Trabajo Fin de Máster - Máster en Técnicas Estadísticas

Estimador presuavizado de Kaplan-Meier con indicador de censura perdido aleatoriamente

Alumna: Beatriz López Calviño

Directores: Ricardo Cao Abad y Ewa Strzalkowska Kominiak

Curso: 2011-2012

Ricardo Cao Abad y Ewa Strzalkowska Kominiak, profesores del Departamento de Estadística e Investigación Operativa de la Universidad de A Coruña,

HACEN CONSTAR

que el presente trabajo titulado *Estimador presuavizado de Kaplan-Meier con indicador de censura perdido aleatoriamente* ha sido realizado por Beatriz López Calviño bajo su dirección, para su presentación como Trabajo Fin de Máster en Técnicas Estadísticas.

Fdo.: Ricardo Cao Abad

Fdo. : Ewa Strzalkowska Kominiak

Agradecimientos

Tras la elaboración de este trabajo fin de máster, quiero expresar mi sincero agradecimiento a todos las personas que me han apoyado durante la realización del mismo

En primer lugar a mis directores de proyecto, a Ricardo Cao Abad por brindarme la oportunidad de realizar este trabajo, así como transmitirme sus conocimientos y apoyarme en todo momento. A Ewa Strzalkowska Kominiak por estar siempre dispuesta a ofrecerme su ayuda y enseñarme a simplificar cada procedimiento matemático

A mis compañeros de la Unidad de Epidemiología Clínica y Bioestadística, Salvador Pita Fernández, Sonia Pértega Díaz, Teresa Seoane Pillado por iniciarme en la investigación biomédica y mostrarme cómo la epidemiología junto con la estadística contribuyen a la toma de decisiones en la práctica clínica. Al resto de miembros, Carmen Varela, Rocío Seijo, Yolanda García por realizar el trabajo de campo de recogida de información y mecanización de los datos de cáncer colorrectal sobre los que se aplica la metodología propuesta en este trabajo.

A mi padre que siempre me acompañará, a mi madre por todas sus enseñanzas y su comprensión, a mis hermanos por su cariño y ánimos para seguir adelante.

A mis amigos por tantos momentos buenos.

Gracias a todos

Índice general

Índice de figuras

Índice de tablas

RESUMEN

En los estudios epidemiológicos se estudia frecuentemente la función de distribución de los tiempos de vida Z, que, debido a limitaciones del tiempo, puede estar censurado por una variable C. Por lo tanto, en estos estudios de investigación médica puede producirse el sesgo de información, debido a que la causa de la muerte no es siempre conocida y por tanto, existe una falta de información de si el tiempo de vida observado es censurado o no. Además, puede ocurrir que el indicador de censura $\delta = 1_{\{T \leq C\}}$ esté perdido.

El objetivo principal de este trabajo es estudiar los estimadores de la función de supervivencia con indicadores de censura perdidos aleatoriamente. Para ello, se recuerdan los estimadores propuestos por Wang and Ng (2008) y se introduce un nuevo estimador basado en el estimador presuavizado de Kaplan-Meier introducido por Cao et al (2005).

Este trabajo se organiza de la siguiente forma. En el capítulo 2, se introduce el estimador de Kaplan-Meier con los indicadores de censura (δ) observados. En la sección 2.2 se recuerda la versión conocida del estimador de Kaplan-Meier, con δ tomando valores 1 si el tiempo de vida es observado y 0 en caso contrario. En la sección 2.3, se presenta el estimador presuavizado de Kaplan-Meier donde $\delta = 1_{\{Z \leq C\}}$ es reemplazado por una función.

El objetivo principal de este trabajo es estimar la función de supervivencia en caso en que δ no es siempre observable. Esto se explica en el capítulo 3, donde se recuerda los métodos ya existentes propuestos por Wang and Ng (2008) (sección 3.1) y se definen dos nuevos estimadores (sección 3.2). Además, en el capítulo 2, se muestra que si todos los indicadores de censura fuesen observables, los estimadores propuestos de la función de supervivencia se reducirían al estimador estándar de Kaplan-Meier y al estimador presuavizado de Kaplan-Meier, respectivamente

En el capítulo 4, se realiza un estudio de simulación, donde se comparan los estimadores propuestos con los definidos por Wang and Ng (2008). Se muestra la distancia de Kolmogorov-Smirnov y el error cuadrático medio integrado (MISE)

Por último, en el capítulo 5, se ilustran las metodologías propuestas mediante el análisis de un conjunto de datos reales correspondiente a un estudio de supervivencia de cáncer colorrectal del

Complexo Hospitalario Universitario de A Coruña.

Keywords: Missing at random; product-limit estimator; random censorship; bandwidth selection

ABSTRACT

In medical studies one is often interested in determinating the distribution function of patient's lifetime Z, which, due to time limitations, may be censored by a variable C. In such studies one may face additional information bias, since the cause of death is not always known and hence there is a lack of information if the observed lifetime is censored or not. More precisely, it may happen that the censoring indicator $\delta = 1_{\{T \leq C\}}$ is missing.

The main objective of this work is to study the survival function estimators with censoring indicators missing at random. For this, we first recall the estimators proposed by Wang and Ng (2008) and we introduce a new estimator based on the presmoth Kaplan-Meier estimator introduced by Cao et al. (2005).

The work is organized as follows. In Chapter 2 we introduce the Kaplan-Meier estimator with observed censoring indicators (δ). In Section 2.2 we recall the well-known version of Kaplan-Meier estimator, where δ takes values 1 if the lifetime is observed and 0 otherwise. In Section 2.3 we present the presmooth Kaplan-Meier estimator where $\delta = 1_{\{Z \leq C\}}$ is replaced by a function.

The main aim of this work is to estimate the survival function when δ is not always observed. This is the subject of Chapter 3, where we recall already existing methods proposed by Wang and Ng (2008) (Section 3.1) and define two new estimators (Section 3.2). Additionally, we show that if all censoring indicators are observed, the two estimators of survival function reduce to standard and presmooth Kaplan-Meier estimators in Chapter 2.

In Chapter 4 we perform a simulation study, where we compare our methods with the estimators proposed by Wang and Ng (2008). We present the Kolmogorov-Smirnov distance and the mean integrated square error (MISE).

In Chapter 5, the proposed methodologies are illustrated by analyzing a real data set corresponding to a survival study of colorectal cancer from A Coruña University Hospital.

Keywords: Missing at random; product-limit estimator; random censorship; bandwidth selection

1 INTRODUCCIÓN

El cáncer es un problema de salud que afecta no sólo a quien lo padece, sino que tiene además un gran impacto en su entorno, debido a la carga psicológica y social que conlleva la enfermedad. El abordaje terapéutico del cáncer colorrectal condiciona tanto la duración como la calidad de vida, por todo ello el sistema sanitario debe intentar maximizar los resultados de las intervenciones en esta patología. (Plan Oncológico de Galicia 2002-2005)

La Organización Mundial de la Salud (OMS) establece como objetivos generales para el control del cáncer:

- Reducir su morbimortalidad.
- Aumentar las tasas de curación.
- Mejorar la calidad de vida, tanto de los enfermos que sobreviven como de los que fallecerán.
- Reducir la carga socioeconómica y psicológica que supone esta enfermedad.

El cáncer colorrectal, en España, contabiliza el 12 % de las defunciones por cáncer en hombres y cerca del 15 % en mujeres según los datos del 2006, constituyendo la segunda localización tumoral en importancia en hombres y en mujeres, con una tendencia temporal ascendente (2,2 % anual en hombres y 0,8 % en mujeres). Se estima que en España el número de casos nuevos por año se sitúa en torno a los 25000 en ambos sexos, frente a 13000 defunciones. La supervivencia ajustada relativa global a los 5 años, según los datos del EUROCARE 4 para casos diagnosticados entre los años 2000 y 2002, se sitúa en el 61,5 % Verdecchia et al. (2007). Es algo mayor en las mujeres que en los hombres y en la localización de colon respecto al recto Coleman et al (2008). En nuestro país se ha registrado un aumento del 2 % en la supervivencia global Berrino et al. (2007).

El cáncer colorrectal se diagnostica habitualmente por manifestaciones clínicas, como resultado de un programa de cribaje o como hallazgo casual. Una vez diagnosticado el tumor, el protocolo de seguimiento de los pacientes con cáncer de colon y recto puede, a su vez, modificar el pronóstico. Es por ello que se considera que estudiar el seguimiento es obligado para conocer con precisión la historia natural de la enfermedad y el pronóstico de la misma.

Entre el inicio de la enfermedad y el diagóstico o tratamiento de la misma transcurre un intervalo de tiempo variable que se conoce como demora. La demora diagnóstica puede verse afectada

por las características de la enfermedad, del paciente y del sistema sanitario. Estudios realizados en nuestro país muestran cómo el tiempo transcurrido entre los primeros síntomas y la primera consulta en el cáncer colorrectal tiene una mediana de 49 días Bernal-Perez et al.(2001). Se han identificado como factores modificadores de la demora diferentes factores relacionados con el paciente y con el sistema sanitario.

Los efectos de la demora en el cáncer colorrectal (CCR) son poco conocidos, como lo son también los factores asociados a ésta, tanto los relacionados con el paciente, como aquellos atribuibles al médico de familia o al ámbito hospitalario. De hecho, algunos planes de salud de diferentes Comunidades Autónomas y el Plan Integral de Cáncer promovido por el Ministerio de Sanidad y Consumo en el 2004, han puesto de manifiesto la existencia de problemas en la continuidad asistencial, tanto entre niveles asistenciales implicados en el proceso diagnóstico, terapéutico y de seguimiento del cáncer o centros hospitalarios, como entre profesionales de distintas especialidades que se pueden traducir en demoras innecesarias. Por ello, el Plan propone reducir el tiempo entre la sospecha, la confirmación diagnóstica y el inicio del tratamiento, de tal manera que todo paciente con sospecha clínica fundada de cáncer, independientemente de su lugar de residencia deba poder efectuar una primera prueba de confirmación diagnóstica en los quince días siguientes al establecimiento de la sospecha. La confirmación de la sospecha clínica deberá efectuarse mediante circuitos prioritarios de acceso a las pruebas diagnósticas.

Para la consecución de estos objetivos es necesario, en primer lugar, conocer en nuestro medio el proceso diagnóstico del CCR, las demoras de las distintas etapas y los factores asociados a cada una de ellas. Ello permitirá identificar puntos de mejora en el proceso asistencial así como establecer estrategias adecuadas para mejorar el diagnóstico de estos pacientes.

Es por ello, que la Unidad de Epidemiología Clínica del Complejo Hospitalario Universitario de A Coruña (CHUAC) ha propuesto la realización de un estudio observacional de seguimiento prospectivo con los siguientes objetivos:

Objetivos principales

- Determinar si la duración de los intervalos de tiempo transcurrido entre el primer síntoma y el diagnóstico, y entre el primer síntoma y el tratamiento, modifica la supervivencia de los pacientes con cáncer de colon y de recto.

- Determinar, en pacientes con cáncer colorrectal no metastático tratados con intención curativa, si diferentes estrategias de seguimiento se asocian con una mejor supervivencia.

Objetivos secundarios

- Determinar, en pacientes con cáncer colorrectal:

    1. La supervivencia global.
    2. La supervivencia específica (mortalidad relacionada con el tumor)
    3. La supervivencia libre de progresión
    4. La supervivencia sin recidiva

- Determinar, de las variables recogidas en este estudio, aquellas que modifican el pronóstico de los pacientes con cáncer colorrectal.

Debido a los objetivos que se abarcan en el proyecto de investigación propuesto, se comprueba que se trata de un estudio de supervivencia, es decir, mide el tiempo que transcurre hasta que sucede el evento de interés (muerte). Determinar la supervivencia del cáncer colorrectal es muy importante y se mide como la probabilidad de permanecer vivo durante un determinado tiempo. La supervivencia al año o a los 5 años es a menudo expresada como indicador de la severidad de una enfermedad y como pronóstico. Usualmente, el pronóstico del cáncer se expresa como el porcentaje de pacientes que sobrevive al menos cinco años después del diagnóstico.

En un estudio de supervivencia, el tiempo de seguimiento puede terminarse cuando se produce la muerte o antes de completarse el periodo de estudio. Si se termina antes, se tienen datos censurados (el paciente abandona el estudio, o el seguimiento se pierde y no tenemos información, o el estudio termina antes de aparecer el evento).

El tiempo de supervivencia se define como el intervalo de tiempo desde el acontecimiento o estado inicial hasta el estado final. Por lo que se debe definir el estado inicial de forma que la fecha en que se produjo el evento sea conocida (fecha de diagnóstico, fecha de la intervención, fecha de inicio de la radioterapia o quimioterapia, etc).

En la práctica es frecuente encontrarse situaciones con observaciones incompletas de los períodos que transcurren entre el tiempo inicial y el tiempo final. Lo anterior es debido a la censura o el truncamiento que son mecanismos, que impiden la observación completa de los tiempos de seguimiento. La censura puede ser de dos tipos: censura de tipo I (se observa a los individuos hasta un tiempo determinado) y censura de tipo II (se observa a los individuos hasta que ocurra un número determinado de fallos o eventos de interés).

La censura de tipo I puede ser: censura por la derecha (si en la última observación del individuo, aún no ha ocurrido el evento que se desea observar), censura por la izquierda (si cuando se realiza la primera observación sobre el individuo ya ha ocurrido el evento que se desea observar) y censura por intervalos (si ocurre el evento de interés entre un instante t_i y un tiempo t_j)

El truncamiento es una condición que presentan ciertos sujetos en el estudio y el investigador no puede considerar su existencia. Cuando los datos presentan truncamiento, solamente los individuos a los que les ocurre algún evento particular, antes del evento de interés o la censura, son considerados en el análisis por el investigador.

El truncamiento también puede ser de dos tipos: truncamiento por la izquierda y truncamiento por la derecha, ver, por ejemplo, Andersen et al. (1993).
El truncamiento por la izquierda (entrada tardía al estudio) se presenta, cuando el individuo comienza a observarse posteriormente al verdadero evento inicial. Si X es el momento de ocurrencia del evento que trunca a los sujetos en estudio e Y el tiempo de vida observado, entonces para muestras truncadas por la izquierda, solo los individuos tales que $Y \leq X$ serán considerados.
El truncamiento por la derecha se presenta, cuando sólo se incluyen los individuos que presentan el evento o fallo de interés y no será considerado ningún sujeto que aún no haya presentado el evento.

En un estudio de supervivencia se debe definir el evento de estudio y determinarse su fecha. Este evento está casi siempre asociado a la muerte del paciente pero podrían ser la fecha de alta, la fecha de remisión de la enfermedad, la fecha de recidiva, la fecha de recaída o fallo, etc.

Al estudiar la supervivencia, el evento considerado no es que se produzca o no la muerte, sino la muerte relacionada con la enfermedad. Es por ello, que se produciría un sesgo de información, si se considerará una muerte no relacionada con la enfermedad. El fallecimiento de un paciente por una causa no relacionada con el evento de interés se debe considerar como censurado y su tiempo de seguimiento como incompleto o perdido.

Se observará el tiempo de supervivencia si el paciente ha fallecido por una causa relacionada con la enfermedad y dicho fallecimiento se ha producido antes del fin del estudio; mientras que se observará el tiempo censurado (incompleto) si el paciente está vivo en la fecha de último contacto o fallece por una causa no relacionada con el evento de interés.

A continuación, en la figura 1.1 se muestra el esquema de un estudio de supervivencia.

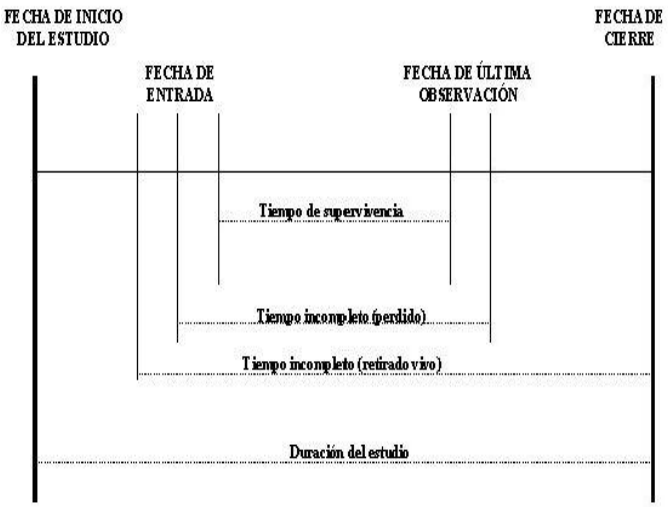

Figura 1.1: Cronograma de un estudio de supervivencia

Actualmente en las bases de datos hospitalarias, resulta complejo conocer si la muerte está o no relacionada con la enfermedad, debido a la falta de definición de muerte relacionada y al no registro obligatorio de la misma, figurando si es exitus y la fecha de muerte.

Surge de este modo el problema de la no respuesta parcial, es decir, ausencia de información sobre el ítem de si la muerte está relacionada con la enfermedad. Ante la no respuesta, se debe conocer el patrón de la pérdida de los datos faltantes para determinar qué método de imputación se debe utilizar. Los patrones de los datos faltantes más frecuentes son el MCAR (datos perdidos completamente al azar), el MAR (perdidos aleatoriamente) y el NMAR (no perdidos aleatoriamente).

En el Registro de Mortalidad de Galicia tienen datos sobre mortalidad por cáncer, aunque este registro tiene una demora de un año y sólo se podrían obtener de los pacientes residentes en la comunidad. El Registro de Mortalidad de Galicia se nutre de datos procedentes de los BED (Boletines Epidemiológicos de Defunción) de Galicia, que son cubiertos por el médico que certifica el éxitus. No se incluyen las defunciones de aquellas personas que fallecen en nuestra comunidad pero tienen su residencia fuera de ella y, por el contrario, recoge las defunciones de personas que fallecen fuera pero tienen su residencia en nuestra comunidad; provocando un sesgo de información al realizar un estudio de seguimiento de los pacientes diagnosticados y tratados en un centro hospitalario. Los datos se recogen en las Direcciones Provinciales del INE (Instituto Nacional de Estadística) que las revisa y codifica y por el IGE (Instituto Galego de Estatística) que incluye datos de población

gallega que fallece fuera de la Comunidad Autónoma cuando son remitidas por el INE o por otras CCAA con convenio antes de cerrar el año. El IGE remite a la Dirección General de Salud Pública de la Consellería de Sanidad los datos de los boletines que esta Dirección General valida y codifica (utilizando la CIE-9. MC. 4^a edición)

El análisis de la mortalidad proporciona datos de gran utilidad, no sólo por el mejor conocimiento de las enfermedades y la evaluación de los programas de control y prevención, sino también para la planificación y contratación de servicios.

