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 común 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).

Las fallas en el corazón están asociadas con factores 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 supervivencia de dicha enfermedad relacionada a otros factores riesgo.

2. Base de datos

Este conjunto de datos contiene los registros médicos de 299 (MLR, 2021) 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.

- `edad`: Edad del paciente (años).
- `sexo`: Mujer u hombre (binaria).
- `anemia`: Disminución de glóbulos rojos o hemoglobina (binaria).
- `diabetes`: Si el paciente tiene diabetes (binaria).
- `fumar`: Tabaquismo, si el paciente fuma o no (binaria).
- `presion alta`: Si el paciente tiene hipertensión (binaria).
- `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 sangre (mEq/L).
- `tiempo`: Periodo de seguimiento (días).
- `deceso`: Si el paciente falleció durante el período de seguimiento (binaria).

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 observar las edades de todos los participantes, el rango de edad va de 40 a casi 100 años; observamos que tenemos más conteos de edades en los 50, también se observa una concentración en los conteos de las edades de 40, 45, 50, 60, 65 y 70.

La Figura 2, presenta también gráficos de barras de todas las variables binarias que tenemos en el estudio que son anemia, diabetes, hipertensión y fumar. En todas, el uno significa presencia y el cero ausencia. En todas las variables mencionadas observamos más presencia de personas que se podrían considerar *sanas*, pues en todas prevalece la ausencia de estos factores de riesgo. Las diferencias más grandes se observan en las variables de hipertensión y de si fumma o no, pues casi el doble de los pacientes tenían ausencia de estos factores. En cuanto a las variables de anemia y diabetes, las diferencias entre la presencia o ausencia, son de unas cuantas decenas.

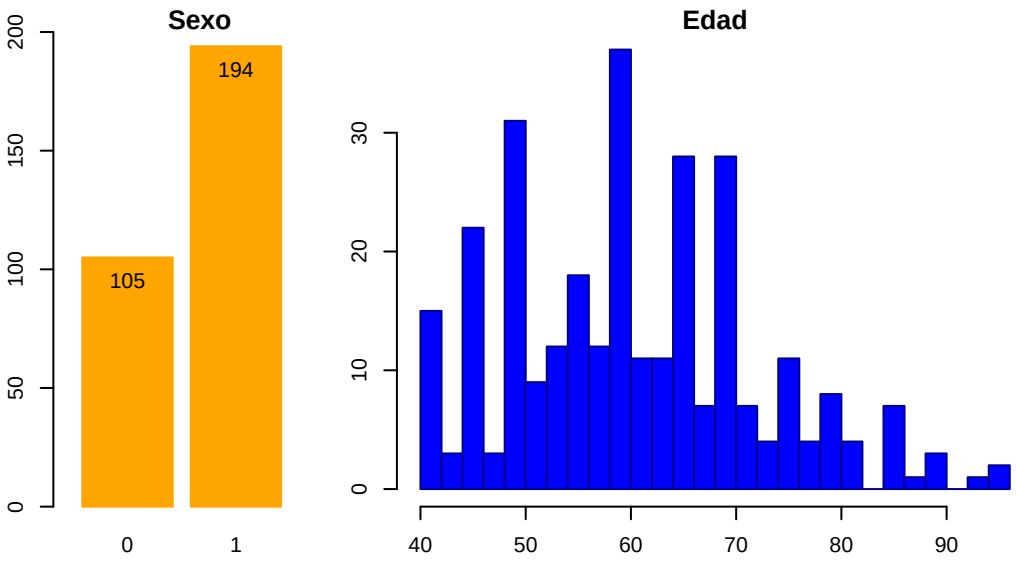

Figura 1: Gráficos de sexo y edad: De lado izquierdo es un gráfico de barras para sexo, donde 1 significa hombre y 0 mujer. De lado derecho, un histograma con las edades de los participantes.

