Análisis de supervivencia de fallas del corazón

Alfie González, Santiago Battezzati & Elena Villalobos

Maestría en Ciencia de datos
Instituto Tecnológico Autónomo de México

27 de Mayo, 2021.

1. Introducción

El término médico adecuado de una falla del corazón, es disfunción sistólica ventricular que se refiere a cuando el ventrículo izquierdo del corazón, muestra una disminución en su funcionalidad. Dicha disminución limitará la cantidad de sangre que bombea el corazón a todo el cuerpo, lo que puede puede producir insuficiencia cardiaca congestiva, infarto al miocardio, entre otras enfermedades vasculares (AEEC, s.f.).

Las fallas en el corazón son la causa de muerte más comun en hombres, mujeres y personas de distintos grupos étnicos en países como Estados Unidos. Además, en México, las enfermedades del corazón han aumentado un 90 % desde 1970 y son la segunda causa de muerte, representando el 17 % de todas las muertes en 2008 (Morales et al, 2014).

Existen muchas causas de fallas del corazón, como la presión alta, la diabetes, el tabaquismo, la alimentación, etc. A pesar de esto, parece que no hay un concenso en las causas de esta enfermedad, por lo que hay varios esfuerzos para determinar las razones principales de esta enfermedad.

1.1. Objetivo

El objetivo del presente trabajo es estudiar variables asociadas a fallas del corazón por lo que buscaremos estimar la mortalidad de dicha enfermedad, relacionada a otros factores riesgo.

2. Base de datos

Este conjunto de datos contiene los registros médicos de 299 pacientes que tuvieron una falla en el corazón. Todos los pacientes tuvieron una disfunción ventricular sistólica izquierda y pertenecen a alguna de las clases 3 o 4, de la clasificación de insuficiencia cardiaca según la NYHA (New York Heart Association). Este estudio se llevó a cabo en Pakistán y tuvo un periodo de seguimiento de 4

a 285 días, con un promedio de 130 días. Cada persona fue diagnosticada por un cirujano médico. Cada paciente tiene las siguientes 13 características clínicas[1].

- `edad`: Edad del paciente (años).
- `sexo`: Mujer u hombre (binaria).
- `anemia`: Disminución de glóbulos rojos o hemoglobina (booleana).
- `diabetes`: Si el paciente tiene diabetes (booleana).
- `fumar`: Tabaquismo, si el paciente fuma o no (booleana).
- `presion alta`: Si el paciente tiene hipertensión (booleana).
- `salida sangre`: Fracción de ejección, porcentaje de sangre que sale del corazón en cada contracción (porcentaje).
- `enzima cpk`: Nivel de la enzima CPK en sangre (mcg/L).
- `plaquetas`: Plaquetas en la sangre (kiloplaquetas/ml).
- `nivel creanitina`: Nivel de creatinina sérica en sangre (mg/dl).
- `nivel sodio`: Nivel de sodio sérico en sanre (mEq/L).
- `tiempo`: Periodo de seguimiento (días).
- `deceso`: Si el paciente falleció durante el período de seguimiento (booleana).

A continuación, se presenta un análisis exploratorio para observar el comportamiento general de las variables.

2.1. Análisis exploratorio

En la Figura 1, se observa de lado izquierdo un gráfico de barras para la variable de sexo, que nos muestra que tenemos casi el doble de hombres que mujeres. En el histograma de lado derecho, podemos las edades de todos los participantes, el rango de edad va de 40 a casi 100; observamos que tenemos más conteos de edades en los 50, también se observa una concentración de conteos en los 40, 45, 50, 60, 65 y 70.

La Figura 2, presenta también gráficos de barras de todas las variables booleanas que tenemos en el estudio que son anemia, diabetes, hipertensión y fumar. En todas, el 1 significa presencia y el cero ausencia. En la mayoría observamos más presencia de personas que se podrían considerar sanas en estas características pues no tiene ni anemia, ni diabetes, ni hipertensión. Así mismo, la variable que más diferencia tiene es la de tabaquismo, pues casi el doble de los sujetos no fuman, en comparación con los que sí fuman.

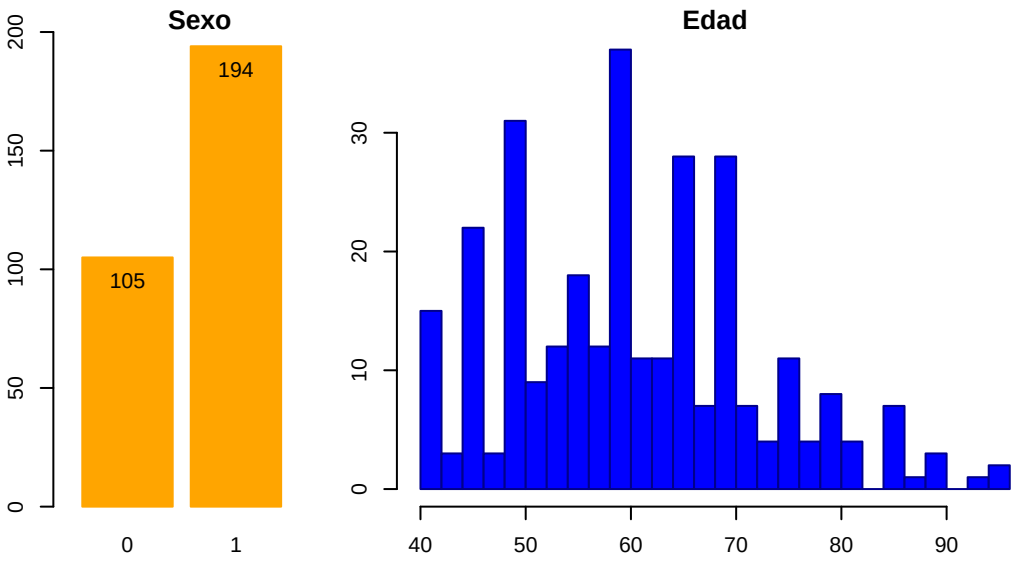

Figura 1: Gráficos de sexo y edad: De lado izquierdo es un gráfico de barras para sexo, done 1 significa hombre y 0 mujer. De lado derecho, un histograma con las edades de los participantes.

