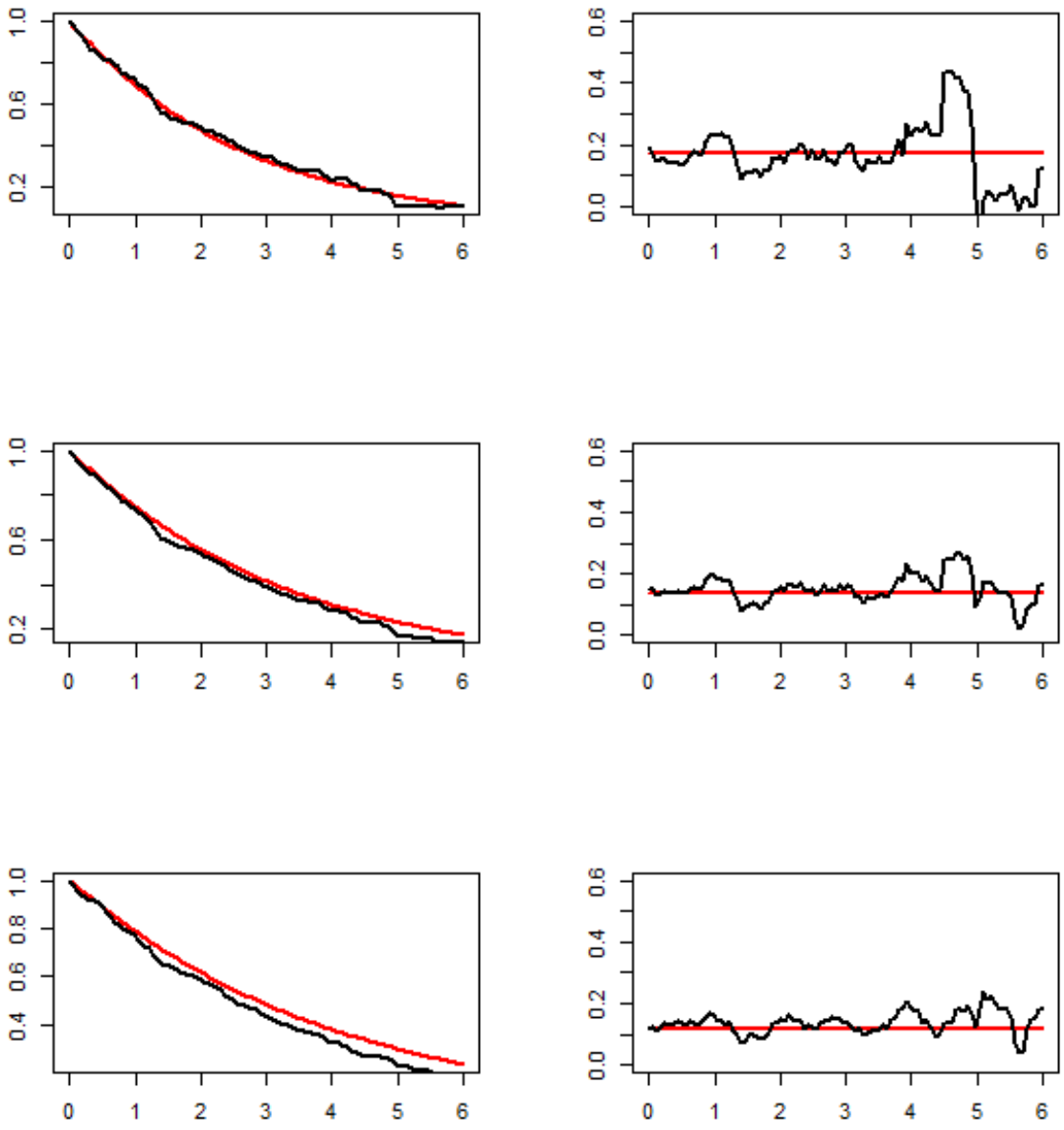


Figura 3.3: Izquierda: Supervivencia condicional (línea roja) y su estimación $\widehat{S}_h^C(t|x)$ (línea negra). Derecha: Probabilidad de mora (línea roja) y su estimación $\widehat{PD}_h^C(t|x)$ (línea negra). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 1.

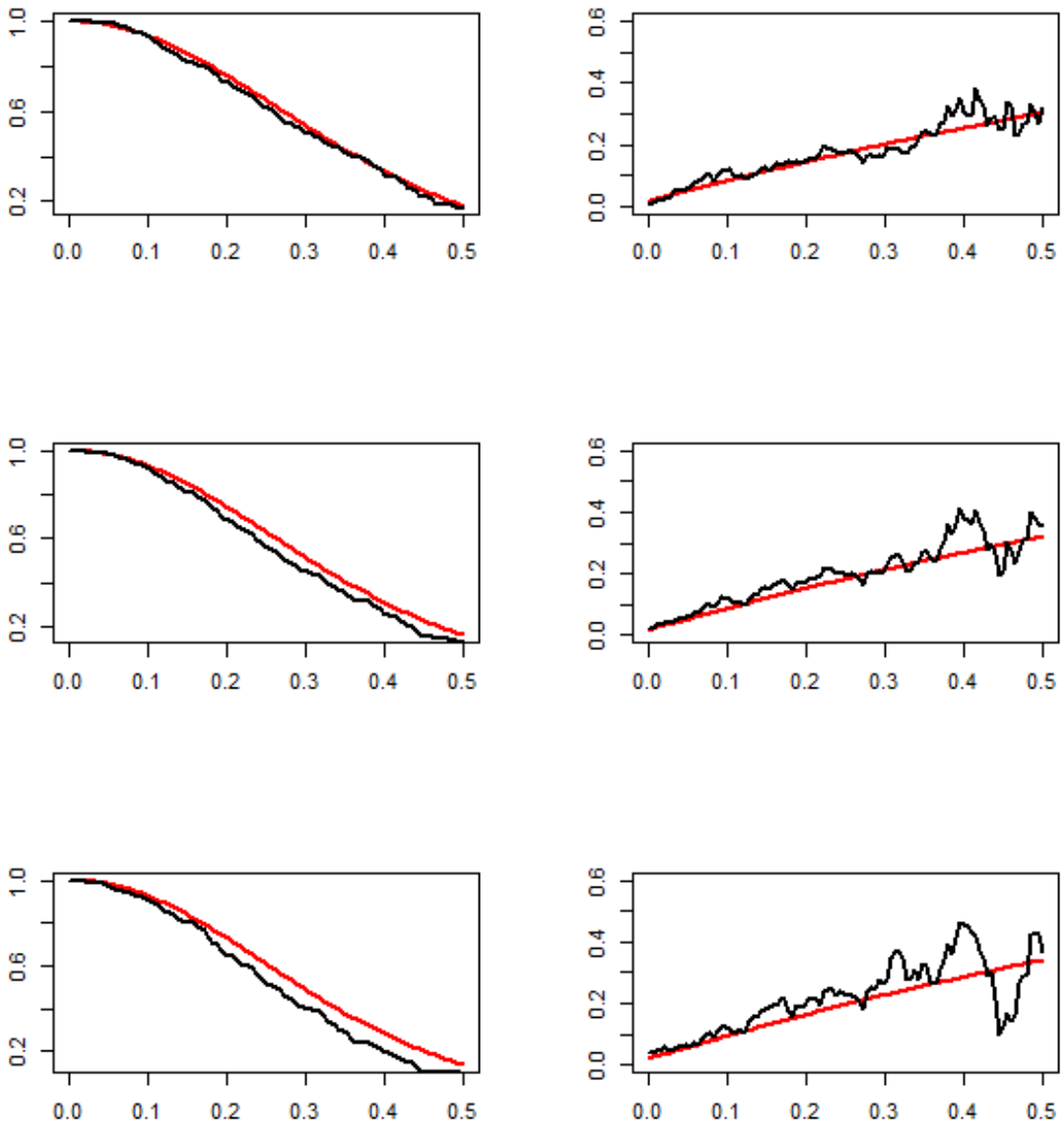


Figura 3.4: Izquierda: Supervivencia condicional (línea roja) y su estimación $\widehat{S}_h^C(t|x)$ (línea negra). Derecha: Probabilidad de mora (línea roja) y su estimación $\widehat{PD}_h^C(t|x)$ (línea negra). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 2.

3.4. Estimador de Cai suavizado

En la sección anterior se vio que el estimador de Cai para la regresión permite obtener estimaciones razonables de la supervivencia y de la PD. De hecho, se obtienen buenas aproximaciones de la verdadera supervivencia. Sin embargo, las estimaciones obtenidas para la probabilidad de mora presentan excesiva variabilidad. Se trata del mismo problema que se observó en el estimador de Beran y, al igual que en ese caso, la solución que se propone consiste en construir la siguiente versión suavizada del estimador:

$$\widehat{S}^C_{h,g}(t|x) \;=\; 1-\sum_{i=1}^{n} s_i \mathbb{K}\left(\frac{t-Z_{(i)}}{g}\right)$$

donde $s_i = \widehat{S}^C_h(Z_{(i)}|x) - \widehat{S}^C_h(Z_{(i-1)}|x)$ siendo $\widehat{S}^C_h(\cdot|x)$ el estimador de Cai de la supervivencia y $\mathbb{K}(t)$ la función de distribución de un núcleo K.

El estimador de la probabilidad de mora a horizonte b de obtiene según la expresión:

$$\widehat{PD}^C_{h,g}(t|x) \;=\; 1-\frac{\widehat{S}^C_{h,g}(t+b|x)}{\widehat{S}^C_{h,g}(t|x)}$$

3.4.1. Aplicación del estimador suavizado a datos simulados

En esta sección se estudia el comportamiento de $\widehat{PD}^C_{h,g}(t|x)$ como estimador de la probabilidad de mora aplicándolo sobre las muestras 1 y 2. El núcleo de suavizado para la covariable utilizado es el de Epanechnikov. La ventana de suavizado h es fijada a los valores óptimos obtenidos en la sección 2.3.1 para cada caso. Para la suavización en la variable tiempo se considera una distribución de núcleo gaussiano y la ventana de suavizado, g, se escoge siguiendo el mismo criterio que se siguió entonces para h: fijado el valor óptimo de h, se obtiene la probabilidad de mora estimada por $\widehat{PD}^C_{h,g}(t|x)$ en una rejilla de valores del parámetro ventana g y se escoge el valor de g que arroja un menor error cuadrático integrado.

Considerando en primer lugar la muestra 1, en la figura 3.5 se muestra la gráfica de la raíz del error cuadrático integrado, $RECI$, como función de g en cada cuantil de la covariable. El valor de la ventana de suavizado g elegido es en el que la función $RECI(h,g)$ alcanza un mínimo.

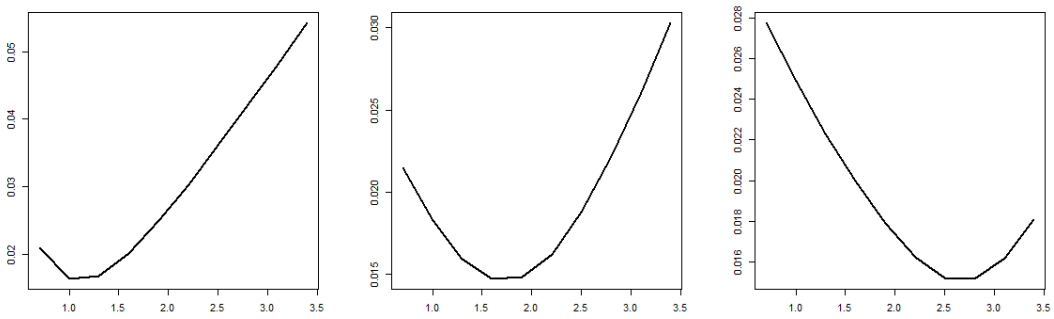


Figura 3.5: $RECI(h,g)$ para el estimador $\widehat{PD}^{C}_{h,g}(t|x)$ con $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$ en la muestra 1.

En la tabla 3.3 se muestran estos valores óptimos de la ventana g, el valor de la ventana h que se consideró y el valor de $RECI$ obtenido. Se observa que el error en la estimación se reduce considerablemente con respecto al error cometido sin suavización.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
h	3.0	3.0	0.5
g	1.0	1.6	2.8
$RECI$	0.016	0.015	0.015

Tabla 3.3: Ventana óptima y $RECI$ obtenido mediante $\widehat{PD}^{C}_{h,g}(t|x)$ en la muestra 1.

En la figura 3.7 se muestran las estimaciones de la supervivencia condicional y la probabilidad de mora obtenidas para los valores de h, g y x dados en la tabla anterior. La estimación para la supervivencia condicional es muy similar a la obtenida en la figura 3.3, es decir, sin suavización. Sin embargo, la mejora en las estimaciones de la probabilidad de mora con respecto a las mostradas en dicha figura son importantes. La estimación de la probabilidad de mora parece ser significativamente mejor si se utiliza la versión suavizada del estimador de Cai.

En la tabla 3.4 se muestra la raíz del error cuadrático integrado cometido en la estimación de la probabilidad de mora en la muestra 1 mediante el estimador basado en el de Cai y

mediante su versión suavizada, quedando patente que el segundo reduce el error cometido mejorando la estimación de la probabilidad de mora.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
$RECI(h)$	0.097	0.047	0.043
$RECI(h,g)$	0.016	0.015	0.015

Tabla 3.4: Valores $RECI$ para $\widehat{PD}_h^C(t|x)$ y para $\widehat{PD}_{h,g}^C(t|x)$ en la muestra 1.

En la figura 3.8 se pueden ver las diferencias entre la estimación de la supervivencia y la PD mediante el estimador de Cai y mediante su versión suavizada para la muestra 1. Las diferencias en la estimación de la supervivencia condicional no son notables. Sin embargo, en la estimación de la probabilidad de mora se observa una mejora muy considerable.

Considerando ahora la muestra 2, en la figura 3.6 se pueden ver las gráficas del error cuadrático integrado como función de g para la estimación de la probabilidad de mora. En la tabla 3.5 se muestra el valor de la ventana óptima, en la que la función $RECI(h,g)$ alcanza un mínimo y el valor de dicho mínimo, para cada cuartil de la covariable. La reducción del error con respecto al cometido mediante el estimador sin suavización en la variable tiempo es notable.

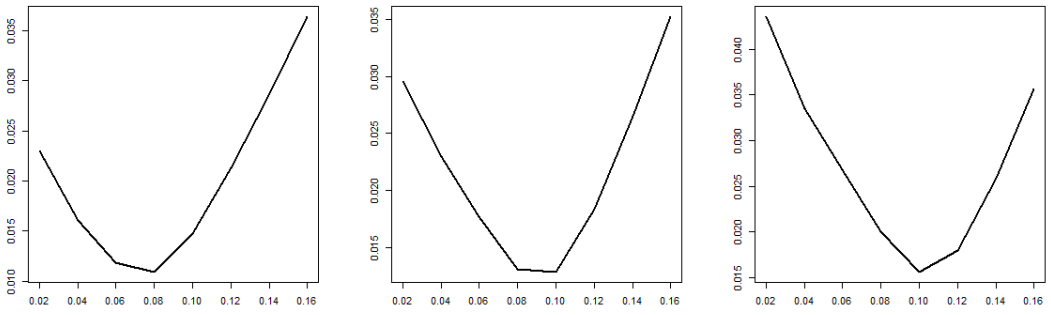


Figura 3.6: $RECI(h,g)$ para el estimador $\widehat{PD}_{h,g}^C(t|x)$ con $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$ en la muestra 2.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
h	0.80	0.80	0.54
g	0.08	0.10	0.10
$RECI$	0.011	0.013	0.016

Tabla 3.5: Ventana óptima y $RECI$ obtenido mediante $\widehat{PD}^{C}_{h,g}(t|x)$ en la muestra 2.

En la figura 3.9 se muestran las gráficas de la supervivencia condicional y la probabilidad de mora estimadas incorporando la suavización en la variable tiempo al estimador de Cai para cada cuartil de la covariable en la muestra 2. Debe destacarse el buen ajuste del estimador a la verdadera curva, tanto para la supervivencia condicional como para la probabilidad de mora.

En la tabla 3.6 se resumen los valores del $RECI$ cometido en la estimación de la probabilidad de mora mediante el estimador basado en Cai y mediante su versión suavizada y la figura 3.10 muestra las diferencias entre uno y otro estimador para la muestra 2. Cabe mencionar que la estimación para la probabilidad de mora con suavización en la variable tiempo es mucho mejor que sin suavización, mientras que para la supervivencia, la estimación suavizada en la variable tiempo no presenta notables diferencias con la estimación sin suavizar.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
$RECI(h)$	0.033	0.044	0.073
$RECI(h,g)$	0.011	0.013	0.016

Tabla 3.6: Valores $RECI$ para $\widehat{PD}^{C}_{h}(t|x)$ y para $\widehat{PD}^{C}_{h,g}(t|x)$ en la muestra 2.

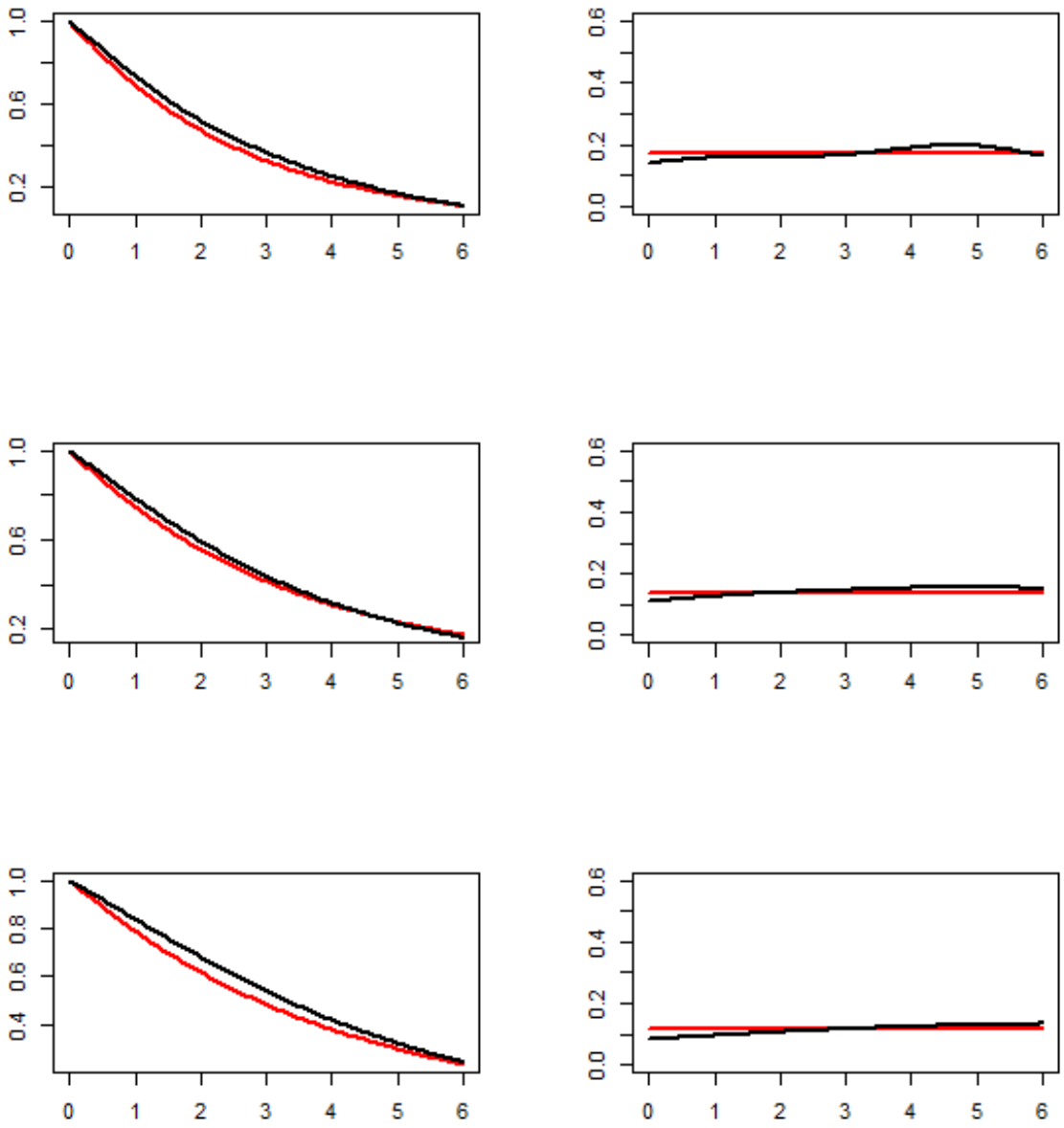


Figura 3.7: Izquierda: Supervivencia condicional (línea roja) y su estimación $\widehat{S}^C_{h,g}(t|x)$ (línea negra). Derecha: Probabilidad de mora (línea roja) y su estimación $\widehat{PD}^C_{h,g}(t|x)$ (línea negra). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 1.

