



UNIVERSIDAD DE LA REPÚBLICA

Facultad de Ciencias Económicas y de Administración

Licenciatura en Estadística

Informe de Pasantía

**Análisis de supervivencia de las tablets del Plan Ceibal
2014-2015**

Guillermina Costabel

Tutores:

Marco Scavino

Natalia da Silva

Montevideo, Noviembre 2019.

UNIVERSIDAD DE LA REPÚBLICA
FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN

El tribunal docente integrado por los abajo firmantes aprueba el trabajo de
Pasantía:

**Análisis de supervivencia de las tablets del Plan Ceibal
2014-2015**

Guillermina Costabel

Tutores:
Marco Scavino
Natalia da Silva
Licenciatura en Estadística

Puntaje

Tribunal

Profesor.....(nombre y firma).

Profesor.....(nombre y firma).

Profesor.....(nombre y firma).

Fecha.....

Índice general

Índice general	III
Índice de figuras	VII
Índice de tablas	XI
1. Introducción	1
1.1. Motivación	2
1.2. Objetivos	3
2. Antecedentes	5
3. Marco teórico	7
3.1. Análisis de supervivencia	7
3.1.1. Datos censurados	8
3.1.2. Diseño de datos de supervivencia	8
3.1.3. Función de supervivencia o confiabilidad	10
3.1.4. Función de riesgo	10
4. Metodología	13
4.1. Notación	13
4.2. Estimación de la función de confiabilidad	14
4.2.1. Intervalos de confianza	15
4.2.2. Comparación de curvas de supervivencia	15

ÍNDICE GENERAL

4.3.	Modelo de Cox	18
4.3.1.	Estimación por máxima verosimilitud	19
4.3.2.	Interpretación del modelo	20
4.3.3.	Contrastes de hipótesis	20
4.3.4.	Estudio de los residuos del modelo	22
4.3.5.	El supuesto de riesgos proporcionales	24
4.3.6.	Curvas de supervivencia ajustadas	25
4.3.7.	Modelo extendido de Cox	25
4.4.	Estimación de la función de riesgo	27
4.4.1.	Elección de la forma del núcleo K	29
4.4.2.	Elección del ancho de banda b	29
4.4.3.	Efecto borde	30
4.4.4.	Intervalos de confianza	30
5.	Descripción y análisis exploratorio de los datos	33
5.1.	Universo de análisis y fuente de datos	34
5.2.	Análisis exploratorio de los datos de las tablets	34
5.2.1.	Primera vida de las tablets	35
5.2.2.	Segunda vida de las tablets	53
5.2.3.	Comparación entre la primera y segunda vida de las tablets .	57
6.	Resultados	63
6.1.	Estimación Kaplan-Meier	64
6.1.1.	Primera vida	64
6.1.2.	Segunda vida y comparación	67
6.2.	Estimación con variables auxiliares	68
6.2.1.	Estimación con el modelo de Cox	73
6.3.	Estimación de la función de riesgo	86
6.3.1.	Primera vida	86
6.3.2.	Segunda vida	89

7. Conclusiones y trabajos a futuro	91
Bibliografía	97
A. Apéndice estadístico	101
A.1. Propiedades del estimador por núcleos de la función de riesgo	101
A.2. Regla de Freedman-Diaconis (F-D)	102
B. Apéndice de resultados	105
B.1. Tiempo de vida hasta la primera defunción según variables sociodemográficas	105
B.2. Estimación Kaplan-Meier de la primera vida de las tablets	107
B.3. Test log-rank y Peto-Peto para comparación de las curvas de supervivencia de la primera y segunda vida	108
B.4. Estimación Kaplan-Meier de la primera vida considerando las variables explicativas	109
B.5. Test log-rank y Peto-Peto para comparación de las curvas de supervivencia correspondientes a las categorías de las variables auxiliares de la primera vida	111
B.6. Estimación de Cox aplicada a diferentes modelos	113
B.7. Función AIC para selección de modelos	114
B.8. Verificación del supuesto de riesgos proporcionales del modelo de Cox aplicado al modelo completo	115
B.9. Comparación de las estimaciones de la función de riesgo con diferentes tipos de núcleos	117

ÍNDICE GENERAL

Índice de figuras

4.1. Núcleo de Epanechnikov.	29
5.1. Estado final de la primera vida de las tablets según año de entrega.	37
5.2. Fechas de inicio de la primera vida de las tablets.	38
5.3. Fechas de finalización de la primera vida de las tablets según estado final.	40
5.4. Duración de la primera vida de las tablets por año de entrega.	41
5.5. Duración de la primera vida de las tablets por año de entrega según estado final.	42
5.6. Causas de defunción de la primera vida de las tablets según año de entrega.	45
5.7. Duración media y mediana de la primera vida de las tablets por causa de defunción.	46
5.8. Distribución de variables sociodemográficas: sexo, área geográfica y contexto sociocultural.	48
5.9. Estado final de la primera vida de las tablets según variables socio-demográficas.	49
5.10. Causas de defunción de la primera vida de las tablets según variables sociodemográficas	51
5.11. Duración de la primera vida de las tablets que presentaron defunciones según variables técnicas.	52
5.12. Estado final de la segunda vida de las tablets según año de entrega. . .	54

ÍNDICE DE FIGURAS

5.13. Duración de la segunda vida de las tablets por año de entrega.	56
5.14. Causas de defunción de la segunda vida de las tablets según año de entrega.	57
5.15. Comparación del estado final de la primera y la segunda vida de las tablets.	58
5.16. Comparación de la proporción de defunciones en distintos intervalos de tiempo de la primera y segunda vida de las tablets.	59
5.17. Comparación de la duración de la primera y segunda vida según estado final de las tablets.	60
5.18. Distribución de las causas de defunción de los equipos, de la primera y segunda vida	61
6.1. Estimación Kaplan-Meier de la función de supervivencia y función de riesgo acumulado de la primera vida de las tablets.	66
6.2. Funciones de supervivencia de la primera y la segunda vida estimadas por el método de Kaplan-Meier.	69
6.3. Funciones de supervivencia estimadas con el método de Kaplan-Meier para las distintas categorías de las variables auxiliares: sexo, área geográfica, contexto y entrada a ST.	71
6.4. Resultado gráfico del ajuste del modelo de Cox al grupo de la primera vida con las variables sexo, área, contexto y ST	77
6.5. Residuos tipo devianza para la estimación del modelo de Cox aplicado a la primera vida con las variables explicativas sexo, área, contexto y ST	79
6.6. Residuos tipo dfbetas para la estimación del modelo de Cox aplicado a la primera vida con las variables explicativas sexo, área, contexto y ST	80
6.7. Residuos tipo martingala para la estimación del modelo de Cox aplicado a la primera vida con las variables explicativas sexo, área, contexto y ST	81

Índice de figuras

6.8. Transformación log-log complementaria para verificar el supuesto de riesgos proporcionales (otra forma).	83
6.9. Estimación por núcleos de la función de riesgo de la primera vida de las tablets	88
6.10. Estimación por núcleos de la función de riesgo de la segunda vida de las tablets	90
B.1. Duración de la primera vida de las tablets que presentaron defunciones según variables sociodemográficas.	106
B.2. Verificación del supuesto de riesgos proporcionales del modelo de Cox, a través de los residuos de Schoenfeld escalados	116
B.3. Comparación de las funciones de riesgo de la primera vida de las tablets utilizando diferentes funciones núcleo	117

ÍNDICE DE FIGURAS

Índice de tablas

3.1. Diseño de datos de tiempos de supervivencia.	9
5.1. Categorías de la variable que contiene las causas de defunción.	44
6.1. Resultado del ajuste de un modelo de Cox al modelo que incluye las variables sexo, área, contexto y ST (modelo completo).	75
6.2. Verificación del supuesto de riesgos proporcionales del modelo de Cox aplicado a la primera vida de las tablets con las variables explicativas sexo, área, contexto y ST.	82
6.3. Estimación de un modelo extendido de Cox	86
B.1. Estimación Kaplan-Meier de la función de supervivencia de la primera vida de las tablets para intervalos de tiempo de 90 días.	107
B.2. Resumen de la estimación Kaplan-Meier de la función de supervivencia de la primera vida de las tablets.	108
B.3. Test log-rank que compara las funciones de supervivencia de la primera y segunda vida.	108
B.4. Test Peto-Peto que compara las funciones de supervivencia de la primera y segunda vida.	109
B.5. Estimación Kaplan-Meier de las funciones de supervivencia de cada categoría de las variables sexo, área, contexto y ST.	110
B.6. Test log-rank que compara las funciones de supervivencia de cada categoría de las variables sexo, área, contexto y ST.	111

ÍNDICE DE TABLAS

B.7. Test Peto-Peto que compara las funciones de supervivencia de cada categoría de las variables sexo, área, contexto y ST.	112
B.8. Resultado del ajuste del modelo de Cox al modelo A.	113
B.9. Resultado del ajuste del modelo de Cox al modelo B.	113
B.10. Resultado del ajuste del modelo de Cox al modelo C.	114
B.11. Comparación de modelos con la función AIC().	114

Capítulo 1

Introducción

Con el propósito de potenciar los procesos de aprendizaje utilizando herramientas tecnológicas de la información y la comunicación (TIC), se implementa en Uruguay en 2007 el Plan Ceibal, como plan de inclusión e igualdad de oportunidades.

Según la Ley Nro. 18.640 (Ceibal, 2010) , de creación del Centro Ceibal, se establece que uno de los principales cometidos del Plan, es “contribuir al ejercicio del derecho a la educación y a la inclusión social mediante acciones que permitan la igualdad de acceso al conocimiento y al desarrollo saludable de la infancia y la adolescencia” .

Desde su creación, cada niño y niña que ingresa al sistema educativo uruguayo cuenta con la posibilidad de acceder a una computadora personal con conexión a Internet en forma gratuita que proviene desde el centro educativo en el cual participa. De esta manera, el aprendizaje se ve ampliado con la posibilidad de, además de lo aprendido en clase, seleccionar información, desarrollar alternativas y saber utilizar la tecnología actual.

La primera entrega de dispositivos se realizó en 2007, en la Escuela Italia de Villa Cardal en el departamento de Florida comenzando de esta manera una serie de etapas que incluyen la implementación del sistema de entregas en primaria, secundaria y UTU en todo el país, la instalación del acceso inalámbrico e internet y el despliegue de las plataformas educativas permitiendo el acceso de forma gratuita a libros de

CAPÍTULO 1. INTRODUCCIÓN

textos como también la inclusión de la familia. Plan Ceibal contribuyó a disminuir la brecha de acceso a la computadora, entre los quintiles de mayor y menor ingreso, consolidando un escenario de equidad que se mantiene estable desde 2010.

Como se menciona en el sitio web de Ceibal (2019), la misión del Plan es la integración de la tecnología al servicio de la educación, con el fin de mejorar su calidad, impulsando la innovación social, la inclusión y el crecimiento personal; mientras que su visión es contribuir con una educación innovadora que permita desarrollar el aprendizaje, la creatividad y el pensamiento crítico.

Con el “Modelo uno a uno” que se implementó, se ha podido contemplar la entrega de una laptop o tablet a cada alumna, alumno y docente de la enseñanza pública básica (Educación Inicial y Primaria y Educación Media). Así mismo, para la permanencia en el tiempo durante la trayectoria del alumno o alumna en su ciclo educativo, el Plan Ceibal planifica el mantenimiento de los servicios de soporte del parque de dispositivos, comprando repuestos, realizando reparaciones de partes y reparaciones de equipos, implementando también un servicio de reparación móvil que permite que todos los dispositivos sean reparados en prácticamente cualquier lugar geográfico del territorio nacional bajo cualquier estado de conservación.

1.1. Motivación

Dado el plan de entregas y recambio, todos los años se otorgan, por parte de Ceibal, equipos nuevos para todos los alumnos y las alumnas de primer y cuarto año de educación primaria y de primero de educación media básica. Por tanto, el conocer la duración aproximada de cada uno de ellos y los tipos de roturas más frecuentes, además del porcentaje de dispositivos que quedan totalmente rotos, ayudaría a tomar decisiones a futuro sobre qué equipo comprar pudiendo así reducir algunos costos y asegurar la equidad en el acceso. Por otra parte, la tarea de soporte es considerablemente importante dentro del presupuesto anual del Plan Ceibal, por tanto

toda información que contribuya a la optimización del uso de los recursos (técnicos, financieros), es fundamental.

Esta investigación contribuiría además con el objetivo propuesto por Ceibal y su política de equidad en el acceso a las TICs generando evidencia empírica con una metodología que permita ampliar los conocimientos generando nuevos antecedentes para luego ser extrapolada a los restantes modelos de equipos del parque.

1.2. Objetivos

El objetivo general de este trabajo es analizar la supervivencia de un modelo de tablet entregado por Ceibal, y sus principales causas de rotura, indagando la relación entre la incidencia del tipo de rotura y la esperanza de vida del equipo con las características sociodemográficas de las alumnas y los alumnos.

En función de éste, surgen los siguientes objetivos específicos:

- Realizar un análisis exploratorio de los datos para identificar posibles variables relevantes para modelizar la supervivencia de las tablets.
- Conocer las principales causas de rotura que presenta este dispositivo a lo largo de su trayectoria y los tiempos asociados a cada una de ellas.
- Estimar la función de supervivencia de la generación de tablets que presentan una primera vida (desde el momento en que el equipo es entregado hasta su primera reparación) y de la generación que presentan una segunda vida (desde que el equipo se vuelve a entregar después de su primera reparación hasta su segunda reparación).
- Comparar el tiempo de vida, la supervivencia y los patrones de rotura de la primera y segunda vida.
- Estimar los efectos de las características sociodemográficas del alumnado y

CAPÍTULO 1. INTRODUCCIÓN

características técnicas de las tablets, sobre el grupo primera vida, aplicando diferentes métodos de análisis.

- Estimar la función de riesgo de la generación de tablets que presentan una primera vida y la de la generación que presentan una segunda vida y compararlas.

En el siguiente repositorio de github se encuentra todo el material necesario para reproducir este trabajo. Debido a la confidencialidad de los datos el repositorio es privado, para acceder al mismo se deberá solicitar el permiso necesario.

<https://github.com/GCost14/TrabajoFinalIEsta.git>

Por otro lado, en este otro repositorio de público acceso se puede encontrar solo un archivo .Rnw con el código de R (R Core Team. R foundation for statistical computing, 2018) utilizado para obtener todos los resultados.

<https://github.com/GCost14/Trabajo-final-grado-Lic.Estadistica.git>

Capítulo 2

Antecedentes

El presente trabajo tiene como antecedente principal el documento de Marconi (2016) donde se estudia la supervivencia de las laptops XO entregadas por Ceibal en el año 2010 a alumnas y alumnos de Educación Primaria Pública de Uruguay teniendo como objetivo analizar las funciones de supervivencia de los equipos de diferentes perfiles sociodemográficos del alumnado. Se utilizó la técnica de análisis de la historia de acontecimientos para la obtención de las tasas de riesgo y estimación de las funciones de supervivencia. Se evidenció que el 70 % de la generación presentó una rotura total siendo la principal causa de rotura el teclado de la laptop. El perfil sociodemográfico de alumnas y alumnos que presenta mayor tasa de riesgo a la rotura total de la XO, son los varones de Montevideo de contextos socioculturales más desfavorables, que han repetido algún año y con registros previos de reparación en servicio técnico.

Por otra parte, el análisis de supervivencia ha sido utilizado en variadas disciplinas, principalmente en el ámbito de la salud, también en problemas de ingeniería, lo que se conoce en la literatura como análisis de confiabilidad.

En un texto dirigido a ingenieros y estadísticos industriales que trabajan con datos de duración de vida de productos, Nelson (2005) presenta entre varias aplicaciones a problemas de ingeniería, un ejemplo sobre datos de fallas de ventiladores de setenta

CAPÍTULO 2. ANTECEDENTES

generadores. Para cada uno se registró el número de horas de funcionamiento desde la primera vez que se puso en servicio, hasta la falla del ventilador o hasta el final del estudio (lo que ocurriera primero). Un problema fue estimar el porcentaje de falla en la garantía. Otro fue determinar si la tasa de rotura de los ventiladores disminuía o aumentaba con la edad; es decir, ¿el problema mejoraría o empeoraría a medida que los ventiladores restantes envejecieran? Esta información ayudó a la gerencia a decidir si reemplazar o no a los ventiladores sin fallas.

Así mismo, entre varios ejemplos de Phillips (2003) es de interés el que refiere a datos de fallas censuradas a la derecha en el que se analizan los tiempos de rotura (en días) de 16 sistemas de telecomunicaciones instalados en el año 1985 y los tiempos de censura de 109 sistemas restantes, totalizando 125 clientes. Los sistemas podrían retirarse del servicio en cualquier momento durante el período de observación a petición del cliente. Tanto como la falta de funcionamiento, un nivel inaceptable de estática, interferencia o ruido en la transmisión del sistema de telecomunicaciones se consideró como una falla del sistema.

Estos datos fueron presentados originalmente por Kim and Proschan (1991) y utilizados para obtener una estimación no paramétrica suave de la función de confiabilidad de los sistemas de telecomunicación, a través del modelo exponencial constante a intervalos.

A nivel nacional, se ha aplicado la técnica de análisis de supervivencia o análisis de confiabilidad en distintos ámbitos. A modo de ejemplos, Triunfo et~al. (2010) estimaron la tasa de mortalidad infantil a partir de los nacimientos ocurridos en Uruguay entre 2002 y 2003 y las defunciones ocurridas en el primer año de vida. Amarante and Dean (2012) estimaron la función de supervivencia y la función de riesgo de duración en el empleo formal al estudiar el mercado de trabajo uruguayo en el período 1997-2009, en base a una muestra de historias laborales de la seguridad social proveniente de los registros administrativos del Banco de Previsión Social (BPS).

Capítulo 3

Marco teórico

Este capítulo ofrece una descripción general del enfoque del estudio de datos llamado análisis de supervivencia. Se incluye el tipo de problema abordado por dicha técnica, la variable de respuesta considerada (Sección 3.1), la necesidad de tener en cuenta datos censurados (Sección 3.1.1), así como lo que representa una función de supervivencia y una función de riesgo (Secciones 3.1.3 y 3.1.4 respectivamente).

3.1. Análisis de supervivencia

En general, el análisis de supervivencia es un conjunto de procedimientos estadísticos para el análisis de datos para los cuales la variable de respuesta de interés es el *tiempo hasta que ocurre un evento*.

El *tiempo* puede estar medido en años, meses, semanas o días desde el comienzo del seguimiento de un individuo (tiempo inicial) hasta que ocurre un evento (tiempo final) y se refiere a él como, **tiempo de supervivencia**, porque es el tiempo que un individuo ha sobrevivido durante un período de seguimiento.

Un **evento** representa un cambio de la unidad de análisis de un estado de origen a otro, tales como muerte, incidencia de enfermedad, recaída de remisión, falla de

CAPÍTULO 3. MARCO TEÓRICO

un equipo, recuperación (por ejemplo, regreso al trabajo) o cualquier experiencia designada de interés que pueda sucederle a un individuo.

Cuando se considera más de un evento (por ejemplo, muerte por una de varias causas), el problema estadístico se puede caracterizar como un evento recurrente o un problema con riesgos competitivos (competing risk).

3.1.1. Datos censurados

Una característica especial de los datos de supervivencia que hace que los métodos estándar para datos completos no sean apropiados, es que contienen observaciones censuradas. En esencia, la censura se produce cuando no se conoce exactamente el tiempo de supervivencia de la unidad de análisis.

Se considera como **censura a la derecha** el caso en el cual no se observa el punto final de interés para cierto individuo. El intervalo de tiempo se ha cortado (es decir, censurado) en el lado derecho, obteniendo un tiempo de supervivencia censurado menor al tiempo de supervivencia real, el cual no se conoce.

3.1.2. Diseño de datos de supervivencia

El diseño de datos expresado en la Tabla 3.1 ayuda a comprender cómo funciona un análisis de supervivencia.

La primer columna de esta tabla muestra tiempos de falla ordenados en forma ascendente. Los mismos se denotan con subíndices entre paréntesis, comenzando con $t_{(0)}$ hasta $t_{(k)}$, siendo k la cantidad de tiempos de fallas de los individuos. Se deben excluir antes del ordenamiento los tiempos asociados a momentos censurados.

En la segunda columna se muestra la frecuencia de los sujetos que experimentaron el evento en cada tiempo de falla, denotada por m_f . La suma de todos los m_f en esta columna indica la cantidad total de fallas respecto al total de individuos

considerados.

La tercera columna proporciona la frecuencia, denotada por q_f , de los individuos censurados en el intervalo de tiempo que comienza con el tiempo de falla $t_{(f)}$ hasta el siguiente tiempo de falla $t_{(f+1)}$. La suma de estos valores da el número total de observaciones censuradas.

La última columna en la tabla muestra el **conjunto de riesgo**. Por definición, el conjunto de riesgo, $R(t_{(f)})$, es el conjunto de individuos que han sobrevivido al menos hasta el tiempo $t_{(f)}$; es decir, cada sujeto en $R(t_{(f)})$ tiene un tiempo de supervivencia que es $t_{(f)}$ o mayor, independientemente de si el individuo ha fallado o está censurado.

Tabla 3.1: Diseño de datos de tiempos de supervivencia.

$t_{(f)}$	m_f	q_f	$R(t_{(f)})$
$t_{(0)} = 0$	$m_0 = 0$	q_0	$R(t_{(0)})$
$t_{(1)}$	m_1	q_1	$R(t_{(1)})$
$t_{(2)}$	m_2	q_2	$R(t_{(2)})$
\vdots	\vdots	\vdots	\vdots
\vdots	\vdots	\vdots	\vdots
\vdots	\vdots	\vdots	\vdots
$t_{(k)}$	m_k	q_k	$R(t_{(k)})$

A continuación se hace referencia a dos términos considerados en cualquier análisis de supervivencia. Estos son la *función de supervivencia*, denotada por $S(t)$, y la *función de riesgo*, denotada por $h(t)$. La primera de ellas también es conocida en la literatura como *función de confiabilidad* (o reliability function), sobre todo en el área de ingeniería.

3.1.3. Función de supervivencia o confiabilidad

Kleinbaum and Klein (2010) introducen la noción de función de confiabilidad o supervivencia como la probabilidad de que un individuo sobreviva más que un tiempo t determinado.

Definición 3.1 *Sea el tiempo de supervivencia T , es decir el tiempo hasta que determinado evento ocurra, una variable aleatoria no negativa con función de distribución absolutamente continua F . La **función de supervivencia** $S(t)$, es la probabilidad de que el tiempo de supervivencia T sea mayor que el tiempo t . Esto es,*

$$S(t) = P(T \geq t) = 1 - F(t). \quad (3.1)$$

3.1.4. Función de riesgo

Se define la **función de riesgo** $h(t)$, como el potencial instantáneo por unidad de tiempo para que ocurra el evento, dado que el individuo ha sobrevivido hasta el tiempo t (Kleinbaum and Klein, 2010).

Definición 3.2 *Sea la variable aleatoria T , el tiempo de supervivencia. Se define la función de riesgo como:*

$$h(t) = \lim_{\delta t \rightarrow 0} \frac{P(t \leq T < t + \delta t \mid T \geq t)}{\delta t}. \quad (3.2)$$

Debido al suceso condicional en la probabilidad, la función de riesgo a veces es llamada tasa de falla condicional.

Horová et~al. (2012) definen la función de riesgo como la probabilidad de que un individuo presente un evento en el tiempo t , condicionado en que haya sobrevivido hasta ese tiempo.

Definición 3.3 *Si la distribución del tiempo de vida F tiene una densidad f , para $S(t) > 0$, la función de riesgo está definida por*

$$h(t) = \frac{f(t)}{S(t)}, \quad (3.3)$$

y la **función de riesgo acumulado** como

$$H(t) = -\log S(t). \quad (3.4)$$

Hay una clara relación entre la función de supervivencia y la función de riesgo.

De hecho, si uno conoce la forma de $S(t)$, puede derivar la $h(t)$ correspondiente, y viceversa a través de las siguientes fórmulas:

$$S(t) = \exp \left[- \int_0^t h(u) du \right], \quad h(t) = - \left[\frac{S'(t)}{S(t)} \right].$$

CAPÍTULO 3. MARCO TEÓRICO

Capítulo 4

Metodología

En este capítulo se presentan diferentes métodos no paramétricos usualmente utilizados para estimar la función de confiabilidad (Secciones 4.2 y 4.3) y la función de riesgo (Sección 4.4) del conjunto de datos con observaciones censuradas a la derecha, junto con respectivas pruebas de hipótesis e intervalos de confianza.

4.1. Notación

En primer lugar se resumen algunos conceptos básicos y se introduce la notación para el modelo de tiempo de vida con censura aleatoria a la derecha (Muller and Wang, 1994).

Sean T_1, T_2, \dots, T_n tiempos de vida independientes e idénticamente distribuidos (i.i.d.) con función de distribución F , y sean C_1, \dots, C_n tiempos de censura i.i.d. con función de distribución G , tal que C y T son independientes.

Se observan los pares

$$(X_i, \delta_i), \quad i = 1, 2, \dots, n,$$

donde $X_i = \min(T_i, C_i)$ y $\delta_i = I_{\{X_i=T_i\}}$, $i = 1, \dots, n$ indica si la observación es

CAPÍTULO 4. METODOLOGÍA

censurada o no. Las X_i son i.i.d. con función de supervivencia S , siendo,

$$S(x) = \bar{F}(x)\bar{G}(x),$$

con $\bar{F}(x) = P(T_1 \geq x)$ y $\bar{G}(x) = P(C_1 \geq x)$ las funciones de supervivencia de $\{T_i\}$ y $\{C_i\}$ respectivamente.

4.2. Estimación de la función de confiabilidad

La estimación no paramétrica más simple de la función de supervivencia para observaciones completas es una función de supervivencia empírica. Kaplan y Meier propusieron una extensión para datos con censuras (?).

(Horová et al., 2012) presentan la estimación de la función de supervivencia, $\hat{S}(x)$, como se explica a continuación.

En primer lugar se deben ordenar los tiempos de falla de forma ascendente. Luego, para cada uno de los tiempos de falla, la probabilidad de supervivencia estimada se calcula de la siguiente manera:

$$\hat{S}_{KM}(x) = \prod_{j:X_{(j)} < x} \left(\frac{k-j}{k-j+1} \right)^{\delta_{(j)}}, \quad (4.1)$$

donde $X_{(j)}$ denota el j -ésimo estadístico de orden de X_1, X_2, \dots, X_k y $\delta_{(j)}$ el indicador del estado de censura correspondiente.

La fórmula general para una probabilidad de supervivencia de Kaplan-Meier en el tiempo de falla $t_{(j)}$ indica la probabilidad de sobrevivir más allá del tiempo de falla anterior $t_{(j-1)}$, multiplicado por la probabilidad condicional de sobrevivir más que $t_{(j)}$, dada la supervivencia hasta al menos el tiempo $t_{(j)}$. Esto es:

$$\hat{S}(t_{(j)}) = \prod_{i=1}^{j-1} \hat{P}[T > t_{(i)} \mid T \geq t_{(i)}] = \hat{S}(t_{(j-1)}) \times \hat{P}(T > t_{(j)} \mid T \geq t_{(j)}).$$

$\hat{S}_{KM}(x)$ es una función escalonada y continua a la derecha. El estimador $\hat{S}_{KM}(x)$ de la función de supervivencia de las observaciones X_1, \dots, X_n se denomina estimador

límite de producto (product-limit) debido a su derivación, como límite, cuando el tiempo se divide en intervalos y la longitud del intervalo tiende a 0.

4.2.1. Intervalos de confianza

Se describe como calcular intervalos de confianza para las curvas estimadas por el método de Kaplan-Meier (Kleinbaum and Klein, 2010).