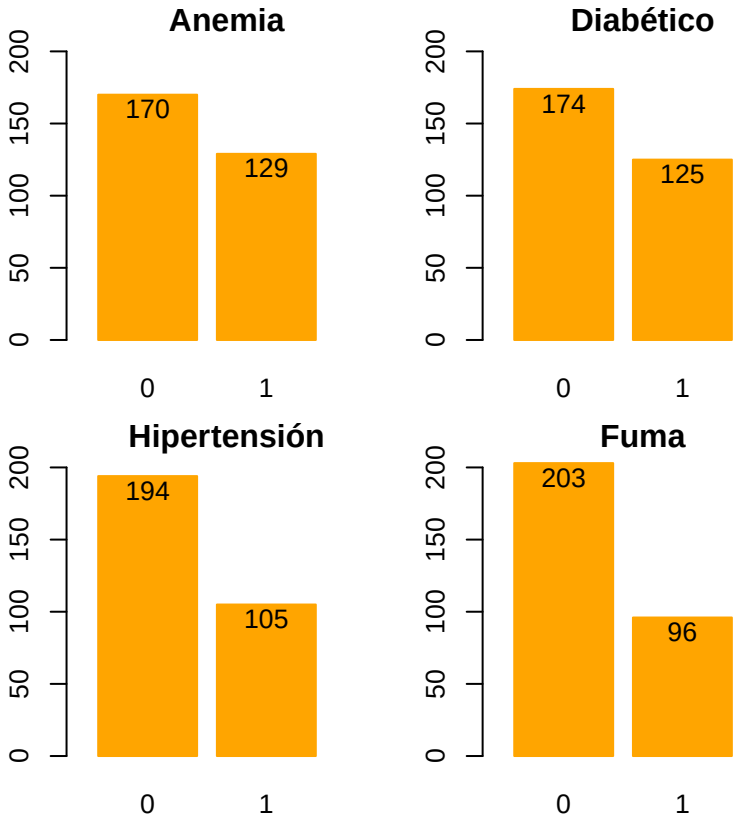

Figura 2: Graficos de barras para las variables de anemia, diabetes, hipertensión y tabaquismo.

La Figura 3 es un gráfico que contiene los scatterplots de las variables continuas que tenemos, así mismo su contraparte muestra las correlaciones. Existen dos colores, el azu, hace referencia a los hombres y el rosa a las mujeres. Además, las variables de salida sangre y plaquetas están en logaritmo para poder apreciar mejor la relación con otras variables. El objetivo de mostrar este gráfico es para apreciar que no existen correlaciones claras entre las variables continuas, que se confirma con el estadístico de correlación, esto sucede para ambos sexos.

