Trabajo Fin de Máster

La correlación de distancias en el Análisis de Supervivencia

María Vidal García

Máster en Técnicas Estadísticas

Curso 2021-2022

Propuesta de Trabajo Fin de Máster

Título en galego: A correlación de distancias na Análise de Supervivencia

Título en español: La correlación de distancias en el Análisis de Supervivencia

English title: Correlation distance in Survival Analysis

Modalidad: Modalidad A

Autor/a: María Vidal García, Universidade de Santiago de Compostela

Director/a: Wenceslao González Manteiga, Universidade de Santiago de Compostela; Manuel Febrero Bande, Universidade de Santiago de Compostela

Breve resumen del trabajo:

El objetivo del trabajo es estudiar los aspectos teóricos detrás de la aplicación de la metodología de correlación de distancias al contexto del Análisis de Supervivencia. Para ello se desarrollan los siguientes apartados:

1. Introducción de los conceptos fundamentales de Análisis de Supervivencia y las técnicas básicas para la estimación de la distribución y de distribución condicionada en presencia de covariables.
2. Revisión de los métodos clásicos para determinar la correlación lineal: el coeficiente de correlación de Pearson y sus limitaciones. Presentación de los resultados fundamentales que sustentan la metodología de correlación de distancias y su aplicación en la caracterización de independencia para datos completos.
3. Adaptación de las técnicas basadas en correlación de distancias al contexto de datos con censura aleatoria por la derecha.
4. Ejemplo de aplicación de los métodos presentados sobre muestras simuladas.

Concluimos revisando posibles líneas de trabajo que abiertas que sería interesante tratar en el futuro.

Don/doña Wenceslao González Manteiga, Catedrático del área de Estadística e Investigación Operativa de la Universidade de Santiago de Compostela, don/doña Manuel Febrero Bande, Catedrático del área de Estadística e Investigación Operativa de la Universidade de Santiago de Compostela, informan que el Trabajo Fin de Máster titulado

La correlación de distancias en el Análisis de Supervivencia

fue realizado bajo su dirección por don/doña María Vidal García para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Santiago, a 2 de febrero de 2022.

El/la director/a:

GONZALEZ MANTEIGA WENCESLAO - 32618861P

Firmado digitalmente por GONZALEZ MANTEIGA WENCESLAO - 32618861P
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-32618861P, givenName=WENCESLAO, sn=GONZALEZ MANTEIGA, cn=GONZALEZ MANTEIGA WENCESLAO - 32618861P
Fecha: 2022.02.02 19:38:37 +01'00'

Don/doña Wenceslao González Manteiga

El/la director/a:

FEBRERO BANDE MANUEL - 34961128R

Firmado digitalmente por FEBRERO BANDE MANUEL - 34961128R
Fecha: 2022.02.02 19:46:23 +01'00'

Don/doña Manuel Febrero Bande

El/la autor/a:

VIDAL GARCIA, MARIA (AUTENTICACIÓN)

Firmado digitalmente por VIDAL GARCIA, MARIA (AUTENTICACIÓN)
Fecha: 2022.02.02 19:19:02 +01'00'

Don/doña María Vidal García

Agradecimientos

Me gustaría agradecer el apoyo a toda mi familia y amigos, especialmente a mis padres, Antonio y María José, a mi hermana Antía y a Gonzalo.

Índice general

Resumen

Resumen en español

El objetivo de este trabajo es estudiar la aplicación de la metodología basada en la correlación de distancias a la hora de caracterizar la independencia entre variables en el contexto del Análisis de Supervivencia.

Comenzamos introduciendo los conceptos más relevantes en Análisis de Supervivencia y revisando los principales métodos de estimación de la función de distribución y la función de distribución condicionada en presencia de covariables.

A continuación planteamos el problema más sencillo de detección de correlación y estudiamos las propiedades del coeficiente de correlación lineal de Pearson. Una vez establecidas las limitaciones de las técnicas clásicas a la hora de identificar relaciones generales de dependencia, introducimos la correlación de distancias y sus principales resultados como solución a dicha problemática. Presentaremos distintos estimadores de dicha magnitud que nos permitirán definir estadísticos de contraste para testar la existencia de cualquier tipo de dependencia.

Finalmente extendemos las técnicas y tests anteriores al contexto de datos censurados por la derecha e ilustramos su implementación de las herramientas presentadas sobre datos simulados.

English abstract

This work aims to study the application of distance correlation methodology to characterize independence in Survival Analysis.

We start by introducing basic concepts of Survival Analysis. Then, we review the most common methods to estimate the distribution function and conditional distribution function when covariates are involved.

Next, we pose the problem of correlation detection and study the properties of Pearson's linear correlation coefficient. Once we have established the limitations of classic techniques to identify the existence of general dependency, as a solution we introduce correlation distance and related results. We pay special attention to its connections with U-statistics theory. Besides its theoretical properties, we explore the construction of general independence tests based on different empirical estimators of correlation distance.

Finally, we study how to extend previously presented techniques and tests to right censored data. To illustrate these methods we simulate several samples and apply all the tools presented throughout the text.

Prefacio

El problema de caracterización de independencia es crucial a la hora de afrontar el análisis de datos en presencia de covariables. Poder confirmar la independencia entre dos variables a menudo nos permite simplificar los modelos empleados o utilizar técnicas que requieren de esta condición.

La solución al problema de caracterización de independencia se encuentra en la metodología de correlación distancias. Esta propuesta, planteada por Gábor J. Székely en 2005, aparece definida por primera vez en Székely y col. (2007) y se encuentra actualmente en plena expansión tanto a nivel teórico como aplicado. En particular, vamos a centrarnos en las recientes propuestas de aplicación de estas técnicas al Análisis de Supervivencia.

En el ámbito del Análisis de Supervivencia, el objeto de estudio suele ser el tiempo hasta un determinado evento de interés. El planteamiento más habitual en presencia de covariables es tratar de determinar si dichas variables o factores de riesgo influyen en cuan pronto observamos dicho evento. La particularidad de esta disciplina es que lo más común es tener limitaciones en la observación del tiempo hasta el evento en la muestra, siendo la más habitual la censura aleatoria por la derecha. El reto a la hora de adaptar la novedosa metodología de correlación de distancias será, por tanto, conseguir trabajar sobre la muestra de observaciones parciales utilizando la mayor cantidad de información posible.

El Análisis de Supervivencia cuenta con numerosas aplicaciones a distintas áreas que se podrían ver beneficiadas por el desarrollo de herramientas para la identificación de independencia. Entre ellas podemos citar investigaciones y seguimientos del ámbito biomédico, análisis de datos genéticos, estudios de riesgo financiero, macroeconómicos o de fiabilidad entre muchos otros. Como ejemplos donde ya se han comenzado a implementar este tipo de soluciones podemos ver Kong y col. (2012), Hua y col. (2015), Li y col. (2012) o Edelmann y col. (2021) entre otros.

Así pues, pasamos a presentar el contenido del trabajo que está organizado como sigue:

- Capítulo 1: Introducción al Análisis de Supervivencia

  En este capítulo se presentan las nociones fundamentales de Análisis de Supervivencia. Comenzamos explicando sus principales elementos de interés y las situaciones de censura y truncamiento. A continuación presentamos distintos métodos de estimación de la función de distribución basados en máxima verosimilitud en el contexto de censura aleatoria por la derecha: los métodos paramétricos (revisando algunas distribuciones notables) y los no paramétricos (Kaplan-Meier y Nelson-Aalen). Finalmente introducimos la presencia de covariables y revisamos los principales métodos paramétricos, semiparamétricos y no paramétricos de estimación de la función de supervivencia condicionada.

- Capítulo 2: Correlación de distancias

  Comenzamos este capítulo revisando el caso de la dependencia lineal y las bondades del coeficiente de correlación de Pearson. A continuación introducimos el contraste de independencia en el caso general y explicamos como se construyen la covarianza de distancias y la correlación de distancias. Tras definir dichas magnitudes, presentamos los estadísticos para estimarlas: vemos primero la propuesta inicial de Székely y col. (2007) para a continuación explorar la versión insesgada del estimador. Esto nos conecta con la teoría de los $\mathcal{U}$-estadísticos que será la que nos permita

adaptar estas técnicas al contexto de censura aleatoria por la derecha como veremos en el capítulo siguiente.

- Capítulo 3: Correlación de distancias en presencia de censura

  En este apartado exploraremos la propuesta de Edelmann y col. (2021) para extender los métodos presentados en el capítulo anterior al caso de censura aleatoria por la derecha. Utiliza el planteamiento de los estadísticos IPCW, que nos permite tomar un $\mathcal{U}$-estadístico ordinario y adaptarlo al caso de censura aleatoria por la derecha. Aplicaremos este razonamiento sobre la versión insesgada de estimador de la correlación de distancias vista en el capítulo anterior para obtener un test de independencia aplicable sobre muestras censuradas por la derecha.

- Capítulo 4: Simulaciones

  Aplicamos las técnicas anteriores sobre conjuntos de datos simulados para ilustrar su implementación explicando como interpretar y utilizar cada propuesta.

Para finalizar en el Capítulo 5 planteamos algunas líneas de trabajo abiertas que esperamos poder abordar en el futuro.

Capítulo 1

Introducción al Análisis de Supervivencia

El Análisis de Supervivencia es una rama de la Estadística versada en estudiar y hacer inferencia sobre el tiempo Y transcurrido hasta que sucede un determinado evento de interés. Dependiendo de la naturaleza de este evento nos encontramos con aplicaciones a distintos ámbitos: el evento puede ser el alta o el fallecimiento de un paciente en estudios biomédicos, el impago de un crédito en trabajos económicos y financieros o el fallo de una máquina en cuestiones relacionadas con la ingeniería. Como muestra de la diversidad de aplicaciones del Análisis de Supervivencia puede consultarse Ferger y col. (2017), donde se revisan aplicaciones al campos de las finanzas, López-Cheda y col. (2018) donde se presentan extensiones de los modelos clásicos para tratar datos de cáncer de colon o López Montoya (2011), donde se hace una breve revisión del uso del Análisis de Supervivencia para predecir la evolución de infraestructuras como las redes de suministro de agua.

Debido a la multitud de aplicaciones en ámbitos tan dispares, la variable Y recibe varios nombres siendo los más comunes *tiempo de fallo* o *tiempo de supervivencia*.

Es evidente que la variable tiempo de fallo Y tal y como la hemos descrito será necesariamente no negativa $Y \geq 0$. En general su distribución puede ser tanto continua como discreta (puede interesarnos saber cuantos días transcurren hasta que se da el evento por ejemplo). En este trabajo nos centraremos únicamente en el caso en que la distribución de Y es absolutamente continua.

Para estudiar el comportamiento de la variable de interés Y recurriremos, como es habitual, a su observación en una muestra aleatoria de la población que queremos estudiar. La principal particularidad del Análisis de Supervivencia, y lo que marca la diferencia con respecto a otras áreas de la Estadística, es que la variable de interés Y no es siempre plenamente observable en la muestra. Para algunos individuos es posible que solo logremos obtener observaciones parciales de Y, con lo que las técnicas que utilicemos para tratar los datos obtenidos deben ser capaces de tener en cuenta esta limitación e integrar la información obtenida en estos casos con el resto de información extraída de las observaciones completas. En función de como sean las restricciones a la hora de observar la variable hablaremos de censura, truncamiento o bien una combinación de ambos.

El objetivo de esta disciplina es desarrollar herramientas análogas a las que se utilizan en el estudio de variables completas (no censuradas o truncadas) que nos permitan estimar el comportamiento del tiempo de fallo Y y realizar regresión e inferencia sobre ella.

Comenzaremos presentando las principales funciones de interés en Análisis de Supervivencia utilizando por simplicidad datos completos. A continuación se describirán las situaciones de censura y truncamiento detallando sus peculiaridades para pasar, en la siguiente sección, a abordar el desarrollo de técnicas específicas para el Análisis de Supervivencia en el contexto de censura aleatoria por la derecha. Los objetivos de estas herramientas son estimar la distribución del tiempo de fallo, comparar distribuciones entre distintos grupos y evaluar el efecto de covariables sobre el tiempo de supervivencia.

En este caso nos interesa especialmente este último aspecto.

Estudiaremos los principales métodos de estimación de la distribución basados en maximizar la verosimilitud aplicando tanto hipótesis paramétricas (veremos las distribuciones notables más empleadas en la literatura) como no paramétricas (estimadores de Kaplan-Meier y Breslow). Finalmente trataremos el estudio del tiempo de fallo Y en presencia de covariables, es decir, las técnicas de regresión con datos censurados. Explicaremos como extender los modelos paramétricos a este contexto, presentaremos los modelos semiparamétricos más utilizados (tiempo de fallo acelerado y riesgos proporcionales de Cox) y modelos no paramétricos como el estimador de Kaplan-Meier adaptado a la presencia de covariables.

1.1. Funciones de interés

Es bien conocido que cualquier variable aleatoria, en este caso el tiempo de fallo Y, está unívocamente determinada por su función de probabilidad acumulada o función de distribución F:

$$F(y) = \mathbb{P}(Y \leq y). \tag{1.1}$$

En Análisis de Supervivencia a menudo nos interesará conocer la probabilidad de que la variable supere un cierto valor (por ejemplo la posibilidad de que un individuo sobreviva más de dos meses después de una operación, o que una pieza tarde más de dos años en averiarse). Por este motivo se suele trabajar directamente con la función opuesta a la de distribución, la función de supervivencia

$$S(y) = \mathbb{P}(Y > y) = 1 - F(y), \tag{1.2}$$

que a la vista de su definición está claro que también determina unívocamente el comportamiento de la variable.

Las propiedades de esta función también son sencillas de deducir a partir de las de la función de distribución F y teniendo en cuenta que $Y \geq 0$:

- $S(0) = 1$ y $\lim_{y\to\infty} S(y) = 0$.
- S monótona decreciente.
- S continua por la derecha.

Las siguientes definiciones son válidas tanto en el caso de tiempos de fallo continuos como discretos. Para el desarrollo ulterior únicamente nos centraremos en los casos en los que la distribución es absolutamente continua y por ello únicamente presentaremos las fórmulas para este caso[1].

En el caso de las variables continuas, es bien conocido que además de con la función de distribución se suele trabajar con su derivada, la función de densidad[2], que describe la probabilidad infinitesimal de fallo en cada punto del dominio:

$$\begin{aligned} f(y) = \frac{dF(y)}{dy} &= \lim_{h\to 0} \frac{F(y+h) - F(y)}{h} = \lim_{h\to 0} \frac{\mathbb{P}(Y \leq y+h) - \mathbb{P}(Y \leq y)}{h} \\ &= \lim_{h\to 0} \frac{\mathbb{P}(y \leq Y \leq y+h)}{h}. \end{aligned} \tag{1.3}$$

Esta magnitud se interpreta como la probabilidad instantánea de que el fallo se produzca en el instante y.

Para aligerar la notación a lo largo de este trabajo identificaremos la probabilidad instantánea anterior con

$$f(y) \equiv \mathbb{P}(Y = y) \tag{1.4}$$

[1] En el caso de variables discretas el papel de la función de densidad lo asume la función de masa de probabilidad, para más información se puede consultar Iglesias Pérez (2021).

[2] La función de densidad $f = F'$ está bien definida únicamente si la variable es absolutamente continua.

donde claramente estamos abusando de la notación pues al tratarse de una variable continua $\mathbb{P}(Y = y) = 0$.

De la definición de función de supervivencia (1.2) se sigue inmediatamente que:

$$f(y) = -\frac{dS(y)}{dy}. \tag{1.5}$$

Supongamos ahora que nos estamos centrando en un individuo particular de la muestra. Sería muy natural que, en cualquier momento del estudio, nos interesase conocer la probabilidad de fallo para ese individuo en ese instante. Al contrario de lo que pueda parecer, esta información no nos la proporciona la probabilidad de fallo instantáneo pues en ese caso no se tiene en cuenta que el individuo ya ha sobrevivido sin experimentar el fallo hasta el instante en cuestión y. La función que nos permite conocer esta información para cada instante es la función de riesgo o razón de fallo que se define como:

$$\lambda(y) = \lim_{h\to 0} \frac{\mathbb{P}(y \leq Y \leq y + h | Y \geq y)}{h} \tag{1.6}$$

y proporciona la probabilidad infinitesimal de fallo en un momento $y + h$ habiendo sobrevivido hasta y.

Como para toda magnitud definida instantáneamente, se puede definir también la función de riesgo acumulado o razón de fallo acumulada que no es más que la acumulación de la probabilidad anterior hasta el instante de interés:

$$\Lambda(y) = \int_0^y \lambda(u)du. \tag{1.7}$$

Existen importantes relaciones entre estas funciones que acabamos de definir. Algunas son obvias sin más que observar la definición y otras, bastante directas también, las detallamos a continuación.

Aplicando la definición de probabilidad condicionada a la definición de función de riesgo y teniendo en cuenta que Y sigue una distribución absolutamente continua:

$$\begin{aligned} \lambda(y) &= \lim_{h\to 0} \frac{\mathbb{P}(y \leq Y \leq y + h | Y \geq y)}{h} = \lim_{h\to 0} \frac{\mathbb{P}(y \leq Y \leq y + h, Y \geq y)}{h \cdot \mathbb{P}(Y \geq y)} \\ &= \frac{1}{\mathbb{P}(Y \geq y)} \lim_{h\to 0} \frac{\mathbb{P}(y \leq Y \leq y + h)}{h} = \frac{f(y)}{S(y^-)} = \frac{f(y)}{S(y)}. \end{aligned} \tag{1.8}$$

De esta igualdad se siguen fácilmente las siguientes sin más que aplicar la relación vista en (1.5):

$$\Lambda(y) = -ln(S(y)), \tag{1.9}$$

$$S(y) = e^{-\Lambda(y)} = e^{-\int_0^y \lambda(s)ds}. \tag{1.10}$$

1.2. Censura y truncamiento

Tal y como se ha comentado en la introducción del capítulo, una situación muy frecuente en Análisis de Supervivencia es aquella en la que existen problemas a la hora de observar la variable tiempo de fallo o tiempo de supervivencia Y en la muestra. Estos problemas pueden ser fundamentalmente de dos tipos: censura o truncamiento.

Censura

Se habla de censura cuando de algunas observaciones solo se sabe que han ocurrido dentro de un intervalo de tiempo determinado mientras que las demás son observadas con exactitud. Este tipo de restricciones en la información accesible está muy relacionado con el diseño del experimento y la recogida de datos. Veamos qué tipos de censura hay y como se formalizan de acuerdo con Klein y Moeschberger (2003).

En general la presencia de censura consiste en la existencia de una variable C que interfiere en la observación de Y en la muestra. La variable C no es más que el tiempo que transcurre hasta que se produce un evento denominado censura que compite con el evento de interés de manera que para cada individuo o bien observamos C y ya no podemos observar Y (observación censurada) o bien observamos Y y ya no podemos observar C (observación no censurada). Al tiempo $C \geq 0$ lo llamaremos tiempo de censura.

Considérese una muestra con n individuos en los que el objetivo es observar las realizaciones de la variable de interés Y. Un tipo de censura muy común, especialmente en trabajos con datos longitudinales como puede ser el seguimiento de pacientes, es la censura por la derecha. En esos casos solo es posible observar el tiempo hasta el evento que se produzca primero: si se produce primero la censura se habla de observación censurada y si se produce primero el evento de interés, de observación no censurada. Las variables observables serían por tanto las siguientes: el tiempo observado

$$T = \text{mín}\{C, Y\} \geq 0, \tag{1.11}$$

y el indicador de no censura, que informa de si la observación es censurada o no

$$\delta = \mathbb{I}(Y \leq C) = \begin{cases} 0 & \text{si la observación está censurada} \\ \\ 1 & \text{si se observa el tiempo de fallo.} \end{cases} \tag{1.12}$$

Por tanto las observaciones que obtendríamos sobre una muestra de tamaño n serían los n pares:

$$\{(T_i, \delta_i)\}_{i=1}^n \tag{1.13}$$

En función de como sea la variable censora C podemos diferenciar distintos tipos de censura por la derecha: se habla de censura tipo I si el tiempo de censura C se corresponde con un valor constante prefijado. Variantes de censura tipo I son aquellas en las que se fijan distintas constantes para distintos grupos de individuos en la muestra (tipo I progresivo) o, incluso, podría darse el caso en el que cada individuo tenga su propio tiempo de censura (tipo I generalizado).

Por su parte, en el caso de censura tipo II el estudio continúa hasta que r individuos experimentan el fallo, siendo $r < n$ un entero fijado de antemano. El tiempo de censura no toma un valor fijo que podamos conocer previamente, sino que depende de la muestra en cuestión. Nuevamente podríamos fijar distintos r_i para varios subgrupos de la muestra y hablaríamos entonces de censura tipo II progresiva.

El caso más general es el caso de censura aleatoria por la derecha. En él se supone que los individuos de la muestra experimentan eventos competitivos Y (evento de interés) y C (censura, cualquier evento que nos impida observar el tiempo de fallo, puede ser de diversa naturaleza), cada uno de los cuáles con su propio comportamiento desconocido. Es crucial en este caso suponer que la censura es no informativa o, dicho de otro modo, que las distribuciones de Y y C son independientes.

Denotaremos por G a la función de distribución de la variable tiempo de censura C, por $S_G = 1 - G$ a la correspondiente función de supervivencia, por $g = G'$ a la función de densidad, y por H a la función de distribución conjunta de Y y C. Entonces la condición de censura independiente se traduce en que

$$H(y, c) = \mathbb{P}(Y \leq y, C \leq c) = \mathbb{P}(Y \leq y)\mathbb{P}(C \leq c) = F(y) \cdot G(c). \tag{1.14}$$

Esta condición es fundamental para todo el desarrollo que presentaremos a continuación. En los casos de censura informativa, la función de verosimilitud cambia y necesitaríamos técnicas más complejas para analizar los datos.

Es muy habitual encontrar simultáneamente censura tipo I y censura aleatoria. Por ejemplo, un estudio sobre el tiempo hasta la muerte por un tipo específico de cáncer posiblemente contemple censura administrativa (tipo I, el tiempo de seguimiento es limitado y no podremos observar eventos

posteriores a la fecha de finalización del estudio) y censura aleatoria (muerte por otras causas, pérdida de seguimiento o *dropout*...). En este caso sería importante comprobar que, por ejemplo, la censura por *dropout* es en efecto independiente del suceso de interés. Podría ser que la evolución acelerada de la enfermedad dificultase la permanencia en el estudio y aumentase la probabilidad de censura por *dropout* al mismo tiempo que las probabilidades de que se experimente el suceso. En ese caso sería importante tener presente esta asociación a la hora de trabajar con los datos.

De forma análoga a la definición de censura por la derecha se define también la censura por la izquierda. En este caso solo es observable el último evento, es decir, observaremos un tiempo de fallo solo cuando el evento de interés se produce después de la censura $Y \geq C$ y viceversa. La información observada puede recogerse nuevamente en forma de dos variables, tiempo observado e indicadora de la censura, definidas como sigue:

$$\begin{aligned} T &= \text{máx}\{Y, C\}, \\ \epsilon &= \mathbb{I}(Y \geq C) = \begin{cases} 0 & \text{si hay censura} \\ \\ 1 & \text{en caso contrario.} \end{cases} \end{aligned} \tag{1.15}$$

Por tanto la información relevante obtenida de una muestra sería $\{(T_i, \epsilon_i)\}_{i=1}^n$.

En caso de que se produzca simultáneamente censura por la derecha y por la izquierda, sigue siendo posible sintetizar la información observada a través de un par de variables (T, χ). Se habla entonces de censura doble. Si denotamos por C_i^R a la variable tiempo de censura por la derecha y C_i^L a la variable tiempo de censura por la izquierda, las variables tiempo observado e indicadora de censura se podrían definir como sigue:

$$\begin{aligned} T &= \text{máx}\{\text{mín}\{Y, C^R\}, C^L\}, \\ \chi &= \begin{cases} 1 & si \quad C^L \leq T \leq C^R \text{ no hay censura,} \\ \\ 0 & si \quad T > C^R \text{ censura por la derecha,} \\ \\ -1 & si \quad T < C^L \text{ censura por la izquierda.} \end{cases} \end{aligned} \tag{1.16}$$

La información obtenida sobre una muestra con n individuos se resumiría en $\{(T_i, \chi_i)\}_{i=1}^n$.

Por último, cabe notar que tanto la censura por la derecha como por la izquierda son casos particulares de censura por intervalo, situación en la que podemos observar el tiempo de fallo exacto de algunos los individuos mientras que para otros solo sabemos que el evento sucede en un intervalo específico. La censura por la derecha e izquierda corresponden a los casos degenerados en los que dicho intervalo es $(-\infty, C^L)$ y (C^R, ∞) respectivamente.

A lo largo de todo este documento nos limitaremos a tratar el caso de censura no informativa, es decir, independiente. En caso de que las distribuciones de C e Y estuviesen relacionadas se introduciría un sesgo en las estimaciones. Habría que desarrollar herramientas específicas para corregir este problema, no ahondaremos en esta cuestión, para más información sobre censura dependendiente se puede consultar Klein y Moeschberger (2003) o Lawless (2011).

A la hora de estudiar como se extienden las herramientas basadas en correlación de distancia a Análisis de Supervivencia en los capítulos siguientes, vamos a restringirnos a la situación de censura

aleatoria por la derecha independiente o no informativa. La censura tipo I estaría incluida como caso particular.

Truncamiento

Hablamos de truncamiento cuando solo son observables aquellos individuos en los que se cumple una determinada condición. Para los restantes no disponemos de información de ningún tipo, ni siquiera podemos constatar su existencia (a diferencia de lo que sucede en caso de censura, donde sí disponemos de información al menos parcial acerca de todos los individuos de la muestra).

Se habla de truncamiento por la izquierda cuando la condición para que un individuo sea observado debe ocurrir antes del evento de interés. Es el tipo más común de truncamiento. Por ejemplo, en un estudio sobre enfermedades neurodegenerativas en los que el evento de interés sea la muerte puede suceder que solo se incluyan aquellos individuos que den muestras de padecer este tipo de problemas desde que se inicia el estudio. Todos los individuos que hayan sucumbido a la enfermedad antes no serán incluidos.

En presencia de truncamiento por la derecha, por su parte, solo serán observables individuos para los que haya tenido lugar el suceso. Por ejemplo en el estudio de enfermedades neurodegenerativas en el que el evento de interés es la muerte, el diagnóstico para confirmar la enfermedad que padece cada individuo de la muestra solo puede ser obtenido mediante la autopsia, así que solo serán incluidos aquellos individuos para los que el fallo ya haya ocurrido.

Es evidente que ambas situaciones pueden darse simultáneamente y se habla entonces de doble truncamiento. Este sería el caso de los estudios de enfermedades que requieren confirmación por autopsia, como sucede en el trabajo sobre el Alzheimer de Rennert (2018) de donde se han extraído los ejemplos anteriores, o de estudios en los que el intervalo de incorporación de individuos depende de características individuales como puede ser el caso de Moreira y Uña-Álvarez (2010) sobre la edad de diagnóstico de cáncer infantil en el norte de Portugal.

1.3. Estimación de la distribución

A continuación estudiaremos distintos estimadores de la función de distribución. Todos ellos pueden obtenerse utilizando métodos de máxima verosimilitud, algunos de ellos imponiendo hipótesis paramétricas y otros serán estimadores no paramétricos. Comentaremos las principales características de ambos tipos.

En cada caso particularizaremos los estimadores para el que será nuestro contexto de interés a la hora de extender las herramientas de correlacion de distancias: la presencia de censura aleatoria por la derecha.

Las hipótesis de partida que asumiremos en todo momento son:

1. La censura es independiente (no informativa), es decir, las distribuciones de Y y C cumplen la condición (1.14) o equivalentemente:

$$H(y,c) = \mathbb{P}(Y \leq y, C \leq c) = \mathbb{P}(Y \leq y)\mathbb{P}(C \leq c) = (1 - S(y))(1 - S_G(c)). \tag{1.17}$$

2. Las distribuciones F del tiempo de fallo y G de la variable censura son comunes para todos los individuos de la muestra.

3. Las distribuciones F y G son absolutamente continuas.

Comenzaremos definiendo la función de verosimilitud, que es la base de nuestro procedimiento de estimación.

1.3.1. Función de verosimilitud

Dada una muestra de una variable aleatoria de interés, en este caso el tiempo de supervivencia Y, la verosimilitud indica la probabilidad de observar dicha muestra concreta condicionada a la distribución desconocida F. A la hora de estimar F, escogeremos aquella que haga las muestra observada más probable.

Llamamos función de verosimilitud a la probabilidad de observar la muestra condicionada a la función de distribución F:

$$\mathcal{L}_n(F) = \mathcal{L}_n(F|\boldsymbol{y}) = \mathbb{P}(\boldsymbol{Y} = \boldsymbol{y}|F) = \mathbb{P}_F(\boldsymbol{Y} = \boldsymbol{y}), \tag{1.18}$$

donde estamos denotando por $\boldsymbol{Y}$ a una muestra aleatoria simple de la variable en cuestión e $\boldsymbol{y}$ a la realización concreta de dicha muestra que hayamos observado. Si la muestra es independiente e idénticamente distribuida la función de verosimilitud se puede expresar como:

$$\mathcal{L}_n(F) = \prod_{i=1}^{n} \mathbb{P}_F(Y = y_i). \tag{1.19}$$

Lo que nos va a interesar a la hora de estimar la función de distribución es maximizar la función anterior sobre F

$$\hat{F} = \arg\max_{F \in \mathcal{F}} \mathcal{L}_n(F), \tag{1.20}$$

donde $\mathcal{F}$ es en principio todo el conjunto de funciones de distribución. Este conjunto puede reducirse asumiendo hipótesis a priori sobre F. Una forma de hacerlo puede ser asumir una forma paramétrica concreta, como veremos en la Sección 1.3.2, pero también se puede obtener un estimador no paramétrico como haremos en la Sección 1.3.3.

El primer reto con el que nos encontramos es que, a diferencia de cuando tratamos con datos completos, en situaciones de censura o truncamiento no vamos a disponer de una muestra aleatoria simple de observaciones de la variable de interés. Comenzaremos entonces viendo como se particulariza la función de verosimilitud para el caso de censura aleatoria por la derecha.

Censura aleatoria por la derecha

En primer lugar recordemos que, por hipótesis, estamos asumiendo que las variables Y y C siguen sendas distribuciones F y G absolutamente continuas e independientes, y contamos con una muestra de observaciones parciales de Y formada por n observaciones independientes e idénticamente distribuidas del par de variables observadas (T, δ) descritas en (1.12).