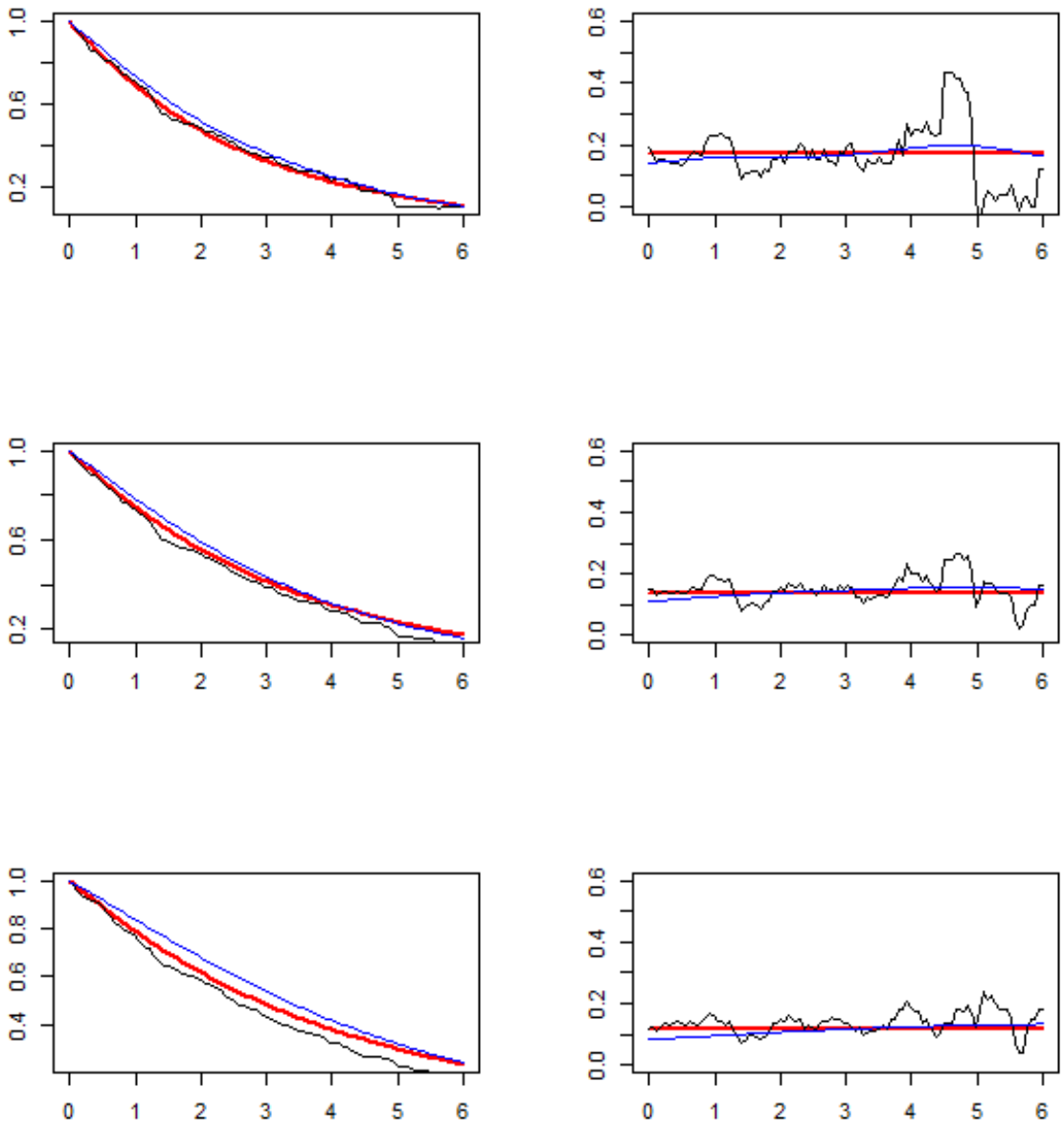


Figura 3.8: Izquierda: Supervivencia condicional (línea roja), su estimación por $\widehat{S}_h^C(t|x)$ (línea negra) y por $\widehat{S}_{h,g}^C(t|x)$ (línea azul). Derecha: Probabilidad de mora (línea roja), su estimación por $\widehat{PD}_h^C(t|x)$ (línea negra) y por $\widehat{PD}_{h,g}^C(t|x)$ (línea azul). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 1.

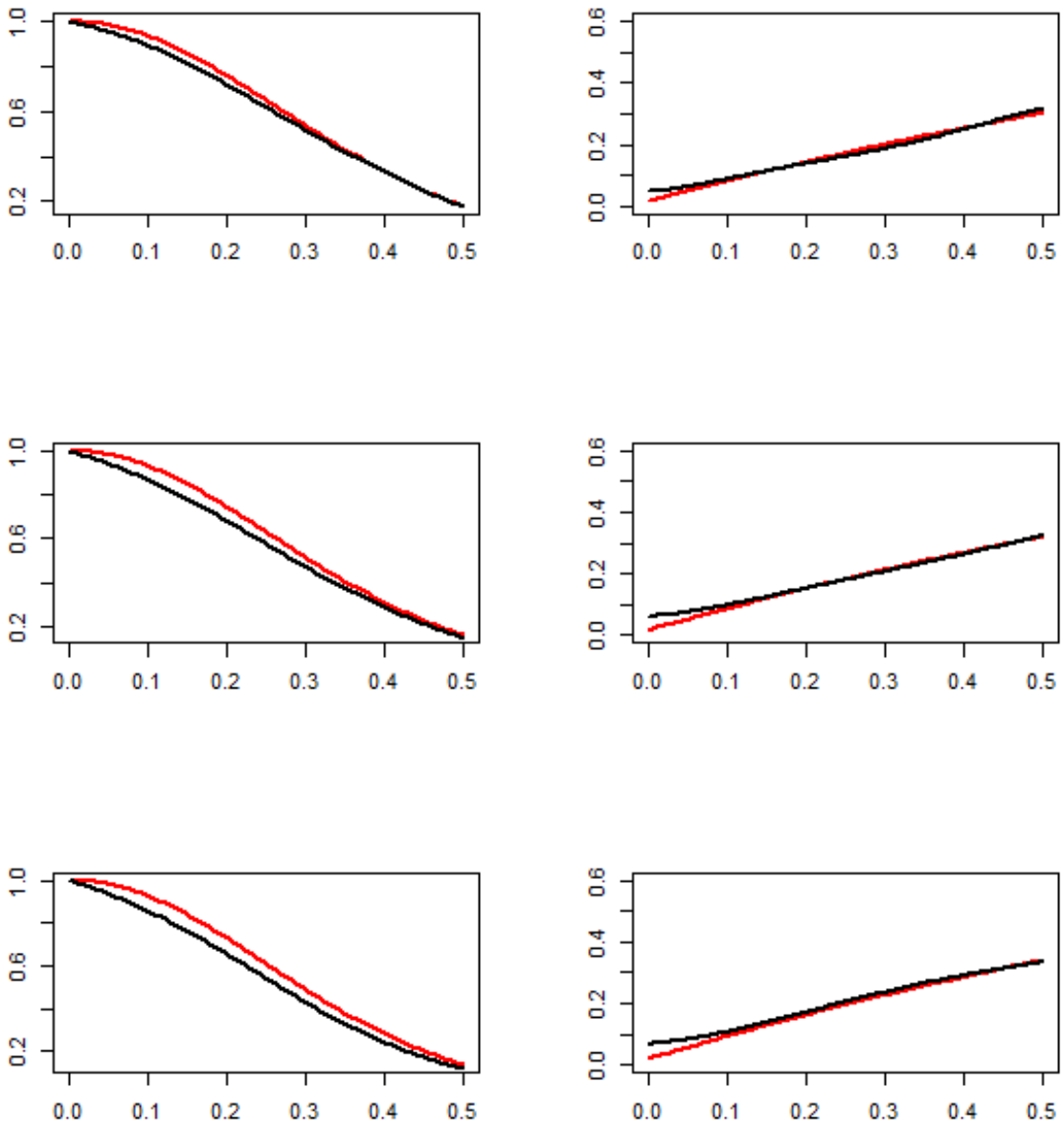


Figura 3.9: Izquierda: Supervivencia condicional (línea roja) y su estimación $\widehat{S}^C_{h,g}(t|x)$ (línea negra). Derecha: Probabilidad de mora (línea roja) y su estimación $\widehat{PD}^C_{h,g}(t|x)$ (línea negra). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 2.

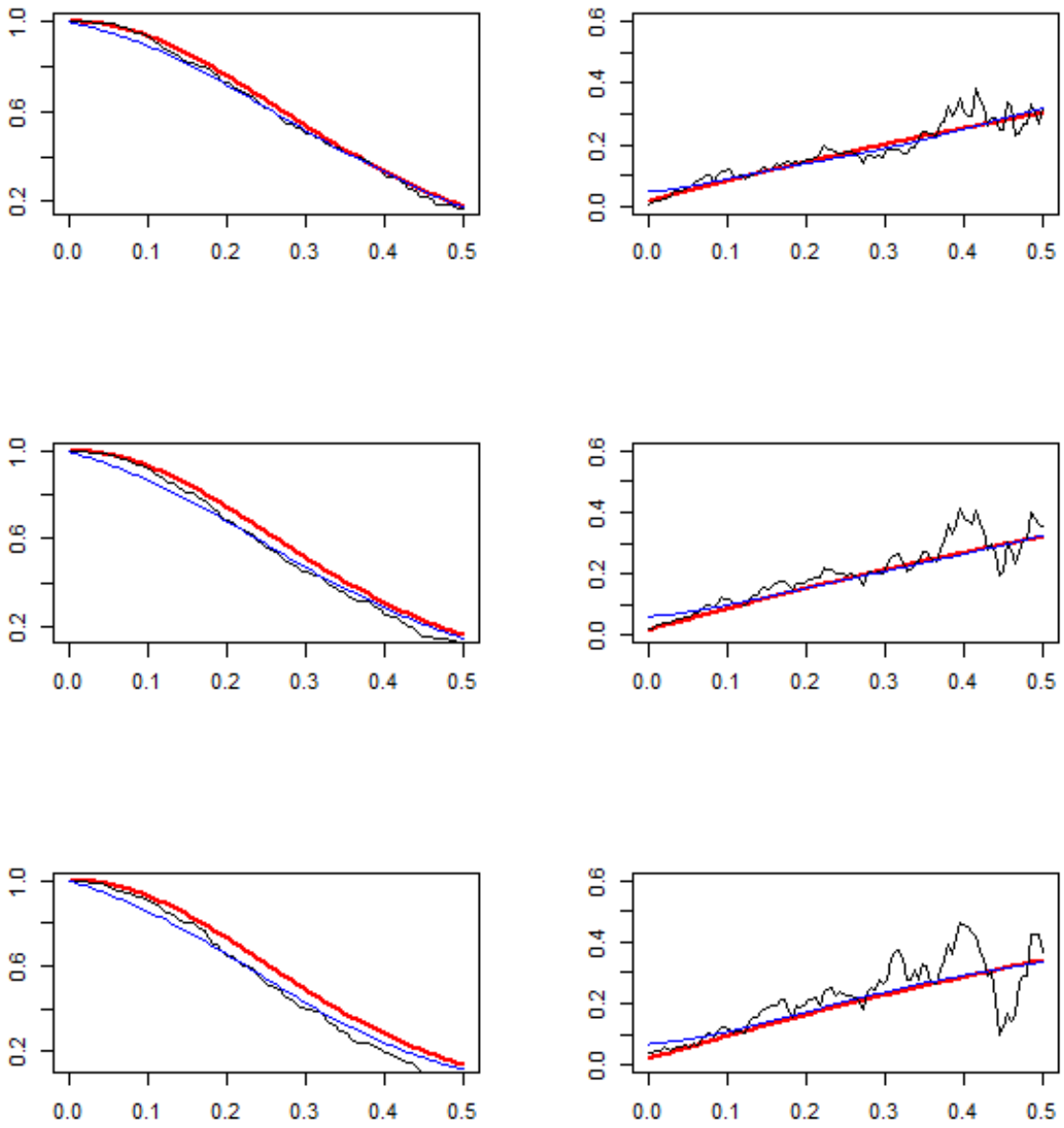


Figura 3.10: Izquierda: Supervivencia condicional (línea roja), su estimación por $\widehat{S}_h^C(t|x)$ (línea negra) y por $\widehat{S}_{h,g}^C(t|x)$ (línea azul). Derecha: Probabilidad de mora (línea roja), su estimación por $\widehat{PD}_h^C(t|x)$ (línea negra) y por $\widehat{PD}_{h,g}^C(t|x)$ (línea azul). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 2.

3.4.2. Discusión sobre la ventana

La ventana de suavización en la variable tiempo, g, se eligió minimizando el error cuadrático integrado y es una ventana global. Del mismo modo que en el estimador de Beran,

se plantea ahora el uso de un parámetro ventana local determinado por el método del k-vecino más próximo o k-NN. La elección del valor del entero k se hace siguiendo el mismo criterio que para la ventana g.

En la tabla 3.7 se muestra la ventana fijada h, el valor óptimo de k y el valor del $RECI$ en la estimación para cada cuartil de la covariable de la muestra 1. En la tabla 3.8 se pueden ver los mismos datos para la muestra 2. En ambos casos es notable el aumento de la raíz de error cuadrático integrado con respecto al suavizado con ventana global. Recuérdese que esto ya se observó en las estimaciones obtenidas a partir del estimador de Beran para la supervivencia.

El error cometido en la estimación de la PD mediante Cai suavizado con ventana k-NN se reduce con respecto al estimador de Cai sin suavización en la variable tiempo, pero en las figuras 3.11 y 3.12, se constata que el comportamiento de este estimador no mejora el del estimador de Cai suavizado con ventana global.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
h	3.0	3.0	0.5
k	40	40	80
$RECI$	0.030	0.019	0.017

Tabla 3.7: Ventana óptima, k óptimo y $RECI$ obtenido mediante $\widehat{PD}^{C}_{h,k}(t|x)$ en la muestra 1.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
h	0.80	0.80	0.54
k	30	30	25
$RECI$	0.021	0.027	0.043

Tabla 3.8: Ventana óptima, k óptimo y $RECI$ obtenido mediante $\widehat{PD}^{C}_{h,k}(t|x)$ en la muestra 2.

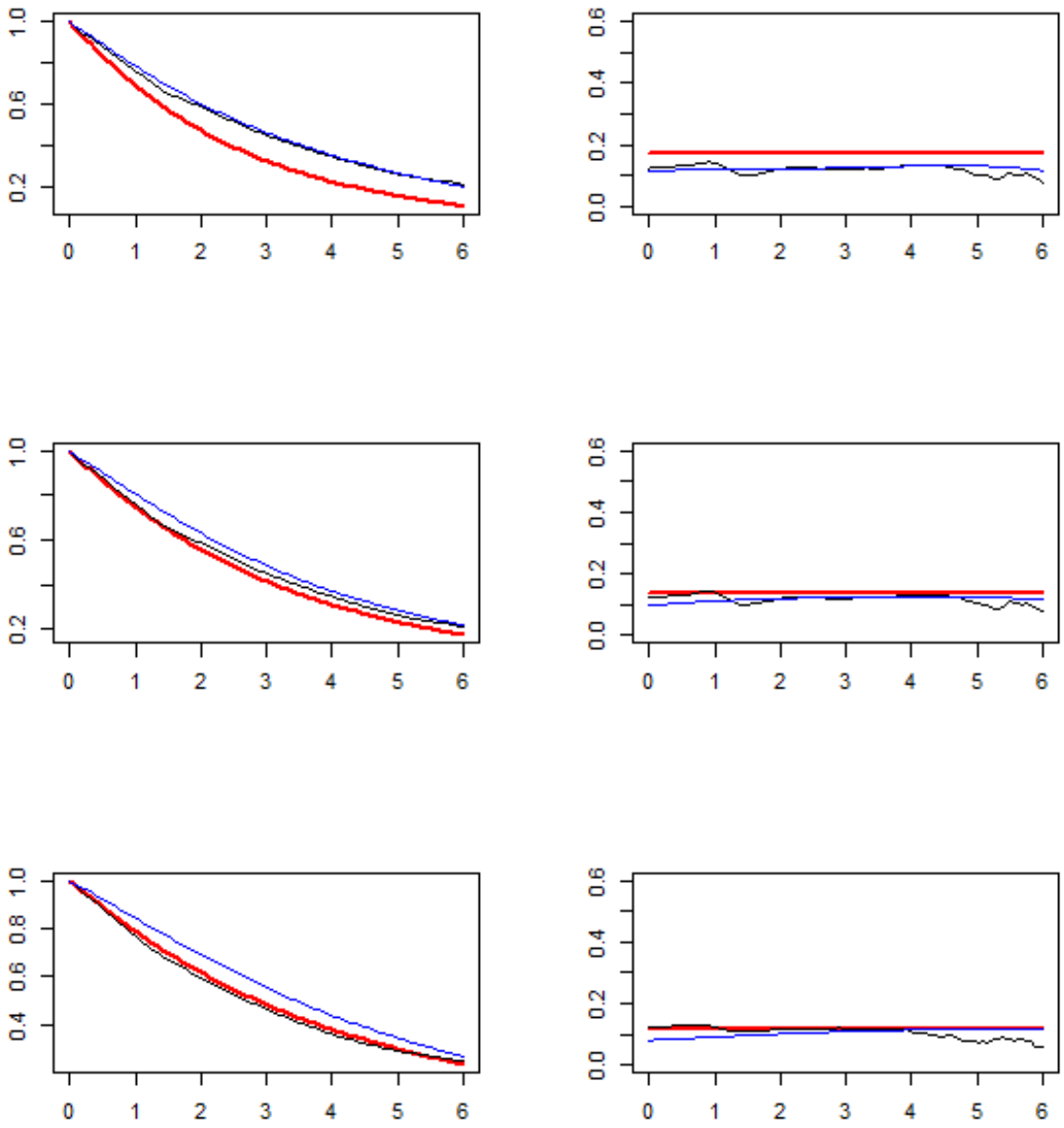


Figura 3.11: Izquierda: Supervivencia condicional (línea roja), su estimación por $\widehat{S}^C_{h,g}(t|x)$ (línea azul) y por $\widehat{S}^C_{h,k}(t|x)$ (línea negra). Derecha: Probabilidad de mora (línea roja), su estimación por $\widehat{PD}^C_{h,g}(t|x)$ (línea azul) y por $\widehat{PD}^C_{h,k}(t|x)$ (línea negra). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 1.