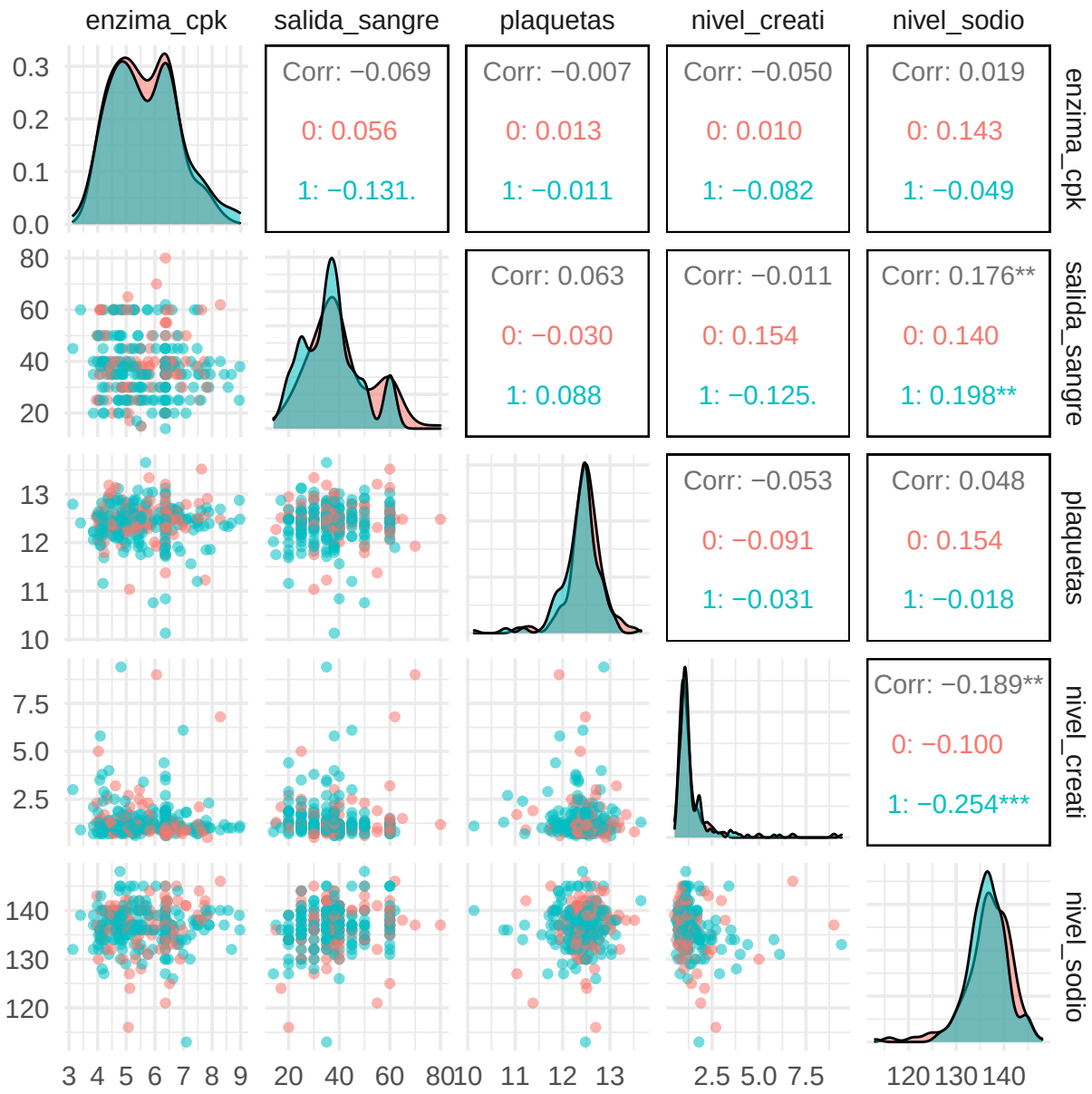

Figura 3: Diagrama de pares para las variables continuas del estudio. En la contraparte inferior son los scatterplots y en la parte superior las correlaciones. En la diagonal se muestran las densidades por variable.

[1]Base de datos obtenidas del Machine Learning Repository: https://archive.ics.uci.edu/

2.2. Datos censurados

La variable evento muerte nos indica si el paciente falleció o no durante el periodo de seguimiento. En la presente base de datos, el 32 % de los pacientes fallecieron durante el periodo de seguimiento del estudio. En la Figura 4 podemos observar una línea que indica el periodo de seguimiento para cada paciente y si es un dato censurado o no, indicado por color. Lo importante de este gráfico es observar la mayoría de los pacientes que fallecieron durante el estudio, lo hicieron en el primer tercio del periodo, que equivale a 90 días aproximadamente. También, existe un conjunto de varios decesos presentados en el periodo de casi 180 días. Por último, parece ser que los pacientes que tuvieron un periodo más largo de seguimiento fueron los que, afortunadamente, no fallecieron.

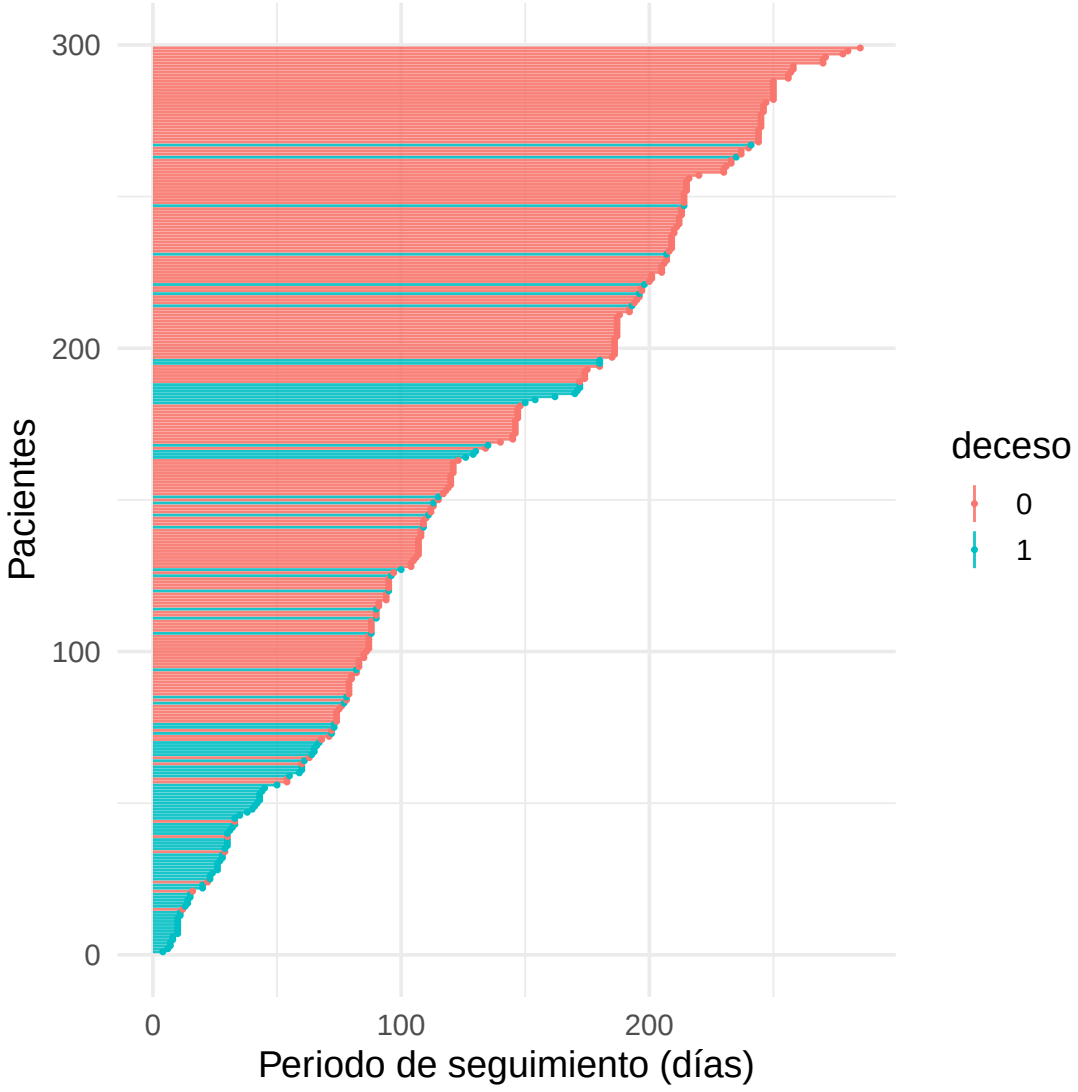

Figura 4: En el eje horizontal tenemos el tiempo de duración del estudio y en el eje vertical cada uno de los pacientes. Los pacientes están ordenados de acuerdo a los días del periodo de seguimiento. El color azul indica si fue un deceso y el rosa lo contrario.