Por tanto, dada una muestra concreta $\{(T_i, \delta_i)\}_{i=1}^n$ si adaptamos la ecuación (1.19) a nuestra muestra obtendremos

$$\mathcal{L}_n(F) = \prod_{i=1}^{n} \mathbb{P}_F((T, \delta) = (T_i, \delta_i)) = \prod_{i=1}^{n} \mathbb{P}_F(T = T_i, \delta = \delta_i). \tag{1.21}$$

Ahora bien, cada uno de los factores del producto anterior se puede descomponer teniendo en cuenta que $\delta \in \{0, 1\}$. Entonces:

$$\mathbb{P}_F(T = T_i, \delta = \delta_i) = \mathbb{P}_F(T = T_i, \delta_i = 1)^{\delta_i} \cdot \mathbb{P}_F(T = T_i, \delta_i = 0)^{1-\delta_i} \tag{1.22}$$

Calculamos las probabilidades conjuntas utilizando las definiciones:

$$\mathbb{P}_F(T = T_i, \delta = 1) = \mathbb{P}_F(T = T_i, Y \leq C) = \mathbb{P}_F(Y = T_i, Y \leq C) = \mathbb{P}_F(Y = T_i, T_i \leq C) \tag{1.23}$$

y teniendo en cuenta que por hipótesis de independencia de Y y C tenemos (1.17) nos queda

$$\mathbb{P}_F(Y = T_i, T_i \leq C) = \mathbb{P}_F(Y = T_i)\mathbb{P}_F(T_i \leq C) = f(T_i)(1 - G(T_i)). \tag{1.24}$$

Procediendo del mismo modo obtenemos que:

$$\begin{aligned}\mathbb{P}_F(T=T_i,\delta=0) &= \mathbb{P}_F(T=T_i,Y>C) = \mathbb{P}_F(C=T_i,Y>T_i) = \mathbb{P}_F(C=T_i,Y>T_i)\\ &= \mathbb{P}_F(C=T_i)\mathbb{P}_F(Y>T_i) = g(T_i)(1-F(T_i))\end{aligned} \tag{1.25}$$

Sustituyendo (1.24) y (1.25) en (1.21):

$$\mathcal{L}_n(F) = \prod_{i=1}^{n} \mathbb{P}_F(T=T_i,\delta=\delta_i) = \prod_{i=1}^{n} \left[f(T_i)(1-G(T_i))\right]^{\delta_i} \cdot \prod_{i=1}^{n} \left[g(T_i)(1-F(T_i))\right]^{1-\delta_i}. \tag{1.26}$$

Como lo que nos interesa es estimar la función de distribución de Y, nos bastará trabajar con la parte de la ecuación de verosimilitud que depende de F, ya sea de forma directa o a través de la función de densidad (su derivada $f = F'$). Por tanto vamos a obviar el resto de los términos que no influyen en la estimación y a considerar

$$\mathcal{L}_n(F) = \prod_{i=1}^{n} f(T_i)^{\delta_i} \cdot (1-F(T_i))^{1-\delta_i} \tag{1.27}$$

o equivalentemente, haciendo uso de la hipótesis de continuidad de Y y de la relación (1.8)

$$\mathcal{L}_n(F) = \prod_{i=1}^{n} \lambda(T_i)^{\delta_i} \cdot S(T_i). \tag{1.28}$$

1.3.2. Estimación paramétrica

El enfoque paramétrico se basa en fijar como hipótesis a priori que la función de distribución tiene una forma específica que depende de una cantidad finita de parámetros. Formalmente, se dice que la función de distribución de interés F pertenece a una familia paramétrica

$$Y \sim F \in \{F_\theta\}_{\theta\in\Theta} \tag{1.29}$$

donde la única incógnita es el valor del parámetro o parámetros que denotaremos por $\theta \in \Theta$.

En esta situación, la estimación de la función de distribución por máxima verosimilitud se reduce a estimar el parámetro desconocido con lo que el problema de maximización sobre la función de distribución pasa a ser un problema de maximización sobre el valor del parámetro. En esta situación, denotamos por

$$\mathcal{L}_n(\theta) \equiv \mathcal{L}_n(F_\theta). \tag{1.30}$$

Censura por la derecha

Las funciones de verosimilitud (1.27) y (1.8) se pueden reescribir como:

$$\begin{aligned}\mathcal{L}_n(\theta) &= \prod_{i=1}^{n} f_\theta(t_i)^{\delta_i} \cdot (1-F_\theta(t_i))^{1-\delta_i},\\ \mathcal{L}_n(\theta) &= \prod_{i=1}^{n} \lambda_\theta(t_i)^{\delta_i} \cdot S(t_i)\end{aligned} \tag{1.31}$$

El estimador del parámetro se obtendría del siguiente modo:

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \mathcal{L}_n(\theta). \tag{1.32}$$

Este problema es equivalente a maximizar el logaritmo de $\mathcal{L}_n(\theta)$ lo que frecuentemente simplifica el problema en la práctica.

Este estimador $\hat{\theta}$ es asintóticamente insesgado y consistente bajo hipótesis de regularidad bastante generales. En general las estimaciones obtenidas utilizando modelos paramétricos son más eficientes que las no paramétricas y presentan mejores propiedades cuando los supuestos del modelo se cumplen.

La principal limitación de estos métodos es que imponer una forma paramétrica para la función de distribución es una condición muy exigente que puede no cumplirse en muchos casos. En particular si no hay mucha literatura sobre la variable a estudiar habría que aplicar primero tests de bondad de ajuste que no son sencillos y pueden no funcionar bien. Si se realiza la estimación suponiendo un modelo paramétrico concreto y esta hipótesis resulta ser falsa los resultados no tienen ninguna validez.

Los métodos no paramétricos, en cambio, apenas imponen restricciones a priori sobre la variable de manera que son de aplicación más general. Trabajan casi en exclusiva con la información proporcionada en la muestra lo que los hace menos eficientes que los análogos paramétricos (que tienen más información que utilizar) pero con las mejoras a nivel computacional son plenamente competitivos y preferibles en aquellos casos en los que no tengamos mucha información sobre la variable a estudiar.

Ambos enfoques presentan ventajas y limitaciones, lo importante es saber en qué casos es más adecuado cada uno de ellos y combinar su uso para aprovechas sus respectivos puntos fuertes.

Modelos paramétricos notables

Vamos a hacer una breve reseña de las distribuciones continuas más comunes en Análisis de Supervivencia de acuerdo con la selección realizada en Iglesias Pérez (2021).

Distribución Exponencial

La distribución exponencial de parámetro γ se denota $Y \sim Exp(\gamma)$. Su función de supervivencia viene dada por:

$$S(y) = e^{-\gamma y}, \tag{1.33}$$

y se caracteriza por ser la única distribución cuya función de riesgo es constante

$$\lambda(y) = \gamma. \tag{1.34}$$

Es una distribución absolutamente continua muy sencilla cuya media y varianza vienen dadas por

$$\begin{aligned} \mathbb{E}(Y) &= \frac{1}{\gamma}, \\ Var(Y) &= \frac{1}{\gamma^2}, \end{aligned} \tag{1.35}$$

sin embargo en la práctica es poco habitual encontrar aplicaciones en las que el riesgo se mantenga constante a lo largo del tiempo.

Distribución Weibull

La distribución Weibull $Y \sim Weibull(\alpha, \gamma)$ generaliza a la distribución exponencial. Se caracteriza por el parámetro de forma $\alpha \in \mathbb{R}$ y el parámetro de tasa (rate) $\gamma > 0$. Su función de supervivencia tiene la siguiente forma:

$$S(y) = e^{-(\gamma y)^{\alpha}}. \tag{1.36}$$

Su media y varianza vienen dadas por las siguientes expresiones:

$$\begin{aligned} \mathbb{E}(Y) &= \frac{1}{\gamma}\Gamma\left(1 + \frac{1}{\alpha}\right), \\ Var(Y) &= \frac{1}{\gamma^2}\left[\Gamma\left(1 + \frac{2}{\alpha}\right) - \Gamma^2\left(1 + \frac{1}{\alpha}\right)\right], \end{aligned} \tag{1.37}$$

donde Γ es la función gamma de Euler.

La correspondiente función de riesgo es

$$\lambda(y) = \alpha\gamma^{\alpha}y^{\alpha-1}. \tag{1.38}$$

Esta distribución es muy utilizada porque permite modelizar tanto situaciones en las que el riesgo crece de forma monótona ($\alpha > 1$) como situaciones en las que el riesgo decrece de forma monótona ($\alpha < 1$). El caso $\alpha = 1$ se corresponde con el caso de la exponencial $W(1, \gamma) \equiv Exp(\gamma)$.

Distribución log-normal
A la hora de aplicar métodos estadísticos generales, lo más habitual cuando hay que establecer hipótesis paramétricas es trabajar con la distribución normal. En el caso del Análisis de Supervivencia esto no es adecuado ya que la variable de interés Y es no negativa. Esto nos permite, sin embargo, extender el dominio también a los valores negativos aplicando una transformación logarítmica, obteniendo así una variable que ya podemos modelizar con una distribución normal.

Una variable Y que cumple que

$$\ln(Y) \sim N(\mu, \sigma), \quad \sigma > 0 \tag{1.39}$$

se dice que sigue una distribución log-normal

$$Y \sim LN(\mu, \sigma). \tag{1.40}$$

Sus funciones de distribución, supervivencia y riesgo se obtienen a partir de las de la normal $X \sim N(\mu, \sigma)$ sin más que tener en cuenta que Y se obtiene como transformación monótona de X

$$Y = e^{X}, \quad X \sim N(\mu, \sigma). \tag{1.41}$$

Distribución log-logística
La distribución log-logística obedece a un planteamiento similar a la log-normal solo que en este caso

$$log(Y) \sim L(\mu, \sigma), \tag{1.42}$$

donde $L(\mu, \sigma)$ es la distribución logística con función de supervivencia dada por

$$S(x) = \frac{1}{1 + e^{\frac{x-\mu}{\sigma}}}. \tag{1.43}$$

En tal caso se dice que el tiempo de supervivencia Y sigue una distribución log-logística y se denota por

$$Y \sim LL(\mu, \sigma). \tag{1.44}$$

Modelos de localización escala
Cualquiera de las distribuciones anteriores (exponencial, Weibull, log-logística y log-normal) admiten una expresión equivalente de la siguiente forma:

$$ln(Y) = \mu + \sigma\varepsilon, \tag{1.45}$$

donde μ es el parámetro de localización, σ el parámetro de escala y ε el error aleatorio que sigue una distribución

- Valor extremo (extreme value distribution) en el caso en el que $Y \sim Weibull(\alpha, \gamma)$. La exponencial sería el caso con $\sigma = 1$.
- Normal estándar en el caso de $Y \sim LN(\mu, \sigma)$.
- Logística en el caso en el que $Y \sim LL(\mu, \sigma)$.

Veremos que esta formulación es particularmente útil en el caso de los modelos de tiempo de fallo acelerado (AFT).

1.3.3. Estimación no paramétrica

Los métodos de estimación no paramétricos trabajan con hipótesis muy poco restrictivas y se basan sobre todo en la información proporcionada por la muestra. Esto los convierte en métodos muy generales y flexibles.

Al depender tanto de la muestra sus resultados estarán supeditados al tamaño muestral n, algo que también sucede en los métodos de estimación paramétricos pero se acentúa en este caso. En general, cuanto más grande sea la muestra más fiables serán los resultados obtenidos.

En el caso de la estimación de la función de distribución con datos completos, el estimador no paramétrico por excelencia es la función de distribución empírica, que se construye a partir de una muestra aleatoria simple $\{Y_i\}_{i=1}^n$ de la variable de interés del siguiente modo

$$\hat{F}(t) = \sum_{i=1}^{n} \mathbb{I}(Y_i \leq t) \quad \Leftrightarrow \quad \hat{S}(t) = \sum_{i=1}^{n} \mathbb{I}(Y_i > t), \tag{1.46}$$

es decir, estima la probabilidad de que la variable sea menor que un determinado valor por la proporción de observaciones de la muestra menores que dicho valor.

En presencia de censura o truncamiento este estimador no es consistente ya que ignora el hecho de que las observaciones son parciales o de que están condicionadas al cumplimiento de una determinada condición.

Veremos como se realiza la estimación de la distribución en este contexto. En general en Análisis de Supervivencia estos problemas se plantean como problemas de estimación de la función de supervivencia $S = 1 - F$ ya que su interpretación suele ser más útil en las aplicaciones.

Censura aleatoria por la derecha

En los casos de censura aleatoria por la derecha la función de distribución empírica en general infraestima el tiempo de supervivencia. El estimador no paramétrico por excelencia de la función de supervivencia en estos casos es el estimador de Kaplan-Meier que pasamos a definir.

Estimador de Kaplan-Meier

Consideremos la muestra $\{T_i, \delta_i\}_{i=1}^n$ ordenada en función de los tiempos observados:

$$T_1 \leq T_2 \leq ... \leq T_n.$$

Denotemos por d_i a la cantidad de fallos que tienen lugar en el instante T_i y por n_i a la cantidad de individuos a riesgo en T_i (individuos que no han experimentado fallo ni censura antes de T_i).

El estimador de Kaplan-Meier se define como sigue

$$\hat{S}(t) = \prod_{T_i \leq t} (1 - h_i), \tag{1.47}$$

donde

$$h_i = \frac{d_i}{n_i}, \tag{1.48}$$

y podemos interpretar h_i como la estimación de la razón de fallo $\lambda(T_i)$.

Si consideramos que no existe la posibilidad de empate en los tiempos observados, una suposición muy razonable en nuestro caso[3], entonces $T_1 < T_2 < ... < T_n$ y el estimador anterior se expresaría como sigue

$$\hat{S}(t) = \prod_{T_i \leq t} \left(\frac{n-i}{n-i+1} \right)^{\delta_i}. \tag{1.49}$$

[3] A nivel teórico, por tratarse Y y C de variables absolutamente continuas e independientes, la probabilidad de empate es cero.

El estimador de Kaplan-Meier se puede obtener como estimador de máxima verosimilitud no paramétrico de la función de riesgo, pero también utilizando el método del producto (motivo por el que también se le llama estimador límite-producto), el método de Efron o un método basado en procesos empíricos.

Para ver las demostraciones en detalle se puede recurrir a Kaplan y Meier (1958). Un esbozo de dichas pruebas acompañado de ejemplos se puede encontrar en Iglesias Pérez (2021).

Para entender de manera intuitiva qué es lo que hace este estimador vamos a explicar el razonamiento detrás del método de deducción de Efron. Dada la muestra ordenada en función de T

$$(T_1, \delta_1), (T_2, \delta_2), ..., (T_n, \delta_n) \tag{1.50}$$

partimos de la situación que tendríamos en el caso de datos completos. En ese caso asignaríamos la misma probabilidad $\frac{1}{n}$ a cada una de las observaciones de la muestra y estimaríamos la función de distribución como hemos visto en (1.46).

En nuestro caso mantendremos esta asignación de pesos hasta encontrarnos con la primera observación censurada T_j. De acuerdo con la definición de censura aleatoria por la derecha en (1.12), que una observación esté censurada significa que lo único que sabemos sobre el tiempo de fallo para ese individuo es que es estrictamente mayor al tiempo observado. Sería lógico por tanto repartir el peso $\frac{1}{n}$ correspondiente al tiempo de fallo del j-ésimo individuo a partes iguales entre todos los posibles valores que puede tomar: $T > T_j$. Sin embargo, recordemos que solo se asigna probabilidad a los valores de las variables observados en la muestra, con lo que repartiremos el peso $\frac{1}{n}$ entre los tiempos de la muestra mayores estrictamente[4] que T_i, que como la muestra está ordenada serán: $\{T_k\}_{k>j}$. Estos tiempos observados T_k pasarán a tener una probabilidad asignada de

$$\frac{1}{n} + \frac{1/n}{\#\{k \in \{1, ..., n\} | k > j\}} = \frac{1}{n} + \frac{1/n}{k-j} = \frac{1}{n} + \frac{1}{n(k-j)}. \tag{1.51}$$

Procederemos del mismo modo cada vez que nos encontremos una observación censurada.

En resumen, el estimador de Kaplan-Meier asigna probabilidad únicamente a los tiempos de fallo de la muestra pero para asignar dichas probabilidades utiliza la información que le proporcionan los tiempos de censura.

Este estimador tiene un problema de definición cuando el mayor tiempo observado en la muestra es un tiempo de censura. En ese caso la probabilidad acumulada en dicha observación no se puede asignar (porque no hay ningún tiempo de fallo posterior) con lo que la función $\hat{S}$ no llega a valer 0 y no es estrictamente hablando una función de supervivencia. Existen varias maneras de solucionar este problema pero la más sencilla es imponer que la última observación sea no censurada.

El estimador de Kaplan-Meier es asintóticamente normal con media la verdadera función de supervivencia $S(t)$. Debido a su sencillez es el estimador más establecido en la literatura.

Estimador de Nelson-Aalen

Un estimador alternativo de $S(t)$ se puede obtener a partir del estimador de Nelson-Aalen de la función de riesgo acumulada utilizando la relación (1.10).

El estimador de Nelson-Aalen se obtiene a partir de la estimación no paramétrica de la razón de fallo detallada en (1.48) sin más que discretizar la relación entre λ y Λ transformando la integral en un sumatorio:

$$\hat{\Lambda} = \sum_{T_i \leq t} \frac{d_i}{n_i}. \tag{1.52}$$

Sustituyendo en (1.10) obtenemos entonces la siguiente estimación de la función de supervivencia

$$\hat{S}_{NA}(t) = e^{-\hat{\Lambda}(t)}. \tag{1.53}$$

[4]En caso de empate entre una muestra censurada y otra no censurada consideraremos primero la no censurada.

1.4. Censura en presencia de covariables

Hasta ahora hemos hablado de como estudiar la variable tiempo de fallo Y a partir de su observación en una muestra. Sin embargo, el objetivo del trabajo es más bien establecer la relación entre la variable respuesta Y y una serie de covariables $X = (X^1, ..., X^p) \in \mathbb{R}^p$. En particular, lo que nos interesa es tomar una herramienta que caracteriza la independencia entre variables (o vectores) aleatorios y ver cómo adaptarla al caso en que una de las variables involucradas sufre censura por la derecha.

Veamos en primer lugar la notación, las distintas covariables se denotarán por $X^{(j)}$, con $1 \leq j \leq p$ y dada una muestra, la observación correspondiente a cada individuo se denotará por un subíndice i como veníamos haciendo hasta ahora:

$$\{X_i, T_i, \delta_i\}_{i=1}^n, \quad X_i = (X_i^1, ..., X_i^p) \tag{1.54}$$

Pasamos a presentar las técnicas clásicas para lidiar con la presencia de covariables en los casos de censura aleatoria por la derecha, es decir, las técnicas de regresión con datos censurados. Los dos modelos más habituales en literatura son el modelo de riesgos proporcionales de Cox y el modelo de tiempo de fallo acelerado (AFT).

En ambos casos, las hipótesis de partida son las siguientes:

1. Las variables tiempo de fallo Y y tiempo de censura C son condicionalmente independientes dada la variable X. Es decir, fijado un valor $X = x$, $Y|x$ y $C|x$ son independientes.

2. Las covariables no varían con el tiempo[5].

La relación entre las covariables y el tiempo de fallo se formaliza a través de la función de distribución condicionada. Para estimarla volvemos a utilizar técnicas basadas en la maximización de la verosimilitud[6], en este caso de la función de verosimilitud condicionada al valor de la covariable.

Veamos cual es su expresión. Para ello establecemos la siguiente notación

$$\begin{aligned} Y|X = x &\sim F(\cdot|x), \\ C|X = x &\sim G(\cdot|x), \end{aligned} \tag{1.55}$$

y análogamente las funciones de densidad serían $f(\cdot|x)$ y $g(\cdot|x)$ respetivamente.

Siguiendo un procedimiento análogo al visto para la estimación de la función de distribución incondicional, la función de verosimilitud condicionada se puede expresar entonces del siguiente modo:

$$\mathcal{L}_n(F) = \prod_{i=1}^n f(T_i|X_i)^{\delta_i}(1 - F(T_i|X_i))^{1-\delta_i} \cdot \prod_{i=1}^n g(T_i|X_i)^{\delta_i}(1 - G(T_i|X_i))^{1-\delta_i}, \tag{1.56}$$

y por la hipótesis 1 de independencia condicional de Y y C, podemos quedarnos solo con la parte que involucra a la distribución de interés F obteniendo

$$\mathcal{L}_n(F) = \prod_{i=1}^n f(T_i|X_i)^{\delta_i}(1 - F(T_i|X_i))^{1-\delta_i}, \tag{1.57}$$

o, utilizando las relaciones detalladas en (1.8) y (1.10)

$$\mathcal{L}_n(F) = \prod_{i=1}^n \lambda(T_i|X_i)^{\delta_i} e^{-\Lambda(T_i|X_i)}. \tag{1.58}$$

[5]El caso en el que varían con el tiempo es una extensión de la teoría que veremos aquí que requiere un tratamiento aparte, ver por ejemplo Therneau y Grambsch (2000) o Lin y Fleming (2012).

[6]Dependendiendo del modelo también se pueden realizar ajustes de otro tipo, por ejemplo por mínimos cuadrados.

1.4.1. Modelos paramétricos

Estos modelos se basan en utilizar familias paramétricas de funciones de distribución y extender estos modelos para incluir covariables. Veamos por ejemplo cómo se extendería un modelo Weibull para considerar la relación entre el tiempo de supervivencia y las covariables tal y como lo formulan en Lawless (2011).

Supongamos que contamos con una variable tiempo de fallo Y que sigue una distribución Weibull $W(\alpha, \gamma)$ conforme hemos descrito en la Sección 1.3.2. Una manera de incorporar el efecto de un vector de covariables $X \in \mathbb{R}^p$ es permitir que alguno de los dos parámetros dependa de él. Lo más habitual es incluir ese efecto en el parámetro γ, de manera que la función de supervivencia condicionada para un tiempo y quedaría

$$S(y|x) = e^{-(\gamma(x)y)^{\alpha}}, y \geq 0. \tag{1.59}$$

En estas circunstancias la proporción entre razones de fallo para dos individuos no depende del tiempo:

$$\frac{\lambda(y|x)}{\lambda(y|x^*)} = \left(\frac{\gamma(x)}{\gamma(x^*)}\right)^{\alpha-1} \frac{\gamma'(x)}{\gamma'(x^*)}. \tag{1.60}$$

En el planteamiento clásico la dependencia se introduce a través de un modelo log-lineal del siguiente modo:

$$\gamma(x) = \beta^t x, \tag{1.61}$$

aunque también se podrían plantear fórmulas más flexibles.

Tal y como sucedía para la estimación de la distribución, el modelo anterior admite una formulación equivalente para $log(Y)$ como sigue

$$\log Y = \mu(x) + \sigma\varepsilon \tag{1.62}$$

donde ya hemos indicado que ε sigue una distribución valor extremo y $\mu(x) = \gamma(x) = x^t\beta$ (un modelo log-lineal sobre $Y|X$ equivale a modelizar linealmente $log(Y|X)$).

En cualquier caso, la principal limitación de estos métodos es evidente: solo son válidos cuando el tiempo de fallo sigue la distribución asumida, en este caso una Weibull. Sería muy valioso, por tanto, contar con herramientas más generales válidas cuando no se cumpla o cuando no sea posible comprobarla.

1.4.2. Modelos semiparamétricos

Estos modelos combinan una parte paramétrica (normalmente el efecto de las covariables) con una parte no paramétrica que dependiendo de nuestros objetivos no siempre es necesario calcular.

El precio por esta flexibilidad es que imponen una forma concreta para la relación entre funciones de supervivencia (AFT) o funciones de riesgo (Cox).

Modelo de tiempo de fallo acelerado (AFT)

El modelo de tiempo de fallo acelerado es un modelo paramétrico libre de distribución, es decir, general para distintas distribuciones que encajen en los supuestos del modelo.

El modelo tiene la siguiente forma

$$S(y|x) = S_0\left(\theta y\right), \tag{1.63}$$

donde $\theta = \theta(x)$ captura el efecto conjunto de las covariables o factor de aceleración y S_0 la función de supervivencia basal que en principio puede ser desconocida. De nuevo lo más habitual es introducir el efecto de las covariables a través de un modelo log-lineal de manera que:

$$S(y|x) = S_0\left(e^{-\beta^t x} y\right), \tag{1.64}$$

es decir, se asume que las covariables influyen en el ritmo de avance de la función de supervivencia. Si $\beta_j > 0$, entonces $e^{-\beta_j}$ se denomina factor de deceleración y significa que el efecto de aumentar la covariable $X^{(j)}$ es que el tiempo se ralentiza. Si $\beta_j > 0$, se acelera. El caso $\beta^j = 0$ sería, lógicamente, que la variable $X^{(j)}$ no tiene efecto (bajo la estructura del modelo lineal, es decir, solo nos sirve para identificar correlación entre la covariable y $log(Y)$).

Este modelo admite una formulación equivalente sobre la variable $log(Y)$:

$$\log Y = \mu(x) + \sigma\varepsilon, \tag{1.65}$$

con $\mu(x) = log(\theta(x))$ es la función de localización que captura el efecto de las covariables, σ se denomina parámetro de escala y ε sigue una distribución independiente de las covariables. Debido a esta formulación, estos modelos también se conocen como modelos de localización-escala.

El modelo equivalente a (1.64) sería aquel en el que el efecto de las covariables sobre $log(Y)$ es lineal:

$$\mu(x) = \beta^t x. \tag{1.66}$$

Aunque en principio estos modelos son semiparamétricos, donde la parte que recoge el efecto de las covariables tiene una forma prefijada y se le da libertad a S_0 (o a la distribución del error dependiendo de la formulación que utilicemos), en la práctica se utilizan a menudo como modelos puramente paramétricos presuponiendo una distribución paramétrica para el tiempo de supervivencia (o su logaritmo). En este sentido son muy útiles las formulaciones de distintas distribuciones en forma de modelo de localización escala presentadas en el apartado 1.3.2.

En cuanto a su aplicación como modelo semiparamétrico, está poco establecida en la literatura. El ajuste más destacable es el de Buckley-James pero presenta importantes inconvenientes en la práctica y no hay estrategias de validación claras.

Para validar el modelo de tiempo de fallo acelerado paramétrico serviría con comprobar que se cumple la hipótesis del modelo (las covariables tienen un efecto multiplicativo sobre la función de riesgo) y realizar tests de bondad de ajuste. Ninguna de estas herramientas nos proporciona, en principio, manera de identificar dependencia en general entre las variables y la respuesta.

Para más detalles sobre la aplicación, validación y ejemplos de ajuste de modelos semiparamétricos de tipo tiempo de fallo acelerado (AFT) o modelos de Cox en el ámbito de la fiabilidad se puede consultar López Montoya (2011).

Modelo de riesgos proporcionales de Cox

El modelo de riesgos proporcionales de Cox es el más común en la literatura. Presupone a priori la siguiente forma semiparamétrica para la función de riesgo o razón de fallo

$$\lambda(y|x) = \lambda_0(y)e^{\beta^t x}, \beta \in \mathbb{R}^p, \tag{1.67}$$

en el que se diferencian los siguientes elementos:

- Parte no paramétrica: el riesgo basal $\lambda_0(t)$

  Es una curva que puede tener cualquier forma y representa la función de riesgo cuando las covariables son cero. Si dichas covariables están centradas, puede interpretarse como la función de riesgo para un individuo promedio. Veremos que no es necesario estimar la forma del riesgo basal para poder estimar el efecto de las covariables sobre el riesgo.

- Parte paramétrica: el efecto de las covariables $e^{\beta^t x}$

  Impone una forma muy específica para el efecto de las covariables sobre el tiempo de fallo. Su forma recuerda a la de un modelo lineal generalizado con función link exponencial (modelo log lineal) para la proporción entre funciones de riesgo y se comporta de manera similar. Deshaciendo

la transformación dada por la función link vemos que, en efecto, es equivalente a modelizar de manera lineal la siguiente magnitud:

$$\ln\left(\frac{\lambda(y|x)}{\lambda_0(y)}\right) = \beta^t x. \tag{1.68}$$

Esta parte paramétrica del modelo es evidentemente la parte más restrictiva.

El nombre del modelo proviene de que como consecuencia de (1.67), la razón entre dos funciones de riesgo de individuos distintos es constante en el tiempo y solo depende de las diferencias entre sus covariables:

$$\frac{\lambda(y|x)}{\lambda(y|x^*)} = \beta^t x, \tag{1.69}$$

a esta magnitud se la denomina riesgo relativo o hazard ratio (HR).

La ecuación anterior permite establecer una interpretación clara para los coeficientes del modelo. En el caso de que la covariable $X^{(j)}$ sea continua, si $X^{(j)}$ aumenta en una unidad y todas las demás se mantienen constantes, el riesgo se multiplica por e^{β_j}. Si $e^{\beta^j} = 1$ significa que el riesgo no depende de la variable $X^{(j)}$. Si la variable es categórica, entonces se establece una categoría de referencia y e^{β_j} es la magnitud por la que se multiplica el riesgo al pasar de dicha categoría a la categoría j.

Como consecuencia de la forma de la función de riesgo condicionada, utilizando la definición (1.7) y la relación (1.10) obtenemos que la función de riesgo acumulado condicionada y la función de supervivencia condicionada también han de tener una forma específica:

$$\begin{aligned} &\Lambda(y|x) = \Lambda_0(y)e^{\beta^t x}, \quad \Lambda_0(y) = \int_0^1 \lambda_0(u)du, \\ &S(y|x) = S_0(y)^{exp(\beta^t x)}, \quad S_0(y) := exp(-\Lambda_0(T)). \end{aligned} \tag{1.70}$$

Gráficamente esto significa que las curvas de la función de supervivencia de dos individuos nunca se van a cruzar y además el logaritmo de sus razones de fallo acumuladas serán paralelas.

Un ejemplo de distribución que encaja en los supuestos del modelo de riesgos proporcionales de Cox es la distribución Weibull (la única distribución que cumple la hipótesis de riesgos proporcionales y de tiempo de fallo acelerado simultáneamente).

La estimación de los coeficientes β se obtiene maximizando la función de verosimilitud parcial, que se obtiene como verosimilitud perfil a partir de la verosimilitud condicional y es el producto de la probabilidad condicional de fallo dadas las observaciones a riesgo en cada tiempo T_i de la muestra. Para maximizar esta función no es necesario conocer la función de riesgo basal $\lambda_0(y)$.

Existen extensiones de este modelo que incorporan efectos más flexibles a la hora de introducir a las covariables. Así, el ajuste en (1.68) se podría modelizar también utilizando splines como ilustra Flores Flores (2011) o modelos aditivos con suavizado como se plantea en Lin y Fleming (2012). Estos ajustes permiten capturar efectos no lineales que la modelización paramétrica no podría, pero lo habitual es usar modelos paramétricos.

Testar efecto de las covariables

De los tres modelos presentados hasta ahora (paramétrico, AFT o Cox) sin duda el más extendido en la literatura es el modelo de Cox. A lo largo de esta sección presentaremos las herramientas clásicas con las que contamos para estudiar el efecto de las covariables centrándonos en este tipo de modelos. Algunas de estas técnicas son aplicables a AFT y a modelos paramétricos.

A raíz de los estimadores del modelo de Cox se puede construir, como sucede en los modelos lineales, tests para evaluar hipótesis simples de la forma

$$H_0 : \beta = \beta_0 \tag{1.71}$$

que se utilizar en general para testar la hipótesis de no efecto $\beta_0 = 0$.

Los tests más utilizados para realizar estos contrastes son el test de Wald, el test de razón de verosimilitudes y el score test. Los dos primeros se pueden adaptar además al caso en el que solo nos interesa ver qué sucede con un subconjunto de parámetros de β.

- El test de Wald se basa en la normalidad asintótica del estimador de máxima verosimilitud, que bajo hipótesis nula cumple
$$(\hat{\beta} - \beta_0)^t I(\hat{\beta})(\hat{\beta} - \beta_0) \sim \chi^2_p, \tag{1.72}$$
donde $I(\beta)$ es la matriz de información de Fisher.

- El test de razón de verosimilitud se plantea en los siguientes términos:
$$\begin{aligned} H_0 &: \text{ Modelo simplificado } \beta = 0 \quad (k \text{ grados de libertad}), \\ H_1 &: \text{ Modelo estimado } \beta = \hat{\beta} \quad (k+p \text{ grados de libertad}), \end{aligned} \tag{1.73}$$
y compara la verosimilitud obtenida bajo ambos supuestos utilizando el siguiente estadístico:
$$2(\ln L(\hat{\beta}) - \ln L(\hat{\beta_0})) \sim \chi^2_p. \tag{1.74}$$

- El score test se basa en comparar el puntaje de eficiencia que se define como
$$U(\beta) = \frac{\partial \mathcal{L}(\beta|x)}{\partial \beta}, \tag{1.75}$$
que también es asintóticamente normal con media nula y matriz de varianzas covarianzas la matriz de información de Fisher, con lo que
$$U(\beta_0)^t I(\beta_0)^{-1} U(\beta_0) \sim \chi^2_p. \tag{1.76}$$

Estos son los enfoques clásicos a la hora de lidiar con la presencia de covariables con respuesta censurada. Los tests anteriores, además de poder aplicarse únicamente cuando se cumplen las hipótesis del modelo, solamente testan la relación bajo la forma paramétrica estricta dictada por el modelo. Por ejemplo, bajo el modelo de Cox de riesgos proporcionales presentado anteriormente si una variable tiene un efecto no lineal sobre $ln\left(\frac{\lambda(y|x)}{\lambda(y|X^*)}\right)$ estos tests no serán capaces de identificarla.