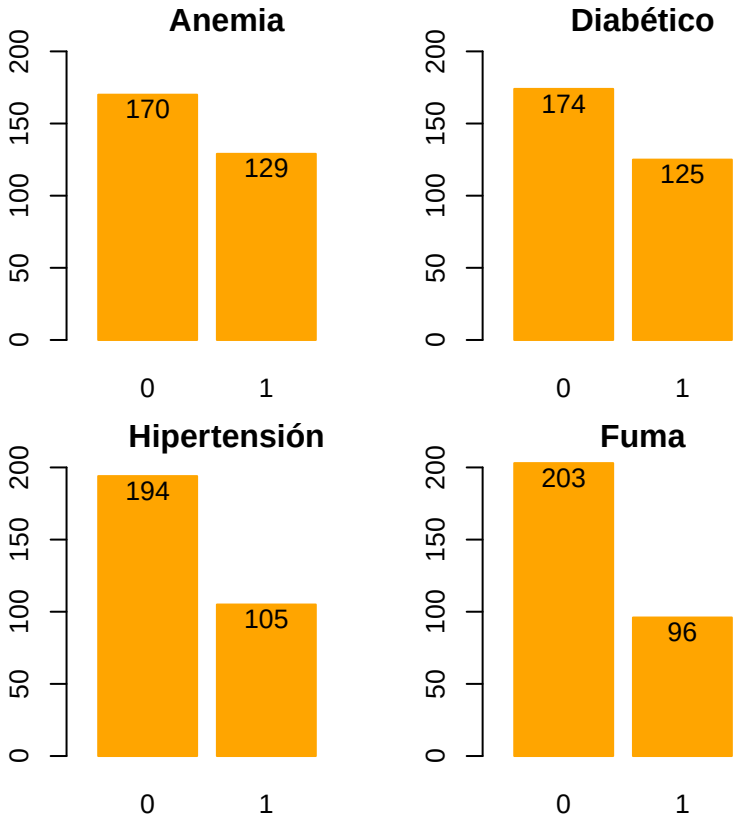

Figura 2: Gráficos de barras para las variables de anemia, diabetes, hipertensión y tabaquismo.