3. Modelado e implementación

3.1. Verificación de supuestos

Se ajustaron diferentes familias de modelos paramétricos (weibull, lognormal y loglogística) para aproximar la supervivencia, resultando el modelo Lognormal con una media de 5.91 y una desviación de 1.91 el que presentó una mejor log-verosimilitud.

Por otra parte, se realizaron gráficas de diagnóstico (Figura 5), en las que también se observa que el modelo log-normal es el que mejor parece adaptarse a los datos.

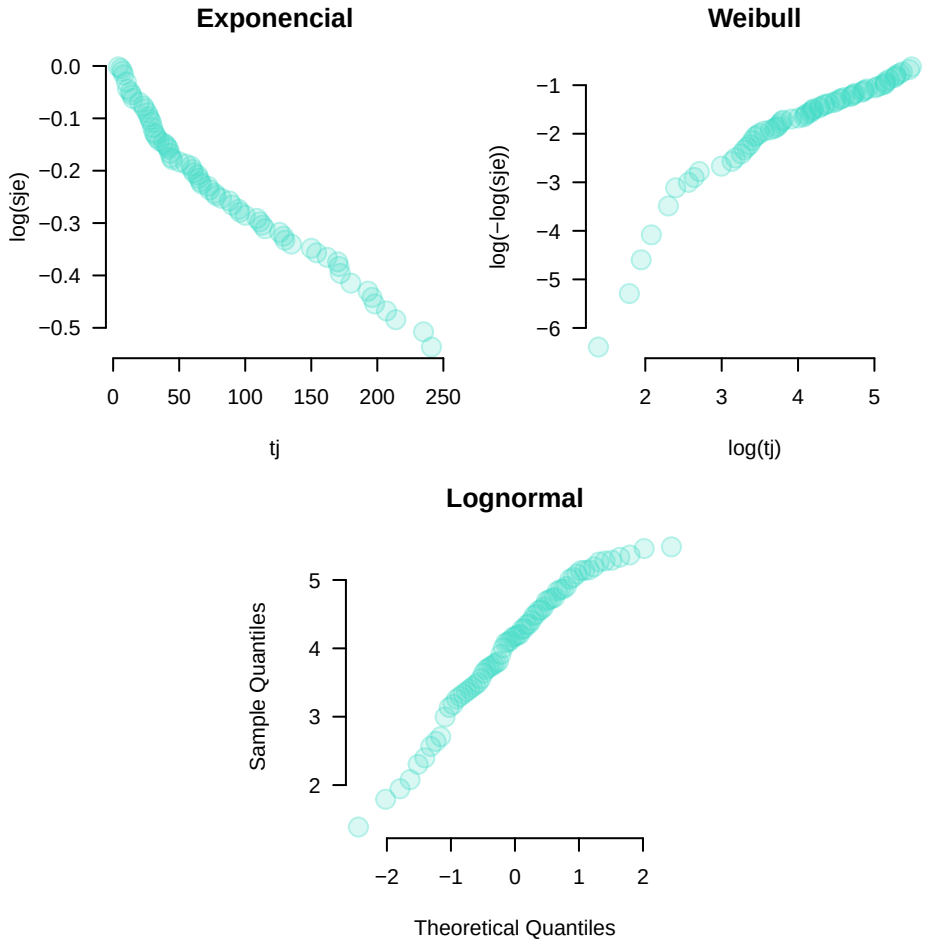

Figura 5: Gráficos de verificación de modelos Exponencial, Weibull y Lognormal.

La siguiente Figura 6 se pueden observar las gráficas de supervivencia paramétrica (a partir del modelo lognormal antes mencionado) y no paramétrica (a partir del Kaplan Mayer). Se observa una superposición entre ambas, lo que da cuenta de la similitud entre ambos modelos.

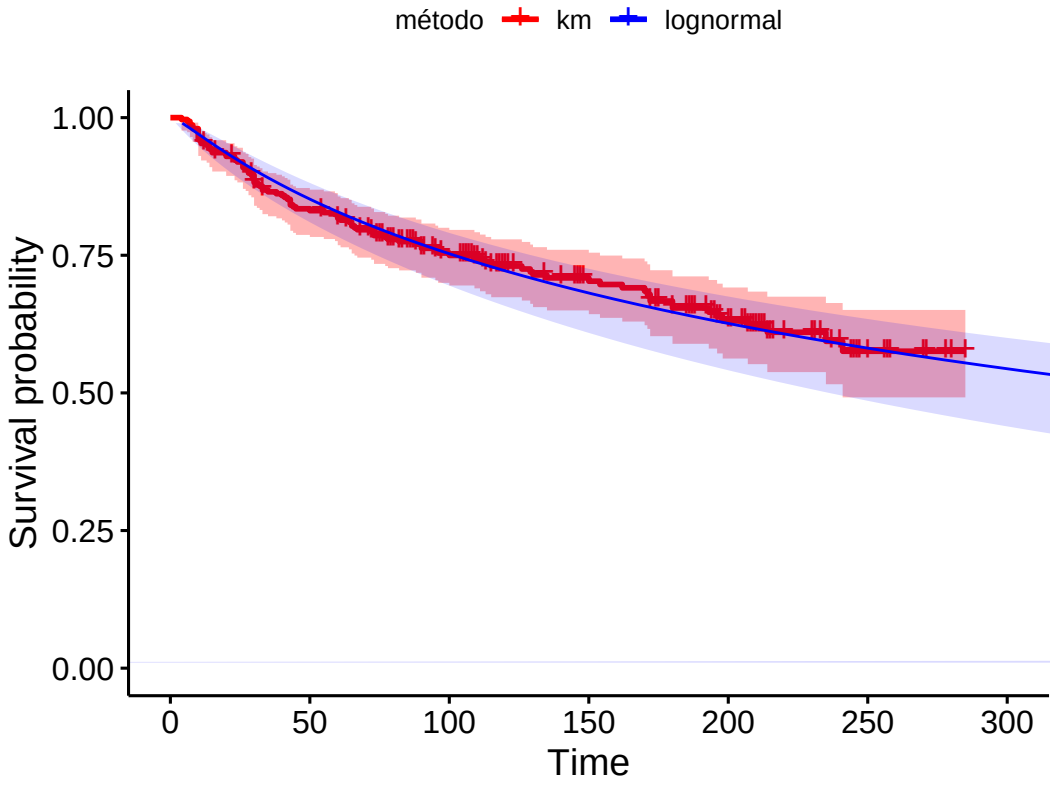

Figura 6: Función de Kaplan Meier y función de supervivencia, supuesto Lognormal.