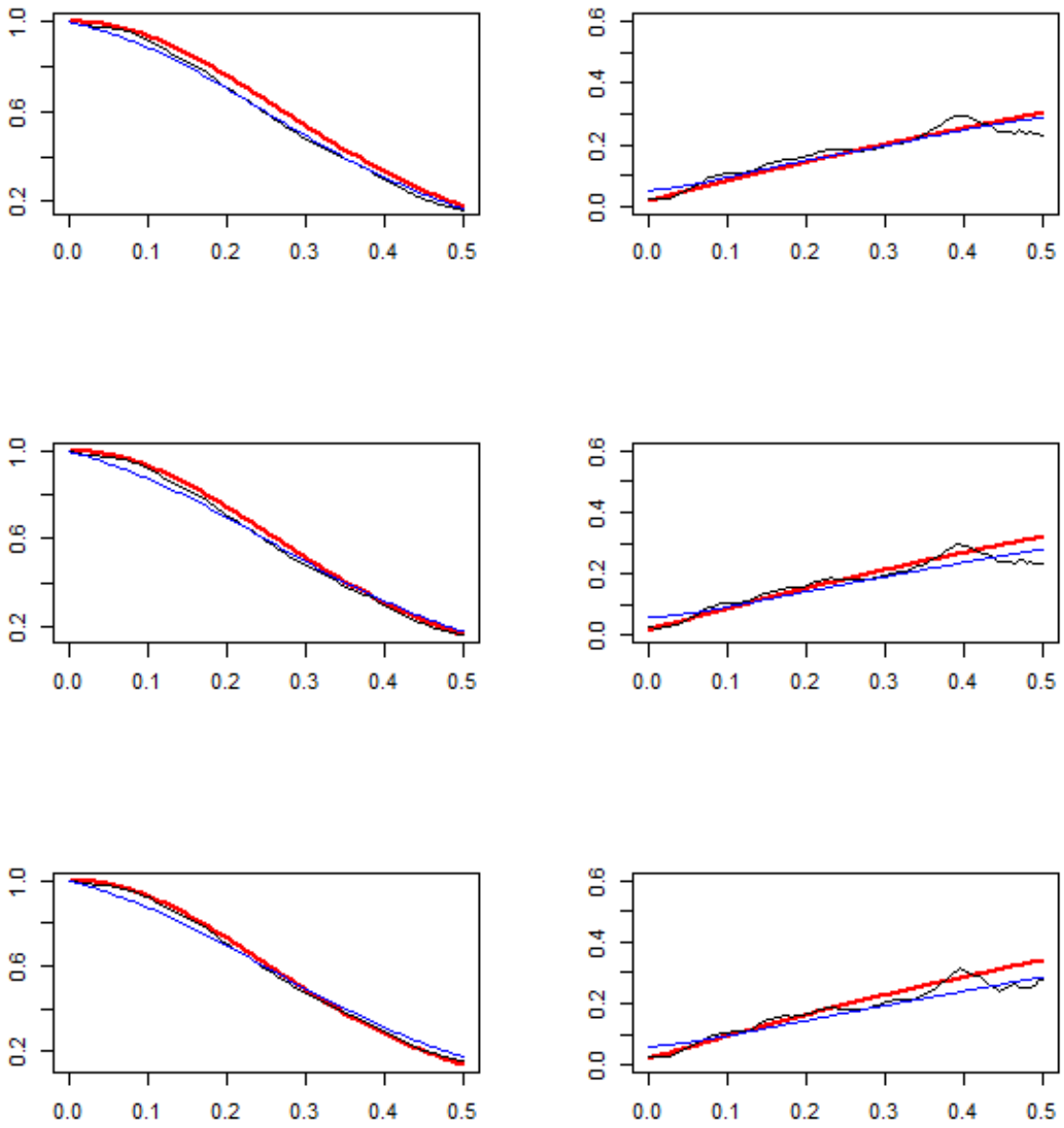


Figura 3.12: Izquierda: Supervivencia condicional (línea roja), su estimación por $\widehat{S}^C_{h,g}(t|x)$ (línea azul) y por $\widehat{S}^C_{h,k}(t|x)$ (línea negra). Derecha: Probabilidad de mora (línea roja), su estimación por $\widehat{PD}^C_{h,g}(t|x)$ (línea azul) y por $\widehat{PD}^C_{h,k}(t|x)$ (línea negra). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 2.

Capítulo 4

Estimador de la PD basado en el estimador de Van Keilegom y Akritas

En Van Keilegom and Akritas (1999) se propone un estimador de la función de distribución condicional basado en estimar la de los residuos de un modelo de regresión no paramétrico que involucra la variable de interés T explicada por la covariable X. A partir del mismo se obtiene, tal y como se hizo anteriormente, un estimador de la función de supervivencia condicional y, por tanto, un estimador de la probabilidad de mora. Mediante este estimador se pretende mejorar la estimación de la PD en la cola derecha. Se utilizará sobre muestras simuladas para comprobar si se consigue dicho objetivo.

4.1. Estimador de la función de distribución condicionada en presencia de censura

El estimador de Beran presenta desventajas en situaciones de alta censura; la ausencia de información en la cola derecha de la distribución lo vuelve inconsistente en esa región. En Van Keilegom and Akritas (1999) se propone un método para estimar la función de distribución condicional $F(t|x)$ de manera que se transfiera información de $F(t|x)$ con x en regiones de baja censura a la cola derecha con alta censura. Para ello, es necesario

asumir el siguiente modelo de regresión no paramétrico sobre las variables involucradas:

$$T = m(X) + \sigma(X)\varepsilon$$

donde $m(x) = E(T|X = x)$ es la curva de regresión desconocida; $\sigma(x)$ es la desviación típica condicional, representando un posible modelo heterocedástico, y ε es la variable error.

Nótese que

$$P(T \leq t|X = x) = P(m(X) + \sigma(X)\varepsilon \leq t|X = x) = P\left(\varepsilon \leq \frac{t - m(x)}{\sigma(x)}\right),$$

con lo que

$$F(t|x) = F_\varepsilon\left(\frac{t - m(x)}{\sigma(x)}\right),$$

donde F_ε denota la función de distribución de la variable error ε. Esta relación entre la función de distribución condicionada de T y F_ε sugiere la siguiente propuesta de estimador para $F(t|x)$.

Sean $\widehat{m}(x)$ y $\widehat{\sigma}(x)$ estimadores consistentes de $m(x)$ y $\sigma(x)$, respectivamente y sea $\widehat{F}_\varepsilon$ el estimador de Kaplan-Meier de F_ε. El estimador de la función de distribución condicionada $F(t|x)$ según este modelo es:

$$\widehat{F}(t|x) = \widehat{F}_\varepsilon\left(\frac{t - \widehat{m}(x)}{\widehat{\sigma}(x)}\right) \tag{4.1}$$

El primer paso para obtener (4.1) será encontrar estimadores adecuados de $m(x)$ y $\sigma(x)$ por lo que, sin pérdida de generalidad, se consideran los funcionales de localización y escala dados por

$$m(x) \;=\; \int_0^1 F^{-1}(s|x)J(s)ds, \tag{4.2}$$

$$\sigma^2(x) \;=\; \int_0^1 F^{-1}(s|x)^2 J(s)ds - m^2(x), \tag{4.3}$$

donde $F^{-1}(s|x) = \text{ínf}\{t : F(t|x) \geq s\}$ es la función cuantil de T condicionada a X y $J(s)$ es tal que $\int_0^1 J(s)ds = 1$. Cuando se elige concretamente $J(s) = 1$, $\forall\ s \in [0, 1]$, entonces, las expresiones (4.2) y (4.3) resultan ser $E(T|X = x)$ y $Var(T|X = x)$, respectivamente. Para otras elecciones de $J(s)$, por ejemplo $J(s) = \frac{1}{1-\alpha} I_{\{\alpha/2 \leq s \leq 1-\alpha 2\}}$, se obtienen versiones truncadas de la media y la varianza condicionadas.

Considérese también el estimador de Beran de $F(t|x)$ dado por

$$\tilde{F}(t|x) = 1 - \prod_{Z_i \leq t} \left(1 - \frac{w_{i,n}(x,h)}{\sum_{j=1}^{n} I_{\{Z_j \geq Z_i\}} w_{j,n}(x,h)}\right)^{\delta_i} \tag{4.4}$$

donde

$$w_{i,n}(x,h) = \frac{K\big((x - X_i)/h\big)}{\sum_{j=1}^{n} K\big((x - X_j)/h\big)}$$

son los pesos de suavización de Nadaraya-Watson, K una función núcleo y $h = h_n$ el parámetro ventana para la suavización en la covariable.

Ahora, sustituyendo el estimador (4.4) en (4.2) y (4.3), se obtienen los estimadores de $m(x)$ y $\sigma^2(x)$ siguientes:

$$\widehat{m}(x) = \int_0^1 \tilde{F}^{-1}(s|x) J(s) ds, \tag{4.5}$$

$$\widehat{\sigma}^2(x) = \int_0^1 \tilde{F}^{-1}(s|x)^2 J(s) ds - \widehat{m}^2(x), \tag{4.6}$$

Finalmente, basta considerar el estimador de Kaplan-Meier de F_ε dado por

$$\widehat{F}_\varepsilon(t) = 1 - \prod_{\widehat{E}_{(i)} \leq t} \left(\frac{n-i}{n-i+1}\right)^{\delta_{(i)}}$$

obtenido a partir de los residuos censurados de la regresión ordenados $(\widehat{E}_{(1)}, ..., \widehat{E}_{(n)})$ y sus correspondientes concomitantes $(\delta_{(1)}, ..., \delta_{(n)})$ siendo

$$\widehat{E}_i = \frac{Z_i - \widehat{m}(X_i)}{\widehat{\sigma}(X_i)}.$$

4.2. Estimación de la probabilidad de mora

Sea $\widehat{S}_h^{VKA}(t|x_0)$ el estimador de la función de supervivencia obtenido a partir del estimador de Van Keilegom de la función de distribución. Esto es,

$$\widehat{S}_h^{VKA}(t|x_0) = 1 - \widehat{F}(t|x_0)$$

donde $\widehat{F}(t|x_0)$ es el estimador de la función de distribución dado en (4.1). Entonces, el estimador de la probabilidad de mora a horizonte b condicionado al valor de la covariable $X = x_0$, obtenido a partir de este estimador de la supervivencia, es:

$$\widehat{PD}_h^{VKA}(t|x_0) = 1 - \frac{\widehat{S}_h^{VKA}(t+b|x_0)}{\widehat{S}_h^{VKA}(t|x_0)} \tag{4.7}$$

4.3. Aplicación del estimador a datos simulados

En esta sección se muestra un análisis del estimador de la PD construido a partir del estimador de la supervivencia propuesto por Van Keilegom y Akritas con corrección del efecto frontera. El estimador permite obtener la probabilidad de mora en las muestras 1 y 2 anteriormente utilizadas.

En primer lugar se establece un criterio para elegir una ventana, en algún sentido óptima, para el suavizado en la covariable X. En la figura 4.1 se muestran las estimaciones de la probabilidad de mora obtenidas para cada una de las muestras a partir del estimador dado en (4.7), pero utilizando los residuos de la regresión teóricos.

Dado que en la muestra 1 la distribución del tiempo de vida del crédito, T, condicionada a la variable X es una exponencial de parámetro conocido, dada X, es posible conocer las verdaderas funciones de regresión y varianza para dichas variables:

$$T|_{X=x} \equiv Exp(1/P(x)) \quad \Rightarrow \quad \begin{cases} m(x) = E(T|X=x) = 1/P(x) \\ \sigma^2(x) = 1/P(x)^2 \end{cases}$$

De este modo, una estimación de la probabilidad de mora obtenida con los residuos teóricos de la regresión es la calculada mediante el estimador de la distribución dado en (4.1) donde $\widehat{m}$, $\widehat{\sigma}^2$ y $\widehat{E}_i$ son sustituidos por m, σ y E_i, siendo

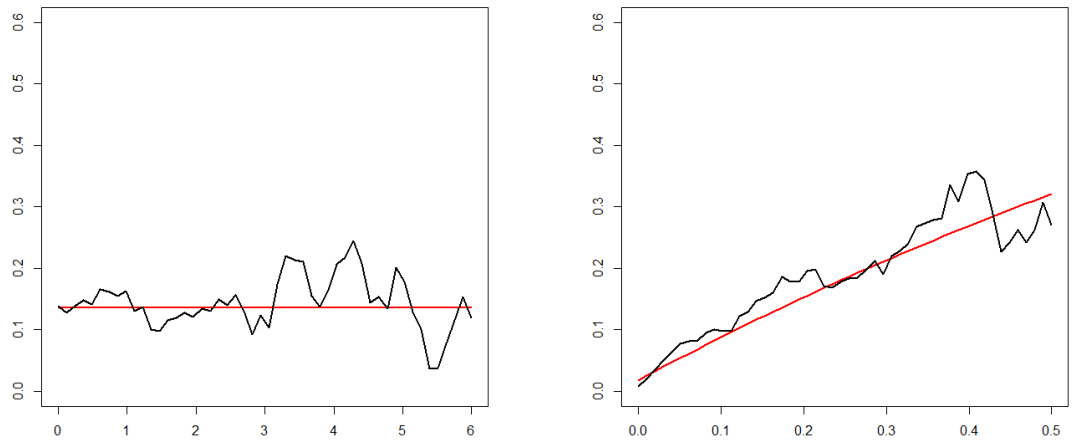
$$E_i = \frac{Z_i - m(X_i)}{\sigma(X_i)}$$

con $i = 1, ..., n$. Esta estimación se muestra en la figura 4.1a.

Análogamente, se conoce la función de regresión $m(x)$ y la función de varianza $\sigma^2(x)$ para la muestra 2 y siguiendo el mismo razonamiento, se obtiene una estimación de la probabilidad de mora para esta muestra calculada a partir de los residuos teóricos de la regresión. Esta estimación se muestra en la figura 4.1b.

En ambos casos se puede ver que el uso de las funciones teóricas de regresión y varianza proporcionan buenas estimaciones de la probabilidad de mora, de donde se puede deducir que cuanto mejor sea la estimación de estas funciones, mejor será la estimación de la PD. Por este motivo, el criterio definido para elegir el parámetro de suavizado es el siguiente:

se estiman m y σ según los estimadores dados en (4.5) y (4.6) para diferentes valores del parámetro ventana h, se calcula para cada una de esas estimaciones el error cuadrático integrado que se comete y, finalmente, se escoge el valor de h que arroja un menor error como parámetro de suavizado en la covariable para la estimar la PD.



(a) Muestra 1.

(b) Muestra 2.

Figura 4.1: Probabilidad de mora teórica (línea roja) y su estimación (línea negra) a partir del estimador de Van Keilegom y Akritas para la supervivencia con residuos teóricos de la regresión, $x = Q_{0.5}$.

En la tabla 4.1 se muestra el valor óptimo de la ventana y la raíz del error cuadrático integrado cometido en la estimación de la probabilidad de mora mediante $\widehat{PD}_h^{VKA}(t|x)$ para dicho valor de h y para cada uno de los cuantiles de la covariable en la muestra 1. En la figura 4.2 se muestran las estimaciones obtenidas en cada caso junto a la verdadera curva de probabilidad de mora, así como la estimación de la supervivencia de la que se obtiene. Tanto en los valores del $RECI$ como en las gráficas, se puede ver que este estimador proporciona curvas razonablemente parecidas a las verdaderas curvas de supervivencia y probabilidad de mora, y la estimación es mejor al aumentar el valor de la covariable.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
h	0.8	0.8	0.8
$RECI$	0.054	0.036	0.037

Tabla 4.1: Ventana óptima y $RECI$ obtenido mediante $\widehat{PD}_h^{VKA}(t|x)$ en la muestra 1.

El parámetro ventana óptimo para la muestra 2 es el presentado en la tabla 4.2, junto con la raíz del error cuadrático integrado cometido en la estimación de la probabilidad de mora para cada cuantil de la covariable. En la figura 4.3 se representan las gráficas de la supervivencia y la probabilidad de mora estimadas frente a las teóricas para cada cuantil de la covariable en la muestra 2. De nuevo, la estimación de la supervivencia mejora al aumentar el valor de la covariable; por el contrario, la estimación de la PD es muy similar para cada cuantil. Tal y como ocurrió en capítulos anteriores, las estimaciones de la probabilidad de mora son razonables, pero presentan excesiva variabilidad. Por este motivo, en la siguiente sección se propone una versión suavizada del estimador de Van Keilegom y Akritas para la supervivencia.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
h	2	2	2
$RECI$	0.039	0.036	0.038

Tabla 4.2: Ventana óptima y $RECI$ obtenido mediante $\widehat{PD}_h^{VKA}(t|x)$ en la muestra 2.