La fórmula de los intervalos con una confianza puntuales del 95% para la probabilidad estimada de Kaplan-Meier, tiene la siguiente fórmula general:

$$\hat{S}_{KM}(t) \pm 1,96 \sqrt{\hat{Var}[\hat{S}_{KM}(t)]}, \quad (4.2)$$

donde $\hat{S}_{KM}(t)$ denota la estimación Kaplan-Meier de la función de supervivencia en el tiempo t y $\hat{Var}[\hat{S}_{KM}(t)]$ denota la estimación de la varianza de $\hat{S}_{KM}(t)$. El enfoque más común utilizado para estimar esta varianza utiliza la *fórmula de Greenwood*:

$$\hat{Var}[\hat{S}_{KM}(t)] = (\hat{S}_{KM}(t))^2 \sum_{j:t_{(j)} \leq t} \left[\frac{m_j}{r_j(r_j - m_j)} \right],$$

siendo $t_{(j)}$ el tiempo de falla ordenado j , m_j el número de fallas en el tiempo $t_{(j)}$, y r_j el número de observaciones en el conjunto de riesgo al tiempo $t_{(j)}$.

La sumatoria representa, en cada tiempo de falla $t_{(j)}$, un promedio ponderado (por $1/r_j$) del riesgo condicional de fallar antes de $t_{(j)}$.

4.2.2. Comparación de curvas de supervivencia

Se describe en primer lugar cómo evaluar si las curvas Kaplan-Meier para dos grupos son estadísticamente equivalentes¹, testeándose lo siguiente:

¹Cuando se afirma que dos o más curvas de Kaplan-Meier son “estadísticamente equivalentes”, se quiere decir que, en base a un procedimiento de testeo que compara dichas curvas de supervivencia, no hay evidencia que indique que sean diferentes.

CAPÍTULO 4. METODOLOGÍA

$$H_0 : \hat{S}_1(t) = \hat{S}_2(t), 0 < t < \infty,$$

$$H_1 : \hat{S}_1(t) \neq \hat{S}_2(t).$$

Para construir el estadístico de contraste basta con calcular el número esperado de fallas y la varianza estimada del número de muertes para uno de los grupos; por ejemplo, para el grupo 1 el número esperado de muertes se calcula de la siguiente manera:

$$\hat{e}_1(t_i) = \frac{r_1(t_i)m(t_i)}{r(t_i)},$$

donde $r(t_i)$ y $m(t_i)$ son el número de individuos en riesgo y el número de muertes (o de ocurrencia del evento de interés) en el momento t_i , y $r_1(t_i)$ es la cantidad de individuos en riesgo del grupo 1 en el tiempo t_i .

La varianza estimada de $m_1(t_i)$ está basada en la distribución hipergeométrica y para el grupo 1 está definida como:

$$\hat{V}(m_1(t_i)) = \frac{r_1(t_i)r_2(t_i)(r(t_i) - m(t_i))}{r^2(t_i)(r(t_i) - 1)},$$

siendo $r_2(t_i)$ la cantidad de individuos en riesgo del grupo 2 en el tiempo t_i .

Finalmente, el estadístico de contraste se define de la siguiente manera:

$$Q = \frac{\left[\sum_{t=1}^k w_i(m_1(t_i) - \hat{e}_1(t_i)) \right]^2}{\sum_{t=1}^k w_i^2 \hat{V}(m_1(t_i))}. \quad (4.3)$$

En la Ecuación 4.3, k es el número de tiempos de ocurrencia de eventos en ambos grupos y w_i denota los pesos que toman valores distintos dependiendo del test utilizado.

El más común de los test es el de Mantel (1966) conocido como test de rangos logarítmicos (o log-rank), el cual está diseñado para verificar igualdad o diferencia en la función de supervivencia en todos los tiempos. En este test los pesos son iguales a 1, es decir, $w_i = 1$.

Otro de los test comúnmente utilizados es el de Peto and Peto (1972), el cual permite verificar igualdad o diferencia de las funciones de supervivencia en los tiempos

4.2. Estimación de la función de confiabilidad

iniciales. En este test los pesos toman la forma:

$$w_i = \tilde{S}(t_{i-1}) \frac{r(t_i)}{r(t_i) - 1},$$

donde $\tilde{S}(t)$ es el estimador de la función de supervivencia definido por:

$$\tilde{S}(t) = \prod_{t_i \leq t} \left(\frac{r(t_i) + 1 - m(t_i)}{r(t_i) + 1} \right).$$

Otra forma de estudiar los test anteriores fue propuesta por Harrington and Fleming (1982) quienes sugirieron pesos de la forma:

$$w_1 = \left[\hat{S}_{KM}(t_{i-1}) \right]^\rho,$$

y haciendo $\rho = 0$ se tiene que $w_i = 1$, test log-rank, y, si $\rho = 1$, se obtiene el test de Peto-Peto.

Puede demostrarse que para un número de observaciones suficientemente grande, la distibución del estadístico Q puede aproximarse, bajo la hipótesis nula que asume que las dos funciones de supervivencia son iguales, con una distribución χ^2_1 (chi-cuadrado de un grado de libertad).

Por otro lado, para el caso de más de dos grupos, se define a continuación solamente la fórmula del estadístico de log-rank, siendo la hipótesis a testear:

$$\begin{aligned} H_0 : \quad & \hat{S}_1(t) = \hat{S}_2(t) = \dots = \hat{S}_G(t), \\ H_1 : \quad & \text{Al menos } \hat{S}_i \neq \hat{S}_l \text{ para cualquier } i, l \in [1, G]. \end{aligned}$$

Sea $i = 1, 2, \dots, G$ la cantidad de grupos y $j = 1, 2, \dots, k$, la cantidad de tiempos de falla diferentes:

- r_{ij} es la cantidad en riesgo en el i -ésimo grupo en el j -ésimo tiempo de falla ordenado,
- m_{ij} es la cantidad de fallas observadas en el i -ésimo grupo en el j -ésimo tiempo de falla ordenado,

CAPÍTULO 4. METODOLOGÍA

- e_{ij} es la cantidad de fallas esperadas en el i -ésimo grupo en el j -ésimo tiempo de falla ordenado, igual a $\left(\frac{r_{ij}}{r_{1j} + r_{2j}}\right)(m_{1j} + m_{2j})$,
- $r_j = \sum_{i=1}^G r_{ij}$,
- $m_j = \sum_{i=1}^G m_{ij}$,
- $O_i - E_i = \sum_{j=1}^k (m_{ij} - e_{ij})$,
- $Var(O_i - E_i) = \sum_{j=1}^k \left(\frac{r_{ij}(r_j - r_{ij})m_{ij}(r_j - m_j)}{r_j^2(r_j - 1)} \right)$,
- $\mathbf{d} = (O_1 - E_1, O_2 - E_2, \dots, O_{G-1} - E_{G-1})'$,
- $\mathbf{V} = ((v_{il}))$ donde $v_{ii} = Var(O_i - E_i)$ y $v_{il} = Cov(O_i - E_i, O_l - E_l)$ para $i = 1, 2, \dots, G-1; l = 1, 2, \dots, G-1$.

Luego, el estadístico log-rank viene dado por la fórmula del producto matricial:

$$Log - rank = \mathbf{d}' \mathbf{V}^{-1} \mathbf{d},$$

que se distribuye aproximadamente χ^2_{G-1} bajo la hipótesis nula de que los G grupos tienen igual curva de supervivencia.

4.3. Modelo de Cox

El modelo de riesgos proporcionales de Cox (1972) establece una expresión del riesgo en el tiempo t para un individuo con un conjunto específico de variables explicativas.

Formalmente, según Peña (2005), el riesgo para un individuo se define de la siguiente manera:

$$h(t; X) = h_0(t) e^{\sum_{i=1}^p \beta_i X_i}, \quad (4.4)$$

donde $X = (X_1, X_2 \dots X_p)$ es el vector de variables explicativas o predictoras para ese individuo.

La fórmula del modelo de Cox dice que el riesgo en el tiempo t es el producto de dos cantidades. La primera de ellas $h_0(t)$, es llamada función de riesgo base y es llamada

así porque la fórmula del modelo cumple la propiedad de que si todas las X 's son iguales a 0, o visto desde otra perspectiva, no hay X 's en el modelo, la fórmula se reduce a esa función. Además es una función que se asume desconocida, la que hace al modelo de Cox un modelo semiparamétrico. Es una función de t que no involucra las X 's. La segunda cantidad es la expresión exponencial a la suma lineal de $\beta_i X_i$, donde la suma es sobre las p variables explicativas X .

Otra propiedad interesante del modelo de Cox, es que, a pesar de que la parte de riesgo base del modelo no está especificada, es posible estimar los β 's en la parte exponencial. A continuación se describe como pueden obtenerse estimaciones de los parámetros del modelo de Cox, utilizando un enfoque basado en la función de verosimilitud.

4.3.1. Estimación por máxima verosimilitud

Las estimaciones de máxima verosimilitud de los parámetros del modelo de Cox se derivan al maximizarse una función de probabilidad, que generalmente se denota como L . La función de verosimilitud es una expresión matemática que describe la probabilidad conjunta de obtener los datos realmente observados, en función de los parámetros desconocidos (β 's) en el modelo considerado.

La fórmula para la función de verosimilitud del modelo de Cox en realidad se llama *función de verosimilitud parcial*. Este término se usa porque la fórmula de verosimilitud considera las probabilidades solo para aquellos individuos que presentan el evento, y no considera explícitamente las probabilidades para aquellos que son censurados.

Entonces, definiendo $\lambda_i = e^{\beta_i X_i}$ y suponiendo que una muerte ha ocurrido en el tiempo t^* , la verosimilitud de que la muerte le ocurra al individuo i -ésimo y no a otro individuo es:

$$L_i(\beta) = \frac{h(t^*, X_i)}{\sum_{l \in R(t^*)} h(t^*, X_l)} = \frac{h_0(t^*) \lambda_i(t^*)}{\sum_{l \in R(t^*)} h_0(t^*) \lambda_l(t^*)} = \frac{\lambda_i(t^*)}{\sum_{l \in R(t^*)} \lambda_l(t^*)}. \quad (4.5)$$

CAPÍTULO 4. METODOLOGÍA

En particular, la verosimilitud parcial puede escribirse como un producto de varias verosimilitudes, una para cada timepo de falla k , y se expresa como:

$$L = L_1 \times L_2 \times L_3 \times \dots \times L_k = \prod_{i=1}^k L_i.$$

Una vez que se forma la función de verosimilitud para un modelo dado, el siguiente paso es maximizar esta función de los parámetros o, de manera equivalente su logaritmo natural (la función de log-verosimilitud).

4.3.2. Interpretación del modelo

La interpretación del modelo de Cox no se hace directamente a través de su parámetro estimado sino del exponencial de dicho parámetro estimado, $e^{\hat{\beta}_i}$, para la variable X_i .

Para variables dicotómicas $e^{\hat{\beta}_i}$ es un estimador de la razón de riesgos y se interpreta como la cantidad de riesgo que se tiene con la presencia de cada covariante en relación a la ausencia del resto de las covariables.

Los intervalos de confianza al 95 % para $e^{\hat{\beta}_i}$ se obtienen mediante:

$$\exp \left[\hat{\beta}_i \pm 1,96 se(\hat{\beta}_i) \right],$$

donde $se(\hat{\beta}_i) = \sqrt{\hat{Var}(\hat{\beta}_i)}$ es el error estándar de $\hat{\beta}_i$.

Para el caso de covariantes continuas, $e^{\hat{\beta}_i}$ representa la razón de riesgos al incrementar en una unidad la covariante X_i .

4.3.3. Contrastes de hipótesis

Una vez que se ajusta el modelo de Cox, existen tres contrastes de hipótesis para verificar la significación del modelo testenado $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$: test de razón de verosimilitud, test de Wald y test de puntajes. A menudo estos test arrojan resultados similares.

El test de Wald tiene una interpretación más directa que el test de verosimilitud y el de puntajes, sin embargo no es invariante ante diferentes parametrizaciones y los otros dos si. Con el test de puntajes solo hace falta maximizar bajo la hipótesis nula, con lo que si hay que realizar el test para varios parámetros es más rápido computacionalmente. Sin embargo, el test de máxima verosimilitud converge más rápido hacia la distribución normal.

4.3.3.1. Test de razón de verosimilitud

Es el que presenta mayor confiabilidad. En este contraste se utiliza el valor de la función de verosimilitud parcial evaluada en $\hat{\beta}$, $L(\hat{\beta})$, y evaluada en β_0 , $L(\hat{\beta}_0)$:

$$\chi_{LR} = -2\{\log(L(\beta_0)) - \log(L(\hat{\beta}))\}, \quad (4.6)$$

que bajo la hipótesis nula sigue una distribución χ_p^2 .

4.3.3.2. Test de Wald

Es quizás el test mas natural debido a que proporciona un contraste por variables en lugar de una medida de significación global. Se basa en que los parámetros $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)$ siguen una distribución aproximadamente normal con media $(\hat{\beta}_1, \dots, \hat{\beta}_p)$ y matriz de varianzas y covarianzas $\hat{\Sigma}_{\hat{\beta}}$. El estadístico se define como:

$$\chi_W = (\hat{\beta} - \beta_0)^t \hat{\Sigma}_{\hat{\beta}}^{-1} (\hat{\beta} - \beta_0), \quad (4.7)$$

que bajo la hipótesis nula sigue una distribución χ_p^2 .

4.3.3.3. Test de puntajes (score test)

En este contraste se utiliza el gradiente (derivadas) del logaritmo de la verosimilitud parcial evaluada en la hipótesis nula y supone que bajo la hipótesis nula el vector de puntajes:

$$\chi_S = \left(\frac{\partial L(\beta_0)}{\partial \beta} \right)^t \left(-\frac{\partial L^2(\beta_0)}{\partial \beta \partial \beta^t} \right)^{-1} \frac{\partial L(\beta_0)}{\partial \beta}, \quad (4.8)$$

CAPÍTULO 4. METODOLOGÍA

es aproximadamente multinormal de medias 0 y matriz de varianzas y covarianzas $\hat{\Sigma}_{\beta}$, con lo que el estadístico sigue también una distribución χ_p^2 .

4.3.4. Estudio de los residuos del modelo

Los residuos se representan gráficamente frente a alguna cantidad, y el patrón observado se utiliza para diagnosticar posibles problemas con el modelo ajustado. Algunos residuos tienen la propiedad adicional de no solo indicar problemas sino también sugerir correcciones.

Existen cuatro tipos de residuos de interés en el modelo de Cox: los residuos de martingala, los de desvíos (devianza), los de puntaje (score) y los de Schoenfeld. De estos cuatro residuos pueden derivarse otros dos: los dfbetas y los residuos escalados de Schoenfeld.

A continuación se definen y explican brevemente cada uno de ellos (Moore, 2016).

4.3.4.1. Residuos martingala

Los residuos martingala son una transformación de los residuos denominados Cox-Snell. Tienen una distribución asimétrica y su esperanza debería ser asintoticamente 0. Son útiles para indicar si con las covariables del modelo se han predicho bien los tiempos de supervivencia. Sirven para analizar transformaciones de las covariables que mejoren dichos residuos. Su fórmula es:

$$R_{M_i} = \delta_i - R_i, \quad (4.9)$$

para $i = 1, \dots, n$, siendo δ_i el indicador del si el sujeto está o no censurado. Donde $R_i = -\log(\hat{S}(t_i, X_i))$ son los denominados residuos de Cox-Snell extendidos. Los residuos de Cox-Snell se utilizan del siguiente modo: si el modelo de riesgos proporcionales estimado es adecuado el gráfico de dichos residuos y su correspondiente curva Kaplan-Meier, $\hat{S}(R)$, aparecerían como una recta de 45 grados.

4.3.4.2. Residuos de desvíos (devianza)

Los residuos de desvíos si se distribuyen de forma simétrica alrededor del cero y sirven para analizar el ajuste del modelo para cada sujeto no censurado y por lo tanto detectar observaciones atípicas. Su fórmula es:

$$R_{D_i} = \text{signo}(R_{M_i}) \sqrt{2[-R_{M_i} - \delta_i \log(\delta_i - R_{M_i})]}, \quad (4.10)$$

para $i = 1, \dots, n$.

4.3.4.3. Residuos de puntajes (scores)

Los residuos de puntajes se calculan para cada sujeto y cada covariante. Los gráficos de estos versus las covariables (una a una) indican la influencia de los individuos en la estimación de los parámetros. Los *dfbeta* y *dfbetas* (el dfbeta estandarizado) se calculan para cada sujeto e indican el cambio aproximado en el vector de parámetros si el sujeto concreto no estuviese en el modelo. Su fórmula es:

$$R_{P_{ij}}(t) = \int_0^t X_{ij}(u) - \bar{X}_{ij}(u) d\hat{M}_i(u), \quad (4.11)$$

para $i = 1, \dots, n$. Donde:

$$\bar{X}_j(u) = \frac{\sum_{i=1}^n J_i(t) X_{ij} e^{\beta X_i(t)}}{\sum_{i=1}^n J_i(t) e^{\beta X_i(t)}} \quad y \quad \hat{M}_i(u) = N_i(u) - \int_0^u J_i(s) e^{\beta X_i(s)} d\hat{H}_0(s),$$

siendo $J_i(t) = 1$ si el individuo i está en riesgo antes del tiempo t , y $N_i(t)$ el indicador del proceso de conteo (*counting process*) de si el individuo i -ésimo ha tenido un evento. Se observa que estos residuos tienen en cuenta variables dependientes del tiempo y su expresión se particulariza cuando las variables son independientes del tiempo.

4.3.4.4. Residuos de Schoenfeld

Los residuos de Schoenfeld se calculan para cada covariante y para cada individuo.

Se definen como:

$$R_{S_{ij}} = \delta_i \left(X_{ij} - \frac{\sum_{l \in R(t_{(i)})} X_{lj} e^{\hat{\beta} X_l}}{\sum_{l \in R(t_{(i)})} e^{\hat{\beta} X_l}} \right), \quad (4.12)$$

para $i = 1, \dots, n$ y $j = 1, \dots, p$, siendo δ_i el indicador de si el sujeto está censurado o no. Por lo tanto estos residuos solo están definidos para los individuos no censurados.

Los residuos de Schoenfeld son útiles para la verificación del supuesto de riesgo proporcional en el modelo de Cox el cual se expresa en la siguiente subsección.

4.3.5. El supuesto de riesgos proporcionales

El supuesto de riesgos proporcionales requiere que el riesgo para un individuo sea proporcional al riesgo para cualquier otro individuo, siendo la constante de proporcionalidad independiente del tiempo. Esto es:

$$\frac{\hat{h}(t, X^*)}{\hat{h}(t, X)} = \frac{\hat{h}_0(t) e^{\sum_{i=1}^p \hat{\beta}_i X_i^*}}{\hat{h}_0(t) e^{\sum_{i=1}^p \hat{\beta}_i X_i}} = e^{\sum_{i=1}^p \hat{\beta}_i (X_i^* - X_i)},$$

donde $X^* = (X_1^*, X_2^*, \dots, X_p^*)$ y $X = (X_1, X_2, \dots, X_p)$ denotan el conjunto de X' s para dos individuos.

Por lo tanto, la expresión final para la relación de riesgo involucra los coeficientes β_i gorro estimados y los valores de X^* y X para cada variable. Sin embargo, debido a que el riesgo base se cancela, la expresión final no depende del tiempo t .

Para una covariante en particular se cumplirá la hipótesis de proporcionalidad de los riesgos si los residuos de Schoenfeld de dicha covariante no están correlacionados con los tiempos de supervivencia. Gráficamente, si se dibujan los residuos de Schoenfeld de la covariante, estos serán horizontales si se cumple la hipótesis de riesgos proporcionales ya que en tal caso los residuos son independientes del tiempo.

Para la realización se deben calcular los residuos de Schoenfeld del modelo de Cox en estudio; ordenar los tiempos de falla etiquetando el orden a cada tiempo; calcular la correlación entre los residuos y la variable de orden creada. Y se realiza el test de si $H_0 : \rho = 0$ para cada covariable por separado. En el caso de aceptar la hipótesis nula de que la correlación es cero, se cumplirá la hipótesis de riesgos proporcionales para la covariable correspondiente.

4.3.6. Curvas de supervivencia ajustadas

Cuando se usa un modelo de Cox para ajustar los datos de supervivencia, se pueden obtener curvas de supervivencia que se ajustan a las variables explicativas utilizadas como predictores.

La fórmula de la función de supervivencia detallada seguidamente, es la base para determinar las curvas de supervivencia ajustadas.

$$S(t, X) = [S_0(t)]^{e^{\sum_{i=1}^p \beta_i X_i}}.$$

Esta fórmula expresa que la función de supervivencia en el tiempo t para un individuo con el vector de covariables X viene dada por una función de supervivencia base $S_0(t)$, aumentada a una potencia igual a la exponencial de la suma de β_i veces X_i .

Luego, la expresión para la función de supervivencia estimada es la siguiente:

$$\hat{S}(t, X) = [\hat{S}_0(t)]^{e^{\sum_{i=1}^p \hat{\beta}_i X_i}}.$$

4.3.7. Modelo extendido de Cox

La fórmula del modelo de Cox establece que el riesgo en el tiempo t es el producto de dos cantidades. La primera de ellas, $h_0(t)$, es llamada función de riesgo base. La segunda cantidad es la expresión exponencial e a la suma lineal de $\beta_i X_i$, sobre las p variables explicativas X . Una característica importante de esta fórmula, que

CAPÍTULO 4. METODOLOGÍA

refiere al supuesto de riesgos proporcionales, es que el riesgo base es una función de t que no involucra las X 's mientras que la expresión exponencial involucra las X 's pero no involucra t . Las X 's en ese caso son llamadas *independientes del tiempo* ya que para un sujeto dado, su valor permanece constante en el tiempo. Es posible, no obstante, considerar X 's que difieran con el tiempo, llamadas variables *dependientes del tiempo*.

Dada una situación de análisis de supervivencia que involucra variables predictoras independientes y dependientes del tiempo, se puede escribir el **modelo extendido de Cox** que incorpora ambos tipos de variables, de la siguiente manera:

$$h(t, X(t)) = h_0(t) e^{\sum_{i=1}^{p_1} \beta_i X_i + \sum_{j=1}^{p_2} \delta_j X_j(t)},$$

siendo

$$X(t) = (\underbrace{X_1, X_2, \dots, X_{p_1}}_{\text{independ. del tiempo}}, \underbrace{X_1(t), X_2(t), \dots, X_{p_2}(t)}_{\text{depend. del tiempo}}).$$

La parte exponencial contiene variables independientes del tiempo, indicados por las X_i , y variables dependientes del tiempo, indicados por las variables $X_j(t)$.

Otra opción es crear variables dependientes del tiempo a partir de variables independientes del tiempo, definiendo términos de productos entre cada variable y alguna función del tiempo. Es decir, si las variables independientes del tiempo se denotan como X_i , se puede definir el término del producto i -ésimo como $X_i \times g_i(t)$, donde $g_i(t)$ es alguna función del tiempo para la variable i -ésima, quedando expresado de la siguiente forma:

$$h(t, X(t)) = h_0(t) e^{\sum_{i=1}^p \beta_i X_i + \sum_{i=1}^p \delta_i X_i g_i(t)}.$$

La forma más simple para $g_i(t)$ es que todos los $g_i(t)$ sean idénticamente 0 en cualquier momento del tiempo; esta sería otra manera de establecer el modelo de riesgos proporcionales original, que no contiene términos dependientes del tiempo. Una se-

gunda opción es dejar $g_i(t) = t$. Esto implica que para cada X_i del modelo, hay una variable dependiente del tiempo correspondiente de la forma $X_i \times t$. Si se quiere centrar en una variable particular independiente del tiempo, por ejemplo, la variable X_L , entonces $g_i(t) = t$ para $i = L$, pero es igual a 0 para todos los demás i . También ese podría elegir $g_i(t)$ como el log de t , de modo que las variables dependientes del tiempo correspondientes tendrán la forma $X_i \times \ln t$. Y otra opción más sería dejar que $g_i(t)$ sea una “función de cabecera” de la forma $g_i(t) = 1$ cuando t está por encima de un tiempo específico, t_0 , y $g_i(t) = 0$ cuando t está por debajo de t_0 .

Al igual que con el modelo de riesgos proporcionales de Cox más simple, los coeficientes de regresión en el modelo extendido, se estiman utilizando un procedimiento de máxima verosimilitud en el que se maximiza la función de verosimilitud parcial L .

Los métodos para hacer inferencias estadísticas son esencialmente los mismos que para el modelo de riesgos proporcionales. Es decir, se pueden usar las pruebas de Wald y/o la razón de verosimilitud y los métodos de intervalo de confianza para muestras grandes.

Una suposición importante del modelo extendido de Cox, es que el efecto de una variable dependiente del tiempo $X_j(t)$ sobre la probabilidad de supervivencia en el tiempo t , depende del valor de esta variable en ese tiempo t , y no en el valor en un momento anterior o posterior.

4.4. Estimación de la función de riesgo

Horová et~al. (2012) introducen la noción de estimación de la función de riesgo como se explica a continuación.

Nelson (1972) propuso estimar la función de riesgo acumulado mediante

$$\mathcal{H}_n(x) = \sum_{X_{(i)} \leq x} \frac{\delta_{(i)}}{n - i + 1}, \quad (4.13)$$

CAPÍTULO 4. METODOLOGÍA

donde n es el número de tiempos de fallas diferentes y $\delta_{(i)}$ es el indicador de censura correspondiente al instante t_i .

Luego, la estimación por núcleo de la función de riesgo h es la siguiente convolución del núcleo K con el estimador de la función de riesgo de Nelson \mathcal{H}_n

$$\hat{h}(x, b) = \frac{1}{b} \int K\left(\frac{x-u}{b}\right) d\mathcal{H}_n(u) \quad (4.14)$$

$$= \frac{1}{b} \sum_{i=1}^n K\left(\frac{x-X_{(i)}}{b}\right) \frac{\delta_{(i)}}{n-i+1}. \quad (4.15)$$

Estas estimaciones (las estimaciones por núcleos) dependen de un parámetro de suavizado llamado ancho de banda (b) que controla la suavidad de la estimación y de un núcleo (K) que desempeña el papel de función de peso. En lo que se refiere a la función del núcleo, un parámetro clave es su orden, que se relaciona tanto con el número de sus momentos de fuga como con el número de derivadas existentes para la curva subyacente que debe estimarse.

El núcleo K de orden 2 presenta las siguientes propiedades:

- Está acotado en el intervalo $[-1, 1]$. ($\text{soporte}(K) = [-1, 1]$) .
- Es una función lipschitziana en ese intervalo, es decir que no oscila demasiado. ($K \in Lip[-1, 1]$) .
- Tiene media igual a cero. $\left(\int_{-1}^1 xK(x)dx = 0\right)$.
- Tiene momento segundo finito. $\left(\int_{-1}^1 x^2K(x)dx = \beta_2(K) \neq 0\right)$.
- Integra 1. $\left(\int_{-1}^1 K(x)dx = 1\right)$.

(Muller and Wang, 1994) proponen una modificación del estimador de suavizado por núcleos de Nelson, en el que se permiten grados de suavizado variables en diferentes puntos, así como la implementación de núcleos en los bordes, definido de la siguiente manera:

$$\hat{h}(x, b) \equiv \hat{h}(t, b(x)) = \frac{1}{b(x)} \sum_{i=1}^k K_x\left(\frac{x-X_{(i)}}{b(x)}\right) \frac{\delta_{(i)}}{k-i+1}. \quad (4.16)$$

4.4. Estimación de la función de riesgo

Aquí, el ancho de banda $b = b(x)$ al igual que el núcleo $K = K_x$ dependen del punto x en donde la estimación es calculada.

Las propiedades del estimador por núcleos y de la modificación propuesta por Muller y Wang, se presentan en el Anexo A.1.

4.4.1. Elección de la forma del núcleo K

Para los núcleos pertenecientes a la clase $S_{0,2}$, el que cumple la condición de optimallidad, es el núcleo de Epanechnikov con el cual se trabajará en el presente trabajo (Horová et~al., 2012).

El núcleo de Epanechnikov se define de la siguiente manera:

$$K(x) = \frac{3}{4}(1 - x^2)I_{[-1,1]}(x).$$

Se puede observar en la Figura 4.1 la forma de dicho núcleo.

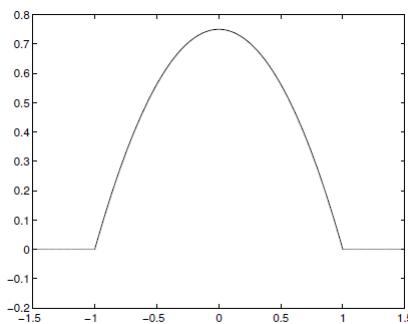


Figura 4.1: Núcleo de Epanechnikov.