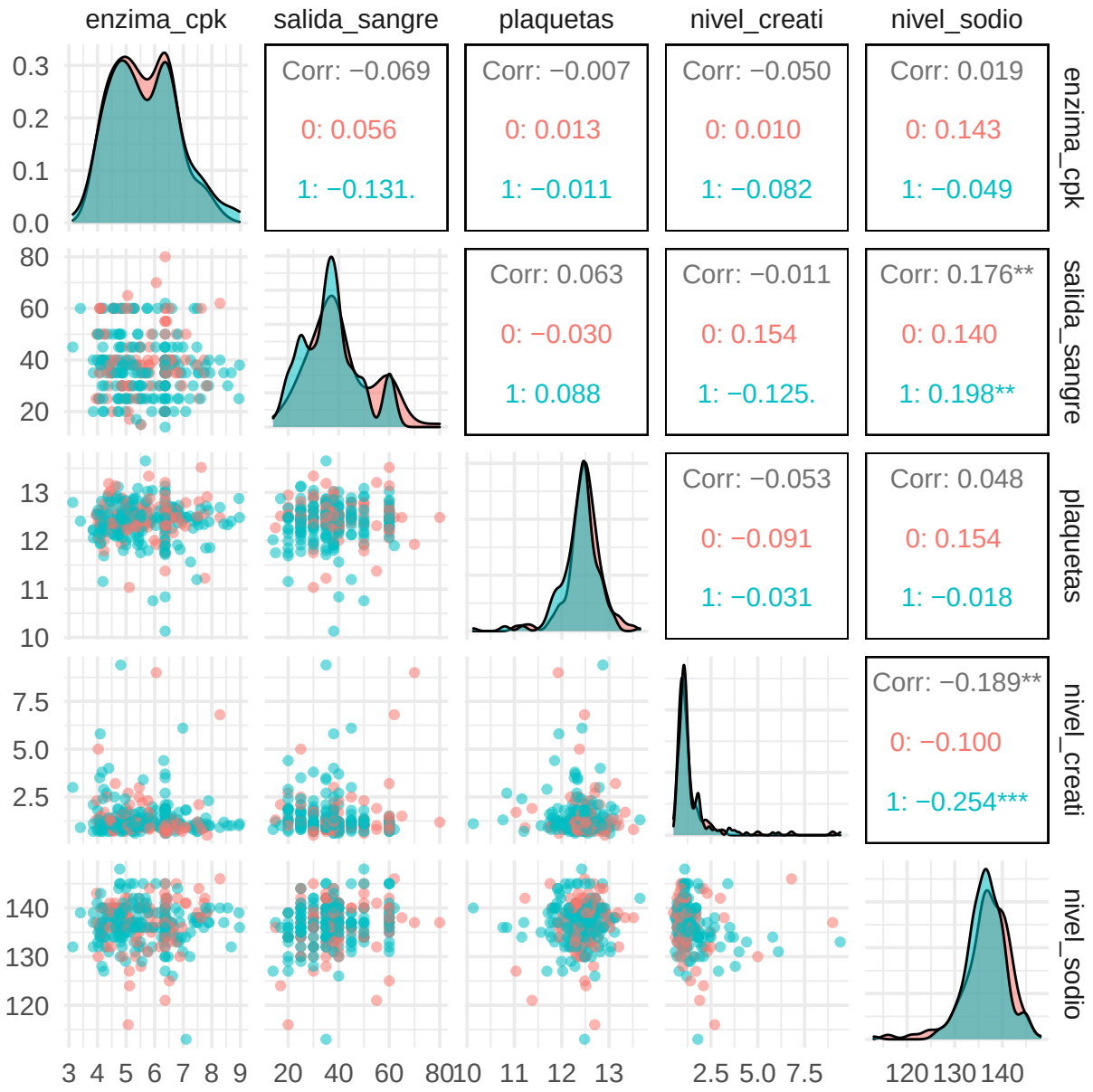

Figura 3: Diagrama de pares para las variables continuas del estudio. En la contraparte inferior son los scatterplots y en la parte superior las correlaciones. En la diagonal se muestran las densidades por variable.

La Figura 3 es un gráfico de pares que presenta información de las variables continuas de nuestra base de datos. El color azul corresponde a los hombres y el rosa a las mujeres. En la parte inferior tenemos los gráficos de dispersión entre las variables de enzima cpk, salida de sangre, plaquetas, nivel de creatinina y nivel de sodio. Asimismo, su contraparte muestra las correlaciones de las variables ya mencionadas, distinguidas también por sexo. Para este gráfico, colocamos las variables de salida sangre y plaquetas en logaritmo para poder apreciar de mejor manera alguna posible

correlación con las otras variables.

En la Figura 3 se puede apreciar que no existen fuertes correlaciones entre las variables continuas, que se confirma con el estadístico de correlación, esto sucede para ambos sexos. Podría ser que la variable de salida de sangre y plaquetas tienen una correlación positiva pero realmente es muy baja. Los otros gráficos de dispersión parecen más un cúmulo de puntos sin correlación. En cuanto a la densidades individuales de cada variable, plaquetas y nivel de sodio, parecen tener un comportamiento un poco similar a una distribución normal. Las distribución de salida de sangre y enzima cpk, parece asemejarse a una distribución bimodal. Por último, el nivel de creatinina, parece tener muchos valores atípicos. Estas densidades individuales tienen un comportamiento similar entre hombres y mujeres.

2.2. Datos censurados

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 se presentó el deceso, indicado por color, cero es que no se observó el evento y uno que sí. En este gráficos los pacientes están ordenados de acuerdo al número de días en el periodo de seguimiento. Se podría decir algunos de estos datos están *censurados por derecha* pues el tiempo observado del estudio es menor al tiempo real en que los partipantes presentaron el evento de interés, que en este caso, es el lamentable deceso.

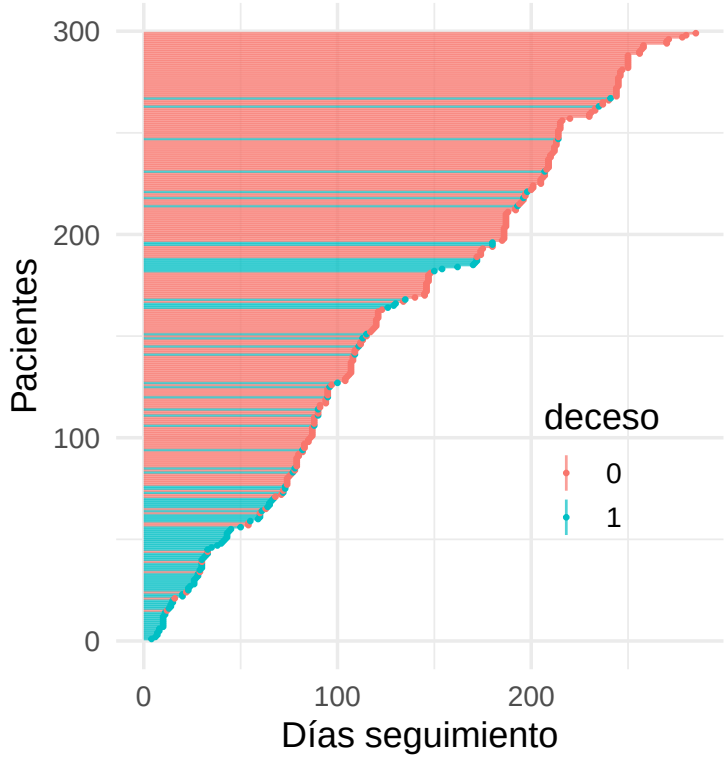

Figura 4: En el eje horizontal tenemos el periodo de seguimiento en días, y en el eje vertical cada uno de los pacientes. Cada línea es la duración de cada participante en el estudio. Están ordenados de acuerdo a número días. El color azul indica si fue un deceso y el rosa los datos censurados.