3.2. Kaplan Meier

Para el presente análisis se utilizó el estimador Kaplan Meier para determinar la supervivencia de las variables categóricas que tenemos en la muestra. Por ejemplo, en la Figura 7 obtuvimos el estimador para todas las variables binarias que tenemos, como lo son diabetes, sexo, fumar, anemia y presión alta. Asimismo, se creó una variable que categoriza el nivel de creatinina en alto o bajo. Esta discretización se realizó basándonos en los umbrales mencionados por Wannamethee et. Al (1997), que consideraban que la concentración de 16 micromol/L (micromoles por litro de sangre) de creatinina, generaba un riesgo de ataque al corazón.

En dicha figura observamos que las supervivencias de diabetes y fumar, no son muy diferenten entre su presencia o ausencia. En cuanto a las variables de anemia y presión alta, parece que la ausencia de estas hace tener una supervivencia más alta, sin embargo, los intervalos de confianza se traslapan para ambas categorías. Por otro lado, las supervivencia que más se separan son las del nivel de creatinina. una supervivencia muy distinta para grupos que padecen alguna enfermedad (como la diabetes o la anemia) que tradicionalmente suelen estar asociadas a los ataques al corazón. Lo mismo sucede con los grupos de fumadores, en comparación con los no fumadores.

De igual manera, la variable ejection fraction se discretizó en tres grupos, siguiendo a Ahmad et al. (2017). En ambos casos, las tablas Kaplan Meyer muestran un grupo de riesgo con una

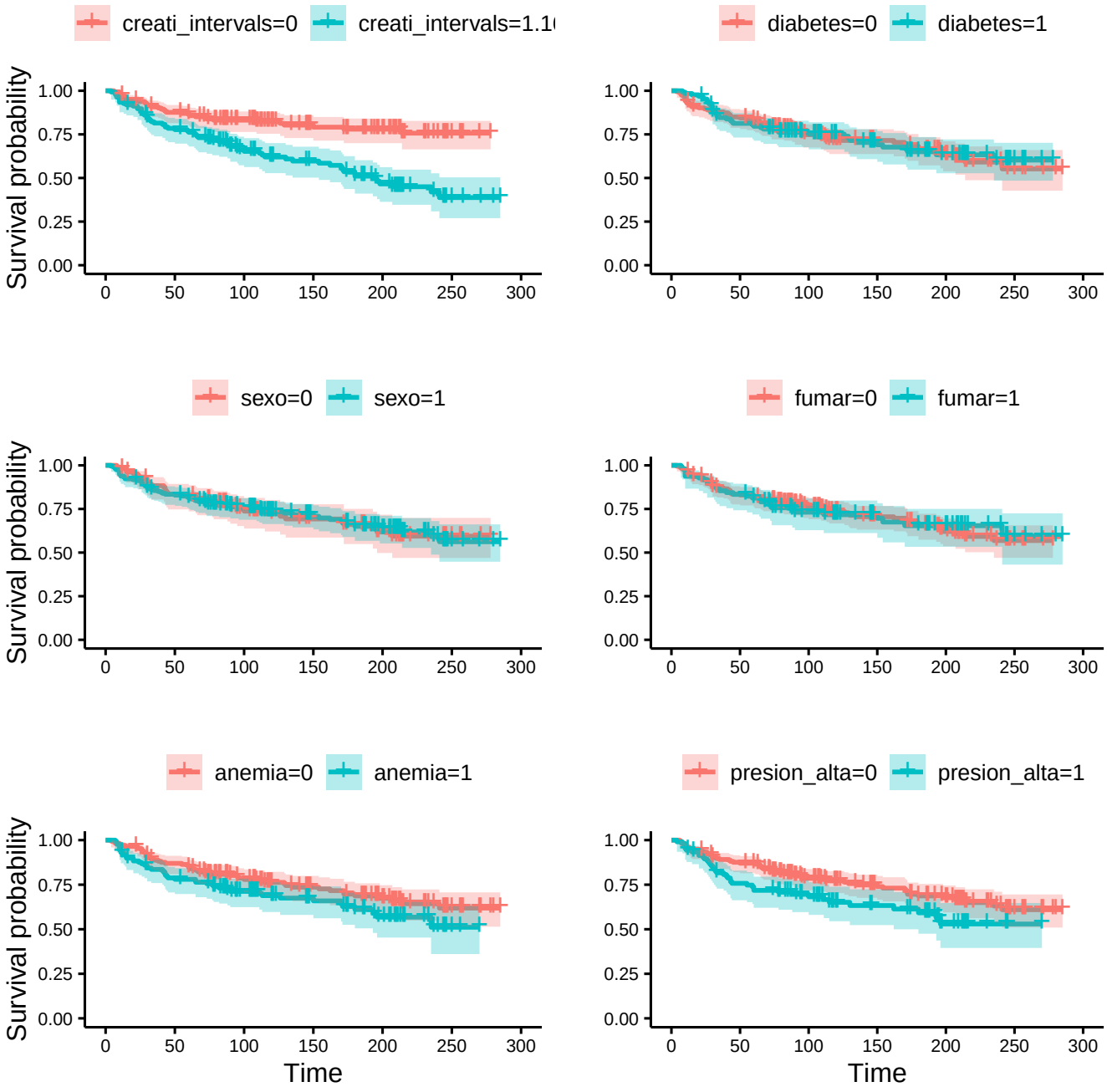

Figura 7: Estimador Kaplan Meier para las variables dicotómicas.

supervivencia considerablemente menor, lo que sugiere que estas variables podrían tener un efecto explicativo relevante.

describan con detalle el modelo, con todas sus especificaciones, que usarán para resolver sus objetivos.