4.4.2. Elección del ancho de banda b

El problema de elegir cuánto suavizar, es decir, cómo elegir el ancho de banda es un problema crucial común en el suavizado por núcleos.

Hay dos posibles métodos en el algoritmo a ser utilizado que contemplan las estimaciones antes mencionadas, en la que se considera un único ancho de banda para

CAPÍTULO 4. METODOLOGÍA

todos los puntos y la variación en la que se considera un ancho de banda variable. La primera de ellas es considerada como el **método global**, el ancho de banda óptimo con dicho método se obtiene al minimizar el Error Cuadrático Medio Integrado (IMSE). Para el **método local**, en donde se utilizan diferentes anchos de banda en cada punto, el ancho de banda óptimo en un punto se obtiene minimizando el Error Cuadrático Medio (MSE) local (Hess and Gentleman, 2014).

La elección del ancho de banda inicial depende de la situación específica, pero es muy importante para obtener el mejor ancho de banda. Un posible valor sugerido por Muller y Wang es $b.inicial = \frac{R}{8*m_u^{1/5}}$, suponiendo que los datos están disponibles en $[0, R]$, y m_u es el número de observaciones sin censura.

4.4.3. Efecto borde

Al igual que anteriormente, se supone que los datos están disponibles en $[0, R]$. Cuando la estimación es cerca de 0 o R , pueden ocurrir lo que comúnmente se denomina efectos de borde. Esto puede llevar a problemas de sesgo o estimaciones negativas de la función de riesgo cerca de esos puntos finales de los datos.

Hay varios métodos para corregir los efectos de borde; uno de ellos se basa en la construcción de núcleos de borde especiales tal y como en el estimador de la Ecuación 4.16.

Al núcleo de Epanechnikov, Muller and Wang (1994) lo definen en los bordes como:

$$\frac{12}{(1+q)^4}(x+1)\left[\frac{x(1-2q)+(3q^2-2q+1)}{2}\right].$$

4.4.4. Intervalos de confianza

Los intervalos de confianza para la estimación de la función de riesgo $\hat{h}^{(v)}(\cdot, b_{opt,v,k})$ se pueden construir tal y como se describe en el libro de Horová et~al. (2012).

4.4. Estimación de la función de riesgo

El intervalo de confianza asintótico $(1 - \alpha)$ para $h^{(v)}(x, b)$ viene dado por

$$\hat{h}^{(v)}(x, b) \pm \left\{ \frac{\hat{h}(x, b)V(K)}{(1 - L_n(x))n\hat{b}^{2v+1}} \right\}^{1/2} \Phi^{-1}(1 - \alpha/2),$$

donde Φ es la función de distribución normal estándar.

CAPÍTULO 4. METODOLOGÍA

Capítulo 5

Descripción y análisis exploratorio de los datos

En este capítulo en primer lugar se define el universo de análisis y la fuente de donde se obtuvieron los datos (Sección 5.1).

Luego se presentan algunos resultados del análisis exploratorio de los datos realizado. Se distingue entre primera vida (Sección 5.2.1) y segunda vida (Sección 5.2.2) de las tablets y para analizar cada una de ellas, se presentan cuadros y gráficos con la cantidad de defunciones y censuras, las fecha de inicio y finalización de los episodios y duración de los mismos, las causas de defunción, y las posibles variables explicativas que se puedan incluir a los modelos, tales como variables sociodemográficas de las alumnas y alumnos: sexo, área geográfica y contexto sociocultural, y variables que cuantifican la entrada de las tablets a servicio técnico.

Las visualizaciones fueron realizadas con el paquete `ggplot2` (Wickham, 2016).

CAPÍTULO 5. DESCRIPCIÓN Y ANÁLISIS EXPLORATORIO DE LOS DATOS

5.1. Universo de análisis y fuente de datos

El universo de análisis considerado en este trabajo refiere a toda la generación del modelo de tablets entregado por Ceibal, pertenecientes a alumnas y alumnos del sistema educativo público que a partir de 2014 ingresaron a educación primaria cursando primer, segundo o tercer grado, en escuelas comunes del todo el país.

Los datos utilizados fueron brindados por el Plan Ceibal tanto de los alumnos y las alumnas matriculadas en primaria a los cuales se les entregó ese modelo de tablet en el 2014 o 2015 y sus respectivas escuelas y características (Fuente CRM), cómo de las bases de datos de las Órdenes de Trabajo (OT) asociadas a las reparaciones de las tablets realizadas, entre el 2014 y 2017, por Servicio Técnico (ST) de Plan Ceibal (Fuente K2B), observando en los registros administrativos el tipo de desperfecto que presentó el equipo y en qué momento del tiempo lo hizo.

5.2. Análisis exploratorio de los datos de las tablets

En primer lugar se definen la primera y segunda vida de las tablets, así como una breve descripción del proceso de creación de dichos subconjuntos de datos.

Se considera la **primera vida** de las tablets, desde el momento en que estas son entregadas a las alumnas y los alumnos de primaria (respetando un proceso planificado por Ceibal) hasta que presentan un primer evento, o son cambiadas por otro dispositivo por parte de Ceibal, sin haber tenido antes ninguna reparación que implique su defunción, considerándose por lo tanto censuradas.

Se considera la **segunda vida** de esta generación de tablets, desde el momento en que estas fueron reparadas luego de su primer defunción y vueltas a entregar a los mismos u otros alumnos o alumnas, hasta que presentan el segundo evento o son

5.2. Análisis exploratorio de los datos de las tablets

censuradas al ser cambiadas por otros dispositivos por parte de Ceibal.

Se comienza trabajando con cinco tablas iniciales: *OT*, *Ventas*, *Personas*, *Relación Extendida* y *Empresas*. Las mismas contienen los datos de las reparaciones realizadas por servicio técnico, las características de la entrega de los equipos, datos de las personas que los recibieron y datos de las instituciones a las que concurren dichas personas. Luego se filtran las cinco tablas para que contengan sólo datos de estudiantes de primaria a los que se les entregó, en 2014 o 2015, ese modelo de tablet, eliminando el resto de las observaciones. Finalmente se crean dos bases para cada una de las vidas de la tablet. La primera de ellas cuenta con 80500 observaciones correspondientes a la primera vida, mientras que la segunda con 48652 correspondientes a la segunda vida.

5.2.1. Primera vida de las tablets

En primer lugar, se debe mencionar que fue necesario depurar la base de datos al encontrarse con casos duplicados y números de series de los equipos que no contaban con registros asociados en algunas de las variables, además de irregularidades y errores de registro. Se realiza para ello un trabajo en la curación de los datos atendiendo las siguientes dimensiones: exactitud, validez, completitud y consistencia. Se analizan por ejemplo, mismas personas con diferentes roles (estudiantes y docentes), duplicados en número de serie y fecha en que fueron entregados, valores perdidos (missing data), registros de consumo de repuestos de otros equipos distintos al modelo en estudio, duraciones negativas o menores a un día de vida, entre otras.

Se obtiene entonces una base con la primera vida de las tablets de 80500 números de series, habiéndose entregadas 43435 de estas en 2014 y 37065 en 2015, 1873 escuelas públicas, a 72429 estudiantes de primero, 8024 estudiantes de segundo y 39 estudiantes de tercero (las restantes fueron entregadas a algunos alumnos de cuarto, quinto y sexto).

CAPÍTULO 5. DESCRIPCIÓN Y ANÁLISIS EXPLORATORIO DE LOS DATOS

5.2.1.1. Defunciones y censuras

Se entiende que una tablet de Ceibal es sobreviviente cuando presenta condiciones técnicas suficientes para ser operativa para el alumno o la alumna. Por el contrario, la *muerte* o *defunción* del dispositivo se considera cuando el mismo presenta roturas que no permiten el correcto funcionamiento, limitando el acceso tanto a la nueva tecnología como a la información. Se diferencian las roturas que implican la defunción de la tablet de otros tipos de rotura que no impiden el funcionamiento de la misma aún cuando tenga desperfectos (véase sección 5.2.1.4).

Por otro lado la *censura* del dispositivo en este estudio significa que si bien las tablets pueden seguir funcionando (no ocurrió el evento aún), la ventana del tiempo de observación se cierra, ya que el dispositivo fue retirado y cambiado por otro por parte de Ceibal, sin haberse observado el evento hasta ese momento. El tiempo de vida real del dispositivo censurado es mayor que el que se considerará, pero desconocido.

Considerando los equipos entregados en 2014 y 2015 como dos generaciones distintas, la mortalidad fue del 64.7% (28113 tablets) para la generación del 2014 y del 67.2% (24922 tablets) para la del 2015.

Las defunciones contemplan dos tipos de registros considerados: el evento registrado por servicio técnico en sus órdenes de trabajo (OT) y el registro por lo que se considera una sustitución, esto es, registros de equipos que fueron entregados para reemplazar otros y no por el sistema normal de entregas¹.

En la Figura 5.1 se presenta un gráfico de mosaico. En (Hofmann, 2000) se presentan los gráficos de mosaicos y se describen los mismos en base a su semejanza

¹En la programación se considera que el equipo murió por sustitución si, ordenados previamente los registros por *Persona*, *FechaVenta* y *Serie*, para dos observaciones de la misma persona, la fecha de cambio de estado de su primer registro es igual a la fecha de venta de su segundo registro y además la variable *Sust* es igual a *Sustitucion* en el último de ellos.

5.2. Análisis exploratorio de los datos de las tablets

gráfica con las tablas de contingencia multivariadas; cada celda de una tabla de correspondencia se visualiza mediante un mosaico, y el tamaño del mismo es directamente proporcional al número de casos en esa celda. Mientras que por ejemplo en un gráfico de barras se comparan valores absolutos, en un mosaico se muestran proporciones relacionadas con los tamaños de muestra de cada subconjunto. La altura de los rectángulos corresponde a la probabilidad marginal de las categorías de la variable y el ancho de los rectángulos corresponde a la probabilidad condicional dentro de los niveles indicados para la otra variable.

En la Figura 5.1 se muestra la proporción entre defunciones y censuras variando con el año de entrega del equipo. Se observa que no hay prácticamente diferencias entre un año y otro, pero condicional al año de entrega, en 2015 hay más defunciones.

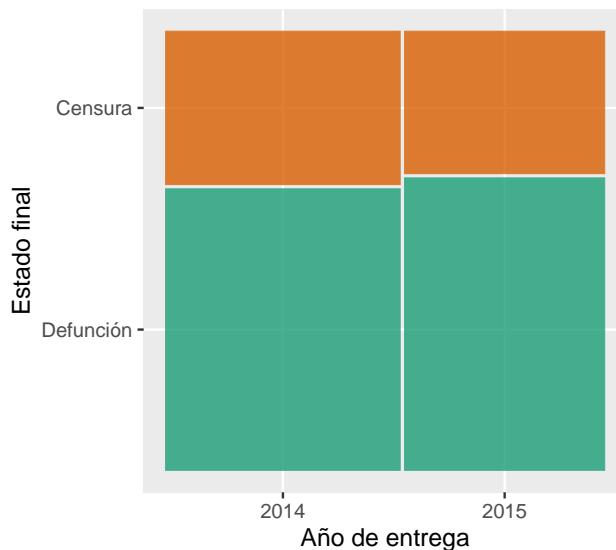


Figura 5.1: Gráfico de mosaico que muestra la proporción de tablets que presentaron defunciones o censuras (estado final) en la primera vida, por separado para la generación de las 43435 tablets entregadas en 2014 y para la generación de las 37065 entregadas en 2015.

CAPÍTULO 5. DESCRIPCIÓN Y ANÁLISIS EXPLORATORIO DE LOS DATOS

5.2.1.2. Fechas de inicio y fechas de finalización

El proceso de entrega de las tablets varía según la planificación de Ceibal dentro de cada año, 2014 y 2015. Algunas pueden haber sido entregadas por ejemplo al comienzo del año, pero otras cuando este ha avanzado más. Lo mismo sucede con el proceso de recambio en 2017. Debido a esto, el período de riesgo es diferente para cada uno de los equipos.

La fecha de inicio de la primera vida de las tablets corresponde a la fecha en que estas fueron entregadas por primera vez a las alumnas y los alumnos, mediante el proceso de entregas de Plan Ceibal.

Se muestra en la Figura 5.2 las fechas de inicio de los episodios, representando el proceso de entregas antes mencionado. Se diferencian claramente dos períodos de entrega, entre julio y noviembre en 2014 y entre fines de mayo y mediados de julio en 2015, correspondientes a las dos generaciones de estudio. El día que se entregaron mas tablets (4154 equipos) fue el 25 de junio del 2015.

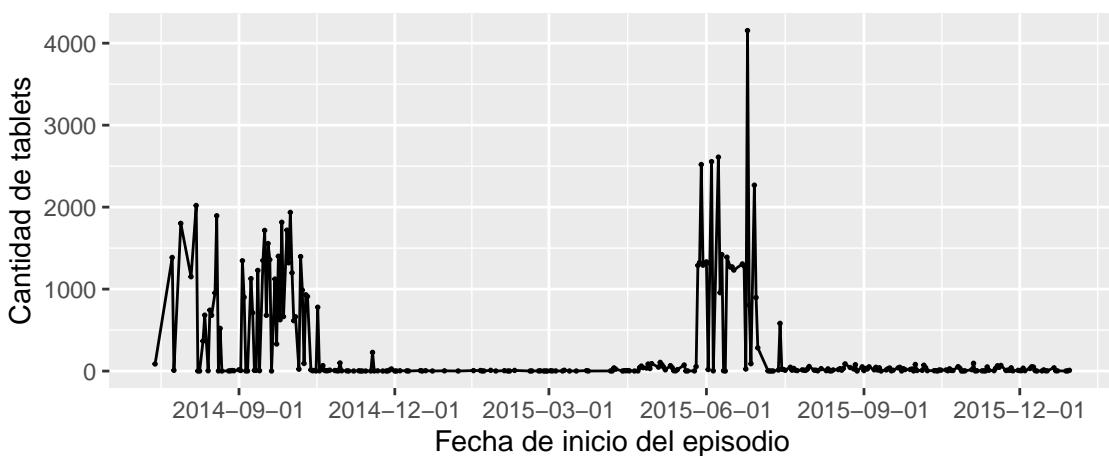


Figura 5.2: Serie temporal con las fechas de inicio de los episodios, es decir las fechas de entrega de las 80500 tablets que dan inicio a la primera vida de las mismas. En el gráfico se diferencian claramente dos períodos de entrega, entre julio y noviembre de 2014 y mayo y julio de 2015.

La fecha de finalización del episodio, para todas las tablets que presentan una de-

5.2. Análisis exploratorio de los datos de las tablets

función, corresponde a la fecha registrada en la OT en la que se comienza a realizar una reparación al equipo que ya no funciona. En el caso de las defunciones por sustitución en particular, la fecha de finalización corresponde a la fecha en que se le entregó a la alumna o el alumno un equipo que sustituye el suyo el cual es retirado para ser reparado por lote.

Considerando que la rotura de la tablet se evidencia cuando el alumno o la alumna lleva su tablet a reparación, se introduce el supuesto de que el intervalo de tiempo entre que se produce efectivamente la rotura y el momento en que se registra en la OT no debiera ser sustancial, ya que Plan Ceibal ha dispuesto varios medios por los cuales los alumnos y las alumnas pueden llevar a reparar sus equipos; el servicio de reparación móvil visita los centros educativos frecuentemente, por otro lado hay centros de reparación fija en distintas ciudades del país a donde el alumnado puede llevar sus equipos o incluso enviarlos a través del Correo Uruguayo para que sean reparados, no generando ninguno de estos medios de reparación un costo económico para el alumno o la alumna.

En cuanto a las tablets que se consideran censuradas, la fecha de finalización de su vida es cuando se le entrega al alumno o la alumna un nuevo modelo de tablet o laptop retirándoles definitivamente el que ya tenían, cerrando así la ventana de observación de las mismas. Corresponde a la fecha en que en los registros de Ceibal, el equipo pasa de estar *entregado*, a estar *devuelto* o *pendiente de devolución*.

En la Figura 5.3 se muestran las frecuencias de las fechas de muerte o censura, es decir las fechas de finalización del episodio, a través de un histograma. Para éste y los próximos histogramas que se presentan en el trabajo, se utiliza el método de selección de ancho de banda de Freedman-Diaconis. (Ver Anexo A.2). Se observa que la mayor cantidad de tablets que tienen sus fechas de finalización el mismo día se concentra en torno a junio de 2016.

CAPÍTULO 5. DESCRIPCIÓN Y ANÁLISIS EXPLORATORIO DE LOS DATOS

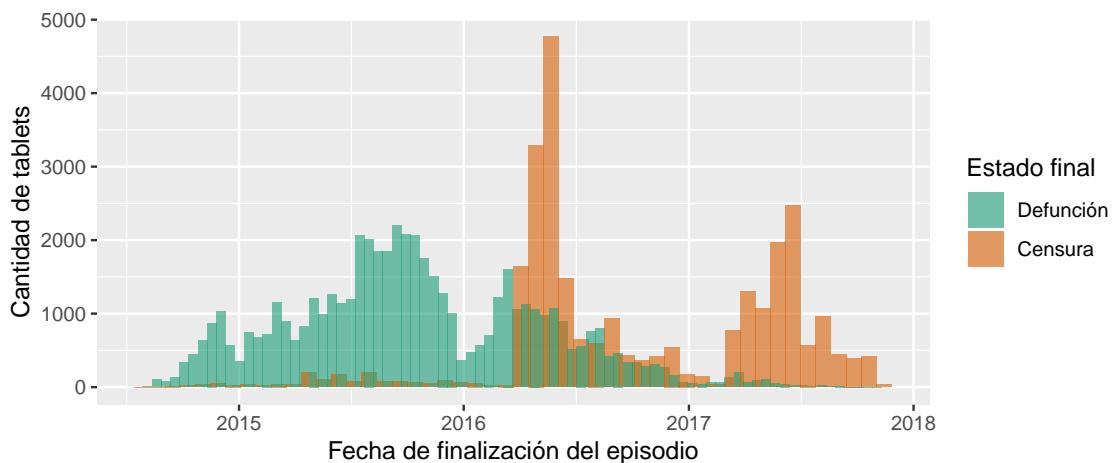


Figura 5.3: Histograma con las fechas de finalización de la primera vida de las 80500 tablets del análisis, con un ancho de banda de Freedman-Diaconis. En color verde se representa la fecha final de las 53035 tablets que presentaron una defunción, y en color naranja la de las restantes 27465 que fueron censuradas.

5.2.1.3. Tiempo de vida hasta la primera defunción o censura

Se define la variable *duración* como el tiempo de vida del equipo, medido como la cantidad de días entre la fecha de entrega del mismo al alumno o la alumna y la fecha registrada como fecha de finalización ya sea por una defunción o no.

Se presenta en la Figura 5.4 la distribución de la duración de todas las tablets entregadas en 2014 y 2015. El gráfico superior presenta la duración de las 43435 tablets entregadas en 2014 que van desde 1 día hasta 1181 días. La mitad de estas tablets tienen una duración de 378 días, en promedio duran 425 días y el 28.9 % tienen una duración entre 500 y 700 días, correspondiendo estas al segundo modo que se observa en dicho gráfico que alcanza los valores más altos. El gráfico inferior presenta las duraciones de las 37065 tablets entregadas en 2015 que van desde 1 día hasta 921 días. La mitad de estas tablets tienen una duración aproximada de 10 meses (320 días) y en promedio duran 356 días. Comparando ambos histogramas, en los dos se distinguen tres modos bien diferentes, pero el correspondiente a la generación del 2014 es más asimétrica a la derecha observándose menor cantidad de

5.2. Análisis exploratorio de los datos de las tablets

duraciones mayores.

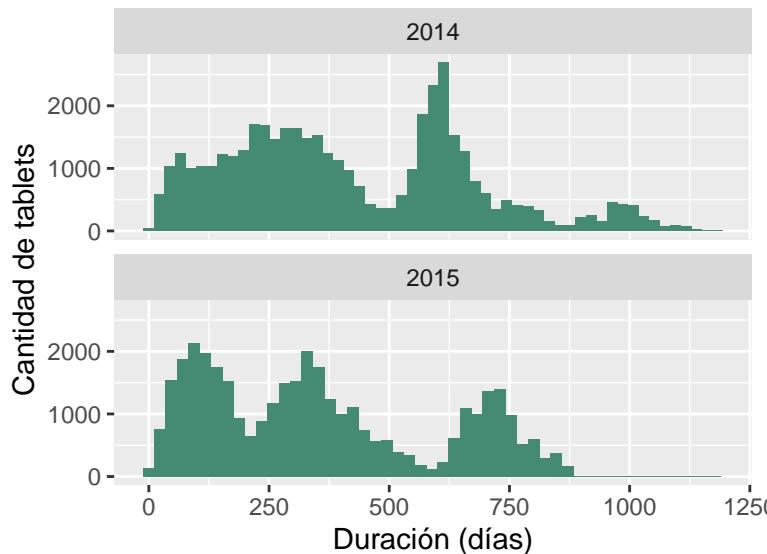


Figura 5.4: Panel superior: histograma con los tiempos de vida de las 43435 tablets entregadas en 2014 agrupadas en intervalos de 22 días según el método de Freedman-Diaconis. Panel inferior: histograma con los tiempos de vida de las 37065 tablets entregadas en 2015 agrupadas en intervalos de 24 días según el método de Freedman-Diaconis.

En la figura 5.5 se refleja también la distribución de la variable duración del total de las tablets pero según si la tablet fue censurada o presentó una defunción, para cada año de entrega de los equipos.

Cabe aclarar para ambos gráficos que, en 2017, a todos y todas las estudiantes se le recambian las tablets por nuevos equipos, por lo que el período de exposición es menor para las que se entregaron en 2015 que para las que se entregaron en 2014, justificando así la mayor duración antes mencionada que se aprecia para los de este último año, principalmente en el panel correspondiente a las censuras.

Se observa que las duraciones tienen comportamientos diferentes según el año de entrega y según si el evento termina con una defunción o con una censura. En cuanto a las censuras, podría deberse al proceso de recambio que determina Ceibal, que puede cambiar de un año a otro ya que es difícil recambiar todas las tablets del

CAPÍTULO 5. DESCRIPCIÓN Y ANÁLISIS EXPLORATORIO DE LOS DATOS

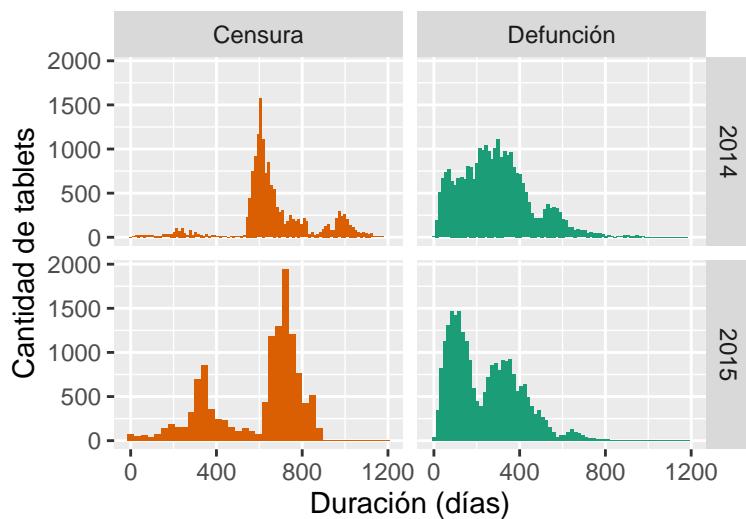


Figura 5.5: Histograma con los tiempos de vida de las tablets según año y estado final. En color verde se representa el tiempo de vida de las 53035 tablets que presentaron una defunción, y en color naranja el tiempo de vida de las 27465 tablets que fueron censuradas. A su vez se diferencia entre la generación del 2014 (paneles superiores) y la del 2015 (paneles inferiores). Los correspondientes anchos de banda de Freedman-Diaconis son igual a 13 y 14 para las censuras y defunciones del 2014 respectivamente, e igual a 31 y 17 para las censuras y defunciones del 2015.

5.2. Análisis exploratorio de los datos de las tablets

país en un mismo día, entonces se hace durante un proceso de varios meses como se ha mencionado.

Si se observan los dos gráficos conjuntamente se puede concluir que para el caso de la generación de 2014, el primer modo que se observaba en la Figura 5.4 corresponde a defunciones y los otros dos a censuras; mientras que para la generación del 2015 el primer modo corresponde a defunciones, el segundo es una mezcla de defunciones y censuras y el tercero representa prácticamente solo censuras. Hay una buena concordancia entre un gráfico y otro.

5.2.1.4. Causas de defunción

Según la información brindada por Plan Ceibal hay en principio cuatro repuestos que implicarían la defunción del equipo, estos son: placa madre, pantalla, touchscreen y módulo completo. Con este último se está considerando la pantalla y el touchscreen como una pieza única. Por otro lado, la carcasa, la cámara, los parlantes, la antena, los botones, el micrófono, el cargador y los cables USB-micro USB, son repuestos que no impiden que el equipo siga funcionando por lo que no se consideran como causas de defunción.

En los registros respectivos a las OT, se encuentra además del consumo de estos repuestos, otros pertenecientes a otros tipos o modelos de dispositivos distintos a la tablet en estudio. Se suponen como errores de tipeo de los técnicos y se consideran también como repuestos que impliquen la defunción del equipo, agrupándolos en las cuatro causas iniciales².

Es necesario definir que cuando los equipos presentan más de una causa de defunción en el mismo registro (día), también se agrupan considerándose como una sola categoría, tal como se muestra en las primeras dos columnas de la Tabla 5.1 en la que se exponen estas causas agrupadas y las cantidades que hubo de cada una de

²Se programan como un factor con cuatro niveles, donde la placa madre tiene el valor 1, la pantalla el 10, el touchscreen el 100 y el 1000 corresponde al módulo completo

CAPÍTULO 5. DESCRIPCIÓN Y ANÁLISIS EXPLORATORIO DE LOS DATOS

ellas.

A su vez se vuelven a agrupar estas categorías computando la causa según un orden de prelación. El orden es el siguiente: 1.Placa Madre; 2.Módulo completo; 3.Otros repuestos; a excepción de cuando se utiliza para reparar la tablet, el módulo en conjunto con la placa, considerando esta categoría como otra causa de defunción diferente. Además los repuestos que implicaron reparar la pantalla y el touchscreen se computan a la causa módulo completo ya que son pocas cantidades. En la columna *Causas agrupadas* de la Tabla 5.1 se refleja la categorización anteriormente explicada.

Tabla 5.1: Categorías de la variable que contiene las causas de defunción.

Cantidad	Causa	Causa agrupada
34135	Placa Madre	Placa
9134	Módulo Completo	Módulo
6420	Módulo Completo + Placa Madre	Módulo + Placa
51	Placa Madre + Placa Madre	Placa
39	Pantalla	Módulo
19	Módulo Completo + Módulo Completo	Módulo
5	Touchscreen + Placa Madre	Placa
5	Módulo Completo + Pantalla	Módulo
5	Módulo Completo + Módulo Completo + Placa Madre + Placa Madre	Módulo + Placa
4	Pantalla + Placa Madre	Placa
2	Touchscreen	Módulo
2	Módulo Completo + Placa Madre + Placa Madre	Módulo + Placa

Las sustituciones de equipos se categorizan en una nueva causa de defunción nombrada *sustitución*. Dado que no se cuenta con la información de una reparación individual que indique una causa específica, pero si se sabe que esos equipos fueron analizados y reparados por lotes que implicaban consumos principalmente del connector (placa madre), se asume por parte del área de datos de Ceibal, que dichas tablets estaban muertas al momento de la sustitución. Hubieron 3219 sustituciones realizadas considerando la generación del 2014 y la del 2015 conjuntamente.