En este gráfico se puede observar que la mayoría de los pacientes que fallecieron durante el

estudio, lo hicieron en un periodo de alrededor de 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 durante el estudio.

3. Modelado e implementación

En el presente estudio se implementaron tres tipos de análisis, uno no paramétrico, uno semi-paramétrico y uno paramétrico, que corresponden al estimador Kaplan Meier, Riesgos Proporcionales y Vida Acelerada, respectivamente.

3.1. 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. En la Figura 5 se pueden apreciar los sub-gráficos de dicho estimador para cada una de 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. Este sub-gráfico se colocó también en la Figura 5 y se encuentra en el extremo superior izquierdo. En este, se puede observar que las probabilidades de supervivencia se diferencían mucho si se tiene una concentración de creatinina alta o no. En otras palabras, si se entra en la categoría de niveles de creatinina alta, la probabilidad de supervivencia es mucho menor, a si tienes los niveles de creatinina bajos. Dicha diferenciación se puede apreciar mejor al final en los periodos más largos.

En los otros sub-gráficos de la Figura 5 también observamos que la probabilidad de supervivencia para la presencia (1) y ausencia (0) de las variables diabetes y tabaquismo, que parecen ser muy similares a lo largo del tiempo, además de sobrelaparse. Algo similar sucede para la variable de sexo, que parece no generar supervivencias diferentes dependiendo de si eres hombre o mujer. En cuanto a las variables de anemia y presión alta, parece que la ausencia genera una mayor probabilidad de supervivencia, sin embargo, hay que notar que los intervalos de confianza se traslapan un poco para ambas categorías.

Un análisis similar se realizó para la variable salida de sangre, que recordemos se refiere al porcentaje de sangre que sale del corazón en cada contracción, la cual se discretizó en tres grupos diferentes, siguiendo los estudios de Ahmad et al. (2017). Este análisis se encuentra en la Figura 6 donde se puede apreciar que existe una mayor probabilidad de supervivencia si se tiene un mayor porcentaje de sangre que sale del corazón. En cambio, la categoría que tiene un porcentaje mejor, tiene una probabilidad de supervivencia mucho menor en comparación con las otras dos categorías.