Debido a todo el complejo proceso especificado anteriormente y a la falta de información de aquellos casos residentes fuera de la comunidad pero tratados en esta, persiste el problema de los datos faltantes de si la causalidad de la muerte está relacionada con el tumor en algunos casos.

Entonces, al realizar un estudio de supervivencia en el ámbito hospitalario, aparecen problemas para poder determinar si un dato es censura o no, produciéndose un sesgo de información. Es por ello que con este trabajo, se pretende minimizar este sesgo estudiando los estimadores de la función de supervivencia con indicadores de censura en datos perdidos aleatoriamente Wang and Ng (2008) y proponer un estimador presuavizado de Kaplan-Meier para datos faltantes aleatoriamente basado en la presuavización del estimador de Kaplan-Meier propuesto por Cao et al. (2005).

En el capítulo 2, se introduce el estimador de Kaplan-Meier para indicadores de censura observados. En la sección 2.2 se recuerda el estimador de Kaplan-Meier cuando el indicador de censura (δ) toma valores 1 ó 0 y en la sección 2.3 se presenta el estimador de Kaplan-Meier presuavizado donde δ es remplazado por una función. El objetivo principal de este trabajo es la estimación de la función de supervivencia con δ faltante, que se presenta en el capítulo 3. Basándose en los resultados del capítulo 2, se presentan, en la sección 3.1 el modelo de Wang y en la sección 3.2 el modelo propuesto. En el capítulo 4, se muestran los resultados de simulación para los distintos estimadores. Por último, en el capítulo 5, se aplica la metodología propuesta a los datos de un estudio de supervivencia de cáncer colorrectal.

2 MODELO DE CENSURA ALEATORIA

2.1. Modelo de censura aleatoria

La censura indica un tipo de pérdida de información en la que la variable de interés es un tiempo de vida. Es decir, se considera que existe censura cuando no se conoce exactamente el tiempo de vida de una muestra de individuos.
En los estudios de supervivencia es frecuente que exista un porcentaje de datos censurados, es decir, si solamente existe una cota para el tiempo de fallo. La censura puede ser de diversos tipos:

- Censura por la izquierda, el tiempo de fallo tiene una cota superior, es decir, el tiempo de vida asociado a un individuo es menor que cierto valor dado. Por lo que, el momento exacto en el que ocurrió el fallo es desconocido, sabiendo tan sólo que ha ocurrido antes de que el individuo se incluya en el estudio.

- Censura por la derecha (el tiempo de fallo tiene una cota inferior), es decir, el tiempo de vida será superior al valor observado. Por lo que, en el momento en que finaliza el estudio hay sujetos para los que no se conoce el instante exacto de fallo, sino que solamente se conoce que ha sido posterior a un momento dado.

- Censura de tipo intervalo (el tiempo de fallo pertenece a cierto intervalo pero se desconoce el momento exacto), es decir, el tiempo de vida pertenece a cierto intervalo. Por lo que, si se encuentra un fallo para un individuo solamente se sabe que el suceso de interés, el fallo, ocurrió entre dos revisiones consecutivas.

Las causas que originan la censura de una observación pueden ser aleatorias o controladas. Esto hace que se distinga entre tres clases de censura:

- Censura tipo I: El suceso se observa si ocurre antes de un momento fijo predeterminado F. En este caso, F es una constante prefijada por el investigador para todas las unidades muestrales. Si no hay pérdidas accidentales, todas las observaciones censuradas son iguales a la longitud del periodo en estudio.

- Censura tipo II: Este tipo de censura surge cuando se fija el final del estudio en el momento en que un número $r < n$ predeterminado de individuos falla. Los tiempos de vida observados

son los r menores valores de la muestra de forma que C se convierte en la variable aleatoria $C = T_{(r)}$

- Censura tipo III: Se fija la duración y los individuos entran a formar parte de la muestra a lo largo de ese periodo. Para los individuos que fallan antes del final del estudio, se conocen exactamente sus tiempos de vida. Para los que no han experimentado el suceso al final del estudio, la censura de sus tiempos de vida es semejante a la de tipo I. En ocasiones, algunos sujetos experimentan otros sucesos independientes del de interés que provocan su eliminación del estudio. Esta situación se denomina también censura aleatoria. En este tipo de censura, C es una variable aleatoria que se supone independiente de la variable de interés.

En el campo de la Medicina (como en nuestra aplicación a datos reales de cáncer colorrectal), la censura por la derecha aparece de modo natural: no es en general factible extender la duración del estudio hasta que todos los pacientes fallen, máxime teniendo en cuenta que, en general, los pacientes entran en el estudio escalonadamente a medida que van apareciendo. Es por ello, que este trabajo se centrará en la censura por la derecha, donde ocurre que los tiempos de vida de los individuos censurados (aquellos en los que no se había producido el suceso final en el momento de finalizar el estudio) es siempre superior al valor observado de la variable.

En el análisis de supervivencia, los datos pueden ser analizados mediante técnicas paramétricas (distribución Exponencial, distribución de Weibull, distribución Lognormal...) y no paramétricas o semiparamétricas (Kaplan-Meier, Regresión de Cox).

Los métodos estadísticos más utilizados en análisis de supervivencia son los no paramétricos. Así, las curvas de supervivencia, por lo general, se producen usando uno de los siguientes métodos: el análisis actuarial o el método del límite de producto de Kaplan-Meier

- El análisis actuarial divide el tiempo en intervalos y calcula la supervivencia en cada intervalo. Debido a que agrupa los tiempos de supervivencia en intervalos, se obtienen aproximaciones. Este método actuarial implica dos premisas en los datos:
  - Todos los abandonos durante un intervalo dado ocurren aleatoriamente durante dicho intervalo.
  - La supervivencia en un período de tiempo es independiente de la supervivencia en los demás períodos, aunque la supervivencia en un tiempo dado depende de la supervivencia en todos los períodos previos.

  La primera premisa es de escasa importancia cuando se analizan intervalos de tiempo cortos; sin embargo, puede haber un sesgo importante cuando los intervalos son grandes, si hay numerosos abandonos o si los abandonos no ocurren a mitad del intervalo.

- El método Kaplan-Meier calcula la supervivencia cada vez que un paciente muere. Este método, se conoce como el "límite-producto". Caracterizándose porque la proporción acumulada que sobrevive se calcula para el tiempo de supervivencia individual de cada paciente y no se agrupan los tiempos de supervivencia en intervalos.
  La probabilidad de ser censurado debe ser independiente del efecto de interés. Es decir, no puede aplicarse el método de Kaplan-Meier con garantías si se sabe que los que se retiran del estudio antes de que acabe son pacientes peculiares, que probablemente tendrán una supervivencia distinta (mejor o peor) de los que son seguidos hasta el final.

Bajo el modelo de censura aleatoria por la derecha se tiene:

Para cada individuo i-ésimo de un total de n, $i = 1, 2, ..., n$ existe un tiempo de fallo T_i (con función de distribución F continua) y un tiempo de censura C_i (con función de distribución G continua), mutuamente independientes, que no son directamente observables. Los vectores aleatorios (T_i, C_i), $i = 1, 2, ..., n$ se suponen igualmente independientes. Los datos se observan en la forma de n pares (Y_i, δ_i), $i = 1, 2, ..., n$ donde $Y_i = \min(T_i, C_i)$ y $\delta_i = 1_{\{Y_i \leq C_i\}}$. Así, si el i-ésimo individuo falla $\delta_i = 1$ y $Y_i = T_i$ corresponde al tiempo de fallo, mientras que si está censurado, $\delta_i = 0$ y $Y_i = T_i$ es el tiempo de censura. La distribución (común) de Y_i, $i = 1, 2, ..., n$ se denota por H y es inmediato demostrar que

$$1 - H(t) = (1 - F(t))(1 - G(t))$$

Podemos suponer, como en el trabajo se estudian tiempos de vida, que las variables aleatorias son positivas. Así se define:

$$a_F = \text{ínf}\{t > 0 : F(t) > 0\}$$

y

$$b_F = \sup\{t > 0 : F(t) < 1\}$$

para representar los extremos inferior y superior del soporte de la función de distribución F. Extendiendo esta notación para las funciones de distribución G y H, se obtiene a_G, b_G, a_H y b_H, respectivamente. Por lo tanto, estos extremos verifican lo siguiente:

$$a_H = \min\{a_F, a_G\}$$

y

$$b_H = \min\{b_F, b_G\}$$

2.2. El estimador de Kaplan-Meier

Si el tiempo de fallo es una variable aleatoria absolutamente continua con función de distribución F y función de densidad f. Se definen una serie de funciones que matemáticamente son equivalentes a la distribución de T, pero permiten destacar aspectos diferentes de ella.
Se define la función de supervivencia,

$$S_F(t) = \mathbb{P}(T > t).$$

El valor de la función de supervivencia en el tiempo t es igual a la probabilidad de que el individuo experimente el fallo con posterioridad al tiempo t. Es, por tanto, el complemento a 1 de la función de distribución, esto es,

$$S_F(t) = 1 - F(t).$$

La función de supervivencia proporciona una sencilla descripción de la progresión temporal de un grupo de individuos hacia el fallo y es útil para comparar, a este respecto, diferentes grupos entre sí. La función de riesgo se define por

$$\lambda_F(t) = \lim_{\triangle t \to 0^+} \frac{\mathbb{P}(t \leq T < t + \triangle t | T \geq t)}{\triangle t}.$$

Intuitivamente, $\lambda_F(t) \triangle t$ sería una aproximación de la probabilidad de que un individuo que no ha fallado antes del tiempo t lo haga en el siguiente periodo de tiempo de duración $\triangle t$. Por lo tanto,

$$\lambda_F(t) = \frac{f(t)}{S_F(t)} = \frac{-dlogS_F(t)}{dt}.$$

La función de riesgo permite describir los cambios temporales de la probabilidad de experimentar un fallo. Relacionada con la función de riesgo está la función de riesgo acumulativa, que se define como

$$\Lambda_F(t) = \int_0^t \lambda_F(u)\, du$$

El estimador de la función de supervivencia para datos censurados aleatoriamente por la derecha más usado es el propuesto por Kaplan y Meier (1958), conocido como estimador límite-producto. Su expresión se puede derivar de la relación entre la función de supervivencia y la razón de fallo acumulada:

$$S_F(t) = 1 - F(t) = \exp\left[-\Lambda_F(t)\right].$$

Si Λ_F presenta discontinuidades, se puede expresar la relación anterior de la siguiente manera:

$$1 - F(t) = \exp[-\Lambda_F^c(t)] \prod_{a_i \in A / a_i \leq t} \left(1 - \Lambda_F\{a_i\}\right), \tag{2.1}$$

donde $\Lambda_F^c(t)$ denota la parte continua de $\Lambda_F(t)$, A el conjunto de puntos donde $\Lambda_F(t)$ tiene discontinuidades de salto y $\Lambda_F\{a_i\} = \Lambda_F(a_i) - \Lambda_F\left(a_i^-\right)$ es la magnitud del salto de $\Lambda_F(t)$ en a_i.

En este modelo de censura se puede escribir la razón de fallo acumulada $\Lambda_F(.)$ en función de cantidades estimables empíricamente. Sea $H^1(t) = \mathbb{P}(Y \leq t, \delta = 1) = \int_0^t (1 - G(\upsilon^-))F(d\upsilon)$, entonces

$$\Lambda_F(t) = \int_0^t \lambda(\upsilon)\, d\upsilon = \int_0^t \frac{dF(\upsilon)}{1 - F(\upsilon^-)} = \int_0^t \frac{1 - G(\upsilon^-)}{1 - H(\upsilon^-)} dF(\upsilon) = \int_0^t \frac{dH^1(\upsilon)}{1 - H(\upsilon^-)}$$

para todo $t < b_H$, donde $H(\upsilon^-)$ significa $\lim_{x \uparrow \upsilon} H(x)$.

Sustituyendo las funciones $H(t)$ y $H^1(t)$ por sus estimaciones empíricas

$$H_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{Y_i \leq t\}}$$

y

$$H_n^1(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{Y_i \leq t\}} \delta_i$$

Se obtiene el estimador no paramétrico de la razón de fallo acumulada, conocido como el estimador de Nelson-Aalen:

$$\Lambda_n^{NA}(t) = \int_0^t \frac{dH_n^1(\upsilon)}{1 - H_n(\upsilon^-)} = \sum_{i=1}^{n} \frac{1_{\{Y_i \leq t, \delta_i = 1\}}}{n\left(1 - H_n\left(Y_i^-\right)\right)} = \sum_{Y_{(i)} \leq t} \frac{\delta[i]}{n - i + 1},$$

donde $Y_{(i)}$, para $i = 1, ..., n$ son las observaciones ordenadas y $\delta_{[i]}$, son los concomitantes correspondientes a los indicadores de no censura.

Respecto a sus propiedades asintóticas, el estimador de Nelson-Aalen es un estimador uniformemente consistente de $\Lambda_F(t)$ en intervalos compactos $[0, \tau]$ tales que $\tau < b_H$. Bajo el modelo de censura aleatoria, puede probarse la convergencia débil del proceso $\sqrt{n}\left(\Lambda_n^{NA}(t) - \Lambda_F(t)\right)$, $0 < t < \tau$.

Finalmente, el estimador de Nelson-Aalen tiene también una interpretación en términos de estimador de máxima verosimilitud. Se puede modificar la expresión del estimador Nelson-Aalen para permitir más de un fallo en un instante t.

Suponiendo que los sucesos ocurren en D tiempos distintos $t_1 < ... < t_D$ y que en cada instante t_i hay d_i sucesos o fallos, siendo N_i el número de individuos en riesgo, es decir, el número de

individuos vivos en t_i. En el caso de que sólo haya un fallo en cada instante, entonces $d_i = \delta_{[i]}$ y $N_i = n - i + 1$. El cociente d_i/N_i proporciona una estimación de la probabilidad condicionada de que un individuo que sobrevive hasta justo antes del instante t_i, falle en el instante t_i.

Por lo tanto, el estimador de Nelson-Aalen se puede escribir de la forma:

$$\Lambda_n^{NA(2)}(t) = \sum_{Y_{(i)} \leq t} \frac{d_i}{Ni}$$

Ahora estamos en disposición de definir los estimadores de $S(t)$, $1 - F_n^{KM}(t)$ y $1 - F_n^{KM(2)}(t)$. Teniendo en cuenta la ecuación (2.1), se observa que Λ_n^{NA} y $\Lambda_n^{NA(2)}$ tienen saltos en datos $Y_{(i)}$ cuando $\delta_{[i]} = 1$. La parte continua es igual a cero. Entonces

$$1 - F_n^{KM}(t) = \prod_{Y_{(i)} \leq t} \left(1 - \frac{\delta_{[i]}}{n - i + 1}\right).$$

De forma similar, si se permiten más fallos en un instante t, el estimador de la función de supervivencia sería

$$1 - F_n^{KM(2)}(t) = \prod_{Y_{(i)} \leq t} \left(1 - \frac{d_i}{N_i}\right).$$

Utilizando la aproximación $e^{-t} \simeq 1 - t$ para t próximo a 0, se obtiene la relación ente el estimador de Kaplan-Meier y el de Nelson-Aalen:

$$1 - F_n^{KM}(t) = \exp\left[-\Lambda_n^{NA}(t)\right] + O_p\left(n^{-1}\right),$$

se puede comprobar fácilmente que

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

si $\delta_{[i]}$ toma los valores 0 ó 1. Así tenemos, la siguiente expresión del estimador de Kaplan-Meier:

$$1 - F_n^{KM}(t) = \prod_{Y_{(i)} \leq t} \left(\frac{n - i}{n - i + 1}\right)^{\delta_{[i]}}, \tag{2.2}$$

equivalente a

$$1 - F_n^{KM}(t) = \prod_{Y_{(i)} \leq t} \left(1 - \frac{\delta_{[i]}}{n - i + 1}\right).$$

Se observa que esta equivalencia, no se tendría si $\delta_{[i]}$ tomase otros valores.

2.2.1. Propiedades del estimador de Kaplan-Meier

El estimador de Kaplan-Meier se caracteriza por su facilidad de cálculo y el hecho de que sea el estimador no paramétrico de máxima verosimilitud para datos censurados (Johansen (1978), Scholz (1980) o Wang (1987)). Además, se reduce al estimador empírico clásico en el caso de no haber censura. Sin embargo, este estimador presenta problemas cuando la hipótesis de independencia entre los tiempos de fallo T y los tiempos de censura C no se verifica.

El estimador de Kaplan-Meier está bien definido para todos los valores de t menores que el mayor tiempo en estudio observado. No obstante, si el mayor valor observado corresponde a un tiempo de vida no censurado, entonces la curva de supervivencia para valores de t posteriores es 0. En cambio, si la última observación es censurada el valor de $1 - F(t)$ para tiempos posteriores es indeterminado porque no se puede saber cuándo este último individuo hubiera fallado de no haber sido censurado. Para solventar lo anterior se propusieron distintas soluciones, como la de Efron (1967) que propuso estimar $1 - F(t)$ para $t > Y_{(n)}$ por el valor 0, es decir, equivaldría a suponer que el individuo con el mayor tiempo de vida fallase inmediatamente después de haber sido censurado, y da lugar a un estimador negativamente sesgado. La solución propuesta por Gill (1980), fue estimar, para $t > Y_{(n)}$, $1 - F(t)$ por $1 - F_n^{KM}\left(Y_{(n)}\right)$, que obtiene valores no siempre igual a 0, sólo si el concomitante $\delta_{[n]} = 1$, es decir, se corresponde a asumir que este individuo fallaría en $t = \infty$ y conduce a un estimador con sesgo positivo. Aunque las dos propuestas tienen las mismas propiedades asintónticas y convergen a la verdadera función de supervivencia, un estudio con tamaños muestrales finitos de ambos estimadores realizado por Klein (1991) revela un mejor comportamiento de la versión de Gill del estimador de Kaplan-Meier.