Al analizar los registros de repuestos utilizados y sustituciones realizadas en la primera vida de las tablets, se destaca que la principal causa de defunción es la rotura

5.2. Análisis exploratorio de los datos de las tablets

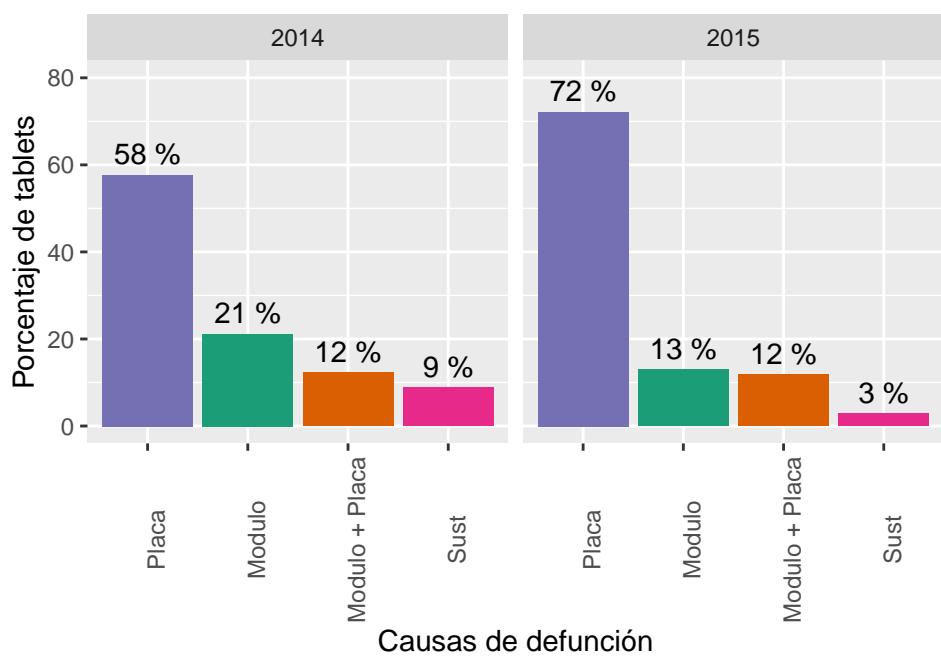


Figura 5.6: Gráfico de barras con el porcentaje de tablets que presentan en su primera vida sustituciones, reparaciones de placa, reparaciones del módulo completo (pantalla + touchscreen) y reparaciones de la placa y el módulo conjuntamente. Se distingue en los paneles entre la generación del 2014 (43435 tablets) y la del 2015 (37065 tablets).

CAPÍTULO 5. DESCRIPCIÓN Y ANÁLISIS EXPLORATORIO DE LOS DATOS

de la placa madre, siendo la que explica el 57.7 % de las defunciones para las entregadas en 2014 y el 72.1 % para las entregadas en 2015, lo cual se muestra en la Figura 5.6. Si se consideraran las tablets a las que se les reparó la placa conjuntamente con el módulo, ese porcentaje sería aún mayor.

En la Figura 5.7 se presentan la duración media y mediana por cada causa de defunción. Sin contar las sustituciones, las que duraron menos en promedio fueron aquellas tablets a las que se les reparó el módulo completo, y las que tuvieron una duración media y una duración mediana mayores, fueron las tablets cuya reparación fue del módulo completo conjuntamente con la placa madre. Las tablets cuya causa de defunción fue la placa madre tienen una mayor duración media respecto a las tablets cuya causa de defunción fue el módulo, pero una menor duración mediana.

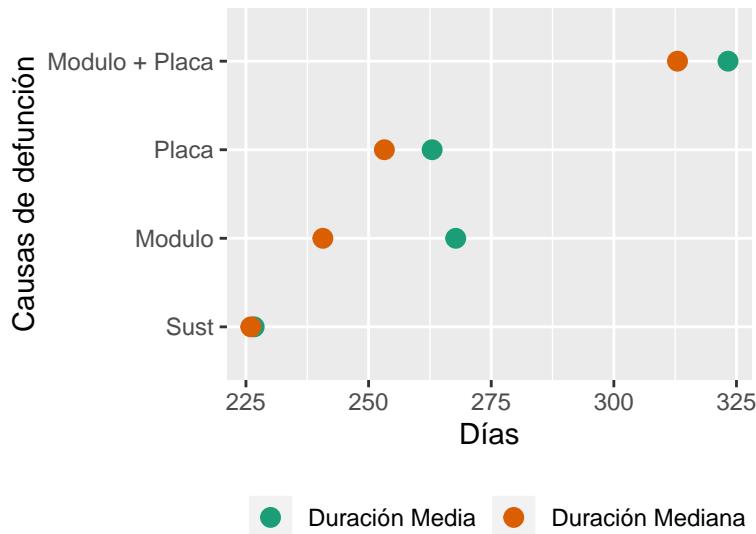


Figura 5.7: Gráfico de puntos que muestra la duración media (color naranja) y la duración mediana (color verde) en días, de las 53035 tablets que presentaron una defunción. Se muestra por separado para las 3219 tablets que presentaron una sustitución, las 34195 a las cuales se les reparó la placa, las 9194 a las cuales se les reparó el módulo y las 6427 que se les reparó el módulo y la placa conjuntamente.

5.2. Análisis exploratorio de los datos de las tablets

5.2.1.5. Variables sociodemográficas a considerar: sexo, área geográfica y contexto sociocultural

Una de las variables que se considera que podría influir en la tasa de rotura de las tablets es el sexo del alumnado, otra es el contexto sociodemográfico de las escuelas a las que ellos concurren. Dicho indicador es determinado por (Anep) y se construye dividiendo el total de escuelas públicas en cinco grupos, de modo que el quintil 1 agrupa al 20 % de las escuelas con mayor nivel de criticidad (contexto más vulnerable) y así sucesivamente hasta el quintil 5 que está conformado por el 20 % de las escuelas con indicadores socioculturales más favorables. Esta clasificación se hace por separado para el conjunto de escuelas urbanas por un lado, y para las escuelas rurales por otro. En este trabajo, los quintiles rurales se agrupan con los urbanos ya que son de dimensiones (cantidad de escuelas en valores absolutos en cada uno de los quintiles) mucho menores a los urbanos. Otra variable que se considera en el análisis es la ubicación geográfica del alumno o alumna identificando si la escuela a la que asiste está ubicada en la capital del país, en el interior urbano o en el ámbito rural.

En la Figura 5.8 se presenta la distribución de estas tres variables sociodemográficas que dan cuenta de la composición del alumnado a quienes se les entregaron las tablets en 2014 y 2015.

En las tres secciones siguientes se analiza como pueden influenciar estas variables sociodemográficas en la proporción de defunciones, en la duración de la primera vida y en las causas de defunción de la misma, tratando de evidenciar si hay o no diferencias entre un perfil y otro.

5.2.1.6. Defunciones y censuras según las variables sociodemográficas

En la Figura 5.9 se compara la distribución de las alumnas y los alumnos que registraron fallas en sus equipos con aquellos que no, según el área geográfica, contexto

CAPÍTULO 5. DESCRIPCIÓN Y ANÁLISIS EXPLORATORIO DE LOS DATOS

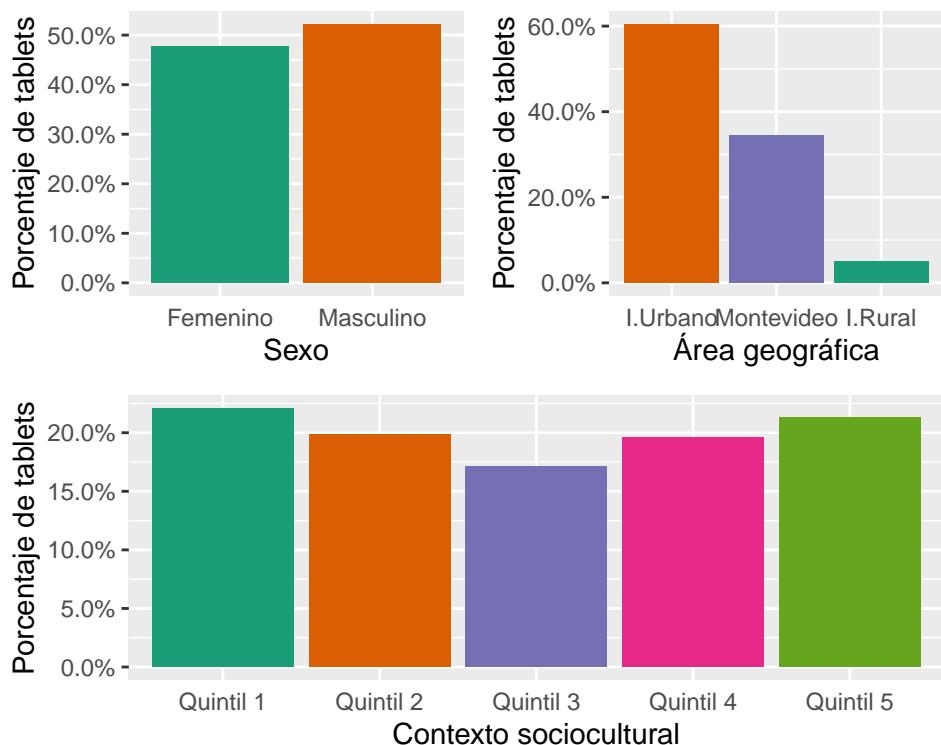


Figura 5.8: Gráficos de barras con los porcentajes de cada categoría de cada una de las variables sociodemográficas consideradas: sexo del alumno o alumna, área geográfica y contexto sociocultural al que pertenece la escuela a la que concurre el alumno o la alumna a quien se le entregó una tablet del modelo en estudio en 2014 y 2015.

5.2. Análisis exploratorio de los datos de las tablets

sociocultural y sexo del mismo/a, sin distinguir una generación de otra ya que la distribución dentro de cada año para cada una de las variables es muy similar para sus respectivas categorías.

Presentan mayor cantidad de defunciones aquellas tablets pertenecientes a alumnos de sexo masculino, de escuelas de contexto sociocultural medio y usuarios que concurren a centros educativos del interior, en particular interior urbano, sin embargo se observa que las diferencias respecto a la cantidad de defunciones entre una categoría y otra de cada una de estas tres variables, no son sustanciales.

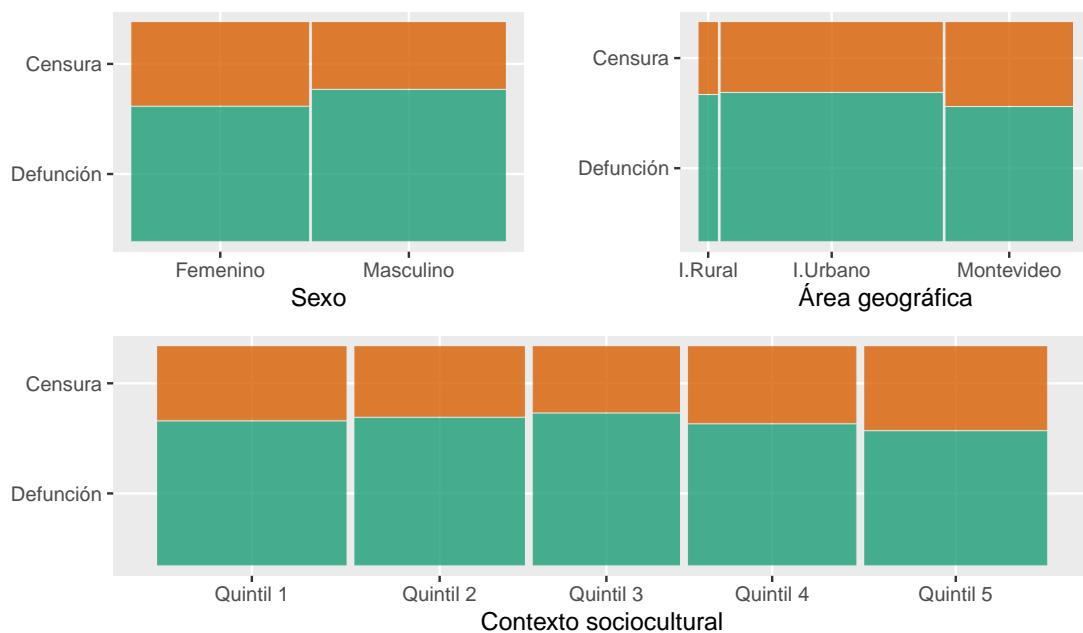


Figura 5.9: Gráficos de mosaicos con la proporción de defunciones o censuras (estado final) de la primera vida de las 80500 tablets, para las diferentes categorías de la variable sexo (panel superior izquierdo), área geográfica (panel superior derecho) y contexto sociocultural (panel inferior).

Al comparar la mortalidad del total de tablets entregadas según el sexo del alumno o la alumna, mueren el 69.6 % de las entregadas a los niños y el 61.8 % de las entregadas a las niñas.

Al comparar la mortalidad según la región geográfica a la que pertenece la escuela,

CAPÍTULO 5. DESCRIPCIÓN Y ANÁLISIS EXPLORATORIO DE LOS DATOS

las tablets de escuelas del Interior Urbano son las que presentaron mayor porcentaje de defunciones (68.2 %) en relación al total de tablets entregadas a dicha área geográfica.

Al comparar la mortalidad según el contexto sociocultural, la cantidad de defunciones ronda entre el 60 % y el 70 % para todos los contextos, siendo aquellas tablets de escuelas pertenecientes al contexto clasificado como *quintil 3*, con un nivel medio de criticidad, las que presentan mayor cantidad de defunciones (69.8 %) y las del contexto menos crítico clasificado como *quintil 5*, las que presentaron menor cantidad (61.6 %) respecto al total entregado.

5.2.1.7. Tiempo de vida hasta la primera defunción según las variables sociodemográficas

Se analiza la duración de las tablets que presentaron una primera defunción según las tres variables sociodemográficas consideradas, a través de gráficos de los cuales se presentan en el Anexo B.1. No se aprecian prácticamente diferencias entre las categorías de cada variable.

5.2.1.8. Causas de defunción según variables sociodemográficas

Para ver si hay algún perfil sociodemográfico del alumnado que determine alguna causa de defunción en particular, se grafican en la Figura 5.10 tres mosaicos en los que se muestra la proporción de tablets que presentan cada una de las causas de defunción definidas, según el sexo, el área geográfica y el contexto sociocultural de las alumnas y alumnos. Al igual que en el análisis conjunto (sin distinción por variables), para todas las categorías, la mayor proporción de defunciones la presentan aquellas tablets a las que se les repara la placa madre. Analizando el contexto, dicha proporción aumenta en forma gradual desde el quintil 1 al quintil 5. En la variable sexo, para ambas categorías, todas las causas de defunción tienen la misma

5.2. Análisis exploratorio de los datos de las tablets

proporción.

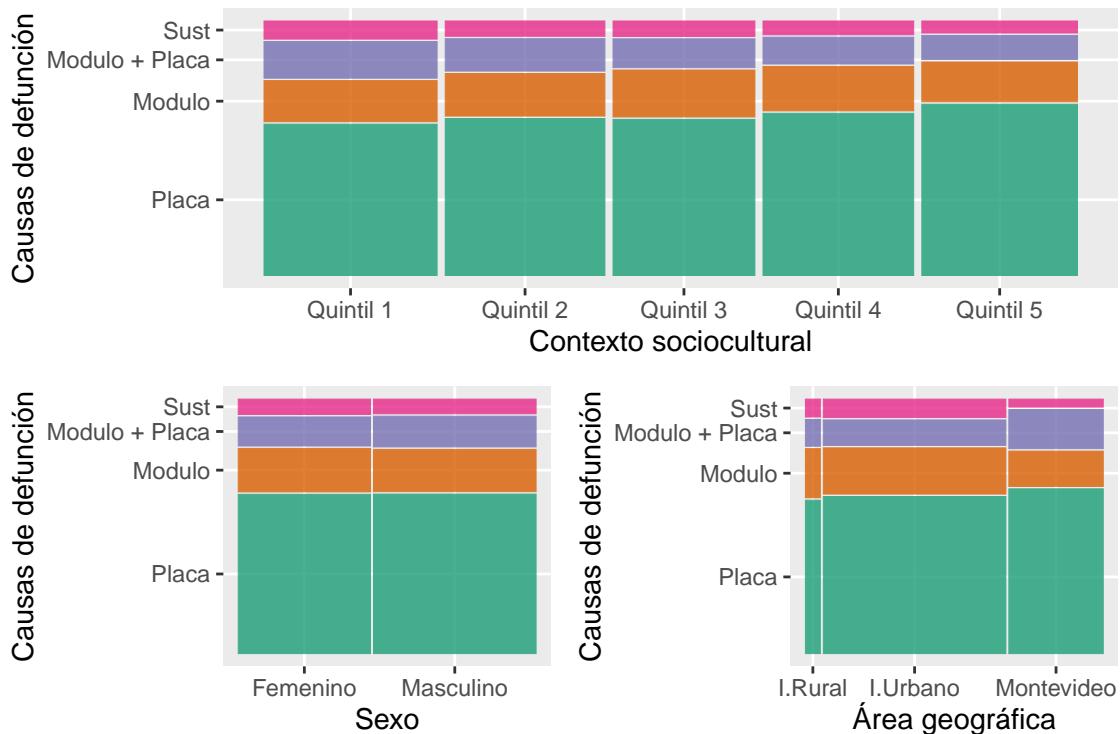


Figura 5.10: Gráficos de mosaico con la proporción las primeras causas de defunción de las 53035 tablets que registraron un evento en su primera vida, según las variables sociodemográficas consideradas en el estudio: sexo (panel inferior izquierdo), área geográfica (panel inferior derecho) y contexto sociocultural (panel superior).

5.2.1.9. Otras variables a considerar: Visitas a servicio técnico y cantidad de repuestos consumidos

Además de considerar las covariables sociodemográficas de la sección anterior, se ha considerado como posible factor que influye la duración del equipo, la cantidad de entradas a servicio técnico, reflejadas en un principio, por un lado en la cantidad de visitas a ST y por otro en la cantidad de fallas registrados durante su primera vida. La primera de ellas cuantifica la cantidad de veces (específicamente días) que el alumno o la alumna llevó a reparar su equipo con el ST de Ceibal. La segunda, cuantifica la cantidad de repuestos que la tablet consumió. La diferencia con la

CAPÍTULO 5. DESCRIPCIÓN Y ANÁLISIS EXPLORATORIO DE LOS DATOS

anterior es que en esta se cuenta por repuestos en vez de por fechas, ya que en un mismo día, en una misma OT, se pueden haber utilizado más de un repuesto para el mismo equipo.

Se analiza la distribución de dichas variables que cuantifican las entradas a servicio técnico y se destaca que el 36.7% (29527) del total de las tablets no entran nunca a ST con una orden de trabajo del tipo reparación, es decir, que luego de que los equipos fueron entregados a los alumnos o las alumnas, no se volvió a tener ningún registro de ellos hasta que fueron recambiados. Considerando sólo las entradas de los equipos que presentaron una defunción, el 97.1% de las tablets no tienen ninguna visita a ST, y por ende, repuestos consumidos, antes de la defunción; es decir, las primeras causas de rotura de estas tablets son causas que les impiden seguir con su funcionamiento. Las demás presentan hasta 3 visitas a ST y hasta 4 reparaciones antes de su defunción.

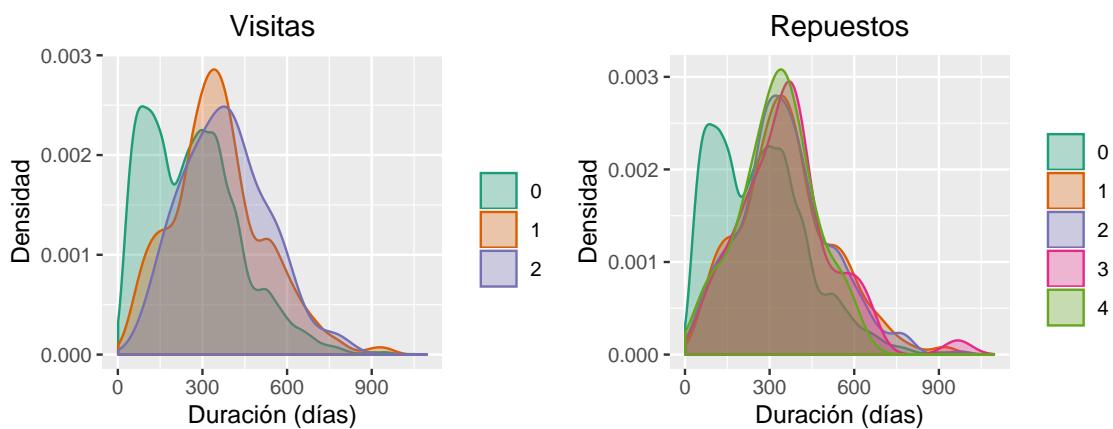


Figura 5.11: Gráficos de densidad con el tiempo de vida de las 53035 tablets que registraron un evento en su primera vida, según la cantidad de visitas a servicio técnico antes de la defunción (panel izquierdo) y la cantidad de repuestos consumidos previo a la defunción (panel derecho).

Por último, en la Figura 5.11 se muestra a través de gráficos de densidad la duración según la cantidad de visitas y según la cantidad de repuestos consumidos, ambas antes de la defunción. Se observa la diferencia de los que van 0 veces a ST previo a

5.2. Análisis exploratorio de los datos de las tablets

su defunción, que tienen una duración menor a la de los demás equipos. Sin embargo no se evidencian prácticamente diferencias en la distribución de la duración si van 1 o 2 veces a ST previo a morir.

Tras este análisis, se decide crear para efectivamente utilizar en los modelos de estimación, una variable dicotómica (llamada *ST*) que vale 1 si el equipo entra a reparación antes de su defunción, por otro motivo que no implique la muerte, y 0 en otro caso.

5.2.2. Segunda vida de las tablets

En esta sección se presentan los resultados del mismo análisis que se ha venido realizando para la primera vida pero para la segunda, de una forma resumida, comenzando con las dimensiones de la base y la proporción de defunciones, luego el tiempo de vida de los equipos y por último las causas de defunción.

De las 53035 tablets que murieron en su primera vida se consideran 48652 que presentan una segunda vida. En el proceso de construcción de la base de la segunda vida de los equipos se dejan fuera 2013 observaciones correspondientes a tablets que fueron retiradas, reparadas por lote (implicando una muerte por lo que se consideró sustitución), pero no se volvieron a entregar a ningún otro alumno o alumna, pasando de 53035 observaciones a 51022. La diferencia con las restantes se debe a que se eliminaron observaciones con duraciones negativas y duraciones menores a un día.

De las 48652 tablets, 25818 pertenecen a la generación del 2014 y 22834 a la del 2015.

5.2.2.1. Defunciones y censuras

De las 48652 tablets que presentan una segunda vida, el 58.8% (28606 tablets) registraron una segunda defunción mientras que para las 41.2% restantes (20046

CAPÍTULO 5. DESCRIPCIÓN Y ANÁLISIS EXPLORATORIO DE LOS DATOS

tablets) no se observó el evento antes de que fueran retiradas por Ceibal pasando a estar fuera del análisis, por lo que se consideran censuradas a la derecha.

Considerando las distintas generaciones, la mortalidad es del 57.5 % (14852 tablets) para las tablets entregadas en 2014 y del 60.2 % (13754 tablets) para las entregadas en 2015, como se observa en la Figura 5.12, donde se grafica un mosaico que muestra la proporción de equipos que presentaron defunciones o censuras en su segunda vida, según el año de entrega de los mismos.

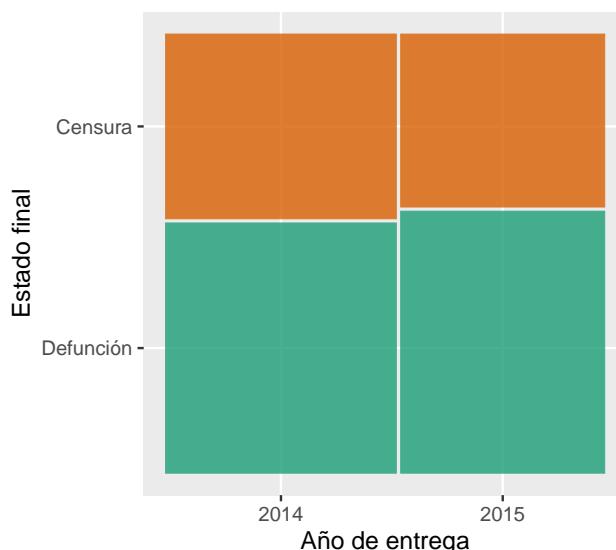


Figura 5.12: Gráfico de mosaico que muestra la proporción de tablets que presentaron defunciones o censuras (estado final) en la segunda vida según el año de entrega, para la generación de las 25818 tablets entregadas en 2014 y la generación de las 22834 entregadas en 2015.

5.2.2.2. Tiempo de vida de las tablets hasta su segunda defunción o censura

Igual que como se había definido anteriormente, el tiempo de vida o duración corresponde a la diferencia entre la fecha de inicio y la fecha de finalización del episodio, ya sea por una defunción o por una censura.

5.2. Análisis exploratorio de los datos de las tablets

Para los equipos que tuvieron una primer defunción por sustitución, la fecha de inicio de la segunda vida es la fecha en que la tablet es entregada por segunda vez a otro alumno u otra alumna. Para el resto de los casos, la fecha en que comienza la segunda vida corresponde a la fecha en que se finalizó la OT correspondiente a la reparación del repuesto que implicó la primer defunción de ese equipo.

La fecha de finalización de los episodios corresponde a la fecha en que el equipo ingresó por segunda vez a servicio técnico o fue retirado para ser reparado por lote, para las defunciones por códigos de OT y defunciones por sustitución respectivamente. Para el caso de las censuras, la fecha de finalización de la segunda vida corresponde a la fecha en que el equipo pasa a estar registrado como devuelto o pendiente de devolución.

Se presenta en la Figura 5.13 a través de un histograma, la distribución de la duración de todas las tablets que presentaron una segunda vida. El panel superior presenta la duración de las 25818 tablets entregadas en 2014, cuyos tiempos de vida van desde 1 día hasta 1046 días. El panel inferior presenta las duraciones de las 22834 tablets entregadas en 2015 que van desde 1 día hasta 844 días aproximadamente. Para ambos años de entrega, se observa como disminuye la cantidad de equipos a medida que se consideran duraciones mayores, siendo más notorio para las entregadas en 2015.

5.2.2.3. Causas de defunción

Al igual que en la primera vida, el equipo puede presentar una defunción por diferentes motivos (o causas); porque se le haya roto la placa, el módulo (incluyendo la pantalla y el touchscreen) o la placa y el módulo a la vez, o porque haya sido sustituido por otro equipo para ser reparado por lote. En la Figura 5.14 se muestra la distribución de estas causas de defunción expresada en términos de porcentajes. Considerando todas las tablets que tienen una segunda vida y vuelven a presentar una defunción, se destaca que la principal causa es nuevamente la rotura de la placa

CAPÍTULO 5. DESCRIPCIÓN Y ANÁLISIS EXPLORATORIO DE LOS DATOS

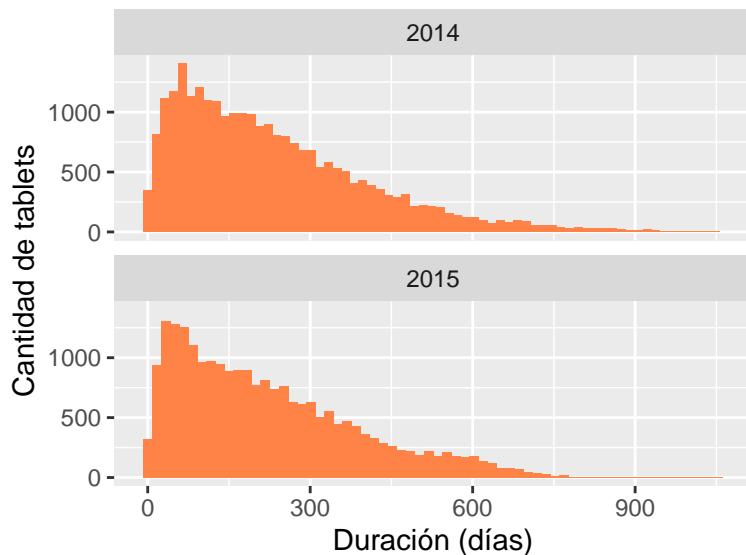


Figura 5.13: Panel superior: Histograma con los tiempos de vida de las 25818 tablets que fueron entregadas en 2014 y tienen una segunda vida, agrupados en intervalos de 16 días según el método de Freedman-Diaconis. Panel inferior: Histograma con los tiempos de vida de las 22834 tablets que fueron entregadas en 2015 y tienen una segunda vida, agrupados en intervalos de 17 días.