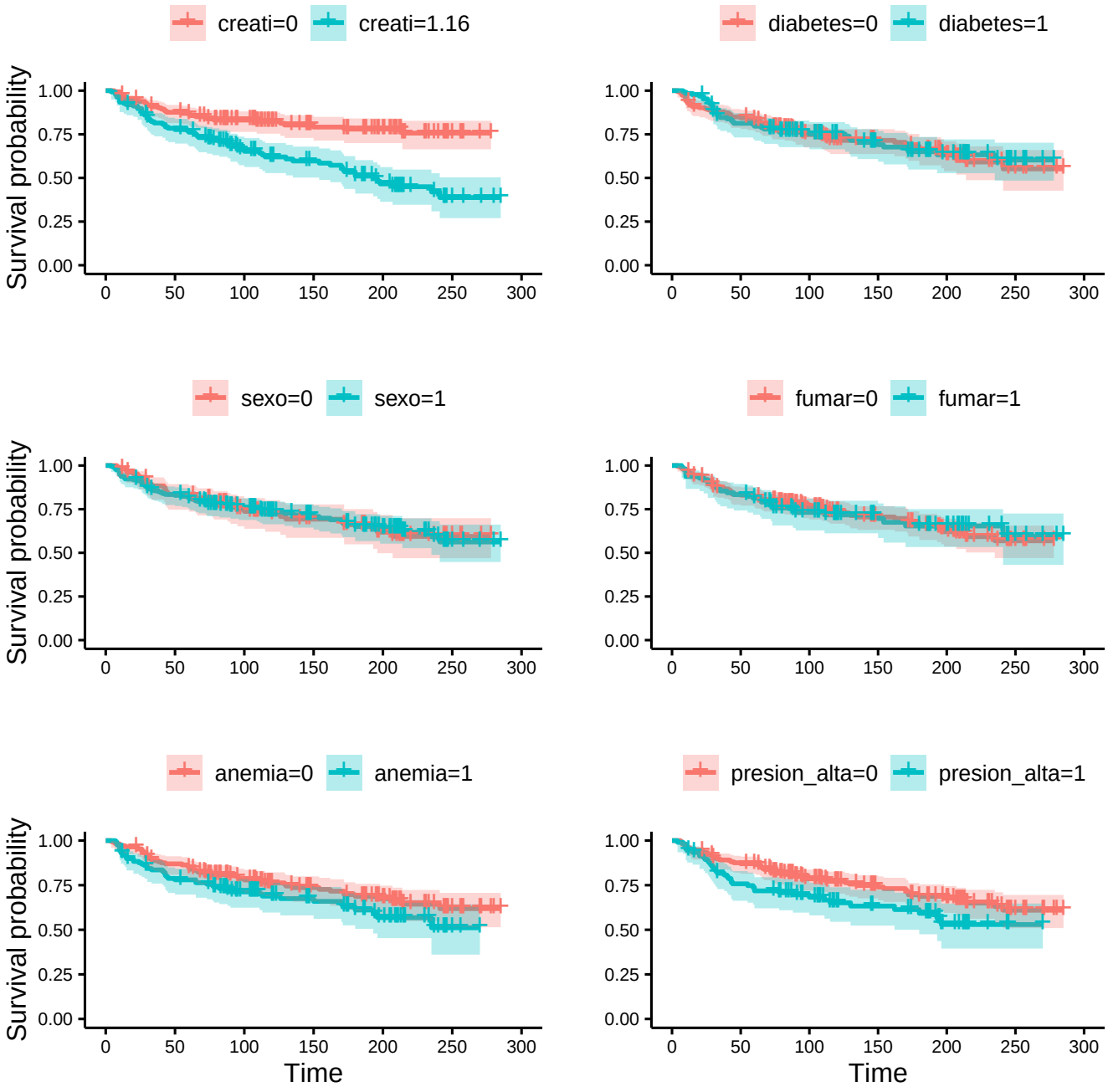

Figura 5: Estimador Kaplan Meier para las variables dicotómicas nivel de creatinina, diabetes, sexo, fumar, anemia y presión alta. Cada variable se referencía en el título de cada sub-gráfico, así como el color que indica cada valor. El cero implica ausencia (rosa) y el 1 presencia (azul), excepto para el nivel de creatinina que 0 es nivel bajo, y 1.16 alto.

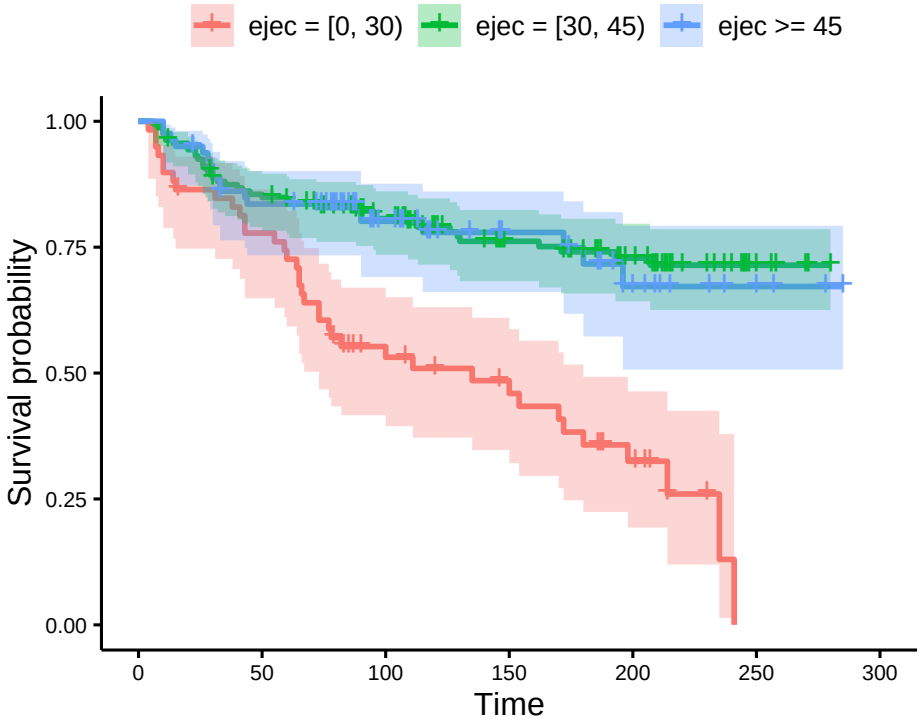

Figura 6: Estimador Kaplan Meier para la variable categórica de salida de sangre. Son tres categorías, la primera pertenece al porcentaje sangre salida del corazón menor de 30, la segunda entre 30 y 45, y la tercera de mayor o igual a 45.