Lo y Singh (1986) establecieron una aproximación fuerte uniforme de la diferencia entre estimador de Kaplan-Meier F_n^{KM} y la función de distribución teórica F como una media de variables aleatorias independientes idénticamente distribuidas y acotadas, más un término despreciable de orden conocido. La importancia de estos desarrollos radica en que permiten trabajar con una suma de variables i.i.d., mucho más manejable que el producto por el que viene dado el estimador F_n^{KM}, y obtener propiedades tales como la normalidad asintótica y la convergencia del proceso.

Las principales propiedades del estimador de Kaplan-Meier son:

Propiedad 2.2.1.1 (Representación casi segura, Lo y Singh (1986))

Bajo la hipótesis de que F y G son continuas, se puede escribir para todo $t \leq T < b_H$:

$$F_n^{KM}(t) - F(t) = n^{-1} \sum_{i=1}^{n} \xi(Y_i, \delta_i, t) + r_n(n),$$

donde

$$\xi(Y, \delta, t) = (1 - F(t)) [g(Y \wedge t) + \frac{1}{1 - H(Y)} 1_{\{Y \leq t, \delta = 1\}}],$$

siendo

$$g(t) = \int_0^t \frac{-H^1(dv)}{(1 - H(v))^2}$$

y

$$\sup_{0 \leq t \leq T} |r_n(t)| = O\left(\left(\frac{\log n}{n}\right)^{3/4}\right) c.s.$$

El orden del término r_n es suficiente para probar la mayoría de las propiedades asintóticas del estimador de Kaplan-Meier, pero se queda pequeño cuando se estudian estimadores de la densidad o razón de fallo derivados de F_n^{KM}. Fue posteriormente mejorado hasta $O\left(n^{-1} \log n\right) c.s.$ por Lo Mack y Wang (1989).

Propiedad 2.2.1.2 (Normalidad asintótica puntual y sobre intervalos compactos Breslow y Crowley (1974))

Sean las funciones de distribución F y G continuas. Entonces:

- Para todo $0 < t < b_H$,

$$\sqrt{n}\left(F_n^{KM}(t) - F(t)\right) \xrightarrow{d} N(0, \sigma(t)),$$

donde

$$\sigma^2(t) = (1 - F(t))^2 \int_o^t (1 - H(v))^{-2} dH^1(v).$$

- El proceso estocástico $\mathbb{X}_n(t) = \sqrt{n}\left(F_n^{KM}(t) - F(t)\right)$ converge globalmente en $D[0, T]$ para cada $T < b_H$ a un proceso Gaussiano Z

$$\mathbb{X}_n = \sqrt{n}\left(F_n^{KM} - F\right) \xrightarrow{d} Z,$$

con media 0 y función de covarianzas:

$$Cov(Z(s), Z(t)) = (1 - F(s))(1 - F(t)) \int_0^{s \wedge t} (1 - H(v))^{-2} dH^1(v),$$

siendo $D[0,T] = \{f \in F([0,T],\mathbb{R}) : f$ continua por la derecha y con discontinuidades, a lo sumo de salto, con la topología de Skorohod$\}$, y $F([0,T],\mathbb{R})$ el conjunto de las funciones que van de $[0,T]$ a $\mathbb{R}$

También a principios de los años 80, diversos autores estudiaron la consistencia del estimador de Kaplan-Meier, entre ellos Földes y Rejtö (1980).

Propiedad 2.2.1.3 (Ley del logaritmo iterado, Csörgo y Horváth (1983))

Si $F(T) < 1$, entonces

$$lim \sup_{n\to\infty} \left(\frac{n}{2\log\log n}\right)^{1/2} \sup_{0\le t\le T} |F_n^{KM}(t) - F(t)| \le \frac{1}{1-H(T)}.$$

Como hemos visto en las propiedades, el estimador de Kaplan-Meier fue estudiado en intervalos compactos $[0,\tau]$ con $T < \tau_H$. En esta situación es necesario mencionar que se puede considerar el estimador de Kaplan-Meier en todo el soporte, dadas ciertas condiciones apropiadas que aseguran que el efecto de censura no domina la variable de interés, (ver, por ejemplo, Gill (1983)). Además, las propiedades de las integrales del estimador de Kaplan-Meier, $\int \phi dF_n^{KM}$, donde ϕ es una función fija, fueron investigadas, tomando ϕ's diferentes, por ejemplo $\phi(x) = x$ o $\phi(x) = 1_{\{x\le t\}}$ podemos obtener estimadores de la media de $F(t)$. En nuestro trabajo, vamos a concentrarnos en los estimadores de la distribución F. Para los estimadores de $\int \phi dF$, ver Stute and Wang (1993) y Stute (1995).

El estimador de Kaplan-Meier salta sólo en los datos completos, por lo que, se considera estudiar la suavización de δ y utilizar una función auxiliar, que no es más que la probabilidad de no censura condicionada al valor observado. En la siguiente sección 2.3 se estudiarán los estimadores presuavizados para datos censurados, en particular, para el caso del estimador presuavizado de la función de fallo acumulativa, estimador que se propone como alternativa al estimador clásico de Nelson-Aalen en Cao et al (2005), obteniéndose una mejor eficiencia relativa considerando el cociente de MISE.

2.3. El estimador presuavizado para datos censurados

El nombre de estimación presuavizada viene del hecho de que la suavización se usa únicamente para obtener una versión suavizada de los pesos de Kaplan-Meier, pero el estimador de la función de distribución no es suave. Se usa para dar pesos, también a los datos censurados ($\delta = 0$).

Sean $T_1, ..., T_n$ variables aleatorias positivas, independientes e idénticamente distribuidas (iid), (tiempos de supervivencia o tiempos de fallo) con función de distribución continua y desconocida F. En el modelo de censura aleatoria por la derecha, estos tiempos de supervivencia están censurados por la derecha por variables aleatorias positivas e i.i.d $C_1, ..., C_n$ con función de distribución continua y desconocida G. Para cada $i = 1, ..., n$ se observa (Y_i, δ_i), donde $Y_i = T_i \wedge C_i$ y $\delta_i = 1_{\{T_i \leq C_i\}}$.

Asumiendo que T_i es independiente de C_i, entonces la función de distribución H de Y_i satisface

$$1 - H(t) = (1 - F(t))(1 - G(t)).$$

La variable δ indica si T está censurado ($\delta = 0$) o no censurado ($\delta = 1$). La probabilidad de no censura es:

$$\gamma = \mathbb{P}(\delta = 1) = \mathbb{E}(\delta) = \mathbb{P}(T \leq C) = \int_0^\infty (1 - G(t))\, dF(t) = H^u(+\infty),$$

donde $H^u(t) = \mathbb{P}(Y \leq t, \delta = 1)$ es la funci ón de subdistribución de las observaciones no censuradas.

Sea $H^u(t) = \int_0^t p(s)\, dH(s)$, donde

$$p(t) = \mathbb{P}(\delta = 1 \mid Y = t) = \mathbb{E}(\delta \mid Y = t).$$

La función p es la probabilidad condicional de que la observación sea no censurada dado $Y = t$. La importancia de la función p es evidente en la siguiente relación:

$$\Lambda_F(t) = \int_0^t \frac{1}{1 - H(s^-)} dH^u(s) = \int_0^t p(s) d\Lambda_H(s),$$

con Λ_F y Λ_H las funciones de riesgo acumulado correspondientes a F y H

A partir de la ecuación anterior se obtiene,

$$1 - F(t) = \exp(-\Lambda_F(t)) = \exp\left(-\int_0^t p(s)\, d\Lambda_H(s)\right)$$

y $\lambda_F(t) = p(t)\, \lambda_H(t)$, siendo λ_F y λ_H las funciones de riesgo.

Se tiene que $p(t) = 1$ en caso de no censura. Si δ es independiente de Y, entonces

$$p(t) = \mathbb{E}(\delta = 1 \mid Y = t) = \mathbb{E}(\delta) = \gamma.$$

Se obtiene el modelo de riesgos proporcionales de Koziol-Green

$$1 - F(t) = exp\left(-\Lambda_F\left(t\right)\right) = exp\left(-\gamma\Lambda_H\left(t\right)\right) = \left(1 - H\left(t\right)\right)^{\gamma}$$

o equivalentemente $1 - G(t) = \left(1 - F(t)\right)^{\beta}$ con $\beta = \left(1 - \gamma\right)/\gamma$.

El estimador clásico para $\Lambda_F\left(t\right)$ es el estimador de Nelson-Aalen

$$\Lambda_n^{NA}\left(t\right) = \sum_{Y_{(i)} \leq t} \frac{\delta_{[i]}}{n - i + 1},$$

donde $Y_{(1)} \leq ... \leq Y_{(n)}$ son los Y_i ordenados y los $\delta_{[i]}$ los concomitantes. A partir del estimador de Nelson-Aalen, se puede considerar la versión puramente empírica de

$$\Lambda_F\left(t\right) = \int_0^t \frac{p\left(s\right)}{1 - H\left(s^{-}\right)} dH\left(s\right),$$

reemplazando H por la función de distribución empírica

$$H_n\left(t\right) = n^{-1} \sum_{i=1}^{n} 1_{\{Y_i \leq t\}}$$

y $p(t)$ por

$$\mathbb{E}\left(Z \mid X = t\right) = \frac{\int z f\left(t, z\right) dz}{f_X\left(t\right)}.$$

Para estimar $p(t)$ basta, por tanto, con estimar $f_X\left(t\right)$ y $f\left(t, z\right)$. El estimador de la función $p(t)$ resultante de reemplazar las cantidades desconocidas por sus estimadores de tipo núcleo en la fórmula de la esperanza condicional fue propuesto por Nadaraya y Watson en 1964.

$$p_n\left(t\right) = \frac{\sum_{i=1}^{n} \delta_i K\left(\frac{t - Y_i}{b_n}\right)}{\sum_{i=1}^{n} K\left(\frac{t - Y_i}{b_n}\right)}, \tag{2.3}$$

con $K(.)$ un núcleo y $b \equiv b_n$, $n = 1, 2, ...$, una sucesión de ventanas ($p_n(.)$ es el estimador tipo núcleo de Nadaraya-Watson de $p(.)$ basado en las respuestas binarias δ_i con covariables T_i, $i = 1, ..., n$). El estimador núcleo de Nadaraya-Watson de la función $p_n\left(t\right)$, es una media (local) ponderada de los valores observados de la variable δ.

A partir de lo anterior se obtienie el estimador presuavizado

$$\Lambda_n^p\left(t\right) = \sum_{Y_{(i)} \leq t} \frac{p_n\left(Y_{(i)}\right)}{n - i + 1}.$$

A partir de $\lambda_F\left(t\right) = p\left(t\right)\lambda_H\left(t\right)$ y teniendo en cuenta que $\lambda_F\left(t\right)$ puede ser estimado por medio

del estimador presuavizado de la función de riesgo, se tiene

$$\lambda_n^P(t) = p_n(t)\,\lambda_n(t),$$

donde $\lambda_n(t)$ es un estimador de $\lambda_H(t)$ (por ejemplo, el estimador núcleo de Watson y Leadbetter).

Este estimador es producto de dos estimadores basados en las observaciones i.i.d. De hecho, $p_n(t)$ está basado en (Y_i, δ_i) $i = 1, ..., n$ y $\lambda_n(t)$ está basado en Y_i, $i = 1, ..., n$.

Cualquier posible estimador no paramétrico para $p(t)$ y para $\lambda_H(t)$ puede usarse como estimador producto en la ecuación

$$\lambda_n^P(t) = p_n(t)\,\lambda_n(t).$$

Usando en la ecuación

$$1 - F(t) = \exp(-\Lambda_F(t)) = \exp\left(-\int_0^t p(s)d\Lambda_H(s)\right),$$

las expresiones $\Lambda_n^{NA}(t) = \sum_{Y_{(i)} \le t} \frac{\delta_{[i]}}{n-i+1}$ y $\Lambda_n^p(t) = \sum_{Y_{(i)} \le t} \frac{p_n\left(Y_{(i)}\right)}{n-i+1}$ y la aproximación $e^{-x} \simeq 1 - x$ para x próximo a 0, entonces se obtienen fácilmente los siguientes estimadores para $1 - F(t) = \exp(-\Lambda_F(t))$, la función de supervivencia en t:

$$1 - F_n^{KM}(t) = \prod_{Y_{(i)} \le t} \left(1 - \frac{\delta_{[i]}}{n-i+1}\right)$$

y

$$1 - F_n^P(t) = \prod_{Y_{(i)} \le t} \left(1 - \frac{p_n\left(Y_{(i)}\right)}{n-i+1}\right).$$

Estos estimadores son el clásico estimador de Kaplan-Meier y el estimador presuavizado propuesto por Cao et al (2005). El estimador presuavizado de la función de distribución en presencia de censura aleatoria por la derecha se obtiene directamente del clásico estimador de Kaplan-Meier, sin más que sustituir $\delta_{[i]}$ por un estimador suave $p_n\left(Y_{(i)}\right)$, como, por ejemplo, p el estimador tipo núcleo de Nadaraya-Watson basados en las respuestas binarias δ_i con covariables Y_i, $i = 1, ..., n$ (estimador que se puede interpretar como el estimador de máxima verosimilitud local de p)

Se tiene que para todo t tal que $H(t) < 1$,

$$1 - F_n^{KM}(t) = \exp\left(-\Lambda_n^{NA}(t)\right) + O_p\left(n^{-1}\right)$$

y también que

$$1 - F_n^P(t) = \exp\left(-\Lambda_n^P(t)\right) + O_p\left(n^{-1}\right).$$

Comprobándose que el error cuadrático medio asintótico (AMSE) de la parte dominante del estimador presuavizado de Nelson-Aalen y del estimador presuavizado de Kaplan-Meier son más pequeños que los AMSE de las correspondientes expresiones de los estimadores de Nelson-Aalen y de Kaplan-Meier.

3 ESTIMADORES DE LA FUNCIÓN DE SUPERVIVENCIA CON INDICADORES DE CENSURA FALTANTES

El análisis estadístico de supervivencia está basado frecuentemente en observaciones censuradas. Bajo la censura aleatoria, Kaplan y Meier (1958) propusieron el estimador de la función de supervivencia, conocido como "límite-producto".

El estimador de Kaplan-Meier utiliza toda la información disponible, casos censurados y no censurados, para realizar la estimación de la función de supervivencia. El estimador en cualquier instante de tiempo se obtiene de la multiplicación de probabilidades condicionales de la supervivencia estimadas.

El estimador de Kaplan-Meier requiere que la variable indicadora de la censura sea siempre observable. Sin embargo, como ya se ha indicado al principio del trabajo, la variable indicadora de la censura, δ, puede ser un dato perdido debido a múltiples razones. Como por ejemplo, en estudios epidemiológicos el certificado de la muerte puede estar perdido debido a la emigración o a la falta de registro en las bases hospitalarias. Así, sabemos que el paciente ha fallecido, pero no si por la causa de enfermedad de interés ($\delta = 1$) o por otras causas no relacionadas ($\delta = 0$).

Es muy importante conocer el tipo de mecanismo que produce los daltos faltantes para poder imputarlos. Los mecanismos que generan ausencia de datos son: MCAR (Missing Completely At Random), MAR (Missing At Random) y MNAR (Missing Not At Random). Ver el capítulo 1 de Litte and Rubin (1987).

El modelo MCAR está basado en que los datos están perdidos completamente al azar, es decir, la probabilidad de que una respuesta a una variable sea dato faltante es independiente tanto del valor de esta variable como del valor de otras variables del conjunto de datos. La ausencia de la información no está originada por ninguna variable presente en la matriz de datos. Se define ξ_i como

una variable indicadora de si δ no está perdido o sí, tomando los valores $\xi_i = 1$ si δ_i no está perdido y $\xi_i = 0$ si δ_i está perdido. Además se define otras variables X que tienen información adicional sobre el paciente, por ejemplo, $Mhosp$ es una variable indicadora de si el exitus se produce en el hospital, que toma los valores $Mhosp = 1$ si el fallecimiento ocurre en el hospital y $Mhosp = 0$ si el fallecimiento sucede fuera del hospital. Entonces en caso de MCAR se tiene,

$$\mathbb{P}\left(\xi = 0 \mid Y, \delta = 1\right) = \mathbb{P}\left(\xi = 0\right).$$

El modelo MAR está basado en que los datos son perdidos aleatoriamente, aunque esta pérdida está asociada a variables presentes, es decir, la probabilidad de que una respuesta sea dato faltante es dependiente de los valores de otras variables del conjunto de datos. La ausencia de datos está asociada a variables presentes en la matriz de datos.

Sea $\xi_i = 0$ si δ_i está perdido y $Mhosp = 1$ si el fallecimiento es en hospital. Entonces en el modelo de MAR se tiene,

$$\mathbb{P}\left(\xi = 0 \mid Y, \delta = 1\right) = \mathbb{P}\left(\xi = 0 \mid Y\right).$$

Los dos mecanismos de datos faltantes mencionados se denominan también ignorables, por cuanto producen efectos que se pueden ignorar si se controla adecuadamente por las variables que determinan la no respuesta.

El esquema de pérdida de datos MNAR está basado en que los datos no son perdidos aleatoriamente, es decir, está asociado a variables conocidas. La probabilidad de que una respuesta sea dato faltante depende de los valores de la respuesta. Este tipo de dato faltante también se denomina no ignorable.

Sea $\xi_i = 0$ si δ_i está perdido y $Mhosp = 1$ si el fallecimiento es en hospital. Entonces en caso de MNAR se tiene,

$$\mathbb{P}\left(\xi = 0 \mid Y, \delta = 1\right) \neq \mathbb{P}\left(\xi = 0 \mid Y\right).$$

Para tratar los datos faltantes, pueden utilizarse los siguientes análisis:

- Análisis con datos completos (Listwise). Consiste en realizar el análisis estadístico únicamente con las observaciones que disponen de información completa para todas las variables. Para ello se eliminan los registros que presentan algún dato faltante. De esta forma, se considera que la submuestra de datos excluidos tiene las mismas características que los datos completos, y que la falta de información se generó de manera aleatoria, lo cual en la mayoría de las situaciones prácticas no se cumple.