El test de razón de verosimilitudes es el más general y el que mejores propiedades estadísticas tiene. Sirve en general para comparar cualesquiera dos modelos anidados.

Existe, sin embargo, una herramiento mucho más interesante relacionada con el modelo de regresión de riesgos proporcionales de Cox para tratar el caso de dependencia general. Para presentarla debemos ver primero como se estima la función de riesgo acumulado condicional bajo este modelo.

El estimador de la función de riesgo acumulado basal se llama estimador de Breslow y es un estimador no paramétrico que se obtiene a partir de la función de verosimilitud condicional:
$$\hat{\lambda}_0(y) = \frac{\sum_{i=1}^n \mathbb{I}(T_i = y, \delta_i = 1)}{\sum_{i=1}^n \mathbb{I}(T_i \geq y) e^{\beta^t X_i}}, \tag{1.77}$$
con lo que
$$\hat{\Lambda}_0(y) = \sum_{i=1}^n \hat{\lambda}_0(y) \mathbb{I}(T_i \leq y) \tag{1.78}$$
y sin más que sustituir en (1.67) y (1.70) nos permite obtener los estimadores de la función de riesgo condicional $\hat{\lambda}_B(y|x)$, de la función de riesgo acumulado condicional $\hat{\Lambda}_B(y|x)$ y de la función de supervivencia condicional $\hat{S}_B(y|x)$.

Las extensiones del modelo de riesgos proporcionales de Cox tales como el uso de residuos para la validación de hipótesis se sustentan en su formulación como proceso de conteo (counting process) tal y como se explica en la sección Extending the Cox Model, páginas 51 a 84 en Lin y Fleming (2012), o en Flores Flores (2011).

En dicha formulación cada individuo i de la muestra se representa por un proceso de conteo donde

- $N_i(y)$ es el número de eventos acumulados hasta tiempo y para el individuo i: 0 o 1.
- $Y_i(y)$ es un indicador de si el individuo está a riesgo en t.

Cada individuo i puede entenderse entonces como una observación de un proceso de Poisson con intensidad constante.

Entre las extensiones del modelo basadas en esta representación se incluye el tratamiento de variables dependientes de tiempo, el tratamiento de modelos estratificados (con estratos dependientes o no de tiempo), con intervalos de riesgo discontinuos o el uso de residuos para validar las hipótesis del modelo. Nos centraremos en este último aspecto. Para más información estas extensiones se puede consultar Therneau y Grambsch (2000), Lin y Fleming (2012) o Iglesias Pérez (2021).

Vamos a centrarnos en el análisis de un tipo de residuos que nos va a permitir detectar cualquier forma de dependencia entre una covariable y la variable respuesta: los residuos martingale o martingala.

La obtención de estos residuos se basan en calcular para cada individuo i de la muestra la diferencia entre el proceso observado y el predicho por el modelo:

$$M_i(y) = \int Y_i(y)(dN_i(y) - d\Lambda_i(y))dy, \tag{1.79}$$

donde el primer término de la resta se correspondería al valor observado y el segundo al esperado.

Como estamos bajo las hipótesis del modelo de Cox podemos estimar la función de riesgo acumulado utilizando el estimador de Breslow, obteniendo el residuo martingale para cada individuo

$$m_i(y) = \hat{M}_i(y) = \int \mathbb{I}(T_i \geq y)(dN_i(y) - d\hat{\Lambda}_i(y))dy, \tag{1.80}$$

donde N_i es la cantidad de fallos que ha experimentado el individuo hasta y, asumiendo que las covariables no varían con el tiempo nos quedará

$$\begin{aligned} m_i(y) &= N_i(y) - e^{\hat{\beta}^t x} \int_0^y \mathbb{I}(T_i \geq s) d\hat{\Lambda}_0(s) ds = N_i - e^{\hat{\beta}^t x} \hat{\Lambda}_0(y) \\ &= N_i - \hat{\Lambda}_i(y|x), \end{aligned} \tag{1.81}$$

y consideramos para cada individuo su residuo martingala

$$m_i = m_i(T_i) = \delta_i - \hat{\Lambda}_i(T_i|X_i) \tag{1.82}$$

Podemos interpretar estos residuos como la diferencia entre la cantidad de fallos que experimenta el individuo hasta T_i (incluido) y la estimación de la cantidad de fallos que sería esperable para dicho individuo utilizando el modelo ajustado.

Las propiedades de estos residuos son:

1. $\mathbb{E}(M_i) = 0$,
2. $Cov(M_i, M_j) = 0$,
3. $\sum \hat{M}_i = 0$,
4. $Cov(\hat{M}_i, \hat{M}_j) = 0$ (consecuencia de la anterior),
5. $\hat{M}_i \in [0, 1]$.

y presentan una notable asimetría.

Se utilizan para descubrir la verdadera forma funcional con la que una variable $X^{(j)}$ está involucrada en el modelo. Lo que se hace es representar los residuos suavizados de un modelo de referencia frente a la covariable $X^{(j)}$. Esto se puede realizar de varias formas.

El enfoque más sencillo se puede encontrar en Therneau y col. (1990). Consiste en tomar como ajuste inicial el modelo nulo (sin efecto de ninguna covariable, $\beta = 0$) y calcular sus residuos martingale. La hipótesis de partida, sumamente flexible, es que el efecto de la variable $X^{(j)}$ sobre la razón de fallo tiene la forma

$$e^{\beta} f(X^{(j)}), \tag{1.83}$$

donde f puede ser cualquier función. Al representar los residuos martingale suavizados m_i frente a las observaciones de la covariable $X_i^{(j)}$, el gráfico representará la forma funcional de f ya que

$$\mathbb{E}(M_i) \approx c f(X_i). \tag{1.84}$$

De este modo podremos transformar los datos de manera adecuada para que la variable transformada pueda entrar con un efecto lineal en el modelo.

Este método inicial presenta algunos problemas, por ejemplo si existe correlación entre variables es posible que identifique relaciones espúreas con la respuesta. Además no está claro como obtener las bandas de confianza para la tendencia representada.

Los mismos autores proponen solución a estos problemas tal y como se detalla en Lin y Fleming (2012) tomando sencillamente como modelo de partida el modelo de Cox ajustado para todas las covariables, obteniendo sus residuos martingale y utilizando estos residuos como respuesta para ajustar un modelo de regresión de Poisson con suavizado en aquellas covariables de las que nos interese ver su forma (se incluye dentro de los modelos lineales generalizados o GAM). En la práctica se sigue utilizando más la representación directa de los residuos con suavizado debido a su sencillez.

En general si en lugar del modelo nulo consideramos otro ajuste de partida, entonces si la covariable no está incluida en el modelo, el gráfico suavizado mostrará la forma del efecto de esa covariable. Si la variable ya está en el modelo y está bien modelada, el gráfico no debería mostrar tendencia.

Este método gráfico es muy informativo ya que además de identificar la independencia nos permite, si hay dependencia, determinar su forma. Sin embargo, presenta varios inconvenientes. En primer lugar solo es aplicable si se cumplen las hipótesis de riesgos proporcionales de Cox. En segundo lugar es un método cualitativo que requiere la interpretación de gráficas, con lo que determinar la independencia de un número moderado de covariables puede ser ya una tarea ardua. En tercer lugar, solo es aplicable para covariables unidimensionales. Por tanto, no es una herramienta que nos vaya a ser útil en contextos de datos de alta dimensión o con covariables multidimensionales. Además, los test construidos en base a estos residuos muestran potencias muy bajas a la hora de identificar dependencias de tipo no monótono entre covariables y respuestas.

1.4.3. Modelos no paramétricos

En Beran (1981) se propone un estimador íntegramente no paramétrico para estimar la función de supervivencia condicionada $S(y|x)$ bajo censura aleatoria. Las únicas hipótesis utilizadas serán la hipótesis de no negatividad de Y y C y la hipótesis de independencia condicional entre $Y|x$ y $C|x$.

El punto de partida de este estimador es la siguiente expresión probada inicialmente por Cox (1972) para el caso en que S tiene un conjunto de discontinuidades $D(x, S)$ finito y extendido después por Beran (1981) al caso general:

$$S(y|x) = exp\left(\int_0^y \frac{dV_c(t|x)}{S_T(t|x)}\right) \prod_{t \in D(x,V)} \left[1 - \frac{V(t^-|x) - V(t|x)}{S_T(t|x)}\right]^{\mathbb{I}(t<y)}, \tag{1.85}$$

donde $S_T(y|x)$ es la función de supervivencia del tiempo observado $T = \text{mín}\{Y, C\}$ condicionado a $X = x$ definida por

$$S_T(y|x) = \mathbb{P}(T > y|X = x), \tag{1.86}$$

y denotamos por $V(y|x)$ a

$$V(y|x) = \mathbb{P}(T > y, \delta = 1|X = x), \tag{1.87}$$

siendo V_c su componente continua y $D(x, V)$ el conjunto de discontinuidades.

Basta entonces con encontrar estimadores no paramétricos de $V(y|x)$ y $S_T(y|x)$ y sustituir en la expresión anterior para poder estimar $S(y|x)$. Utilizando los estimadores propuestos por Stone (1977) que son sumas ponderadas de las observaciones muestrales cuyos pesos $W_{n,i}(x) \geq 0$ dependen de la covariable y suman 1, el estimador buscado quedaría del siguiente modo:

$$\hat{S}(y|x) = \prod_{t \in D(x,\hat{V})} \left(1 - \frac{\sum_{i=1}^{n} W_{n,i}(x)\mathbb{I}(T_i \geq t)}{\sum_{i=1}^{n} W_{n,i}(x)\mathbb{I}(T_i \geq t, \delta_i = 1)}\right)^{\mathbb{I}(t<y)}, \tag{1.88}$$

y se conoce como estimador de Beran.

Los resultados de consistencia para los estimadores $\hat{V}$ y $\hat{S_T}$ garantizan la consistencia de este estimador.

Además, en Gonzalez-Manteiga y Cadarso-Suarez (1994) a partir de la expresión del estimador Kaplan-Meier generalizado (GKM), que coincide con el estimador de Beran cuando los pesos utilizados suman 1, propone una representación casi segura del mismo como suma de variables aleatorias independendientes. Esto permite obtener resultados de convergencia asintótica a una normal para los estimadores de $S(y|x)$ tanto en el caso de datos completos como bajo censura.

Capítulo 2

Correlación de distancias

En este capítulo vamos a introducir una nueva metodología para caracterizar la independencia entre variables aleatorias de dimensión arbitraria. Lo primero que hemos de hacer, por tanto, es definir de manera precisa qué significa que dos variables o vectores aleatorios sean independientes.

Se dice que X e Y son independientes cuando la probabilidad de observar una realización concreta del par (X, Y) es igual al producto de las probabilidades de observar el valor de cada variable por separado:

$$\mathbb{P}((X, Y) = (x, y)) = \mathbb{P}(X = x)\mathbb{P}(Y = y). \tag{2.1}$$

Es decir, el valor que tome una de las variables no afecta a la otra.

Esta condición de independencia se puede reexpresar de muchas formas distintas. La más habitual quizás sea en términos de la función de distribución. Denotamos por F_X y f_X a las funciones de distribución y de densidad (si existe) de la variable X, análogamente para la variable Y.

La condición de independencia en términos de la función de distribución es que la función de distribución de Y condicionada a X sea igual a la distribución marginal de Y (o viceversa). Es decir:

$$F_{Y|X} = F_Y, \tag{2.2}$$

o equivalentemente que la distribución conjunta de ambas variables es producto de las marginales:

$$F_{Y|X}(y|x) = \frac{F_{X,Y}(x, y)}{F_X(x)} = F_Y(y) \quad \Leftrightarrow \quad F_{X,Y}(x, y) = F_X(x)F_Y(y). \tag{2.3}$$

En la práctica, para variables continuas suele presentarse muchas veces en términos de la función de densidad:

$$f_{X,Y}(x, y) = f_X(x)f_Y(y). \tag{2.4}$$

Recordemos que las funciones característica y de distribución existen siempre, no así la de densidad. En general a lo largo de este trabajo consideraremos variables continuas para las que no habrá problema.

2.1. Dependencia lineal: correlación de Pearson

De todas las relaciones de dependencia que pueden existir entre dos variables aleatorias, la más sencilla, la más estudiada y para la que contamos con herramientas más consolidadas es la relación lineal o correlación.

Si denotamos por $X, Y \in \mathbb{R}$ dos variables aleatorias arbitrarias, la medida básica de la relación lineal existente entre ellas es la llamada covarianza, que mide la variabilidad conjunta entre ambas variables del siguiente modo:

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y). \tag{2.5}$$

Cuando esta magnitud toma el valor cero, decimos entonces que las variables X e Y son incorreladas o no tienen relación lineal entre sí. Es evidente, en vista de la expresión anterior, que cualquier par de variables aleatorias independientes será incorrelado, mas no se cumple en general la implicación opuesta pues existen muchos tipos de dependencia distintos de la lineal.

De la definición (2.5) también se sigue, sin más que usar las propiedades básicas de la esperanza, que esta magnitud es invariante ante cambios de localización, pero no de escala.

En cuanto a su interpretación, el signo nos indica si la relación lineal es directa o inversa, pero la magnitud es difícil de interpretar debido a que sus unidades son las unidades de X multiplicadas por las unidades de Y. Además esto hace que sea difícil comparar la correlación entre dos pares de variables distintos.

Para solventar los problemas de interpretación y comparabilidad, nos basta con reescalar el valor anterior teniendo en cuenta la desviación típica de las variables involucradas, es decir, la dispersión intrínseca a cada una de ellas.

La variabilidad propia de cada variable involucrada se obtiene calculando la varianza, es decir, la covarianza de una variable consigo misma

$$Var(X) = Cov(X, X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2. \tag{2.6}$$

Esta magnitud tendrá como unidades las unidades de X^2. Si queremos una medida de dispersión en las unidades de la variable original debemos calcular su raíz cuadrada, que llamaremos desviación típica, desviación estándar o error típico:

$$\sigma_X = \sqrt{Var(X)}. \tag{2.7}$$

Análogamente se obtendrían la varianza y la desviación típica de Y.

De esta forma, reescalando la covarianza con ayuda del error típico, definimos el coeficiente de correlación lineal de Pearson:

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]}{\sigma_X \sigma_Y}, \tag{2.8}$$

una magnitud que solventa algunos de los problemas que nos planteaba la covarianza entre variables.

Es evidente a partir de la definición (2.8) que el coeficiente de correlación de Pearson es invariante ante cambios de escala y, ahora sí, también de localización. En cuanto a su interpretación, el signo nos sigue indicando si la relación lineal es directa o inversa, pero ahora además se trata de una magnitud adimensional que toma valores en el intervalo acotado $[-1, 1]$ de manera que cuanto más próximo a 1 está su valor absoluto más fuerte es la correlación, con lo que es mucho más interpretable y se puede comparar fácilmente el coeficiente de correlación obtenido para distintos pares de variables.

Podemos decir, por tanto, que contamos con una medida de la correlación lineal entre variables aleatorias muy sencilla de calcular, de interpretar y con propiedades muy deseables a la hora de trabajar con ella.

Las limitaciones de la correlación de Pearson como medida de dependencia se pueden resumir en dos aspectos: solo está definida para dos variables aleatorias y solo caracteriza la dependencia lineal.

En el primer aspecto, si tratamos con un vector aleatorio $X \in \mathbb{R}^p, p > 1$ podemos utilizar la correlación de Pearson para determinar el grado de dependencia lineal entre cada par de componentes del mismo. Del mismo modo, podemos medir el grado de relación lineal entre pares de componentes de dos vectores aleatorios de dimensiones arbitrarias $X \in \mathbb{R}^p$ e $Y \in \mathbb{R}^q$, pero no contamos con una medida global de la correlación entre ambos.

En cuanto a la segunda cuestión, es interesante destacar el caso de la población normal bivariante. Si trabajamos con un vector normal bivariante

$$(X, Y) \sim N_2(\mu, \Sigma) \tag{2.9}$$

entonces es bien sabido que la distribución marginal de cada una de sus componentes es normal y además dichas componentes son independientes si, y solo si, son incorreladas. Entonces en este caso

tan particular el coeficiente de correlación de Pearson es una magnitud que nos permite caracterizar la independencia entre X e Y y, en caso de que exista dependencia, determinar su magnitud.

La carencia de medidas más generales que nos permitan testar dependencia entre variables aleatorias que sigan cualquier distribución y, más aún, entre vectores u objetos aleatorios más genéricos no ha pasado inadvertida en la literatura. Sin embargo, los intentos clásicos de testar la independencia se han revelado poco eficaces, bien porque requieren de hipótesis muy restrictivas a la hora de trabajar con datos reales, porque no son efectivos a la hora de trabajar con datos de alta dimensión (muy comunes en genética) o, en general, porque son poco sensibles ante determinadas formas de dependencia (por ejemplo dependencia no monótona). Un primer acercamiento al problema utilizando la metodología de correlación de distancia se puede encontrar en Bakirov y col. (2006).

El objetivo en Székely y col. (2007) es presentar una medida de dependencia entre dos vectores aleatorios de dimensiones arbitrarias, siendo dicha magnitud en muchos sentidos una extensión de las ideas subyacentes a la correlación lineal de Pearson similar en interpretación y propiedades.

2.2. Dependencia general: correlación de distancias

El desarrollo fundamental de la propuesta para caracterizar la independencia basada en la distancia entre objetos estadísticos se puede encontrar en Székely y col. (2007). En dicho trabajo se propone la correlación de distancias $\mathcal{R}$ como medida de la dependencia entre dos variables o vectores aleatorios de dimensiones arbitrarias, generalizando así el coeficiente de correlación lineal de Pearson en varios sentidos. A continuación veremos brevemente la idea que hay detrás de su definición y probaremos que cumple las siguientes propiedades:

- $\mathcal{R}(X,Y)$ está definida para $X \in \mathbb{R}^p$ e $Y \in \mathbb{R}^q$ variables/vectores aleatorios de dimensiones arbitrarias $p \geq 1, q \geq 1$. Hablaremos indistintamente de variables o vectores aleatorios para abreviar, en cualquier caso estamos considerando cualquier dimensión $p \geq 1$.
- $\mathcal{R}(X,Y)$ caracteriza la dependencia entre X e Y:

$$\mathcal{R}(X,Y) = 0 \quad \Leftrightarrow \quad \text{X e Y son independientes.} \tag{2.10}$$

- $\mathcal{R}(X,Y) \in [0,1]$.
- $\mathcal{R}(X,Y)$ es invariante ante cambios de localización y escala.

Para el caso de la normal bivariante el coeficiente de correlación de Pearson ya funcionaba perfectamente, con lo que el objetivo es que $\mathcal{R}$ herede en lo posible este buen comportamiento. En particular, veremos que se cumple

$$\mathcal{R}(X,Y) \leq |\rho(X,Y)|, \tag{2.11}$$

lo que garantiza que en todos los casos en los que se detecta independencia con la correlación de Pearson se detectará también utilizando $\mathcal{R}$. Además indica que como medida de la dependencia lineal la correlación de distancias aspira a ser como mucho igual de eficaz que la correlación de Pearson.

2.2.1. Distancia entre funciones características

Al inicio del capítulo hemos visto distintas formas de expresar la condición de independencia siendo la más general y de las más utilizadas (2.3). En Székely y col. (2007) se parte de una expresión equivalente de dicha condición en términos de funciones características.

Dado un vector aleatorio $X \in \mathbb{R}^p$, la función característica de X es una función compleja $\varphi : \mathbb{R}^p \longrightarrow \mathbb{C}$ que se define como sigue:

$$\varphi_X(t) = \mathbb{E}\left(e^{i\langle t, X\rangle}\right), t \in \mathbb{R}^p, \tag{2.12}$$

donde trivialmente en el caso de una variable aleatoria ($p = 1$) la expresión anterior se reduce a

$$\varphi_X(t) = \mathbb{E}\left(e^{itX}\right), t \in \mathbb{R}. \tag{2.13}$$

Recordemos que la función característica, que existe siempre, determina la distribución de probabilidad de X existiendo una relación biunívoca entre φ_X y F_X (aún cuando alguna de ellas no se pueda expresar en términos de funciones sencillas). En caso de que exista, la función de densidad f_X también está, por tanto, caracterizada por φ_X, en particular una es transformada de Fourier de la otra. Análogamente denotaremos por φ_Y a la función característica de Y y por $\varphi_{X,Y} : \mathbb{R}^p \times \mathbb{R}^q \longrightarrow \mathbb{C}$ a la función característica del par (X, Y) definida como sigue:

$$\varphi_{X,Y}(t,s) = \mathbb{E}\left(e^{i\langle t,X\rangle + i\langle s,Y\rangle}\right), (t,s) \in \mathbb{R}^p \times \mathbb{R}^q \equiv \mathbb{R}^{p+q}. \tag{2.14}$$

Es evidente a la vista de las relaciones anteriores que (2.3) es equivalente a

$$\varphi_{X,Y} = \varphi_X \varphi_Y, \tag{2.15}$$

Nuestro objetivo es entonces encontrar un estadístico que nos permita realizar el siguiente contraste de hipótesis:

$$\begin{cases} H_0: & \varphi_{X,Y} = \varphi_X \varphi_Y \quad \text{(independencia)}, \\ H_1: & \varphi_{X,Y} \neq \varphi_X \varphi_Y \quad \text{(dependencia)}. \end{cases} \tag{2.16}$$

Determinar si existe dependencia entre X e Y se reduce entonces a encontrar una forma adecuada de medir la distancia entre las funciones características φ_X y φ_Y, de forma que satisfaga una serie de exigencias que veremos a continuación. Una vez establezcamos una forma adecuada de medir esta distancia hemos de plantear un estadístico de contraste evaluable sobre la muestra.

La idea de construir estadísticos evaluados no sobre la muestra sino sobre distancias entre elementos de la muestra es la clave del desarrollo propuesto en Székely y col. (2007) y, más en general, del desarrollo de los llamados Energy Statistics en Székely y Rizzo (2017), un marco teórico global que permite proponer técnicas no solo para identificar la independencia sino para otras muchas aplicaciones como contrastes de bondad de ajuste, clutering o comparaciones de muestras entre otros.

Queremos por tanto definir una distancia entre funciones complejas φ_X y φ_Y que sea adecuada a nuestros propósitos. Vamos a considerar el espacio L_2 ponderado de funciones complejas definidas en $\mathbb{R}^p \times \mathbb{R}^q \equiv \mathbb{R}^{p+q}$ dotado de la norma $\|\cdot\|_w$ que hemos de definir de manera conveniente.

Sea $g : \mathbb{R}^{p+q} \longrightarrow \mathbb{C}$ una función cualquiera en el espacio anterior, en general una norma L_2 ponderada obedece a la siguiente expresión:

$$\|g(t,s)\|_w^2 = \int_{\mathbb{R}^{p+q}} |g(t,s)|^2 w(t,x) dt ds \tag{2.17}$$

donde $|g(t,s)| = g(t,s)g(\bar{t},s) = \sqrt{Re(g(t,s))^2 + Im(g(t,s))^2}$ denota el módulo complejo y $w(t,s)$ define la función de pesos, una función positiva tal que la integral anterior existe.

La elección de la función de pesos no es única, sin embargo la elección de los autores en Székely y col. (2007) presenta importantes ventajas tales como que la norma resultante presenta propiedades deseables y sus análogos muestrales admiten una interesante interpretación como veremos más adelante.

El proceso de selección se enfoca hacia lograr que la magnitud

$$\begin{aligned} \mathcal{V}^2(X,Y;w) :&= \|\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)\|_w^2 \\ &= \int_{\mathbb{R}^{p+q}} |\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)|^2 w(t,s) dt ds, \end{aligned} \tag{2.18}$$

cumpla:

- $\mathcal{V}^2(X,Y;w)$ debe estar bien definida en todo $(s,t) \in \mathbb{R}^{p+q}$, es decir, la integral anterior ha de ser finita.

- Queremos que caracterice la independencia, es decir,

$$\mathcal{V}^2(X,Y;w)=0 \quad \Leftrightarrow \quad \text{X e Y son independientes.} \tag{2.19}$$

- Impondremos que en caso de dependencia, la magnitud anterior tome valores estrictamente positivos $\mathcal{V}^2(X,Y;w)>0$.

Así tendríamos una medida que sería análoga a la covarianza en el caso de la dependencia lineal, con la salvedad de que ahora el signo es siempre positivo y no es informativo (lo cual es lógico, en el caso del análisis de la existencia de correlación los datos pueden apartarse de la hipótesis de incorrelación únicamente en dos sentidos: o bien la correlación es positiva o bien es negativa, con lo que se puede diseñar un estadístico que asigne un signo diferente en cada uno de estos casos. En el caso de dependencia general, sin embargo, existen infinitas formas en que la función puede apartarse de la independencia).

Las condiciones anteriores suponen restricciones importantes a la hora de elegir la función de pesos, pero no son las únicas. Otras restricciones que afectarán a la elección de $w(t,s)$ tienen que ver con que, del mismo modo que a partir de la covarianza definíamos el coeficiente de correlación de Pearson, también ahora nos interesará definir una medida análoga reescalando $\mathcal{V}^2(X,Y;w)$ de forma que se tenga en cuenta la variabilidad intrínseca de cada variable

$$\mathcal{R}_w := \frac{\mathcal{V}(X,Y;w)}{\mathcal{V}(X;w)\mathcal{V}(Y;w)}, \tag{2.20}$$

donde la notación anterior significa respectivamente

$$\begin{aligned}
&\mathcal{V}(X,Y;w)=\sqrt{\mathcal{V}^2(X,Y;w)},\\
&\mathcal{V}(X;w)=\sqrt{\mathcal{V}^2(X,X;w)}\\
&=\left(\int_{\mathbb{R}^{2p}}|\varphi_{X,X}(t,s)-\varphi_X(t)\varphi_X(s)|^2 w(t,s)dtds\right)^{\frac{1}{2}},\\
&\mathcal{V}(Y;w)=\sqrt{\mathcal{V}^2(Y,Y;w)}\\
&=\left(\int_{\mathbb{R}^{2q}}|\varphi_{Y,Y}(t,s)-\varphi_Y(t)\varphi_Y(s)|^2 w(t,s)dtds\right)^{\frac{1}{2}}.
\end{aligned} \tag{2.21}$$

Igual que en el caso de la correlación lineal de Pearson, $\mathcal{R}_w$ es una medida adimensional, que toma valores entre 0 y 1, e impondremos además que sea invariante ante cambios de escala.

Teniendo en cuenta estas condiciones, se prueba que si $w(t,s)$ es una función integrable, entonces $\mathcal{R}_w$ puede tomar valores arbitariamente próximos a cero en casos de dependencia. La solución entonces es escoger una función de persos $w(t,s)$ no integrable. De este modo, a la vista de la expresión (2.18) es evidente que la integral $\mathcal{V}^2(X,Y;w)$ solo se anulará cuando lo haga el módulo de la diferencia entre la función característica conjunta y el producto de las individuales

$$|\varphi_{X,Y}(t,s)-\varphi_X(t)\varphi_Y(s)|=0, \tag{2.22}$$

es decir, en caso de independencia. Sin embargo, hemos de tener cuidado, pues en caso de dependencia necesitamos que dicha medida $\mathcal{V}^2(X,Y;w)$ esté bien definida, es decir, que tome un valor (positivo) finito.

Los autores proponen como elección natural la siguiente función de pesos que, si bien no es única, veremos que conduce a expresiones sencillas de los estadísticos empíricos

$$w(t,s) := \frac{1}{c_p c_q |t|^{1+p}|s|^{1+q}}, \tag{2.23}$$

donde $|\cdot|$ es la norma euclídea (cuando exista riesgo de confusión especificaremos la dimensión con un subíndice, por ejemplo $|t|_p$) y c_p y c_q son dos constantes dadas por

$$c_d = \frac{\pi^{\frac{1+d}{2}}}{\Gamma\left(\frac{1+d}{2}\right)}. \tag{2.24}$$

La primera aparición de esta función de pesos se remonta a la propuesta de Feuerverger (1993) de construcción de un test basado en la distancia entre funciones de distribución para testar la dependencia bivariante. Para más detalles sobre la justificación de esta elección de la función de pesos se puede consultar Székely y col. (2007). En Bakirov y col. (2006) se puede encontrar otra posible elección de la función de pesos w pero no conduce a una expresión tan interesante del estadístico muestral.

En estas condiciones se prueba que $\mathcal{V}^2(X, Y; w)$ está bien definida (la integral existe) siempre que las variables involucradas tengan media finita[1], o equivalentemente:

$$\mathbb{E}(|X|_p + |Y|_q) < \infty. \tag{2.25}$$

Hemos definido por tanto la norma $\|\cdot\|_w$ asociada al espacio L_2 ponderado de las funciones complejas que toman valores en $\mathbb{R}^{p+q}$, donde mediremos la distancia entre la función característica conjunta de (X, Y) y el producto de las marginales. Hemos visto también que el comportamiento de esta norma satisface las exigencias detalladas siempre que los vectores aleatorios X e Y tengan momento de orden 1 finito. En lo sucesivo, denotaremos por $\|\cdot\| \equiv \|\cdot\|_w$ a dicha norma L_2 ponderada donde w es la función de pesos definida en (2.23).

2.2.2. Covarianza y correlación de distancias

Con lo dispuesto en el apartado anterior estamos en condiciones de definir la covarianza de distancias y la correlación de distancias.

Definición 2.2.1 (Covarianza de distancias)**.** Se define la covarianza de distancias $dCov$ entre dos vectores aleatorios de dimensiones arbitrarias $X \in \mathbb{R}^p$ e $Y \in \mathbb{R}^q$ tales que sus momentos de primer orden son finitos $\mathbb{E}(|X|_p) < \infty$, $\mathbb{E}(|Y|_q) < \infty$ como la magnitud no negativa $\mathcal{V}(X, Y)$ dada por

$$\begin{aligned} \mathcal{V}^2(X, Y) &= \|\varphi_{X,Y}(t, s) - \varphi_X(t)\varphi_Y(s)\|^2 \\ &= \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|\varphi_{X,Y}(t, s) - \varphi_X(t)\varphi_Y(s)|^2}{|t|^{1+p}|s|^{1+q}} dtds. \end{aligned} \tag{2.26}$$

Para aligerar la notación, denotaremos por

$$dw = \frac{1}{c_p c_q |t|^{1+p}|s|^{1+q}} dtds, \tag{2.27}$$

y la definición anterior se puede reescribir como

$$\mathcal{V}^2(X, Y) = \int_{\mathbb{R}^{p+q}} |\varphi_{X,Y}(t, s) - \varphi_X(t)\varphi_Y(s)|^2 dw. \tag{2.28}$$

Igual que sucedía en el caso lineal en (2.6), a partir de la covarianza de distancias se puede definir la varianza de distancias $dVar$ de un vector aleatorio X, que no es más que la covarianza de distancias de un vector consigo mismo:

$$\mathcal{V}^2(X) := \mathcal{V}^2(X, X) = \|\varphi_{X,X}(t, s) - \varphi_X(t)\varphi_X(s)\|^2. \tag{2.29}$$

Respetando de nuevo la analogía con el caso de correlación lineal, definimos a partir de la covarianza de distancias la correlación de distancias.

[1] Esta condición se puede relajar a $\mathbb{E}(|X|_p^\alpha + |Y|_q^\alpha) < \infty$ y trabajaríamos entonces con medidas dependientes de la distancia α. Para más detalles consultar Székely y col. (2007), sección 3.1.