3.2. Vida acelerada

Se evaluaron diversos modelos probando diferentes combinaciones de variables, así como aquel modelo que consideraba todas las variables. El modelo que se presenta a continuación fue uno de los que consideramos más parsimoniosos y que considera las variables que aportan significancia al modelo.

En el siguiente bloque se presenta un resumen de la evaluación del mejor modelo. En este se puede observar que los coeficientes de edad, salida de sangre y nivel de creatinina, son significativas por lo que sus coeficientes son diferentes de cero, es decir, tienen un impacto en los tiempos de supervivencia. Por otro lado, las variables de enzima cpk y presión alta, no tuvieron significancia en este modelo. En evaluaciones de modelos anteriores, quitamos las variables de enzima cpk y presión alta del modelo, sin embargo, eso generaba que la log-verosimilitud de modelo no fuera tan adecuada, en comparación a este modelo. Adicionalmente, notamos que las variables categóricas no tenían significancia cuando se las agregaba a los modelos. La única variable binaria que parece tener un efecto es la de presión alta. Las otras variables como sexo, diabetes y fumar, casi en ningún modelo tenían significancia.

```
##
## 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
##
## 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
```

3.3. Riesgos proporcionales

Para el modelo de riesgos proporcionales, los coeficientes que tienen un efecto en los tiempos de supervivencia son la edad, la salida de sangre la presión alta y el nivel de creatinina. En este caso, la variable de enzima cpk no tiene efecto en la supervivencia. Si observamos, en este modelo la variable de enzima cpk sigue sin tener efecto. Por otro lado la variable de presión alta, para este modelo tiene significacia de 0.01, por lo que si entra en la zona de rechazo por lo que puedo decir que tiene una influencia en los tiempos de supervivencia.

En cuanto a los demás modelos evaluados, este fue de la misma manera el que tenía las variables que consideramos más adecuadas dadas las métricas de log-verosimilitud. Además, se confirmó de la misma manera que las variables binarias como sexo, fumar y diabetes no tuvieron efecto en la supervivencia. Otra variable que salió significativa en ambos modelos fue la del nivel de creatinina, que desde antes que habíamos realizado una categorización entre alta y baja, habíamos observado que generaba una diferencia en la probabilidad de supervivencia. En estos modelos, se confirmó que tomarla en cuenta como variable continua, también es significativa en el modelo y que su coeficiente es diferente de cero.

```
## 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
```

3.4. Validación de modelos

Para la realizar la validación de los modelos ya mencionábamos que realizamos la evaluación de diferentes combinaciones de variables, y concluimos que el modelo que contempla edad, proporción de salida de sangre, nivel de creatinina y presión alta, son las que tienen un efecto en los tiempos de supervivencia. Para continuar con la elección y validación de modelos presentamos la Figura 7 que muestra la probabilidad de supervivencia de los mejores modelos para la estimación paramétrica y semiparamétrica, que incluyen los regresores ya mencionados. En esta figura también se agregó el estimador Kaplan Meier para poder comparar de manera empírica con las diferentes curvas de supervivencia.

En esta figura se puede observar como la curva del modelo de vida acelerada, decae drásticamente y no coincide en nada con el estimador Kaplan Meier. En cambio, el modelo de riesgos