- Análisis con datos disponibles (Pairwise deletion). Utilizar en el análisis de cada variable todos los datos de que se disponga. Con este método se obtienen buenos resultados únicamente en el caso de estar bajo un proceso de no respuesta de tipo MCAR.

- Ponderación. Incrementa los pesos de los datos que se tiene la información, de modo que representen a los que no se les conoce el dato. El objetivo de esta técnica es mejorar la precisión de las estimaciones y reducir el sesgo que introducen los que no tienen información, ya que el resultado final presupone que todos los sujetos no tienen datos perdidos.

En el análisis de la supervivencia, si la variable aleatoria indicadora de censura es un dato perdido, un método es ignorar estos datos perdidos y utilizar el estimador de Kaplan-Meier (análisis con datos completos, Listwise). Sin embargo, en estos casos, el estimador es altamente ineficiente si existe un grado significativo de datos perdidos. Cuando los indicadores de censura son datos perdidos bajo censura aleatoria, los datos observados son $(Y_i, \xi_i, \delta_i\xi_i)$, siendo los Y_i siempre observables y $\xi_i = 0$ si δ_i está perdido, en este caso $\delta_i\xi_i = 0$ y $\xi_i = 1$ si δ_i no está perdido, en este caso $\delta_i\xi_i = \delta_i$. Suponiendo que δ está perdido aleatoriamente (MAR), lo que implica que δ y ξ son independientemente condicionadas a Y. Entonces,

$$\mathbb{P}(\xi = 0 \mid Y, \delta) = \mathbb{P}(\xi = 0 \mid Y).$$

3.1. Modelo de Wang

Se han desarrollado métodos para estimar una función de supervivencia con indicadores de censura perdida aleatoriamente. Los métodos resultantes permiten el uso de la imputación y la ponderación.

El análisis estadístico de los datos de tiempo de vida o tiempo de fallo están frecuentemente basados en observaciones censuradas. Bajo censura aleatoria, Kaplan y Meier (1958) sugirieron un estimador límite-producto de la función de supervivencia. Para describir el estimador límite-producto, se denota con T la variable aleatoria que representa el tiempo de vida con función de distribución F, y por C la variable aleatoria que describe la censura por la derecha con función de distribución G. Asumiendo que T es independiente de C. Bajo censura aleatoria, se observa $(Y_i, \xi_i, \delta_i\xi_i)$, siendo los Y_i siempre observables y $\xi_i = 0$ si δ_i está perdido, en este caso $\delta_i\xi_i = 0$ y $\xi_i = 1$ si δ_i no está perdido, en este caso $\delta_i\xi_i = \delta_i$, donde $Y = T \wedge C$ y $\delta = 1_{\{T \leq C\}}$, con $1_{\{.\}}$ función indicadora. Se supondrá que los datos consisten en observaciones independientes e idénticamente distribuidas (Y_i, δ_i) para $i = 1, 2, ..., n$. Recordamos que en esta situación, Kaplan y Meier (1958) definieron el estimador límite-producto de la función de supervivencia $S(t) = 1 - F(t)$ como

$$\hat{S}_{KM}(t) = \prod_{i:Y_{(i)} \leq t} \left(\frac{n-i}{n-i+1} \right)^{\delta_{[i]}}, \tag{3.1}$$

donde $Y_{(i)}$ son los datos ordenados y $\delta_{[i]}$ sus concomitantes.

El estimador límite-producto requiere que el indicador de censura sea siempre observado. Sin embargo, el indicador de censura δ (o causa del fallo de información) puede estar perdido por diversas razones. Cuando los indicadores de censura están perdidos bajo censura aleatoria, los datos observados son $(Y_i, \xi_i, \delta_i \xi_i)$, donde los Y_i son siempre observados, $\xi_i = 0$ si δ_i está perdido, en este caso $\delta_i \xi_i = 0$ y $\xi_i = 1$ si δ_i no está perdido, en este caso $\delta_i \xi_i = \delta_i$.

Asumiendo que δ está perdido aleatoriamente (MAR) la suposición de MAR implica que ξ y δ son independientes condicionadas a Y. Por lo tanto, $\mathbb{P}(\xi = 1 \mid Y, \delta) = \mathbb{P}(\xi = 1 \mid Y)$. MAR es una suposición común para el análisis estadístico con datos perdidos y es razonable en muchas situaciones prácticas, ver el capítulo 1 de Little and Rubin (1987).

Es sabido que la imputación y los enfoques de ponderación son aplicados habitualmente a problemas de regresión con variables respuesta perdidas o covariables perdidas. Teniendo en cuenta que, según Dikta (1998), la función de supervivencia $S(t)$ se puede representar como un funcional de $m(y) = \mathbb{E}[\delta \mid Y = y]$, una función de regresión del indicador δ sobre Y. Entonces, se puede utilizar la imputación y ponderación junto con el estimador límite-producto, para obtener un estimador de la función de supervivencia, $S(t)$

A continuación, se construyen estimadores asintóticamente eficientes a través de los métodos de imputación y ponderación. El resultado de la aproximación utilizando la ponderación fue introducido por Robins y Rotnitzky (1992).

Sea H la función de distribución de Y y $H^1(t) = \mathbb{P}(Y \leq t, \delta = 1)$. La función de riesgo acumulado $\Lambda(t)$ correspondiente a F está dada por

$$\Lambda(t) = \int_0^t \frac{1}{1 - F(\upsilon^-)} dF(\upsilon) = \int_0^t \frac{1}{1 - H(\upsilon^-)} dH^1(\upsilon).$$

Si todos los δ_i fuesen observables, entonces $H^1(\upsilon)$ se podría estimar con $H_n^1(\upsilon) = \frac{1}{n} \sum_{i=1}^n 1_{\{Y_i \leq t\}} \delta_i$. Puesto que δ_i sólo es observable cuando $\xi = 1$, se tiene que encontrar otro estimador de $H^1(\upsilon)$, basado en la función $m(y) = \mathbb{P}(\delta = 1 \mid Y = y)$.

Según Dikta (1998), se tiene $H^1(t) = \mathbb{P}(\delta = 1, Y \leq t) = \int_0^t m(\upsilon) dH(\upsilon)$, donde $m(y) = \mathbb{P}(\delta = 1 \mid Y = y) = \mathbb{E}[\delta \mid Y = y]$. De una forma similar, a como se realizó en la sección 2.3 del

capítulo 2, se utiliza la suavización. Aunque en dicha sección se utilizaba para dar pesos también los datos con $\delta_i = 0$, pero aquí utilizamos métodos similares para estimar δ cuando está perdida. Entonces, se tiene que

$$\Lambda(t) = \int_0^t \frac{m(\upsilon)}{1 - H(\upsilon^-)} dH(\upsilon).$$

Sea $H_n(t) = n^{-1} \sum_{i=1}^n 1\{Y_i \leq t\}$, $H_n(t-) = \lim_{y \uparrow t} H_n(y)$. Si se puede definir un estimador de $m(y)$, $m_n(y)$, a partir de los datos observados $(Y_i, \xi_i, \delta_i \xi_i)$, $i = 1, 2, ..., n$. Entonces $\Lambda(t)$, se puede estimar como

$$\Lambda_n(t) = \int_0^t \frac{m_n(\upsilon)}{1 - H_n(\upsilon^-)} dH_n(\upsilon) = \sum_{i:Y_i \leq t} \frac{m_n(Y_i)}{(1 - H_n(Y_i^-))n}$$

A continuación, sean $Y_{(i)}$ los datos ordenados. Entonces, podemos escribir $\Lambda_n(t)$ como

$$\Lambda_n(t) = \sum_{i:Y_{(i)} \leq t} \frac{m_n(Y_{(i)})}{n - i + 1}.$$

Así, $S(t) = \exp\{-\Lambda(t)\}$ puede estimarse por $\exp(-\Lambda_n(t))$. Utilizando la aproximación $\exp(-x) \simeq 1 - x$, se tiene que

$$\exp(-\Lambda_n(t)) = \prod_{i:Y_{(i)} \leq t} \left(\exp\left\{ -\frac{1}{n - i + 1} \right\} \right)^{m_n(Y_{(i)})} \simeq \prod_{i:Y_{(i)} \leq t} \left(\frac{n - i}{n - i + 1} \right)^{m_n(Y_{(i)})}.$$

Por lo que se puede considerar el estimador límite-producto,

$$S_n(t) = \prod_{i:Y_{(i)} \leq t} \left(\frac{n - i}{n - i + 1} \right)^{m_n(Y_{(i)})}.$$

A continuación, se presentan cuatro estimadores diferentes de la función $m(y)$, propuestos por Wang and Ng (2008).

Primero, se utiliza el enfoque de ponderación para estimar m. Sea $\pi(y) = \mathbb{P}(\xi = 1 \mid Y = y)$ se estima mediante $\pi_n(y)$, que es el estimador núcleo de regresión de Nadaraya-Watson que tiene la siguiente forma:

$$\pi_n(y) = \frac{\sum_{i=1}^n \xi_i K\left(\frac{y - Y_i}{b_n}\right)}{\sum_{i=1}^n K\left(\frac{y - Y_i}{b_n}\right)},$$

donde K es la función núcleo y b_n la sucesión de ventanas de suavización.

Se tiene que,

$$\hat{m}_n(y) = \frac{\sum_{i=1}^n \left(\frac{\xi_i \delta_i}{\pi_n(Y_i)}\right) K\left(\frac{y - Y_i}{h_n}\right)}{\sum_{i=1}^n \left(\frac{\xi_i}{\pi_n(Y_i)}\right) K\left(\frac{y - Y_i}{h_n}\right)},$$

donde K es la función núcleo y h_n es la sucesión de ventanas.

El primer estimador $\hat{S}_{n,W}(t)$, se obtiene sustituyendo m_n por $\hat{m}_n$ en $S_n(t)$. Se tiene de esta forma el primer estimador ponderado, $\hat{S}_{n,W}(t)$. Es decir, es el estimador de Kaplan-Meier, reemplazando $\delta_{[i]}$ en $\hat{S}_{KM}(t) = \prod_{i:Y_{(i)}\leq t}\left(\frac{n-i}{n-i+1}\right)^{\delta_{[i]}}$ por el estimador $\hat{m}_n\left(Y_{(i)}\right)$. Por lo tanto,

$$\hat{S}_{n,W}\left(t\right) = \prod_{i:Y_{(i)}\leq t}\left(\frac{n-i}{n-i+1}\right)^{\hat{m}_n\left(Y_{(i)}\right)}.$$

Observándose que si $\xi_{[i]} = 1$ $i = 1, 2, ..., n$, es decir, si se observan todos los $\delta_{[i]}$, se tiene que $\hat{m}_n(Y_{(i)}) = p_n(Y_{(i)})$, $i = 1, 2, ..., n$. Por lo que el estimador anterior, $\hat{S}_{n,W}\left(t\right)$, se reduce al estimador presuavizado de Kaplan-Meier (2.3) propuesto por Cao et al (2005), definido en el capítulo 2 en la sección 2.3 de este trabajo.

Se define el segundo estimador de la función $m(y)$, modificando el estimador anterior, $\hat{S}_{n,W}\left(t\right)$, al reemplazar los δ_i perdidos por $\hat{m}_n\left(Y_i\right)$. Esto permite obtener un estimador imputado, dado por

$$\hat{S}_{n,I}\left(t\right) = \prod_{i:Y_{(i)}\leq t}\left(\frac{n-i}{n-i+1}\right)^{\xi_{[i]}\delta_{[i]}+\left(1-\xi_{[i]}\right)\hat{m}_n\left(Y_{(i)}\right)},$$

donde $Y_{(i)}$ son los datos ordenados y $\delta_{[i]}$, $\xi_{[i]}$ los concomitantes.

Este estimador puede ser motivado por el hecho de que $E[\xi\delta + (1-\xi)m(Y)] = E[\delta]$ bajo MAR. Observándose que si $\xi_{[i]} = 1$, $i = 1, 2, ..., n$, es decir, si se observan todos los $\delta_{[i]}$, se tiene que $\xi_{[i]}\delta_{[i]} + \left(1-\xi_{[i]}\right)\hat{m}_n\left(Y_{(i)}\right) = \delta_{[i]}$ $i = 1, 2, ..., n$. Por lo que el estimador propuesto, $\hat{S}_{n,I}\left(t\right)$, se reduce al estimador de Kaplan-Meier (2.2), definido en el capítulo 2 en la sección 2.2 del presente trabajo.

La tercera opción es reemplazar, $\hat{m}_n$ en $\hat{S}_{n,I}\left(t\right)$ con

$$\tilde{m}_n\left(y\right) = \frac{\sum_{i=1}^{n}\xi_i\delta_i K\left(\frac{y-Y_i}{h_n}\right)}{\sum_{i=1}^{n}\xi_i K\left(\frac{y-Y_i}{h_n}\right)}.$$

Se obtendría otro estimador imputado $\tilde{S}_{N,I}\left(t\right)$, con la siguiente expresión

$$\tilde{S}_{N,I}\left(t\right) = \prod_{i:Y_{(i)}\leq t}\left(\frac{n-i}{n-i+1}\right)^{\xi_{[i]}\delta_{[i]}+\left(1-\xi_{[i]}\right)\tilde{m}_n\left(Y_{(i)}\right)},$$

observándose que si $\xi_{[i]} = 1$ $i = 1, 2, ..., n$, es decir, si se observan todos los $\delta_{[i]}$, se tiene que $\xi_{[i]}\delta_{[i]} + \left(1-\xi_{[i]}\right)\tilde{m}_n\left(Y_{(i)}\right) = \delta_{[i]}$ $i = 1, 2, ..., n$, por lo que el estimador propuesto, $\tilde{S}_{N,I}\left(t\right)$, se

reduce al estimador de Kaplan-Meier (2.2), definido en el capítulo 2 en la sección 2.2 del presente trabajo.

Sea $\pi(y) = \mathbb{P}(\xi = 1 \mid Y = y)$. Bajo MAR se tiene que, $\mathbb{E}\left[\frac{\xi\delta}{\pi(Y)+\left(\frac{1-\xi}{\pi(Y)}\right)m(Y)}\right] = \mathbb{E}[\delta]$ y $\pi_n(y)$ es el estimador núcleo de la función de regresión de $\pi(y)$. A partir de esto, se define otro estimador ponderado como,

$$\tilde{S}_{n,W}(t) = \prod_{i:Y_{(i)}\leq t}\left(\frac{n-i}{n-i+1}\right)^{\frac{\xi_{[i]}\delta_{[i]}}{\pi_n\left(Y_{(i)}\right)}+\left(1-\frac{\xi_{[i]}}{\pi_n\left(Y_{(i)}\right)}\right)\tilde{m}_n\left(Y_{(i)}\right)}.$$

Obsérvese que si $\xi_{[i]} = 1$ $i = 1, 2, ..., n$, es decir, si se observan todos los $\delta_{[i]}$, se tiene que $\frac{\xi_{[i]}\delta_{[i]}}{\pi_n\left(Y_{(i)}\right)}+\left(1-\frac{\xi_{[i]}}{\pi_n\left(Y_{(i)}\right)}\right)\tilde{m}_n\left(Y_{(i)}\right) = \delta_{[i]}$ $i = 1, 2, ..., n$, por lo que el estimador propuesto, $\tilde{S}_{n,W}(t)$, se reduce al estimador de Kaplan-Meier (2.2), definido en el capítulo 2 en la sección 2.2 de este trabajo.

Este estimador $\tilde{S}_{n,W}(t)$ está mal definido para el valor $t \geq Y_{(n)}$, si $\xi_{[n]} = 1$ y $\delta_{[n]} = 0$. De modo que, $\xi_{[n]}\delta_{[n]} = 0$ y teniendo en cuenta que si la tasa de perdidos es positiva, $\pi_n(Y_{(i)}) < 1$, se tiene que $\left(1-\frac{\xi_{[i]}}{\pi_n\left(Y_{(i)}\right)}\right) < 0$. Así que, si $\xi_{[n]} = 1$ y $\delta_{[n]} = 0$, se tiene

$$\frac{\xi_{[n]}\delta_{[n]}}{\pi_n\left(Y_{(n)}\right)}+\left(1-\frac{\xi_{[n]}}{\pi_n\left(Y_{(n)}\right)}\right)\tilde{m}_n\left(Y_{(n)}\right) < 0.$$

Además, cuando $i = n$ $\left(\frac{n-i}{n-i+1}\right)$, toma el valor $\left(\frac{n-n}{n-n+1}\right) = 0$.

Si $t \geq Y_{(n)}$, entonces el último factor en la definición del estimador es

$$\left(\frac{n-n}{n-n+1}\right)^{\frac{\xi_{[n]}\delta_{[n]}}{\pi_n\left(Y_{(n)}\right)}+\left(1-\frac{\xi_{[n]}}{\pi_n\left(Y_{(n)}\right)}\right)\tilde{m}_n\left(Y_{(n)}\right)} = 0^k,$$

siendo $k < 0$. Por lo tanto

$$\tilde{S}_{n,W}(t) = \prod_{i:Y_{(i)}\leq t}\left(\frac{n-i}{n-i+1}\right)^{\frac{\xi_{[i]}\delta_{[i]}}{\pi_n\left(Y_{(i)}\right)}+\left(1-\frac{\xi_{[i]}}{\pi_n\left(Y_{(i)}\right)}\right)\tilde{m}_n\left(Y_{(i)}\right)} = \infty.$$

Lo anterior se corrrespondería a asumir que el individuo con tiempo de vida observado más grande fallaría en $t = \infty$ y esto conduciría a obtener un estimador con sesgo positivo.

De los cuatro estimadores propuestos por Wang et al (2008), los dos estimadores imputados y el segundo estimador ponderado se reducen al estimador de Kaplan-Meier y $\hat{S}_{n,W}(t)$ se reduce al estimador de Kaplan-Meier suavizado cuando los indicadores de censura son observados com-

pletamente. En este caso, sin embargo, la distribución asintótica de $\hat{S}_{n,W}(t)$, se reduce a la del estimador de Kaplan-Meier.