5.2. Análisis exploratorio de los datos de las tablets

madre, siendo la que explica el 62.1 % de las defunciones para las entregadas en 2014 y el 70.3 % para las entregadas en 2015 tal como se muestra en la figura.

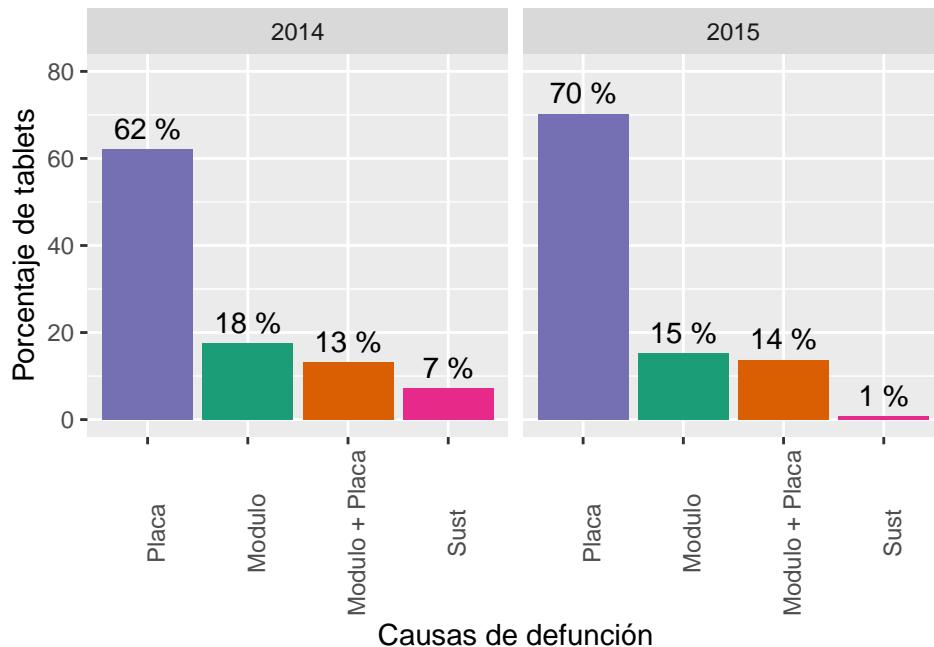


Figura 5.14: Gráfico de barras del porcentaje de tablets que presentan en su segunda vida sustituciones, reparaciones de placa, reparaciones del módulo y reparaciones de placa y módulo conjuntamente según año de entrega del equipo, para la generación del 2014 (25818 tablets) y la del 2015 (22834 tablets).

5.2.3. Comparación entre la primera y segunda vida de las tablets

En la presente sección se realiza una comparación entre la primera y segunda vida de las tablets, para observar si hay o no diferencias entre sus causas de defunción y entre sus tiempos de vida, analizando algunas medidas de resumen e indicadores.

En primer lugar se muestra en la Figura 5.15 un mosaico en el que se presenta la proporción de defunciones y censuras para la primera y segunda vida. Como se ha mencionado, los mosaicos tienen en cuenta la cantidad de observaciones en cada subconjunto a comparar, por eso el azulejo correspondiente a la primera vida es más

CAPÍTULO 5. DESCRIPCIÓN Y ANÁLISIS EXPLORATORIO DE LOS DATOS

ancho. Se observa además que la proporción de defunciones es mayor en la primera vida que en la segunda.

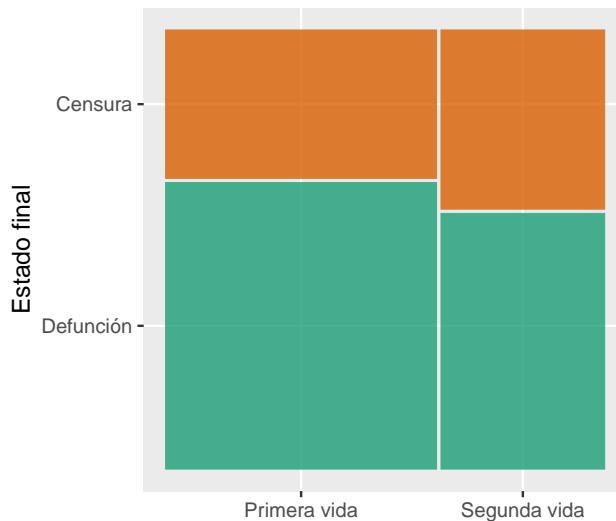


Figura 5.15: Gráfico de mosaico que muestra la proporción de tablets que presentan defunciones o son censuradas (estado final) para las 80500 tablets de la primera vida y para las 48652 tablets de la segunda vida por separado.

Luego en la Figura 5.16 se muestra un gráfico de barras que indica para cada una de las vidas (primera y segunda) cuántas de las tablets que presentaron defunciones lo hicieron en distintos intervalos de tiempo. En los intervalos se dividen las tablets que murieron en los primeros tres meses, entre los tres y seis meses, entre los seis meses y el año, entre uno y dos años y luego de los dos años. Observando el gráfico correspondiente a la primera vida, la mayor cantidad de defunciones suceden en el intervalo de tiempo entre 6 meses y 1 año (20178 tablets). Para la segunda vida sin embargo, las defunciones suceden mayormente en los primeros tres meses (9979).

Por último se presenta, en la Figura 5.17, la distribución de la duración de cada una de las vidas de las tablets (primera y segunda vida), según si fueron censuradas o presentaron una defunción. Se observan en el panel correspondiente a los tiempos de vida hasta la primera defunción, dos modos bien definidos mientras que para las defunciones de la segunda vida hay solo uno, concentrándose la mayoría de las

5.2. Análisis exploratorio de los datos de las tablets

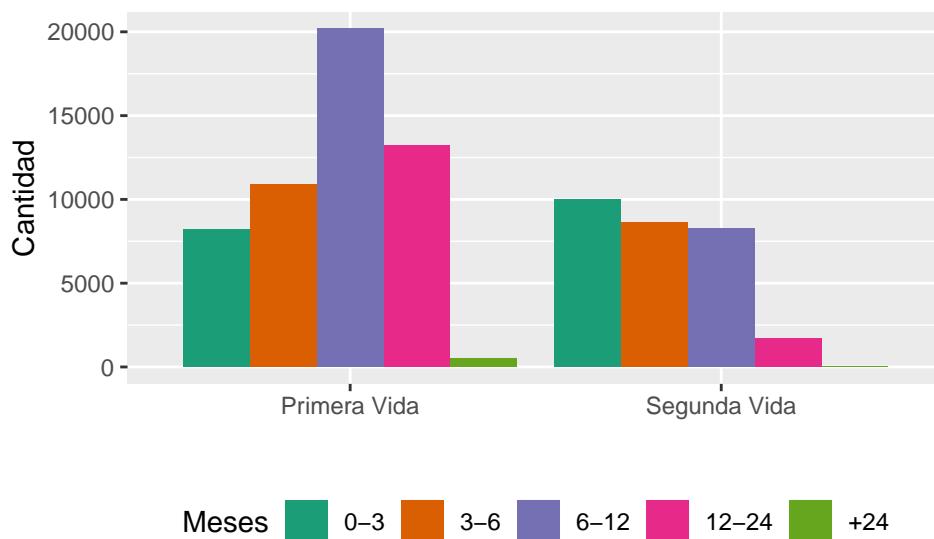


Figura 5.16: Gráficos de barras que muestra, para cada una de las vidas (primera y segunda) de las tablets, la cantidad de equipos que presentan una defunción en distintos intervalos de tiempo. (Entre 0 y 3 meses, entre 3 y 6 meses, entre 6 meses y un año, entre 1 y 2 años, en más de 2 años)

muertes de los equipos en los primeros 200 días. Por otro lado, se puede apreciar en general que los tiempos de vida hasta las segundas defunciones son menores que los tiempos hasta las primeras defunciones.

Sin embargo, en proporción, para la generación de tablets entregadas en 2014, el 51.6 % de las tablets de la primera vida y el 20 % de las de la segunda vida, llegan al año de vida; mientras que para la generación del 2015, el 40.3 % de las de la primera y el 19.3 % de las de la segunda vida lo hacen.

Se quiere indagar además si las causas de defunción de la primera y segunda vida son similares, por lo que se realiza un gráfico de mosaico en el que se muestra la probabilidad condicional de presentar una defunción por determinada causa en la segunda vida habiendo dejado de funcionar por esa misma u otra causa en la primera vida (Figura 5.18). Como se ha mencionado, para la segunda vida sigue siendo la placa la principal causa de defunción (66 %) de las cuales 68.4 % en su primera vida también habían muerto por haberseles roto dicha pieza.

CAPÍTULO 5. DESCRIPCIÓN Y ANÁLISIS EXPLORATORIO DE LOS DATOS

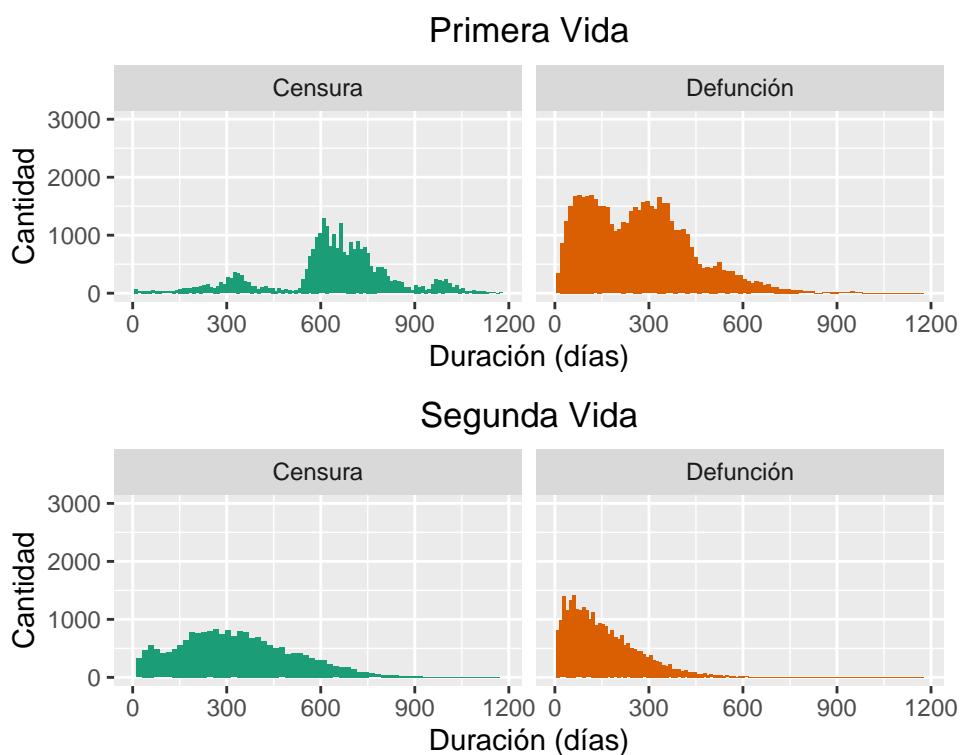


Figura 5.17: Histogramas con los tiempos de vida de las tablets. Se diferencia entre la primera vida (paneles superiores) y la segunda vida (paneles inferiores). A su vez, en color verde se representa el tiempo de vida de las tablets que presentan una defunción, y en color naranja el tiempo de las que son censuradas en las respectivas vidas.

5.2. Análisis exploratorio de los datos de las tablets

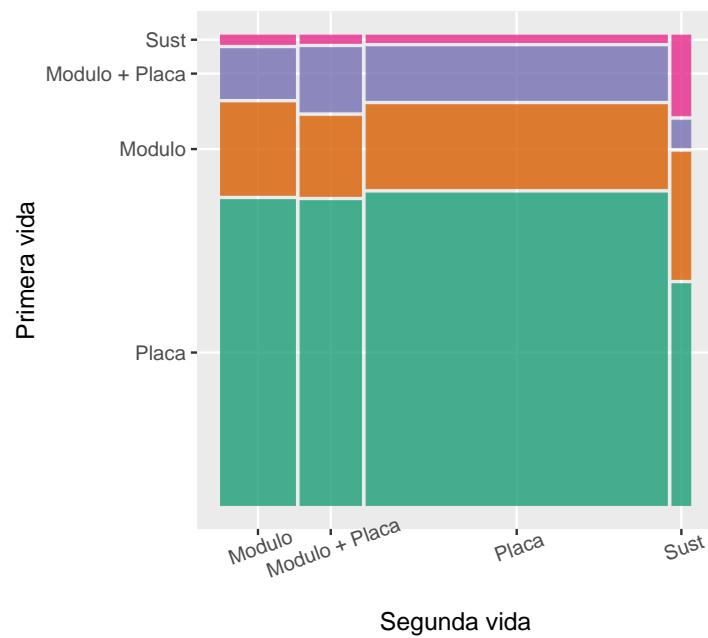


Figura 5.18: Gráfico de mosaico con la distribución de la causas de defunción de la segunda vida condicional a las causas de defunción de la primera vida

CAPÍTULO 5. DESCRIPCIÓN Y ANÁLISIS EXPLORATORIO DE LOS DATOS

Capítulo 6

Resultados

En este capítulo se presenta el resultado de la estimación de la función de confiabilidad a través del método de Kaplan-Meier con su respectivo intervalo de confianza (Sección 6.1). Se presentan los resultados para la primera y segunda vida de las tablets por separado y luego una comparación de ellos utilizando los paquetes **survival** (Therneau and Grambsch, 2000) y **survminer** (Kassambara and Kosinski, 2018) del software estadístico R.

Luego, se presenta para la primera vida, la comparación y testeo de curvas de confiabilidad estimadas de distintos grupos, correspondientes a las categorías de las variables sociodemográficas y las categorías de las variables técnicas mencionadas en el capítulo anterior (Sección 6.2). En esta parte del estudio se introduce el modelo de regresión de Cox como una alternativa de estimación de la función de supervivencia con la influencia de variables explicativas (Sección 4.3). Aquí nuevamente se utilizan funciones de los paquetes **survival** para estimar los modelos y **survminer** para la realización de las visualizaciones.

Por último se presenta la estimación núcleo de la función de riesgo, para la primera y para la segunda vida, siguiendo la metodología planteada en el Capítulo 4, tanto para la elección del núcleo como para el ancho de banda en el interior y en el borde (Sección 4.4). En esta parte del capítulo se utiliza el paquete **muhaz** (Hess and

CAPÍTULO 6. RESULTADOS

Gentleman, 2014) para obtener dichos resultados.

6.1. Estimación Kaplan-Meier

6.1.1. Primera vida

Se consideran las 80500 tablets que tienen una primera vida, y se realiza una estimación de la función de supervivencia o confiabilidad para un modelo con datos censurados a la derecha.

La cantidad de eventos (defunciones) registrados, tal como se ha mencionado anteriormente en el análisis descriptivo, asciende a 53035, lo cual representa el 65.9 % del total de las tablets.

La media de la función de supervivencia estimada, se basan en un estimador truncado. Es decir, si la última observación no es una muerte, entonces la estimación de la curva de supervivencia no va a cero y la media no está definida. La estimación de la misma se realiza integrando la función de supervivencia estimada hasta un valor T :

$$\hat{\mu} = \int_0^T \hat{S}_{KM}(t) dt.$$

Una opción es establecer el límite superior en una constante, en este trabajo se define T como el tiempo de seguimiento máximo observado durante el estudio.

En el presente análisis, la vida media o media de la supervivencia es estimada en 538.6 días, con límite superior igual a 1181.1 días y una desviación estandar de 1.7 días.

La vida mediana de las tablets, la cual representa el tiempo en el cual el 50 % de la generación ha sobrevivido, es estimada en 366 días. Su intervalo de confianza al 95 % es entre 364 y 369.2 días.

6.1. Estimación Kaplan-Meier

En el Anexo B.2 se muestra la información de la estimación de la función de supervivencia a través del método de Kaplan-Meier para cada tiempo de supervivencia y una visualización del resumen de la misma.

Se presenta en la Figura 6.1 (panel izquierdo), el gráfico de la función de supervivencia estimada por el método de Kaplan-Meier¹, para el total de las tablets consideradas en el grupo primera vida, teniendo en cuenta la censura por la derecha e ignorando los distintos grupos de riesgo que se determinan por las distintas covariables que se especifican más adelante. Conjuntamente se grafica su respectivo intervalo de confianza al 95 % calculado con una aproximación logarítmica². En el panel derecho se traza una función arbitraria que define una transformación de la curva de supervivencia, llamada función de riesgo acumulado, $f(S(t)) = -\log(S(t))$.

Observando la curva de confiabilidad que se ha obtenido, se ve como en los primeros 450 días de vida de los dispositivos, la supervivencia de los mismos disminuye progresivamente, hasta mantenerse más constante a partir de ese tiempo, de hecho más de la mitad de las tablets ya han muerto. Como la estimación de la curva de supervivencia no vale cero en ningún momento, se sabe que el último dato registrado es un tiempo de censura. Otra aclaración pertinente es que los intervalos de confianza son casi inperceptibles y es debido al gran volumen de datos con el que se está trabajando. Con el gráfico de la función de riesgo acumulado, que muestra la probabilidad de defunción del equipo en el tiempo, se puede confirmar que el número de fallos se acumula con los días. Y por último, al observar las dos curvas en conjunto se puede comprobar que dichas funciones son equivalentes entre si como se mencionó en la metodología.

¹Este y los siguientes gráficos en los que se dibujan las diferentes curvas de supervivencia, fueron obtenidos con la función `ggsurvplot()` del paquete `survminer`.

²La aproximación logarítmica es la opción por defecto utilizada para calcular los límites de confianza. Esta opción calcula los intervalos en función del riesgo acumulado o el $\log(\text{supervivencia})$.

CAPÍTULO 6. RESULTADOS

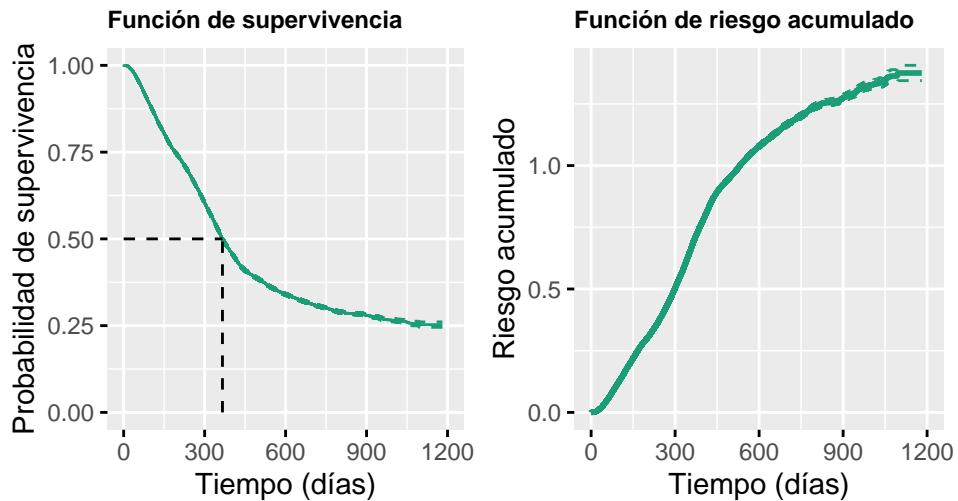


Figura 6.1: Panel izquierdo: curva de supervivencia (línea verde continua) y su intervalo de confianza (líneas verdes punteadas), estimada a través del método de Kaplan-Meier para las 80500 tablets que tienen una primera vida. La línea negra punteada, indica el valor de la mediana de supervivencia igual a 366 días. Panel derecho: curva de la función de riesgo acumulado, asociada a las probabilidades de supervivencia de la primera vida de las tablets.

6.1.2. Segunda vida y comparación

Luego de que la tablet presenta la rotura que le impide seguir funcionando, es reparada por servicio técnico y vuelta a entregar al alumno o alumna, comenzando lo que se considera su segunda vida. Entonces, ¿cuanto durará ese equipo luego de su primera reparación con repuestos que implican su defunción? ¿Valdrá la pena arreglarlo? Con el fin de investigar estos interrogantes, se considera como otro grupo a analizar y comparar con la primera vida, la segunda vida de las tablets.

Se realiza entonces la estimación de la función de supervivencia de las 48652 tablets que presentan una segunda vida, a través de la estimación Kaplan-Meier, considerando nuevamente datos censurados a la derecha y no incluyendo aún ninguna variable explicativa.

Los resultados obtenidos son: el 58.8 % (28606) del total de las tablets registraron una segunda defunción mientras que las 41.2 % restantes (20046 tablets) fueron censuradas. La media de la supervivencia es estimada, con límite superior igual a 1046 días, en 411.9 días con desvío estándar de 2.6 días. La vida mediana es estimada en 239.8 días, su intervalo de confianza al 95 % es entre 237 y 243.8 días.

Siguiendo el procedimiento, se comparan las funciones de supervivencia de la primera y segunda vida, con el fin de obtener una aproximación de cuan conveniente es reparar los equipos en función de cuanto sobreviven una vez realizada la reparación, respecto a lo que sobrevivieron antes de la primer rotura que impidió seguir su funcionamiento.

Para ello se construye un gráfico donde se observa la estimación de Kaplan-Meier de las funciones de supervivencia, con sus respectivos intervalos de confianza al 95 %, para la primera y segunda vida, considerándolas como dos grupos diferentes. Conjuntamente con el gráfico, en la Figura 6.2, se presenta la tabla de riesgos que muestra, para cada una de las vidas, el número absoluto de equipos en riesgo en cada tiempo de supervivencia, el número acumulado de eventos y el número acumulado

CAPÍTULO 6. RESULTADOS

de censuras.

El gráfico parece indicar que las tablets del grupo de la primera vida, tienen un mejor pronóstico de supervivencia que las del grupo de la segunda vida, teniendo esta última una pendiente mucho mayor. Quiere decir que luego de que los equipos son reparados después de su primer defunción, vuelven a romperse rápidamente. Las curvas tienen una diferencia uniforme hasta los primeros 400 días aproximadamente, manteniendo una amplia distancia que luego va disminuyendo.

Una prueba de significación de esta diferencia es proporcionada por los test de G-rho de Fleming y Harrington, los cuales incluyen el test log-rank y el test de Peto-Peto. A priori no se tiene información de que prueba podría proporcionar el mayor poder estadístico, ni como se podría violar la hipótesis nula, por lo que se implementan los dos test mencionados y se presentan resultados comparativos. (Ver Anexo B.3)

El valor del estadístico log-rank es 2933 y bajo la hipótesis nula la distribución asintótica de este estadístico es χ^2_1 , el valor del estadístico de Peto-Peto es 4284 y bajo la hipótesis nula este estadístico también se distribuye asintóticamente χ^2_1 . Como el *p*-valor es menor a 0.05 en ambos casos, hay evidencia para rechazar la hipótesis nula de igualdad de funciones de supervivencia (para un nivel de significación del 5%). Entonces se puede concluir que la primera vida y la segunda vida tienen curvas de supervivencia Kaplan-Meier significativamente diferentes siendo las tablets del grupo primera vida las que tienen mayor probabilidad de sobrevivir a medida que pasa el tiempo.

6.2. Estimación con variables auxiliares

Se vuelve a considerar solamente la primera vida de las tablets y se complementa el análisis realizado agregando al modelo cuatro variables explicativas como posibles predictoras del tiempo de supervivencia *T*, donde *T* indica cantidad de días hasta que el equipo presenta por primera vez una rotura que impide continuar con su

6.2. Estimación con variables auxiliares

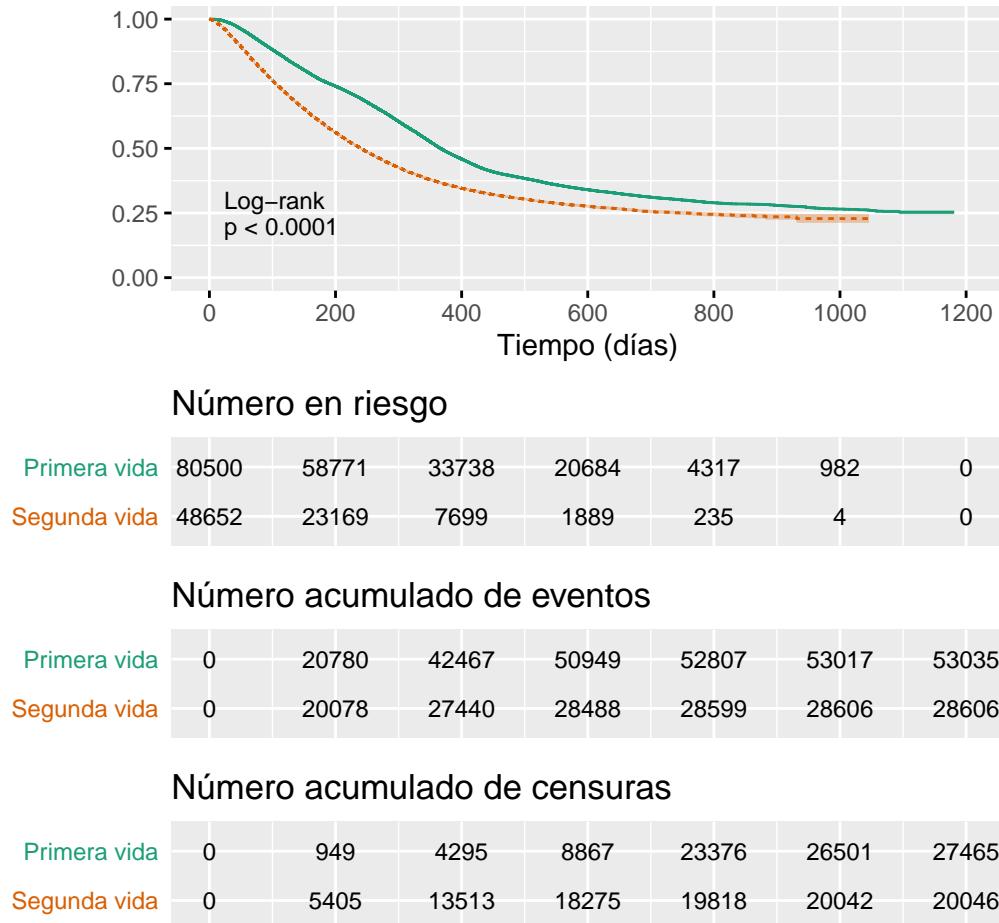


Figura 6.2: Curvas de supervivencia estimadas por el método de Kaplan-Meier de las 80500 tablets pertenecientes al grupo primera vida (línea continua color verde) y de las 48652 tablets pertenecientes al grupo segunda vida (línea punteada color naranja). En el eje Y se muestra la probabilidad de supervivencia y en el eje X el tiempo. El sombreado de cada curva indica su intervalo de confianza al 95 %. Además para cada una de las vidas (o grupos), se muestran para intervalos de tiempo de 200 días, el número de observaciones en el conjunto de riesgo, el número acumulado de eventos y el número acumulado de censuras. En la parte inferior izquierda del gráfico se muestra el valor p calculado con el test de igualdad de funciones de supervivencia de log-rank, con pesos igual a 1.

CAPÍTULO 6. RESULTADOS

funcionamiento.

Las variables consideradas (todas variables categóricas), como ya fue expuesto en las secciones 5.2.1.5 y 5.2.1.9 , corresponden a características sociodemográficas de las alumnas y los alumnos y la entrada a servicio técnico de Ceibal. Se dividen de la siguiente manera:

- Sexo del alumno o alumna. Dos grupos: femenino y masculino.
- Área geográfica en la que se encuentra la escuela. Tres grupos: interior rural, interior urbano y Montevideo.
- Contexto en el que se encuentra la escuela: Cinco grupos: desde quintil 1 hasta 5.
- Entradas de la tablet a servicio técnico. Dos grupos: el equipo fue reparado antes de morir, el equipo no tuvo ninguna reparación previo a su defunción. Dicho de otra forma, el equipo entró o no entró a servicio técnico (ST) antes de su defunción.