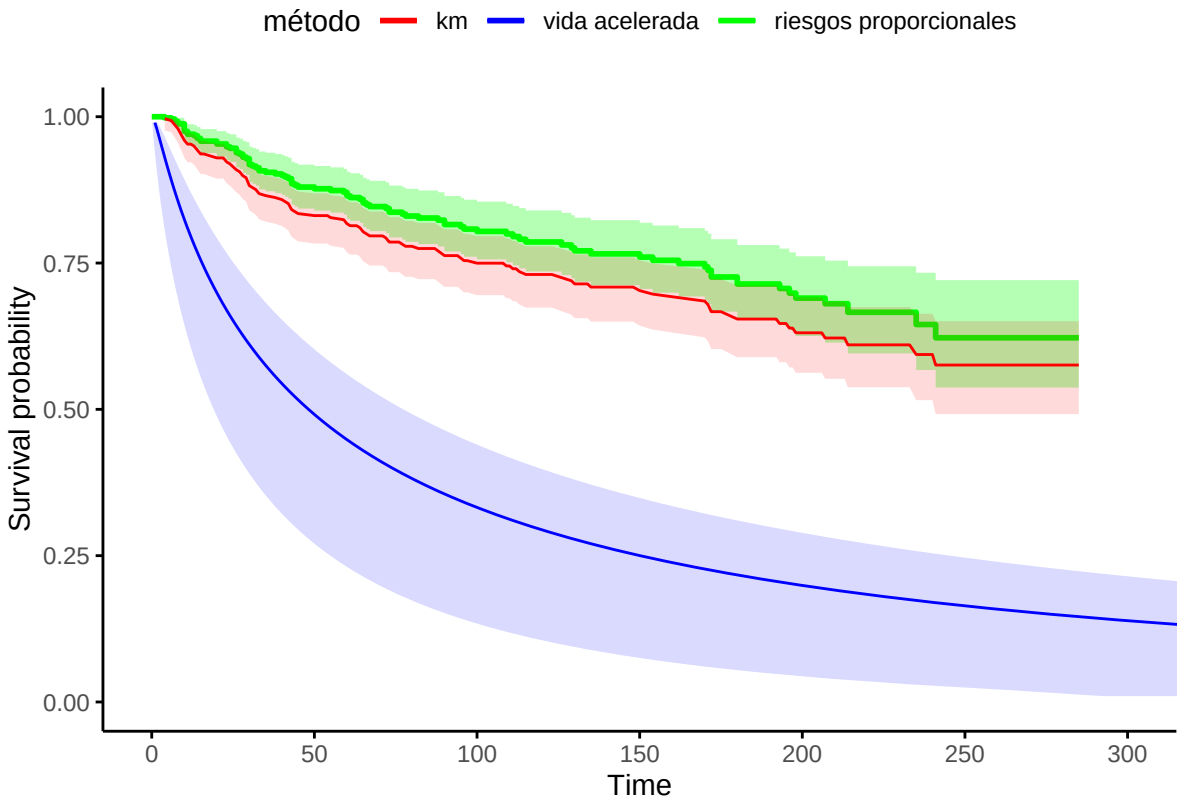

Figura 7: Curvas de supervivencia para el estimador KM, para el modelo de vida acelerada y para el modelo de riesgos proporcionales, estos dos últimos considerando las mejores variables regresoras.

proporcionales parece acercarse mejor a la curva empírica de los tiempos de supervivencia. Esto puede ser un indicio de que para nuestros datos conviene usar un modelo semi-paramétrico, pues este parece adecuarse mejor.

Otra manera de validar los mejores modelos es tomando en cuenta los residuales, la Figura 8, muestra tres diferentes gráficos que muestran el comportamiento de los residuales para el modelo de vida acelerada. De este gráfico lo primero que tenemos que observar es que la distribución de los residuales no sigue una distribución normal y que muchos de éstos se encuentran jalados a valores negativos. Además, tiene un pico muy alto en el histograma que se acerca a cero. En el gráfico superior, se observa que también todos los residuales tienden a ser negativos para casi todos los periodos de tiempo, lo cual no es un buen indicio. Por último, el gráfico de ACF, muestra la autorrelación de estos mismos y se observa que en un punto se pasa de las bandas azules y en otros se acercan mucho, por lo que no es un buen indicio de este modelo.

Por último, mostraremos los residuales para el modelo de riesgos proporcionales, mejor conocidos como residuos Cox-Snell. En la Figura 9, se muestra un gráfico de diagnóstico donde si el modelo de riesgos proporcionales ajusta a los datos, la gráfica debe ser una línea recta que pasa por el origen. Para este caso, se observa que los residuales coinciden en varios puntos con las tasa de riesgo acumulada, solo existen algunos residuales en la parte central del gráfico que no coinciden