Los estimadores propuestos extenderse a más situaciones reales. Frecuentemente, en análisis de supervivencia y estudios biomédicos, la información de covariables es recogida cuando algunos indicadores de censura son perdidos. Una posible extensión sería estimar la función de riesgo condicional y la función de supervivencia condicionada para incorporar la información relativa a covariables con un método de suavización núcleo cuando los indicadores de censura son perdidos aleatoriamente.

3.2. Modelo propuesto

A continuación se propone un estimador presuavizado de Kaplan-Meier con el indicador de censura desconocida aleatoriamente.

Por lo que, bajo la censura aleatoria, definimos el estimador límite-producto de la función de supervivencia (Kaplan-Meier). Se denota con T la variable aleatoria que representa el tiempo de vida con función de distribución F, y por C la variable aleatoria que describe la censura por la derecha con función de distribución G. Asumiendo que T es independiente de C, bajo censura aleatoria, se observa $(Y_i, \xi_i, \delta_i\xi_i)$, siendo los Y_i siempre observables y $\xi_i = 0$ si δ_i está perdido, en este caso $\delta_i\xi_i = 0$ y $\xi_i = 1$ si δ_i no está perdido, en este caso $\delta_i\xi_i = \delta_i$, donde $Y = T \wedge C$ y $\delta = 1_{\{T \leq C\}}$, con $1_{\{.\}}$ la función indicadora. Se supone que los datos consisten en observaciones independientes e idénticamente distribuidas (Y_i, δ_i) para $i = 1, 2, ..., n$. Suponiendo que δ está perdido aleatoriamente (MAR), lo que implica que δ y ξ son independientemente condicionadas dado Y. Entonces, $\mathbb{P}(\xi = 1 \mid Y, \delta) = \mathbb{P}(\xi = 1 \mid Y)$.

Se necesita estimar $\mathbb{P}(\delta = 1 \mid Y = y)$. Se tiene que,

$$\mathbb{P}(\delta = 1 \mid Y = y) = \mathbb{P}(\delta = 1 \mid Y = y, \xi = 1)\mathbb{P}(\xi = 1 \mid Y = y) + \mathbb{P}(\delta = 0 \mid Y = y, \xi = 0)\mathbb{P}(\xi = 0 \mid Y = y). \tag{3.2}$$

Las funciones $\mathbb{P}(\delta = 1 \mid Y = y, \xi = 1)$, $\mathbb{P}(\xi = 1 \mid Y = y)$ y $\mathbb{P}(\xi = 0 \mid Y = y)$ son fácilmente estimables, por lo que necesitamos trabajar con $\mathbb{P}(\delta = 0 \mid Y = y, \xi = 0)$

$$\mathbb{P}(\delta = 0 \mid Y = y, \xi = 0) = \frac{\mathbb{P}(\delta = 0, Y = y, \xi = 0)}{\mathbb{P}(\xi = 0, Y = y)} = \frac{\mathbb{P}(\xi = 0 \mid Y = y, \delta = 0)\mathbb{P}(Y = y, \delta = 0)}{\mathbb{P}(\xi = 0 \mid Y = y)\mathbb{P}(Y = y)} =$$

$$= \frac{\mathbb{P}(Y = y, \delta = 0)}{\mathbb{P}(Y = y)} = \mathbb{P}(\delta = 0 \mid Y = y) = 1 - \mathbb{P}(\delta = 1 \mid Y = y). \tag{3.3}$$

Sustituyendo $\mathbb{P}(\delta = 0 \mid Y = y, \xi = 0)$ en (3.1) con (3.2), obtenemos:

$$\mathbb{P}(\delta = 1 \mid Y = y) = \mathbb{P}(\delta = 1 \mid Y = y, \xi = 1)\mathbb{P}(\xi = 1 \mid Y = y) + (1 - \mathbb{P}(\delta = 1 \mid Y = y))\mathbb{P}(\xi = 0 \mid Y = y).$$

Entonces,

$$\mathbb{P}(\delta = 1 \mid Y = y) + \mathbb{P}(\delta = 1 \mid Y = y)\mathbb{P}(\xi = 0 \mid Y = y) - \mathbb{P}(\xi = 0 \mid Y = y) =$$

$$= \mathbb{P}(\delta = 1 \mid Y = y, \xi = 1)\mathbb{P}(\xi = 1 \mid Y = y),$$

que es equivalente a

$$\mathbb{P}(\delta = 1 \mid Y = y)[1 + \mathbb{P}(\xi = 0 \mid Y = y)] = \mathbb{P}(\delta = 1 \mid Y = y, \xi = 1)\mathbb{P}(\xi = 1 \mid Y = y) + \mathbb{P}(\xi = 0 \mid Y = y).$$

Entonces,

$$\mathbb{P}(\delta = 1 \mid Y = y) = \frac{\mathbb{P}(\delta = 1 \mid Y = y, \xi = 1)\mathbb{P}(\xi = 1 \mid Y = y) + \mathbb{P}(\xi = 0 \mid Y = y)}{1 + \mathbb{P}(\xi = 0 \mid Y = y)} =$$

$$= \frac{\mathbb{P}(\delta = 1 \mid Y = y, \xi = 1)\mathbb{P}(\xi = 1 \mid Y = y) + \mathbb{P}(\xi = 0 \mid Y = y)}{1 + (1 - \mathbb{P}(\xi = 1 \mid Y = y))} =$$

$$= \frac{\mathbb{P}(\delta = 1 \mid Y = y, \xi = 1)\mathbb{P}(\xi = 1 \mid Y = y) + \mathbb{P}(\xi = 0 \mid Y = y)}{2 - \mathbb{P}(\xi = 1 \mid Y = y)}.$$

Sea $\delta(y) := \mathbb{P}(\delta = 1 \mid Y = y)$, $\pi(y) = \mathbb{P}(\xi = 1 \mid Y = y)$ y $\gamma(y) = \mathbb{P}(\delta = 1 \mid Y = y, \xi = 1)$. Entonces,

$$\delta(y) = \frac{\gamma(y)\pi(y) + (1 - \pi(y))}{2 - \pi(y)}$$

Utilizando el método plug-in, que consiste en representar una función desconocida a partir de funciones estimables y después reemplazar las cantidades poblacionales por sus estimaciones (γ por $\hat{\gamma}$ y π por $\hat{\pi}$), podemos definir el estimador $\hat{\delta}(y)$ de $\delta(y)$.

Sea K el núcleo y h_n, b_n dos ventanas, se define el estimador $\pi_n(y)$ de $\pi(y)$ como,

$$\pi_n(y) = \frac{\sum_{i=1}^{n} \xi_i K\left(\frac{y-Y_i}{b_n}\right)}{\sum_{i=1}^{n} K\left(\frac{y-Y_i}{b_n}\right)}$$

y el estimador $\hat{\gamma}(y)$ de $\gamma(y)$ como,

$$\hat{\gamma}(y) = \frac{\sum_{i=1}^{n} \delta_i \xi_i K\left(\frac{y-Y_i}{h_n}\right)}{\sum_{i=1}^{n} \xi_i K\left(\frac{y-Y_i}{h_n}\right)}.$$

Entonces, el estimador propuesto es

$$\hat{\delta}(y) = \frac{\hat{\gamma}(y)\hat{\pi}(y) + (1 - \hat{\pi}(y))}{2 - \hat{\pi}(y)}.$$

Así el estimador de Kaplan-Meier con el indicador de censura perdida tiene la siguiente forma,

$$1 - \hat{F}_n^{KM}(t) = \prod_{Y_{(i)} \leq t} \left(\frac{n-i}{n-i+1}\right)^{\hat{\delta}(Y_{(i)})}.$$

Se utiliza el enfoque de la ponderación para obtener este estimador, el cual tiene una expresión similar a los estimadores propuestos por Wang and Ng (2008). Por lo tanto, la estimación de la función de supervivencia con el estimador anterior es la siguiente:

$$\hat{S}_{n,P}(t) = \prod_{Y_{(i)} \leq t} \left(\frac{n-i}{n-i+1}\right)^{\hat{\delta}(Y_{(i)})}.$$

Si $\xi_i = 1$ para cada $i = 1, 2, ..., n$, se tiene que $\pi_n(y) = 1$, por lo tanto, $\hat{\delta}(y) = \hat{\gamma}(y)$. Además, se observa que

$$\hat{\gamma}(y) = \frac{\sum_{i=1}^{n} \delta_i K\left(\frac{y-Y_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{y-Y_i}{h_n}\right)},$$

si $\xi_i = 1$ para todo $i = 1, 2, ..., n$. Entonces, $\hat{\gamma}(y) = p_n(y)$, por lo que el estimador anterior, $\hat{S}_{n,P}(t)$, se reduce al estimador presuavizado de Kaplan-Meier (2.3) propuesto por Cao et al (2005), definido en la sección 2.3 del capítulo 2 de este trabajo.

A continuación, se propone un segundo estimador, reemplazando sólo cada δ_i perdido por $\hat{\delta}(Y_{(i)})$. Esto permite obtener un estimador de $\delta(y)$ con la siguiente expresión:

$$\hat{\delta}^*(Y_{(i)}) = \delta_{[i]}\xi_{[i]} + (1 - \xi_{[i]})\hat{\delta}(Y_{(i)}).$$

El segundo estimador propuesto para la función de supervivencia es

$$\hat{S}^{*}_{n,P}(t) = \prod_{Y_{(i)} \leq t} \left(\frac{n-i}{n-i+1} \right)^{\hat{\delta}^{*}(Y_{(i)})} .$$

Observándose que si $\xi_{[i]} = 1 \ i = 1, 2, ..., n$, es decir, si se observan todos los $\delta_{[i]}$, se tiene que $\xi_{[i]}\delta_{[i]} + \left(1 - \xi_{[i]}\right) \hat{\delta}\left(Y_{(i)}\right) = \delta_{[i]} \ i = 1, 2, ..., n$. Por lo que este estimador propuesto, $\hat{S}^{*}_{n,P}(t)$, se reduce al estimador de Kaplan-Meier (2.2), definido en la sección 2.2 del capítulo 2 del presente trabajo.

4 SIMULACIÓN

Se ha realizado un estudio de simulación para evaluar, en una muestra de tamaño finito las propiedades de los estimadores propuestos y comparar los resultados con el estimador de Kaplan-Meier y con los estimadores propuestos por Wang and Ng (2008) bajo MAR (Missing At Random), es decir, los datos son perdidos aleatoriamente aunque esta pérdida está asociada a variables presentes. Para ello, se calcula la distancia de Kolmogorov-Smirnov y el error cuadrático medio integrado (MISE). El estimador de Kaplan-Meier puede servir como "gold standard", aunque es practicamente inalcanzable debido a los indicadores de censura perdida.

En las simulaciones, el tiempo de vida Y y la variable de censura C se generan a partir de las distribuciones exponenciales exp(1) y exp(1/4) para un 20 % de censura, exp(1) y exp(2/3) para un 40 % de censura y de exp(1) y exp(7/3) para una 70 % de censura, respectivamente. El tamaño muestral escogido fue $n = 30, 60, 100, 200$. El mecanismo de perdidos sigue el modelo $logit(\pi(y)) = \theta_1 + \theta_2 y$ con diferentes valores de $\theta = (\theta_1, \theta_2)$. Para un 20 % de censura, θ fue $(1{,}25, 0{,}13)$ y $(0{,}5, -0{,}10)$ para una tasa de indicadores de censura perdidos de un 0,2 y un 0,4, respectivamente. Para una censura del 40 % de censura, θ fue $(1{,}25, 0{,}15)$ y $(0{,}70, -0{,}28)$, para una tasa de indicadores de censura perdidos de un 0,2 y un 0,4, respectivamente. Para un 70 % de censura, θ fue $(1{,}40, -0{,}12)$ y $(0{,}45, -0{,}18)$ para una tasa de indicadores de censura perdidos de un 0,20 y 0,40, respectivamente.

Se calculan los siguientes estimadores de la función de supervivencia:

- Estimador de Kaplan-Meier que tiene la siguiente expresión:

$$S_n^{KM}(t) = \prod_{Y_{(i)} \leq t} \left(\frac{n-i}{n-i+1} \right)^{\delta_{[i]}},$$

  donde $Y_{(i)}$ son los datos ordenados y $\delta_{[i]}$ sus concomitantes.

- Estimadores propuestos:

  - El estimador $\hat{S}_{n,P}(t)$, que si $\xi_i = 1$ para cada $i = 1, 2, ..., n$, se reduce al estimador presuavizado de Kaplan-Meier (2.3) propuesto por Cao et al (2005), con la siguiente expresión:

$$\hat{S}_{n,P}(t) = \prod_{Y_{(i)} \leq t} \left(\frac{n-i}{n-i+1} \right)^{\hat{\delta}(Y_{(i)})},$$

donde los $Y_{(i)}$ representan los datos ordenados

- El estimador $\hat{S}^*_{n,P}(t)$ que si $\xi_i = 1$ para cada $i = 1, 2, ..., n$, se reduce al estimador de Kaplan-Meier (2.2), con la siguiente expresión:

$$\hat{S}^*_{n,P}(t) = \prod_{Y_{(i)} \leq t} \left(\frac{n-i}{n-i+1} \right)^{\hat{\delta}^*(Y_{(i)})}.$$

- Estimadores propuestos por Wang and Ng (2008):

  - El primer estimador ponderado $\hat{S}_{n,W}(t)$, que si $\xi_{[i]} = 1$ $i = 1, 2, ..., n$, se reduce al estimador presuavizado de Kaplan-Meier (2.3) propuesto por Cao et al (2005), con la siguiente expresión:

$$\hat{S}_{n,W}(t) = \prod_{i:Y_{(i)} \leq t} \left(\frac{n-i}{n-i+1} \right)^{\hat{m}_n(Y_{(i)})}.$$

  - El estimador $\hat{S}_{n,I}(t)$, que si $\xi_{[i]} = 1$ $i = 1, 2, ..., n$, se reduce al estimador de Kaplan-Meier (2.2), con la siguiente expresión:

$$\hat{S}_{n,I}(t) = \prod_{i:Y_{(i)} \leq t} \left(\frac{n-i}{n-i+1} \right)^{\xi_{[i]}\delta_{[i]} + (1-\xi_{[i]})\hat{m}_n(Y_{(i)})},$$

    donde $\delta_{[i]}$, $\xi_{[i]}$ son los concomitantes.

  - El estimador imputado $\tilde{S}_{N,I}(t)$, que si $\xi_{[i]} = 1$ $i = 1, 2, ..., n$, se reduce al estimador de Kaplan-Meier (2.2), con la siguiente expresión

$$\tilde{S}_{N,I}(t) = \prod_{i:Y_{(i)} \leq t} \left(\frac{n-i}{n-i+1} \right)^{\xi_{[i]}\delta_{[i]} + (1-\xi_{[i]})\tilde{m}_n(Y_{(i)})},$$

    donde $\delta_{[i]}$, $\xi_{[i]}$ son los concomitantes.

Para obtener $\hat{\delta}(Y_{(i)})$ y $\hat{\delta}^*(Y_{(i)})$, definidos como:

$$\hat{\delta}^*(Y_{(i)}) = \delta_{[i]}\xi_{[i]} + (1-\xi_{[i]})\hat{\delta}(Y_{(i)})$$

y

$$\hat{\delta}(y) = \frac{\hat{\gamma}(y)\hat{\pi}(y) + (1 - \hat{\pi}(y))}{2 - \hat{\pi}(y)}.$$

Se necesita calcular:

$$\pi_n(y) = \frac{\sum_{i=1}^{n} \xi_i W\left(\frac{y-Y_i}{b_n}\right)}{\sum_{i=1}^{n} W\left(\frac{y-Y_i}{b_n}\right)}$$

y

$$\hat{\gamma}(y) = \frac{\sum_{i=1}^{n} \delta_i \xi_i K\left(\frac{y-Y_i}{h_n}\right)}{\sum_{i=1}^{n} \xi_i K\left(\frac{y-Y_i}{h_n}\right)},$$

estimadores de $\pi(y) = \mathbb{P}(\xi = 1 \mid Y = y)$ y $\gamma(y) = \mathbb{P}(\delta = 1 \mid Y = y, \xi = 1)$, respectivamente.

También se necesitan obtener $\hat{m}_n(y)$ y $\tilde{m}_n(y)$, que tienen las siguientes expresiones:

$$\hat{m}_n(y) = \frac{\sum_{i=1}^{n} \left(\frac{\xi_i \delta_i}{\pi_n(Y_i)}\right) K\left(\frac{y-Y_i}{h_n}\right)}{\sum_{i=1}^{n} \left(\frac{\xi_i}{\pi_n(Y_i)}\right) k\left(\frac{y-Y_i}{h_n}\right)}$$

y

$$\tilde{m}_n(y) = \frac{\sum_{i=1}^{n} \xi_i \delta_i K\left(\frac{y-Y_i}{h_n}\right)}{\sum_{i=1}^{n} \xi_i K\left(\frac{y-Y_i}{h_n}\right)}.$$

Se tiene que $\hat{\gamma}(y)$=$\tilde{m}_n(y)$. Finalmente se obtienen los estimadores $\pi_n(y)$, $\hat{\gamma}(y)$ y $\hat{m}_n(y)$ a partir del estimador de Nadaraya-Watson, tomando el núcleo gaussiano y usando para cada uno, el parámetro de suavización mediante validación cruzada.

Teniendo en cuenta que el estimador tipo núcleo hereda las propiedades de suavidad del núcleo, se ha elegido el núcleo gaussiano, es decir, la función de densidad de una variable aleatoria que siga una distribución $N(0,1)$, es decir, $K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right)$.
Para calcular el estimador propuesto, se define la función núcleo K de manera que satisfaga las siguientes condiciones:

$\int K(u)\, du = 1$, $\int uK(u)\, du = 0$, $\int u^2 K(u)\, du = \sigma_k^2 > 0$.

El método de validación cruzada selecciona como parámetro de suavizado el valor h que minimiza la siguiente función:

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{m}_{-(i),K}(X_i)\right)^2,$$

donde $\hat{m}_{-(i),K}$ denota el estimador de Nadaraya-Watson construido a partir de la muestra original después de eliminar el par (X_i, Y_i).

El ajuste más adecuado lo logrará aquella $\hat{S}_{KM}(x)$ que logre el mejor balance entre sesgo y varianza. Se puede cuantificar a través de la distancia de Kolmogorov-Smirnov y a través del Error Cuadrático Medio Integrado (MISE).