Definición 2.2.2 (Correlación de distancias)**.** Se define la correlación de distancias *dCor* entre dos vectores aleatorios no degenerados de dimensiones arbitrarias $X \in \mathbb{R}^p$ e $Y \in \mathbb{R}^q$ tales que sus momentos de primer orden son finitos $\mathbb{E}(|X|_p) < \infty$, $\mathbb{E}(|Y|_q) < \infty$ como la raíz cuadrada de la siguiente magnitud

$$\mathcal{R}^2(X,Y) = \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}. \tag{2.30}$$

De nuevo es inmediato que el objeto definido es, por definición, siempre no negativo

$$\mathcal{R}^2(X,Y) \geq 0. \tag{2.31}$$

Observación 2.2.3. En el caso de que las variables o vectores aleatorios X e Y sean degenerados se define por separado ya que $\sqrt{\mathcal{V}(X)\mathcal{V}(Y)} = 0$. En esta situación se considera que $\mathcal{R}^2(X,Y) = 0$.

Observación 2.2.4. Las versiones poblacionales de *dCov* y *dCor* pueden definirse sin funciones características, para más detalles puede consultarse el trabajo de Lyons (2013) donde desarrollan la teoría para espacios métricos y caracterizan aquellos para los que la covarianza de distancias caracteriza la independencia: los espacios métricos de tipo fuertemente negativo.

2.2.3. Caso normal bivariante

En el caso de la normal bivariante, como ya hemos dicho, lo que nos interesa es que la correlación de distancias se parezca todo lo posible al coeficiente de correlación lineal de pearson.

En este sentido el resultado más destacado sobre la relación entre ambas magnitudes es el siguiente:

Teorema 2.2.5. *Si X e Y son normales estándar con correlación dada por $\rho = \rho(X,Y)$, entonces*

$$\mathcal{R}(X,Y) \leq |\rho|, \tag{2.32}$$

y la igualdad se tiene en los casos en los que $\rho \in \{-1, 1\}$.

Demostración. La demostración de este resultado se obtiene de manera sencilla sin más que introducir la función característica de las variables aleatorias normales estándar en la definición de correlación de distancias y utilizar el desarrollo en serie de Taylor de la exponencial para ver que $\mathcal{R}_n^2(X,Y)$ se reduce al producto de ρ^2 por un término creciente acotado por 1. □

2.2.4. Estadísticos empíricos

Las magnitudes anteriores, *dCov*, *dVar* y *dCor* han sido escogidas de manera que satisfagan las propiedades exigidas en la Sección 2.2.1 con lo que su buen comportamiento está garantizado. Sin embargo, en la práctica vamos a necesitar trabajar con estadísticos empíricos, es decir, obtener versiones de dichas magnitudes que sean evaluables sobre nuestra muestra.

La principal diferencia entre los estadísticos que obtendremos con respecto a los enfoque clásicos es que en lugar de ser funciones de la muestra serán funciones evaluadas sobre distancias entre elementos de la muestra (imitando el planteamiento teórico). Por este motivo se les conoce como estadísticos dependientes de distancia. Esta diferencia será clave a la hora de comparar objetos aleatorios de distinta naturaleza (siempre que tomen valores en un espacio métrico). Veremos en la sección siguiente que esta idea conecta los estadísticos obtenidos con una teoría más general, la de los Energy Statistics.

Veamos entonces como definir los estadísticos muestrales de la covarianza de distancias y la correlación de distancias. Cabe destacar que para obtener estos estimadores es necesario contar con más de una observación de cada una de las variables involucradas.

Consideremos una muestra aleatoria simple $(\boldsymbol{X}, \boldsymbol{Y}) = \{(X_i, Y_i)\}_{i=1}^n$ de observaciones completas de los vectores aleatorios $X \in \mathbb{R}^p$ e $Y \in \mathbb{R}^q$. Vamos a suponer que nuestras variables se encuentran en espacios euclídeos dotados con la norma euclídea correspondiente a su dimensión $|\cdot|$.

Tomamos la matriz de distancias doblemente centrada para cada una de las variables observadas, esto es, calculamos la distancia entre cada par de observaciones y las centramos por filas y por columnas para obtener una matriz cuya media por filas, por columnas y global es cero. Por ejemplo para la variable $X \in \mathbb{R}^p$ la matriz de distancias doblemente centrada $A = (a_{kl}) \in \mathcal{M}_{n\times n}(\mathbb{R})$ se obtendría del siguiente modo:

$$\begin{aligned} &a_{kl} = |X_k - X_l|_p, \\ &\bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}, \quad \bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl}, \quad \bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}, \\ &A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot} \end{aligned} \tag{2.33}$$

De un modo similar obtendríamos la matriz de distancias doblemente centrada $B = (b_{kl}) \in \mathcal{M}_{n\times n}(\mathbb{R})$ para la variable Y definiendo por $b_{kl} = |Y_k - Y_l|_q$ la distancia en norma euclídea entre las observaciones de Y y procediendo de manera similar.

Contamos entonces con dos matrices cuadradas A, B de dimensión $n \times n$ y obtenemos a partir de ellas la covarianza de distancias empírica.

Definición 2.2.6 (Covarianza de distancias empírica)**.** Se define la covarianza de distancias empírica $\mathcal{V}_n(\boldsymbol{X}, \boldsymbol{Y})$ como la raíz cuadrada del número no negativo siguiente:

$$\mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}, \tag{2.34}$$

siendo A_{kl} y B_{kl} las matrices de distancias doblemente centradas obtenidas a partir de los vectores de observaciones $\boldsymbol{X}$ e $\boldsymbol{Y}$ respectivamente.

Igual que definíamos la varianza de distancias a partir de la covarianza, podemos hacer lo mismo con las magnitudes empíricas, con lo que la varianza de distancias empírica de una variable, por ejemplo de X, se definiría como sigue:

$$\mathcal{V}_n(\boldsymbol{X}) = \sqrt{\mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{X})} = \sqrt{\frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}^2}. \tag{2.35}$$

Es evidente en vista de la definición anterior que la varianza de distancias empírica será cero si y solo si $A_{kl} = 0 \quad \forall k, l = 1, ...n$, y es sencillo ver que esto solo sucede si $a_{kl} = |X_k - X_l|_p = 0$ para todo $k, l = 1, ...n$. Es decir, si todas las observaciones $X_1, ..., X_n$ son idénticas.

Las propiedades expecíficas de la varianza de distancias se resumen en el siguiente enunciado[2].

Teorema 2.2.7 (Propiedades de dVar)**.** *Las siguientes propiedades se cumplen para vectores aleatorios con momentos de primer orden finitos:*

1. $dVar(X) \equiv \mathcal{V}(X) = 0 \quad \Leftrightarrow \quad X = \mathbb{E}[X]$ *casi seguro.*
2. $dVar(a + bCX) = |b|dVar(X)$ *para cualquier* $a \in \mathbb{R}^p$, $b \in \mathbb{R}$ *y* $C \in \mathcal{M}_{p\times p}(\mathbb{R})$.
3. $dVar(X + Y) \leq dVar(X) + dVar(Y)$ *para vectores aleatorios independientes* $X \in \mathbb{R}^p$ *e* $Y \in \mathbb{R}^p$.

De manera similar a lo que sucedía en el caso de la dependencia lineal, podemos también reescalar las covarianzas de distancias empíricas para obtener la correlación de distancias empírica:

[2]La demostración se puede encontrar en Székely y col. (2007), teorema 4.

Definición 2.2.8 (Correlación de distancias empírica)**.** La correlación de distancias empírica $\mathcal{R}_n(\boldsymbol{X}, \boldsymbol{Y})$ para vectores aleatorios no degenerados se define como la raíz cuadrada de la siguiente magnitud:

$$\mathcal{R}_n^2(\boldsymbol{X}, \boldsymbol{Y}) = \begin{cases} \frac{\mathcal{V}_n^2(X,Y)}{\sqrt{\mathcal{V}_n^2(X)\mathcal{V}_n^2(Y)}} & si \quad \mathcal{V}_n^2(\boldsymbol{X})\mathcal{V}_n^2(\boldsymbol{Y}) > 0, \\ 0 & si \quad \mathcal{V}_n^2(\boldsymbol{X})\mathcal{V}_n^2(\boldsymbol{Y}) = 0. \end{cases} \tag{2.36}$$

Es fácil ver que tanto $\mathcal{V}_n^2(X, Y)$ como $\mathcal{R}_n^2(X, Y)$ son invariantes ante movimientos rígidos y $\mathcal{R}_n^2(X, Y)$ es además invariante a cambios de escala como indican en Székely y Rizzo (2017).

Propiedades

En este apartado enunciaremos algunas propiedades generales de los anteriores estadísticos basados en distancias entre observaciones y, más importante, presentaremos los principales resultados que establecen su convergencia al correspondiente parámetro poblacional.

Veamos en primer lugar las propiedades de la covarianza de distancias empírica. Lo primero que podríamos pensar es que, vista la definición de la covarianza de distancias en (2.18)

$$\mathcal{V}^2(X, Y) = \|\varphi_{X,Y}(t, s) - \varphi_X(t)\varphi_Y(s)\|, \tag{2.37}$$

una manera intuitiva de definir la covarianza de distancias empírica podría ser introduciendo en la expresión anterior las funciones características empíricas obtenidas a partir de la muestra observada. Entonces si definimos dichas funciones características del siguiente modo:

$$\begin{aligned} \varphi_X^n(t) &= \frac{1}{n}\sum_{i=1}^{n} e^{i\langle t, X_i\rangle}, \\ \varphi_Y^n(s) &= \frac{1}{n}\sum_{i=1}^{n} e^{i\langle s, Y_i\rangle} \end{aligned} \tag{2.38}$$

y la función característica conjunta empírica como

$$\varphi_{X,Y}^n(t, s) = \frac{1}{n}\sum_{i=1}^{n} e^{i\langle t, X_i\rangle + i\langle s, Y_i\rangle}, \tag{2.39}$$

podríamos definir la covarianza de distancias empírica del siguiente modo:

$$\mathcal{V}_n(X, Y) = \left\|\varphi_{X,Y}^n(t, s) - \varphi_X^n(t)\varphi_Y^n(s)\right\|. \tag{2.40}$$

El primer resultado importante es que la definición de la covarianza de distancias empírica es consistente con esta segunda definición (gracias a la forma en la que hemos definido la norma $\|\cdot\|_w$).

Teorema 2.2.9. *Sea* $(\boldsymbol{X}, \boldsymbol{Y})$ *una muestra de la distribución conjunta de (X,Y) variables aleatorias de dimensión arbitraria, entonces la magnitud* $\mathcal{V}_n^2$ *definida en* (2.18) *cumple que*

$$\mathcal{V}_n^2(X, Y) = \|\varphi_{X,Y}{}^n(t, s) - \varphi_X{}^n(t)\varphi_Y{}^n(s)\|^2. \tag{2.41}$$

Este resultado no solo valida la elección de la función de pesos sino que permite concluir de forma inmediata que $\mathcal{V}_n^2$ es un estimador no negativo. Para más detalles sobre la demostración se puede consultar el Teorema 1 en Székely y col. (2007).

En cuanto a las principales propiedades de la correlación de distancias y su estimador empírico se resumen en el teorema siguiente[3].

[3]Para más detalles sobre la demostración se puede consultar en Székely y col. (2007), Teorema 3.

Teorema 2.2.10 (Propiedades de dCor)**.** *Sean X e Y variables aleatorias de dimensión arbitraria, entonces*

1. *Si $\mathbb{E}(|X|_p + |Y|_q) < \infty$, entonces $0 \leq \mathcal{R} \leq 1$ y*

$$\mathcal{R}(X,Y) = 0 \quad \Leftrightarrow \quad X \text{ e } Y \text{ son independientes.} \tag{2.42}$$

2. $0 \leq \mathcal{R}_n \leq 1$.

3. *Si $\mathcal{R}_n(\boldsymbol{X}, \boldsymbol{Y}) = 1$, entonces existe $a \in \mathbb{R}^q$, $b \in \mathbb{R}, b \neq 0$ y $C \in \mathcal{M}_{p\times q}(\mathbb{R})$ una matriz ortogonal tal que*

$$\boldsymbol{Y} = a + b\boldsymbol{X}C. \tag{2.43}$$

Veamos ahora que, en efecto, los estadísticos muestrales definidos se aproximan al valor teórico para tamaños muestrales elevados. La consistencia de la covarianza de distancias empírica se tiene bajo condiciones muy generales.

Teorema 2.2.11. *Si $\mathbb{E}(|X|_p) < \infty$ y $\mathbb{E}(|Y|_q) < \infty$, entonces se tiene el siguiente resultado de convergencia casi segura:*

$$\mathcal{V}_n(X,Y) \xrightarrow[c.s.]{} \mathcal{V}(X,Y) \quad \Leftrightarrow \quad \lim_{n\to\infty} \mathcal{V}_n(X,Y) = \mathcal{V}(X,Y). \tag{2.44}$$

Como corolario se tiene la convergencia casi segura de la correlación de distancias empírica en las condiciones del teorema anterior:

$$\mathcal{R}_n(X,Y) \xrightarrow[c.s.]{} \mathcal{R}(X,Y) \quad \Leftrightarrow \quad \lim_{n\to\infty} \mathcal{R}_n(X,Y) = \mathcal{R}(X,Y). \tag{2.45}$$

En resumen, tras un proceso de doble centrado sobre la matriz de distancias de las variables involucradas (distinto del habitual ya que trabajamos con las distancias entre elementos sin elevar al cuadrado[4]) el producto escalar por filas de las matrices resultantes adquiere la propiedad de que el caso en el que vale cero tiene una significación especial: se corresponde con el caso de independencia.

El estimador $\mathcal{V}_n^2$ de $dCov^2$ que acabamos de presentar es bajo hipótesis de independencia un V-estadístico, con lo que muchos de los resultado asintóticos probados tanto para él como para el estadístico $n\mathcal{V}_n^2$, que será en el que nos basaremos para contrastar la hipótesis de independencia, pueden deducirse a partir de la teoría asintótica existente para este tipo de estadísticos.

2.2.5. Test de independencia

Por último veamos las propiedades asintóticas del estadístico

$$\frac{n\mathcal{V}_n^2}{S_2}, \tag{2.46}$$

donde

$$S_2 = \frac{1}{n^2} \sum_{k,l=1}^{n} |X_k - X_l|_p \sum_{k,l=1}^{n} |Y_k - Y_l|_q \tag{2.47}$$

ya que en base a estos resultados construiremos un test de independencia consistente contra cualquier alternativa de dependencia tal y como queríamos inicialmente.

Teorema 2.2.12. *Si $\mathbb{E}(|X|_p + |Y|_q) < \infty$, entonces:*

[4]De hecho las propiedades del estimador son válidas elevando las distancias a cualquier magnitud $0 < \alpha < 2$.

1. *Si X e Y son independientes (bajo la hipótesis nula), entonces*

$$\frac{n\mathcal{V}_n^2}{S_2} \underset{d}{\to} Q, \tag{2.48}$$

donde Q es una forma cuadrática no negativa de variables aleatorias centradas y $\mathbb{E}[Q] = 1$.

2. *Si X e Y son dependientes (hipótesis alternativa), entonces*

$$\frac{n\mathcal{V}_n^2}{S_2} \underset{p}{\to} \infty. \tag{2.49}$$

Vemos que con estos resultados tenemos un test que rechaza independencia para valores grandes de $n\mathcal{V}_n^2$ consistente ante cualquier alternativa de dependencia.

Es más, se prueba que el test así construido, es decir, el test que rechaza la hipótesis de independencia cuando

$$\sqrt{\frac{n\mathcal{V}_n^2}{S_2}} \geq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \tag{2.50}$$

es en general conservador y tiene un nivel de significación asintótico de α como máximo.

Otra forma de ajustar la distribución del test es plantear un test de permutaciones en el que se estima el p-valor del estadístico de contraste simulando remuestras a partir de la original bajo la hipótesis nula y evaluando sobre ellas el estadístico de contraste. Este planteamiento será el que herede el test de independencia con datos censurados propuesto por Edelmann y col. (2021) que presentaremos en la Sección 3.3.

2.2.6. Estadísticos muestrales: Energy Statistics

La novedad de la propuesta basada en correlación de distancias con respecto a enfoques previos es que los estadísticos ahora no son funciones evaluadas directamente en la muestra, sino funciones evaluadas sobre una métrica definida en el espacio de muestras de tamaño n. Este importante matiz es el que nos permite evaluar la dependencia entre objetos estadísticos distintos, siempre que pertenezcan a un espacio métrico, es decir, exista una distancia asociada a ellos.

El resultado que vincula a las magnitudes presentadas con la teoría de los energy statistics (o estadísticos de energía) es el siguiente.

Sean (X, Y), (X', Y') y (X'', Y'') tres copias de pares de variables independientes e idénticamente distribuidos, entonces se prueba que

$$\mathcal{V}^2(X, Y) = \mathbb{E}[|X - X'||Y - Y'|] + \mathbb{E}[|X - X'|]\mathbb{E}[|Y - Y'|] - 2\mathbb{E}[|X - X'||Y - Y''|], \tag{2.51}$$

a partir de resultados relacionados con la elección de la función de pesos w utilizando Fubini y las condiciones de independencia que obtenemos como consecuencia de contar con una muestra aleatoria simple.

2.2.7. Relación con U-estadísticos

En esta sección vamos a demostrar que existe un estimador insesgado y consistente de $\mathcal{V}^2(X, Y)$. Veremos que es un $\mathcal{U}$-estadístico, lo que nos permitiría utilizar la teoría asintótica desarrollada para probar algunos resultados ya vistos. La razón por la que hacemos hincapié en esta relación, además de su evidente interés, es que nos será útil a la hora de adaptar los estadísticos de correlación de distancias y covarianza de distancias al contexto de datos censurados.

Comenzaremos por obtener una versión insesgada del estimador del cuadrado de la covarianza de distancias siguiendo el desarrollo propuesto en Huo y Székely (2016). Para introducirla necesitamos algunas definiciones previas ya que el proceso de doble centrado de distancias que utilizamos para construir el estimador presentado en la Definición 2.2.1 será ahora sustituido por el $\mathcal{U}$-centrado de las matrices de distancias involucradas.

Definición 2.2.13 (Matriz $\mathcal{U}$-centrada)**.** Sea $A = (a_{ij}) \in \mathcal{M}_{n\times n}(\mathbb{R}), n \geq 2$ una matriz simétrica, cuadrada y cuyos elementos de la diagonal son cero. Se define la matriz $\mathcal{U}$-centrada $\tilde{A}$ como la matriz cuyos elementos $\tilde{A}_{ij}$ vienen dados por la siguiente expresión:

$$\tilde{A}_{ij} = \begin{cases} a_{ij} - \frac{1}{n-2}\sum_{l=1}^{n} a_{il} - \frac{1}{n-2}\sum_{k=1}^{n} a_{kj} + \frac{1}{(n-1)(n-2)}\sum_{k,l=1}^{n} a_{kl}, & si \quad i \neq j \\ \\ 0, & si \quad i = j \end{cases} \tag{2.52}$$

En nuestro caso, dada la muestra $(\boldsymbol{X}, \boldsymbol{Y}) = \{X_i, Y_i\}_{i=1}^{n}$ de los vectores aleatorios $X \in \mathbb{R}^p$ e $Y \in \mathbb{R}^q$, consideramos sus matrices de distancias A y B en norma euclídea. Estas matrices se encuentran en las hipótesis exigidas por la definición anterior, con lo que podemos calcular sus versiones $\mathcal{U}$-centradas: $\tilde{A}$ y $\tilde{B}$. Tenemos entonces el resultado enunciado en la siguiente proposición.

Proposición 2.2.14. *En las condiciones anteriores, si* $\mathbb{E}(|X|_p + |Y|_q) < 1$*, para cualquier* $n > 3$ *la magnitud*

$$(\tilde{A} \cdot \tilde{B}) := \frac{1}{n(n-3)} \sum_{i\neq j} \tilde{A}_{ij}\tilde{B}_{ij} \tag{2.53}$$

es un estimador insesgado del cuadrado de la covarianza de distancias $\mathcal{V}^2(X, Y)$*.*

La demostración del resultado anterior se puede encontrar en Székely y Rizzo (2014).

Pasamos a trabajar por tanto con el siguiente estimador, que denotaremos por Ω_n:

$$\Omega_n := (\tilde{A} \cdot \tilde{B}), \tag{2.54}$$

y se prueba que se puede reexpresar en los siguientes términos

$$\Omega_n = \frac{1}{n(n-3)} \sum_{i\neq j} a_{ij}b_{ij} - \frac{2}{n(n-2)(n-3)} \sum_{i=1}^{n} a_{i\cdot}b_{i\cdot} + \frac{1}{n(n-1)(n-2)(n-3)} a_{\cdot\cdot}b_{\cdot\cdot} \tag{2.55}$$

Utilizaremos esta expresión de Ω_n para probar que es un $\mathcal{U}$-estadístico. Para ello utilizaremos el teorema 2.2.16, que establece una condición necesaria y suficiente para probar que un estadístico es $\mathcal{U}$-estadístico: la invariancia jackknife.

Definición 2.2.15 (Invariancia jackknife)**.** Sea $r \geq 1$ un entero arbitrario y $f_n : \mathbb{R}^n \longrightarrow \mathbb{R}$ un estadístico definido sobre cualquier muestra $(x_1, ..., x_n)$ con $n \geq r$. Sea f_{n-1}^i el estadístico evaluado sobre la muestra jackknife, obtenida eliminando la i-ésima observación de la muestra original (con $f_0^i := f_1(x_1)$). f_n es jackknife invariante de orden r si

$$n \cdot f_n(x_1, ..., x_n) = \sum_{i=1}^{n} f_{n-1}^i(x_1, ..., x_n) \tag{2.56}$$

para todo $n \geq r$ siendo r el menor entero positivo para el que se cumple esta propiedad.

Cabe notar que esta propiedad establece una relación entre los valores del estadístico evaluado sobre muestras de diferente tamaño.

El siguiente teorema establece que una condición necesaria y suficiente para que Ω_n sea un $\mathcal{U}$-estadístico de orden r es que Ω_n sea jackknife invariante de orden r.

Teorema 2.2.16. *$\Omega_n(x_1, ..., x_n)$ es un $\mathcal{U}$-estadístico de orden $r \in \mathbb{N}$ si, y solo si se cumple la siguiente igualdad para todo $n \geq r$*

$$n \cdot \Omega_n(x_1, ..., x_n) = \sum_{i=1}^{n} \Omega^i_{n-1}(x_1, ..., x_n) \tag{2.57}$$

Los detalles de la demostración se pueden consultar en Huo y Székely (2016), teorema 3.1.

Finalmente, estamos en condiciones de probar que, en efecto, Ω_n es un $\mathcal{U}$-estadístico.

Teorema 2.2.17. *El estimador Ω_n es un $\mathcal{U}$-estadístico con función kernel (o función núcleo) igual al producto interior definido en* (2.53) *evaluado con $n = 4$.*

Demostración. La demostración se hará recurriendo al Teorema 2.2.16. Por tanto, lo primero que debemos hacer es hallar una expresión para el estadístico evaluado sobre la muestra reducida $\Omega^i_{n-1} = \Omega_{n-1}(x_1, ...x_{i-1}, x_{i+1}, ..., x_n)$.

Para ello basta aplicar la fórmula (2.55), con lo que si para cada $1 \leq i \leq n$, con $n > 4$ denotamos por $a^i_{k\cdot}, b^i_{k\cdot}, a^i_{\cdot\cdot}$ y $b^i_{\cdot\cdot}$ a las sumas correspondientes sobre la muestra reducida, obtendríamos:

$$\begin{aligned}\Omega^i_{n-1} = \frac{1}{(n-1)(n-4)} \sum_{k\neq j, k\neq i, j\neq i} a_{kj}b_{kj} \quad - \quad \frac{2}{(n-1)(n-3)(n-4)} \sum_{k=1, k\neq i}^{n} a^i_{k\cdot}b^i_{k\cdot} \\ + \frac{1}{(n-1)(n-2)(n-3)(n-4)} a^i_{\cdot\cdot}b^i_{\cdot\cdot}.\end{aligned} \tag{2.58}$$

Por tanto, teniendo en cuenta las siguientes igualdades

$$\begin{aligned} a^i_{k\cdot} &= a_{k\cdot} - a_{ik}, \\ b^i_{k\cdot} &= b_{k\cdot} - b_{ik}, \\ a^i_{\cdot\cdot} &= a_{\cdot\cdot} - 2a_{\cdot k}, \\ b^i_{\cdot\cdot} &= b_{\cdot\cdot} - 2b_{\cdot k}. \end{aligned} \tag{2.59}$$

entonces

$$\begin{aligned} \sum_{k=1, k\neq i}^{n} a^i_{k\cdot}b^i_{k\cdot} &= \sum_{k=1, k\neq i}^{n} (a_{k\cdot} - a_{ki})(b_{k\cdot} - b_{ki}), \\ a^i_{\cdot\cdot}b^i_{\cdot\cdot} &= \sum_{i=1}^{n} (a_{\cdot\cdot} - 2a_{\cdot i})(b_{\cdot\cdot} - 2b_{\cdot i}) \end{aligned} \tag{2.60}$$

y teniendo en cuenta que

$$\sum_{k\neq j, k\neq i, j\neq i} a_{kj}b_{kj} = (n-2) \sum_{k\neq j} a_{kj}b_{kj}, \tag{2.61}$$

tendríamos el resultado buscado. □

La principal ventaja de utilizar un $\mathcal{U}$-estadístico es que se pueden emplear los teoremas clásicos de Hoeffding (1948) entre otros para obtener sus distribuciones asintóticas.

Para trabajar posteriormente con el estimador consistente de la covarianza de distancias al cuadrado es conveniente conocer la expresión explícita de su núcleo $h((X_1, Y_1), (X_2, Y_2), (X_3, Y_3), (X_4, Y_4))$.

En Huang y Huo (2017) obtienen la expresión de este núcleo para un estimador insesgado y consistente de $\mathcal{V}^2_n(X, Y)$ que tenga la siguiente forma:

$$\hat{\Omega} = \frac{n}{n-3}T_1 + \frac{n^3}{(n-1)(n-2)(n-3)}T_2 - \frac{2n^2}{(n-2)(n-3)}T_3, \tag{2.62}$$

donde

$$\begin{aligned} T_1 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ij}, \\ T_2 &= \frac{1}{n^4}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}\right), \\ T_3 &= \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} a_{ij}b_{ik}. \end{aligned} \tag{2.63}$$

Basta con que probemos entonces que $\Omega_n = \hat{\Omega}$.

Demostración. En cuanto al primer sumando en (2.54) es suficiente notar que $a_{ii} = b_{ii} = 0$ para cualquier $i \in \{1, ...n\}$ por ser las matrices $A = (a_{ij})$ y $B = (b_{ij})$ matrices de distancias. Por tanto es evidente que:

$$\frac{1}{n-3}\sum_{i\neq j}^{n} a_{ij}b_{ij} = \frac{1}{n-3}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ij} = \frac{n}{n-3}T_1. \tag{2.64}$$

Para el segundo sumando basta con utilizar las definiciones de las sumas por columnas y las propiedades de los sumatorios:

$$\begin{aligned} \frac{2}{n(n-2)(n-3)}\sum_{i=1}^{n} a_{i\cdot}b_{i\cdot} &= \frac{2}{n(n-2)(n-3)}\sum_{i=1}^{n}\left(\sum_{j=1}^{n} a_{ij}\sum_{k=1}^{n} b_{ik}\right) \\ &= \frac{2n^2}{(n-2)(n-3)}\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} a_{ij}b_{ik} \\ &= \frac{2n^2}{(n-2)(n-3)}T_3. \end{aligned} \tag{2.65}$$

Y por último, para el tercer sumando nos basta de nuevo con utilizar las definiciones de los sumatorios dobles $a_{\cdot\cdot}$ y $b_{\cdot\cdot}$:

$$\begin{aligned} \frac{a_{\cdot\cdot}b_{\cdot\cdot}}{n(n-1)(n-2)(n-3)} &= \frac{1}{n(n-1)(n-2)(n-3)}\frac{n^3}{n^3}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\right)\left(\sum_{i=1}^{n}\sum_{k=1}^{n} b_{ik}\right) \\ &= \frac{n^3}{n(n-1)(n-2)(n-3)}T_3. \end{aligned} \tag{2.66}$$

□

Por tanto, estamos en las condiciones especificadas en Huang y Huo (2017) y se tiene que el kernel del $\mathcal{U}$-estadístico $\hat{\Omega}$ sería:

$$\begin{aligned} &h_4((X_1,Y_1),(X_2,Y_2),(X_3,Y_3),(X_4,Y_4)) \\ &= \frac{1}{4}\sum_{\substack{1\leq i,j\leq 4 \\ i\neq j}} a_{ij}b_{ij} - \frac{1}{4}\sum_{i=1}^{4}\left(\sum_{j=1,j\neq i}^{4} a_{ij}\sum_{j=1,j\neq i}^{4} b_{ij}\right) + \frac{1}{24}\sum_{\substack{1\leq i,j\leq 4 \\ i\neq j}} a_{ij}\sum_{\substack{1\leq i,j\leq 4 \\ i\neq j}} b_{ij}, \end{aligned} \tag{2.67}$$

que vemos que coincide con la definición de Ω_n cuando $n = 4$.

Muchas de las expresiones detalladas en este apartado serán utilizadas en el capítulo siguiente cuando veamos como adaptar la correlación de distancias al contexto de datos censurados.

Capítulo 3

Correlación de distancias en presencia de censura

En esta sección vamos a tratar de extender las nociones de covarianza de distancias y correlación de distancias para medir la dependencia entre un vector de covariables $\boldsymbol{X} \in \mathbb{R}^p$ y la variable tiempo de supervivencia $Y \in \mathbb{R}^+$ sujeta a censura aleatoria por la derecha.

Este tipo de contextos es habitual en Análisis de Supervivencia, donde a menudo se intenta establecer la relación entre una serie de factores de interés y un tiempo de supervivencia. Por ejemplo, podría interesarnos establecer la relación entre la edad, el peso, el tabaquismo... y la supervivencia ante un determinado tipo de cáncer. Además de las dificultades intrínsecas de tratar con datos de supervivencia vistas en el Capítulo 1, en algunas aplicaciones se suma que los datos pueden ser de alta dimensión o contener información sobre covariables multidimensionales. Por ejemplo, en genética puede interesarnos detectar qué localizaciones del genoma, de entre decenas de miles, son relevantes a la hora de desarrollar una patología concreta y hacerlo utilizando datos de apenas unos cientos o miles de pacientes.

Como ejemplos en los que ha sido relevante estudiar la dependencia en contextos de Análisis de Supervivencia tenemos los trabajos de Edelmann y col. (2020), Li y col. (2012), Hua y col. (2015) o Kong y col. (2012) entre otros

En la Sección 1.4 hemos mencionado algunas de las técnicas clásicas para lidiar con la presencia de covariables en situaciones con respuesta censurada y nos hemos detenido especialmente en sus limitaciones. Los primeros intentos de desarrollar herramientas más generales se pueden encontrar en Peng y Fine (2008) y Wood (2013).

El desarrollo en el que nos centraremos se puede encontrar en Edelmann y col. (2021). En él se presentar una herramienta general para testar relaciones genéricas entre variables bajo censura aleatoria por la derecha. Se obtendrá una versión empírica e insesgada de la correlación de distancias evaluable sobre muestras censuradas que satisface las propiedades necesarias para poder derivar tests de independencia válidos para detectar cualquier tipo de relación entre el tiempo de supervivencia y las covariables.

El esquema general en base al que vamos a presentar los resultados es el siguiente: comenzaremos viendo como adaptar un $\mathcal{U}$-estadístico a la presencia de censura presentando los IPCW $\mathcal{U}$-estadísticos propuestos por Datta y col. (2010). Una vez hayamos comprobado que esta versión conserva las principales propiedades del $\mathcal{U}$-estadístico original, en particular su esperanza, pasaremos a aplicar este procedimiento sobre el estimador insesgado $\hat{\Omega}$ de $\mathcal{V}_n^2$ definido en (2.62), obteniendo el estimador que llamaremos $\hat{\Omega}_n^C$. A partir de esa primera versión del estimador derivaremos otra más eficiente computacionalmente y con ella construiremos un test de independencia basado en permutaciones que nos permita detectar relaciones entre la respuesta censurada y las covariables.