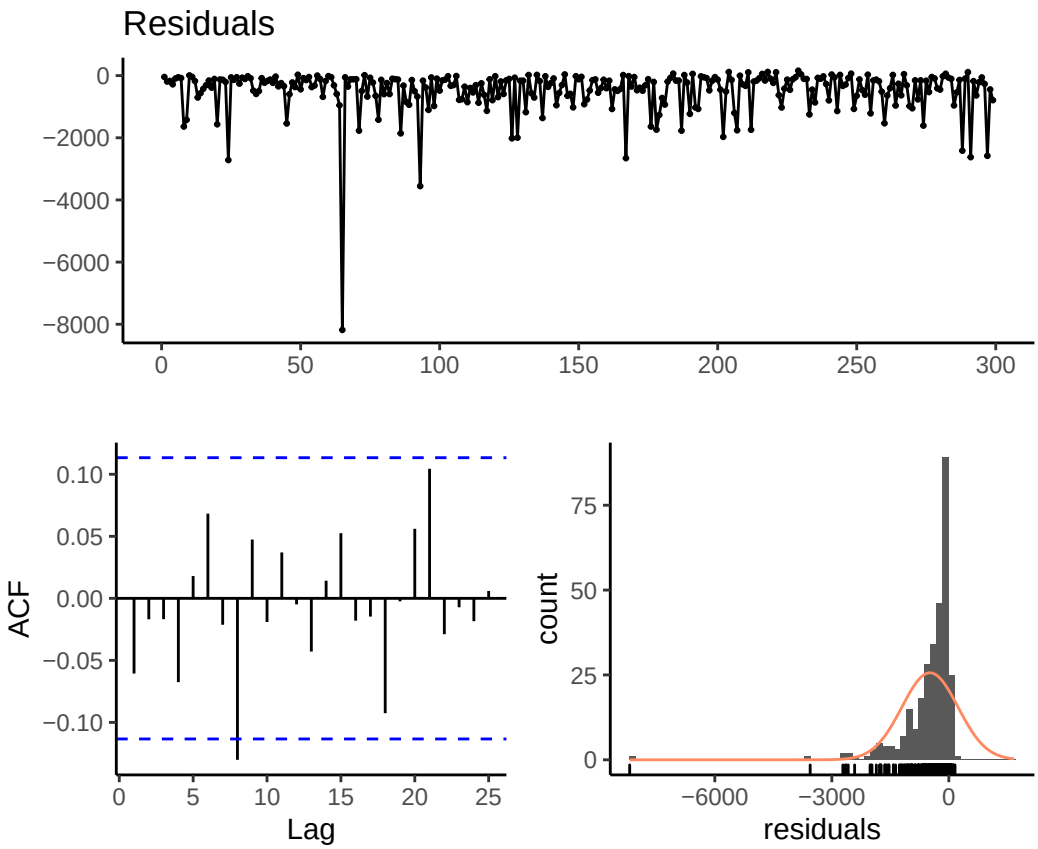

Figura 8: Gráficos de residuales para el modelo de vida acelerada.

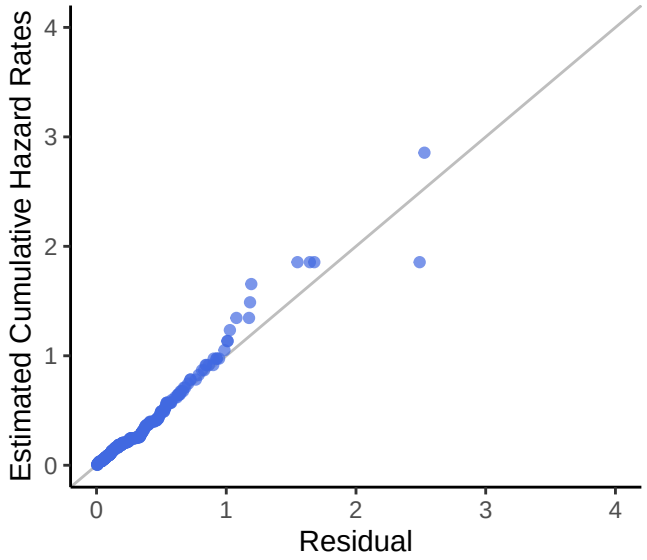

Figura 9: Residuales para el modelo de riesgos proporcionales.

a la perfección con la línea recta, sin embargo, realmente no son muchas las observaciones que no coinciden.

4. Resultados y conclusiones

En el presente proyecto se encontró que las variables que tienen un efecto sigficativo sobre los tiempos de supervivencia fueron la edad, proporción de salida de sangre, nivel de creatinina y presión alta. El modelo que mejor parece adecuarse a nuestros datos fue el de riesgos proporcionales, pues como vimos parece acercarse más a la supervivencia de KM y tiene mejor comportamiento en sus residuales. El modelo de vida acelerada, es uno completamente paramétrico que puede que no se se este adecuado a nuestros datos, tal como lo vimos en sus residuales. Además, evaluamos familias de modelos para la regresión de vida acelerada y aún así, el modelo semiparamétrico funciona mucho mejor.

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.

MLR: Machine Learning repository, Irvine University. (2021). [https://archive.ics.uci.edu].

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.