La distancia de Kolmogorov-Smirnov está dada por $D = max(D^+, D^-)$, siendo D^+ y D^- las distancias de Kolmogorov unilaterales definidas como

$$D^+ = \sup_y (S_n(y) - S(y), 0) = \max_{1 \leq i \leq n} \left(S_n(y_{(i)}) - S\left(y_{(i)}\right), 0\right),$$

$$D^- = \sup_y (S(y) - S_n(y), 0) = \max_{1 \leq i \leq n} \left(S\left(y_{(i)}\right) - S_n(y_{(i-1)}), 0\right),$$

siendo S_n el estimador de la función de supervivencia y S la función real de supervivencia, definidas en una muestra aleatoria $(y_1, y_2, ..., y_n)$ de la variable aleatoria Y tiempo de vida

El error cuadrático medio integrado (MISE) es uno de los criterios de error más empleados y se define como

$$MISE = \mathbb{E}\left[\int (S_n(y) - S(y))^2 \, dy\right],$$

siendo S_n el estimador de la función de supervivencia y S la función real de supervivencia, definidas en una muestra aleatoria $(y_1, y_2, ..., y_n)$ de la variable aleatoria Y tiempo de vida.

Para llevar a cabo la simulación, se generan 500 muestras aleatorias de Monte Carlo de tamaño $n = 30, 60, 100, 200$ bajo cada diferente combinación de tasas de censura y proporción de indicadores de censura perdidos. Para los 500 valores simulados de los estimadores, se calcula la distancia de Kolmogorov-Smirnov y el MISE. El MISE fue calculado sobre el intervalo $[0, 2]$.

Para cada una de las simulaciones se siguieron los mismos pasos que a continuación se describen:

1. Se generan los tiempos de vida y los tiempos de censura para los 3 escenarios: 20 %, 40 % y 70 % de censura. Se ordenan los tiempos de vida, calculando a partir de ellos la función de supervivencia como la $\exp(-\lambda Y_{(i)})$, con $\lambda = 1$.

2. Para cada uno de los escenarios de censura, se calcula:

   a) $Y = T \wedge C$ y $\delta = 1_{\{Y \leq C\}}$.

   b) Para cada uno de los 2 escenarios de tasas de pérdida (0.2 y 0.4), se calcula:

1) A partir de la función $logit(\pi(y)) = \theta_1 + \theta_2$, se tiene $\pi_1(y)$ y $\pi_2(y)$ para una tasa de perdidos de 0,2 y 0,4, respectivamente. Estas funciones generan ξ, siendo $\xi = 0$ si δ_i está perdido y $\xi = 1$ si δ_i no está perdido.
2) Se obtienen los tiempos de vida ordenados $Y_{(i)}$ y sus concomitantes $\delta_{[i]}$ y $\xi_{[i]}$.
3) A continuación, mediante la validación cruzada se generan los parámetros de suavización para los estimadores $\pi_n(y)$, $\hat{\gamma}(y)$ y $\hat{m}_n(y)$, los cuáles se obtienen a partir del estimador de Nadaraya-Watson, con función núcleo gaussiano.
4) Se calculan los estimadores π_n, $\hat{\gamma}$ y, a partir de ellos, el estimador $\hat{\delta}$, que tiene la siguiente forma:
$$\hat{\delta}(y) = \frac{\hat{\gamma}(y)\hat{\pi}(y) + (1 - \hat{\pi}(y))}{2 - \hat{\pi}(y)}.$$
5) A continuación se calculan los estimadores $\hat{S}_n^{KM}$, $\hat{S}_n^{P}$, $\hat{S}_n^{P*}$, $\hat{S}_{n,W}$, $\hat{S}_{n,I}$, $\tilde{S}_{N,I}$.
6) Se calcula la distancia de Kolmogorov-Smirnov y el MISE para cada uno de los estimadores obtenidos, para los 3 escenarios de censura, para las 2 tasas de datos perdidos y para cada tamaño muestral $n = 30, 60, 100, 200$.

Todos los cálculos y gráficos fueron realizados con el programa R v2.12.2.

En las tablas 4.1 y 4.2 de esta sección se muestran la distancia de Kolmogorov-Smirnor y el error cuadrático medio integrado para cada uno de los estimadores en los tres escenarios de censura y para las dos tasas de perdidos ($\pi_1(y)$ y $\pi_2(y)$ correspondientes a 0,2 y 0,4, respectivamente). Se observa que los estimadores propuestos tienen una distancia de Kolmogorov-Smirnov y MISE similares, cercanos a los propuestos por Wang and Ng (2008) y cercanos al estimador de Kaplan-Meier. Esto sugiere que los estimadores obtienen buenos y similares resultados.

Tabla 4.1: Distancia de Kolmogorov-Smirnov bajo MAR

		20 % censura		40 % censura		70 % censura	
n	Estimadores	$\pi_1(Y)$	$\pi_2(Y)$	$\pi_1(Y)$	$\pi_2(Y)$	$\pi_1(Y)$	$\pi_2(Y)$
	$\hat{S}_n^{KM}$	0,16689	0,16688	0,19693	0,19693	0,27763	0,27763
	$\hat{S}_n^{P}$	0,16525	0,18216	0,17785	0,17781	0,33193	0,33080
	$\hat{S}_n^{P*}$	0,16722	0,17195	0,19500	0,19618	0,28491	0,28966
30	$\hat{S}_{n,W}$	0,15926	0,16292	0,18670	0,19370	0,34712	0,35380
	$\hat{S}_{n,I}$	0,16733	0,16983	0,19837	0,20572	0,29715	0,31874
	$\tilde{S}_{N,I}$	0,16730	0,16924	0,19815	0,20544	0,29675	0,31754
	$\hat{S}_n^{KM}$	0,11716	0,11716	0,13886	0,13886	0,23220	0,23219
	$\hat{S}_n^{P}$	0,12485	0,14777	0,12597	0,12987	0,27187	0,27473
	$\hat{S}_n^{P*}$	0,11811	0,12287	0,13954	0,13879	0,23802	0,23787
60	$\hat{S}_{n,W}$	0,11255	0,11382	0,12925	0,13606	0,28145	0,28596
	$\hat{S}_{n,I}$	0,11785	0,11762	0,14109	0,14427	0,24671	0,25946
	$\tilde{S}_{N,I}$	0,11757	0,11713	0,14097	0,14455	0,24726	0,25964
	$\hat{S}_n^{KM}$	0,09319	0,09319	0,11302	0,11302	0,20497	0,20497
	$\hat{S}_n^{P}$	0,10590	0,13380	0,10080	0,10421	0,23096	0,23615
	$\hat{S}_n^{P*}$	0,09470	0,10217	0,11189	0,11080	0,20669	0,21103
100	$\hat{S}_{n,W}$	0,08865	0,09153	0,10221	0,11408	0,23833	0,24418
	$\hat{S}_{n,I}$	0,09367	0,09554	0,11278	0,10602	0,21602	0,22836
	$\tilde{S}_{N,I}$	0,09334	0,09509	0,11269	0,11451	0,21679	0,22936
	$\hat{S}_n^{KM}$	0,06496	0,06496	0,08021	0,08021	0,16880	0,16880
	$\hat{S}_n^{P}$	0,08367	0,11522	0,07404	0,07826	0,19030	0,19888
	$\hat{S}_n^{P*}$	0,06650	0,07542	0,08073	0,07843	0,17041	0,17627
200	$\hat{S}_{n,W}$	0,06205	0,06374	0,07434	0,07573	0,19426	0,19741
	$\hat{S}_{n,I}$	0,06546	0,06615	0,08175	0,08125	0,17638	0,18642
	$\tilde{S}_{N,I}$	0,06514	0,06544	0,08151	0,08133	0,17723	0,18749

Tabla 4.2: Error Cuadrático Medio Integrado (MISE) bajo MAR

		20 % censura		40 % censura		70 % censura	
n	Estimadores	$\pi_1(Y)$	$\pi_2(Y)$	$\pi_1(Y)$	$\pi_2(Y)$	$\pi_1(Y)$	$\pi_2(Y)$
	$\hat{S}_n^{KM}$	0,01474	0,01474	0,02079	0,02079	0,07231	0,07231
	$\hat{S}_n^{P}$	0,01728	0,02301	0,01731	0,01717	0,06062	0,06320
	$\hat{S}_n^{P*}$	0,01542	0,01691	0,02068	0,02092	0,06390	0,05637
30	$\hat{S}_{n,W}$	0,01475	0,01557	0,02048	0,02234	0,06928	0,07579
	$\hat{S}_{n,I}$	0,01545	0,01615	0,02183	0,02446	0,07517	0,08361
	$\tilde{S}_{N,I}$	0,01544	0,01601	0,02183	0,02440	0,07539	0,08343
	$\hat{S}_n^{KM}$	0,00694	0,00694	0,00922	0,00922	0,04053	0,04053
	$\hat{S}_n^{P}$	0,01002	0,01654	0,00806	0,00865	0,04005	0,04601
	$\hat{S}_n^{P*}$	0,00721	0,00865	0,00961	0,00972	0,03861	0,03389
60	$\hat{S}_{n,W}$	0,00680	0,00708	0,00899	0,01027	0,04295	0,04594
	$\hat{S}_{n,I}$	0,00715	0,00736	0,00995	0,01103	0,04429	0,04681
	$\tilde{S}_{N,I}$	0,00712	0,00730	0,00993	0,01098	0,04490	0,04748
	$\hat{S}_n^{KM}$	0,00441	0,00441	0,00595	0,00595	0,02658	0,02658
	$\hat{S}_n^{P}$	0,00815	0,01535	0,00521	0,00584	0,02849	0,03592
	$\hat{S}_n^{P*}$	0,00476	0,00651	0,00594	0,00598	0,02517	0,02449
100	$\hat{S}_{n,W}$	0,00434	0,00471	0,00554	0,00609	0,02886	0,03144
	$\hat{S}_{n,I}$	0,00456	0,00490	0,00612	0,00661	0,02912	0,03252
	$\tilde{S}_{N,I}$	0,00453	0,00485	0,00611	0,00656	0,02952	0,03316
	$\hat{S}_n^{KM}$	0,00214	0,00214	0,00294	0,00294	0,01504	0,01504
	$\hat{S}_n^{P}$	0,00546	0,01250	0,00286	0,00346	0,01931	0,02801
	$\hat{S}_n^{P*}$	0,00233	0,00367	0,00309	0,00290	0,01443	0,01577
200	$\hat{S}_{n,W}$	0,00213	0,00231	0,00294	0,00295	0,01633	0,01765
	$\hat{S}_{n,I}$	0,00224	0,00239	0,00323	0,00318	0,01619	0,01834
	$\tilde{S}_{N,I}$	0,00222	0,00235	0,00321	0,00313	0,01646	0,01873

En las tablas 4.1 y 4.2, se muestra el comportamiento de los estimadores. En general $\hat{S}_n^{P*}$ es el mejor estimador (salvo $\hat{S}_n^{KM}$ que es inobservable en la práctica con datos perdidos) para censuras del 40 % y 70 %, siendo ligeramente peor que los estimadores $\hat{S}_{n,W}$, $\hat{S}_{n,I}$ y $\hat{S}_{N,I}$ para el 20 % de censura. En ese caso de bajo porcentaje de censura, $\hat{S}_{n,W}$ es el que mejor comportamiento presenta. Un hecho importante es que el error cometido por $\hat{S}_n^{P}$ no decrece al decrecer el porcentaje

de censura. Concretamente, los errores de $\hat{S}_n^P$ para un 40 % de censura son frecuentemente más pequeños (en ocasiones incluso mucho más pequeños) que para un 20 % de censura.

En las tablas 4.1 y 4.2 se observa para el estimador $\hat{S}_n^P$, que aumentando el porcentaje de censura del 20 % al 40 % no aumenta la distancia de Kolmogorov-Smirnov ni el MISE.
Debido a lo anterior se realizan las simulaciones para una tasa de perdidos de 0, $\pi_0(Y)$. En las tablas 4.3 y 4.4 figuran las distancias de Kolmogorov-Smirnov y MISE para una tasa de perdidos de 0, $\pi_0(Y)$, para los tres escenarios de censura y para $n = 30, 60, 100, 200$. Se observa que los estimadores propuestos $\hat{S}_n^P$ y $\hat{S}_n^{P*}$ obtienen buenos resultados, al incrementarse las distancias de Kolmogorov-Smirnov y los valores de MISE al aumentar la tasa de censura. Se comprueba como $\hat{S}_n^{P*}$ se reduce al estimador de Kaplan-Meier, obteniendo resultados similares entre ellos.

Tabla 4.3: Distancia de Kolmogorov-Smirnov bajo MAR sin datos perdidos

		20 % censura	40 % censura	70 % censura
n	Estimadores	$\pi_0(Y)$	$\pi_0(Y)$	$\pi_0(Y)$
30	$\hat{S}_n^{KM}$	0,16615	0,20002	0,28760
	$\hat{S}_n^P$	0,15552	0,18172	0,34023
	$\hat{S}_n^{P*}$	0,16615	0,20002	0,28760
60	$\hat{S}_n^{KM}$	0,11993	0,14424	0,23563
	$\hat{S}_n^P$	0,11195	0,12909	0,28092
	$\hat{S}_n^{P*}$	0,11993	0,14424	0,23563
100	$\hat{S}_n^{KM}$	0,09411	0,11464	0,19516
	$\hat{S}_n^P$	0,08802	0,10228	0,24119
	$\hat{S}_n^{P*}$	0,09411	0,11454	0,19516
200	$\hat{S}_n^{KM}$	0,06496	0,08021	0,16881
	$\hat{S}_n^P$	0,06057	0,07109	0,19314
	$\hat{S}_n^{P*}$	0,06496	0,08021	0,16881

Tabla 4.4: Error Cuadrático Medio Integrado (MISE) bajo MAR sin datos perdidos

		20 % censura	40 % censura	70 % censura
n	Estimadores	$\pi_0(Y)$	$\pi_0(Y)$	$\pi_0(Y)$
30	$\hat{S}_n^{KM}$	0,01454	0,02033	0,07442
	$\hat{S}_n^{P}$	0,01352	0,01827	0,06523
	$\hat{S}_n^{P*}$	0,01454	0,02033	0,07443
60	$\hat{S}_n^{KM}$	0,00731	0,01006	0,04352
	$\hat{S}_n^{P}$	0,00679	0,00881	0,04163
	$\hat{S}_n^{P*}$	0,00731	0,01006	0,04352
100	$\hat{S}_n^{KM}$	0,00455	0,00604	0,02619
	$\hat{S}_n^{P}$	0,00422	0,00517	0,02818
	$\hat{S}_n^{P*}$	0,00428	0,00527	0,02619
200	$\hat{S}_n^{KM}$	0,00214	0,00294	0,01504
	$\hat{S}_n^{P}$	0,00199	0,00253	0,01566
	$\hat{S}_n^{P*}$	0,00214	0,00294	0,01504

A continuación se muestra en la figura 4.1, el gráfico de la función de supervivencia y las curvas de $\hat{S}_n^{KM}$, $\hat{S}_n^{P}$ y $\hat{S}_n^{P*}$. Se observa como las curvas de los estimadores propuestos se aproximan a la curva de supervivencia real y las curvas de los estimadores propuestos se solapan entre sí. Esto sugiere que las estimaciones realizadas son buenas, y que los resultados de los estimadores propuestos son similares en términos de sesgo.

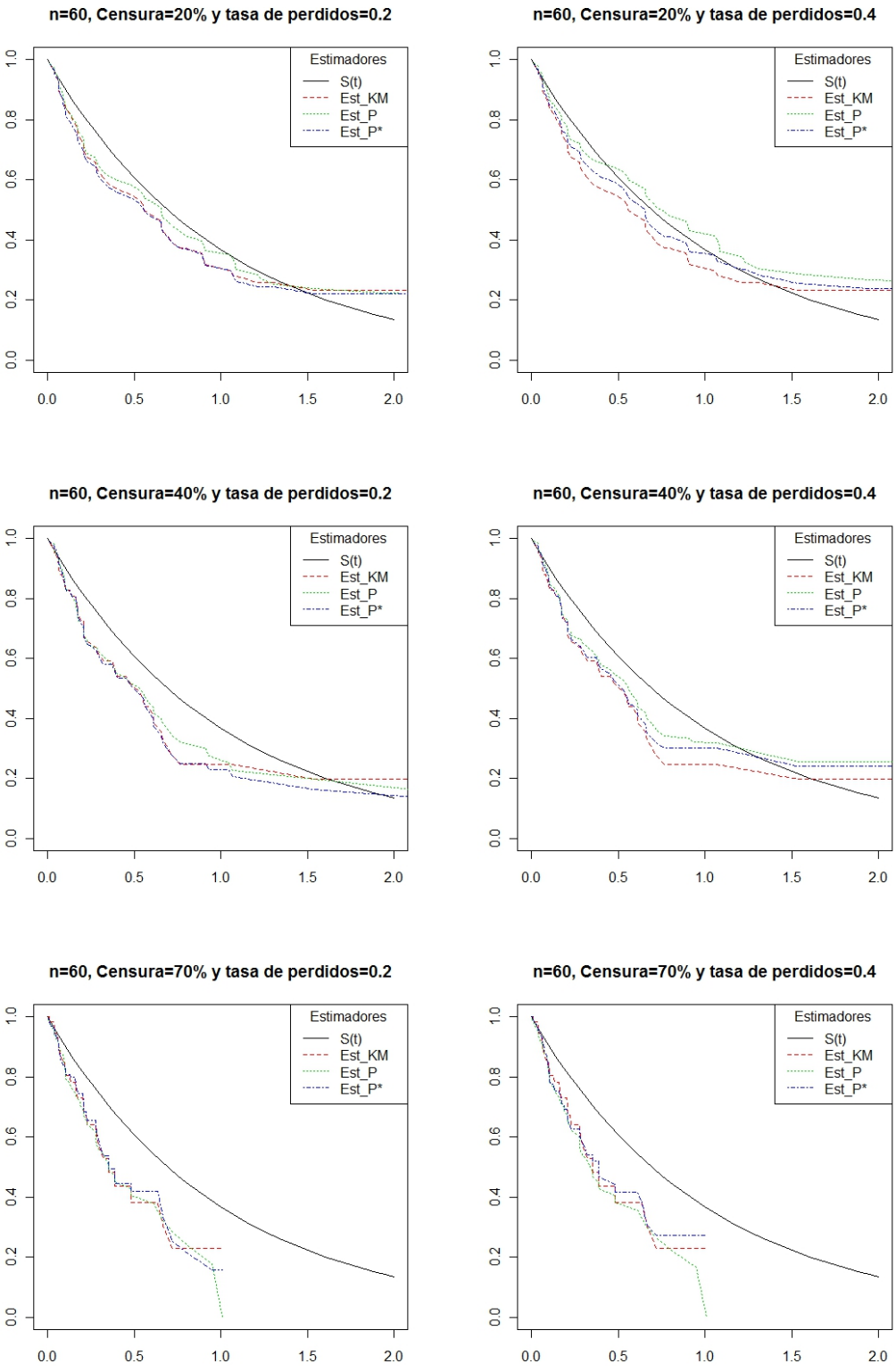

Figura 4.1: Estimaciones de $\hat{S}_n^{KM}$, $\hat{S}_n^{P}$ y $\hat{S}_n^{P*}$ y la función de supervivencia real $S(t)$

5 APLICACIÓN A DATOS REALES

5.1. Diseño del estudio

Se trata de un estudio realizado en el Complexo Hospitalario Universitario de A Coruña (CHUAC) en el que se pretenden estudiar los siguientes objetivos:

- Determinar si la duración del intervalo de tiempo transcurrido entre el primer síntoma y el diagnóstico, y entre el primer síntoma y el tratamiento, modifica la supervivencia de los pacientes con cáncer de colon y de recto.
- Determinar, en pacientes con cáncer colorrectal no metastático tratados con intención curativa, si diferentes estrategias de seguimiento se asocian con una mejor supervivencia.
- Determinar, en pacientes con cáncer colorrectal; la supervivencia global, la supervivencia específica (mortalidad relacionada con el tumor), la supervivencia libre de progresión y la supervivencia sin recidiva
- Determinar, de las variables recogidas en este estudio, aquellas que modifican el pronóstico de los pacientes con cáncer colorrectal.