3.1. IPCW U-estadístico

Dado un $\mathcal{U}$-estadístico definido para datos completos en presencia de covariables, veamos como podemos extenderlo al caso en el que una de las covariables presenta censura aleatoria por la derecha.

Supongamos que tenemos un $\mathcal{U}$-estadístico genérico de orden m definido sobre la muestra de datos completos $\{X_i, Y_i\}_{i=1}^n$:

$$U_n = \binom{n}{m}^{-1} \sum_{1\leq i_1<\ldots<i_m\leq n} h_m((X_{i_1}, Y_{i_1}), ..., (X_{i_m}, Y_{i_m})), \tag{3.1}$$

donde h_m es la función kernel simétrica asociada. Supongamos además que U_n es un estimador insesgado de la magnitud teórica θ.

Consideremos ahora una muestra con censura aleatoria por la derecha sobre la variable tiempo de fallo $\{(X_i, T_i, \delta_i)\}_{i=1}^n$ de acuerdo con (1.12). Asumiremos que la censura es independiente (no informativa) y que la razón de riesgo de la variable censura

$$\lambda_C(t) = \lim_{h\to 0} \frac{\mathbb{P}(t \leq C \leq t+h | C \geq t)}{h} \tag{3.2}$$

es absolutamente continua.

En estas circunstancias, lo que nos interesa es definir un estadístico que

1. Incorpore la censura aleatoria, de manera que esté definido sobre la muestra obtenida bajo este supuesto.
2. Conserve su media, de manera que si el $\mathcal{U}$-estadístico es un estimador insesgado del parámetro θ, el nuevo IPCW $\mathcal{U}$-estadístico también lo sea.
3. Sería conveniente que fuera también un $\mathcal{U}$-estadístico para poder aprovechar toda la teoría desarrollada sobre ellos.

Veamos que todos esto se consigue con el IPCW $\mathcal{U}$-estadístico propuesto por Datta y col. (2010), que se construye utilizando una reponderación que conserva la media. Veamos como funciona esta reponderación tal y como se detalla en Datta (2005).

La idea clave es observar que la esperanza de los tiempos observados ponderados por la inversa de la función de supervivencia en cada tiempo T_i de fallo es justo la esperanza de la variable de interés Y:

$$\mathbb{E}\left[\frac{\delta T}{S_C(T^-)}\right] = \mathbb{E}\left[\mathbb{E}\left(\frac{\delta Y}{S_C(Y^-)}\right)\right] = \mathbb{E}\left[\mathbb{E}\left(\frac{\delta Y}{S_C(Y^-)}|Y\right)\right], \tag{3.3}$$

donde estamos utilizando que si $\delta = 1$ entonces $T = Y$ en la primera igualdad. Entonces, sin más que aplicar la definición de δ en (1.12) vemos que

$$\mathbb{E}\left[\mathbb{E}\left(\frac{\delta Y}{S_C(Y^-)}|Y\right)\right] = \mathbb{E}\left[\frac{Y}{S_C(Y^-)}\mathbb{E}\left(\mathbb{I}(C \geq Y|Y)\right)\right]. \tag{3.4}$$

Y teniendo en cuenta que

$$\mathbb{E}\left(\mathbb{I}(C \geq Y|Y)\right) = \mathbb{P}[\mathbb{I}(C \geq Y|Y) = 1] = \mathbb{P}[C \geq Y] = S_C(Y) \tag{3.5}$$

y que como la distribución G es absolutamente continua

$$S_C(Y^-) = S_C(Y), \tag{3.6}$$

obtendríamos que

$$\mathbb{E}\left(\mathbb{I}(C \geq Y|Y)\right) = \mathbb{E}\left[\frac{Y}{S_C(Y)}S_C(Y)\right] = \mathbb{E}(Y). \tag{3.7}$$

Recapitulando acabamos de ver que en efecto

$$\mathbb{E}\left[\frac{\delta T}{S_C(T^-)}\right] = \mathbb{E}(Y). \tag{3.8}$$

Esto nos permite aproximar la media de los tiempos observados $\{Y_i\}_{i=1}^n$ como un promedio ponderado de los tiempos observados

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \frac{\delta_i}{S_C(T_i)} T_i, \tag{3.9}$$

y será posible trasladar esta idea al caso de nuestro $\mathcal{U}$-estadístico $\hat{\Omega}$.

De esta forma, utilizando el proceso de reponderación anterior, podemos definir entonces una versión IPCW del $\mathcal{U}$-estadístico U_n definido en (3.1) de manera que siga siendo estimador insesgado de θ y además esté evaluado sobre las variables observables (T, δ) en lugar de Y:

$$U_n^c = \binom{n}{m}^{-1} \sum_{1\leq i_1<\ldots<i_m\leq n} \frac{\left(\prod_{k\in\{i_1,\ldots i_m\}} \delta_k\right) h_m((X_{i_1}, Y_{i_1}), ..., (X_{i_m}, Y_{i_m}))}{\prod_{k\in\{i_1,\ldots i_m\}} S_C(T_k{}^-)}. \tag{3.10}$$

No solo eso, sino que es inmediato que este estimador es a su vez un $\mathcal{U}$-estadístico también de orden m cuya función núcleo o kernel evaluado sobre la muestra observada $\{(X_i, T_i, \delta_i)\}_{i=1}^n$ viene dado por

$$\mathcal{H}((X_{i_1}, T_{i_1}, \delta_{i_1}), ..., (X_{i_m}, T_{i_m}, \delta_{i_m})) = \frac{\left(\prod_{k\in\{i_1,\ldots i_m\}} \delta_k\right) h(X_{i_1}, ..., X_{i_m})}{\prod_{k\in\{i_1,\ldots i_m\}} S_C(T_k{}^-)}, \tag{3.11}$$

donde nuevamente trabajar con $T_k{}^-$ en vez de con T_k solamente afecta en caso de empates.

En la práctica, como no conocemos la función de supervivencia de la variable tiempo de censura S_C tendremos que estimarla utilizando el estimador de Kaplan-Meier correspondiente, que se define igual que en (1.47) pero intercambiando el papel de las observaciones censuradas y no censuradas:

$$\hat{S}_C(t) = \prod_{T_i\leq t} \left(1 - \frac{c_i}{n_i}\right), \tag{3.12}$$

donde ahora c_i es la cantidad de observaciones censuradas en el instante T_i (bajo supuestos de no empate: 0 o 1).

Por tanto, el IPCW $\mathcal{U}$-estadístico para datos con censura aleatoria por la derecha será

$$\hat{U}_n^c = \binom{n}{m}^{-1} \sum_{1\leq i_1<\ldots<i_m\leq n} \frac{\left(\prod_{k\in\{i_1,\ldots i_m\}} \delta_k\right) h_m((X_{i_1}, T_{i_1}), ..., (X_{i_m}, T_{i_m}))}{\prod_{k\in\{i_1,\ldots i_m\}} \hat{S}_C(T_k{}^-)} \tag{3.13}$$

y no es técnicamente hablando un $\mathcal{U}$-estadístico ya que para evaluarlo sobre cualquier combinación $\{i_1, ...i_m\}$ necesitaremos estimar $\hat{S}_C$ y para ello utilizamos todas las observaciones de la muestra.

El hecho de que esta expresión sea un promedio ponderado facilita el estudio de sus propiedades asintóticas. Así se obtendría fácilmente su insesgadez, su normalidad asintótica y otras propiedades que utilizaremos a continuación. Para más detalles se puede consultar Datta y col. (2010).

3.2. Estimador IPCW de la covarianza de distancias

Partimos entonces del estimador insesgado de la covarianza de distancias al cuadrado $\hat{\Omega}$, que de acuerdo a lo presentado en el capítulo anterior será un $\mathcal{U}$-estadístico de orden 4 de la forma

$$\hat{\Omega} = \binom{n}{4}^{-1} \sum_{1\leq i_1<\ldots<i_4\leq n} \frac{\delta_{i_1}}{\hat{S}_C(T_{i_1}{}^-)} \cdot \ldots \cdot \frac{\delta_{i_4}}{\hat{S}_C(T_{i_4}{}^-)} h((X_{i_1}, Y_{i_1}), ..., (X_{i_4}, Y_{i_4})), \tag{3.14}$$

donde la función kernel h está definida en (2.67). Su versión IPCW pasará a ser, de acuerdo con el desarrollo que acabamos de presentar

$$\hat{\Omega}^C = \binom{n}{4}^{-1} \sum_{1\leq i_1<\ldots<i_4\leq n} \frac{\delta_{i_1}}{\hat{S}_C(T_{i_1}{}^-)} \cdot \frac{\delta_{i_2}}{\hat{S}_C(T_{i_2}{}^-)} \cdot \frac{\delta_{i_3}}{\hat{S}_C(T_{i_3}{}^-)} \cdot \frac{\delta_{i_4}}{\hat{S}_C(T_{i_4}{}^-)} h((X_{i_1},T_{i_1}),...,(X_{i_4},T_{i_4})). \tag{3.15}$$

Por el método de reponderación empleado sabemos entonces que $\hat{\Omega}^C$ sigue siendo un estimador insesgado y consistente para $\mathcal{V}^2(X,Y)$.

3.2.1. Reducción de la complejidad computacional

La expresión anterior tiene una complejidad computacional $\mathcal{O}(n^4)$, lo que supone que en la práctica no es empleable. A continuación vemos que es posible reescribir la fórmula anterior de manera que sea calculable en tiempo $\mathcal{O}(n^2)$, lo que haría a este estadístico competitivo en la implementación práctica. Para ello necesitamos algunos resultados preliminares que pasamos a presentar.

Si consideramos la matrix de distancias $D = (d_{ij})_{i,j=1}^n$ donde $d_{ij} = |T_i - T_j|$, el kernel evaluado sobre $\{(X_i,T_i)\}_{i=1}^4$ quedaría, utilizando su definición en (2.67):

$$\begin{aligned} h((X_1,T_1),(X_2,T_2),(X_3,T_3),(X_4,T_4)) &= \frac{1}{4} \sum_{\substack{1\leq i,j\leq 4 \\ i\neq j}} a_{ij}d_{ij} \\ - \frac{1}{4}\sum_{i=1}^4 \left(\sum_{j=1,j\neq i}^4 a_{ij} \sum_{j=1,j\neq i}^4 d_{ij} \right) &+ \frac{1}{24} \sum_{\substack{1\leq i,j\leq 4 \\ i\neq j}} a_{ij} \sum_{\substack{1\leq i,j\leq 4 \\ i\neq j}} d_{ij}, \end{aligned} \tag{3.16}$$

El primer lema que nos ocupa nos permite reexpresar el kernel en términos de una colección de subíndices $\{i,j,k,l\}$.

Lema 3.2.1. *La función kernel descrita en* (2.67) *se puede reexpresar del siguiente modo:*

$$\begin{aligned} &h_4((X_1,Y_1),(X_2,Y_2),(X_3,Y_3),(X_4,Y_4)) \\ &= \sum_{\substack{1\leq i,j\leq 4 \\ i\neq j}} a_{ij}b_{ij} \quad - \quad \frac{1}{12} \sum_{\substack{1\leq i,j,k\leq 4 \\ i,j,k \text{ distintos}}} a_{ij}b_{ik} \quad + \quad \frac{1}{24} \sum_{\substack{1\leq i,j,k,l\leq 4 \\ i,j,k,l \text{ distintos}}} a_{ij}b_{ik}, \end{aligned} \tag{3.17}$$

donde $a_{ij} = \|X_i - X_j\|_p$ *y* $b_{ij} = \|Y_i - Y_j\|_q$ *para cualquier par* $i,j \in \{1,2,...,n\}$.

Demostración. En primer lugar es evidente que la expresión (2.67) se puede reescribir en los siguientes términos:

$$\begin{aligned} &h_4((X_1,Y_1),(X_2,Y_2),(X_3,Y_3),(X_4,Y_4)) \\ &= \frac{1}{4} \sum_{\substack{1\leq i,j\leq 4 \\ i\neq j}} a_{ij}b_{ij} - \frac{1}{4} \sum_{\substack{1\leq i,j,k\leq 4 \\ i\neq j,\, j\neq k}} a_{ij}b_{ik} + \frac{1}{24} \sum_{\substack{1\leq i,j,k,l\leq 4 \\ i\neq j,\, l\neq k}} a_{ij}b_{kl}, \end{aligned} \tag{3.18}$$

Fijémonos en el último sumando. Notemos que tomar subíndices $\{1\leq i,j,k,l\leq 4 | i\neq j, k\neq l\}$ equivale a tomar cuatro números en el conjunto $\{1,2,3,4\}$ con reemplazamiento de manera que los dos primeros no sean iguales, el par (i,j), y los dos últimos tampocos, el par (k,l). Esto reduce las posibilidades a tres opciones mutuamente excluyentes:

- Los cuatro índices i,j,k,l son distintos.

- Los pares (i,j) y (k,l) tienen solo un elemento en común de manera que las opciones se podrían expresar en términos de tres subíndices del siguiente modo:

$$(i,j,k,l) \in \{(i,j,i,k),(i,j,k,i),(i,j,j,k),(i,j,k,j)\},$$

  con i,j,k distintos entre sí.

- Los pares (i,j) y (k,l) tienen dos elementos en común, de manera que $k,l \in \{(i,j),(j,i)\}$.

Por tanto y como los subconjuntos descritos son evidentemente disjuntos, el el último sumando de la expresión anterior se puede reescribir como

$$\begin{aligned}\sum_{\substack{1\leq i,j,k,l\leq 4\\ i\neq j, l\neq k}} a_{ij}b_{kl} &= \sum_{\substack{1\leq i,j,k,l\leq 4\\ \text{todos distintos}}} a_{ij}b_{kl} + \sum_{\substack{1\leq i,j,k\leq 4\\ \text{todos distintos}}} (a_{ij}b_{ik}+a_{ij}b_{ki}+a_{ij}b_{jk}+a_{ij}b_{kj}) \\ &\quad + \sum_{\substack{1\leq i,j\leq 4\\ \text{distintos}}} (a_{ij}b_{ij}+a_{ij}b_{ji}),\end{aligned} \tag{3.19}$$

y aplicando ahora argumentos de simetría obtenemos:

$$\sum_{\substack{1\leq i,j,k,l\leq 4\\ i\neq j, l\neq k}} a_{ij}b_{kl} = \sum_{\substack{1\leq i,j,k,l\leq 4\\ \text{todos distintos}}} a_{ij}b_{kl} + 4\sum_{\substack{1\leq i,j,k\leq 4\\ \text{todos distintos}}} a_{ij}b_{jk} + 2\sum_{\substack{1\leq i,j\leq 4\\ \text{distintos}}} a_{ij}b_{ij}. \tag{3.20}$$

Procediendo de manera similar para el segundo sumando de (3.18), basta observar que la selección de subíndices $(i,j,k) \in \{1\leq i,j,k\leq 4 | i\neq j, j\neq k\}$ equivale a seleccionar tres números (i,j,k) del conjunto $1,2,3,4$ con reemplazamiento de manera que el primero es distinto del segundo y el segundo es distinto del tercero. Es evidente que esto nos deja únicamente dos opciones que se excluyen mutuamente:

- Los tres subíndices (i,j,k) son diferentes.

- El primero es igual al tercero y solo difiere el segundo: $(i,j,k)=(i,j,i)$.

Por tanto, podemos reexpresar:

$$\sum_{\substack{1\leq i,j,k\leq 4\\ i\neq j, j\neq k}} a_{ij}b_{ik} = \sum_{\substack{1\leq i,j,k\leq 4\\ \text{todos distintos}}} a_{ij}b_{ik} + \sum_{\substack{1\leq i,j\leq 4\\ i\neq j}} a_{ij}b_{ij}. \tag{3.21}$$

Sustituyendo (3.21) y (3.20) en (3.18), obtenemos la expresión buscada:

$$\begin{aligned}&h_4((X_1,Y_1),(X_2,Y_2),(X_3,Y_3),(X_4,Y_4)) \\ &= \frac{1}{24}\left[2\sum_{\substack{1\leq i,j\leq 4\\ i\neq j}} a_{ij}b_{ij} - 2\sum_{\substack{1\leq i,j,k\leq 4\\ i,j,k \text{ distintos}}} a_{ij}b_{ik} + \sum_{\substack{1\leq i,j,k,l\leq 4\\ i,j,k \text{ distintos}}} a_{ij}b_{ik}\right]\end{aligned} \tag{3.22}$$

□

Para mayor simplicidad se puede demostrar que, si denotamos por

$$\begin{aligned} W_i &= \frac{\delta_i}{\hat{S}(T_i^-)} \quad \text{(peso asignado a } T_i\text{)}, \\ a_{ij} &= \|X_i - X_j\|, \\ d_{ij} &= |T_i - T_j|, \\ m &:= \sum_{i=1}^{n} W_i, \\ M &:= \sum_{i=1}^{n} {W_i}^2, \end{aligned} \tag{3.23}$$

el estimador $\hat{\Omega}^c$ admite la siguiente expresión de complejidad $\mathcal{O}(n^2)$

$$\begin{aligned} \tilde{\Omega} = \frac{1}{n(n-1)(n-2)(n-3)} &\left(\sum_{i,j=1}^{n} W_i W_j a_{ij} d_{ij} (m^2 - m(W_i + W_j) - M + 2W_i W_j) \right) \\ &- 2\sum_{i=1}^{n} W_i (m + W_i) \left(\sum_{j=1}^{n} W_j a_{ij} \right) \left(\sum_{j=1}^{n} W_j d_{ij} \right) + 2\sum_{i=1}^{n} W_i \left(\sum_{j=1}^{n} {W_j}^2 a_{ij} \right) \left(\sum_{j=1}^{n} W_j d_{ij} \right) \\ &+ 2\sum_{i=1}^{n} W_i \left(\sum_{j=1}^{n} W_j a_{ij} \right) \left(\sum_{j=1}^{n} {W_j}^2 d_{ij} \right) + \left(\sum_{i,j=1}^{n} W_i W_j a_{ij} \right) \left(\sum_{i,j=1}^{n} W_i W_j d_{ij} \right). \end{aligned} \tag{3.24}$$

Para detalles sobre la demostración se puede consultar la información complementaria de Edelmann y col. (2021).

Por último, presentamos resultados relativos a las buenas propiedades asintóticas de $\tilde{\Omega}$. ESalgunas condiciones técnicas relativas a:

1. La existencia de momentos de segundo orden de las covariables y del tiempo de supervivencia:

$$\mathbb{E}[\|X\|^2] < \infty, \quad \mathbb{E}[|Y|^2] < \infty, \tag{3.25}$$

   lo que garantiza la existencia del momento de segundo orden de la función kernel y, por tanto, la normalidad asintótica del estimador $\tilde{\Omega}$ en ausencia de censura.

2. Condiciones sobre la existencia de momentos de segundo orden de la función kernel ponderada, garantizando así la normalidad asintótica del estimador $\tilde{\Omega}$ en presencia de censura.

3. La tercera condición atañe a la función de pesos y garantiza la normalidad asintótica cuando en lugar de utilizar la verdadera función de supervivencia de C utilizamos su estimador de Kaplan-Meier $\hat{S}_C$.

Bajo dichas condiciones tenemos el siguiente resultado sobre la distribución asintótica del IPCW estimador de $\mathcal{V}^2(X, Y)$.

Teorema 3.2.2. *Bajo ciertas condiciones técnicas, siendo*

$$\sigma^2 = 16Var\left(\frac{h_1(X,T)\delta}{S_C(T^-)} + \int_{[0,\infty)} w(t) dM^C(t) \right), \tag{3.26}$$

entonces, para $n \to \infty$,

$$\tilde{\Omega} \underset{P}{\to} \mathcal{V}^2(X, Y). \tag{3.27}$$

Si además $\sigma^2 > 0$ *entonces*

$$\sqrt{n}(\tilde{\Omega} - \mathcal{V}^2(X,Y)) \underset{D}{\longrightarrow} \mathcal{N}(0,\sigma^2), \tag{3.28}$$

si $\sigma^2 = 0$*, entonces*

$$\sqrt{n}(\tilde{\Omega} - \mathcal{V}^2(X,Y)) \underset{P}{\longrightarrow} 0. \tag{3.29}$$

En el caso de que X e Y sean independientes se cumple que $\sigma^2 = 0$ y, por tanto,

$$\tilde{\Omega} \underset{P}{\longrightarrow} 0. \tag{3.30}$$

Dada la condición (1) se cumplen (2) y (3) sin más que garantizar que λ_C es acotado y los soportes de Y y C son de la forma $[0,\tau_Y]$, $[0,\tau_C]$ cumpliendo $\tau_Y < \tau_C$.

Una cuestión problemática es que la condición (2) nunca se cumple en aquellos casos en los que el soporte de la función de supervivencia es mayor que el de la función de censura, $\tau_Y > \tau_C$ lo cual es muy habitual en estudios clínicos. La solución que proponen Edelmann y col. (2021) en esos casos es truncar los tiempos de supervivencia escogiendo una constante no negativa adecuada y trabajar con ellos en lugar de con los tiempos originales. Una elección natural para dicha constante en muchos casos puede ser el instante de finalización del estudio.

3.3. Test de independencia

A raíz de las propiedades presentadas en el Teorema 3.2.2, se deriva directamente la posibilidad de plantear un test de permutaciones para contrastar la independencia entre un vector de covariables X y el tiempo de supervivencia Y que sea evaluables sobre una muestra $\{(X_i,T_i,\delta_i)\}_{i=1}^n$ obtenida bajo censura aleatoria por la derecha.

Ese test consiste en asociar de manera aleatoria (sin repeticiones) las observaciones $\{X_i\}_{i=1}^n$ y $\{(T_i,\delta_i)\}_{i=1}^n$ y evaluar sobre la muestra obtenida $\{(X_i,T_{\pi(i)},\delta_{\pi(i)})\}_{i=1}^n$ el estimador propuesto $\tilde{\Omega}$:

$$\tilde{\Omega}_\pi = \tilde{\Omega}(\{(X_i,T_{\pi(i)},\delta_{\pi(i)})\}_{i=1}^n). \tag{3.31}$$

Si repetimos este procedimiento B veces, siendo B un número razonablemente alto, podemos aproximar el p-valor del estimador evaluado sobre la muestra original $\tilde{\Omega}$ como la proporción de veces que hemos obtenido un valor superior al evaluar el estadístico sobre las permutaciones aleatorias π de la muestra:

$$p = \frac{\#\{j \in \{1,...,B\} | \tilde{\Omega}_{\pi_j}\} \geq \tilde{\Omega} + 1}{B+1}. \tag{3.32}$$

Bajo la hipótesis nula de independencia, es de suponer que la mitad de los valores obtenidos al simular quede. por encima del valor $\tilde{\Omega}$. Si por el contrario existe dependencia lo esperable será que la correlación de distancias obtenida sobre la muestra original sea en general mayor que la obtenida bajo censura, lo que conduciría a p-valores pequeños de acuerdo con la fórmula anterior y, por tanto, al rechazo de la hipótesis nula tal y como queremos.

Capítulo 4

Simulaciones

A continuación vamos a ilustrar el comportamiento de las herramientas presentadas a lo largo del trabajo sobre muestras de ejemplo obtenidas por simulación.

4.1. Ejemplo de iteración

Para que los resultados sean un poco más robustos, en lugar de simular una muestra y analizar sus resultados vamos a generar B simulaciones y promediar los resultados. Como se trata de un ejemplo no tomaremos un número de repeticiones B muy grande.

Comenzaremos explicando qué sucede en cada una de las iteraciones y cómo interpretar los resultados. Esto nos permite ilustrar con gráficas los resultados obtenidos. En la sección siguiente presentamos los resultados obtenidos con B simulaciones.

Para comparar el comportamiento del estimador propuesto por Edelmann y col. (2021) que hemos presentado en el Capítulo 3 con las herramientas clásicas y las técnicas de correlación de distancias para datos completos, vamos a simular distintas relaciones de dependencia entre una covariable continua $z \in \mathbb{R}^N$ y los tiempos de supervivencia observados que denotaremos por $y \in \mathbb{R}^N$.

El vector de covariables z lo generaremos en cada caso tomando $N = 100$ realizaciones independientes e idénticamente distribuidas que siguen la distribución

$$z \sim N(0, 1). \tag{4.1}$$

A partir de él vamos a generar cuatro vectores de tiempos de supervivencia no negativos $\{y_i\}_{i=1}^4$ con las siguientes estructuras de relación:

1. $y_1 = cos(z)$
2. $y_2 = \log |z|$
3. $y_3 = \sqrt{z}$
4. $y_4 \sim N(0, 1)$ Independencia

El código correspondiente al ejemplo que desgranamos aquí y con el que hemos generado todas las gráficas de esta sección se pueden encontrar en el Anexo B.1. En él se puede ver que a la generación de los tiempos de supervivencia y_i se le han aplicado algunas correcciones técnicas para garantizar que sean siempre positivos.

La relación de dependencia entre cada uno de estos tiempos de supervivencia y la covariable z se puede observar en la Figura 4.1. Es evidente que los tres primeros casos se corresponden a situaciones de dependencia entre la covariable y la respuesta y el último a una situación de independencia. Representamos en rojo el ajuste lineal de estos datos y en línea discontinua azul un estimador lowess

	ρ_{z,y_i}	Test F p-value	$dCor(z, y_i)$	Test dCor p-value
$\boldsymbol{y_1}$	-0.11208	0.26689	0.52523	0.002
$\boldsymbol{y_2}$	0.09080	0.36894	0.51195	0.002
$\boldsymbol{y_3}$	0.97005	0.00000	0.99610	0.002
$\boldsymbol{y_4}$	0.03216	0.75076	0.12497	0.138

Cuadro 4.1: Medidas de correlación y dependencia de los datos completos

	F-test rechaza H_0 (incorrelación)	dCor rechaza acepta H_0 (independencia)
$\boldsymbol{y_1}$	FALSE	TRUE
$\boldsymbol{y_2}$	FALSE	TRUE
$\boldsymbol{y_3}$	TRUE	TRUE
$\boldsymbol{y_4}$	FALSE	FALSE

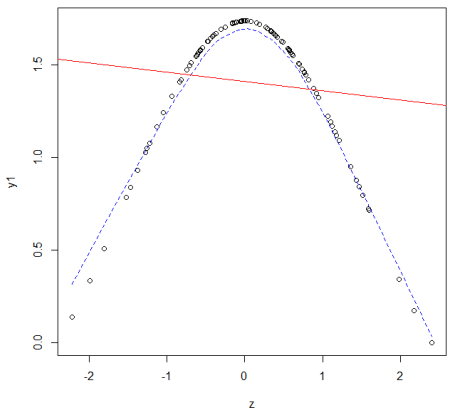
Cuadro 4.2: De acuerdo con el test F sobre el ajuste lineal y con el test de permutaciones basado en distance covariance que utiliza el estimador con corrección de sesgo, vemos en qué casos se detecta correlación (F-test) e independencia (dCor test) a un nivel de significación del 5 %.

local para acentuar la forma de dependencia que captaría un estimador más flexible (no paramétrico). La dependencia en el caso (3) es bastante parecida al ajuste lineal, por lo que es previsible que la correlación de Pearson logre capturarla razonablemente bien, mientras que en los casos (1) y (2) la recta ajustada apenas guarda relación con la asociación observable en los datos.

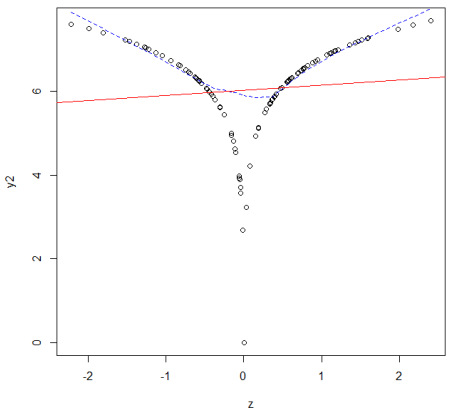
Cada muestra $\{(y_i, z)\}_{i=1}^4$ está constituida por N pares de datos completos, es decir, lo que obtendríamos si fuésemos capaces de observar todos los tiempos de fallo (sin censura). Por tanto, podremos analizar sus relaciones empleando todas las herramientas presentadas en el Capítulo 2: la correlación de Pearson clásica que denotaremos por $\rho(z, y_i)$ y las técnicas de correlación de distancias en datos completos. En particular trabajaremos con el estimador de $dCor$ insesgado presentado en (2.62) para facilitar la comparación posterior con las técnicas específicas para datos censurados por la derecha, y lo denotaremos por $dCor(z, y_i)$. En cuanto a los tests de independencia, realizaremos un test F clásico para testar la existencia de correlación y el test de independencia basado en permutaciones sobre el estadístico $dCor(z, y_i)$ presentado en la Sección 2.2.5.

En la Tabla 4.1 presentamos las magnitudes obtenidas sobre la muestra simulada y los p-valores asociados a los tests. Estos últimos se traducen en decisiones acerca de si se acepta o no la hipótesis de incorrelación e independencia resumidas en la Tabla 4.2 (asumiendo un nivel de significación del 5 %). Vemos que mientras que el test F clásico solo detecta correlación cuando la asociación se parece a una dependencia lineal (como es el caso de y_3), el test basado en dCor identifica correctamente todas las situaciones de dependencia e independencia.

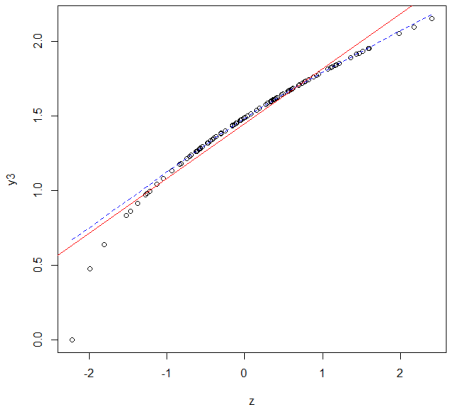
Para poder poner a prueba la correlación de distancias en el contexto del Análisis de Supervivencia



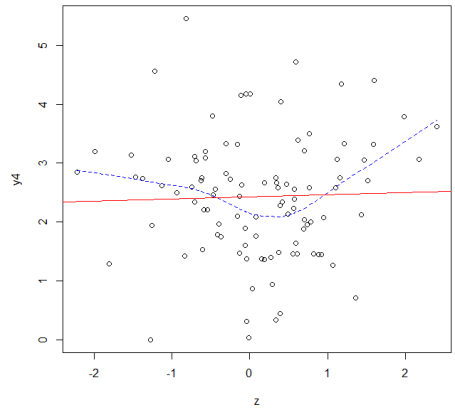
(a) $y_1 = cos(z)$



(b) $y_2 = log|z|$



(c) $y_3 = z^2$



(d) $y_4 \sim N_{100}(\mathbf{0}, I_{100})$

Figura 4.1: Relación entre tiempos simulados $\{y_i\}_{i=1}^5$ y la covariable z. En rojo se presenta un ajuste lineal de los datos y en línea discontinua negra un estimador lowess no paramétrico.

tal y como hemos planteado en el Capítulo 3, no basta haber simulado los tiempos de supervivencia, es necesario introducir en la muestra censura aleatoria por la derecha. Para ello vamos a simular los tiempos de censura a partir de exponenciales independientes de la distribución del vector de covariables z.

Como es previsible que la proporción de censura afecte a los estadísticos, para cada y_i simularemos tres vectores procedentes de distintas distribuciones c_{i1}, c_{i2}, c_{i3} para lograr distintos niveles de censura y ver el efecto que puede tener en la detección de la dependencia.