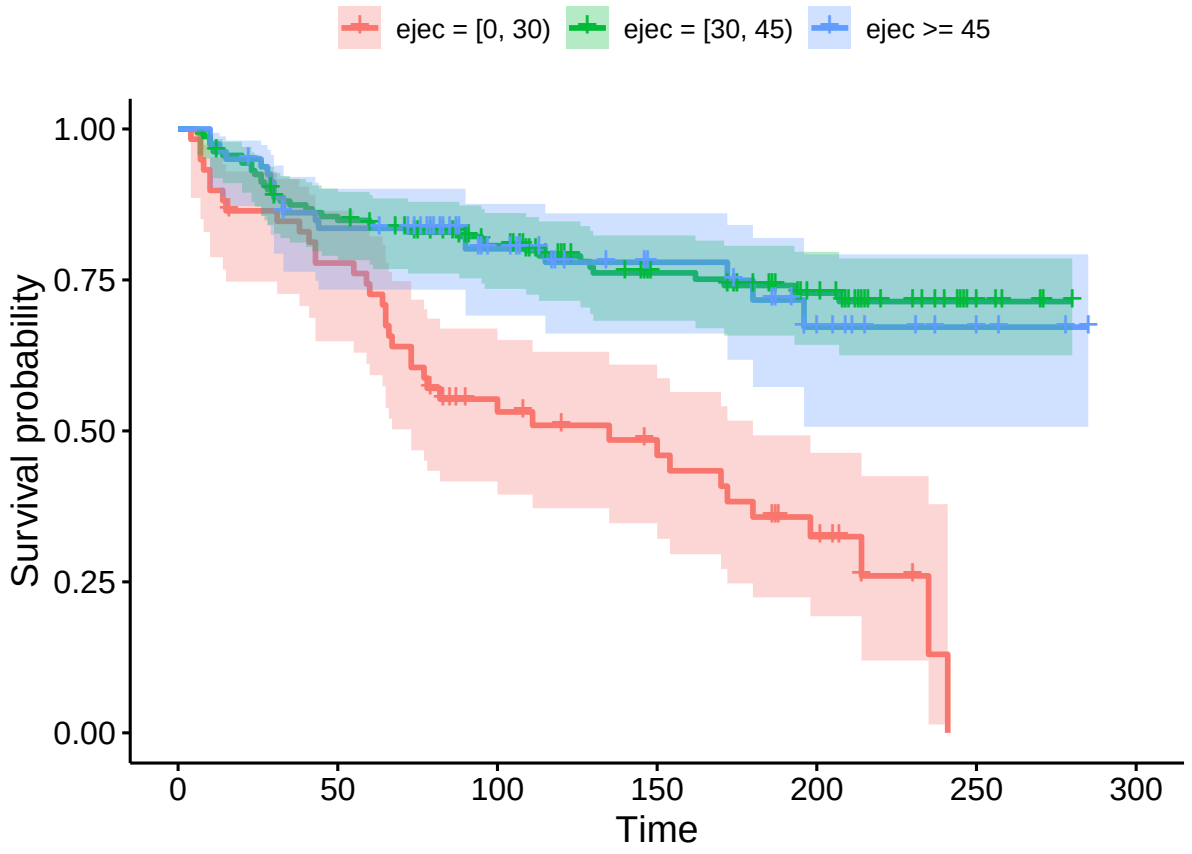

Figura 8: Estimador Kaplan Meier para la variable categórica de salida de sangre.

4. Resultados

Interpretación de resultados: presenten un resumen de sus estimadores (puntuales y por intervalo) e interpreten en el contexto del problema. Planteen pruebas de hipótesis y tomen decisiones. Hagan uso de sus resultados para responder a los objetivos planteados e incluyan predicciones.

Lo que presentaremos a continuación, serán los resultados de los mejores modelos con las variables que resultaron ser signficativas. Realizamos una comprobación de supuestos, por lo que para el modelo de vida acelerada el modelo evaluado fue el de lognormal.

4.1. Vida acelerada

Para este modelo se observa que los coeficientes de edad, la enzima cpk, presión alta y el nivel de creatinina, tienen un impacto negativo en la supervivencia. Mientras que la única variable que tiene un efecto positivo en la supervivencia fue salida de sangre. Sin embargo, los únicos coeficientes significativos fueron edad, salida de sangre, y nivel de creatinina.