Se realiza un estudio observacional de cohortes con seguimiento prospectivo, incluyendo a los pacientes con confirmación anatomopatológica de cáncer colorrectal (CIE-9 153 y 154), casos incidentes de cáncer colorrectal diagnosticados durante el periodo 2006-2012; excluyéndose los casos prevalentes o recurrentes, casos de cáncer múltiple, atendidos en hospitales privados, detectados por cribado de cáncer colorrectal y diagnosticados en otros hospitales pero referidos a los hospitales incluidos en el estudio.

Se estudian n=1407 pacientes que permiten detectar un riesgo relativo R.R $\geq 1,25$, asumiendo un 50 % de exposición y un porcentaje de censura del 50 % (nivel de confianza 95 %; potencia estadística 80 %).

Se realizará una entrevista con los pacientes y revisión de historias clínicas de atención primaria y especializada. Realizándose un seguimiento de los pacientes para estudiar su supervivencia y la presencia de eventos (progresión/recidiva/muerte) durante el seguimiento. Para ello se recogerán variables de identificación, demográficas del paciente, antecedentes familiares de cáncer, comorbilidad (índice de comorbilidad de Charlson), primeros síntomas, motivo de consulta, exploraciones

y pruebas pre-diagnóstica. Se estudiará la demora (primeros síntomas - primera consulta primaria; primeros síntomas - primera consulta especializada; primeros síntomas - diagnóstico anatomopatológico; primeros síntomas-tratamiento). Características del tumor, tratamiento quirúrgico, tratamiento oncológico. Así como pruebas realizadas en el seguimiento y mortalidad relacionada o no con el tumor, supervivencia global, supervivencia específica, supervivencia específica, supervivencia libre de progresión, supervivencia libre de recidiva.

Este estudio se realiza contemplando los siguientes aspectos ético-legales: aprobación por el CEIC Galicia (2004/159), se solicita consentimiento informado, se realiza según las normas de buena práctica clínica de la declaración de Helsinki y se respeta la ley de protección de datos 15/1999.

5.2. Resultados

5.2.1. Características generales de la muestra

En el estudio observacional de seguimiento prospectivo, se estudian a los pacientes diagnosticados de cáncer colorrectal durante el periodo 2006-2012, siendo el 60,9 % hombres con una edad de $70,0 \pm 11,1$ años y una comorbilidad segn el Índice Charlson ajustado por edad de $3,6 \pm 1,8$

La localizaciones más frecuentes del tumor son colon sigmoide excepto unión recto sigmoidea (31,3 %) y recto-ampolla rectal (24,5 %). Se objetiva el grado tumoral más frecuente T3 (56,3 %), N0 (53,4 %) y M0 (68,8 %).

El 49,6 % tiene metástasis, siendo las más frecuentes la ganglionar (53,3 %) y hepáticas (13,7 %). Tienen infiltraciones en vísceras vecinas el 14,9 %, siendo localizándose las vísceras afectadas la pared abdominal (51,5 %) e intestino (21,6 %). Tuvieron permeación vascular el 24,3 %, linfática el 42,2 % y nerviosa el 21,9 %. Sólo el 3,1 % tiene infiltración en bordes quirúrgicos.

El 91,3 % tuvo tratamiento quirúrgico, siendo el más habitual la resección sin colostomía en el 60,2 % y la resección con colostomía en el 14,6 %. El tipo de intervención quirúrgica más frecuente es la programada en el 90,3 % de los casos, siendo el abordaje quirúrgico habitual el abierto (laparotomía). La intención del tratamiento quirúrgico es el radical-erradicador en el 84,9 % y las técnicas quirúrgicas más frecuentes son la resección anterior del sigma (Operación de Dixon) en el 22,7 % y la hemicolectomía derecha en el 22,2 %. El tipo de anastomosis en la mayoría de los casos fue la termino-terminal (50,1 %) y el procedimiento de anastomosis más realizado es la mecánica en el 55,3 %. Se realiza resección de vísceras en el 17,3 % siendo las vísceras más frecuentes las de intestino (24,0 %) y la de próstata (23,1 %). La resección de metástasis se realizó en el 2,5 %, siendo hepáticas en el 89,3 %. Los pacientes tienen morbiidad quirúgica en el 24,1 %, siendo sistémicas en

el 34,6 % e infección de herida en el 12,0 %.

Un 14,3 % de los pacientes son reintervenidos, siendo esta reintervención urgente en el 22,7 % y programada en el 14,3 %. La causa de reintervención urgente más frecuente es la deshicencia de sutura y de la programada es el cierre de colostomía-ileostomía.

Según el tratamieno oncológico, los pacientes toman quimioterapia previa a la cirugía (12,3 %) y post-cirugía (39,4 %). Además, reciben radioterapia previa a la cirugía (10,0 %) y post-cirugía el 8,7 %.

Una vez los pacientes se consideren libres de enfermedad, es decir, tengan un TAC negativo y/o resonancia magnética negativa y/o colonoscopia negativa, se inicia seguimiento. Un 55,7 % está libre de enfermedad con un tiempo de seguimiento medio de $1,6 \pm 1,1$ años (rango=$(0 - 5,48)$), teniendo eventos en el seguimiento un 5,6 % recidiva, 4,1 % metástasis y un 3,9 % nuevas neoplasias. El 38,3 % de los pacientes son exitus, siendo la causa de la muerte relacionada con el tumor en el 71,8 % de los casos y existiendo una tasa de perdidos del 0,035 (3,5 %).

5.2.2. Supervivencia de los pacientes con cáncer colorrectal

A continuación se estudian dos bases de datos de estudios de supervivencia de cáncer colorrectal. El primero es descrito anteriormente con un tiempo medio de seguimiento de $1,6 \pm 1,1$ años. En la otra base de datos se estudian $n = 1111$ pacientes diagnósticados de cáncer colorrectal, considerando las mismas variables que en la anterior base, que tienen un tiempo medio de seguimiento de $6,0 \pm 4,9$ años.

En algunos estudios, se estima la función de supervivencia utilizando el estimador de Kaplan-Meier para los datos en los que el indicador de censura está perdido, no sabiendo si el fallecimiento es producido por el evento de interés o no, por lo que se está asumiendo que los datos perdidos son no censurados. Esta suposición provoca un sesgo en la estimación de la función de supervivencia.

En la figura 5.1 se muestran las curvas del estimador de Kaplan-Meier para los datos observados, $EstKM$ y para todos los datos, es decir, suponiendo que los indicadores de censura perdidos son no censurados, $EstKM_C$. En ambas figuras, se comprueba que la utilización del estimador de Kaplan-Meier en observaciones no completas, indica una disminución de la supervivencia, provocando un sesgo positivo.

Al comparar las gráficas, teniendo en cuenta que la de la izquierda se corresponde con una tasa de perdidos de 0,035 y la de la derecha con una tasa de perdidos de 0,42. Se observa que al incre-

mentarse la tasa de perdidos aumenta el sesgo, implicando una disminución de la supervivencia.

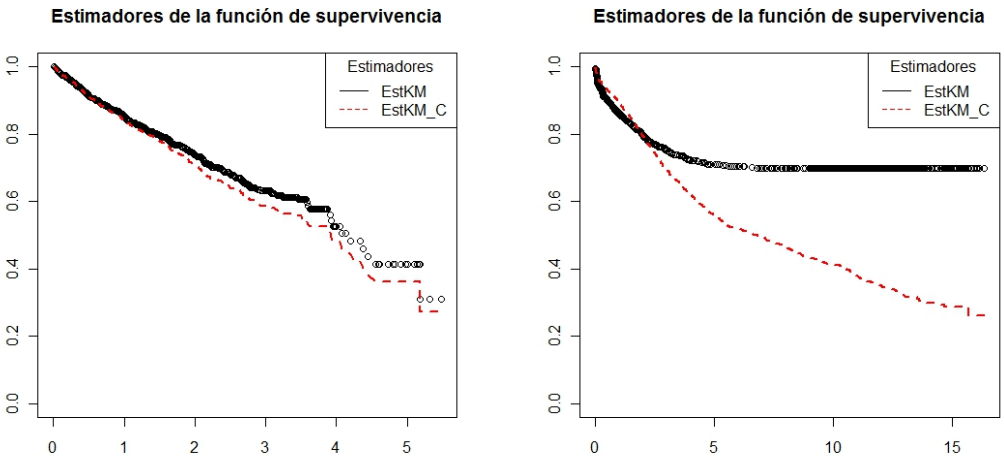

Figura 5.1: Estimación de Kaplan-Meier para datos observados y suponiendo los datos perdidos no censurados (tasa de perdidos=0,035 vs. 0,42)

Por lo que se reafirma que el estimador de Kaplan-Meier sólo se puede utilizar para observaciones completas

A continuación, en la figura 5.2, se muestran como se comportan los estimadores propuestos y los estimadores de Wang and Ng (2008) para la función de supervivencia, aplicados a los datos con una tasa de perdidos de 0,035.

En las gráficas superiores se representan el estimador de Kaplan-Meier para los datos observados junto con los dos estimadores propuestos y con los estimadores definidos por Wang and Ng (2008). En la primera figura, se observa que ambos estimadores propuestos tienen un comportamiento similar, aproximándose más al estimador de Kaplan-Meier el estimador $\hat{S}_n^{P*}$. En la gráfica donde se representan los estimadores de Wang and Ng (2008) también se observan resultados similares entre ellos y próximos a los obtenidos con el estimador de Kaplan-Meier para los datos observados. Aproximándose más al estimador de Kaplan-Meier y los estimadores $\hat{S}_{n,I}$, $\tilde{S}_{N,I}$.

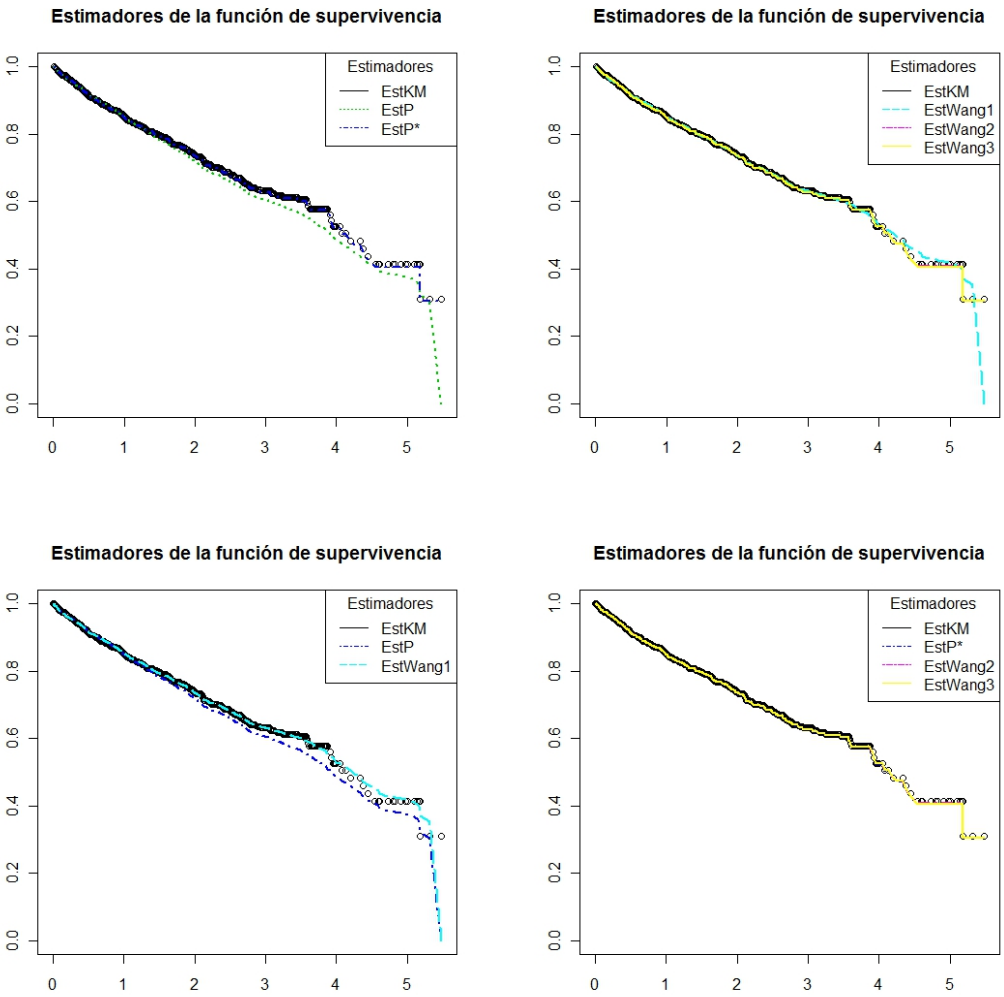

Figura 5.2: Estimaciones de la función de supervivencia con indicador de censura perdido aleatoriamente (tasa de perdidos=0,035)

En las gráficas inferiores se comparan los estimadores estudiados. En la de la izquierda, se comparan los estimadores $\hat{S}_n^P$ y $\hat{S}_{n,W}$, obteniendo resultados muy similares entre ellos y comprobándose como estos estimadores, que se reducen al estimador presuavizado de Kaplan-Meier, tienden a cero. En la gráfica de la derecha, se comparan los estimadores $\hat{S}_n^{P*}$, $\hat{S}_{n,I}$ y $\tilde{S}_{N,I}$, obteniendo resultados muy similares entre ellos y con el estimador de Kaplan-Meier.

Para una tasa de perdidos de 0,035 resultan buenos estimadores de la función de supervivencia todos los estimadores propuestos, siendo $\hat{S}_n^{P*}$, $\hat{S}_{n,I}$ y $\tilde{S}_{N,I}$ más próximos al estimador de Kaplan-Meier. Es decir, entre los estimadores propuestos en este trabajo para estos datos de cáncer

colorrectal con una tasa de perdidos de 0,035, parece comportarse mejor el estimador $\hat{S}_n^{P*}$.

A continuación, en la figura 5.3, se muestran como se comportan los estimadores propuestos y los estimadores de Wang and Ng (2008) para la función de supervivencia, aplicados a los datos con una tasa de perdidos de 0,42.