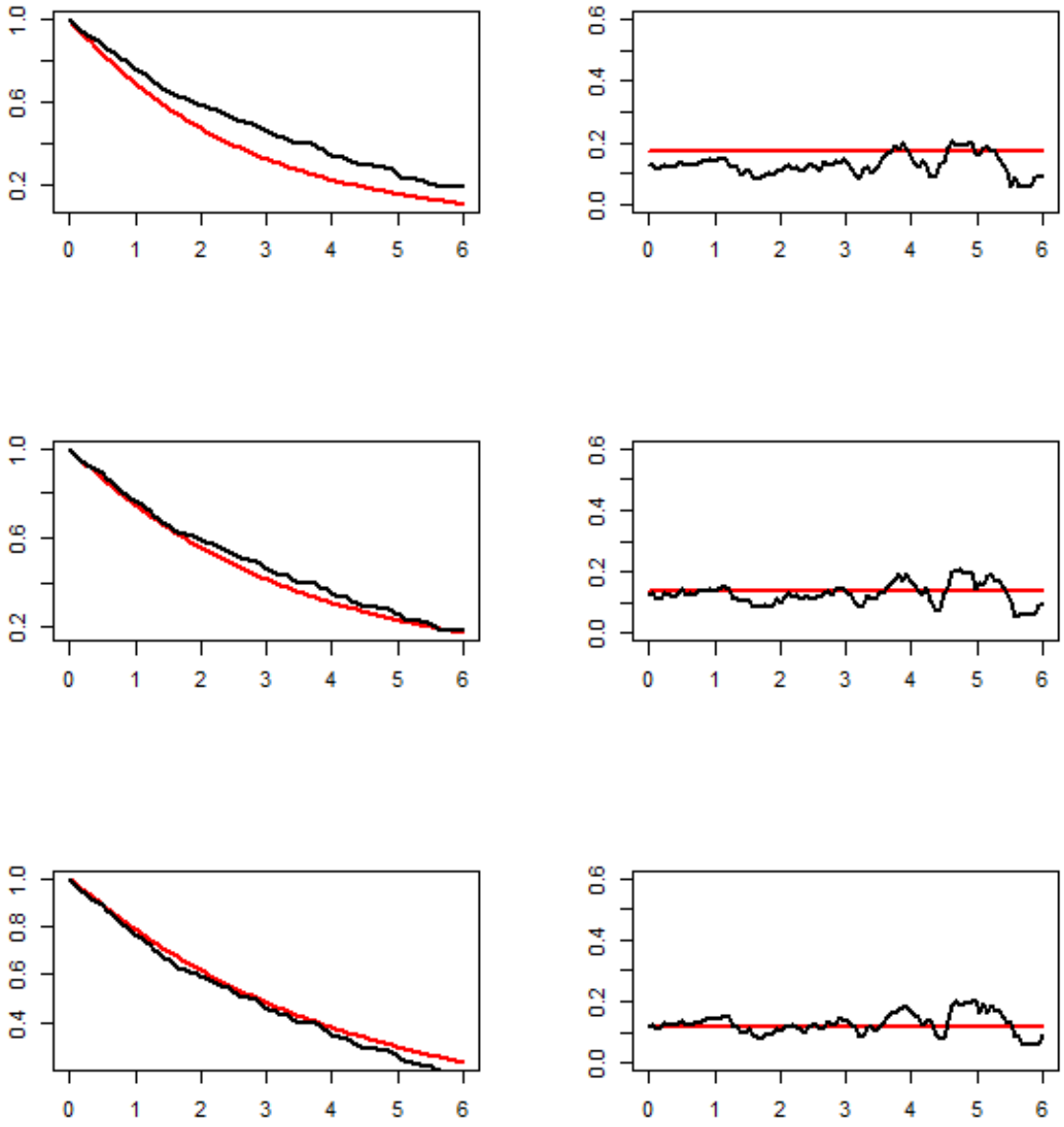


Figura 4.2: Izquierda: Supervivencia condicional (línea roja) y su estimación $\widehat{S}_h^{VKA}(t|x)$ (línea negra). Derecha: Probabilidad de mora (línea roja) y su estimación $\widehat{PD}_h^{VKA}(t|x)$ (línea negra). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 1.

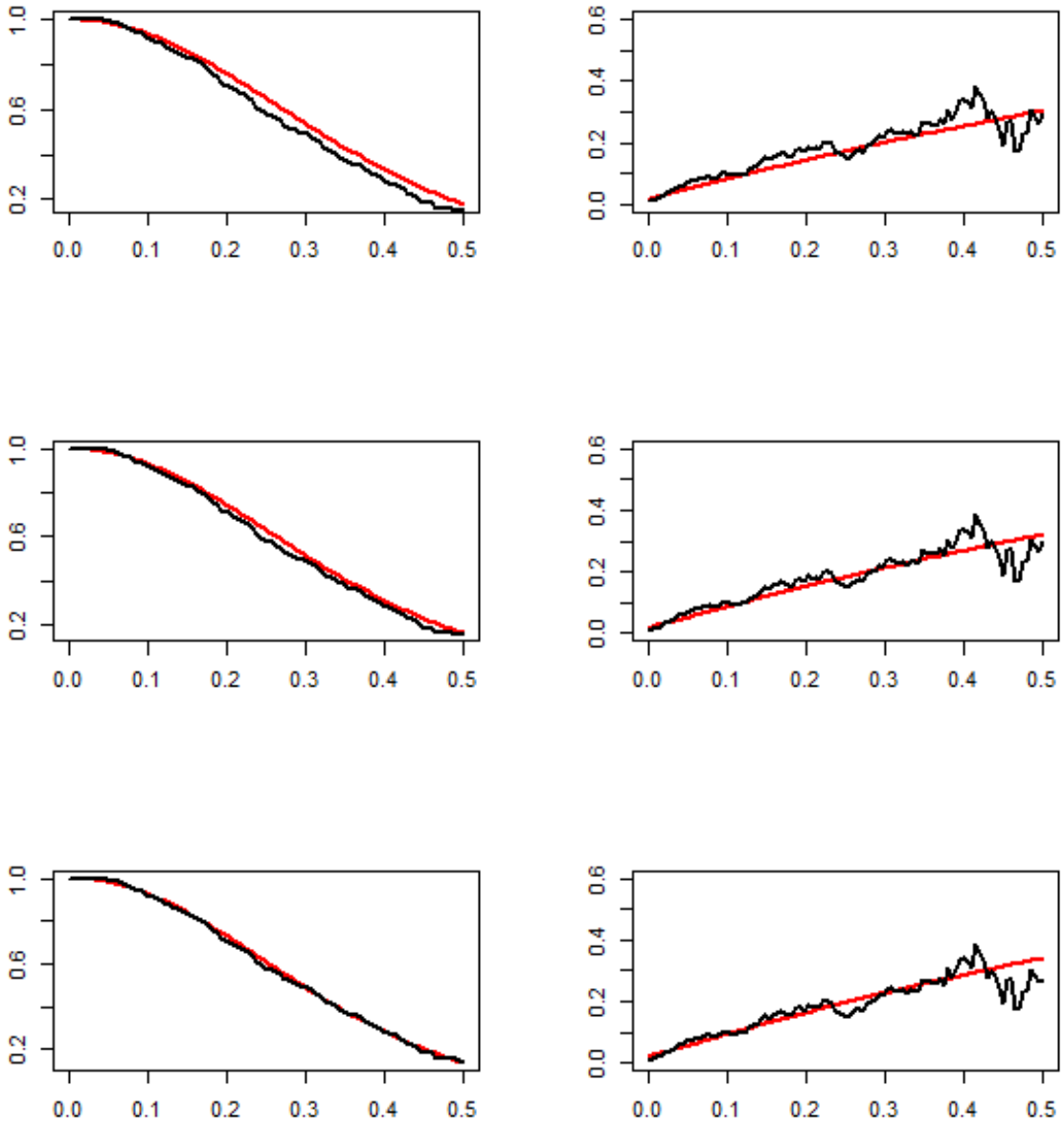


Figura 4.3: Izquierda: Supervivencia condicional (línea roja) y su estimación $\widehat{S}_h^{VKA}(t|x)$ (línea negra). Derecha: Probabilidad de mora (línea roja) y su estimación $\widehat{PD}_h^{VKA}(t|x)$ (línea negra). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 2.

4.4. Estimador de Van Keilegom y Akritas suavizado

En capítulos anteriores se observó que suavizar el estimador de la supervivencia en la variable tiempo disminuye notablemente el error de estimación de la probabilidad de mora. Esta propuesta resultó exitosa tanto para el estimador de Beran de la supervivencia como para el estimador de Cai. Por ello, en esta sección se procede de forma análoga con el estimador de Van Keilegom y Akritas, construyendo una nueva versión suavizada del estimador dada por la siguiente expresión:

$$\widehat{S}_{h,g}^{VKA}(t|x) \quad = \quad 1 - \sum_{i=1}^{n} s_i \mathbb{K}\left(\frac{t - Z_{(i)}}{g}\right)$$

donde $s_i = \widehat{S}_h^{VKA}(Z_{(i)}|x) - \widehat{S}_h^{VKA}(Z_{(i-1)}|x)$ siendo $\widehat{S}_h^{VKA}(\cdot|x)$ el estimador de Van Keilegom y Akritas de la supervivencia y $\mathbb{K}(t)$ la función de distribución de un núcleo K.

El estimador de la probabilidad de mora a horizonte b de obtiene según la expresión:

$$\widehat{PD}_{h,g}^{VKA}(t|x) \quad = \quad 1 - \frac{\widehat{S}_{h,g}^{VKA}(t+b|x)}{\widehat{S}_{h,g}^{VKA}(t|x)}$$

4.4.1. Aplicación del estimador suavizado a datos simulados

A continuación, se analiza el estimador de la probabilidad de mora $\widehat{PD}_{h,g}^{VKA}(t|x)$ aplicándolo sobre las muestras 1 y 2. En la suavización para la covariable se utiliza el núcleo de Epanechnikov y la ventana de suavizado h se fija a los valores óptimos obtenidos en la sección 4.3 para cada caso. Para la suavización en la variable tiempo se considera una distribución de núcleo gaussiano y la ventana de suavizado, g, se escoge del siguiente modo: fijado el valor óptimo de h, se obtiene la probabilidad de mora estimada por $\widehat{PD}_{h,g}^{VKA}(t|x)$ en una rejilla de valores del parámetro ventana g y se elige el valor de g que arroja un menor error cuadrático integrado.

Considerando en primer lugar la muestra 1, en la figura 4.4 se presenta la gráfica de la raíz del error cuadrático integrado, $RECI$, como función de g en cada cuantil de la covariable. El valor óptimo de la ventana de suavizado g es en el que la función $RECI(h, g)$ alcanza un mínimo.

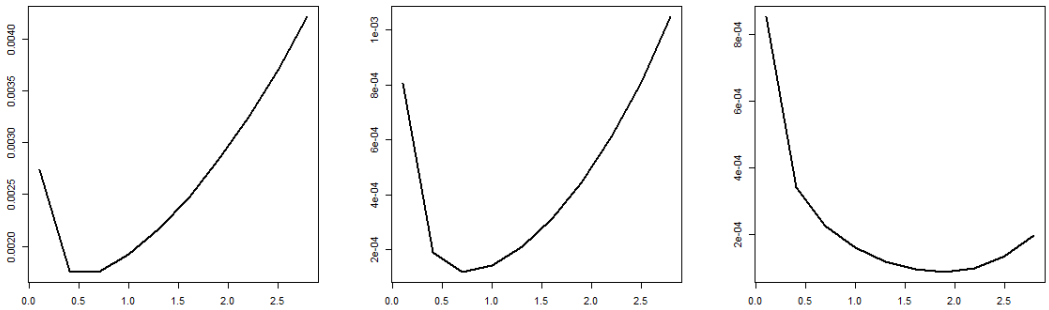


Figura 4.4: $RECI(h)$ para el estimador $\widehat{PD}_h^{VKA}(t|x)$ con $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$ en la muestra 1.

En la tabla 4.4 se muestra el valor de las ventanas h y g, parámetros de suavizado en la covariable y en la variable tiempo, respectivamente, así como la raíz del error cuadrático integrado cometido en la estimación de la probabilidad de mora mediante dichas ventanas para cada cuartil de la covariable. El error cometido mediante este estimador suavizado en el tiempo se reduce notablemente con respecto al estimador sin suavización. En la figura 4.6 se representan gráficamente la función de supervivencia condicional y la probabilidad de mora estimadas mediante el estimador de Van Keilegom y Akritas frente a las curvas teóricas para los tres cuantiles de la covariable en la muestra 1. Tal y como ocurría sin suavización, las estimaciones tanto de la supervivencia como de la PD son peores en el primero de los cuartiles que en los otros dos, pero, en cualquier caso, con esta propuesta ha sido posible eliminar la excesiva variabilidad que presentaba el estimador original.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
h	0.8	0.8	0.8
g	0.7	0.7	1.9
$RECI$	0.042	0.010	0.009

Tabla 4.3: Ventana óptima y $RECI$ obtenido mediante $\widehat{PD}_{h,g}^{VKA}(t|x)$ en la muestra 1.

En la tabla 4.4 se muestra la raíz del error cuadrático integrado cometido en la estimación de la probabilidad de mora en la muestra 1 mediante el estimador basado en el de Van

Keilegom y Akritas para la supervivencia y mediante su versión suavizada. Es claro que el error de estimación se reduce de forma considerable. En la figura 4.7 se puede constatar este hecho, pues se presentan la supervivencia y probabilidad de mora teóricas para la muestra 1 frente a las estimaciones mediante $\widehat{PD}_h^{VKA}(t|x)$ y mediante $\widehat{PD}_{h,g}^{VKA}(t|x)$. Aunque las diferencias entre uno y otro estimador de la función de supervivencia no son notables, la estimación suavizada de la probabilidad de mora es mucho más razonable que la estimación sin suavizar en la variable tiempo.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
$RECI(h)$	0.054	0.036	0.037
$RECI(h,g)$	0.042	0.010	0.009

Tabla 4.4: Valores $RECI$ para $\widehat{PD}_h^{VKA}(t|x)$ y para $\widehat{PD}_{h,g}^{VKA}(t|x)$ en la muestra 1.

Considerando ahora la muestra 2, en la figura 4.5 se muestra la curva del $RECI$ como función del parámetro de suavizado g en la estimación de la PD mediante $\widehat{PD}_h^{VKA}(t|x)$ para cada cuantil de la covariable. Se observa que en los tres casos, tal función tiene forma parabólica y presenta un mínimo. El valor de g en el que se alcanza dicho mínimo es el utilizado para estimar la supervivencia condicional y la probabilidad de mora.

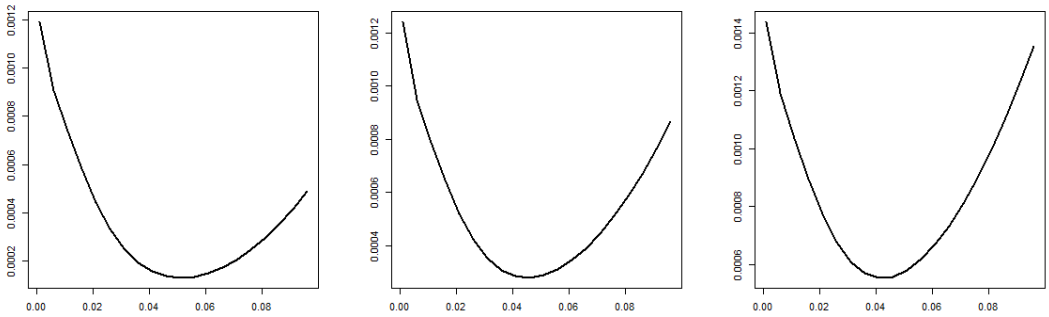


Figura 4.5: $RECI(h)$ para el estimador $\widehat{PD}_h^{VKA}(t|x)$ con $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$ en la muestra 2.

En la tabla 4.5 se muestran los valores óptimos de h y g y la raíz del error cuadrático integrado correspondiente a la estimación de la PD con tales parámetros ventana. En la

figura 4.8 se representan la supervivencia condicional y la probabilidad de mora teóricas y estimadas mediante el estimador de Van Keilegom y Akritas para cada cuantil de la covariable y los valores de las ventanas dados en esta tabla.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
h	2	2	2
g	0.051	0.046	0.041
$RECI$	0.011	0.017	0.023

Tabla 4.5: Ventana óptima y $RECI$ obtenido mediante $\widehat{PD}_{h,g}^{VKA}(t|x)$ en la muestra 2.

A diferencia de lo que ocurría en la muestra 1, en este caso el error cometido en la estimación aumenta en los cuantiles altos de la covariable. Sin embargo, el aumento del error no es tan acusado y la estimación de la probabilidad de mora es muy similar a la verdadera curva. De hecho, en la figura 4.8 se puede ver que la diferencia entre las estimaciones de la probabilidad de mora en uno y otro cuartil es sutil y en los tres casos la estimación empeora ligeramente en la cola derecha de la distribución.

En la tabla 4.6 se muestra la raíz del error cuadrático integrado cometido en la estimación de la probabilidad de mora en la muestra 2 mediante el estimador basado en el de Van Keilegom y Akritas para la supervivencia y mediante su versión suavizada para cada cuartil. Es evidente que el error de estimación se ve reducido al aplicar la suavización en la variable tiempo en el estimador de la supervivencia, aunque esta reducción es menor en la cola derecha de la distribución.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
$RECI(h)$	0.038	0.036	0.038
$RECI(h,g)$	0.011	0.017	0.023

Tabla 4.6: Valores $RECI$ para $\widehat{PD}_{h}^{VKA}(t|x)$ y para $\widehat{PD}_{h,g}^{VKA}(t|x)$ en la muestra 2.

En la figura 4.9 se muestran la función de supervivencia y la probabilidad de mora estimadas para cada cuartil de la covariable en la muestra 2 mediante los estimadores $\widehat{PD}_h^{VKA}(t|x)$ y $\widehat{PD}_{h,g}^{VKA}(t|x)$. En esta figura queda constatada la mejora que supone en la estimación de la probabilidad de mora el uso del estimador suavizado; aunque en la función de supervivencia la diferencia entre uno y otro estimador es imperceptible.

Anteriormente se vio que la elección del parámetro de suavizado en tiempo mediante el método del k vecino más próximo no mejoraba la estimación con una ventana global. Por este motivo, en este apartado se descarta dicho método y no se plantea una suavización con ventana local para el estimador propuesto por Van Keilegom y Akritas.