Desde el punto de vista tecnológico, el desgaste que la máquina sufre con el tiempo es un componente que podría aumentar el riesgo de experimentar roturas en las tablets. En cambio, desde el punto de vista de las características sociodemográficas, no se cuenta con la certeza de que dichos factores influyan de manera diferente en el riesgo de rotura.

Para ello se estima a través del método de Kaplan-Meier la curva de supervivencia para las diferentes variables, con la finalidad de encontrar alguna diferencia considerable entre estas. Se presentan dichos gráficos en la Figura 6.3.

A primera vista los diferentes sexos presentan una clara diferencia entre sus curvas de supervivencia, con un comportamiento paralelo en donde los alumnos de sexo masculino presentan mayor riesgo. Esto puede estar asociado a la educación que se le sigue dando a niñas y niños, siendo las primeras mayormente estimuladas para el cuidado de las cosas.

6.2. Estimación con variables auxiliares

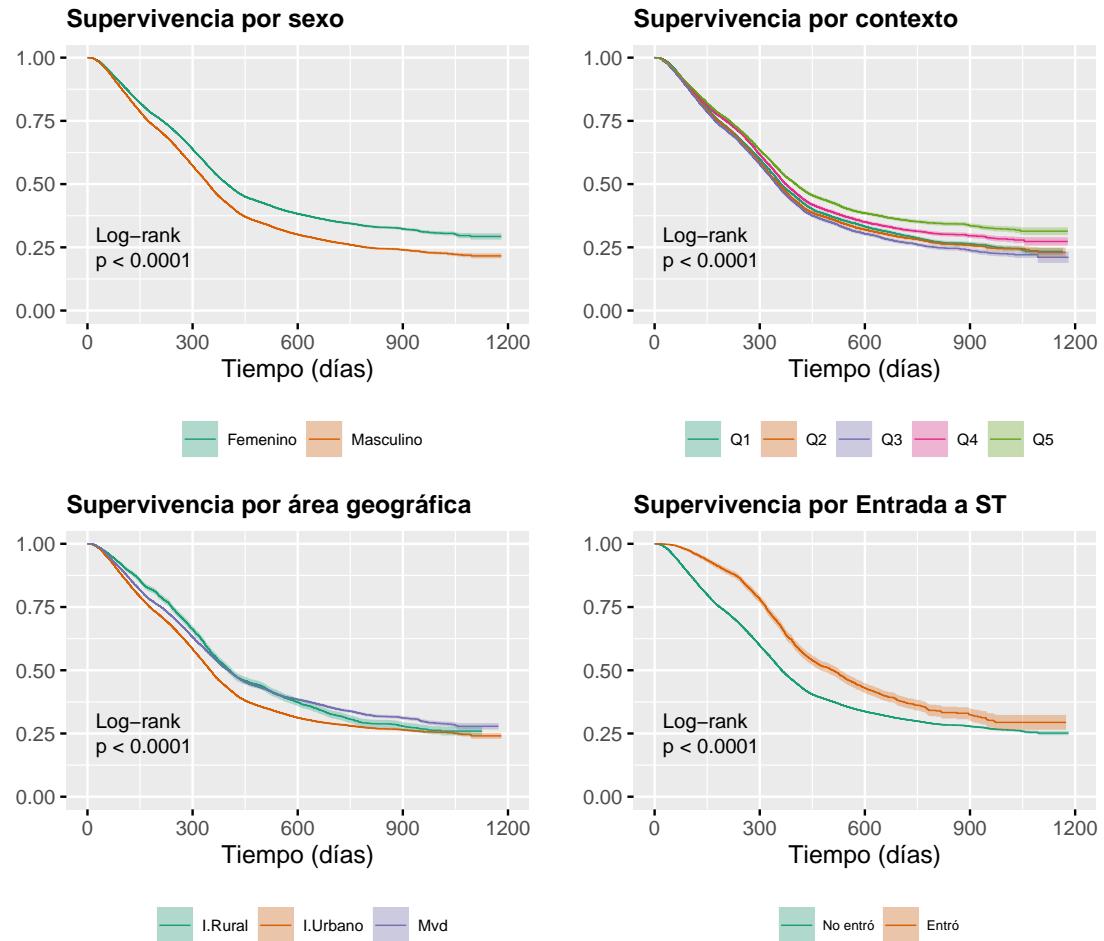


Figura 6.3: Gráficos de las funciones de supervivencia estimadas por el método de Kaplan-Meier para cada una de las variables auxiliares: sexo (panel superior izquierdo), área geográfica (panel inferior izquierdo), contexto (panel superior derecho) y entrada a ST (panel inferior derecho). Para cada una de las curvas se dibuja además, con un sombreado del mismo color de la misma, su intervalo de confianza al 95 %. Para cada variable se muestra también, el p-valor de los test de igualdad de curvas de supervivencia calculado con el método log-rank.

CAPÍTULO 6. RESULTADOS

Las curvas por área geográfica muestran que las tablets entregadas en escuelas urbanas del interior del país son más vulnerables en relación a las entregadas en el resto del país, cuyas curvas son más parecidas y se entremezclan solapándose en varios momentos del tiempo los intervalos de confianza. El resultado de la curva del área rural podría estar asociado a que dicho sector tal vez cuenta con menor acceso a los centros de reparación de Ceibal.

El gráfico de los diferentes contextos socioculturales muestra también tendencias similares casi paralelas. La supervivencia de las tablets del contexto menos crítico, es mayor respecto a los demás, y las del quintil 3 son aquellas que presentan mayor riesgo de dejar de funcionar a lo largo del tiempo. El contar con otras alternativas tecnológicas parece hacer que el alumnado de contextos más favorables cuiden más sus equipos y por lo tanto tengan roturas más tardías. Para aquellas niñas y niños de nivel más crítico, que probablemente no cuentan con el acceso a otras tecnologías, también parece hacer que el cuidado sea mayor.

Se observa además en estos tres casos, como en los primeros 400 días aproximadamente, la supervivencia de estas tablets disminuye progresivamente hasta mantenerse más estable a partir de ese tiempo.

Por último, las tablets sin ninguna reparación antes de su primera defunción tienen un menor riesgo de dejar de funcionar. Para los equipos que nunca fueron a servicio técnico, la probabilidad de sobrevivir disminuye rápidamente al inicio de la vida, a diferencia de los otros, que tienen un mejor pronóstico. Al final del período de análisis, las curvas son más similares, mostrando una probabilidad de sobrevivir más constante.

Para testear la significancia de la diferencia entre las categorías de cada una de las variables y los supuestos antes mencionados, se realizan las respectivas pruebas de hipótesis. En todos los casos el *p*-valor es prácticamente cero por lo que hay evidencia para rechazar la hipótesis nula que establece que todas las curvas de supervivencia de la misma variable son iguales, concluyendo que las funciones de supervivencia de

cada grupo son significativamente diferentes.

En los Anexos B.4 y B.5 se muestra información descriptiva sobre las diferentes curvas de Kaplan-Meier junto con los resultados de las pruebas de log-rank y de Peto-Peto, ambas con un nivel de confianza del 95 %.

6.2.1. Estimación con el modelo de Cox

Como alternativa a la estimación Kaplan-Meier se aplicará el modelo de riesgos proporcionales de Cox, usualmente conocido como un modelo semi paramétrico, ya que como se ha explicado en la Sección 4.3, es paramétrico porque especifica un modelo de regresión con una forma funcional determinada y es no paramétrico en cuanto que no especifica la forma exacta de la distribución de los tiempos de supervivencia.

En primer término se consideran cuatro modelos distintos. En el *Modelo A* y el *Modelo B* se incluyen únicamente las covariables *sexo* y *contexto* respectivamente. El en *Modelo C* se incluyen ambas covariables, *sexo* y *contexto*. Y, el *Modelo D* es un modelo completo que incluye las cuatro covariables que se han utilizado en este trabajo: *sexo*, *área*, *contexto* y *ST*. Se elijen respectivamente las categorías *Femenino*, *I.Urbano*, *Quintil 5* y *0* como niveles de referencia de cada una de las variables.

$$\hat{S}(t, X) = \left[\hat{S}_0(t) \right]^{exp(\hat{\beta}_{sexo} sexo)} \quad (\text{A})$$

$$\hat{S}(t, X) = \left[\hat{S}_0(t) \right]^{exp(\hat{\beta}_{contexto} contexto)} \quad (\text{B})$$

$$\hat{S}(t, X) = \left[\hat{S}_0(t) \right]^{exp(\hat{\beta}_{sexo} sexo + \hat{\beta}_{contexto} contexto)} \quad (\text{C})$$

$$\hat{S}(t, X) = \left[\hat{S}_0(t) \right]^{exp(\hat{\beta}_{sexo} sexo + \hat{\beta}_{área} área + \hat{\beta}_{contexto} contexto + \hat{\beta}_{ST} ST)} \quad (\text{D})$$

En el Anexo B.6 se presenta el resultado de la estimación de cada uno de ellos a través de un modelo de Cox.

Para saber cual de estos modelos candidatos que incluyen distintas variables se podría elegir, se utiliza una cantidad conocida como el *Criterio de Información de*

CAPÍTULO 6. RESULTADOS

Akaike, o *AIC*. Este es una medida de la bondad de ajuste de un modelo estadístico. Se puede decir que describe la relación entre el sesgo y varianza en la construcción del modelo (el *AIC* no es una prueba del modelo en el sentido de la prueba de hipótesis, sino que es una prueba entre los modelos, una herramienta para la selección de los mismos). Dicha cantidad está dada por $AIC = -2\ln(L(\hat{\beta})) + kp$ donde $\ln(L(\hat{\beta}))$ es el máximo valor del logaritmo de la función de verosimilitud para el modelo estimado, p es el número de parámetros en el modelo y $k = 2$. Los valores de *AIC* equilibran dos cantidades que son propiedades de un modelo, la bondad de ajuste y el número de parámetros (que entra como un término de penalización). Por lo tanto, un modelo “bueno” es aquel que se ajusta bien a los datos (valor pequeño de la bondad de ajuste) con pocos parámetros, de modo que los valores más pequeños de *AIC* deberían en teoría indicar mejores modelos.

El mejor ajuste utilizando el criterio *AIC*, es el *Modelo D* (ver resultados en Anexo B.7), el cual incluye todas las variables explicativas. El *Modelo C*, que incluye tanto *sexo* como *contexto*, es una segunda opción cercana.

Aplicando la estimación de Cox al *Modelo D* o modelo completo, se obtienen con la función `coxph()` del paquete `survival` los resultados que se expresan en la Tabla 6.1. En ella se especifican los parámetros estimados $\hat{\beta}_i$, cuyas estimaciones son estimaciones de máxima verosimilitud parcial y para cada covariante, el parámetro β_i es la relación de riesgo para el efecto de ese parámetro en la supervivencia, ajustándose para las otras covariantes. Para variables dummy o categóricas representa el efecto del nivel correspondiente en comparación con la categoría de referencia. También se pueden mostrar esos resultados como un diagrama de bosque, lo cual se muestra en la Figura 6.4³. Este es un gráfico de las estimaciones de los coeficientes y los intervalos de confianza al 95 %, cada uno con respecto a un nivel de referencia. Por ejemplo, se puede ver que las tablets de las niñas (la referencia) tienen menos riesgo que las de los niños, las del interior rural y Montevideo tienen una mejor probabi-

³Dicho gráfico se realiza con la función `ggforest()` del paquete `survminer`.

6.2. Estimación con variables auxiliares

Tabla 6.1: Resultado del ajuste de un modelo de Cox al modelo que incluye las variables sexo, área, contexto y ST (modelo completo).

	coef	exp(coef)	se(coef)	lower .95	upper .95	z	Pr(> z)
SexoMasculino	0.22	1.25	0.01	1.23	1.27	25.21	0.00
AreaI.Rural	-0.17	0.85	0.02	0.81	0.88	-8.30	0.00
AreaMontevideo	-0.19	0.83	0.01	0.81	0.85	-19.39	0.00
ContextoQuintil 1	0.18	1.20	0.01	1.17	1.23	13.24	0.00
ContextoQuintil 2	0.18	1.20	0.01	1.17	1.23	13.31	0.00
ContextoQuintil 3	0.20	1.22	0.01	1.19	1.26	14.31	0.00
ContextoQuintil 4	0.09	1.09	0.01	1.06	1.12	6.22	0.00
ST1	-0.35	0.70	0.03	0.67	0.74	-13.46	0.00

Concordance = 0.549 (se = 0.001)

Likelihood ratio test = 1541 on 8 df, p= <2e-16

Wald test = 1509 on 8 df, p= <2e-16

Score (logrank) test = 1515 on 8 df, p= <2e-16

CAPÍTULO 6. RESULTADOS

lidad de supervivencia que las del interior urbano (la referencia). Además se puede ver que los quintiles mas críticos (quintiles 1, 2 y 3) tienen peores resultados que los quintiles menos críticos (quintiles 4 y 5) y que entrar a servicio técnico una vez antes de morir reduce el riesgo respecto a las tablets que no entran (la referencia). Estas conclusiones concuerdan con las estimaciones obtenidas anteriormente a través del método de Kaplan-Meier, particularmente se puede comparar con la Figura 6.3 y observar esa relación.

Por otro lado, en la columna $Pr(>|z|)$ de la tabla con los resultados de la estimación del modelo de Cox aplicado al modelo completo, se pueden observar las variables que son significativas y las que no, ya que z es el test de Wald, asintóticamente normal bajo la hipótesis de nulidad del coeficiente. En este caso todos los coeficientes son significativos al 1%.

Además se muestran en el resultado, estadísticos de Concordancia y R^2 , así como los estadísticos de prueba, grados de libertad y p-valores para la significación del modelo de los test de hipótesis de razón de verosimilitud, de Wald y de puntajes (ver metodología, Sección 4.3.3). Los tres criterios son asintóticamente equivalentes, con sus p-valores significativos. Por tanto, se puede afirmar que el modelo permite explicar la variable tiempo de supervivencia considerada.

Otra información importante que amplía lo expresado en párrafos anteriores, es la estimación de los riesgos relativos. $\exp(\text{coef})$ permite una interpretación de los efectos de las variables explicativas sobre la función de riesgo $h(t)$, y se acompaña de un intervalo de confianza al 95 %. Se observa por ejemplo que los equipos entregados a niños tienen 1.2 veces el riesgo de morir en relación con los equipos entregados a niñas. Pertenercer al área interior rural hace que la defunción del equipo tenga un riesgo de 0.8 veces el riesgo de muerte de los que pertenecen al interior urbano. En cuanto al contexto en donde se entrega el equipo, pertenecer al quintil 3 es 1.2 veces más riesgoso que pertenecer al quintil 5. Finalmente, haber entrado a servicio técnico antes, hace que la muerte tenga un riesgo de 0.7 veces el riesgo de muerte

6.2. Estimación con variables auxiliares

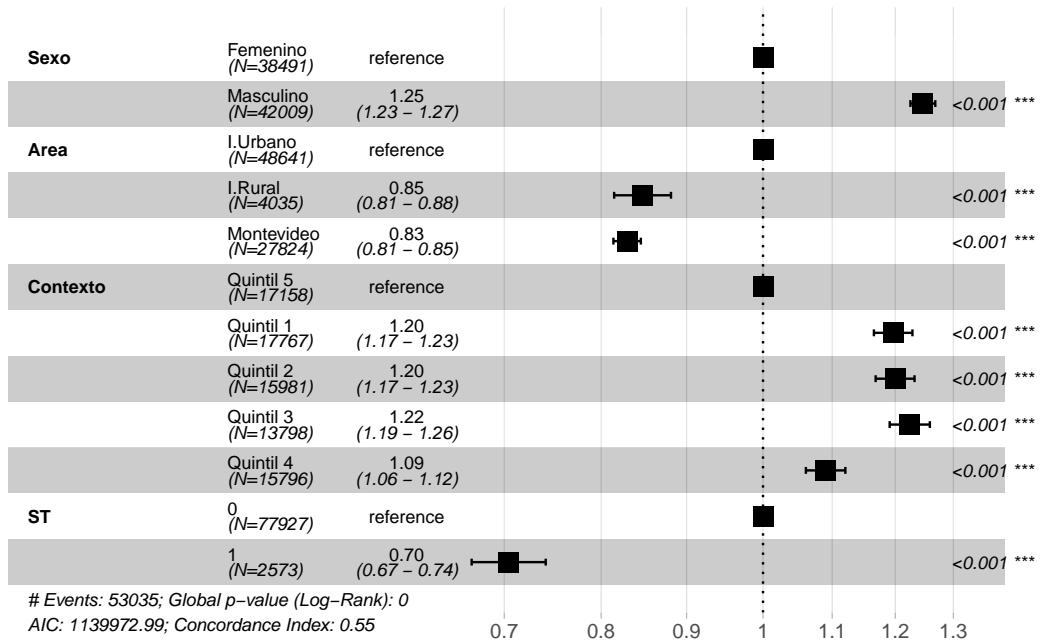


Figura 6.4: Gráfico de bosque que muestra de forma sintetizada el resultado del ajuste del modelo de regresión Cox aplicado al grupo de la primera vida, considerando las variables explicativas sexo, área geográfica, contexto sociocultural y entrada a ST. Para cada una de las variables indica: las categorías, la cantidad de observaciones en cada una de ellas y los coeficientes estimados con sus respectivos intervalos de confianza al 95 %, de forma numérica y de forma gráfica, ubicándolos a la derecha o la izquierda de una recta vertical en el valor 1, según si la probabilidad de sobrevivir es mayor o menor respecto a la categoría de referencia de cada variable. También se muestran la cantidad total de eventos, el p-valor del test log-rank global y de cada variable, el valor de AIC y el índice de concordancia.

CAPÍTULO 6. RESULTADOS

de los que no entraron ninguna vez.

6.2.1.1. Diagnóstico del modelo

El modelo de Cox hace varias suposiciones. Por lo tanto es importante evaluar si el modelo de regresión de Cox ajustado describe adecuadamente los datos. En concreto se debe comprobar:

- Suposición de riesgos proporcionales.
- Si existen observaciones influyentes (o valores atípicos).
- Detectar la no linealidad de los efectos de las variables predictoras en la función de riesgo.

Para verificar estas suposiciones del modelo, se utilizan diferentes tipos de residuos. Los residuos a considerar son:

- Residuos de *Schoenfeld* vs *tiempo* para verificar la suposición de riesgos proporcionales.
- Residuos de *Schoenfeld* vs *tiempo* o *martingala* vs *identificador de la observación* de cada predictora para evaluar la no linealidad.
- Desviación residual (transformación simétrica de los residuos de martingala) para examinar observaciones influyentes.

Observaciones influyentes

La detección de observaciones influyentes se realiza mediante métodos gráficos. En la Figura 6.5 se muestran los residuos tipo devianza⁴, en la cual se evidencia que no existe ningún individuo que esté influenciando el ajuste del modelo.

⁴Este y los siguientes gráficos en los que se dibujan los diferentes residuos que permiten evaluar el modelo son obtenidos con la función `ggcoxdiagnositcs()` del paquete `survminer` incluyendo en el argumento `type` el tipo de residuo que corresponda.

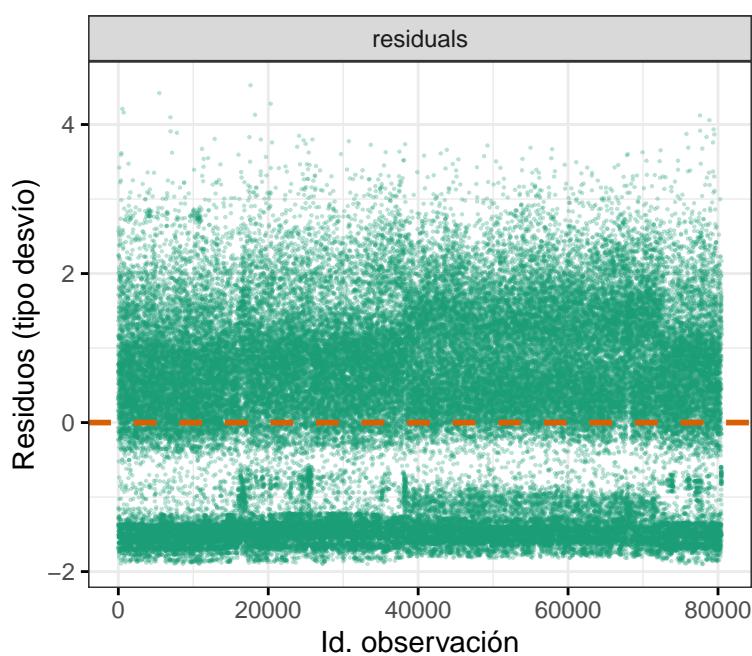


Figura 6.5: Gráfico de los residuos tipo devianza del modelo de Cox aplicado a la primera vida con las variables explicativas sexo, área, contexto y ST que permite detectar observaciones influyentes. En el eje Y se presentan los residuos y en el eje X la identificación de cada observación. Se agrega la línea horizontal que resalta el nivel $y = 0$.

CAPÍTULO 6. RESULTADOS

También se pueden generar gráficos de influencia sobre la estimación de cada coeficiente, los cuales se obtienen utilizando los residuos dfbetas y se muestran en la Figura 6.6.

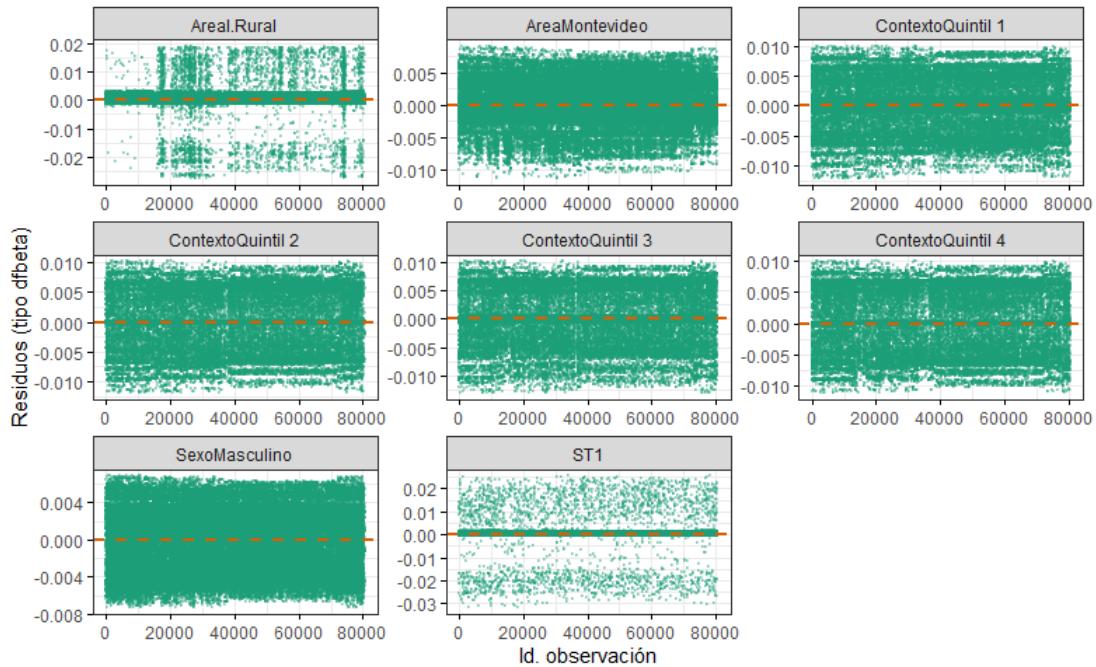


Figura 6.6: Gráfico de los residuos tipo dfbetas del modelo de Cox aplicado a la primera vida con las variables explicativas sexo, área, contexto y ST que permite detectar observaciones influyentes por separado para cada coeficiente estimado (uno en cada panel, dejando fuera los niveles de referencia). En el eje Y se presentan los residuos y en el eje X la identificación de cada observación. En cada panel se agrega la línea horizontal que resalta el nivel $y = 0$.

El patrón de residuos es bastante simétrico respecto de cero. Además no hay residuos con valores claramente fuera del rango (-0.03, 0.03), indicando que no hay ninguna observación que pueda identificarse como anómala. En caso de existir alguna se debería eliminar y ajustar el nuevo modelo.

No linealidad

El test de no linealidad sólo se puede interpretar para variables de tipo numérico, donde se pueden plantear diferentes tendencias en la respuesta en función de

6.2. Estimación con variables auxiliares

diferentes transformaciones numéricas de la predictora.

En este caso, que son todas variables categóricas, se puede evaluar el modelo creando residuos martingala y graficándolos frente a las predictoras seleccionadas. Los resultados que se exponen en la Figura 6.7, muestran que los residuos se distribuyen uniformemente sobre los valores de cada covariable.

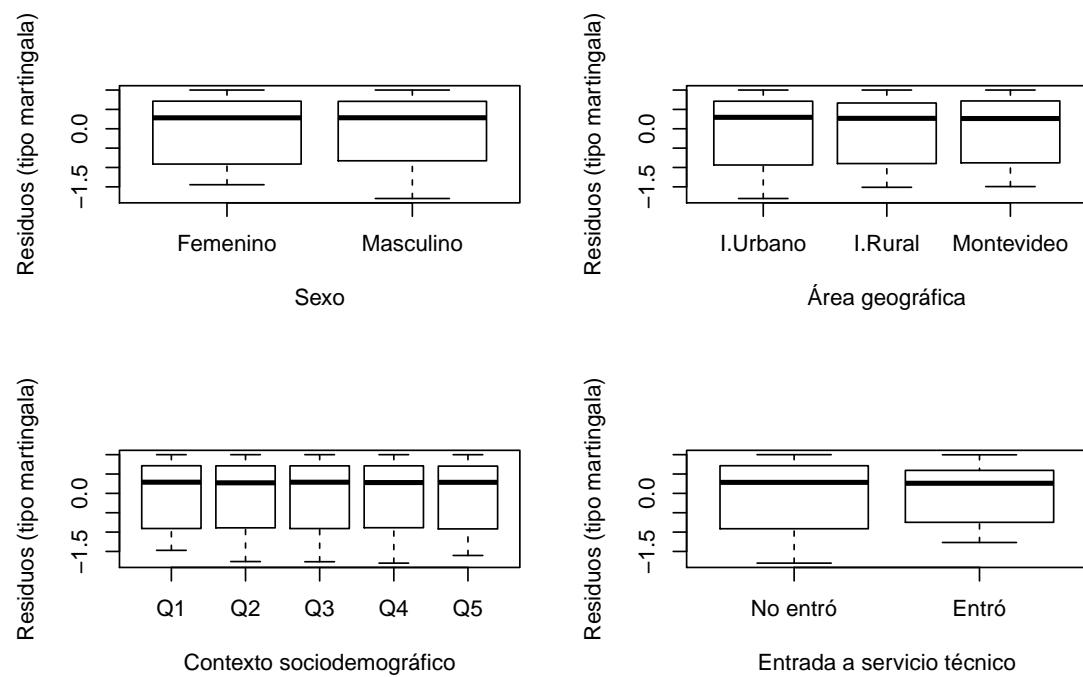


Figura 6.7: Gráfico de cajas en el que se dibuja para cada categoría de cada covariable (en paneles diferentes), los residuos tipo martingala, lo cual permite corroborar el supuesto de linealidad de los efectos de cada predictora. En el eje Y de cada una se presentan los residuos y en el eje X las categorías de la variable correspondiente.

Riesgos proporcionales

La validación de riesgos proporcionales se realiza con la prueba de hipótesis de correlación para los residuos Schoenfeld escalados de cada covariable con alguna transformación del tiempo. El valor predeterminado de dicha transformación (y utilizado en este trabajo) se basa en la estimación de Kaplan-Meier de la función de supervivencia.

CAPÍTULO 6. RESULTADOS

La suposición de proporcionalidad es que la tasa de riesgo de un individuo es relativamente constante en el tiempo, y esto es lo que prueba la función `cox.zph()` del paquete `survival`. La salida de dicha función, la cual se muestra en la Tabla 6.2, es una matriz con una fila para cada variable, y adicionalmente una última fila para la prueba global para el modelo en su conjunto. Las columnas de la matriz contienen el coeficiente de correlación entre el tiempo de supervivencia transformado y los residuos de Schoenfeld escalados *rho*, el estadístico de prueba chi-cuadrado *chisq* y su respectivo *p*-valor de dos lados (o colas). Para la prueba global no existe una correlación adecuada, por lo que se ingresa un *NA* en la matriz como marcador de posición.