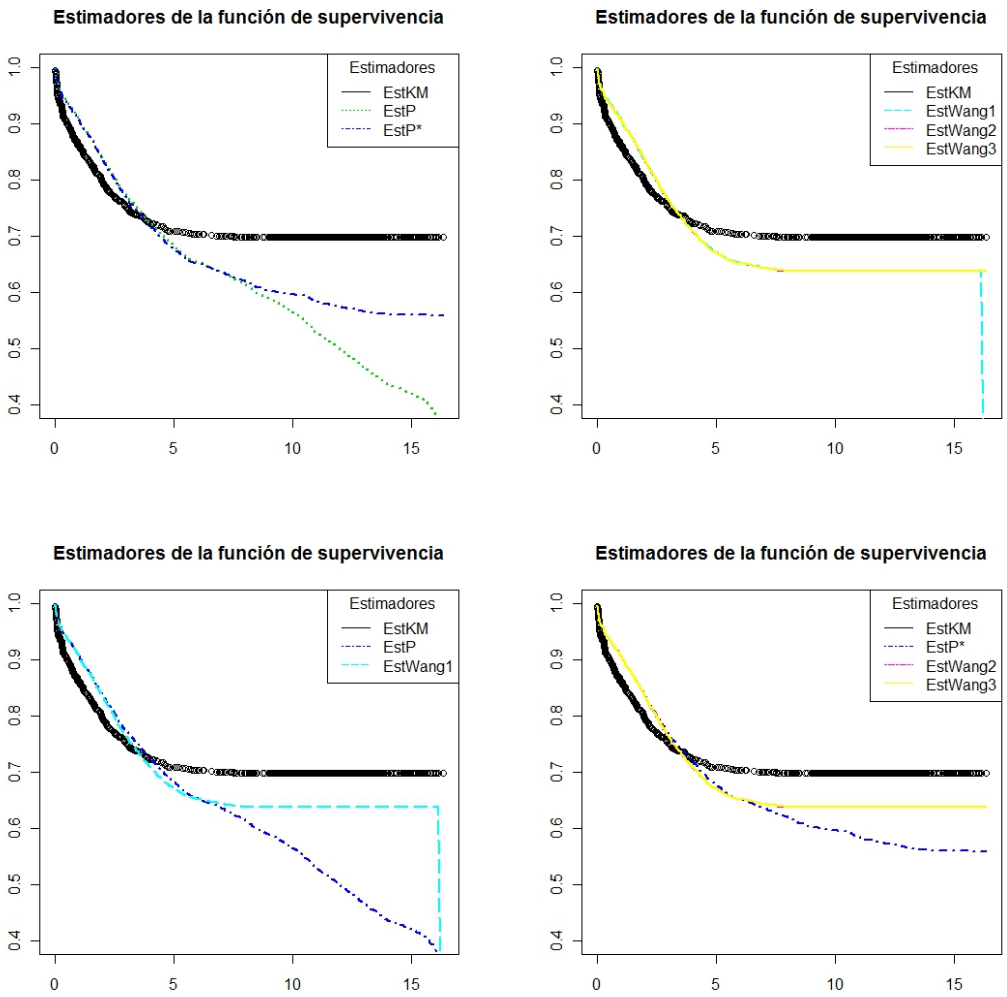

Figura 5.3: Estimaciones de la función de supervivencia con indicador de censura perdido aleatoriamente (tasa de perdidos=0,42)

En las gráficas superiores se representan el estimador de Kaplan-Meier para los datos observados junto con los dos estimadores propuestos y con los estimadores definidos por Wang and Ng (2008). En la primera figura, se observa que ambos estimadores propuestos infraestiman la super-

vivencia, teniendo mejor comportamiento el estimador $\hat{S}_n^{P*}$ obteniendo resultados más próximos al estimador de Kaplan-Meier. Recordamos que el estimador $\hat{S}_n^{P*}$ se reduce al estimador de Kaplan-Meier (2.2), definido en el capítulo 2 en la sección 2.2 de este trabajo, si $\xi_{[i]} = 1\ i = 1, 2, ..., n$. En la gráfica donde se representan los estimadores de Wang and Ng (2008) también se observan resultados similares entre $\hat{S}_{n,I}$, $\tilde{S}_{N,I}$ y próximos a los obtenidos con el estimador de Kaplan-Meier para los datos observados. Estando más alejados los obtenidos con $\hat{S}_{n,W}$, estimador que se reduce al estimador presuavizado de Kaplan-Meier si $\xi_{[i]} = 1\ i = 1, 2, ..., n$

En las gráficas inferiores se comparan los estimadores estudiados. En la de la izquierda, se comparan los estimadores $\hat{S}_n^P$ y $\hat{S}_{n,W}$, obteniendo resultados muy similares entre ellos y comprobándose como estos estimadores, que se reducen al estimador presuavizado de Kaplan-Meier, tienden a cero En la gráfica de la derecha, se comparan los estimadores $\hat{S}_n^{P*}$, $\hat{S}_{n,I}$ y $\tilde{S}_{N,I}$, obteniendo resultados muy similares entre ellos y con el estimador de Kaplan-Meier.

Para una tasa de perdidos de 0,42 los estimadores de la función de supervivencia $\hat{S}_n^{P*}$, $\hat{S}_{n,I}$, $\tilde{S}_{N,I}$ obtienen mejores resultados y más próximos al estimador de Kaplan-Meier que $\hat{S}_n^P$ y $\hat{S}_{n,W}$. Es decir, entre los estimadores propuestos en este trabajo para estos datos de cáncer colorrectal con una tasa de perdidos de 0,42, parece comportarse mejor el estimador $\hat{S}_n^{P*}$, aunque infraestime la supervivencia

Al aplicar los estimadores propuestos en dos bases de datos de cáncer colorrectal con tasas de perdidos de 0,035 y 0,42, repectivamente, se comprueba que estos estimadores obtienen buenos resultados, siendo mejores los de $\hat{S}_n^{P*}$ que los de $\hat{S}_n^P$ y comportándose de una forma más próxima al estimador de Kaplan-Meier cuando la tasa de perdidos es pequeña.

A continuación se estudian modelos de Cox para las distintas bases de datos, con tasas de perdidos 0,035 y 0,42, respectivamente. El modelo de Cox permite estimar la función de supervivencia ajustando por covariables a partir de los datos observados, ver, por ejemplo, Cox and Oakes (1984) y Härdle et al. (2004) .
Es por ello, que a partir de los estimadores propuestos $\hat{\delta}(y)$ y $\hat{\delta}^*(y)$, siendo $\hat{\delta}(y) = \frac{\hat{\gamma}(y)\hat{\pi}(y)+(1-\hat{\pi}(y))}{2-\hat{\pi}(y)}$ y $\hat{\delta}^*(Y_{(i)}) = \delta_{[i]}\xi_{[i]} + (1 - \xi_{[i]})\hat{\delta}(Y_{(i)})$, se generan los valores de δ cuando están perdidos. Es decir, sólo se reemplazan los valores perdidos de δ y para los no perdidos se toma el valor de δ observado.

Una vez obtenidos se realiza un análisis de regresión de Cox para determinar que variables se asocian con una mayor probabilidad de fallecer por causa del tumor de cáncer colorrectal. Este análisis se realiza en las dos bases de datos para comprobar el comportamiento de los estimadores

de δ con tasas de perdidos 0,035 y 0,42, respectivamente y se compara con el modelo de Cox obtenido con los datos de δ en los casos completos, es decir, donde el indicador de censura está no perdido.

Para los datos de cáncer colorrectal con una tasa de perdidos de 0,035 se obtuvieron los siguientes resultados:

En las tablas 5.1 y 5.2 figuran las estimaciones de los modelos de regresión de Cox comprobándose que ambos obtienen resultados similares, por lo que cuando se reemplazan los valores perdidos de δ por los generados por el estimador $\hat{S}_n^P$ se tienen resultados similares a los que se obtienen sólo con los δ observados

Tabla 5.1: Modelo de regresión de Cox generando δ si $\xi = 0$ con $\hat{S}_n^P$ con una tasa de perdidos de 0,035

Variables	B	Exp(B)	E.T	Z	p
edad	0,0151	1,0152	0,0056	2,705	0,00682
sexo (hombre vs. mujer)	0,0330	1,0333	0,1218	0,269	0,78764
tratamiento quirúrgico (sí)	-2,1597	0,1154	0,1419	-15,223	$< 2e-16$

Tabla 5.2: Modelo de regresión de Cox de las observaciones completas con una tasa de perdidos de 0,035

Variables	B	Exp(B)	E.T	Z	p
edad	0,0277	1,0281	0,0059	4,702	2,58e-06
sexo (hombre vs. mujer)	-0,1343	0,8743	0,1249	-1,075	0,282
tratamiento quirúrgico (sí)	-1,4623	0,2317	0,1659	-8,814	$< 2e-16$

A continuación se muestran los resultados obtenidos en los modelos de regresión de Cox, para los datos de cáncer colorrectal con una tasa de perdidos de 0,42. En las tablas 5.5 y 5.6 figuran las estimaciones de los modelos de regresión de Cox, comprobándose que en ambos se obtienen resultados similares. Por lo que para una tasa de perdidos de 0,42, el estimador $\hat{S}_n^P$ obtiene resultados próximos a los que se tienen al estudiar sólo las observaciones completas.

Tabla 5.3: Modelo de regresión de Cox generando δ si $\xi = 0$ con $\hat{S}_n^P$ con una tasa de perdidos de 0,42

Variables	B	Exp(B)	E.T	Z	p
edad	0,0177	1,0178	0,0046	3,834	0,000126
sexo (hombre vs. mujer)	-0,1204	0,8866	0,1023	-1,177	0,239115
tratamiento (curativo vs. paliativo)	1.4832	4,4072	0,1218	12,178	$< 2e-16$

Tabla 5.4: Modelo de regresión de Cox de las observaciones completas con una tasa de perdidos de 0,42

Variables	B	Exp(B)	E.T	Z	p
edad	0,0215	1,0218	0,0072	3,004	0,00267
sexo (hombre vs. mujer)	-0,1160	0,8905	0,1506	-0,770	0,44143
tratamiento (curativo vs. paliativo)	0,8706	2,3884	0,1514	5,750	8,95e-09

Finalmente, se observa que el comportamiento de los modelos de Cox para diferentes tasas de perdidos, tras generar el indicador de censura sólo cuando está perdido, es similar que el comportamiento de los modelos de Cox sólo con la submuestra de datos observados completamente.

6 CONCLUSIONES

6.1. Estimadores de la función de supervivencia con indicador de censura desconocida

- El estimador de Kaplan-Meier requiere que la variable censura sea siempre observable. Sin embargo, la variable de la censura, δ, puede ser un dato perdido debido a múltiples razones. Así en los estudios de supervivencia, puede ocurrir que sepamos que un paciente ha fallecido y no si la causa ha sido por el evento de interés ($\delta = 1$) o por otras causas no relacionadas ($\delta = 0$)

- Se han desarrollado métodos para estimar la función de supervivencia con indicadores de censura perdida aleatoriamente.

  - Métodos que permiten el uso de la imputación y de la ponderación (estimadores propuestos por Wang and Ng (2008))

    - $\hat{S}_{n,I}(t)$ y $\tilde{S}_{N,I}(t)$ estimadores imputados que se reducen al estimador de Kaplan-Meier

    - $\hat{S}_{n,W}(t)$ estimador que se reduce al estimador de Kaplan-Meier suavizado cuando los indicadores de censura son observados completamente

  - Estimadores presuavizados de Kaplan-Meier con la causa de censura desconocida aleatoriamente, calculados utilizando el método plug-in, que consiste en representar una función desconocida a partir de funciones estimables (estimadores propuestos)

    - $\hat{S}_n^P(t)$ estimador que se reduce al estimador presuavizado de Kaplan-Meier cuando los indicadores de censura son observados completamente.

- $\hat{S}_n^{P*}(t)$ estimador que se reduce al estimador de Kaplan-Meier cuando los indicadores de censura son observados completamente.

6.2. Simulaciones

- Se realiza un estudio de simulación para comparar los resultados de los estimadores propuestos y de los propuestos por Wang et al.(2008) con el estimador de Kaplan-Meier, bajo MAR (Missing At Random).

- Se calculan los estimadores propuestos a partir del estimador de Nadaraya-Watson, tomando el núcleo gaussiano y calculando para cada uno, el parámetro de suavización mediante la validación cruzada.

  - Entre los estimadores $\hat{S}_n^P$ y $\hat{S}_{n,W}$, que se reducen al estimador presuavizado de Kaplan-Meier cuando los indicadores de censura son observados completamente, se tiene que $\hat{S}_n^P(t)$ se comporta mejor que el estimador $\hat{S}_{n,W}(t)$ propuesto por Wang et al. (2008), obteniendo distancias de Kolmogorov-Smirnov y valores de MISE más pequeños cuando la censura es de 40 % y 70 %. Aunque el estimador de Wang et al (2008) da mejores resultados que el estimador propuesto cuando la censura es del 20 %.

  - Entre los estimadores $\hat{S}_n^{P*}$, $\hat{S}_{n,I}$ y $\tilde{S}_{N,I}$, que se reducen al estimador de Kaplan-Meier cuando los indicadores de censura son observados completamente, se tiene que $\hat{S}_n^{P*}(t)$ se comporta mejor que los estimadores $\hat{S}_{n,I}(t)$ y $\tilde{S}_{N,I}(t)$ propuestos por Wang et al. (2008), obteniendo distancias de Kolmogorov-Smirnov y valores de MISE más pequeños.

6.3. Aplicación a datos reales

- Entre los estimadores propuestos en este trabajo para estos datos de cáncer colorrectal con una tasa de perdidos de 0,035, parece comportarse mejor el estimador $\hat{S}_n^{P*}(t)$.

- Entre los estimadores propuestos en este trabajo para estos datos de cáncer colorrectal con una tasa de perdidos de 0,42, parece comportarse mejor el estimador $\hat{S}_n^{P*}(t)$.

- Para distintas tasas de perdidos los estimadores propuestos obtienen buenos resultados, siendo mejores los de $\hat{S}_n^{P*}(t)$ que los de $\hat{S}_n^{P}(t)$ y comportándose de una forma más próxima al estimador de Kaplan-Meier cuando la tasa de perdidos es pequeña.

Bibliografía

Andersen PK, Borgan Ø, Gill RD & Keiging N (1993) *Statistical Models Based on Counting Processes*. N.Y.: Springer-Verlag

Bernal-Perez M, Gomez-Bernal FJ & Gomez-Bernal GJ (2001) *Tiempos de demora en el diagnóstico del cáncer* Atención Primaria, 27(2). 79-85 págs

Berrino F, De-Angelis R, Sant M, Rosso S, Lasota MB, Coebergh JW, Santaquilani M & the EUROCARE working group (2007) *Survival for eight major cancers and all cancer combined for European adults diagnosed in 1995-99: results of the EUROCARE-4 study*, Lancet Oncol. 8(9):773-83 págs

Cao R, López-de-Ullibarri I, Janssen P & Veraverbeke N (2005) Presmoothed Kaplan-Meier and Nelson-Aalen estimators Journal of Nonparametric Statistics. 17(1): 31-56 págs

Coleman MP, Quaresma M, Berrino F, Lutz JM, De Angelis R & Capocaccia R (2008) Cancer survival in five continents: a worldwide population-based study (CONCORD) Lancet Oncol. 9(8): 730-53 págs

Cox DR & Oakes D (1984) Analysis of Survival Data Chapman Hall.

Dikta, D (1998) On semiparametric random censorship models J. Statist. Plann. Inference 66, 253-279 págs

Gill R (1983) *Large sample behaviour of the product-limit estimators on the whole line* The Annals of Statistics. 11(1),49-58 págs

Härdle W, Müller, Sperlich S & Werwatz A. (2004). *Non- and Semiparametric Models* Springer Series in Statistics: Berlin

Kaplan EL & Meier P (1958) *Nonparametric estimation form incomplete observations* J Am Stat Assoc. 53: 457-481 págs

Klein JP (1991). Small-sample moments of some estimators of the variance of the Kaplan-Meier and Nelson-Aalen estimators. Scandinavian Journal of Statistics,18, 333-340 págs

Little RJ A & Rubin DB(1987) *Statistical Analysis with Missing Data* John Wiley & Sons, New York

Macdonald S, Macleod U, Campbell NC, Weller D & Mitchell E (2006) *Systematic review of factors influencing patient and practioner delay in diagnosis of upper gastrointestinal cancer* Br J Cancer. 94 (9):1272-80 págs

Maglinte DD, O'Connor K, Bessette J, Chernish SM & Kelvin F.M. (1991) *The role of the physician in the late diagnosis of primary malignant tumors of the small intestine* Am J Gastroenterol 86(3): 304-8 págs

Mariscal M, Llorca J, Prieto D & Delgado-Rodriguez M (1996) *Determinants of the interval between the onset of symptoms and first medical visit in patients with digestive tract cancer* Int J Oncol 8: 941-949 págs

Pita Fernández S (1995) *Análisis de supervivencia* Cad Aten Primaria; 2: 130-135 págs

Robins J & Rotnitzky A (1992) *Recovery of information and adjustment for dependent censoring using surrogate markers.* AIDS Epidemiology-Methodological Issues. (Edited by Jewell N, Dietz K & Farewell V), 297-334 págs. Birlzhauser, Boston

Stute W & Wang JL (1993) *The strong law under random censorship* The Annals of Statistics 21(3), 1591-1607 págs

Stute W (1995) *The central limit theorem under random censorship* Annals Statiscs 23 (2), 422-439 págs

Verdecchia A, Francisi S, Brenner H, Gatta G, Micheli A & Mangore L (2007). Recent survival in Europe a 2000-02 period analysis of EUROCARE-4 data. Lancet Oncol. 8(9)784-96 págs

Wang Q & Ng KW (2008) *Asymptotically efficient product-limit estimators with censoring indicators missing at random* Statistica Sinica 18, 749-768 págs

Verdecchia A Francisi S, Brenner H, Gatta G, Micheli A & Mangore L (2007) *Recent survival in Europe a 2000-02 period analysis of EUROCARE-4 data* Lancet Oncol, 8(9):748-96 págs

8 Anexos

8.1. Autorización del Comité de Ética de Investigación de Galicia

Conselleria de Sanidade
SERGAS
División de Formación e Productos Sanitarios

Edificio Administrativo San Lázaro
15703 Santiago de Compostela
Teléfono (981) 54 28 32 - Fax (981) 54 18 04

www.sergas.es

XACOBEO 2004
Galicia

Comité Ético de Investigación Clínica
Edificio Administrativo San Lázaro, s/n
15780 Santiago de Compostela
Telf. (981) 54 64 25 — Fax (981) 54 18 04
E-mail: ceic@sergas.es

Informe del Comité Ético de Investigación Clínica de Galicia

D. Miguel Amor Otero, Secretario del Comité Ético de Investigación Clínica de Galicia

CERTIFICA:

Que este Comité ha evaluado en su reunión del 29 de julio de 2004 la propuesta del Salvador Pita Fernández para que se realice el estudio titulado "*Demora diagnóstica en el cáncer de colon y recto*", con nuestro número de registro: ***2004/159***, y considera que:

Se cumplen los requisitos éticos aplicables a este tipo de estudios, están justificados los riesgos y molestias previsibles para el sujeto y es adecuado el procedimiento para obtener el consentimiento informado.

Y que este Comité acepta, de conformidad con sus Procedimientos Normalizados de Trabajo, que dicho estudio sea realizado en el Centro/s C.H. Juan Canalejo por Salvador Pieta Fernández como investigador/es principal/es.

Lo que firmo en Santiago de Compostela a 30 de julio de 2004.

Firmado:
Miguel Amor Otero

NOTA genérica: Debido a las connotaciones éticas y la especial naturaleza del consentimiento informado, es exigible que, con anterioridad al reclutamiento de pacientes, esté disponible una versión fidedigna y redactada en gallego normativo del mismo (hojas de información y de firmas). Garantizándose así, el derecho del paciente al acceso a la información en los idiomas oficiales de Galicia y la completa comprensión del consentimiento informado.

Figura 8.1: Autorización del Comité de Ética de Investigación de Galicia (CEIC Galicia)