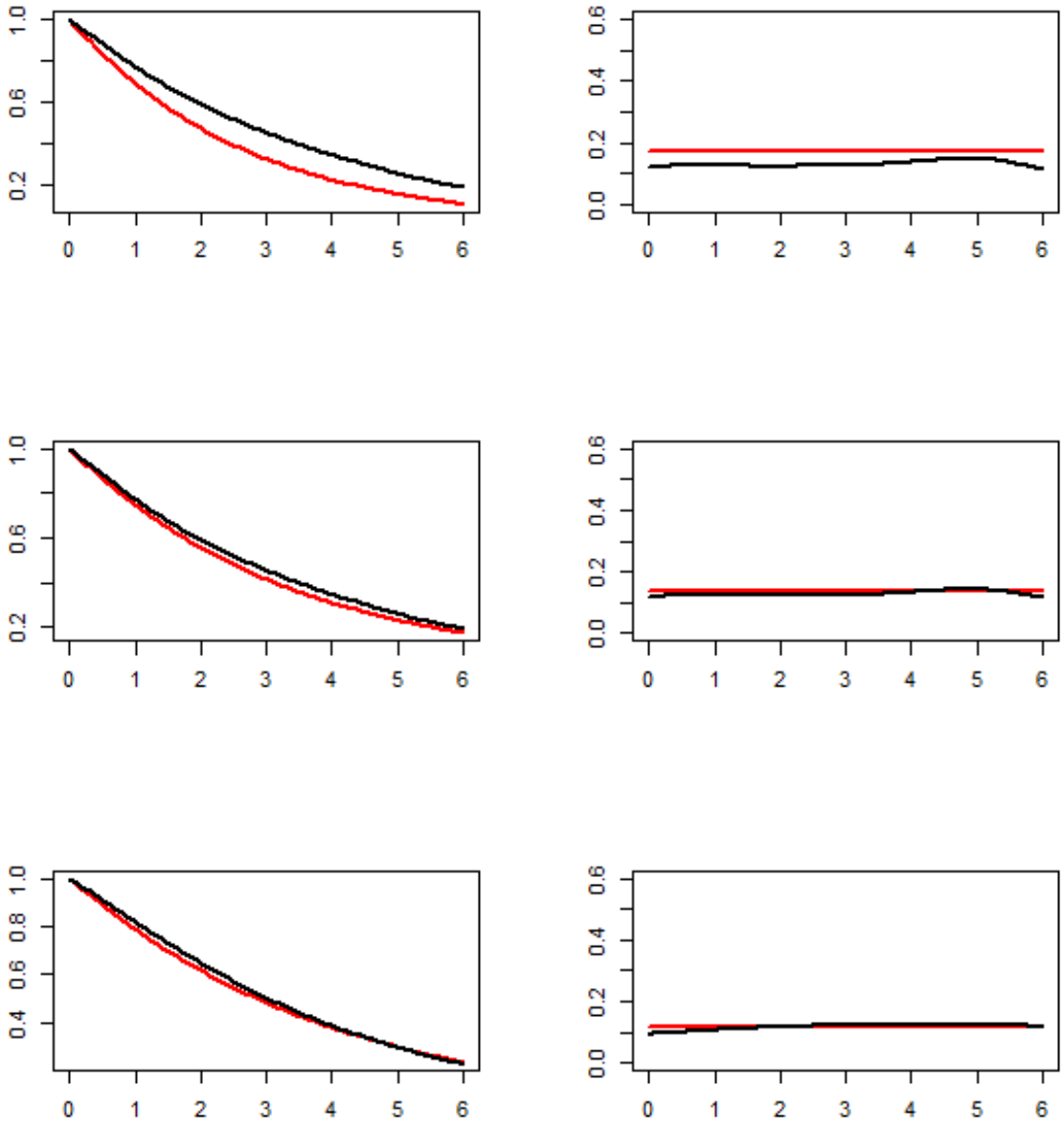


Figura 4.6: Izquierda: Supervivencia condicional (línea roja) y su estimación $\widehat{S}_{h,g}^{VKA}(t|x)$ (línea negra). Derecha: Probabilidad de mora (línea roja) y su estimación $\widehat{PD}_{h,g}^{VKA}(t|x)$ (línea negra). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 1.

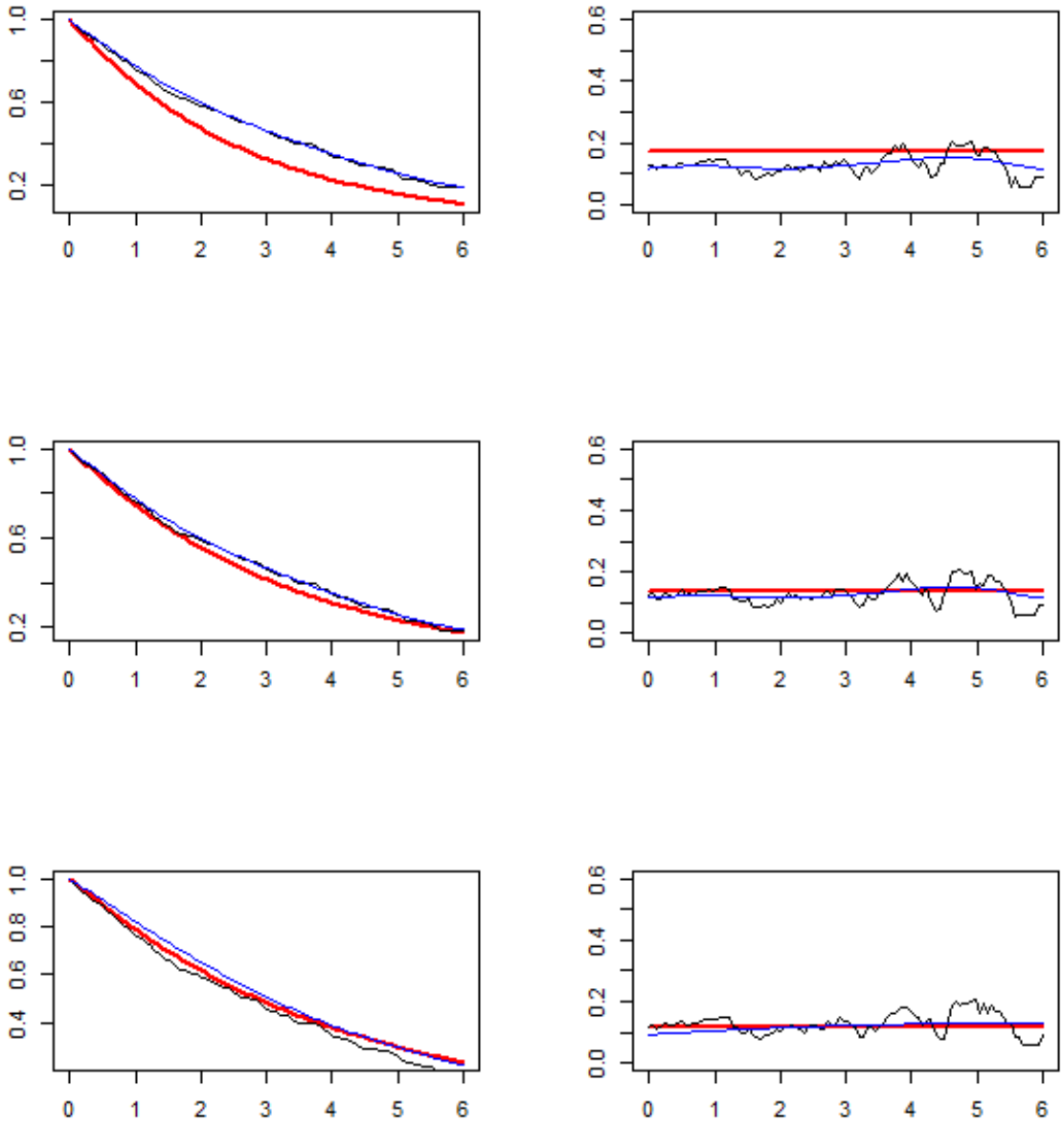


Figura 4.7: Izquierda: Supervivencia condicional (línea roja), su estimación por $\widehat{S}_h^{VKA}(t|x)$ (línea negra) y por $\widehat{S}_{h,g}^{VKA}(t|x)$ (línea azul). Derecha: Probabilidad de mora (línea roja), su estimación por $\widehat{PD}_h^{VKA}(t|x)$ (línea negra) y por $\widehat{PD}_{h,g}^{VKA}(t|x)$ (línea azul). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 1.

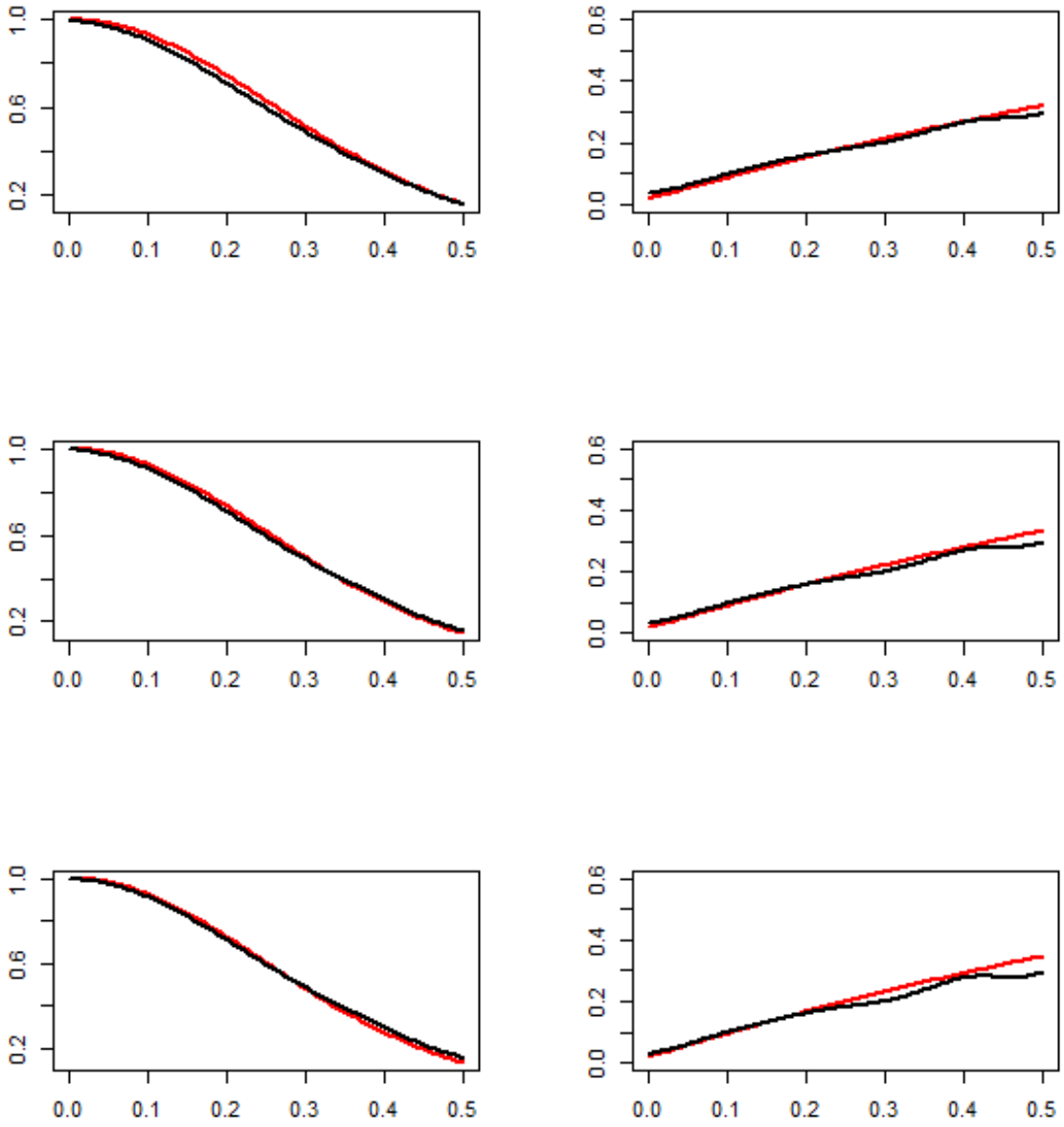


Figura 4.8: Izquierda: Supervivencia condicional (línea roja) y su estimación $\widehat{S}_{h,g}^{VKA}(t|x)$ (línea negra). Derecha: Probabilidad de mora (línea roja) y su estimación $\widehat{PD}_{h,g}^{VKA}(t|x)$ (línea negra). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 2.

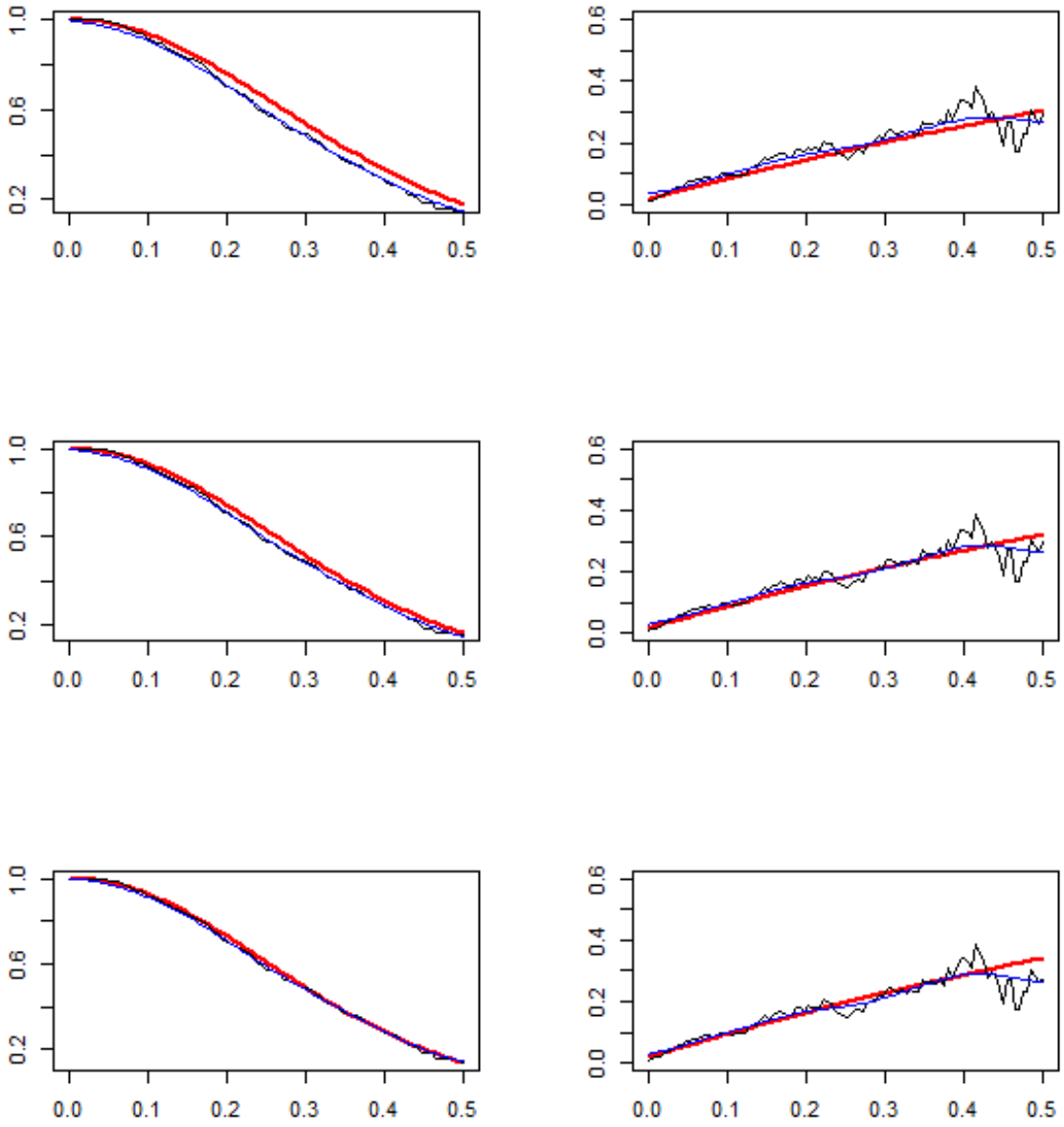


Figura 4.9: Izquierda: Supervivencia condicional (línea roja), su estimación por $\widehat{S}_h^{VKA}(t|x)$ (línea negra) y por $\widehat{S}_{h,g}^{VKA}(t|x)$ (línea azul). Derecha: Probabilidad de mora (línea roja), su estimación por $\widehat{PD}_h^{VKA}(t|x)$ (línea negra) y por $\widehat{PD}_{h,g}^{VKA}(t|x)$ (línea azul). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 2.

Capítulo 5

Comparación de los estimadores de la PD basados en los estimadores de Beran, Cai y Van Keilegom y Akritas

En este capítulo se compara el comportamiento de los estimadores propuestos por Beran, Cai y Van Keilegom y Akritas para estimar la función de supervivencia condicional, así como el comportamiento de los estimadores de la probabilidad de mora construidos a partir de los mismos en las muestras 1 y 2. En primer lugar, considérense los estimadores sin suavización en la variable tiempo.

En la figura 5.1, se muestran las gráficas de la supervivencia condicional y la probabilidad de mora estimadas por estos tres métodos para la muestra 1. El estimador de Cai proporciona una buena estimación de la supervivencia condicionada a cualquiera de los tres cuartiles de la covariable X, mientras que el estimador de Beran resulta ligeramente peor en el primer cuartil, mejorando notablemente el ajuste a la verdadera curva en el tercero. La estimación de la supervivencia condicional obtenida con el estimador de Van Keilegom y Akritas es muy similar a la obtenida con Beran para los tres cuartiles. Por otro lado, el estimador de Beran de la supervivencia condicional parece proporcionar mejores estimaciones de la probabilidad de mora que el estimador de Cai, el cual presenta en este caso mayor variabilidad, especialmente en la cola derecha. De nuevo, la estimación de la PD

obtenida a partir del estimador de Van Keilegom y Akritas de la supervivencia es similar a la obtenida con Beran. Las tres estimaciones presentan una variabilidad más acusada en la cola derecha, pero el estimador de Cai es con diferencia el que peores resultados arroja en este caso.

Los valores de la raíz del error cuadrático integrado para la estimación de la PD mostrados en la tabla 5.1 constatan estas observaciones: el error cometido mediante el estimador $\widehat{PD}_h^B(t|x)$ para la estimación de la PD en la muestra 1 es menor que el cometido mediante $\widehat{PD}_h^C(t|x)$, quedándose también ligeramente por debajo del cometido con $\widehat{PD}_h^{VKA}(t|x)$. En los tres casos se reduce conforme aumenta el valor de la covariable.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
$RECI_{\widehat{PD}_h^B(t\|x)}$	0.050	0.031	0.031
$RECI_{\widehat{PD}_h^C(t\|x)}$	0.097	0.047	0.043
$RECI_{\widehat{PD}_h^{VKA}(t\|x)}$	0.054	0.036	0.037

Tabla 5.1: Valores de $RECI$ para $\widehat{PD}_h^B(t|x)$, $\widehat{PD}_h^C(t|x)$ y $\widehat{PD}_h^{VKA}(t|x)$ en la muestra 1.

En la figura 5.2 aparecen las gráficas de la supervivencia condicional y la PD obtenidas mediante los tres estimadores para la muestra 2. Es claro que, en este caso, las estimaciones de Beran y Van Keilegom y Akritas de la supervivencia condicional se ajustan mucho mejor a la verdadera curva que la estimación de Cai, especialmente para el tercer cuartil. En la estimación de la probabilidad de mora los tres métodos presentan los mismos problemas: excesiva variabilidad sobretodo en la cola derecha. Sin embargo, en esta muestra los estimadores $\widehat{PD}_h^B(t|x)$ y $\widehat{PD}_h^{VKA}(t|x)$ parecen dar mejores resultados que el construido a partir del estimador de Cai. En efecto, en la tabla 5.2 se puede comprobar que el error medio cometido en la estimación mediante $\widehat{PD}_h^B(t|x)$ es similar al cometido en la estimación mediante $\widehat{PD}_h^{VKA}(t|x)$ y en ambos casos menor al cometido con el estimador $\widehat{PD}_h^C(t|x)$. Aunque el valor del $RECI$ es menor en la estimación condicionada al cuantil $Q_{0.25}$ para el estimador de Cai, el error cometido mediante $\widehat{PD}_h^C(t|x)$ aumenta ligeramente en la estimación para el cuantil $Q_{0.5}$ y es notablemente mayor para el cuantil $Q_{0.75}$ que el

error cometido con los otros dos estimadores.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
$RECI_{\widehat{PD}_h^B(t\|x)}$	0.035	0.034	0.037
$RECI_{\widehat{PD}_h^C(t\|x)}$	0.033	0.044	0.073
$RECI_{\widehat{PD}_h^{VKA}(t\|x)}$	0.039	0.036	0.038

Tabla 5.2: Valores de $RECI$ para $\widehat{PD}_h^B(t|x)$, $\widehat{PD}_h^C(t|x)$ y $\widehat{PD}_h^{VKA}(t|x)$ en la muestra 2.