Tabla 6.2: Verificación del supuesto de riesgos proporcionales del modelo de Cox aplicado a la primera vida de las tablets con las variables explicativas sexo, área, contexto y ST.

	rho	chisq	p
SexoMasculino	0.009	3.948	0.047
AreaI.Rural	0.038	76.825	0.000
AreaMontevideo	0.006	2.436	0.119
ContextoQuintil 1	-0.011	6.279	0.012
ContextoQuintil 2	0.002	0.145	0.703
ContextoQuintil 3	-0.002	0.199	0.656
ContextoQuintil 4	0.000	0.010	0.920
ST1	0.078	324.054	0.000
GLOBAL		418.698	0.000

Todos los *p*-valores (tanto los individuales de cada variable como el global para el modelo) son inferiores a 0.05 por lo que hay evidencia para rechazar la hipótesis nula no cumpliéndose el supuesto de riesgos proporcionales, tener valores *p* muy pequeños indica que hay coeficientes dependientes del tiempo.

6.2. Estimación con variables auxiliares

El supuesto de riesgos proporcionales también se puede verificar de forma gráfica dibujando para cada coeficiente los residuos de Schoenfeld escalados con la función `ggcoxzph()` del paquete `survminer`. Dicho resultado se muestra en el Anexo B.8 lo cual aporta la misma información que lo anterior.

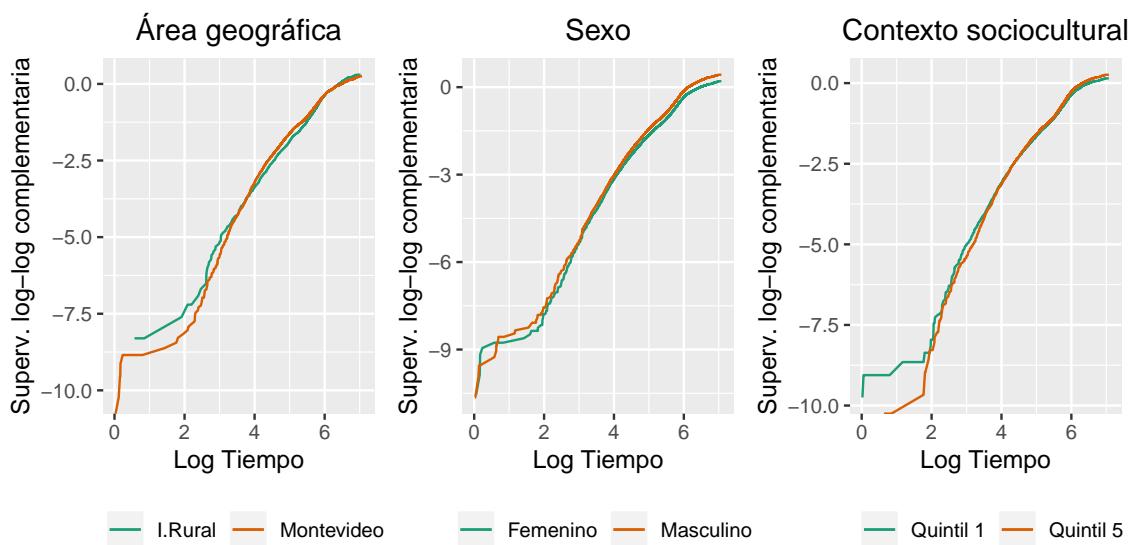


Figura 6.8: Gráficos de líneas utilizados para evaluar el supuesto de riesgo proporcionales. En el eje Y se grafica la transformación log-log complementaria de las tasas de riesgo de la primera vida de las tablets considerando las variables auxiliares sexo, área y contexto y en el eje X el logaritmo del tiempo. En el panel izquierdo se comparan las categorías interior rural (línea verde) y Montevideo (línea naranja) de la variable área, en el panel del medio las categorías femenino (línea verde) y masculino (línea naranja) de la variable sexo y en el panel derecho, las categorías quintil 1 y quintil 5 de la variable contexto.

Si se comparan los tiempos de supervivencia entre dos grupos, hay una gráfica simple que puede ayudar a evaluar el supuesto de riesgos proporcionales. La función $g(u) = \log - \log(u)$ se llama transformación $\log - \log$ complementaria, y tiene el efecto de cambiar el rango de $(0, 1)$ para u a $(-\infty, +\infty)$ para $g(u)$. Un gráfico de $g[S_1(t)]$ y $g[S_0(t)]$ versus t o $\log(t)$ producirá curvas paralelas separadas por β si el supuesto de riesgos proporcionales es correcto.

CAPÍTULO 6. RESULTADOS

Por ejemplo, en la Figura 6.8 se compara el área interior rural con Montevideo, por otro lado el quintil 1 con el 5 y ambos sexos, a través de la transformación de las tasas de riesgo mencionadas anteriormente. Para realizar dichas gráficas primero se crean las curvas de supervivencia pero considerando solamente el subconjunto de filas correspondientes a los datos de cada categoría de las que se quiere comparar, luego se realiza la transformación mencionada en el párrafo anterior. Para los tres casos, las curvas son claramente no paralelas (se cruzan) evidenciando que las relaciones de riesgo no son constantes en el tiempo, por lo tanto, no sería apropiado usar un modelo de riesgos proporcionales de Cox para esta situación.

Si una covariante rompe la suposición, es posible que deba corregirse ya que existen coeficientes dependientes del tiempo. Sin hacer algo al respecto, se podrían invalidar los resultados, de una manera similar a cómo se podrían romper los supuestos de regresión lineal. Entonces es natural preguntar en este punto, si el modelo de riesgos proporcionales de Cox no es apropiado ¿cómo se debe llevar a cabo el análisis?. Hay varias opciones que podrían servir, tales como:

- Analizar estratificando en las variables auxiliares; es decir, en lugar de ajustarse un modelo, obtener las curvas de Kaplan-Meier para cada grupo de cada variable por separado. Esta alternativa ya se realizó mas bien como análisis exploratorio en la Sección 6.2.
- Ajustar el modelo de Cox a dos subconjuntos diferentes según el momento en que la razón de riesgo cambia, para obtener dos estimaciones de razón de riesgo diferentes, una para cada uno de estos dos períodos, o directamente comenzar el análisis luego de cierta cantidad de días.
- Ajustar un modelo de Cox modificado que incluya variables dependientes del tiempo que midan la interacción de las covariables con el tiempo. Este modelo se llama modelo extendido de Cox.

6.2.1.2. Modelo extendido de Cox

El correspondiente modelo de Cox extendido, puede usarse para verificar el supuesto de riesgos proporcionales para las variables independientes del tiempo. Además, se puede usar para obtener una fórmula de razón de riesgo que considere los efectos de las variables que no satisfacen dicho supuesto como es en este caso.

En esta investigación, a modo de ejemplo, se considera en lugar del modelo completo con el que se venía trabajando, un nuevo modelo que incluye solo la covariable que indica si el equipo entró a servicio técnico antes de su defunción o no (ST), y se extenderá aplicando una función del tiempo, $g_i(t) = t$. Entonces, el modelo extendido de Cox en este caso toma la forma:

$$h(t, X(t)) = h_0 \exp [\beta_{ST} ST + \delta_{ST} (ST \times t)] .$$

Para ejecutar un modelo extendido de Cox en R, el conjunto de datos debe estar en el formato del proceso de conteo (inicio, detención). Esto se puede lograr con la función `survSplit()` del paquete `survival`. Dicha función crea múltiples observaciones para individuos en riesgo en múltiples puntos de tiempo, los cuales son proporcionados por el usuario. La opción más general es un vector de puntos de corte de tiempo que incluye todos los tiempos de eventos en los datos.

Otra forma es utilizar la función `coxphw()` del paquete `coxphw` (Dunkler et al., 2018) la cual realiza una regresión ponderada de Cox, proporcionando estimaciones imparciales de las razones de riesgo promedio, también en caso de riesgos no proporcionales. Los efectos dependientes del tiempo se pueden estimar convenientemente incluyendo interacciones de covariables con funciones arbitrarias del tiempo, con o sin el uso de la opción de ponderación.

En la Tabla 6.3 se presenta el resultado de dicha estimación.

El estadístico z de la prueba de Wald de 501.3 y el valor p significativo ($p = 0$)

CAPÍTULO 6. RESULTADOS

Tabla 6.3: Estimación de un modelo extendido de Cox

	coef	exp(coef)	se(coef)	z	p
ST1	-1.13	0.32	0.05	-22.37	0.00
Duracion:ST1	0.00	1.00	0.00	19.66	0.00

Wald Chi-square=501.3418 on 2df, p=0, n=80500

sugiere que la razón de riesgo para ST no es igual a lo largo del tiempo. En otras palabras, el valor p significativo proporciona evidencia de que la suposición de riesgo proporcional se viola.

6.3. Estimación de la función de riesgo

La función de riesgo (Sección 3.1.4) es especialmente útil para describir los cambios temporales de la probabilidad de experimentar un fallo. Para obtener desde otro enfoque, alguna idea de como se comporta el riesgo de presentar una defunción, es que se hace hincapié en esta función estimándola a través del método propuesto en la metodología (Sección 4.4).

6.3.1. Primera vida

Se aplica, utilizando la función `muhaz()` del paquete `muhaz`, la estimación por núcleos de la función de riesgo de la primera vida de todas las tablets (incluyendo las observaciones censuradas), con un núcleo de Epanechnikov. Se realiza la comparación con otras posibles funciones núcleos: rectangular, bicoadrática y tricoadrático, sin observarse diferencias notorias entre ellas, por lo que se decide seguir trabajando con el de Epanechnikov (ver Anexo B.9).

Otra consideración importante en la estimación no paramétrica de curvas, es la selección del ancho de ventana (o bandwith). Una alternativa es utilizar una *ventana global*, es decir el mismo parámetro ventana para todos los puntos en los que se hará

6.3. Estimación de la función de riesgo

la estimación.

Puede ser que este método no presente buenos resultados. Entonces, en regiones de baja densidad de datos es preferible elegir un parámetro ventana grande, mientras que en aquellas áreas en las que existen un mayor número de datos, es preferible elegir parámetros ventana pequeños, pues evidenciarán más claramente las características locales de la función a estimar. Con esta filosofía se construyen los estimadores de *ventana local*, que utilizan una ventana diferente dependiendo del punto concreto donde se realice la estimación.

En el panel superior de la Figura 6.9 se presenta la estimación de la función de riesgo de las 80500 tablets que presentan una primera vida, utilizando un método de selección de ancho de banda global, cuyo valor óptimo igual a 17.1 se obtiene minimizando el IMSE, aplicando correcciones de borde izquierda y derecha y un núcleo de Epanechnikov. En el panel inferior se presenta la estimación de la función de riesgo de la primera vida pero utilizando un método de selección de ancho de ventana local, en donde el valor óptimo en cada punto se obtiene minimizando el MSE local, también con corrección en ambos bordes y núcleo de Epanechnikov.

Para ambos casos, el tiempo mínimo es por defecto 0 y el tiempo máximo es por defecto 1174.1 días, correspondiendo al momento en que 10 tablets permanecen en riesgo. El ancho de banda inicial (pilot bandwidth) utilizado en la minimización del ECM es por defecto el recomendado por Muller y Wang, expresado en la metodología, igual a 16.7. Por último, el ancho de banda utilizado para suavizar los anchos de banda locales (smoothing bandwidth), es 5 veces el ancho de banda inicial, es decir 83.3.

Para la comparación de las estimaciones de la función de riesgo suavizado, también se grafica la curva de estimación exponencial por tramos obtenida con la función `pehaz()` del paquete `muhaz`. Esta, da una idea de las características de los datos sin que el tratamiento de los mismos implique un suavizado. Éste método de estimación paramétrica divide el dominio de tiempo en intervalos de igual ancho y luego calcula

CAPÍTULO 6. RESULTADOS

el riesgo en cada intervalo como el número de eventos dividido entre el tiempo total de seguimiento en ese intervalo. No se agregan detalles de este modelo, ya que solo se utiliza de forma comparativa y de ayuda para confirmar que se están generando estimaciones realistas de la función de riesgo referida. Se utiliza el ancho del intervalo por defecto igual a $\frac{(tiempo.max - tiempo.min)}{(8 * (nu)^{0,2})} = \frac{(1181,1 - 0)}{8 * (53035)^{0,2}} = 16,8$.

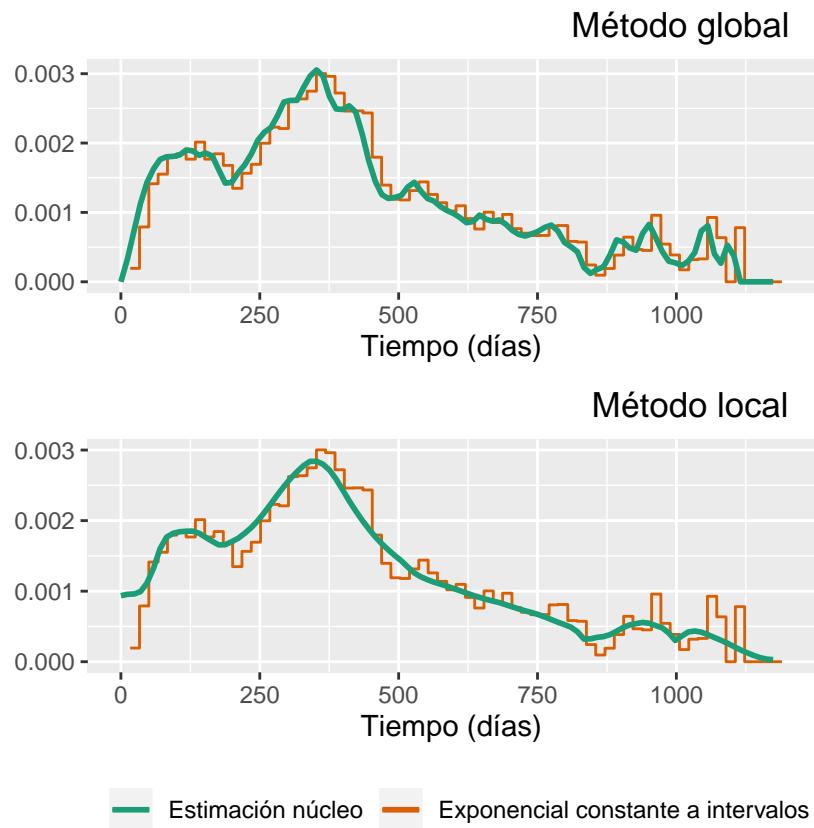


Figura 6.9: Gráficos de la estimación por núcleos de la función de riesgo de la primera vida de las tablets (líneas en color verde). En el panel superior, se utiliza un ancho de ventana global, siendo el ancho óptimo igual a 17.1 mientras que en el panel inferior se realiza la estimación con un ancho de ventana local. Para ambos casos el ancho de banda inicial es igual a 16.7 y se aplica corrección de límites en el borde izquierdo y el derecho, el núcleo utilizado es el de Epanechnikov. Las líneas en color naranja representan la estimación de la función de riesgo utilizando el modelo exponencial constante a intervalos.

Se observa un crecimiento parcial del riesgo en los primeros meses de vida de los

6.3. Estimación de la función de riesgo

equipos, que luego se incrementa, alcanzándose la mayor tasa de riesgo en aproximadamente los 350 días. A medida que transcurre el tiempo, el riesgo disminuye, es decir que la probabilidad de que las tablets se rompan dejando de funcionar dado que no se han roto hasta ese momento, va siendo cada vez menor. Al comparar la estimación obtenida seleccionando una ventana global con ancho de banda variables, se observa que la curva obtenida con ventanas locales es más suavizada, destacando que se asemeja menos a la estimación exponencial constante a intervalos.

6.3.2. Segunda vida

Se vuelven a realizar las mismas estimaciones que las realizadas para el grupo primera vida, pero considerando las 48652 tablets que tienen una segunda vida después de su primer defunción. Se presentan los resultados en la Figura 6.10⁵.

En la figura se observa como el riesgo de presentar una segunda defunción es muy alto al principio de la segunda vida de las tablets, y ese riesgo va disminuyendo con el paso de los días hasta volverse más insignificante. Se advierte nuevamente que con una ventana global la estimación es más rugosa asemejándose a los picos de la función escalonada, y con una ventana local, es más suavizada.

⁵Se utiliza la función `ggplot()` para obtener dicha visualización.

CAPÍTULO 6. RESULTADOS

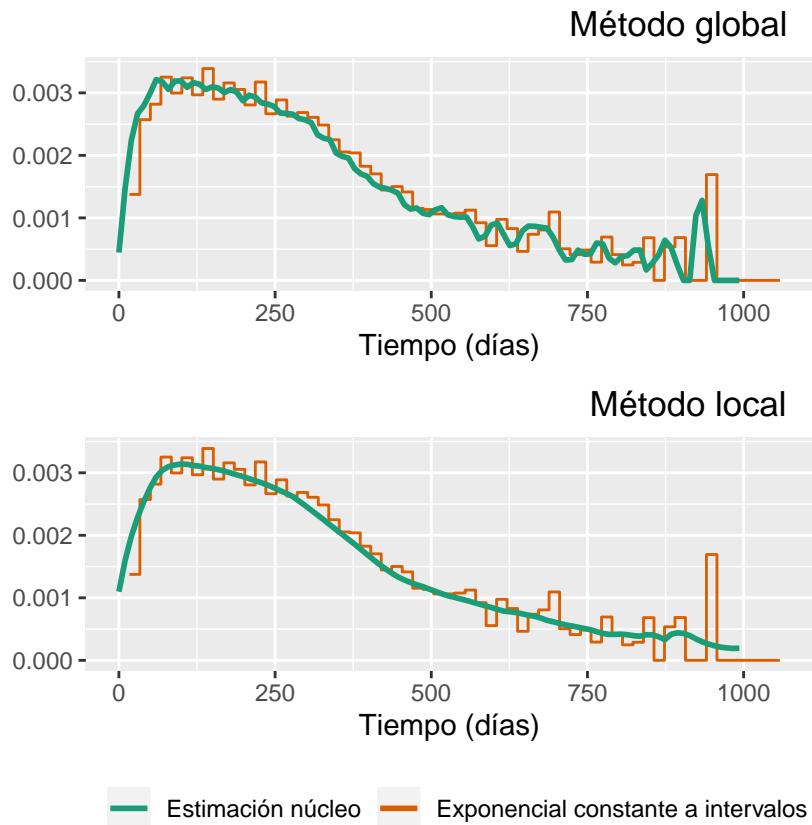


Figura 6.10: Gráficos de la estimación por núcleos de la función de riesgo de la segunda vida de las tablets (líneas color verde). En el panel superior, se utiliza un ancho de ventana global, siendo el ancho óptimo igual a 16.3 mientras que en el panel inferior se realiza la estimación con un ancho de ventana local. Para ambos casos el ancho de banda inicial es igual a 15.9 y se aplica corrección de límites en el borde izquierdo y el derecho, el núcleo utilizado es el de Epanechnikov. Las líneas en color naranja representan la estimación de la función de riesgo utilizando el modelo exponencial constante a intervalos.

Capítulo 7

Conclusiones y trabajos a futuro

A partir del análisis exploratorio realizado con los datos correspondientes a la primera y segunda vida del equipo, se pudo observar principalmente que no existen diferencias notorias entre la generación de tablets entregadas en 2014 y la generación de tablets entregadas en 2015 y que la primer vida presenta un mejor pronóstico que la segunda.

La proporción de defunciones para las 80500 tablets que tuvieron una primera vida ascendió a 65% para la generación del 2014 y a 67% para la del 2015. Para las 48652 tablets que presentaron una segunda vida, la mortalidad fue del 58% para las tablets entregadas en 2014 y del 60% para las entregadas en 2015.

Los tiempos de vida de las tablets de la primera vida tuvieron comportamientos diferentes según el año de entrega, siendo la generación del 2015 la que tuvo duraciones menores, pero también con un período de exposición menor. Más específicamente: las tablets entregadas en 2014 tuvieron una duración máxima de un poco mas de tres años (1181 días) y una duración media de un año y dos meses (425 días), en cambio la duración máxima de las entregadas en 2015 fue de dos años y medio (921 días) y la duración media de un año (356 días). Los tiempos de vida hasta las segundas defunciones fueron menores que los tiempos hasta las primeras defunciones, concentrándose la mayor cantidad de defunciones en los primeros seis meses. La du-

CAPÍTULO 7. CONCLUSIONES Y TRABAJOS A FUTURO

ración máxima de las tablets entregadas en 2014 fue de casi tres años (1046 días) y de las entregadas en 2015 fue de dos años y cuatro meses (844 días).

La rotura de la placa madre fue la principal causa de defunción para este modelo específico de tablet. Para la primera vida, este repuesto se reparó en un 58 % de los equipos entregados en 2014 y en un 72 % de los entregadas en 2015. Para la segunda vida, en un 62 % de los del 2014 y en un 70 % de los del 2015.

Al analizar la distribución de las variables sociodemográficas correspondientes a la primera vida: sexo, área geográfica y contexto sociocultural, se pudo apreciar que no hubo diferencias notorias respecto al tiempo de vida y a las causas de defunción, pero si respecto a la proporción de defunciones entre una categoría y otra de cada una de estas variables. Dejaron de funcionar el 70 % de las entregadas a los niños y el 62 % de las entregadas a niñas; el área geográfica que presentó mayor cantidad de defunciones respecto al total de tablets entregadas es el de las escuelas del interior urbano (68 %); y aquellas tablets de escuelas pertenecientes al contexto con un nivel medio de criticidad son las que presentaron mayor cantidad de defunciones (70 %) y los del contexto menos crítico son las que presentaron menor cantidad respecto al total entregado (62 %) .

Al analizar las variables que cuantificaron las entradas a servicio técnico, se destacó que del 37 % del total de las tablets de la primera vida, no se tuvo ningún registro de ellas desde que se entregaron hasta que fueron recambiadas. Considerando sólo las entradas de los equipos que presentaron una defunción, el 97 % no tuvo ninguna orden de trabajo antes de la que indicó defunción.

Por otro lado, al utilizar el método de Kaplan-Meier para estimar la función de supervivencia para las 80500 tablets que presentaron una primera vida, los resultados fueron los siguientes: 53035 tablets (66 %) presentaron una defunción, el resto fueron censuradas; la vida media fue estimada (con límite superior igual a 1181 días) en 538.6 días y la vida mediana en 366 días, entre 364 y 369 días con 95 % de confianza.

Al analizar dicha estimación gráficamente se observó como a medida que pasó el

tiempo, la supervivencia de las tablets disminuyó rápidamente, hasta los 450 días aproximadamente.

Para la estimación de la función de supervivencia de las 48652 tablets que tuvieron una segunda vida, los resultados fueron los siguientes: 28606 tablets (59 %) presentaron un segundo evento, las restantes 20046 fueron censuradas; la vida media se estimó con un límite superior igual a 1046 en 411.9 días y la vida mediana en 240 días entre 237 y 244 días con 95 % de confianza.

Al comparar ambas curvas de supervivencia, resultaron significativamente diferentes y la primera vida tuvo un mejor pronóstico que la segunda durante todo el período de análisis.

A futuro sería importante incluir y tener en cuenta el precio de los diferentes reuestos utilizados, mano de obra, etc., para obtener una mejor valoración de la conveniencia de reparar los equipos, en función no sólo de cuanto sobreviven una vez realizada la reparación, si no de los costos que esta implica, respecto a la compra de un nuevo dispositivo.

Por otro lado, para analizar como influían las variables explicativas (sexo del alumno, área geográfica y contexto sociodemográfico al que pertenece la escuela y st que indica si el equipo entró a servicio técnico antes de la defunción o no) en la supervivencia del equipo, se realizó una estimación Kaplan-Meier de la función de confiabilidad de cada categoría de estas variables por separado. A través de los test de log-rank y Peto-Peto se concluyó para todos los casos, que no hubo evidencia para aceptar la hipótesis nula de igualdad de curvas de supervivencia. En particular, se observó que las tablets de las niñas sobrevivieron mas que las de los niños a lo largo del tiempo, es la variable que presentó mayor diferencia entre sus curvas. Las tablets entregadas en escuelas del interior rural y escuelas de Montevideo tuvieron una mayor probabilidad de sobrevivir que las del interior urbano. Los quintiles más críticos (quintiles 1, 2 y 3) tuvieron peores resultados que los quintiles menos críticos (quintiles 4 y 5) cuya supervivencia fue mayor, los equipos entregados en el quintil 3 presentaron

CAPÍTULO 7. CONCLUSIONES Y TRABAJOS A FUTURO

el mayor riesgo de dejar de funcionar durante el período de análisis. Las tablets sin ninguna reparación antes de su primera defunción tuvieron una mayor supervivencia a lo largo del tiempo, para los equipos que nunca fueron a servicio técnico, la probabilidad de sobrevivir disminuyó rápidamente principalmente al inicio de la vida.

Otra forma de analizar lo planteado anteriormente fue aplicando un modelo de riesgos proporcionales de Cox. Se obtuvieron resultados similares a los de la estimación de Kaplan-Meier manteniendo la relación entre las categorías de cada variable y la probabilidad de sobrevivir a lo largo del tiempo durante el período de análisis. Sin embargo al realizar el diagnóstico y verificación de los supuestos del modelo, resultó que no se cumplía el supuesto de riesgos proporcionales, de forma global y para ninguna de las variables explicativas.

Una alternativa fue utilizar un modelo extendido de Cox, en el que se consideraron las variables explicativas como dependientes del tiempo. A futuro sería interesante profundizar en este tipo de modelos y probar con diferentes interacciones entre las variables que no cumplieron el supuesto de riesgos proporcionales y el tiempo.

Otra alternativa a futuro sería incluir en el modelo de Cox además de las variables explicativas, interacciones entre ellas.

Finalmente, se realizó la estimación de la función de riesgo. Para ello se utilizó una estimación no paramétrica por núcleos obteniendo una curva en la que el riesgo fue mayor al comienzo de la vida de las tablets, alcanzando el nivel más alto de la tasa a los 350 días aproximadamente. A medida que transcurrió el tiempo, el riesgo disminuyó, es decir que la probabilidad de que las tablets se rompieran dejando de funcionar, dado que no se rompieron hasta ese momento, fue cada vez menor. Para la estimación de la función de riesgo de la generación de tablets que presentaron una segunda vida, fue aún más notorio el crecimiento del riesgo de presentar una defunción, alcanzando el nivel más alto en los primeros 100 días aproximadamente y luego comenzando a disminuir progresivamente.

Por último, se proponen otras interrogantes e ideas con las que sería interesante trabajar en un futuro como forma de ampliar esta investigación.

- Incluir indicadores de uso del equipo que permitan valorar si usarlas mas o menos influye en la supervivencia. También contribuiría a obtener información más precisa del momento en que finaliza la vida de cada una de las tablets.
- Levantar el supuesto de que el intervalo de tiempo entre que el equipo deja de funcionar y es reparado no es sustancial y analizar con mayor profundidad dichos casos. ¿Cuánto tiempo pasan los niños y las niñas exactamente sin usar la máquina entre que se les rompió y la pueden reparar efectivamente? ¿Cómo afecta el hecho de no contar con el equipo por determinado tiempo al rendimiento de los alumnos y alumnas o a la planificación de las maestras y maestros?
- Realizar comparaciones con otros modelos de equipos entregados por el Plan Ceibal. El tipo de dispositivo ¿es un factor que cambia los resultado del análisis o no? ¿Qué equipo presenta mayor probabilidad de supervivencia?

CAPÍTULO 7. CONCLUSIONES Y TRABAJOS A FUTURO

Bibliografía

Amarante, V., and A. Dean. 2012. Dinámica del mercado laboral uruguayo. *Serie Documentos de Trabajo/FCEA-IE; DT17/12.*

Anep. Anep: Definiciones. <https://www.anep.edu.uy/observatorio/paginas/definiciones.html>.

Ceibal. 2010. Promoción de la salud y la educación en la niñez y la adolescencia. ley nº 18.640.