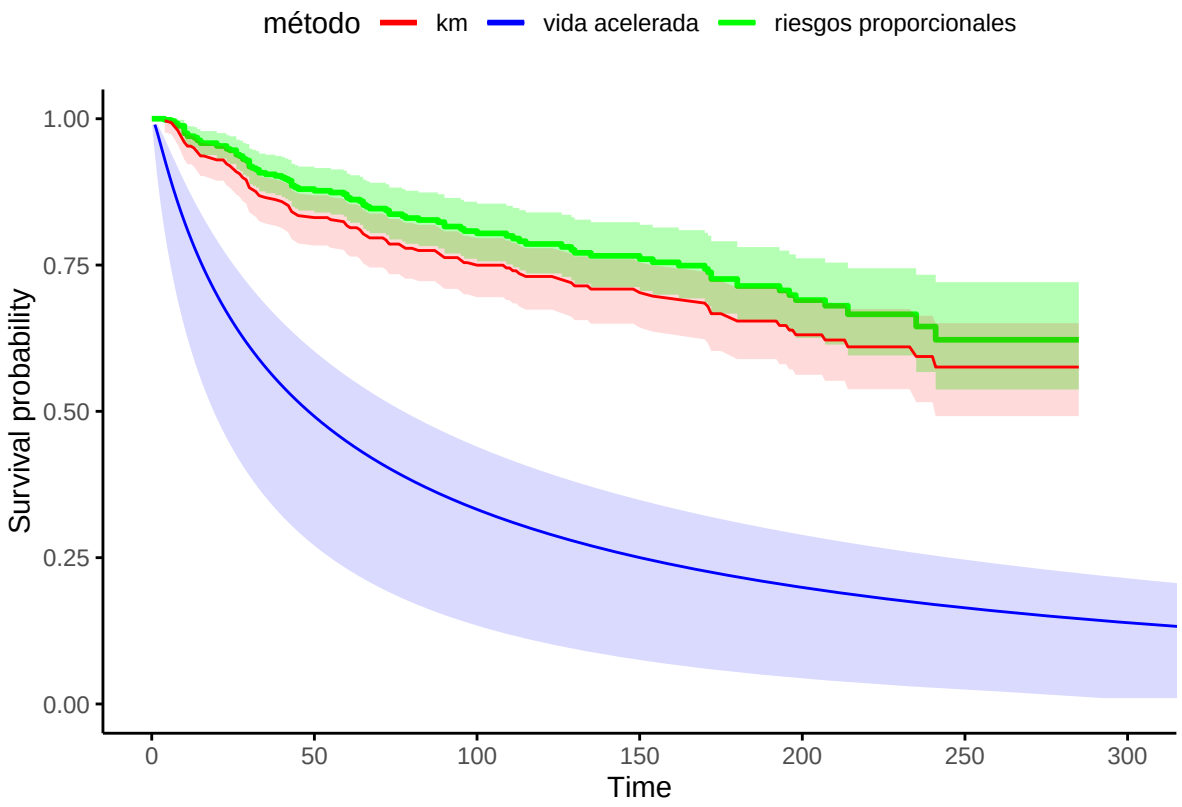

Figura 9: En el eje horizontal tenemos el tiempo de duración del estudio y en el eje vertical cada uno de los pacientes. Los pacientes están ordenados de acuerdo a los días del periodo de seguimiento. El color azul indica si fue un deceso y el rosa lo contrario.

```
##
## Call:
## survreg(formula = t ~ age + enzima_cpk + salida_sangre + presion_alta +
##     nivel_creati, data = corazones, dist = "lognormal")
##                   Value Std. Error     z       p
## (Intercept)    7.932465   0.763538 10.39 < 2e-16
## age           -0.048353   0.010689 -4.52 6.1e-06
## enzima_cpk    -0.000172   0.000115 -1.50   0.134
## salida_sangre  0.048181   0.011446  4.21 2.6e-05
## presion_alta  -0.499714   0.258672 -1.93   0.053
## nivel_creati  -0.418039   0.105802 -3.95 7.8e-05
## Log(scale)     0.521346   0.080294  6.49 8.4e-11
##
## Scale= 1.68
##
```

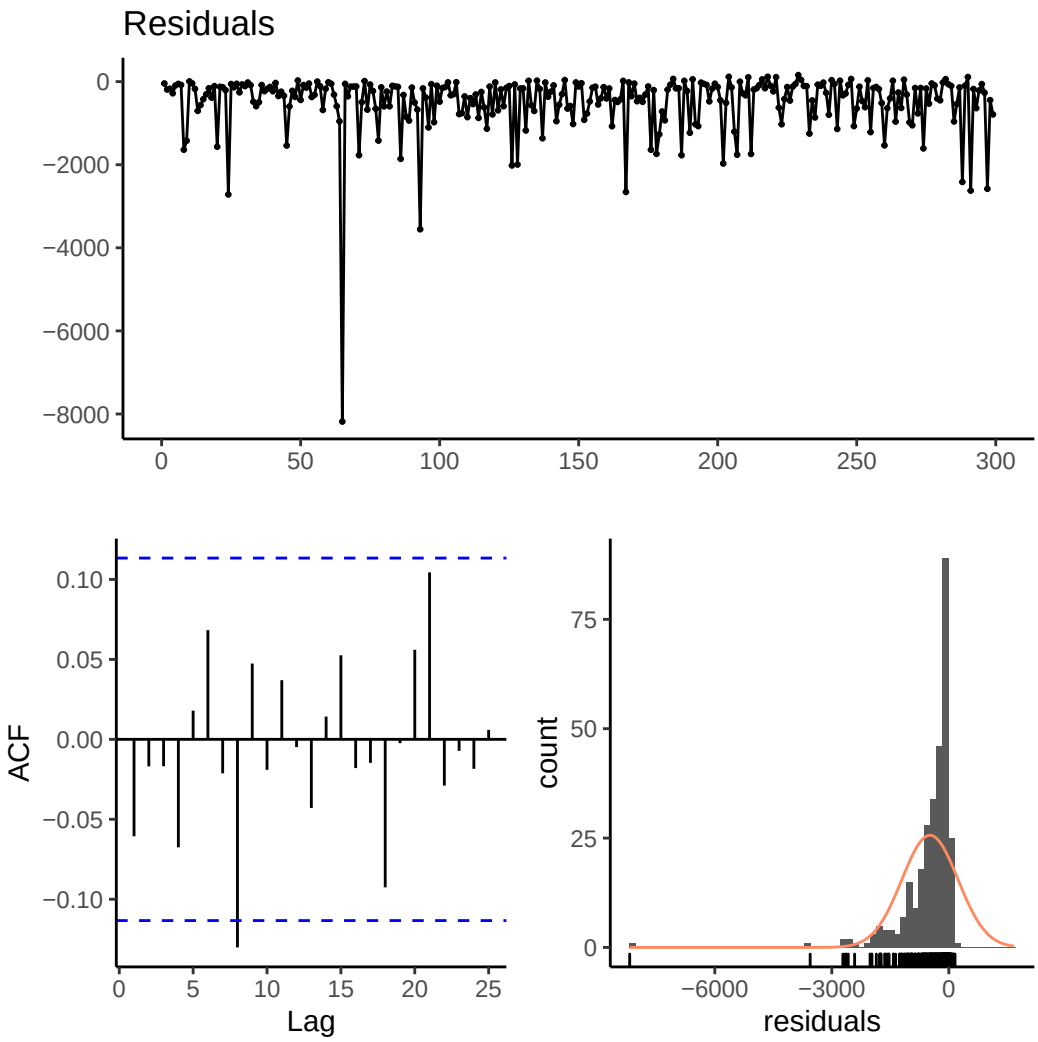

Figura 10: En el eje horizontal tenemos el tiempo de duración del estudio y en el eje vertical cada uno de los pacientes. Los pacientes están ordenados de acuerdo a los días del periodo de seguimiento. El color azul indica si fue un deceso y el rosa lo contrario.

```
## Log Normal distribution
## Loglik(model)= -635.6   Loglik(intercept only)= -666.3
##  Chisq= 61.3 on 5 degrees of freedom, p= 6.5e-12
## Number of Newton-Raphson Iterations: 4
## n= 299
```