En capítulos anteriores quedó patente que la suavización propuesta para la variable tiempo proporcionaba mejores resultados en la estimación de la probabilidad de mora. En la figura 5.3 se puede comparar el comportamiento de los estimadores de Beran, Cai, y Van Keilegom y Akritas suavizados en la variable tiempo para estimar la supervivencia condicional en la muestra 1. En la mediana de la covariable las estimaciones de la supervivencia condicional obtenidas por los tres métodos son razonables y muy similares. Sin embargo, en el cuartil $Q_{0.25}$ los estimadores de Beran y Van Keilegom y Akritas sobrestiman la verdadera curva de supervivencia, mientras que para el cuantil $Q_{0.75}$ de X en la muestra 1 es el estimador de Cai el que se aleja de ella. En la misma figura se pueden ver las estimaciones de la probabilidad de mora mediante estos estimadores suavizados; los resultados obtenidos con los tres métodos son razonables y tienden a mejorar al aumentar el valor de la covariable X, pues en el primer cuartil, como consecuencia de la sobrestimación de la supervivencia, se produce una infraestimación de la PD.

En la tabla 5.3 se puede comprobar que en el primer cuartil de X el error medio cometido al estimar la probabilidad de mora mediante $\widehat{PD}_{h,g}^C(t|x)$ es menor que mediante $\widehat{PD}_{h,g}^B(t|x)$ o $\widehat{PD}_{h,g}^{VKA}(t|x)$. Aunque el valor del $RECI$ para el segundo y tercer cuartil de la covariable es menor para los dos últimos estimadores, resulta ser mucho mayor para el primer cuartil. Podría inferirse de esta tabla que el comportamiento del estimador $\widehat{PD}_{h,g}^C(t|x)$ en la muestra 1 se mantiene más constante al variar el valor x al que se condiciona; mientras que los estimadores $\widehat{PD}_{h,g}^B(t|x)$ y $\widehat{PD}_{h,g}^{VKA}(t|x)$ se ven más influenciados por él.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
$RECI_{\widehat{PD}^{B}_{h,g}(t\|x)}$	0.037	0.011	0.008
$RECI_{\widehat{PD}^{C}_{h,g}(t\|x)}$	0.016	0.015	0.015
$RECI_{\widehat{PD}^{VKA}_{h,g}(t\|x)}$	0.042	0.010	0.009

Tabla 5.3: Valores de $RECI$ para $\widehat{PD}^{B}_{h,g}(t|x)$, $\widehat{PD}^{C}_{h,g}(t|x)$ y $\widehat{PD}^{VKA}_{h,g}(t|x)$ en la muestra 1.

En la figura 5.4 aparecen las gráficas de las supervivencias condicionales teóricas y estimadas para los tres cuartiles de la covariable X en la muestra 2, así como las estimaciones de la probabilidad de mora obtenidas a partir de los estimadores de Beran, Cai y Van Keilegom y Akritas suavizados en la variable tiempo. Nótese que, al igual que ocurría antes, las estimaciones de la supervivencia condicional obtenidas mediante los estimadores de Beran y Van Keilegom y Akritas suavizados mejoran al aumentar el valor de la covariable al que se condiciona, mientras que la estimación obtenida con el estimador de Cai suavizado es muy similar para los tres cuantiles de X. Por el contrario, las estimaciones de la probabilidad de mora condicionada al primer y segundo cuantil son muy similares con los tres métodos y ajustan muy bien la verdadera curva de probabilidad. En el caso del tercer cuantil, el estimador de Cai suavizado mantiene un buen ajuste, mientras que las estimaciones obtenidas a partir de Beran y Van Keilegom son ligeramente peores. Esto se constata con los valores del $RECI$ mostrados en la tabla 5.4, pues no hay diferencias significativas entre los errores de estimación para el primer y segundo cuantil. Sin embargo, el error cometido para el cuantil $Q_{0.75}$ es mayor con $\widehat{PD}^{B}_{h,g}(t|x)$ y $\widehat{PD}^{VKA}_{h,g}(t|x)$.

x	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
$RECI_{\widehat{PD}^{B}_{h,g}(t\|x)}$	0.010	0.012	0.021
$RECI_{\widehat{PD}^{C}_{h,g}(t\|x)}$	0.011	0.013	0.016
$RECI_{\widehat{PD}^{VKA}_{h,g}(t\|x)}$	0.011	0.017	0.023

Tabla 5.4: Valores de $RECI$ para $\widehat{PD}^{B}_{h,g}(t|x)$, $\widehat{PD}^{C}_{h,g}(t|x)$ y $\widehat{PD}^{VKA}_{h,g}(t|x)$ en la muestra 2.

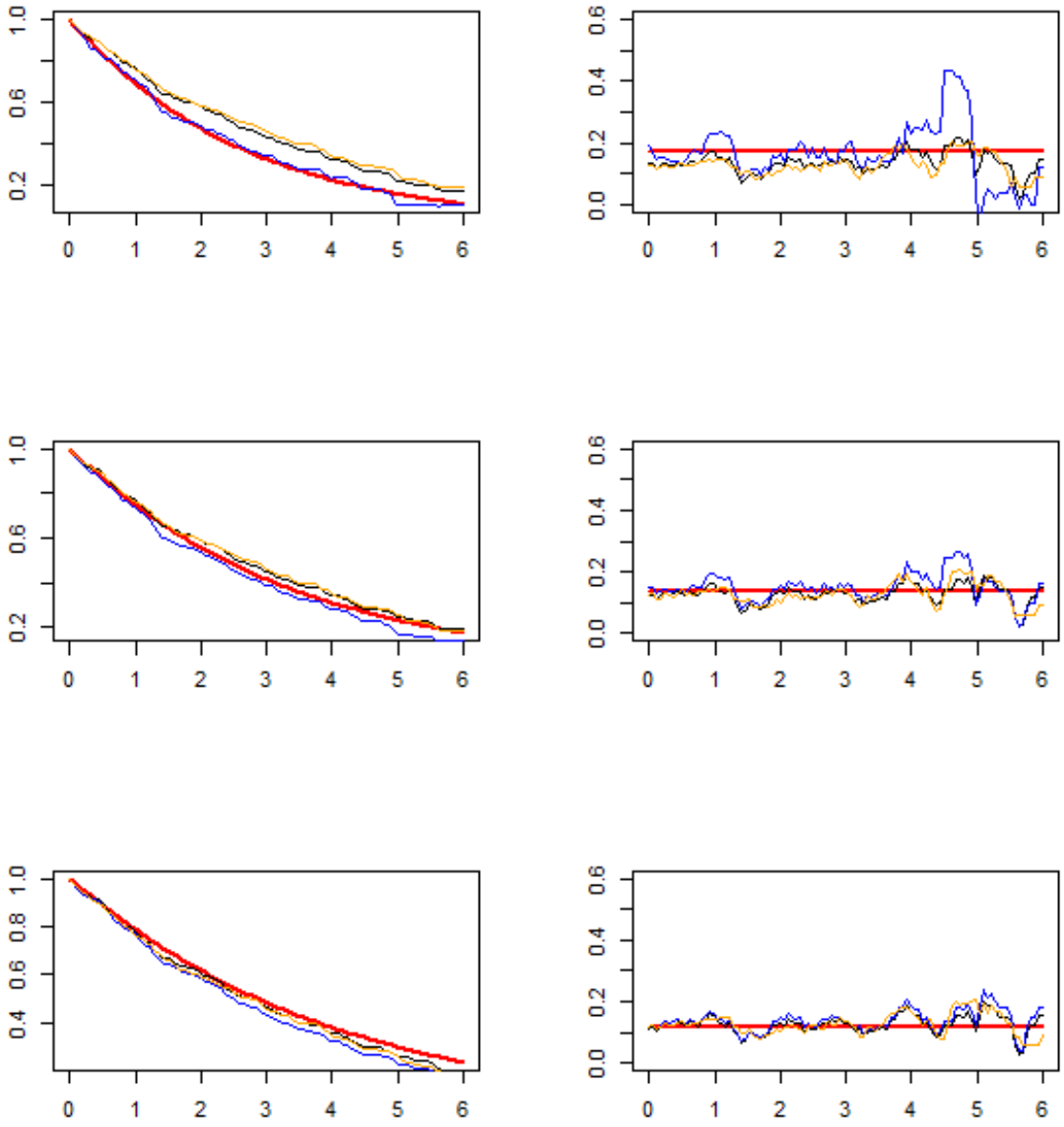


Figura 5.1: Izquierda: Supervivencia condicional (línea roja), su estimación $\widehat{S}_h^B(t|x)$ (línea negra), $\widehat{S}_h^C(t|x)$ (línea azul) y $\widehat{S}_h^{VKA}(t|x)$ (línea naranja). Derecha: Probabilidad de mora (línea roja), su estimación $\widehat{PD}_h^B(t|x)$ (línea negra), $\widehat{PD}_h^C(t|x)$ (línea azul) y $\widehat{PD}_h^{VKA}(t|x)$ (línea naranja). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 1.

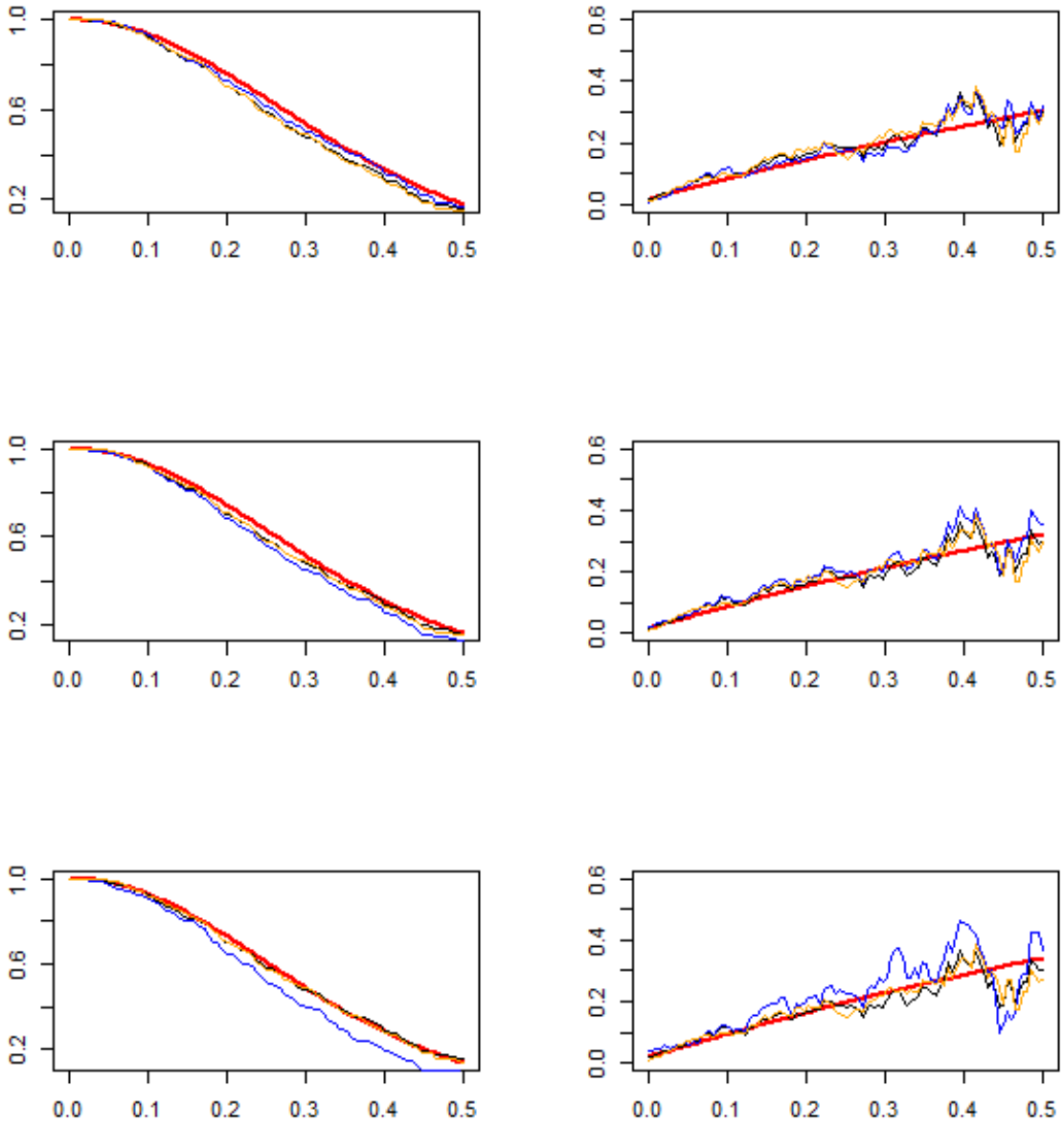


Figura 5.2: Izquierda: Supervivencia condicional (línea roja), su estimación $\widehat{S}_h^B(t|x)$ (línea negra), $\widehat{S}_h^C(t|x)$ (línea azul) y $\widehat{S}_h^{VKA}(t|x)$ (línea naranja) . Derecha: Probabilidad de mora (línea roja), su estimación $\widehat{PD}_h^B(t|x)$ (línea negra), $\widehat{PD}_h^C(t|x)$ (línea azul) y $\widehat{PD}_h^{VKA}(t|x)$ (línea naranja). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 2.

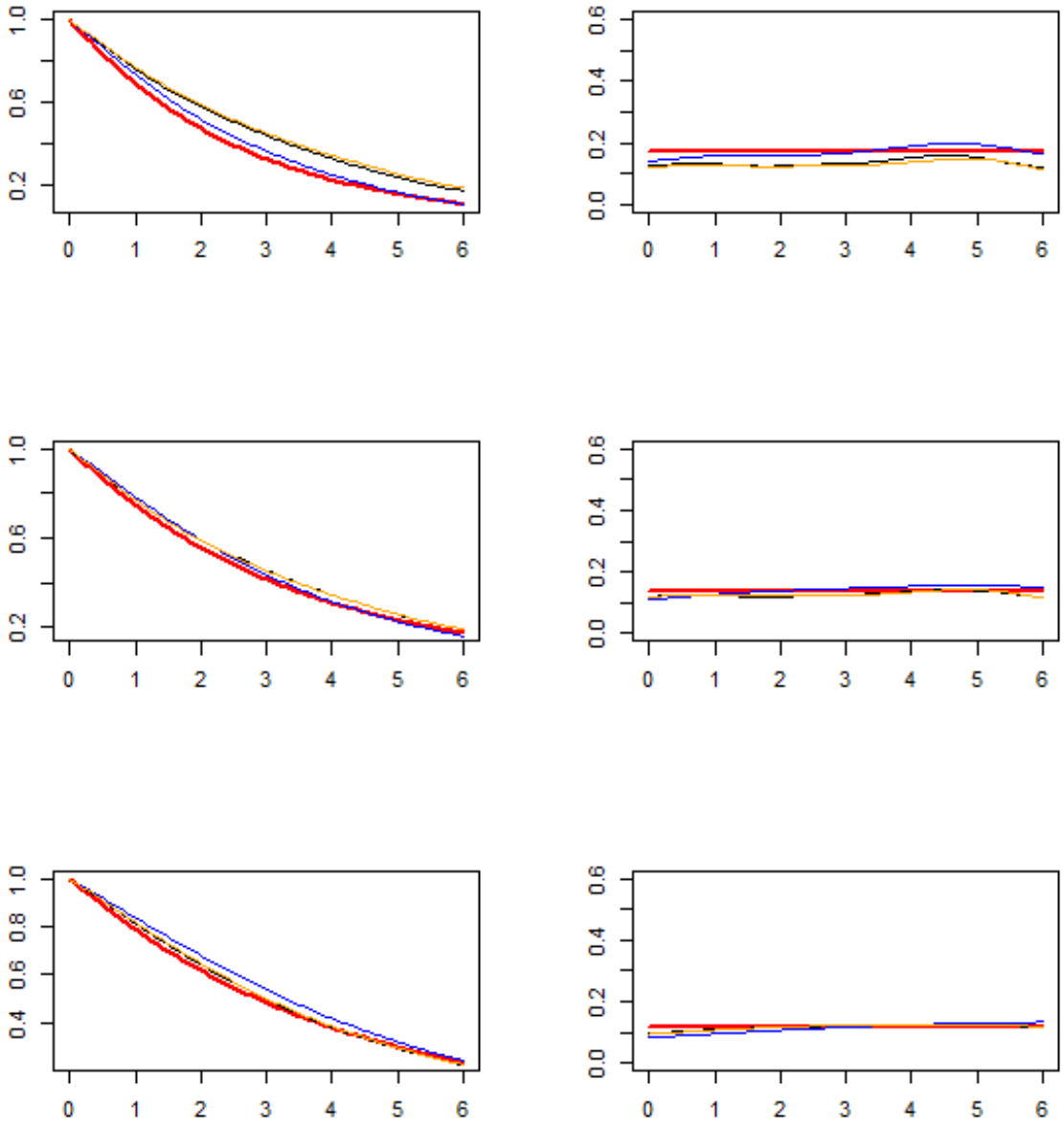


Figura 5.3: Izquierda: Supervivencia condicional (línea roja), su estimación $\widehat{S}^{B}_{h,g}(t|x)$ (línea negra), $\widehat{S}^{C}_{h,g}(t|x)$ (línea azul) y $\widehat{S}^{VKA}_{h,g}(t|x)$ (línea naranja). Derecha: Probabilidad de mora (línea roja), su estimación $\widehat{PD}^{B}_{h,g}(t|x)$ (línea negra), $\widehat{PD}^{C}_{h,g}(t|x)$ (línea azul) y $\widehat{PD}^{VKA}_{h,g}(t|x)$ (línea naranja). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 1.