Ceibal. 2019. ¿Qué es el Plan Ceibal? Recuperado de <http://www.ceibal.edu.uy/es/institucional>.

Cox, D. R. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 34 (2): 187–202.

Dunkler, D., M. Ploner, M. Schemper, and G. Heinze. 2018. Weighted Cox regression using the R package coxphw. *Journal of Statistical Software* 84 (2): 1–26.

Harrington, D. P., and T. R. Fleming. 1982. A class of rank test procedures for censored survival data. *Biometrika* 69 (3): 553–566.

Hess, K., and R. Gentleman. 2014. *muhaz: hazard function estimation in survival analysis*. R package version 1.2.6.

Hofmann, H. 2000. Exploring categorical data: interactive mosaic plots. *Metrika* 51 (1): 11–26.

BIBLIOGRAFÍA

- Horová, I., J. Kolácek, and J. Zelinka. 2012. *Kernel smoothing in Matlab: theory and practice of kernel smoothing*. World scientific.
- Kassambara, A., and M. Kosinski. 2018. *survminer: drawing survival curves using 'ggplot2'*. R package version 0.4.3.
- Kim, J. S., and F. Proschan. 1991. Piecewise exponential estimator of the survivor function. *IEEE Transactions on Reliability* 40 (2): 134–139.
- Kleinbaum, D. G., and M. Klein. 2010. *Survival analysis*, Vol. 3. New York: Springer.
- Mantel, N. 1966. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep* 50: 163–170.
- Marconi, C. 2016. Supervivencia de las laptops XO: ¿las características sociodemográficas de los alumnos inciden en la igualdad de acceso? *Maestría en Demografía y Estudios de Población*.
- Moore, D. F. 2016. *Applied survival analysis using R*. Springer.
- Muller, H. G., and J. L. Wang. 1994. Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics*.
- Nelson, W. B. 2005. *Applied life data analysis*, Vol. 577. John Wiley & Sons.
- Nelson, Wayne. 1972. Theory and applications of hazard plotting for censored failure data. *Technometrics* 14 (4): 945–966.
- Peña, R. E. B. 2005. Análisis de sobrevivencia utilizando el lenguaje R. *Paipa, Boyacá, Colombia*.
- Peto, R., and J. Peto. 1972. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society: Series A (General)* 135 (2): 185–198.
- Phillips, M. J. 2003. Statistical methods for reliability data analysis. In *Handbook of reliability engineering*, 475–492. Springer.

Bibliografía

- R Core Team. R foundation for statistical computing. 2018. *R: a language and environment for statistical computing*. Vienna, Austria. <https://www.R-project.org/>.
- Therneau, T. M., and P. M. Grambsch. 2000. *Modeling survival data: extending the Cox model*. New York: Springer.
- Triunfo, P., R. Todd Jewell, et al. 2010. Mortalidad infantil en Uruguay: un análisis de supervivencia. *Cuadernos de Economía* 29 (53).
- Wickham, H. 2016. *ggplot2: elegant graphics for data analysis*. Springer.

BIBLIOGRAFÍA

Apéndice A

Apéndice estadístico

A.1. Propiedades del estimador por núcleos de la función de riesgo

Las propiedades de este estimador han sido estudiadas por Muller and Wang (1994).

- Sea $[0, S]$, $S > 0$, un intervalo para el cual $L(S) < 1$.
- $\lambda \in C^{k_0}[0, S]$, $k_0 \geq 2$, $k \leq k_0$, es decir λ es una función continua y con derivadas continuas de orden k_0 .
- $K^{(v)} \in S_{v,k}^0$, siendo

$$S_{v,k} = \left\{ \begin{array}{l} K \in Lip[-1, 1], \quad \text{soporte}(K) = [-1, 1] \\ \int_{-1}^1 x^j K(x) dx = \begin{cases} 0, & 0 \leq j < k, \quad j \neq v \\ (-1)^v v!, & j = v \\ \beta_k(K) \neq 0, & j = k. \end{cases} \end{array} \right.$$

- Sea $b = b(n)$ el ancho de banda que satisface que

$$\begin{aligned} \lim_{n \rightarrow \infty} b &= 0, & \lim_{n \rightarrow \infty} b^{2v+1} n &= \infty, \\ \lim_{n \rightarrow \infty} nb^{k+1}(\log n)^{-1} &= \infty, & \lim_{n \rightarrow \infty} nb(\log n)^{-2} &= \infty. \end{aligned}$$

APÉNDICE A. APÉNDICE ESTADÍSTICO

Luego, el *Error Cuadrático Medio* en el punto x puede ser expresado como

$$MSE\{\hat{\lambda}^{(v)}(x, b)\} = \underbrace{\left[b^{k-v} \lambda^{(k)}(x) \left\{ \frac{(-1)^k \beta_k}{k!} + o(1) \right\} \right]^2}_{\text{sesgo}^2\{\hat{\lambda}^{(v)}(x, b)\}} + \underbrace{\frac{1}{nb^{2v+1}} \left\{ \frac{\lambda(x)V(K^{(v)})}{\bar{L}(x)} + o(1) \right\}}_{\text{var}\{\hat{\lambda}^{(v)}(x, b)\}}.$$

La calidad global de esta estimación se puede describir por medio del *Error Cuadrático Medio Integrado* ($MISE\{\hat{\lambda}^{(v)}(\cdot, b)\}$).

El *Error Cuadrático Medio Integrado Asintótico*, $AMISE\{\hat{\lambda}^{(v)}(\cdot, b)\}$ toma la forma:

$$AMISE\{\hat{\lambda}^{(v)}(\cdot, b)\} = b^{2(k-v)} \beta_k^2 D_k + \frac{V(K^{(v)})\mathcal{L}}{nb^{2v+1}},$$

donde

$$\begin{aligned} \beta_k &= \beta_k(K^{(v)}) = \int_{-1}^1 x^k K^{(v)}(x) dx, \\ D_k &= \int_0^S \left(\frac{\lambda^{(k)}(x)}{k!} \right)^2 dx, \\ \mathcal{L} &= \int_0^S \frac{\lambda(x)}{\bar{L}(x)} dx. \end{aligned}$$

Entonces, el ancho de banda asintóticamente óptimo $b_{opt,v,k}$ que minimiza el $AMISE\{\hat{\lambda}^{(v)}(\cdot, b)\}$ sobre el conjunto B_n de anchos de banda aceptables con respecto a b viene dado por

$$b_{opt,v,k}^{2k+1} = \frac{\mathcal{L}(2v+1)}{2n(k-v)D_k} \gamma_{v,k}^{2k+1}.$$

A.2. Regla de Freedman-Diaconis (F-D)

El histograma es el estimador de la función de densidad más sencillo. Depende de dos parámetros: el ancho de ventana b y el origen x_0 . Para optimizar la ventana b de manera que el sesgo y la varianza sean relativamente pequeñas, existen varios métodos. Uno de ellos es la regla de Freedman-Diaconis donde se define el ancho de ventana óptimo, b^* , como:

$$b^* = 2 * IQ * n^{-1/3},$$

A.2. Regla de Freedman-Diaconis (F-D)

siendo IQ el rango intercuartílico de la muestra. Éste método fue propuesto como una modificación para datos no normales, menos sensible a los datos atípicos.

APÉNDICE A. APÉNDICE ESTADÍSTICO

Apéndice B

Apéndice de resultados

B.1. Tiempo de vida hasta la primera defunción según variables sociodemográficas

En la Figura B.1 se muestran los gráficos de densidad de la duración de las tablets que presentan una primera defunción según las variables sociodemográficas: sexo, área y contexto.

Al observar el gráfico correspondientes al área geográfica, se observa que la categoría que considera las escuelas rurales, tiene una duración diferente al resto, siendo el primer modo que se distingue en todas las categorías, un poco “más bajo” para este subconjunto. En general, no hay diferencias entre los tiempos de vida hasta la primera defunción correspondientes a cada categoría de cada variable.

APÉNDICE B. APÉNDICE DE RESULTADOS

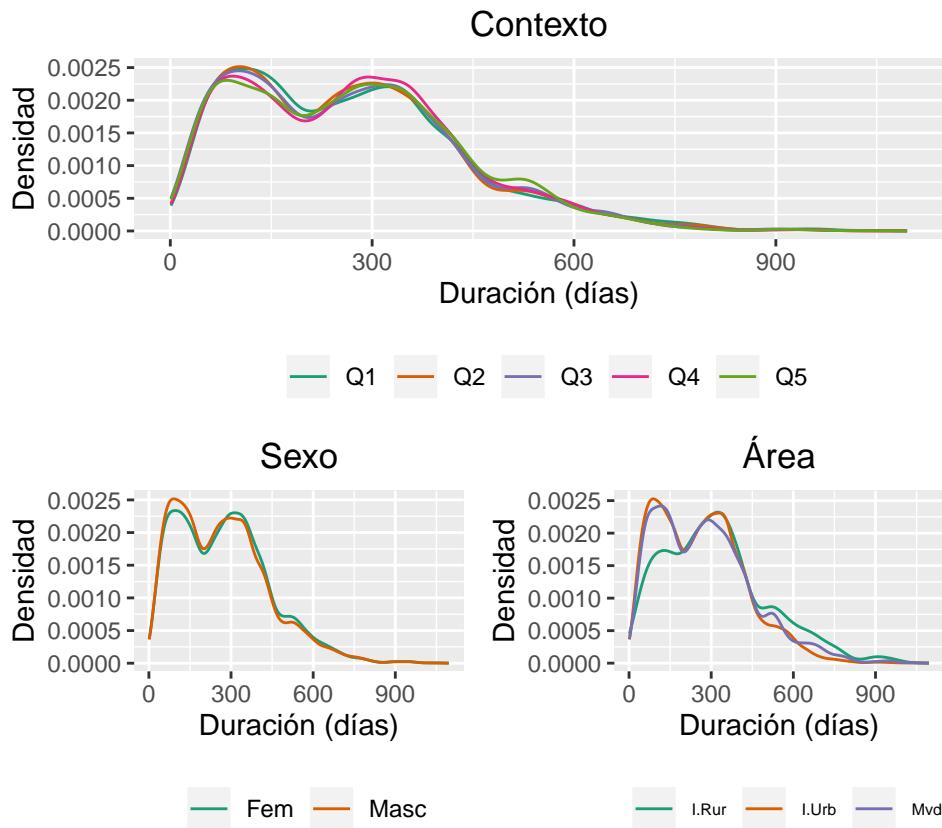


Figura B.1: Gráficos de densidad con el tiempo de vida de las 53035 tablets que registraron un evento en su primera vida, según las variables sociodemográficas consideradas en el análisis: sexo (panel inferior izquierdo), área geográfica (panel inferior derecho) y contexto sociocultural (panel superior).

B.2. Estimación Kaplan-Meier de la primera vida de las tablets

Se presenta en la Tabla B.1, la salida de R al aplicar la función `survfit()` para estimar a través del método de Kaplan-Meier la función de supervivencia de la primera vida de las tablets, sólo que en vez de mostrar los resultados para cada tiempo en los que ocurre al menos un evento, se muestra para intervalos de 90 días. En dicha tabla se muestra el tiempo, la cantidad de tablets en el conjunto de riesgo, la cantidad de eventos, la probabilidad de sobrevivir, el error estándar y los intervalos de confianza al 95 %.

Tabla B.1: Estimación Kaplan-Meier de la función de supervivencia de la primera vida de las tablets para intervalos de tiempo de 90 días.

Tiempo	Riesgo	Eventos	Superv.	Err. std.	IC inferior	IC superior
0	80500	0	1.00	0.0000	1.00	1.00
90	71973	8196	0.90	0.0011	0.90	0.90
180	60633	10896	0.76	0.0015	0.76	0.76
270	50884	8840	0.65	0.0017	0.65	0.65
360	38021	10741	0.51	0.0018	0.51	0.51
450	29692	7415	0.41	0.0018	0.41	0.41
540	25909	3293	0.36	0.0017	0.36	0.37
630	17081	2056	0.33	0.0017	0.33	0.33
720	8963	1008	0.31	0.0018	0.30	0.31
810	3929	380	0.29	0.0019	0.28	0.29
900	2485	91	0.28	0.0021	0.28	0.28
990	1157	98	0.27	0.0024	0.26	0.27
1080	225	19	0.26	0.0034	0.25	0.26
1170	11	2	0.25	0.0040	0.25	0.26

APÉNDICE B. APÉNDICE DE RESULTADOS

A continuación, como forma de resumir la estimación mencionada anteriormente, se presentan en la Tabla B.2 la cantidad de eventos, la media calculada con el límite superior igual al máximo de la duración, la mediana y los límites del intervalo de confianza.

Tabla B.2: Resumen de la estimación Kaplan-Meier de la función de supervivencia de la primera vida de las tablets.

n	Eventos	*Media rest.	*se(media rest.)	Mediana	0.95LCL	095UCL	
80500	53035	538.61		1.67	366.01	363.99	369.16

*Media restringida con límite superior = 1181

B.3. Test log-rank y Peto-Peto para comparación de las curvas de supervivencia de la primera y segunda vida

Se presentan los resultados de la salida de R de la función `survdiff()` del paquete `survival` que comprueba si hay una diferencia entre dos o más curvas de supervivencia utilizando la familia de pruebas G-rho. Se presenta el resultado con la comparación de la curva de supervivencia de la primera vida y la de la segunda vida; en primer lugar el test log-rank (Tabla B.3) y luego el test Peto-Peto (Tabla B.4).

Tabla B.3: Test log-rank que compara las funciones de supervivencia de la primera y segunda vida.

N	Observado	Esperado	$(O-E)^2/E$	$(O-E)^2/Var$
80500	53035	59782	761	2933
48652	28606	21859	2083	2933

Chisq= 2933 con 1 grado de libertad, p= <2e-16

B.4. Estimación Kaplan-Meier de la primera vida considerando las variables explicativas

Tabla B.4: Test Peto-Peto que compara las funciones de supervivencia de la primera y segunda vida.

N	Observado	Esperado	$(O-E)^2/E$	$(O-E)^2/Var$
80500	32626	38586	921	4284
48652	21759	15799	2248	4284

Chisq= 4284 con 1 grado de libertad, p= <2e-16

B.4. Estimación Kaplan-Meier de la primera vida considerando las variables explicativas

Se presentan en la Tabla B.5 la cantidad de eventos, la media calculada con el límite igual al máximo de la duración, la mediana y los límites del intervalo de confianza correspondientes a las estimaciones a través del método de Kaplan-Meier de las funciones de supervivencia para cada categoría de las variables sociodemográficas utilizadas en el análisis: sexo, área geográfica, contexto sociodemográfico y entrada a servicio técnico, considerando el grupo primera vida, obtenidas con un `summary()` del objeto de supervivencia obtenido con la función `survfit()` del paquete `survival`.

APÉNDICE B. APÉNDICE DE RESULTADOS

Tabla B.5: Estimación Kaplan-Meier de las funciones de supervivencia de cada categoría de las variables sexo, área, contexto y ST.

	n	Eventos	Media rest.	se(media)	Mediana	0.95LCL	095UCL
Femenino	38491	23788	580.37	2.50	398.71	394.23	403.08
Masculino	42009	29247	500.31	2.21	345.01	342.27	348.07
I.Rural	48641	33172	515.59	2.11	350.79	348.01	353.10
I.Urbano	4035	2711	566.33	6.72	406.18	393.00	416.17
Montevideo	27824	17152	571.58	2.88	400.98	396.16	406.25
Quintil 1	17158	10569	587.30	3.72	402.08	396.16	408.99
Quintil 2	17767	11754	525.96	3.47	360.21	356.79	365.21
Quintil 3	15981	10844	517.44	3.66	353.11	348.87	358.85
Quintil 4	13798	9633	500.53	3.92	347.10	342.13	353.00
Quintil 5	15796	10235	552.97	3.80	374.71	368.85	379.96
ST=0	77927	51514	534.41	1.69	362.95	360.21	365.00
ST=1	2573	1521	640.09	8.76	507.95	473.25	531.82

B.5. Test log-rank y Peto-Peto para comparación de las curvas de supervivencia correspondientes a las categorías de las variables auxiliares de la primera vida

B.5. Test log-rank y Peto-Peto para comparación de las curvas de supervivencia correspondientes a las categorías de las variables auxiliares de la primera vida

En las Tablas B.6 y B.7 se presentan los resultados de las pruebas log-rank y Peto-Peto respectivamente, obtenidos con la función `survdiff()` en los que se testeó por separado para cada variable, si las funciones de supervivencia de cada categoría son iguales.

Tabla B.6: Test log-rank que compara las funciones de supervivencia de cada categoría de las variables sexo, área, contexto y ST.

	N	Observado	Esperado	$(O-E)^2/E$	$(O-E)^2/Var$	chisq
Sexo=Femenino	38491	23788	26702	318	641	641
Sexo=Masculino	42009	29247	26333	322	641	641
Area=I.Urbano	48641	33172	30987	154	371	372
Area=I.Rural	4035	2711	2971	23	24	372
Area=Montevideo	27824	17152	19077	194	304	372
Contexto=Quintil 1	17158	10569	11997	170	220	320
Contexto=Quintil 2	17767	11754	11449	8	10	320
Contexto=Quintil 3	15981	10844	10250	34	43	320
Contexto=Quintil 4	13798	9633	8727	94	113	320
Contexto=Quintil 5	15796	10235	10612	13	17	320
ST=0	77927	51514	50952	6	158	158
ST=1	2573	1521	2083	151	158	158

APÉNDICE B. APÉNDICE DE RESULTADOS

Tabla B.7: Test Peto-Peto que compara las funciones de supervivencia de cada categoría de las variables sexo, área, contexto y ST.

	N	Observado	Esperado	$(O-E)^2/E$	$(O-E)^2/Var$	chisq
Sexo=Femenino	38491	15638	17521	202	559	559
Sexo=Masculino	42009	19545	17662	201	559	559
Area=I.Urbano	48641	22225	20709	111	374	377
Area=I.Rural	4035	1665	1925	35	52	377
Area=Montevideo	27824	11293	12549	126	271	377
Contexto=Quintil 1	17158	6975	7847	97	174	266
Contexto=Quintil 2	17767	7808	7638	4	7	266
Contexto=Quintil 3	15981	7225	6837	22	38	266
Contexto=Quintil 4	13798	6427	5842	59	97	266
Contexto=Quintil 5	15796	6748	7019	10	18	266
ST=0	77927	34338	33836	7	276	276
ST=1	2573	844	1346	187	276	276

B.6. Estimación de Cox aplicada a diferentes modelos

Se muestran en las tablas a continuación (Tablas B.8, B.9 y B.10) los resultados de la estimación realizada con el modelo de Cox sobre el *Modelo A*, *Modelo B* y *Modelo C* obtenidas con la función `coxph()` del paquete `survival`.

Tabla B.8: Resultado del ajuste del modelo de Cox al modelo A.

	coef	exp(coef)	se(coef)	z	p
SexoMasculino	0.22	1.25	0.01	25.27	0.00

Likelihood ratio test=642.2 on 1 df, p=<2e-16

n= 80500, number of events= 53035

Tabla B.9: Resultado del ajuste del modelo de Cox al modelo B.

	coef	exp(coef)	se(coef)	z	p
ContextoQuintil 1	0.15	1.17	0.01	11.43	0.00
ContextoQuintil 2	0.18	1.20	0.01	13.41	0.00
ContextoQuintil 3	0.23	1.25	0.01	16.02	0.00
ContextoQuintil 4	0.09	1.09	0.01	6.54	0.00

Likelihood ratio test=323.9 on 4 df, p=<2e-16

n= 80500, number of events= 53035

APÉNDICE B. APÉNDICE DE RESULTADOS

Tabla B.10: Resultado del ajuste del modelo de Cox al modelo C.

	coef	exp(coef)	se(coef)	z	p
SexoMasculino	0.22	1.24	0.01	25.02	0.00
ContextoQuintil 1	0.15	1.16	0.01	10.93	0.00
ContextoQuintil 2	0.18	1.20	0.01	13.04	0.00
ContextoQuintil 3	0.22	1.25	0.01	15.82	0.00
ContextoQuintil 4	0.09	1.09	0.01	6.34	0.00

Likelihood ratio test=953.3 on 5 df, p=<2e-16

n= 80500, number of events= 53035

B.7. Función AIC para selección de modelos

Se presentan en la Tabla B.11 los resultados de la salida de R de la función `AIC()` para el *Modelo A*, *Modelo B*, *Modelo C* y para el *Modelo completo*.

Tabla B.11: Comparación de modelos con la función `AIC()`.

	df	AIC
mA	1	1140857.59
mB	4	1141181.88
mC	5	1140554.49
mcom	8	1139972.99

B.8. Verificación del supuesto de riesgos proporcionales del modelo de Cox aplicado al modelo completo

En la Figura B.2¹ se presentan los resultados de la verificación del supuesto de riesgos proporcionales del modelo de regresión de Cox, aplicado al modelo que incluye todas las variables explicativas (modelo completo), a través de una visualización gráfica en la que para cada coeficiente, además de los residuos de Schoenfeld escalados, se ajusta una curva por el sistema de alisado de splines, junto con dos líneas adicionales a ± 2 el error estándar, proporcionando una estimación del coeficiente $\beta(t)$. Si se cumple la hipótesis de riesgos proporcionales, los residuos deberían agruparse de forma aleatoria a ambos lados del valor 0 del eje Y , y la curva ajustada (es decir $\beta(t)$) debería ser próxima a una línea recta.

¹Se utiliza la función `ggcoxzph()` del paquete `survminer` para obtener dicha visualización.

APÉNDICE B. APÉNDICE DE RESULTADOS

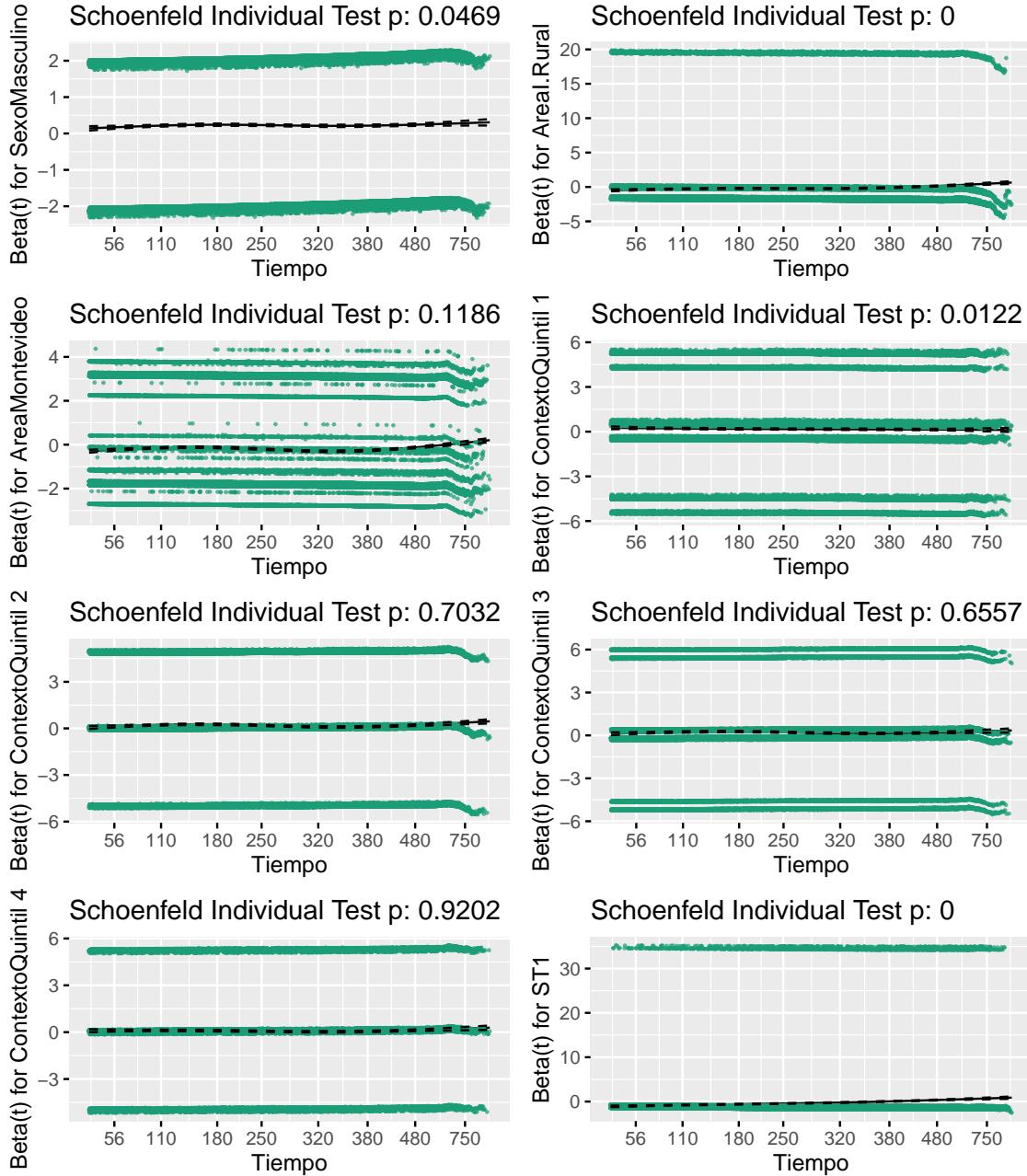


Figura B.2: Gráfico de los residuos de Schoenfeld escalados contra el tiempo transformado para cada covariable en un modelo de Cox ajustado a los datos de la primera vida de las tablets con las variables explicativas sexo, área, contexto y ST. En cada panel, la línea continua es una spline de suavizado ajustada y las líneas discontinuas representan una banda de ± 2 el error estándar alrededor del ajuste.

B.9. Comparación de las estimaciones de la función de riesgo con diferentes tipos de núcleos

B.9. Comparación de las estimaciones de la función de riesgo con diferentes tipos de núcleos

Se presentan en la Figura B.3 los resultados gráficos de la estimación núcleo de la función de riesgo de las tablets de la primera vida obtenida con la función `muhaz()` del paquete `muhaz` considerando las funciones núcleos de Epanechnikov, rectangular, bicoadrática y tricuadrática. No se observan diferencias notorias entre ellas.

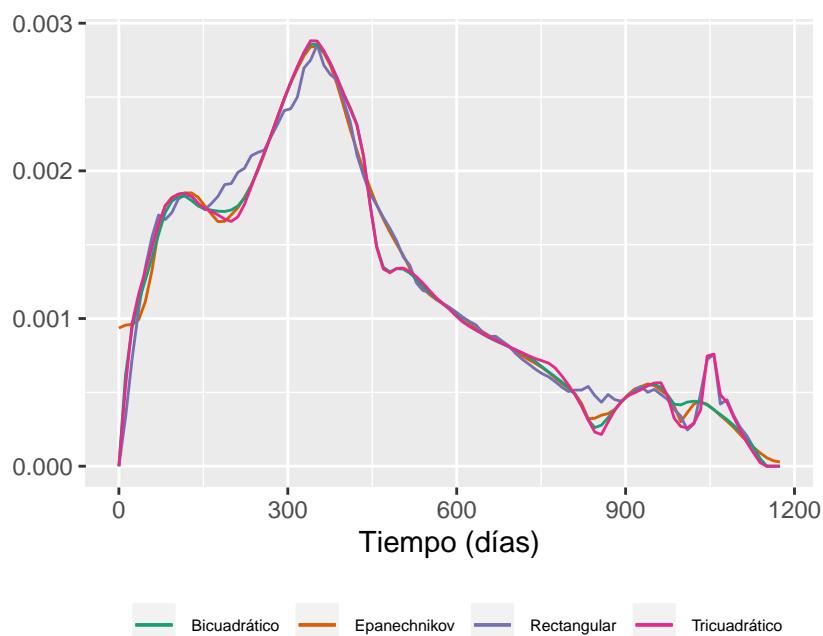


Figura B.3: Gráficos de la función de riesgo de las 80500 tablets que tienen una primera vida, estimadas a través del método por núcleos utilizando diferentes funciones núcleos. La curva color naranja es estimada con un núcleo de Epanechnikov, la de color violeta utiliza un núcleo rectangular, la de color verde un núcleo bicoadrático y la rosada uno tricuadrático.