4.2. Riesgos proporcionales

Para el modelo de riesgos proporcionales, los coeficientes tienen un efecto

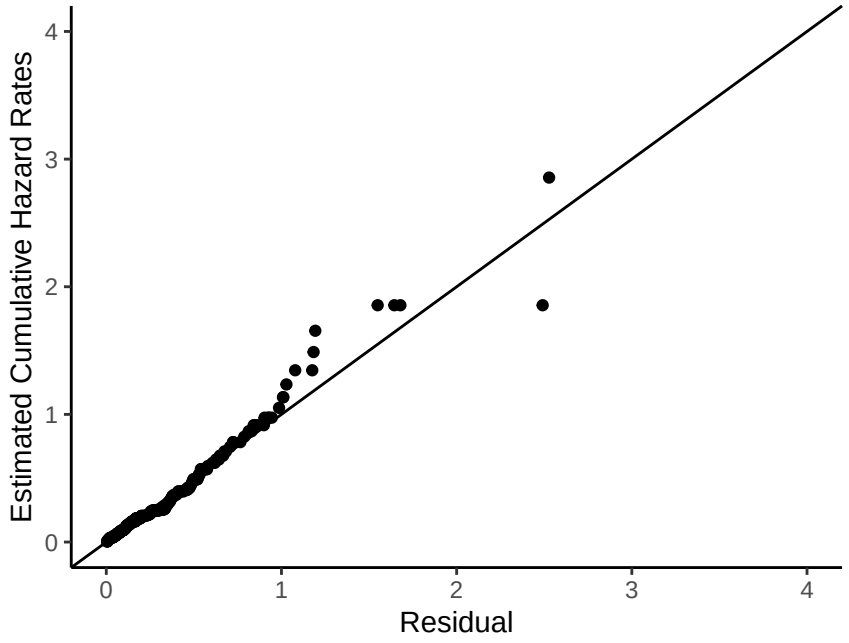

Figura 11: En el eje horizontal tenemos el tiempo de duración del estudio y en el eje vertical cada uno de los pacientes. Los pacientes están ordenados de acuerdo a los días del periodo de seguimiento. El color azul indica si fue un deceso y el rosa lo contrario.

```
## Call:
## coxph(formula = t ~ age + enzima_cpk + salida_sangre + presion_alta +
##     nivel_creati, data = corazones)
##
##   n= 299, number of events= 96
##
##                     coef  exp(coef)   se(coef)      z Pr(>|z|)
## age            4.434e-02  1.045e+00  8.912e-03  4.975 6.51e-07 ***
## enzima_cpk     1.634e-04  1.000e+00  9.709e-05  1.683   0.0924 .
## salida_sangre -5.023e-02  9.510e-01  1.004e-02 -5.006 5.57e-07 ***
## presion_alta   4.954e-01  1.641e+00  2.120e-01  2.337   0.0195 *
## nivel_creati   3.564e-01  1.428e+00  6.646e-02  5.363 8.20e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
## age               1.045     0.9566    1.0272    1.0638
## enzima_cpk        1.000     0.9998    1.0000    1.0004
## salida_sangre     0.951     1.0515    0.9325    0.9699
## presion_alta      1.641     0.6094    1.0831    2.4865
```

```
## nivel_creati      1.428     0.7002    1.2538    1.6269
##
## Concordance= 0.734  (se = 0.028 )
## Likelihood ratio test= 73.63  on 5 df,   p=2e-14
## Wald test            = 80.53  on 5 df,   p=6e-16
## Score (logrank) test = 80  on 5 df,   p=8e-16
```

5. Discusión

Referencias

Incluyan una lista de las fuentes que consultaron para hacer su trabajo, desde páginas de internet, libros, revistas o apuntes de clase.

Asociación Española de Enfermería en Cardiología (s.f.). [www.enfermeriaencardiologia.com]

Chicco D, Jurman G (2020), "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone". BMC Medical Informatics and Decision Making 20, 16.

Morales LS, Flores YN, Leng M, Sportiche N, GallegosCarrillo K, Salmerón J. (2014). Risk factors for cardiovascular disease among Mexican-American adults in the United States and Mexico: a comparative study. Salud Publica Mex 56:197–205.

Tanvir Ahmad, Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab, and Muhammad Ali Raza (2017), "Survival analysis of heart failure patients: a case study". PLoS ONE 12(7), 0181001.

Wannamethee S G, Shaper A G, Perry I J (1997), Serum creatinine concentration and risk of cardiovascular disease: a possible marker for increased risk of stroke, Stroke, Mar;28(3):557-63. doi: 10.1161/01.str.28.3.557.

Apéndice

Apéndice. Incluyan si quieren, todo el código utilizado. Por favor no incluyen código dentro de ninguna de las secciones anteriores. NOTA: Las gráfica que consideren útiles las pueden incluir en cualquiera de las secciones de la i-iv con comentarios para que el lector vea lo que ustedes quieren que vean. Las gráficas que no sean indispensables las pueden mandar al apéndice.

4) Preparen una presentación de 15 minutos más o menos, el formato es libre. Todos los integrantes tienen que hablar y la calificación de la presentación será individual, mientras que la calificación del trabajo será por equipo. Se penalizará a aquellos equipos que se tarden más del tiempo asignado originalmente en su presentación.