Las distribuciones en cuestión son las siguientes:

- y_1:

$$\begin{aligned} c_{11} &= 1{,}3 + S_{11}, S_{11} \sim Exp_N(1), \\ c_{12} &= 0{,}7 + S_{12}, S_{12} \sim Exp_N(1), \\ c_{13} &= 0{,}2 + S_{13}, S_{13} \sim Exp_N(1). \end{aligned} \tag{4.2}$$

- y_2:

$$\begin{aligned} c_{21} &= S_{21}, S_{21} \sim Exp_N(0{,}03), \\ c_{22} &= S_{22}, S_{22} \sim Exp_N(0{,}13), \\ c_{23} &= 0{,}5 + S_{23}, S_{23} \sim Exp_N(0{,}35). \end{aligned} \tag{4.3}$$

- y_3:

$$\begin{aligned} c_{31} &= S_{31}, S_{31} \sim Exp_N(0{,}2), \\ c_{32} &= S_{32}, S_{32} \sim Exp_N(0{,}6), \\ c_{33} &= S_{33}, S_{33} \sim Exp_N(1). \end{aligned} \tag{4.4}$$

- y_4:

$$\begin{aligned} c_{41} &= 2{,}5 + S_{41}, S_{41} \sim Exp_N(0{,}3), \\ c_{42} &= 0{,}3 + S_{42}, S_{42} \sim Exp_N(0{,}3), \\ c_{43} &= S_{43}, S_{43} \sim Exp_N(0{,}6). \end{aligned} \tag{4.5}$$

Al ser tiempos generados a partir de variaciones de la distribución exponencial tenemos garantizado que los tiempos de censura serán siempre positivos.

Una vez tenemos estos vectores de tiempos de censura $\{c_{ij}\}_{i,j=1}^{3,4}$ definimos los tiempos observados de acuerdo con la definición vista en (1.12):

$$y_{ij} = \text{mín}\{y_i, c_{ij}\}, \quad \delta_{ij} = \mathbb{I}(y_i \leq c_{ij}) \tag{4.6}$$

entendiendo las operaciones anteriores como operaciones elemento a elemento. Como los tiempos de censura han sido originados a partir de distribuciones totalmente independientes de los tiempo de fallo, la muestra observada (y_{ij}, δ_{ij}) presenta censura aleatoria por la derecha.

Las distribuciones que usamos para generar cada c_{ij} han sido definidas ad hoc para lograr diferentes niveles de censura:

- y_{i1} tiene poca censura (alrededor de un 20 %).
- y_{i2} tiene alrededor de la mitad de las observaciones censuradas (alrededor de un 50 %).
- y_{i2} tiene mucha censura (sobre un 70-80 %).

	$i=1$	$i=2$	$i=3$	$i=4$
$\boldsymbol{c_{i1}}$	0.19	0.15	0.19	0.05
$\boldsymbol{c_{i2}}$	0.49	0.52	0.49	0.46
$\boldsymbol{c_{i3}}$	0.71	0.82	0.72	0.77

Cuadro 4.3: Proporción de censura obtenida en la iteración de ejemplo

Las proporciones de censura obtenidas en este caso se pueden observar en la Tabla 4.3. Evidentemente estos valores no se tienen por qué cumplir para cada muestra simulada, por eso en la Tabla 4.3 hay algunos valores que se alejan bastante del rango que les correspondería. La Tabla 4.8 recoge los niveles promedios de censura obtenidos con B iteraciones y vemos que se aproximan a los valores buscados. A la hora de controlar los niveles de censura nos hemos encontrado con algunas dificultades técnicas que iremos comentando a lo largo de esta sección.

Gráficamente vemos como varía la muestra obtenida en esta iteración al modificar la proporción de censura en las Figuras 4.2, 4.3, 4.4 y 4.5.

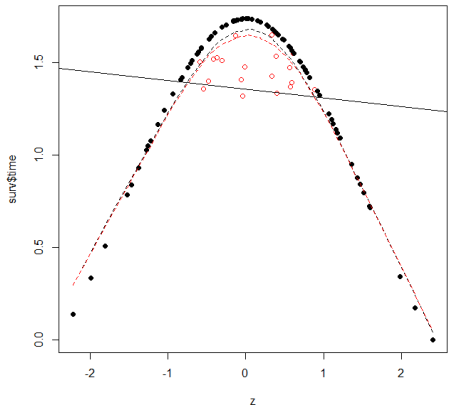
Obviando las observaciones censuradas: El primer enfoque a la hora de tratar las muestras con censura podría ser intentar utilizar las técnicas para datos completos empleando solo los datos correspondientes a tiempos de fallo, es decir, omitir las observaciones censuradas. Bajo este enfoque es de prever que cuanto mayor sea la proporción de censura peor se detecte la correlación o la dependencia. En nuestro caso, quizás el ejemplo más claro sea el caso de la relación logarítmica (y_{21}, δ_{21}). En la Figura 4.3a vemos que al haber tan pocas observaciones censuradas el hecho de obviarlas apenas perjudica a los estimadores, la asociación sigue estando muy patente. Si pasamos a considerar el caso de mucha censura como sucede en la muestra (y_{23}, δ_{23}), representada en la Figura 4.3c, vemos que cada vez la relación es más difícil de identificar y el estimador lineal varía ostensiblemente respecto al ajuste con poca censura.

Podemos observar los cambios en función de la proporción de censura en cada caso en las Figuras 4.2, 4.3, 4.4 y 4.5. En general la relación sigue siendo muy evidente ya que no hemos introducido ruido.

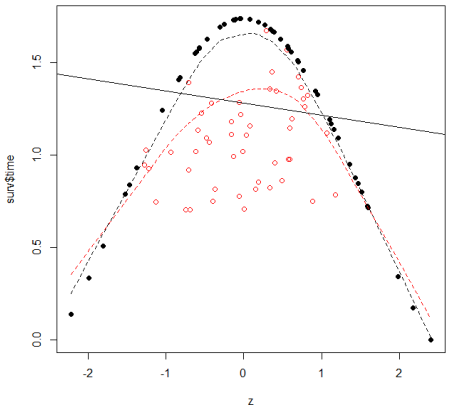
Veamos cuáles son los valores obtenidos para el coeficiente de correlación lineal de Pearson $\rho(y_{ij}, z)$ y la correlación de distancias $dCor(y_{ij}, z)$ cuando obviamos las observaciones censuradas, y asimismo cuáles son los p-valores de los tests asociados. Los resultados los podemos observar en la Tabla 4.4 redondeados a cinco cifras decimales (excepto el p-valor del test de independencia ya que la función implementada en el paquete dcortools da por defecto únicamente tres cifras decimales).

Modelo de Cox y análisis de residuos martingale: El siguiente paso lógico en nuestro análisis es utilizar, finalmente, técnicas específicas para Análisis de Supervivencia. En este sentido podemos utilizar recursos clásicos como ajustar un modelo de Cox y aplicar los tests clásicos para detectar correlación, analizar la relación entre los residuos del modelo y la covariable (que vuelve a ser un contexto de datos completos) o utilizar el estimador IPCW insesgado de la correlación de distancias bajo censura propuesto en el Capítulo 3 y test de independencia asociado.

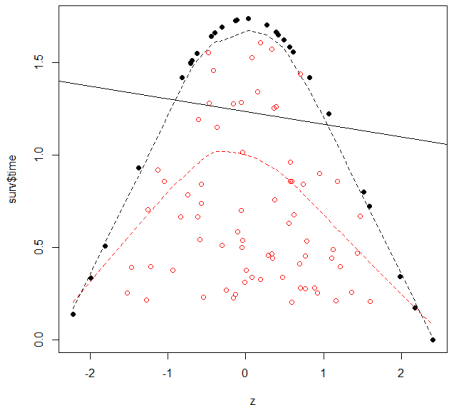
Veamos en primer lugar las opciones que nos ofrece el modelo de Cox semiparamétrico con un ajuste log-lineal sobre las covariables (el modelo clásico). Los primero que haremos es aprovechar las propiedades de los residuos martingale expuestas en el Capítulo 1. En él se explica que al representar los residuos resultantes de ajustar un modelo de Cox vacío (sin covariable) respecto a una covariable el gráfico indica la forma funcional correcta en la que esa variable debería entrar en el modelo. Así,



(a) Poca censura: c_{11}

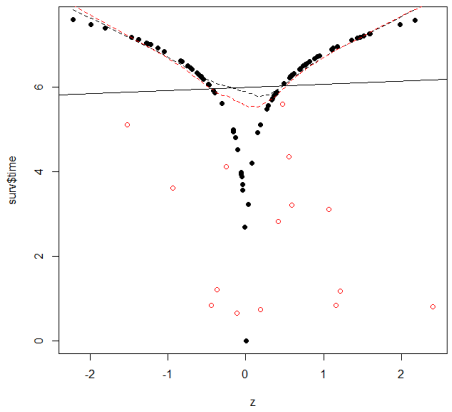


(b) Censura intermedia: c_{12}

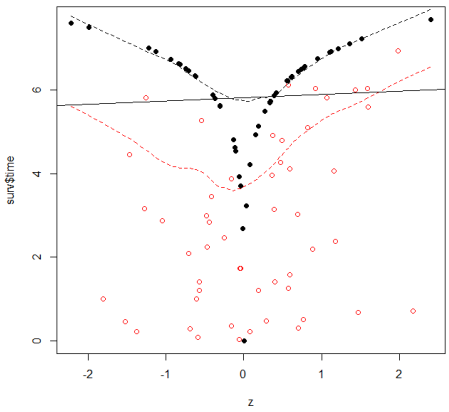


(c) Mucha censura: c_{13}

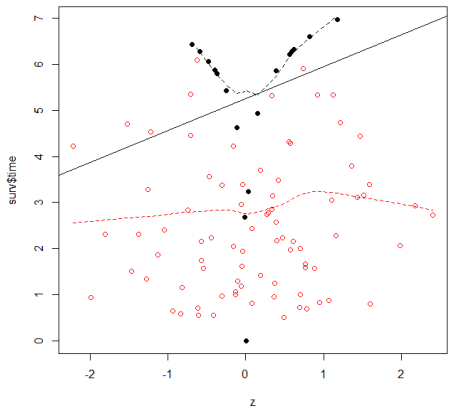
Figura 4.2: Muestras correspondientes a los tiempos de supervivencia simulados $y_1 = cos(z)$ frente a z. En rojo las observaciones censuradas en cada muestra (y_1, δ_{1j}) con distintos niveles de censura c_{1j}. En negro las observaciones no censuradas. La línea negra es un ajuste lineal empleando solo las observaciones no censuradas. La línea discontinua negra es un estimador lowess utilizando las observaciones no censuradas. La línea discontinua roja es un estimador lowess utilizando todas las observaciones.



(a) Poca censura: c_{21}

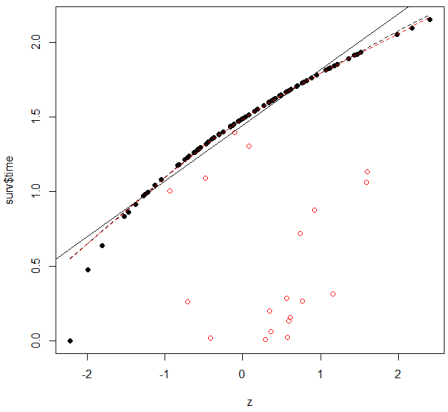


(b) Censura intermedia: c_{22}

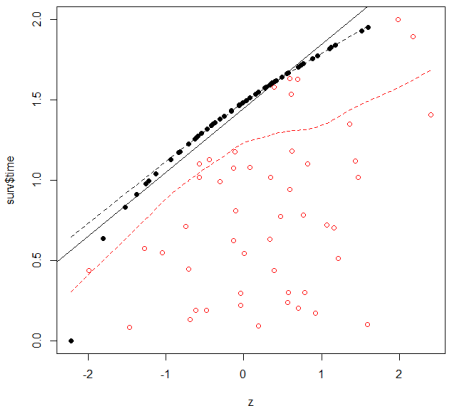


(c) Mucha censura: c_{23}

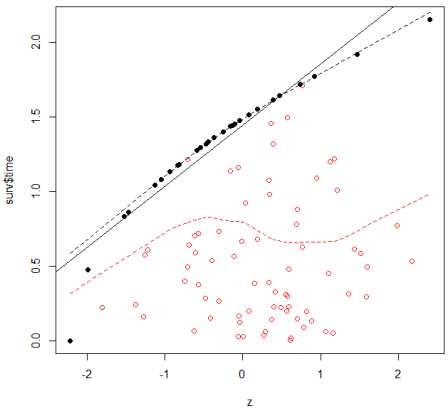
Figura 4.3: Muestras correspondientes a los tiempos de supervivencia simulados $y_2 = log|z|$ frente a z. En rojo las observaciones censuradas en cada muestra (y_2, δ_{2j}) con distintos niveles de censura c_{2j}. La línea negra es un ajuste lineal empleando solo las observaciones no censuradas. La línea discontinua negra es un estimador lowess utilizando las observaciones no censuradas. La línea discontinua roja es un estimador lowess utilizando todas las observaciones.



(a) Poca censura: c_{31}

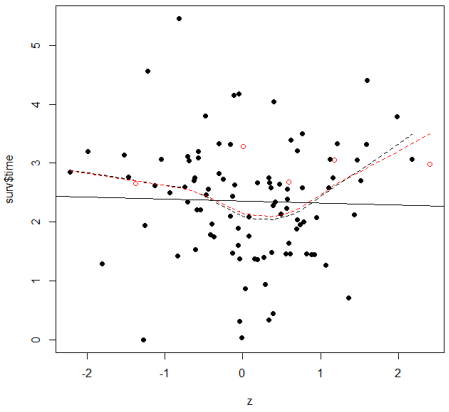


(b) Censura intermedia: c_{32}

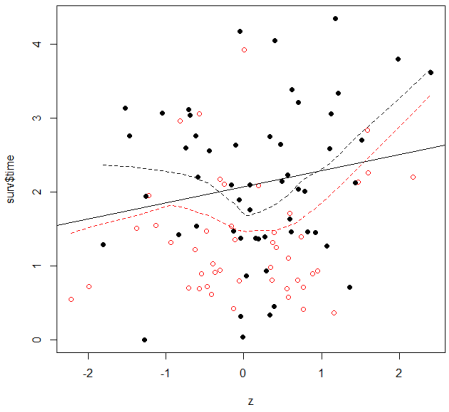


(c) Mucha censura: c_{33}

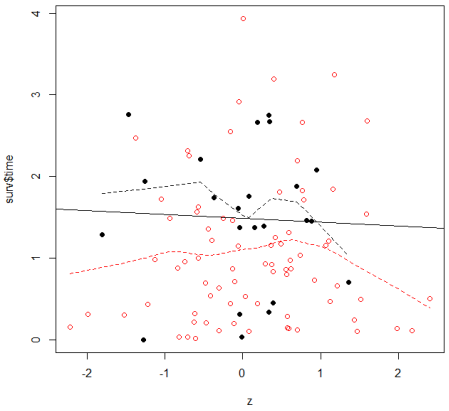
Figura 4.4: Muestras correspondientes a los tiempos de supervivencia simulados $y_3 = z^2$ frente a z. En rojo las observaciones censuradas en cada muestra (y_3, δ_{3j}) con distintos niveles de censura c_{3j}. En negro las observaciones no censuradas. La línea negra es un ajuste lineal empleando solo las observaciones no censuradas. La línea discontinua negra es un estimador lowess utilizando las observaciones no censuradas. La línea discontinua roja es un estimador lowess utilizando todas las observaciones.



(a) Poca censura: c_{41}



(b) Censura intermedia: c_{42}



(c) Mucha censura: c_{43}

Figura 4.5: Muestras correspondientes a los tiempos de supervivencia simulados $y_4 \sim N_{100}(0, I)$ frente a z. En rojo las observaciones censuradas en cada muestra (y_4, δ_{4j}) con distintos niveles de censura c_{4j}. En negro las observaciones no censuradas. La línea negra es un ajuste lineal empleando solo las observaciones no censuradas. La línea discontinua negra es un estimador lowess utilizando las observaciones no censuradas. La línea discontinua roja es un estimador lowess utilizando todas las observaciones.

	$\rho_{z,y_{ij}}$	Test F p-value	$dCor(z, y_{ij})$	Test dCor p-value
$\boldsymbol{y_{11}}$	-0.10748	0.33955	0.51147	0.002
$\boldsymbol{y_{12}}$	-0.14327	0.31588	0.53332	0.002
$\boldsymbol{y_{13}}$	-0.14075	0.46646	0.51506	0.004
$\boldsymbol{y_{21}}$	0.05238	0.63397	0.50360	0.002
$\boldsymbol{y_{22}}$	0.04768	0.74762	0.51285	0.002
$\boldsymbol{y_{23}}$	0.21177	0.39890	0.49270	0.024
$\boldsymbol{y_{31}}$	0.96791	0.00000	0.99544	0.002
$\boldsymbol{y_{32}}$	0.96267	0.00000	0.99640	0.002
$\boldsymbol{y_{33}}$	0.95031	0.00000	0.99250	0.002
$\boldsymbol{y_{41}}$	-0.02959	0.77589	0.12427	0.110
$\boldsymbol{y_{42}}$	0.18328	0.18466	0.12808	0.196
$\boldsymbol{y_{43}}$	-0.04316	0.84499	-0.28338	0.962

Cuadro 4.4: Medidas de correlación y dependencia obviando las observaciones censuradas.

	$\rho_{z,m_{ij}}$	Test F p-value	$dCor(z, m_{ij})$	Test dCor p-value
$\boldsymbol{y_{11}}$	0.07273	0.47207	0.48245	0.002
$\boldsymbol{y_{12}}$	0.12949	0.19912	0.42294	0.002
$\boldsymbol{y_{13}}$	0.03617	0.72091	0.26448	0.004
$\boldsymbol{y_{21}}$	-0.00175	0.98624	0.47237	0.002
$\boldsymbol{y_{22}}$	-0.13405	0.18362	0.36798	0.002
$\boldsymbol{y_{23}}$	-0.13184	0.19103	0.19122	0.020
$\boldsymbol{y_{31}}$	-0.85027	0.00000	0.87620	0.002
$\boldsymbol{y_{32}}$	-0.70428	0.00000	0.68239	0.002
$\boldsymbol{y_{33}}$	-0.58545	0.00000	0.53549	0.002
$\boldsymbol{y_{41}}$	0.00098	0.99232	0.13191	0.116
$\boldsymbol{y_{42}}$	-0.13293	0.18735	0.13905	0.098
$\boldsymbol{y_{43}}$	-0.06278	0.53493	-0.09595	0.704

Cuadro 4.5: Medidas de correlación y dependencia de los residuos del modelo de Cox vacío y la covariable

podemos utilizar las mismas herramientas que hemos empleado hasta ahora para datos completos para detectar asociación entre los residuos del modelo m_{ij} y la covariable z.

Gráficamente podemos observar estas relaciones en la segunda columna de las Figuras 4.6, 4.7, 4.8 y 4.9. Incluimos un ajuste lineal que es evidente que, excepto en el caso (2), no se adapta en absoluto a la relación entre los residuos m_{ij} y z.

Los resultados numéricos pueden observarse en la Tabla 4.5. El coeficiente de Pearson solo captará la existencia de efectos lineales y para capturar estructuras de dependencia más complejas debemos recurrir a los estimadores de la correlación de distancias. En efecto, en vista de los p-valores en dicha tabla solo se detecta correlación en el caso (3) mientras que se detecta dependencia en todos excepto en el (4).

Ajustamos a continuación un modelo de Cox esta vez sí sobre la covariable z. Al representar los residuos de este modelo ${m_{ij}}^z$ deberíamos haber eliminado la componente lineal de los gráficos de residuos del modelo anterior. Gráficamente la representación de los nuevos residuos frente a la explicativa en la tercera columna de los gráficos 4.6, 4.7, 4.8 y 4.9. Vemos que en los casos en los que la dependencia de los residuos del modelo anterior tenía una componente lineal importante el patrón dibujado por los residuos del nuevo modelo varía notablemente respecto al anterior (este cambio solo es perceptible en la Figura 4.8).

Si testamos nuevamente la existencia de correlación y de dependencia entre los residuos de este nuevo modelo y la covariable, ahora únicamente el test de independencia podría capturar la asociación no modelizada. Para testar la hipótesis de no efecto de la covariable bajo la estructura del modelo

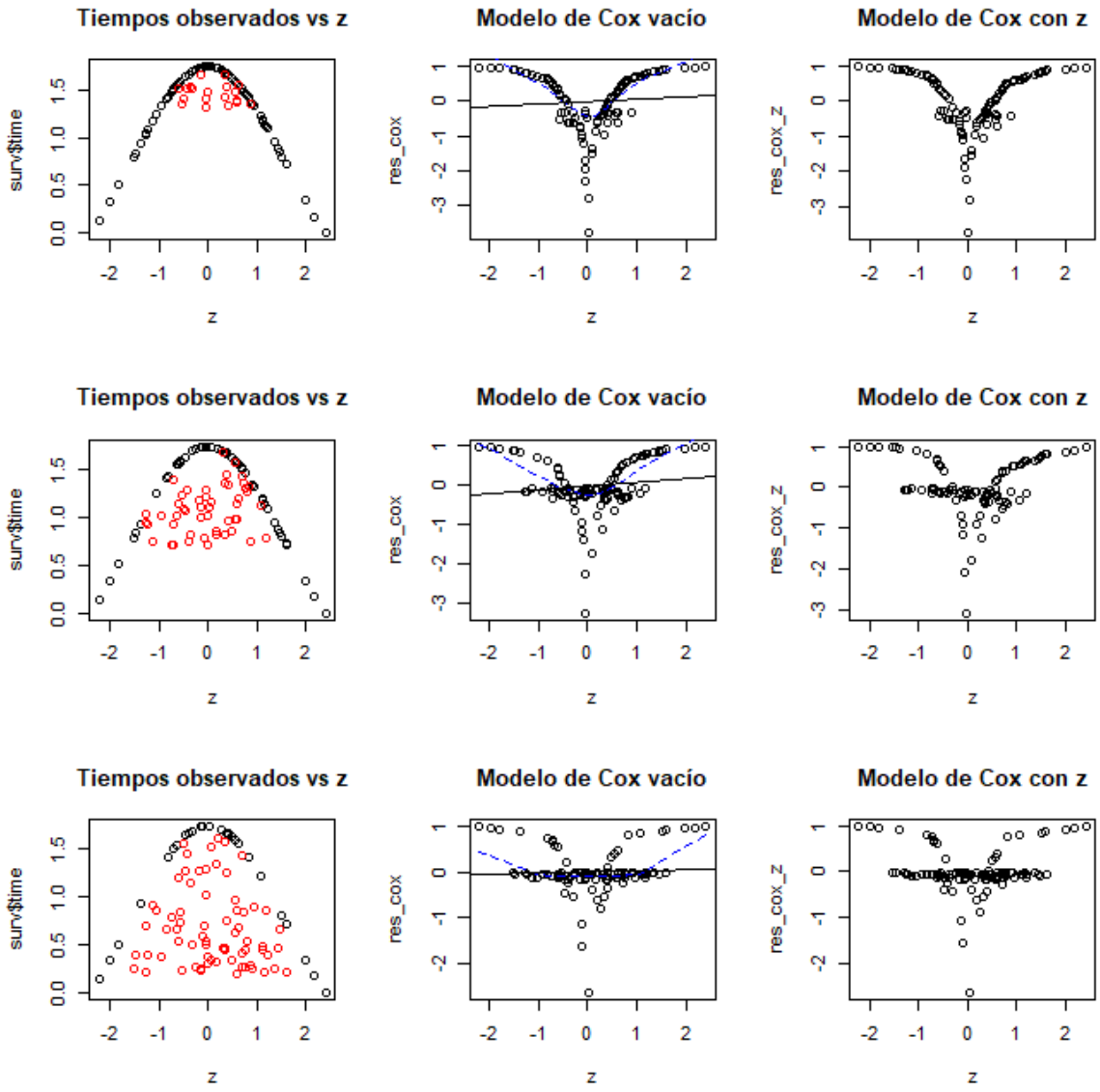


Figura 4.6: En la primera columna se representan los tiempos observados $\boldsymbol{y_{11}}$, $\boldsymbol{y_{12}}$ e $\boldsymbol{y_{13}}$ frente a la covariable $\boldsymbol{z}$. En negro se dibujan los tiempos de fallo y en rojo los tiempos censurados. Vemos que la proporción de censura cambia sensiblemente en los tres casos. En la segunda columna se representan los residuos martingale del modelo de Cox vacío (sin covariable) frente a la covariable z. En línea sólida negra se representa el ajuste lineal para dichos puntos y en línea discontinua negra un estimador lowess no paramétrico. En la tercera columna se representan los residuos martingale del modelo de Cox sobre la covariable z frente a z.

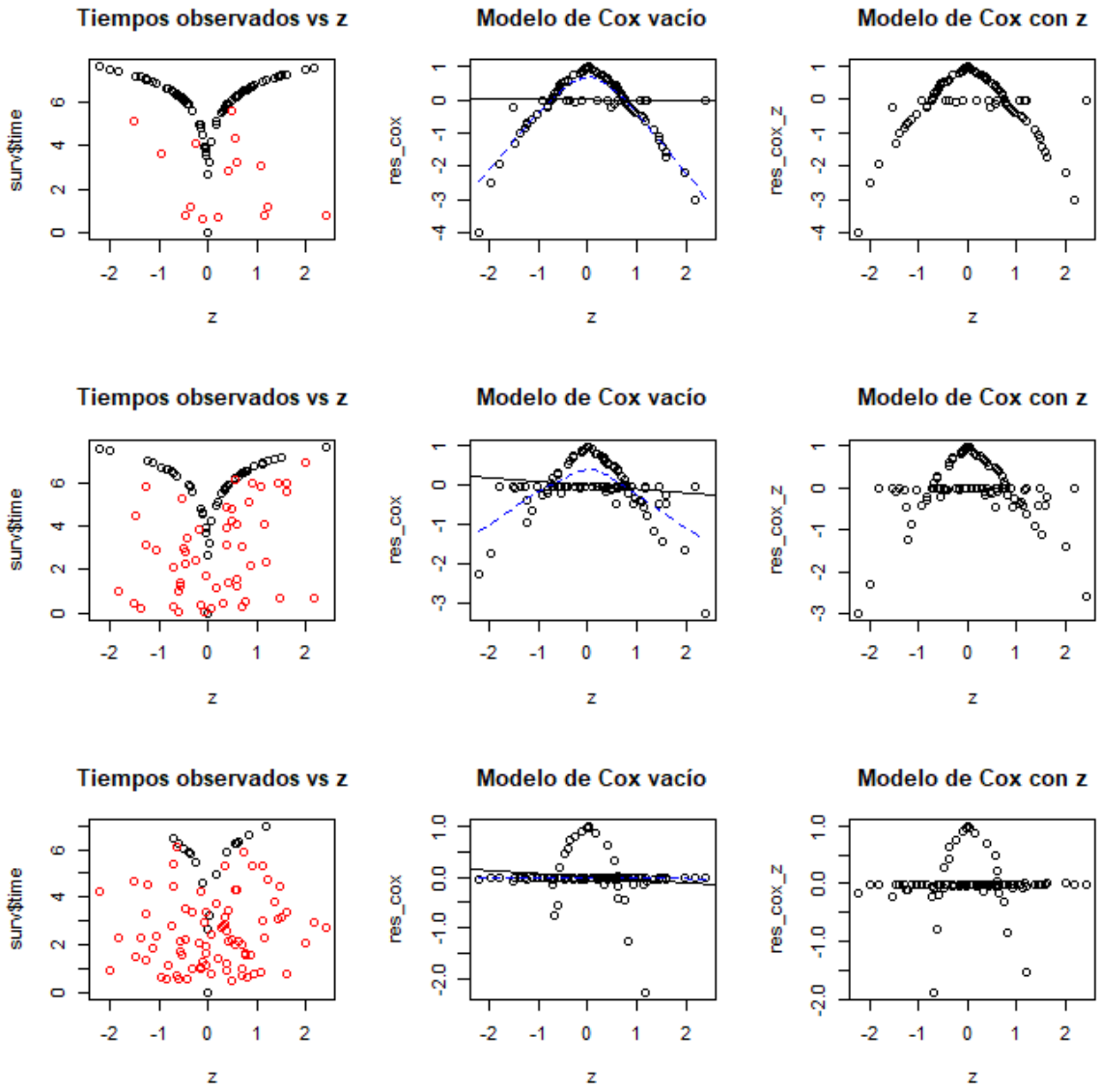


Figura 4.7: En la primera columna se representan los tiempos observados $\boldsymbol{y_{21}}$, $\boldsymbol{y_{22}}$ e $\boldsymbol{y_{23}}$ frente a la covariable $\boldsymbol{z}$. En negro se dibujan los tiempos de fallo y en rojo los tiempos censurados. Vemos que la proporción de censura cambia sensiblemente en los tres casos. En la segunda columna se representan los residuos martingale del modelo de Cox vacío (sin covariable) frente a la covariable z. En línea sólida negra se representa el ajuste lineal para dichos puntos y en línea discontinua negra un estimador lowess no paramétrico. En la tercera columna se representan los residuos martingale del modelo de Cox sobre la covariable z frente a z.

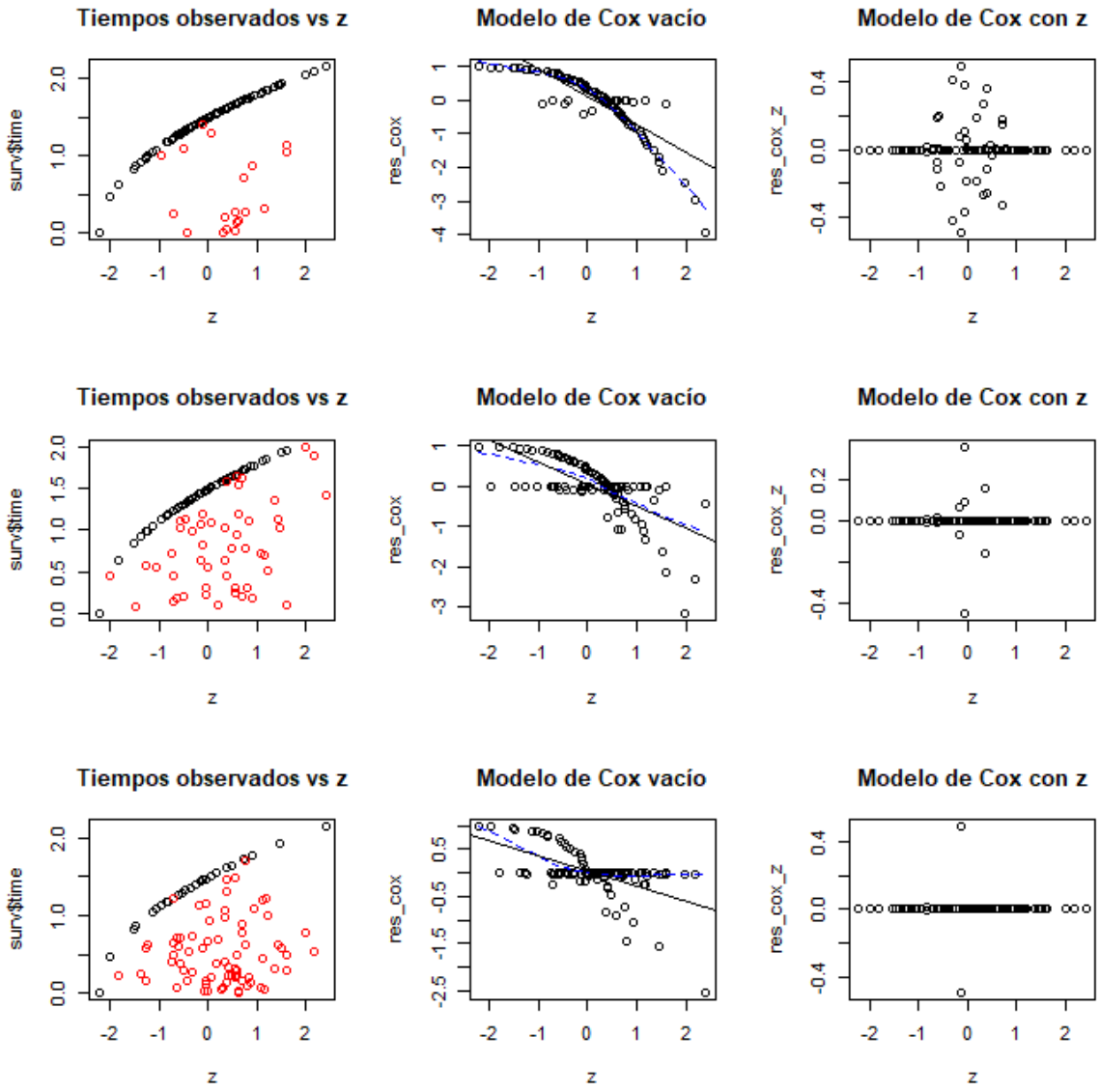


Figura 4.8: En la primera columna se representan los tiempos observados $\boldsymbol{y_{31}}$, $\boldsymbol{y_{32}}$ e $\boldsymbol{y_{33}}$ frente a la covariable $\boldsymbol{z}$. En negro se dibujan los tiempos de fallo y en rojo los tiempos censurados. Vemos que la proporción de censura cambia sensiblemente en los tres casos. En la segunda columna se representan los residuos martingale del modelo de Cox vacío (sin covariable) frente a la covariable z. En línea sólida negra se representa el ajuste lineal para dichos puntos y en línea discontinua negra un estimador lowess no paramétrico. En la tercera columna se representan los residuos martingale del modelo de Cox sobre la covariable z frente a z.