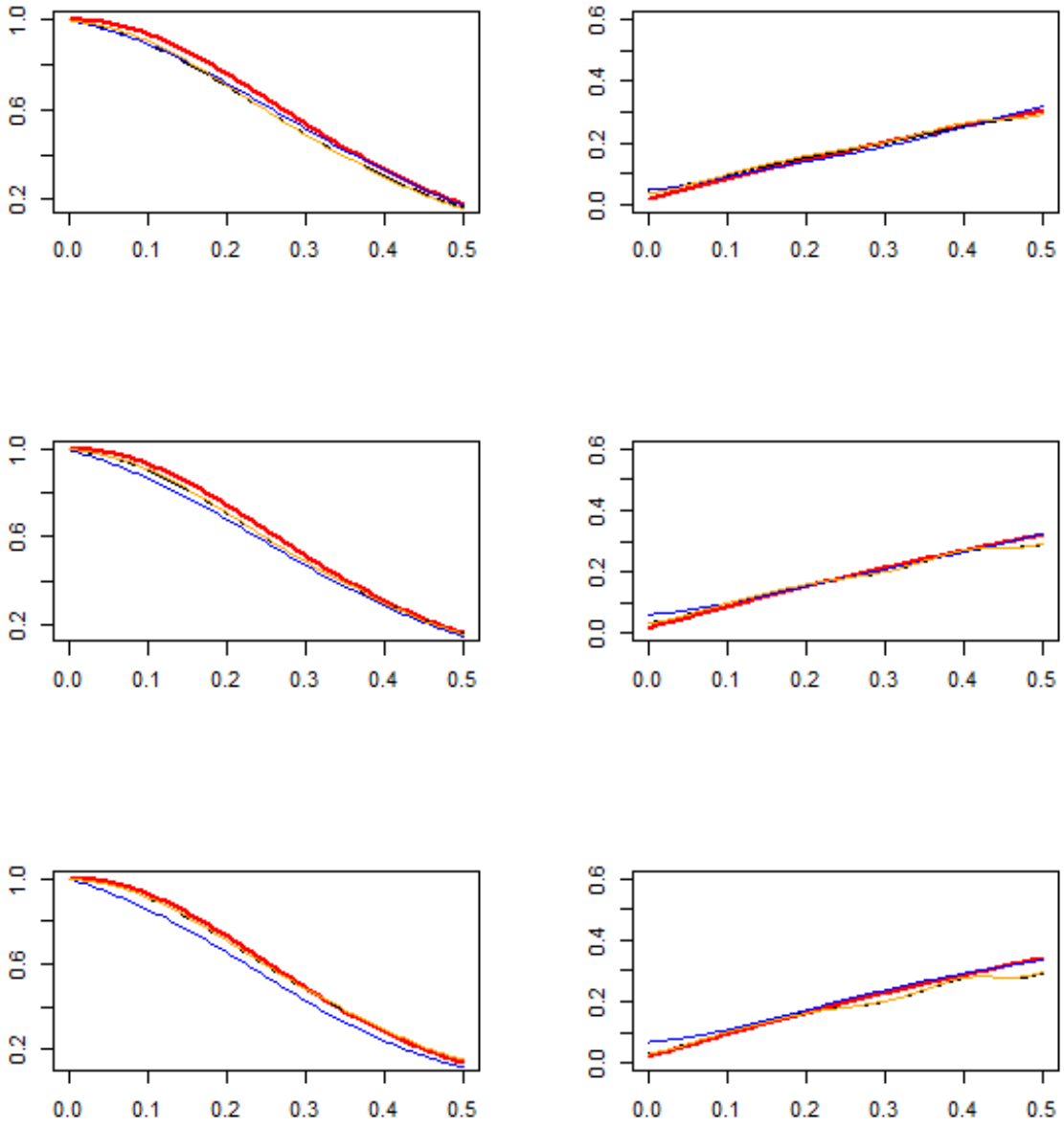


Figura 5.4: Izquierda: Supervivencia condicional (línea roja), su estimación $\widehat{S}^{B}_{h,g}(t|x)$ (línea negra), $\widehat{S}^{C}_{h,g}(t|x)$ (línea azul) y $\widehat{S}^{VKA}_{h,g}(t|x)$ (línea naranja). Derecha: Probabilidad de mora (línea roja), su estimación $\widehat{PD}^{B}_{h,g}(t|x)$ (línea negra), $\widehat{PD}^{C}_{h,g}(t|x)$ (línea azul) y $\widehat{PD}^{VKA}_{h,g}(t|x)$ (línea naranja). De arriba a abajo: $x = Q_{0.25}, Q_{0.5}, Q_{0.75}$. Muestra 2.

5.1. Análisis del Error Cuadrático Medio Integrado

Las comparaciones entre unos y otros estimadores expuestas hasta este momento se han realizado atendiendo a los errores cuadráticos integrados cometidos en la estimación de la PD para las dos muestras consideradas. Es claro que un análisis del error cuadrático medio integrado, es decir, el que se obtiene promediando los errores cuadráticos integrados cometidos en la estimación de la PD en un número suficientemente grande de muestras de cada modelo, daría mucha más información sobre el comportamiento de estos estimadores. En las tablas 5.5 y 5.6 se muestran, precisamente, los errores cuadráticos medios integrados (en adelante $ECMI$) y su raíz ($RECMI$) obtenidos al promediar los ECI cometidos en la estimación de la probabilidad de mora en veinticinco muestras de cada uno de los modelos (tiempos de vida y censura exponenciales para el modelo 1 y tiempos de vida y censura con distribución Weibull para el modelo 2, según lo especificado en la sección 2.3.1) mediante cada uno de los estimadores suavizados en t.

	$\widehat{\mathbf{PD}}^{\mathbf{B}}_{\mathbf{h,g}}(\mathbf{t}\|\mathbf{x})$			$\widehat{\mathbf{PD}}^{\mathbf{C}}_{\mathbf{h,g}}(\mathbf{t}\|\mathbf{x})$			$\widehat{\mathbf{PD}}^{\mathbf{VKA}}_{\mathbf{h,g}}(\mathbf{t}\|\mathbf{x})$		
$\mathbf{x}$	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
$\mathbf{ECMI \cdot 10^{-4}}$	21.90	4.68	0.925	6.07	1.96	1.50	27.56	4.53	0.94
$\mathbf{RECMI \cdot 10^{-2}}$	4.68	2.16	0.961	2.46	1.40	1.22	5.25	2.13	0.97

Tabla 5.5: Valores de $ECMI$ y $RECMI$ para el modelo 1.

	$\widehat{\mathbf{PD}}^{\mathbf{B}}_{\mathbf{h,g}}(\mathbf{t}\|\mathbf{x})$			$\widehat{\mathbf{PD}}^{\mathbf{C}}_{\mathbf{h,g}}(\mathbf{t}\|\mathbf{x})$			$\widehat{\mathbf{PD}}^{\mathbf{VKA}}_{\mathbf{h,g}}(\mathbf{t}\|\mathbf{x})$		
$\mathbf{x}$	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
$\mathbf{ECMI \cdot 10^{-4}}$	1.42	3.09	6.31	2.59	1.61	2.84	2.72	3.14	4.98
$\mathbf{RECMI \cdot 10^{-2}}$	1.19	1.76	2.51	1.61	1.27	1.69	1.65	1.77	2.23

Tabla 5.6: Valores de $ECMI$ y $RECMI$ para el modelo 2.

Los valores del $ECMI$ mostrados en las tablas anteriores se obtienen promediando los errores cuadráticos integrados obtenidos en la estimación de la probabilidad de mora en cada una de las muestras simuladas. Para estimar la PD en cada una de esas 25 muestras se fija la ventana de suavizado en la covariable, h, a un valor razonable y se busca el valor óptimo de la ventana de suavizado en t, g, siguiendo el mismo criterio de selección que en secciones anteriores: se obtiene la estimación en una rejilla de valores de g y se escoge aquel que arroja un menor ECI.

En la tabla 5.5 se muestran los $ECMI$ para el modelo 1 y se puede observar que para los tres estimadores el error cuadrático medio cometido en la estimación de la probabilidad de mora condicional disminuye al aumentar el valor de la covariable al que se condiciona, siendo el estimador de Beran el que presenta un menor error en el tercer cuartil. Sin embargo, el estimador de Cai parece ser la mejor opción. Este estimador presenta, en líneas generales, un menor error cuadrático medio y, además, un comportamiento más constante a lo largo de los valores de la covariable, pues sus valores del $ECMI$ presentan menor variabilidad entre unos y otros cuartiles que los de los otros estimadores.

Los errores cuadráticos medios integrados obtenidos para el modelo 2 con cada uno de los estimadores se muestran en la tabla 5.6. En este caso, el error cometido al estimar la probabilidad de mora condicional aumenta conforme aumenta el valor de x al que se condiciona. El estimador más adecuado para este modelo parece ser, de nuevo, el estimador basado en el de Cai para la supervivencia, pues presenta un menor $ECMI$ que los otros dos estimadores para el segundo y tercer cuartil. Sin embargo, cabe destacar que el estimador de Beran parece ser una mejor opción para estimar la PD condicionada a valores pequeños de la covariable. Por el contrario, el estimador basado en Van Keilegom y Akritas presenta un error cuadrático medio integrado superior al de los estimadores anteriores para cualquiera de los tres cuartiles.

De forma general, el estimador de la PD basado en el de Cai para la supervivencia parece ser la mejor opción en ambos modelos, en cuanto a sus valores del $ECMI$. Aunque en función del valor de x al que se condicione, el estimador de Beran podría arrojar mejores resultados; concretamente, resulta tener mejor comportamiento para valores grandes de la covariable en el modelo 1 y para valores pequeños en el modelo 2.

5.2. Tiempos de computación

Un aspecto importante a tener en cuenta en la implementación de cualquier técnica matemática es el tiempo de CPU que invierte, es decir, el tiempo que emplea la unidad central de procesamiento para procesar las instrucciones de dicho programa. El tiempo de CPU puede considerarse, por tanto, una medida de eficiencia computacional y, en este sentido, se comparará la eficiencia de cada uno de los estimadores presentados.

En las tablas 5.7 y 5.8 se muestra el tiempo que tardan en ejecutarse las implementaciones para estimar la probabilidad de mora mediante los estimadores sin suavización en t. En la tabla 5.7 se anota el tiempo de CPU en segundos necesario para obtener una estimación puntual, en un tiempo t_0, para diferentes tamaños muestrales y considerando un mismo parámetro ventana en los tres estimadores. Se puede ver cómo los estimadores de Beran y Cai para la PD apenas se ven afectados por el aumento del tamaño muestral; además, en estas condiciones son muy semejantes en tiempos de CPU. Por el contrario, el estimador de Van Keilegom y Akritas para la PD es del orden de cientos de veces más lento que el de Beran, llegando a invertir unos nueve minutos en la ejecución para $n = 400$. En la tabla 5.8 se muestra el tiempo de CPU en segundos necesario para obtener una estimación de la probabilidad de mora mediante los tres estimadores sin suavización en tiempo, con tamaño muestral $n = 400$ y en una rejilla de tiempos de tamaño cien. Tanto el estimador de Beran como el de Cai aumentan su tiempo de CPU con respecto a la estimación puntual, mientras que el estimador de Van Keilegom y Akritas para la PD no se ve afectado por el aumento de valores de t en los que obtener la estimación.

En las tablas 5.9 y 5.10 se muestran los tiempos de CPU en segundos invertidos por los tres estimadores suavizados en la variable tiempo para obtener una estimación de la probabilidad de mora bajo diferentes condiciones. En la tabla 5.9 se pueden ver los tiempos de CPU necesarios para obtener una estimación puntual, en un tiempo t_0 con cada estimador y diferentes tamaños muestrales. El estimador de Van Keilegom y Akritas suavizado en el tiempo se ve terriblemente afectado por el aumento del tamaño muestral, aunque su eficiencia es muy semejante a la del estimador sin suavización. Los estimadores de Beran y Cai suavizados en t presentan, de nuevo, tiempos de CPU similares y la suavización provoca que ambos se vean influenciados por el tamaño muestral, aumentando

ligeramente el tiempo de computación al aumentar el valor de n. En la tabla 5.10 se muestran los tiempos de CPU invertidos en la estimación de la PD en una rejilla de tiempos de tamaño cien y un tamaño muestral de $n = 400$ para cada uno de los estimadores suavizados. Tal y como ocurrió con su versión no suavizada en t, el estimador de Van Keilegom y Akritas no aumenta el tiempo de CPU al aumentar el número de valores de t en los que obtener la estimación de la probabilidad de mora, pues en ambos casos emplea aproximadamente trece minutos. Lo mismo ocurre con el estimador de Beran: el tiempo de CPU que el estimador de Beran suavizado invierte para estimar la PD en un punto t_0 con $n = 400$ resulta ser aproximadamente el mismo que el tiempo que invierte en obtener la PD en una rejilla de cien valores. Por el contrario, el tiempo de CPU del estimador de Cai es lineal en el tamaño de la rejilla, invirtiendo unos 190 segundos en estimar 100 valores de la PD frente a los 1.8 segundos que invierte en estimar un único valor.

La construcción de cada uno de los tres estimadores aquí presentados es diferente, por lo que es lógico que la implementación de cada uno de ellos también lo sea y, como consecuencia, es de esperar que el tiempo de CPU varíe de unos a otros. Sin embargo, algunas de las diferencias observadas en las tablas anteriores resultan algo sorprendentes. Por ejemplo, resulta extraño que los estimadores de Beran y Cai para la PD tengan un tiempo de computación tan similar bajo ciertas condiciones, pero que este tiempo se dispare para el estimador de Cai sobre una rejilla de tamaño cien. Una explicación para esto podría estar en la forma de implementarlo: mientras que el programa para el estimador de Beran puede vectorizarse en t, y con ello optimizarse, esto no fue posible para el estimador de Cai, pues para obtener la estimación de $PD(t|x)$ es necesario pasar del conjunto de datos original $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$ al conjunto $\{(X_i, I_{\{Z>t\}}, \delta_i)\}_{i=1}^n$ y esto ha de hacerse para cada tiempo t donde se quiere hallar la estimación.

Por otro lado, sorprende también el tiempo de CPU empleado por la implementación del estimador de Van Keilegom y Akritas y la poca influencia que tienen en él tanto la suavización como el tamaño de la rejilla de tiempos. Esto se debe a que la parte realmente lenta es el cálculo de los residuos ajustados de la regresión, que dependen directamente de la muestra, y por tanto del tamaño muestral, pero no dependen del valor de t donde se pretende obtener la estimación y tampoco de la suavización en la variable tiempo.

	$\widehat{\mathbf{PD}}_{\mathbf{h}}^{\mathbf{B}}(\mathbf{t}\|\mathbf{x})$	$\widehat{\mathbf{PD}}_{\mathbf{h}}^{\mathbf{C}}(\mathbf{t}\|\mathbf{x})$	$\widehat{\mathbf{PD}}_{\mathbf{h}}^{\mathbf{VKA}}(\mathbf{t}\|\mathbf{x})$
$n = 50$	0.05	0.07	0.90
$n = 100$	0.07	0.08	5.00
$n = 200$	0.06	0.08	46.00
$n = 400$	0.07	0.09	564.48

Tabla 5.7: Tiempo de CPU en segundos para la estimación en $\mathbf{t} = \mathbf{t_0}$, $\mathbf{x} = \mathbf{Q_{0.5}}$.

	$\widehat{\mathbf{PD}}_{\mathbf{h}}^{\mathbf{B}}(\mathbf{t}\|\mathbf{x})$	$\widehat{\mathbf{PD}}_{\mathbf{h}}^{\mathbf{C}}(\mathbf{t}\|\mathbf{x})$	$\widehat{\mathbf{PD}}_{\mathbf{h}}^{\mathbf{VKA}}(\mathbf{t}\|\mathbf{x})$
$n = 400$	0.4	0.5	570.0

Tabla 5.8: Tiempo de CPU en segundos para la estimación en $\mathbf{t} \in \{\mathbf{t_1}, ..., \mathbf{t_{100}}\}$, $\mathbf{x} = \mathbf{Q_{0.5}}$.

	$\widehat{\mathbf{PD}}_{\mathbf{h,g}}^{\mathbf{B}}(\mathbf{t}\|\mathbf{x})$	$\widehat{\mathbf{PD}}_{\mathbf{h,g}}^{\mathbf{C}}(\mathbf{t}\|\mathbf{x})$	$\widehat{\mathbf{PD}}_{\mathbf{h,g}}^{\mathbf{VKA}}(\mathbf{t}\|\mathbf{x})$
$n = 50$	0.07	0.07	0.7
$n = 100$	0.12	0.20	5.00
$n = 200$	0.3	0.4	45.0
$n = 400$	1.6	1.8	768.0

Tabla 5.9: Tiempo de CPU en segundos para la estimación en $\mathbf{t} = \mathbf{t_0}$, $\mathbf{x} = \mathbf{Q_{0.5}}$.

	$\widehat{\mathbf{PD}}_{\mathbf{h,g}}^{\mathbf{B}}(\mathbf{t}\|\mathbf{x})$	$\widehat{\mathbf{PD}}_{\mathbf{h,g}}^{\mathbf{C}}(\mathbf{t}\|\mathbf{x})$	$\widehat{\mathbf{PD}}_{\mathbf{h,g}}^{\mathbf{VKA}}(\mathbf{t}\|\mathbf{x})$
$n = 400$	1.7	191.5	750.1

Tabla 5.10: Tiempo de CPU en segundos para la estimación en $\mathbf{t} \in \{\mathbf{t_1}, ..., \mathbf{t_{100}}\}$, $\mathbf{x} = \mathbf{Q_{0.5}}$.

Capítulo 6

Análisis con datos reales

En este capítulo se estima la probabilidad de mora condicionada a la puntuación crediticia de una muestra de 10000 préstamos personales con un 92.8 % de censura. Se trata de créditos personales de una entidad financiera española concedidos entre julio de 2004 y noviembre de 2006[1]. Estos datos ya fueron utilizados en Devia (2016).

Se dispone de una muestra aleatoria simple de la terna (X, Z, δ), esto es, $\{(X_i, Z_i, \delta_i)\}_{i=1}^{n}$ con $n = 10000$ siendo:

- X es la puntuación crediticia que toma valores entre 0 y 1 donde los valores más altos indican mayor solvencia,
- Z es el tiempo de vida observado del crédito medido en meses y toma valores entre 0 y 30,
- δ es el indicador de no censura.

En la figura 6.1 se muestran los histogramas del tiempo de vida observado en el grupo de datos censurados y en el grupo de datos no censurados. Se puede ver que la muestra no censurada presenta entre 0 y 6 la mayor proporción de tiempos de vida; de hecho, la mediana de esta variable en la muestra es 5.2 y el primer y tercer cuartil son 2.8 y 10.5, respectivamente. La mediana del tiempo de vida en la muestra censurada asciende a 10.9,

[1]La proporción de créditos morosos ha sido modificada con respecto a los datos originales por motivos de confidencialidad.

siendo 6.7 el primer cuartil y 20.0 el tercero. De esto se puede deducir que, en general, los créditos que resultan morosos tienen tiempos de vida inferiores.