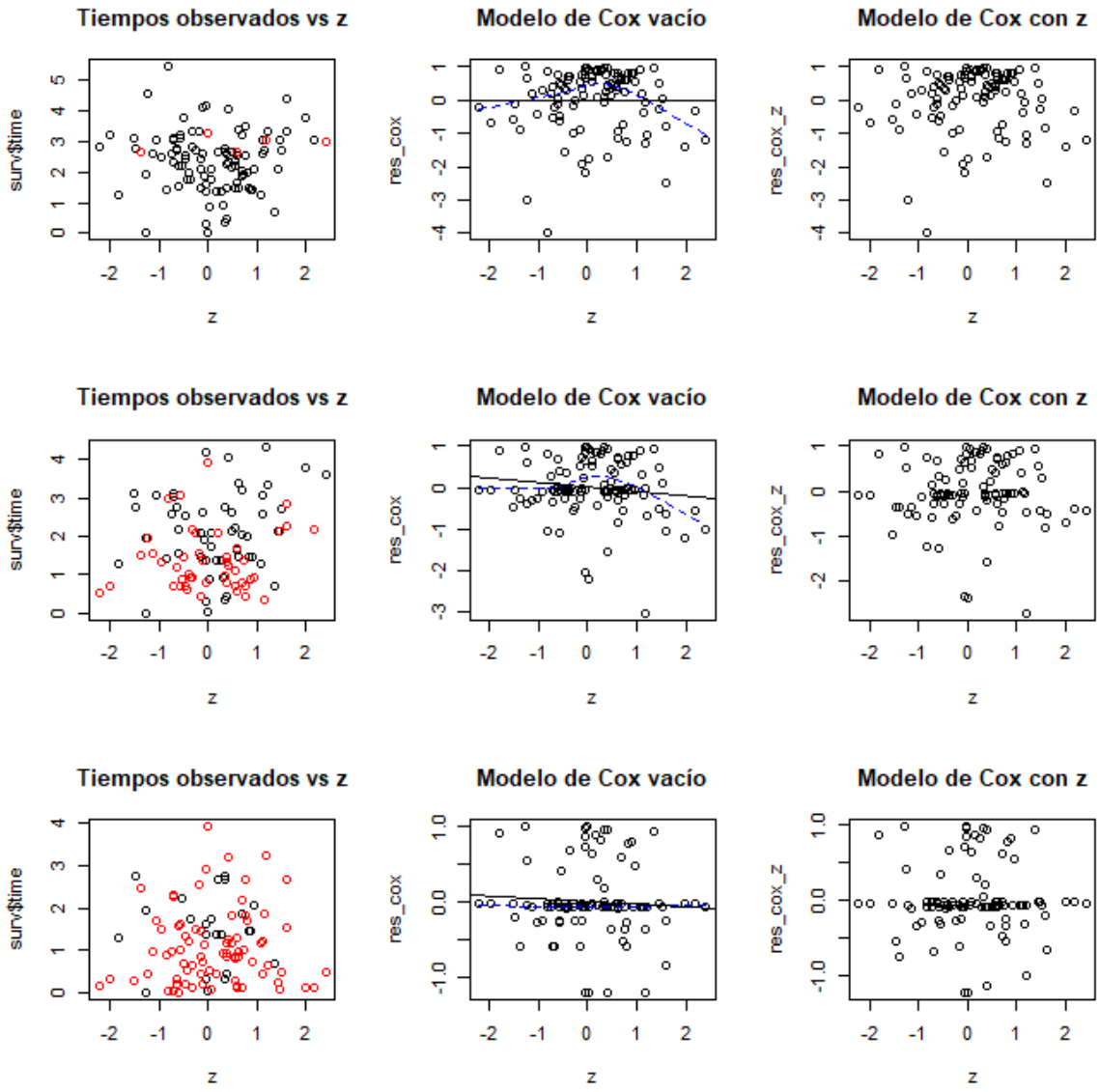


Figura 4.9: En la primera columna se representan los tiempos observados $\boldsymbol{y_{41}}$, $\boldsymbol{y_{42}}$ e $\boldsymbol{y_{43}}$ frente a la covariable $\boldsymbol{z}$. En negro se dibujan los tiempos de fallo y en rojo los tiempos censurados. Vemos que la proporción de censura cambia sensiblemente en los tres casos. En la segunda columna se representan los residuos martingale del modelo de Cox vacío (sin covariable) frente a la covariable z. En línea sólida negra se representa el ajuste lineal para dichos puntos y en línea discontinua negra un estimador lowess no paramétrico. En la tercera columna se representan los residuos martingale del modelo de Cox sobre la covariable z frente a z.

	Test LR p-value	Score Test p-value	Test Wald p-value	$dCor(\boldsymbol{z}, \boldsymbol{y_{ij}}^{\boldsymbol{z}})$	Test dCor p-value
$\boldsymbol{y_{11}}$	0.17204	0.17437	0.17535	0.48103	0.002
$\boldsymbol{y_{12}}$	0.02543	0.02794	0.02915	0.41659	0.002
$\boldsymbol{y_{13}}$	0.54945	0.55029	0.55056	0.26335	0.004
$\boldsymbol{y_{21}}$	0.99080	0.99080	0.99080	0.47235	0.002
$\boldsymbol{y_{22}}$	0.38781	0.38677	0.38796	0.35236	0.002
$\boldsymbol{y_{23}}$	0.11838	0.11144	0.11599	0.16824	0.034
$\boldsymbol{y_{31}}$	0.00000	0.00000	0.00336	0.14546	0.094
$\boldsymbol{y_{32}}$	0.00000	0.00000	0.13868	0.10350	0.206
$\boldsymbol{y_{33}}$	0.00000	0.00000	0.53261	0.09578	0.234
$\boldsymbol{y_{41}}$	0.99325	0.99325	0.99325	0.13191	0.092
$\boldsymbol{y_{42}}$	0.23592	0.23818	0.23902	0.09732	0.164
$\boldsymbol{y_{43}}$	0.50735	0.50735	0.50557	-0.09083	0.680

Cuadro 4.6: Medidas de correlación y dependencia de los residuos del modelo de Cox sobre la covariable z y la propia covariable.

de Cox se pueden utilizar los tests de no efecto presentados en el Capítulo 1: el test de razón de verosimilitudes (LR o loglikelihood), el score test y el test de Wald. Estos resultados se pueden apreciar en la Tabla 4.6. Vemos que los tests de razón de verosimilitudes y score ofrecen resultados coherentes con lo que acabamos de ver, aceptando la hipótesis de no efecto en todos los casos excepto para y_2, con una excepción para y_1 bajo niveles de censura intermedios (este efecto se trata simplemente de una particularidad de esta muestra concreta, veremos que al analizar los resultados ofrecidos por la simulación completa no se observa). El test de Wald, por su parte, se desmarca aceptando la hipótesis de no efecto de la covariable en prácticamente todos los supuestos y en este caso, veremos que este comportamiento se mantiene al realizar la simulación completa.

En cuanto a la correlación de distancias, en este caso se acepta la hipótesis nula de independencia al 5 % tanto para y_3 como para y_4 sea cual sea el nivel de censura. Esto indica que en los casos en los que la relación entre la covariable y el tiempo de fallo es muy parecida a una relación lineal, al introducir el efecto de dicha covariable utilizando un modelo de Cox la correlación de distancias ya no es capaz de detectar la dependencia restante.

dCor **en el contexto de datos censurados:** Por último, vamos a utilizar el estimador IPCW de la correlación de distancias y el test basado en él conforme hemos explicado en el Capítulo 3. Este estimador se evalúa directamente sobre la muestra censurada y esto supone una pequeña restricción práctica. Si recordamos la expresión 3.13, el estimador IPCW de la correlación de distancias es un

	$dCor^c((y_{ij}, \delta_{ij}), z)$	Test $dCor^c$ p-value
$\boldsymbol{y_{11}}$	0.52823	0.002
$\boldsymbol{y_{12}}$	0.53674	0.002
$\boldsymbol{y_{13}}$	0.44187	0.004
$\boldsymbol{y_{21}}$	0.50007	0.002
$\boldsymbol{y_{22}}$	0.47398	0.002
$\boldsymbol{y_{23}}$	0.36133	0.034
$\boldsymbol{y_{31}}$	0.99624	0.002
$\boldsymbol{y_{32}}$	0.91585	0.002
$\boldsymbol{y_{33}}$	0.92268	0.002
$\boldsymbol{y_{41}}$	0.11883	0.150
$\boldsymbol{y_{42}}$	0.15826	0.166
$\boldsymbol{y_{43}}$	-0.18573	0.840

Cuadro 4.7: Medidas de correlación y dependencia de los datos completos

$\mathcal{U}$-estadístico con una función kernel de orden 4 ponderada por unos pesos que tienen en cuenta la censura y la estimación de la supervivencia. Para evaluar este estadístico, que denotaremos por $dCor^c$, es necesario que en cada muestra haya al menos cuatro observaciones no censuradas. Esta condición puede dar problemas en la práctica si estamos trabajando con porcentajes de censura muy altos porque puede suceder que en alguna de las muestras simuladas haya menos de cuatro fallos. Para evitar problemas hemos seleccionado las distribuciones de los tiempos de censura buscando minimizar el número de muestras para las que no podemos calcular el IPCW. Si en una muestra se da este problema, al estimador de esa muestra se le asigna un valor nulo y se hace el promedio sin tenerlo en cuenta. Para controlar que la cantidad de veces que esto sucede sea pequeña en la simulación con B muestras, hemos implementado un contador y fijado una tolerancia del 2 % (veremos que nos quedamos muy lejos de ese valor).

Los resultados obtenidos para el estimador IPCW de la correlación de distancias y el p-valor del test de permutaciones asociado a él se pueden observar en la Tabla 4.7.

Cabe notar que en este caso nuevamente se identifica correctamente la dependencia en los tres primeros casos y la independencia en el último caso.

Antes de pasar a los resultados obtenidos en la simulación completa, cabe mencionar que en las tablas de presentación de los resultados se muestran a menudo estimadores de la correlación de distancias que toman valores negativos. Recordemos que, si bien en el capítulo 2 hemos definido $\mathcal{V}^2$ y $\mathcal{V}_n^2$ como magnitudes estrictamente no negativas, en nuestra implementación estamos trabajando con el estimador insesgado $\tilde{\Omega}$ que sí puede ser menor que cero.

	$i=1$	$i=2$	$i=3$	$i=4$
$\boldsymbol{c_{i1}}$	0.23656	0.19868	0.26608	0.17196
$\boldsymbol{c_{i2}}$	0.50928	0.56228	0.52948	0.45816
$\boldsymbol{c_{i3}}$	0.69704	0.78758	0.73200	0.72052

Cuadro 4.8: Niveles de censura promedio en cada uno de los 12 tipos de muestra simulados

	$\bar{\rho}(z, y_i)$	Test F $\bar{p}$-value	$d\bar{C}or(z, y_i)$	Test dCor $\bar{p}$-value
$\boldsymbol{y_1}$	-0.00836	0.33041	0.54807	0.002
$\boldsymbol{y_2}$	0.00252	0.50525	0.52494	0.002
$\boldsymbol{y_3}$	0.96907	0.00000	0.99564	0.002
$\boldsymbol{y_4}$	0.00126	0.50785	-0.01110	0.498

Cuadro 4.9: Medidas de correlación y dependencia de los datos completos obtenidos promediando las medidas obtenidas para cada una de las B iteraciones.

4.2. Simulación

Los resultados obtenidos en el apartado anterior han servido para ilustrar que significa cada una de las magnitudes calculadas y apoyar su explicación con representaciones gráficas. Lo que vamos a hacer ahora es repetir todo el proceso y los cálculos anteriores B veces y obtener los estimadores de interés y los p-valores de los tests realizados promediando los resultados de estas B iteraciones para tener resultados un poco más sólidos. Vamos a fijar $B = 250$.

El código correspondiente a esta simulación se encuentra disponible para consulta en el Anexo B. Vamos a presentar directamente los resultados obtenidos.

En la Tabla 4.8 se presentan los niveles de censura obtenidos en cada caso promediando la proporción de censura de cada muestra simulada. Como vemos los valores se mueven en los rangos que habíamos planteado inicialmente.

En la Tabla 4.9 vemos el promedio de correlaciones de distancias $\bar{\rho}(z, y_i)$, los $\bar{p}$-valores promedio obtenidos al testar la existencia de correlación utilizando el test F, el promedio de los estimadores de la correlación de distancias $d\bar{C}or(z, y_i)$ y el promedio de $\bar{p}$-valores del test de independencia a él asociado. Los resultados obtenidos muestran valores altos de correlación únicamente para los tiempos de supervivencia y_3 tal y como habíamos explicado en el ejemplo de iteración anterior. Las discrepancias respecto a la Tabla 4.1 pueden deberse en parte a las modificaciones introducidas para garantizar que los tiempos de supervivencia sean positivos. Lo que nos interesa es comparar el resultado para datos completos con los obtenidos para las demás técnicas. La correlación de distancias presenta valores altos en todos los casos excepto en el último. Los niveles de significación obtenidos concuerdan con estas observaciones.

En la Tabla 4.10 se presentan medidas análogas a las de la tabla anterior pero teniendo en cuenta únicamente las observaciones sin censurar dentro de cada muestra simulada. En este caso es eviden-

	ρ	F Test p-value	$dCor$ insesgado	Test dCor p-value
$\boldsymbol{y_{11}}$	-0.00675	0.36730	0.51906	0.00200
$\boldsymbol{y_{12}}$	-0.00971	0.39942	0.51702	0.00228
$\boldsymbol{y_{13}}$	-0.01071	0.39665	0.51694	0.00715
$\boldsymbol{y_{21}}$	0.00222	0.51452	0.52994	0.00200
$\boldsymbol{y_{22}}$	-0.00133	0.47380	0.54766	0.00260
$\boldsymbol{y_{23}}$	-0.00156	0.46274	0.57731	0.02561
$\boldsymbol{y_{31}}$	0.96544	0.00000	0.99522	0.00200
$\boldsymbol{y_{32}}$	0.96122	0.00000	0.99432	0.00200
$\boldsymbol{y_{33}}$	0.95658	0.00000	0.99317	0.00200
$\boldsymbol{y_{41}}$	-0.00524	0.50101	-0.00569	0.47928
$\boldsymbol{y_{42}}$	-0.01215	0.48719	-0.01147	0.48757
$\boldsymbol{y_{43}}$	0.00436	0.50393	-0.03285	0.50768

Cuadro 4.10: Promedios obtenidos sobre las muestras simuladas obviando la censura del coeficiente de correlación de Pearson, de la correlación de distancias y de los p-valores obtenidos con el F test y el test de independencia.

te que para niveles de censura bajos tanto los estadísticos como los p-valores se parecen más a los resultados de la Tabla 4.9. En cualquier caso el efecto del nivel de censura sobre los p-valores no es muy pronunciado con lo que la decisión sobre el contraste en esta situación donde la dependencia es tan clara no se modifica: el F-test solo rechaza la hipótesis nula de incorrelación en las muestras obtenidas a partir de y_3 mientras que el test de independencia rechaza correctamente la hipótesis nula de independencia en las muestras obtenidas a partir de y_1, y_2 e y_3.

La Tabla 4.11 es un resumen de los resultados obtenidos analizando los residuos de los modelos de Cox: uno vacío (ajustado sin covariables) y el otro ajustado con covariables. Vemos que a nivel cualitativo el resultado del F-test no cambia en cuanto a aceptar/rechazar correlación. Del mismo modo, las decisiones del test realizado sobre los residuos del modelo de Cox vacío m_{ij} tampoco varían aunque vemos que en el caso y_{23} (alta proporción de censura) el p-valor está mucho más próximo a aceptar la hipótesis de incorrelación, lo que ilustra que la cantidad de censura podría, en efecto, llegar a modificar el resultado del test. Al residuos m_{ij}^z del modelo de Cox sobre la variable z solo podemos aspirar a captar la dependencia no recogida por el modelo, que vemos que en el caso de los tiempos de supervivencia generados a partir de y_3 ya no será lo suficientemente importante como para que se detecte independencia. Evidentemente tampoco se detecta en el caso de independencia y_4.

Por último, la Tabla 4.12 ofrece los resultados obtenidos promediando los valores del estimador IPCW $dCor^c(z, y_{ij})$ y el p-valor del test de independencia asociado al él. Lo primero que vemos es que a un nivel de significación del 5 % seguimos identificando de forma correcta las situaciones de

	$\rho(z, m_{ij})$	Ftest p-val	$dCor(z, m_{ij})$	$dCor$ test p-val	$dCor(z, m^z{}_{ij})$	$dCor^z$ test p-val
$\boldsymbol{y_{11}}$	0.00503	0.51416	0.49512	0.00200	0.48598	0.00200
$\boldsymbol{y_{12}}$	0.00575	0.43826	0.43400	0.00200	0.41958	0.00202
$\boldsymbol{y_{13}}$	0.00030	0.40992	0.34929	0.00283	0.33171	0.00290
$\boldsymbol{y_{21}}$	-0.00794	0.30528	0.48624	0.00201	0.46626	0.00200
$\boldsymbol{y_{22}}$	-0.00505	0.33907	0.34677	0.00325	0.32872	0.00416
$\boldsymbol{y_{23}}$	0.00041	0.43343	0.23739	0.03268	0.23497	0.02358
$\boldsymbol{y_{31}}$	-0.80293	0.00000	0.82443	0.00200	0.02007	0.42168
$\boldsymbol{y_{32}}$	-0.67957	0.00000	0.66226	0.00200	0.03217	0.40297
$\boldsymbol{y_{33}}$	-0.54965	0.00000	0.50162	0.00201	0.01654	0.44901
$\boldsymbol{y_{41}}$	0.00346	0.52075	-0.01512	0.50482	-0.07663	0.69764
$\boldsymbol{y_{42}}$	0.00478	0.49129	-0.01420	0.48991	-0.07072	0.67031
$\boldsymbol{y_{43}}$	-0.00392	0.48084	-0.00748	0.47979	-0.04155	0.58514

Cuadro 4.11: Residuos de los modelo de Cox: correlación y tests de incorrelación sobre el modelo vacío y correlación de distancias y test de independencia tanto sobre el modelo vacío como sobre el modelo lineal sobre la covariable z.

	$dCor^c(z, y_{ij})$	Test $dCor^c$ p-value
$\boldsymbol{y_{11}}$	0.54507	0.00200
$\boldsymbol{y_{12}}$	0.54094	0.00205
$\boldsymbol{y_{13}}$	0.52650	0.00427
$\boldsymbol{y_{21}}$	0.52242	0.00200
$\boldsymbol{y_{22}}$	0.50213	0.00335
$\boldsymbol{y_{23}}$	0.41387	0.07644
$\boldsymbol{y_{31}}$	0.98750	0.00200
$\boldsymbol{y_{32}}$	0.97452	0.00200
$\boldsymbol{y_{33}}$	0.92335	0.00200
$\boldsymbol{y_{41}}$	-0.00555	0.48006
$\boldsymbol{y_{42}}$	-0.01479	0.49406
$\boldsymbol{y_{43}}$	-0.01082	0.48376

Cuadro 4.12: Estimador IPCW de la correlación de distancias y p-valor correspondiente al test de independencia basado en él.

dependencia/independencia con la excepción del caso y_{23}, el caso al que los niveles altos de censura le afectan de manera más acusada. Además si nos fijamos en los valores concretos vemos que están sumamente próximos a los valores obtenidos para datos completos de la Tabla 4.9, compitiendo con los obtenidos obviando las observaciones censuradas y mejorando las estimaciones obtenidas con los residuos del modelo de Cox. El motivo por el que obviar las observaciones censuradas funciona tan bien en este caso podría ser que al tratarse de relaciones tan limpias (no hemos introducido error ni hay efectos mezclados) aun teniendo menos datos el desempeño de los estimadores es bueno.

En conclusión:

- Vemos que tanto los valores de la correlación de distancias obtenidos obviando la censura como los obtenidos utilizando el estimador IPCW están bastante próximos a los valores teóricos. Los obtenidos analizando los residuos del modelo de Cox vacío se quedan un poco más lejos aunque su rendimiento es aceptable en los casos de poca censura y empeora notablemente a medida que esta aumenta.

- Los resultados obtenidos utilizando solo los datos no censurados (Tabla 4.10) y los obtenidos con el estimador IPCW (Tabla 4.12) son muy similares entre sí. De hecho el primer enfoque capta la dependencia incluso mejor. Hemos de tener en cuenta que al no haber introducido error ni efectos cruzados la dependencia se marca notablemente incluso censurando una parte importante de la muestra. En cualquier caso, de los dos el más afectado por niveles altos de censura parece ser el estimador IPCW.

- Los resultados basados en la correlación de Pearson en general solo presentan un comportamiento aceptable identificando independencia en el caso de la relación (3) debido a su gran similitud con la relación lineal tal y como cabía esperar.
- El coeficiente de correlación de Pearson de los residuos del modelo de Cox funciona peor que el que obvia la censura a la hora de detectar relaciones de tipo lineal o similares.

En definitiva, en este ejemplo hemos visto como se implementan las herramientas basadas en correlación de distancias y son capaces de detectar tipos de dependencia más general que la correlación de Pearson. En el caso de datos censurados, lo que mejor resultado nos ha dado es obviar las observaciones censuradas (quizás debido a la pureza de las relaciones en nuestro ejemplo) pero el estimador IPCW resulta ser plenamente competitivo y aproxima razonablemente bien los valores teóricos. El análisis de residuos de un modelo de Cox vacío con técnicas de correlación de distancias también resulta útil para detectar asociación, aunque estima peor la dependencia especialmente cuando hay mucha censura. Por último, el análisis de residuos de un modelo de Cox sobre la covariable permite identificar en qué casos existen efectos no lineales que se quedan sin capturar en el modelo ajustado.

Capítulo 5

Trabajo Futuro

A lo largo de este texto hemos intentado ofrecer una visión global sobre como resolver el problema de la caracterización de la independencia en el contexto del Análisis de Supervivencia. Para ello hemos intentado seleccionar los trabajos más relevantes e ir guiando al lector a través del desarrollo teórico que conduce a las propuestas de Edelmann y col. (2021).

Hemos intentado que se trate de un trabajo autocontenido incluyendo las cuestiones básicas sobre Análisis de Supervivencia y ahondando en cuestiones específicas relativas a la correlación de distancias (conexión con $\mathcal{U}$-estadísticos, propuestas IPCW para datos censurados) para poder enlazar mejor todo el contenido.

Las aplicaciones de la correlación de distancias se encuentran en franca expansión actualmente. Existen multitud de líneas de investigación abiertas y las propuestas de extensión a Análisis de Supervivencia son pocas y muy recientes. Relacionado de manera directa con el trabajo presentado se podría plantear:

- Extensión de los estadísticos y tests de independencia vistos al caso de truncamiento.
- Extensión de los estadísticos y tests de independencia vistos al caso de censura informativa o dependiente.
- Adaptación de las técnicas de correlación de distancias a situaciones específicas con covariables categóricas: como definir las distancias de manera conveniente.
- Diseño de un estudio de simulación exhaustivo que permita explorar la influencia de los niveles de censura o la dimensión de la covariable sobre la implementación de estas herramientas.

Más en general, como hemos mencionado la correlación de distancias no es más que una de las herramientas englobadas dentro de los denominados Energy-Statistics (estadísticos de energía). Del mismo modo que se ha extendido el test de independencia al caso de censura aleatoria por la derecha se puede pensar en extender otros tests basados en estadísticos de energía a ese u otros contextos específicos del Análisis de Supervivencia, por ejemplo tests de bondad de ajuste o tests de comparación de poblaciones.

Por último, sería interesante utilizar estos métodos para analizar bases de datos reales de interés en algún ámbito concreto, especialmente bases de datos de alta dimensión o con covariables multidimensionales.

Apéndice A

Implementación en R

Existen dos paquetes fundamentalmente para implementar las técnicas presentadas en este trabajo disponibles para el software estadístico R.

A.1. Paquete dcortools

El paquete utilizado en las simulaciones es el paquete **dcortools**, un paquete centrado en implementar de manera sencilla y eficiente las técnicas que hemos presentado a lo largo del trabajo.

En particular las funciones que hemos empleado han sido:

- **distcor**: permite calcular la correlación de distancias entre dos vectores X e Y. Con el argumento *bias.correct = TRUE* permite obtener el estimador insesgado presentado en la Sección 2.2.7.
- **distcov.test**: permite realizar tests de correlación de distancias. Por defecto utiliza el argumento *method = "permutation"* con lo que implementa el test presentado en la Sección 2.2.5. Si además fijamos el argumento *bias.correct = TRUE* se toma como estadístico de contraste el estimador insesgado de la covarianza de distancias presentado en la Sección 2.2.7.
- **ipcw.dcor**: permite calcular el estimador IPCW de la correlación de distancias derivado de 2.62. Al menos uno de los argumentos ha de ser una muestra censurada (a diferencia de las funciones anteriores).
- **ipcw.dcov.test**: realiza el contraste de independencia presentado en la Sección 3.3. En este caso no es necesario especificar nada, por defecto el cálculo del p-valor se hace por permutaciones y el estadístico de contraste es el estimador IPCW de la covarianza de distancias (insesgado).

Para más información sobre las opciones que ofrece este paquete se puede consultar Jochen Fiedler (2021).

A.2. Paquete energy

El paquete energy es un paquete más general que no solo implementa los métodos de correlación de distancias vistos sino que incluye funciones para implementar la mayor parte de Energy Statistics propuestos en Székely y Rizzo (2017).

Incluye varias funciones que permiten realizar los cálculos que hemos realizado con dcortools con la excepción del cálculo directo de los estimadores IPCW sobre muestras censuradas.

Para más información se puede consultarGabor Szekely (2022).

Apéndice B

Código

B.1. Código iteración de ejemplo

```
# Ejemplo ilustrativo:

library(dcortools)
library(survival)
library(mgcv)

set.seed(1)

B = 1
k = 1
N = 100
rel = 4
L = 3*rel

prop_cens = as.data.frame(matrix(nrow=L,ncol=B))
rownames(prop_cens) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
"cens33","cens41","cens42","cens43")

ctrol_elim = as.data.frame(matrix(0,(rel*3),1))
rownames(ctrol_elim) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
"cens33","cens41","cens42","cens43")

# Sin censura:

pearson_cor = as.data.frame(matrix(nrow=rel,ncol=B))
rownames(pearson_cor) = c("y1","y2","y3","y4")

Ftest_pvalue = as.data.frame(matrix(nrow=rel,ncol=B))
rownames(Ftest_pvalue) = c("y1","y2","y3","y4")

distcor = as.data.frame(matrix(nrow=rel,ncol=B))
rownames(distcor) = c("y1","y2","y3","y4")

distcor_pvalue = as.data.frame(matrix(nrow=rel,ncol=B))
rownames(distcor_pvalue) = c("y1","y2","y3","y4")
```

```
# Con censura:

# Usando las mismas herramientas que antes para las obs falladas:

pearson_cor_cens = as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(pearson_cor_cens) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
     "cens33","cens41","cens42","cens43")

Ftest_pvalue_cens = as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(Ftest_pvalue_cens) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
     "cens33","cens41","cens42","cens43")

distcor_cens = as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(distcor_cens) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
     "cens33","cens41","cens42","cens43")

distcor_pvalue_cens = as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(distcor_pvalue_cens) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
     "cens33","cens41","cens42","cens43")

# Usando herramientas especificas de modelos de supervivencia con censura
    aleat por la derecha:

# Propuesta IPCW de Edelmann:

ipcw_dcorr = as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(ipcw_dcorr) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
     "cens33","cens41","cens42","cens43")
ipcw_pvalue = as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(ipcw_pvalue) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
     "cens33","cens41","cens42","cens43")

# Sobre los residuos de un ajuste de Cox:

# Residuos martingale del modelo vacio:

pearson_cor_res <-  as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(pearson_cor_res) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
     "cens33","cens41","cens42","cens43")

pearson_pvalue_res <-  as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(pearson_pvalue_res) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
     "cens33","cens41","cens42","cens43")

dcor_res <-  as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(dcor_res) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
```

```
    "cens33","cens41","cens42","cens43")

dcor_pvalue_res <-  as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(dcor_pvalue_res) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
    "cens33","cens41","cens42","cens43")

# Tests para un modelo de Cox sobre la covariable z:

logtest_pvalue <-  as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(logtest_pvalue) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
    "cens33","cens41","cens42","cens43")

sctest_pvalue <-  as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(sctest_pvalue) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
    "cens33","cens41","cens42","cens43")

wald_pvalue <-  as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(wald_pvalue) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
    "cens33","cens41","cens42","cens43")

dcor_res_z <-  as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(dcor_res_z) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
    "cens33","cens41","cens42","cens43")

dcov_res_z_pvalue <-  as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(dcov_res_z_pvalue) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
    "cens33","cens41","cens42","cens43")


set.seed(1)

z <- rnorm(N)

y1 <- cos(z) - min(0,cos(z)) + rep(0.00001,N) # Relacion coseno (y >= 0)

cens11 <- 1.3 + rexp(N)   # Poca censura
cens12 <- 0.7 + rexp(N)   # Censura intermedia
cens13 <- 0.2 + rexp(N)         # Mucha censura

delta <- as.numeric(y1 < cens11)
time <- sapply(1:N, function(u) min(y1[u], cens11[u]))
surv11 <- as.data.frame(cbind(time, delta))
prop_cens[1,k] = (sum(surv11$delta == 0))/N

delta <- as.numeric(y1 < cens12)
time <- sapply(1:N, function(u) min(y1[u], cens12[u]))
surv12 <- as.data.frame(cbind(time, delta))
prop_cens[2,k] = (sum(surv12$delta == 0))/N

delta <- as.numeric(y1 < cens13)
```

```
time <- sapply(1:N, function(u) min(y1[u], cens13[u]))
surv13 <- as.data.frame(cbind(time, delta))
prop_cens[3,k] = (sum(surv13$delta == 0))/N

y2 <- log(abs(z)) - min(0,log(abs(z))) + rep(0.00001,N)  # Relacion
    logaritmica (y >= 0)

cens21 <- rexp(N, rate = 0.03)     # Poca censura
cens22 <- rexp(N, rate = 0.13 )   # Censura intermedia
cens23 <- 0.5 + rexp(N, rate = 0.35)     # Mucha censura

delta <- as.numeric(y2 < cens21)
time <- sapply(1:N, function(u) min(y2[u], cens21[u]))
surv21 <- as.data.frame(cbind(time, delta))
prop_cens[4,k] = (sum(surv21$delta == 0))/N

delta <- as.numeric(y2 < cens22)
time <- sapply(1:N, function(u) min(y2[u], cens22[u]))
surv22 <- as.data.frame(cbind(time, delta))
prop_cens[5,k] = (sum(surv22$delta == 0))/N

delta <- as.numeric(y2 < cens23)
time <- sapply(1:N, function(u) min(y2[u], cens23[u]))
surv23 <- as.data.frame(cbind(time, delta))
prop_cens[6,k] = (sum(surv23$delta == 0))/N

y3 <- sqrt(z - min(0,z) + rep(0.00001,N) ) # Relacion raiz cuadrada

cens31 <- rexp(N, rate = 0.2)  # Poca censura
cens32 <- rexp(N, rate = 0.6)  # Censura intermedia
cens33 <- rexp(N)              # Mucha censura

delta <- as.numeric(y3 < cens31)
time <- sapply(1:N, function(u) min(y3[u], cens31[u]))
surv31 <- as.data.frame(cbind(time, delta))
prop_cens[7,k] = (sum(surv31$delta == 0))/N

delta <- as.numeric(y3 < cens32)
time <- sapply(1:N, function(u) min(y3[u], cens32[u]))
surv32 <- as.data.frame(cbind(time, delta))
prop_cens[8,k] = (sum(surv32$delta == 0))/N

delta <- as.numeric(y3 < cens33)
time <- sapply(1:N, function(u) min(y3[u], cens33[u]))
surv33 <- as.data.frame(cbind(time, delta))
prop_cens[9,k] = (sum(surv33$delta == 0))/N

aux = rnorm(N)
y4 <- aux - min(0,aux) + rep(0.00001,N) # Independencia

cens41 <- 2.5 + rexp(N, rate = 0.3) # Poca censura
cens42 <- 0.3 + rexp(N, rate = 0.3) # Censura intermedia
cens43 <- rexp(N, rate = 0.6) # Mucha censura

delta <- as.numeric(y4 < cens41)
time <- sapply(1:N, function(u) min(y4[u], cens41[u]))
```

```
surv41 <- as.data.frame(cbind(time, delta))
prop_cens[10,k] = (sum(surv41$delta == 0))/N

delta <- as.numeric(y4 < cens42)
time <- sapply(1:N, function(u) min(y4[u], cens42[u]))
surv42 <- as.data.frame(cbind(time, delta))
prop_cens[11,k] = (sum(surv42$delta == 0))/N

delta <- as.numeric(y4 < cens43)
time <- sapply(1:N, function(u) min(y4[u], cens43[u]))
surv43 <- as.data.frame(cbind(time, delta))
prop_cens[12,k] = (sum(surv43$delta == 0))/N


for (j in 1:L){
  if (prop_cens[j,k] >= ((N-4)/N)){
    prop_cens[j,k] <- NA # Con menos de cuatro observaciones no podemos
        evaluar el estimador (u-estadistico de orden 4)
    ctrol_elim[j,] <- ctrol_elim[j,] + 1
  }
}
#prop_cens[is.na(prop_cens)] = 1


# Medimos la asociaci?n en cada caso:

# Sin censura:

times = as.data.frame(cbind(y1,y2,y3,y4))
times_names = c("y1","y2","y3","y4")

for (j in 1:rel){

  windows()
  plot(z,times[,j], ylab = times_names[j])
  lines(lowess(z,times[,j]), col="blue", lty=2)

  pearson_cor[j,k] = cor(z,times[,j])
  m0 <- lm(times[,j] ~ z)
  abline(m0, col="red")
  Ftest_pvalue[j,k] = 1 -
      pf(summary(m0)$fstatistic[1],summary(m0)$fstatistic[2],summary(m0)$fstatistic[3])

  distcor[j,k] = distcor(z, times[,j], bias.corr = TRUE)
  distcor_pvalue[j,k] = distcov.test(z,times[,j], bias.corr = TRUE)$pvalue

}

# Con censura:

sample =
    list(surv11,surv12,surv13,surv21,surv22,surv23,surv31,surv32,surv33,surv41,
     surv42,surv43)

# Herramientas generales:
```

```
for (j in 1:L){
  surv = sample[[j]]
  pearson_cor_cens[j,k] <- cor(z[which(surv$delta ==
      1)],surv[which(surv$delta == 1),]$time)
  m1 <- lm(surv[which(surv$delta == 1),]$time ~ z[which(surv$delta == 1)])
  Ftest_pvalue_cens[j,k] = 1 -
      pf(summary(m1)$fstatistic[1],summary(m1)$fstatistic[2],summary(m1)$fstatistic[3])

  windows()
  plot(z,surv$time, col="red")
  points(z[which(surv$delta == 1)],surv[which(surv$delta == 1),]$time,pch=19)
  lines(lowess(z[which(surv$delta == 1)], surv[which(surv$delta ==
      1),]$time),lty=2)
  lines(lowess(z, surv$time),lty=2,col="red")
  abline(m1, col="black")

  distcor_cens[j,k] = distcor(z[which(surv$delta == 1)],
      surv[which(surv$delta == 1),]$time, bias.corr = TRUE)
  distcor_pvalue_cens[j,k] = distcov.test(z[which(surv$delta == 1)],
      surv[which(surv$delta == 1),]$time, bias.corr = TRUE)$pvalue
}

# Herramientas analisis de supervivencia

for (j in 1:L){
  surv = sample[[j]]
  if(is.na(prop_cens[j,k])){
    ipcw_dcor[j,k] <- NA
    ipcw_pvalue[j,k] <- NA
    rechazo_ipcw_dcor[j,k] <- NA
  }else{
    surv = sample[[j]]
    #print(ipcw.dcor(sample[[j]], z))
    ipcw_dcorr[j,k] = ipcw.dcor(surv, z)
    ipcw_pvalue[j,k] = ipcw.dcov.test(surv, z)$pvalue

    if (sum(j == c(1,4,7,10,13)) > 0){
      windows()
      par(mfrow=c(3,3))
    }

    plot(z, surv$time, main="Tiempos␣observados␣vs␣z")
    points(z[which(surv$delta == 0)],surv[which(surv$delta ==
        0),]$time,col="red")
  }

  # Utilizando los residuos de un modelo de Cox:

  # Modelo de Cox vacio: residuos martingale

  mod_cox <- coxph(Surv(surv$time, surv$delta) ~ 1)
  res_cox <- resid(mod_cox)
  plot(z, res_cox, main = "Modelo␣de␣Cox␣vacio")
  abline(lm(res_cox ~ z), col="red")
  lines(lowess(z, res_cox),lty=2, col="blue")
```

```
  pearson_cor_res[j,k] <- cor(z,res_cox)
  m_cox0 <- lm(res_cox ~ z)
  abline(m_cox0)
  pearson_pvalue_res[j,k] <- 1 -
      pf(summary(m_cox0)$fstatistic[1],summary(m_cox0)$fstatistic[2],
    summary(m_cox0)$fstatistic[3])

  dcor_res[j,k] <- distcor(res_cox,z, bias.corr = TRUE)
  dcor_pvalue_res[j,k] <- distcov.test(z,res_cox, bias.corr = TRUE)$pvalue

  # Modelo de Cox con covariable z:

  mod_cox_z <- coxph(Surv(surv$time, surv$delta) ~ z)
  res_cox_z <- resid(mod_cox_z)
  plot(z, res_cox_z, main = "Modelo de Cox con z")

  logtest_pvalue[j,k] <- summary(mod_cox_z)$logtest[3]
  sctest_pvalue[j,k] <- summary(mod_cox_z)$sctest[3]
  wald_pvalue[j,k] <- summary(mod_cox_z)$waldtest[3]
  dcor_res_z[j,k] <- distcor(res_cox_z,z, bias.corr = TRUE)
  dcov_res_z_pvalue[j,k] <- distcov.test(z,res_cox_z, bias.corr = TRUE)$pvalue
}


##############################################################################
# RESULTADOS:

# Proporcion de censura:
rowMeans(prop_cens, na.rm = TRUE)

# Controlamos que se vaya a poder calcular el estimador ipcw.dcor en todos los
# casos (i. e. que haya al menos 4 fallos en cada muestra simulada)
t(ctrol_elim)

# Ejemplo de que con menos de cuatro fallos no funciona:
# times0 <- c(1:5)
# z0 <- c(2:6)
# for (i in 1:6){
#   delta0 <- c(rep(0,i-1),rep(1,5-(i-1)))
#   surv0 <- as.data.frame(cbind(times0,delta0)
#   print("Indicador de censura:")
#   print(delta0)
#   print(ipcw.dcor(surv0, z0))
# }
# Solo funciona para i = 1, 2 (cinco y cuatro fallos respectivamente)

# Resultados sin censura
pearson_cor_m <- rowMeans(pearson_cor)
Ftest_pvalue_m <- rowMeans(Ftest_pvalue) # Se acepta la incorrelaci?n en los
    dos primeros casos
distcor_m <- rowMeans(distcor)
distcor_pvalue_m <- rowMeans(distcor_pvalue) # Se rechaza la independencia en
    todos los casos excepto el ?ltimo

# Resultados utilizando la muestra censurada:
```

```
# Herramientas generales para datos completos:

pearson_cor_cens_m <- rowMeans(pearson_cor_cens)
Ftest_pvalue_cens_m <- rowMeans(Ftest_pvalue_cens)
distcor_cens_m <- rowMeans(distcor_cens)
distcor_pvalue_cens_m <- rowMeans(distcor_pvalue_cens)

# Propuesta IPCW de Edelmann

ipcw_dcorr_m <- rowMeans(ipcw_dcorr, na.rm = TRUE)
ipcw_pvalue_m <- rowMeans(ipcw_pvalue, na.rm = TRUE)

# Resultados residuos del modelo de Cox:

# Modelo vacio (residuos martingale):

pearson_cor_res_m <-  rowMeans(pearson_cor_res)
pearson_pvalue_res_m <-  rowMeans(pearson_pvalue_res)
dcor_res_m <- rowMeans(dcor_res)
dcor_pvalue_res <-  rowMeans(dcor_pvalue_res)

# Modelo con covariable z:

logtest_pvalue_m <- rowMeans(logtest_pvalue,na.rm = TRUE)
sctest_pvalue_m <- rowMeans(sctest_pvalue,na.rm = TRUE)
wald_pvalue_m <- rowMeans(wald_pvalue,na.rm = TRUE)

dcor_res_z_m <- rowMeans(dcor_res_z,na.rm = TRUE)
dcovres_res_z_pvalue_m <- rowMeans(dcov_res_z_pvalue,na.rm = TRUE)


# RESULTADOS FINALES:

result_sin_cens <-
    cbind(pearson_cor_m,Ftest_pvalue_m,distcor_m,distcor_pvalue_m)
result_con_cens <-
    cbind(pearson_cor_cens_m,Ftest_pvalue_cens_m,distcor_cens_m,distcor_pvalue_cens_m,
     ipcw_dcorr_m,ipcw_pvalue_m,pearson_cor_res_m,pearson_pvalue_res_m,dcor_res_m,
     dcor_pvalue_res,logtest_pvalue_m,sctest_pvalue_m,wald_pvalue_m,dcor_res_z_m,
     dcovres_res_z_pvalue_m)

head(result_sin_cens)
head(result_con_cens)

# Rechazo incorrelacion/independencia sin censura:
result_sin_cens[,c(2,4)] < 0.05

# rechazo incorrelacion/independencia con censura:
result_con_cens[,c(2,4,6,8,10,11,12,13,15)] < 0.05
```

B.2. Código simulaciones

```
# SIMULACIONES:

B = 250
N = 100
rel = 4
L = 3*rel

prop_cens = as.data.frame(matrix(nrow=L,ncol=B))
rownames(prop_cens) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
     "cens33","cens41","cens42","cens43")

ctrol_elim = as.data.frame(matrix(0,(rel*3),1))
rownames(ctrol_elim) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
     "cens33","cens41","cens42","cens43")

# Sin censura:

pearson_cor = as.data.frame(matrix(nrow=rel,ncol=B))
rownames(pearson_cor) = c("y1","y2","y3","y4")

Ftest_pvalue = as.data.frame(matrix(nrow=rel,ncol=B))
rownames(Ftest_pvalue) = c("y1","y2","y3","y4")

distcor = as.data.frame(matrix(nrow=rel,ncol=B))
rownames(distcor) = c("y1","y2","y3","y4")

distcor_pvalue = as.data.frame(matrix(nrow=rel,ncol=B))
rownames(distcor_pvalue) = c("y1","y2","y3","y4")

# Con censura:

# Usando las mismas herramientas que antes para las obs falladas:

pearson_cor_cens = as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(pearson_cor_cens) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
     "cens33","cens41","cens42","cens43")

Ftest_pvalue_cens = as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(Ftest_pvalue_cens) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
     "cens33","cens41","cens42","cens43")

distcor_cens = as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(distcor_cens) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
     "cens33","cens41","cens42","cens43")

distcor_pvalue_cens = as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(distcor_pvalue_cens) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
     "cens33","cens41","cens42","cens43")
```

```
# Usando herramientas especificas de modelos de supervivencia con censura
    aleatoria por la derecha:

# Propuesta IPCW de Edelmann:

ipcw_dcorr = as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(ipcw_dcorr) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
    "cens33","cens41","cens42","cens43")
ipcw_pvalue = as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(ipcw_pvalue) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
    "cens33","cens41","cens42","cens43")

# Sobre los residuos de un ajuste de Cox:

# Residuos martingale del modelo vacio:

pearson_cor_res <-  as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(pearson_cor_res) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
    "cens33","cens41","cens42","cens43")

pearson_pvalue_res <-  as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(pearson_pvalue_res) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
    "cens33","cens41","cens42","cens43")

dcor_res <-  as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(dcor_res) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
    "cens33","cens41","cens42","cens43")

dcor_pvalue_res <-  as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(dcor_pvalue_res) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
    "cens33","cens41","cens42","cens43")

# Tests para un modelo de Cox sobre la covariable z:

logtest_pvalue <-  as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(logtest_pvalue) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
    "cens33","cens41","cens42","cens43")

sctest_pvalue <-  as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(sctest_pvalue) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
    "cens33","cens41","cens42","cens43")

wald_pvalue <-  as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(wald_pvalue) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
    "cens33","cens41","cens42","cens43")
```

```
dcor_res_z <-  as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(dcor_res_z) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
    "cens33","cens41","cens42","cens43")

dcov_res_z_pvalue <-  as.data.frame(matrix(nrow=(rel*3),ncol=B))
rownames(dcov_res_z_pvalue) =
    c("cens11","cens12","cens13","cens21","cens22","cens23","cens31","cens32",
    "cens33","cens41","cens42","cens43")


set.seed(24)
for (k in 1:B){

  z <- rnorm(N)

  y1 <- cos(z) - min(0,cos(z)) + rep(0.00001,N) # Relaci?n coseno (y >= 0)

  cens11 <- 1.3 + rexp(N)   # Poca censura
  cens12 <- 0.7 + rexp(N)         # Censura intermedia
  cens13 <- 0.15 + rexp(N)         # Mucha censura

  delta <- as.numeric(y1 < cens11)
  time <- sapply(1:N, function(u) min(y1[u], cens11[u]))
  surv11 <- as.data.frame(cbind(time, delta))
  prop_cens[1,k] = (sum(surv11$delta == 0))/N

  delta <- as.numeric(y1 < cens12)
  time <- sapply(1:N, function(u) min(y1[u], cens12[u]))
  surv12 <- as.data.frame(cbind(time, delta))
  prop_cens[2,k] = (sum(surv12$delta == 0))/N

  delta <- as.numeric(y1 < cens13)
  time <- sapply(1:N, function(u) min(y1[u], cens13[u]))
  surv13 <- as.data.frame(cbind(time, delta))
  prop_cens[3,k] = (sum(surv13$delta == 0))/N

  y2 <- log(abs(z)) - min(0,log(abs(z))) + rep(0.00001,N)  # Relaci?n
      logar?tmica (y >= 0)

  cens21 <- rexp(N, rate = 0.05)      # Poca censura
  cens22 <- rexp(N, rate = 0.2 )    # Censura intermedia
  cens23 <- rexp(N, rate = 0.4)      # Mucha censura

  delta <- as.numeric(y2 < cens21)
  time <- sapply(1:N, function(u) min(y2[u], cens21[u]))
  surv21 <- as.data.frame(cbind(time, delta))
  prop_cens[4,k] = (sum(surv21$delta == 0))/N

  delta <- as.numeric(y2 < cens22)
  time <- sapply(1:N, function(u) min(y2[u], cens22[u]))
  surv22 <- as.data.frame(cbind(time, delta))
  prop_cens[5,k] = (sum(surv22$delta == 0))/N

  delta <- as.numeric(y2 < cens23)
  time <- sapply(1:N, function(u) min(y2[u], cens23[u]))
```

```
surv23 <- as.data.frame(cbind(time, delta))
prop_cens[6,k] = (sum(surv23$delta == 0))/N

y3 <- sqrt(z - min(0,z) + rep(0.00001,N) ) # Relaci?n cuadr?tica

cens31 <- rexp(N, rate = 0.2)  # Poca censura
cens32 <- rexp(N, rate = 0.5)  # Censura intermedia
cens33 <- rexp(N, rate = 0.9)              # Mucha censura

delta <- as.numeric(y3 < cens31)
time <- sapply(1:N, function(u) min(y3[u], cens31[u]))
surv31 <- as.data.frame(cbind(time, delta))
prop_cens[7,k] = (sum(surv31$delta == 0))/N

delta <- as.numeric(y3 < cens32)
time <- sapply(1:N, function(u) min(y3[u], cens32[u]))
surv32 <- as.data.frame(cbind(time, delta))
prop_cens[8,k] = (sum(surv32$delta == 0))/N

delta <- as.numeric(y3 < cens33)
time <- sapply(1:N, function(u) min(y3[u], cens33[u]))
surv33 <- as.data.frame(cbind(time, delta))
prop_cens[9,k] = (sum(surv33$delta == 0))/N

aux = rnorm(N)
y4 <- aux - min(0,aux) + rep(0.00001,N)

cens41 <- 2 + rexp(N, rate = 0.3) # Poca censura
cens42 <- 0.3 + rexp(N, rate = 0.3) # Censura intermedia
cens43 <- rexp(N, rate = 0.6) # Mucha censura

delta <- as.numeric(y4 < cens41)
time <- sapply(1:N, function(u) min(y4[u], cens41[u]))
surv41 <- as.data.frame(cbind(time, delta))
prop_cens[10,k] = (sum(surv41$delta == 0))/N

delta <- as.numeric(y4 < cens42)
time <- sapply(1:N, function(u) min(y4[u], cens42[u]))
surv42 <- as.data.frame(cbind(time, delta))
prop_cens[11,k] = (sum(surv42$delta == 0))/N

delta <- as.numeric(y4 < cens43)
time <- sapply(1:N, function(u) min(y4[u], cens43[u]))
surv43 <- as.data.frame(cbind(time, delta))
prop_cens[12,k] = (sum(surv43$delta == 0))/N


for (j in 1:L){
  if (prop_cens[j,k] >= ((N-4)/N)){
    prop_cens[j,k] <- NA # Con menos de cuatro observaciones no podemos
    # evaluar el estimador (u-estadistico de orden 4)
    ctrol_elim[j,] <- ctrol_elim[j,] + 1
  }
}

# Medimos la asociacion en cada caso:
```

```
# Sin censura:

times = as.data.frame(cbind(y1,y2,y3,y4))

for (j in 1:rel){

  pearson_cor[j,k] = cor(z,times[,j])
  m0 <- lm(times[,j] ~ z)
  Ftest_pvalue[j,k] = 1 -
      pf(summary(m0)$fstatistic[1],summary(m0)$fstatistic[2],summary(m0)$fstatistic[3])

  distcor[j,k] = distcor(z, times[,j], bias.corr = TRUE)
  distcor_pvalue[j,k] = distcov.test(z,times[,j], bias.corr = TRUE)$pvalue

}

# Con censura:

sample =
    list(surv11,surv12,surv13,surv21,surv22,surv23,surv31,surv32,surv33,surv41,
  surv42,surv43)

# Herramientas generales:

for (j in 1:L){
  surv = sample[[j]]
  pearson_cor_cens[j,k] <- cor(z[which(surv$delta ==
      1)],surv[which(surv$delta == 1),]$time)
  m1 <- lm(surv[which(surv$delta == 1),]$time ~ z[which(surv$delta == 1)])
  Ftest_pvalue_cens[j,k] = 1 -
      pf(summary(m1)$fstatistic[1],summary(m1)$fstatistic[2],summary(m1)$fstatistic[3])

  distcor_cens[j,k] = distcor(z[which(surv$delta == 1)],
      surv[which(surv$delta == 1),]$time, bias.corr = TRUE)
  distcor_pvalue_cens[j,k] = distcov.test(z[which(surv$delta == 1)],
      surv[which(surv$delta == 1),]$time, bias.corr = TRUE)$pvalue
}

# Herramientas analisis de supervivencia

for (j in 1:L){
  surv = sample[[j]]
  if(is.na(prop_cens[j,k])){
    ipcw_dcorr[j,k] <- NA
    ipcw_pvalue[j,k] <- NA
  }else{
    surv = sample[[j]]
    #print(ipcw.dcor(sample[[j]], z))
    ipcw_dcorr[j,k] = ipcw.dcor(surv, z)
    ipcw_pvalue[j,k] = ipcw.dcov.test(surv, z)$pvalue

  }

  # Utilizando los residuos de un modelo de Cox:
```

```
    # Modelo de Cox vacio: residuos martingale

    mod_cox <- coxph(Surv(surv$time, surv$delta) ~ 1)
    res_cox <- resid(mod_cox)

    pearson_cor_res[j,k] <- cor(z,res_cox)
    m_cox0 <- lm(res_cox ~ z)
    pearson_pvalue_res[j,k] <- 1 -
        pf(summary(m_cox0)$fstatistic[1],summary(m_cox0)$fstatistic[2],
         summary(m_cox0)$fstatistic[3])

    dcor_res[j,k] <- distcor(res_cox,z, bias.corr = TRUE)
    dcor_pvalue_res[j,k] <- distcov.test(z,res_cox, bias.corr = TRUE)$pvalue

    # Modelo de Cox con covariable z:

    mod_cox_z <- coxph(Surv(surv$time, surv$delta) ~ z)
    res_cox_z <- resid(mod_cox_z)

    logtest_pvalue[j,k] <- summary(mod_cox_z)$logtest[3]
    sctest_pvalue[j,k] <- summary(mod_cox_z)$sctest[3]
    wald_pvalue[j,k] <- summary(mod_cox_z)$waldtest[3]
    dcor_res_z[j,k] <- distcor(res_cox_z,z, bias.corr = TRUE)
    dcov_res_z_pvalue[j,k] <- distcov.test(z,res_cox_z, bias.corr =
        TRUE)$pvalue

  }

print(k) # Contador para ver el avance del algoritmo

}

##############################################################################

# RESULTADOS:

# Proporcion de censura:
prop_cens[,1:10]
rowMeans(prop_cens, na.rm = TRUE)

# Controlamos que se vaya a poder calcular el estimador ipcw.dcor en todos los
# casos (i. e. que haya al menos 4 fallos en cada muestra simulada)
t(ctrol_elim)

# Ejemplo de que con menos de cuatro fallos no funciona:
# times0 <- c(1:5)
# z0 <- c(2:6)
# for (i in 1:6){
#   delta0 <- c(rep(0,i-1),rep(1,5-(i-1)))
#   surv0 <- as.data.frame(cbind(times0,delta0)
#   print("Indicador de censura:")
#   print(delta0)
#   print(ipcw.dcor(surv0, z0))
# }
# Solo funciona para i = 1, 2 (cinco y cuatro fallos respectivamente)
```

```
# Resultados sin censura
pearson_cor_m <- rowMeans(pearson_cor, na.rm = TRUE)
Ftest_pvalue_m <- rowMeans(Ftest_pvalue, na.rm = TRUE) # Se acepta la
    incorrelaci?n en los dos primeros casos
distcor_m <- rowMeans(distcor, na.rm = TRUE)
distcor_pvalue_m <- rowMeans(distcor_pvalue, na.rm = TRUE) # Se rechaza la
    independencia en todos los casos excepto el ?ltimo

# Resultados utilizando la muestra censurada:

# Herramientas generales para datos completos:

pearson_cor_cens_m <- rowMeans(pearson_cor_cens, na.rm = TRUE)
Ftest_pvalue_cens_m <- rowMeans(Ftest_pvalue_cens, na.rm = TRUE)
distcor_cens_m <- rowMeans(distcor_cens, na.rm = TRUE)
distcor_pvalue_cens_m <- rowMeans(distcor_pvalue_cens, na.rm = TRUE)

# Propuesta IPCW de Edelmann

ipcw_dcorr_m <- rowMeans(ipcw_dcorr, na.rm = TRUE)
ipcw_pvalue_m <- rowMeans(ipcw_pvalue, na.rm = TRUE)

# Resultados residuos del modelo de Cox:

# Modelo vacio (residuos martingale):

pearson_cor_res_m <-  rowMeans(pearson_cor_res, na.rm = TRUE)
pearson_pvalue_res_m <-  rowMeans(pearson_pvalue_res, na.rm = TRUE)
dcor_res_m <- rowMeans(dcor_res, na.rm = TRUE)
dcor_pvalue_res_m <-  rowMeans(dcor_pvalue_res, na.rm = TRUE)

# Modelo con covariable z:

logtest_pvalue_m <- rowMeans(logtest_pvalue,na.rm = TRUE)
sctest_pvalue_m <- rowMeans(sctest_pvalue,na.rm = TRUE)
wald_pvalue_m <- rowMeans(wald_pvalue,na.rm = TRUE)

dcor_res_z_m <- rowMeans(dcor_res_z,na.rm = TRUE)
dcov_res_z_pvalue_m <- rowMeans(dcov_res_z_pvalue,na.rm = TRUE)


# RESULTADOS FINALES:

result_sin_cens <-
    cbind(pearson_cor_m,Ftest_pvalue_m,distcor_m,distcor_pvalue_m)
result_con_cens <-
    cbind(pearson_cor_cens_m,Ftest_pvalue_cens_m,distcor_cens_m,distcor_pvalue_cens_m,
     ipcw_dcorr_m,ipcw_pvalue_m,pearson_cor_res_m,pearson_pvalue_res_m,dcor_res_m,dcor_pvalue_

head(result_sin_cens)
head(result_con_cens)

# Rechazo incorrelacion/independencia sin censura:
result_sin_cens[,c(2,4)] < 0.05

# rechazo incorrelacion/independencia con censura:
```

Bibliografía

Bakirov, Nail K, Maria L Rizzo y Gábor J Székely (2006). «A multivariate nonparametric test of independence». En: *Journal of multivariate analysis* 97(8), págs. 1742-1756.

Beran, Rudolf (ene. de 1981). «Nonparametric regression with randomly censored survival data». En.

Cox, David R (1972). «Regression models and life-tables». En: *Journal of the Royal Statistical Society: Series B (Methodological)* 34(2), págs. 187-202.

Datta, Somnath (2005). «Estimating the mean life time using right censored data». En: *Statistical Methodology* 2(1), págs. 65-69.

Datta, Somnath, Dipankar Bandyopadhyay y Glen A Satten (2010). «Inverse Probability of Censoring Weighted U-statistics for Right-Censored Data with an Application to Testing Hypotheses». En: *Scandinavian Journal of Statistics* 37(4), págs. 680-700.

Edelmann, Dominic, Thomas Welchowski y Axel Benner (2021). «A consistent version of distance covariance for right-censored survival data and its application in hypothesis testing». En: *Biometrics*.

Edelmann, Dominic y col. (2020). «Marginal variable screening for survival endpoints». En: *Biometrical Journal* 62(3), págs. 610-626.

Ferger, Dietmar y col. (2017). *From statistics to mathematical finance*. Springer.

Feuerverger, Andrey (1993). «A consistent test for bivariate dependence». En: *International Statistical Review/Revue Internationale de Statistique*, págs. 419-433.

Flores Flores, Claudio Jaime (2011). «Modelo de regresión de Cox usando splines». En.

Gabor Szekely, Martin Rizzo y (2022). *energy: E-Statistics: Multivariate Inference via the Energy of Data*. R package version 1.7-9.

Gonzalez-Manteiga, Wenceslao y Carmen Cadarso-Suarez (1994). «Asymptotic properties of a generalized Kaplan-Meier estimator with some applications». En: *Communications in Statistics-Theory and Methods* 4(1), págs. 65-78.

Hoeffding, Wassily (1948). «A Class of Statistics with Asymptotically Normal Distribution». En: *The Annals of Mathematical Statistics* 19(3), págs. 293-325. DOI: 10.1214/aoms/1177730196. URL: https://doi.org/10.1214/aoms/1177730196.

Hua, Wen-Yu y col. (2015). «Multiple comparison procedures for neuroimaging genomewide association studies». En: *Biostatistics* 16(1), págs. 17-30.

Huang, Cheng y Xiaoming Huo (2017). «A statistically and numerically efficient independence test based on random projections and distance covariance». En: *arXiv preprint arXiv:1701.06054*.

Huo, Xiaoming y Gábor J Székely (2016). «Fast computing for distance covariance». En: *Technometrics* 58(4), págs. 435-447.

Iglesias Pérez María del Carmen y de Uña Álvarez, Jacobo (feb. de 2021). *Análisis de Supervivencia*.

Jochen Fiedler, Dominic Edelmann y (2021). *dcortools: Providing Fast and Flexible Functions for Distance Correlation Analysis*. R package version 0.1.2.

Kaplan, Edward L y Paul Meier (1958). «Nonparametric estimation from incomplete observations». En: *Journal of the American statistical association* 53(282), págs. 457-481.

Klein, John P y Melvin L Moeschberger (2003). *Survival analysis: techniques for censored and truncated data*. Vol. 1230. Springer.

Kong, Jing y col. (2012). «Using distance correlation and SS-ANOVA to assess associations of familial relationships, lifestyle factors, diseases, and mortality». En: *Proceedings of the National Academy of Sciences* 109(50), págs. 20352-20357.

Lawless, Jerald F (2011). *Statistical models and methods for lifetime data*. Vol. 362. John Wiley & Sons.

Li, Runze, Wei Zhong y Liping Zhu (2012). «Feature screening via distance correlation learning». En: *Journal of the American Statistical Association* 107(499), págs. 1129-1139.

Lin, Danyu y Thomas R Fleming (2012). *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis: Survival Analysis*. Vol. 123. Springer Science & Business Media.

López Montoya, Antonio Jesús (2011). «Comparación de dos modelos de regresión en fiabilidad». Tesis de mtría. Máster Universitario en Estadística Aplicada. Universidad de Granada.

López-Cheda, Ana y col. (2018). «Nonparametric inference in mixture cure models». En: *Multidisciplinary Digital Publishing Institute Proceedings* 2(18), pág. 1181.

Lyons, Russell (2013). «Distance covariance in metric spaces». En: *The Annals of Probability* 41(5), págs. 3284-3305.

Moreira, Carla y Jacobo de Uña-Álvarez (2010). «A semiparametric estimator of survival for doubly truncated data». En: *Statistics in Medicine* 29(30), págs. 3147-3159.

Peng, L y JP Fine (2008). «Nonparametric tests for continuous covariate effects with multistate survival data». En: *Biometrics* 64(4), págs. 1080-1089.

Rennert, Lior (2018). «Statistical Methods for Truncated Survival Data». Tesis doct. University of Pennsylvania.

Stone, Charles J (1977). «Consistent nonparametric regression». En: *The annals of statistics*, págs. 595-620.

Székely, Gábor J y Maria L Rizzo (2014). «Partial distance correlation with methods for dissimilarities». En: *The Annals of Statistics* 42(6), págs. 2382-2412.

Székely, Gábor J y Maria L Rizzo (2017). «The energy of data». En: *Annual Review of Statistics and Its Application* 4, págs. 447-479.

Székely, Gábor J, Maria L Rizzo y Nail K Bakirov (2007). «Measuring and testing dependence by correlation of distances». En: *The annals of statistics* 35(6), págs. 2769-2794.

Therneau, Terry M y Patricia M Grambsch (2000). «The cox model». En: *Modeling survival data: extending the Cox model*. Springer, págs. 39-77.

Therneau, Terry M, Patricia M Grambsch y Thomas R Fleming (1990). «Martingale-based residuals for survival models». En: *Biometrika* 77(1), págs. 147-160.

Wood, Simon N (2013). «On p-values for smooth components of an extended generalized additive model». En: *Biometrika* 100(1), págs. 221-228.