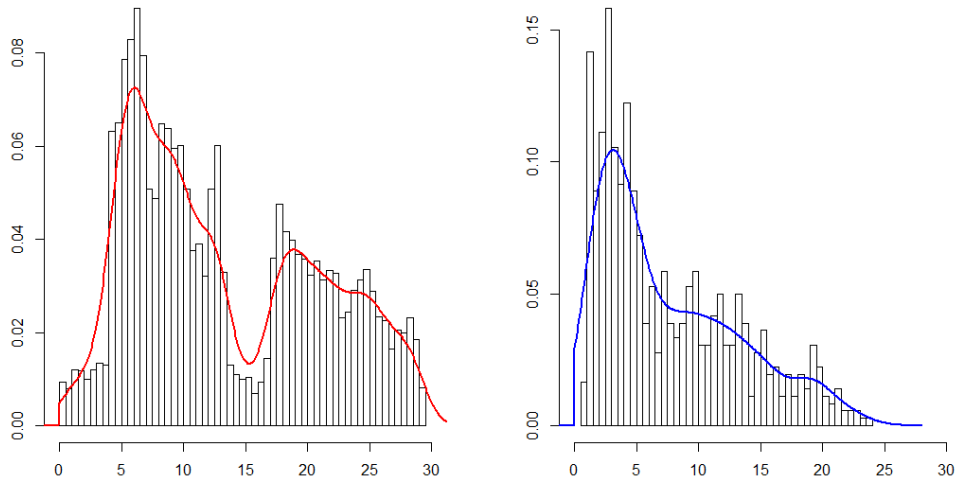


Figura 6.1: Izquierda: Histograma de $\{Z_i : \ \delta_i = 0\}_{i=1}^n$ (muestra censurada) y su densidad estimada. Derecha: Histograma de $\{Z_i : \ \delta_i = 1\}_{i=1}^n$ (muestra no censurada) y su densidad estimada.

En la figura 6.2 se muestran los histogramas de la puntuación crediticia de créditos censurados y no censurados. Como era de esperar, los créditos censurados, es decir, los que no han caído en impago durante su estudio, tienen valores de la puntuación crediticia muy altos, encontrándose el 75 % de los datos entre 0.88 y 0.97 y siendo su mediana igual a 0.94. Por el contrario, la puntuación crediticia de créditos morosos es más variable y en general toma valores más pequeños, aunque sus valores medios siguen siendo altos. Esto es razonable, teniendo en cuenta que son puntuaciones crediticias correspondientes a clientes con créditos ya concedidos por la entidad financiera. Concretamente, su primer cuartil es 0.61 y el tercero 0.76, siendo la mediana 0.70.

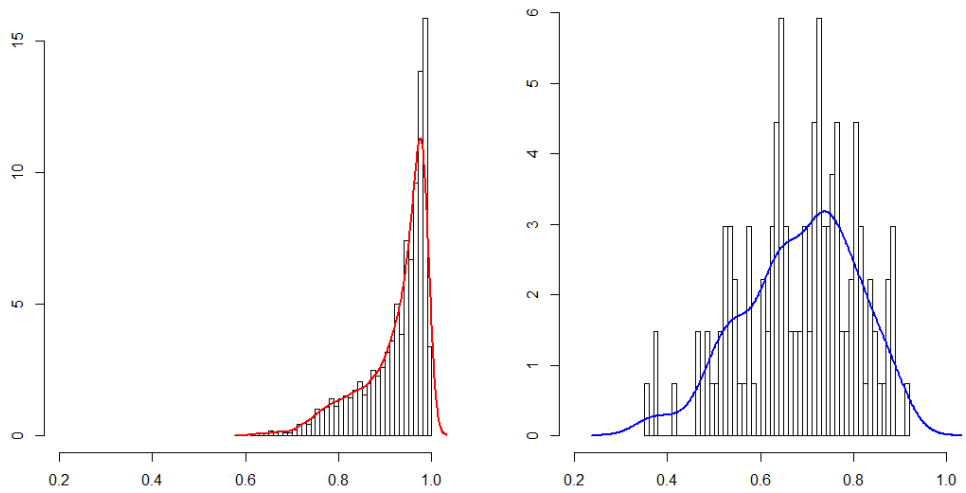


Figura 6.2: Izquierda: Histograma de $\{X_i : \delta_i = 0\}_{i=1}^n$ (muestra censurada) y su densidad estimada. Derecha: Histograma de $\{X_i : \delta_i = 1\}_{i=1}^n$ (muestra no censurada) y su densidad estimada.

A continuación se obtiene la estimación de la probabilidad de mora a horizonte b, con b igual a un mes, en una rejilla de tiempos en el intervalo $[0, 25]$ mediante el estimador de Beran suavizado en t para la supervivencia condicional, puesto que la suavización en el tiempo proporciona buenas estimaciones de la PD y, computacionalmente, el estimador de Beran es el más eficiente de los tres presentados. Dado que la ventana de suavizado en la covariable h tiene un efecto bastante leve en la estimación de esta probabilidad, se fija a un valor razonable y se obtiene la estimación de la PD para varios valores de la ventana de suavizado en el tiempo, g, en los tres cuartiles de la covariable X; concretamente, para $x = 0.87$, $x = 0.95$ y $x = 0.97$. En la figura 6.3 se muestran estas estimaciones.

Se puede observar que la estimación de la probabilidad de mora es muy similar en los tres cuartiles de la covariable, y en los tres casos una mayor suavización en la variable tiempo, dada por un valor más alto de g, parece proporcionar mejores resultados.

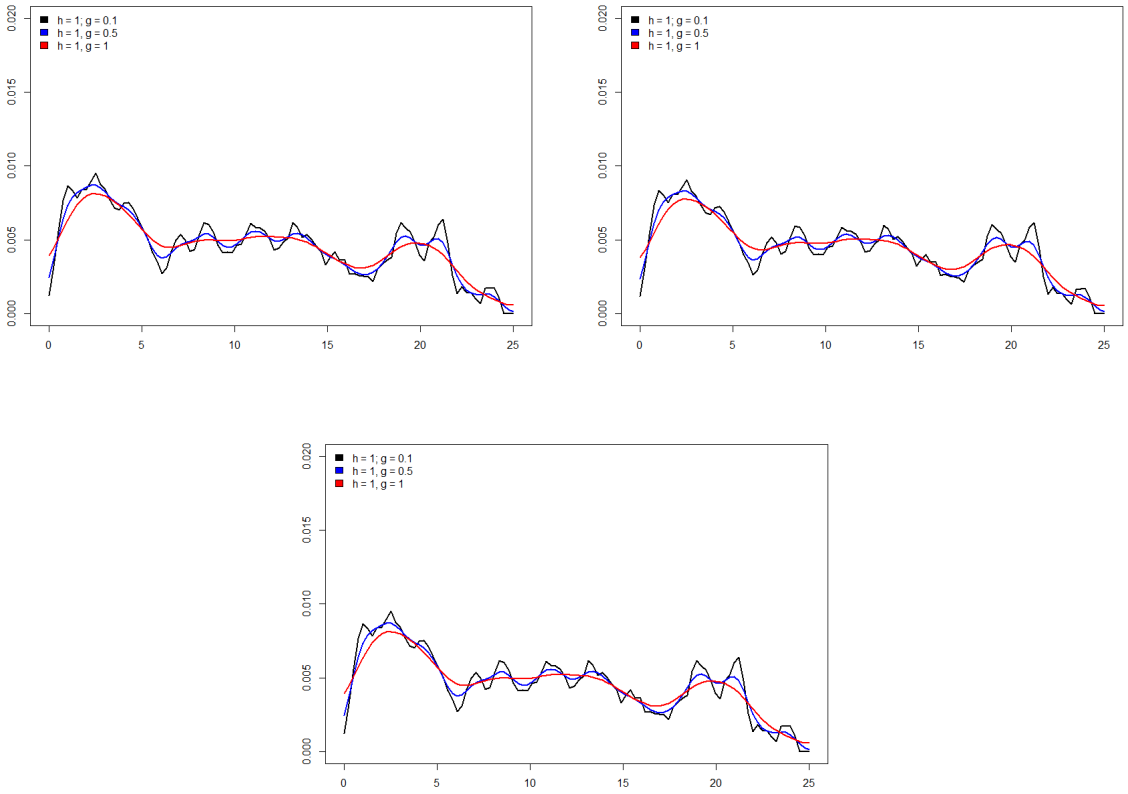


Figura 6.3: Estimación de $PD(t|x)$ a horizonte $b = 1$ mediante $\widehat{PD}^{B}_{h,g}$ en la muestra de créditos personales para distintos parámetros de suavizado en t y distintos valores de la covariable: $x = 0.87$ (arriba izquierda), $x = 0.95$ (arriba derecha) y $x = 0.97$ (abajo).

En la figura 6.4 se muestran las estimaciones de la PD para cada cuartil obtenidas para el mayor valor de la ventana g considerado. Cabe destacar que la probabilidad de mora tiene una tendencia, en general, decreciente y cercana a cero en todo punto. De lo primero se deduce que la probabilidad de caer en impago se reduce conforme pasan los meses tras contraer la deuda. Lo segundo es razonable, dado que se está calculando la probabilidad de mora para valores notablemente altos de la covariable, que indican mayor solvencia del acreditado. Se aprecian pocas diferencias entre las curvas de probabilidad de mora para cada uno de los cuartiles y no es posible determinar si esto se debe a que realmente son curvas similares o a que el estimador propuesto no es capaz de detectar las diferencias existentes. En cualquier caso, se observa que la curva de probabilidad de mora correspondiente al primer cuartil queda por encima de las otras dos curvas, indicando mayor probabilidad de caer en impago en todo tiempo t. Esto es de esperar dado que el primer

cuartil se corresponde con un menor valor de la covariable y, por tanto, menor capacidad de hacer frente a la deuda. Los cuartiles segundo y tercero de la puntuación crediticia toman valores cercanos, de ahí que las curvas de probabilidad de mora correspondientes sean aún más similares, quedando la obtenida para la mediana ligeramente por encima de la otra a lo largo de todos los valores de tiempo.

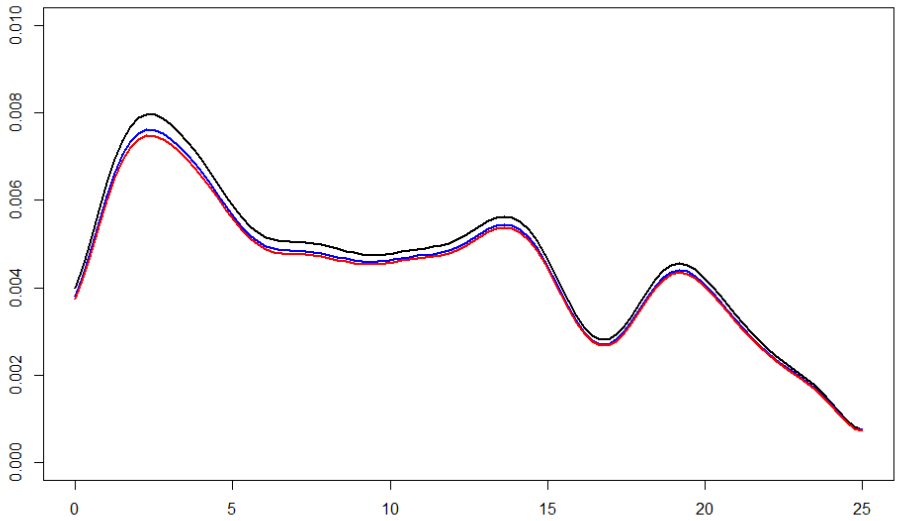


Figura 6.4: Estimación de $PD(t|x)$ a horizonte $b = 1$ en la muestra de créditos personales mediante $\widehat{PD}^{B}_{h,g}$ para $x = 0.87$ (línea negra), $x = 0.95$ (línea azul) y $x = 0.97$ (línea roja).

Capítulo 7

Conclusiones y trabajo futuro

Al comienzo de este trabajo se presentó una expresión para la probabilidad de mora a horizonte b condicionada a una puntuación crediticia, x, en términos de las funciones de supervivencia en tiempos t y $t + b$. Se presentaron tres estimadores de la supervivencia condicional que, mediante dicha transformación, pasaron a ser estimadores de la PD: el estimador límite-producto generalizado de Beran, el estimador de Cai y el estimador de Van Keilegom y Akritas. Además, se propusieron versiones suavizadas en la variable tiempo para cada uno de ellos. Los estimadores de la probabilidad de mora resultantes fueron aplicados sobre dos muestras de datos simulados de dos modelos diferentes: uno con tiempos de vida y censura exponenciales y otro con tiempos de vida y censura con distribución Weibull. Para la muestra procedente del modelo exponencial, el estimador de Beran resultó ser el más adecuado. Lo mismo ocurrió con la muestra del modelo Weibull, aunque aquí su comportamiento fue muy similar al del estimador de Van Keilegom y Akritas, mejorando ambos los resultados obtenidos mediante el estimador de Cai. Por tanto, para las muestras consideradas, el estimador de Beran para la PD arrojó los mejores resultados. Por el contrario, fue el estimador de Cai suavizado en tiempo el que proporcionó las mejores estimaciones para la PD en la muestra del modelo exponencial, especialmente para valores pequeños de la covariable. En la muestra procedente del modelo Weibull los tres estimadores suavizados tuvieron un comportamiento muy similar; sin embargo, para valores altos de la covariable la estimación obtenida mediante el estimador de Cai suavizado en el tiempo fue mejor que mediante los otros métodos, por lo que podría decirse que

condicionalmente a las muestras consideradas, el estimador de Cai suavizado en t resultó ser la mejor opción entre los tres estimadores suavizados. Además, se hizo evidente que la suavización en la variable tiempo de cualquiera de los tres estimadores de la supervivencia proporcionaba mejores resultados para estimar la probabilidad de mora que cualquiera de los tres estimadores sin suavizar en estas muestras.

Sin más análisis, es posible afirmar con certeza que la suavización propuesta en la variable tiempo para cada uno de los estimadores de la función de supervivencia condicional resulta en estimadores de la probabilidad de mora considerablemente mejores que aquellos sin suavizar; se elimina el exceso de variabilidad, se reduce la rugosidad y se proporciona un mejor ajuste a la verdadera curva de probabilidad reduciendo el error cuadrático integrado.

Con la intención de comparar el comportamiento de cada uno de los estimadores suavizados en t, se realizó un breve análisis del error cuadrático medio integrado, concluyéndose que para estos modelos el estimador que presenta un mejor comportamiento es el basado en el estimador de Cai para la supervivencia. Aunque cabe destacar que para valores grandes de la covariable en el modelo 1 o valores pequeños en el modelo 2, el estimador de basado en el de Beran proporciona mejores resultados. De todos modos, las principales diferencias entre los estimadores se encontraron en su eficiencia computacional, siendo el estimador de Beran suavizado el más rápido y el estimador de Van Keilegom y Akritas el más lento de los tres.

Otro análisis que podría arrojar información interesante acerca de los estimadores es una comparación de su comportamiento en la cola derecha de la distribución. Recuérdese que el estimador propuesto por Van Keilegom y Akritas para la función de distribución (y, por tanto, para la función de supervivencia condicional) se introdujo por proporcionar buenos resultados precisamente en esta zona, incluso en situaciones de alta censura; sería interesante estudiar si el estimador de la PD hereda esas buenas propiedades del estimador de la supervivencia de Van Keilegom y Akritas. Para ello, bastaría estimar la probabilidad de mora de tiempos en la cola derecha de la distribución para un número suficientemente grande de muestras de cada uno de los modelos aquí considerados para, posteriormente, aproximar el error cuadrático medio a partir del promedio de los cuadrados de los errores de estimación para esas muestras. Este procedimiento daría una idea de en qué medida

cada uno de los estimadores es capaz de aproximar el verdadero valor de la PD en esta zona conflictiva, donde, evidentemente, la censura es mayor. Sin embargo, los tiempos de computación son un inconveniente para realizar este análisis, que podrá ser abordado en el futuro.

No se analizó en esta memoria, quedando pendiente para futuros trabajos en el tema, el efecto que la elección de la función núcleo puede tener en la estimación de la probabilidad de mora, tanto para la suavización en la covariable como para la suavización en la variable tiempo. Tampoco se estableció ningún criterio automático para la selección de la ventana óptima en ninguna de las dos suavizaciones.

La elección de los parámetros de suavizado se realizó en todos los casos en base al error cuadrático integrado. En los casos en los que este dependía tanto de la ventana de suavizado en la covariable como de la ventana de suavizado en el tiempo, la primera era fijada a un valor razonable (no siempre un mínimo del ECI) y se elegía un valor para la segunda que arrojase un menor error. Los métodos para hallar o aproximar los puntos en los que una función bivariante alcanza un mínimo también podrían ser técnicas a considerar en el futuro, puesto que podría incurrirse en error al suponer que la elección de la ventana de suavizado en t no está influida por la ventana de suavizado en x.

Bibliografía

Allen, L. N. and Rose, L. C. (2006). Financial survival analysis of defaulted debtors. Journal of the Operational Research Society, 57(6):630–636.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical report, University of California.

Cai, Z. (2003). Weighted local linear approach to censored nonparametric regression. In Akritas, M. G. and Politis, D. N., editors, Recent Advances and Trends in Nonparametric Statistics, page 217–231.

Cao, R., Vilar, J. M., and Devia, A. (2009). Modelling consumer credit risk via survival analysis (with discussion). Statistics and Operations Research Transactions, 33(1):3–30.

Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional kaplan-meier estimate. The Annals of Statistics, 17(3):1157–1167.

Devia, A. (2016). Contribuciones al análisis estadístico del riesgo de crédito. Tesis doctoral, Universidade da Coruña.

Glennon, D. and Nigro, P. (2005). Measuring the default risk of small business loans: a survival analysis approach. Journal of money, credit and banking, (37):923–947.

Hanson, S. G. and Schuermann, T. (2004). Estimating probabilities of default. Staff Report Federal Reserve Bank of New York, (190):923–947.

Iglesias-Pérez, M. C. and González-Manteiga, W. (1999). Strong representation of a generalized product-limit estimator for truncated and censored data with some applications. Journal of Nonparametric Statistics, 10(3):213–244.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of American Statistical Association, (53):457–481.

Masry, E. (1996). Multivariate regression estimation. local polynomial fitting for time series. Stochastic Processes and their Applications, (65):81–101.

Naraim, B. (1992). Survival analysis and the credit granting decision. In Thomas, L. C., Crook, J. N., and Edelman, D. B., editors, Credit Scorig and Credit Control, Oxford University Press, pages 109–121.

Parzen, E. (1962). On estimation of a probability density function and mode. The Annals of Mathematical Statistics, (33):1065–1076.

Rosenblatt, M. (1956). Remarks on some nonparametric estimate of a density function. The Annals of Mathematical Statistics, (27):832–837.

Silverman, B. W. (1986). Density estimation for statistics and data analysis. Monographs on Statistics and Applied Probability, Chapman and Hall.

Stute, W. (1999). Nonlinear censored regression. Statistica Sinica, (9):1089–1102.

Van Keilegom, I. and Akritas, M. (1999). Transfer of tail information in censored regression models. The Annals of Statistics, 27(5):1745–1784.

Van Keilegom, I., Akritas, M. G., and Veraverbeke, N. (2001). Estimation of the conditional distribution in regression with censored data: a comparative study. Computational Statistics and Data Analysis, (35):487–500.