Analisis del riesgo de renuncia con R

Aldo Enrique Chong Valetin

2023-01-04

En este documento se hace una demostracion de analisis de superviviencia en R. Trabajamos con la base de datos "Employee Turnover", disponible en https://www.kaggle.com/datasets/davinwijaya/employee-turnover. Esta base de datos tiene informacion sobre empleados de distintas industrias de Rusia y Ucrania. La base de datos contiene variables como Experiencia, Evento de renuncia, genero, edad, industria y algunas otras. Se tiene que el evento de interes es el evento de renuncia y el tiempo de superviviencia asociado son los años de experiencia del empleado. El dataset incluye 1129 observaciones. Damos un vistazo a la base de datos para familiarizarnos con ella.

```
#Cargamos los datos y la libreria que vamos a utilizar

library(survival)
data = read.csv("turnover.csv")

#Vemos las primeras observaciones
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
head(data)
```

```
##        stag event gender age        industry profession     traffic coach
## 1  7.030801     1      m  35           Banks         HR rabrecNErab    no
## 2 22.965092     1      m  33           Banks         HR       empjs    no
## 3 15.934292     1      f  35 PowerGeneration         HR rabrecNErab    no
## 4 15.934292     1      f  35 PowerGeneration         HR rabrecNErab    no
## 5  8.410678     1      m  32          Retail Commercial       youjs   yes
## 6  8.969199     1      f  42     manufacture         HR       empjs   yes
##   head_gender greywage way extraversion independ selfcontrol anxiety novator
## 1           f    white bus          6.2      4.1         5.7     7.1     8.3
```

```
## 2          m    white bus         6.2      4.1        5.7      7.1      8.3
## 3          m    white bus         6.2      6.2        2.6      4.8      8.3
## 4          m    white bus         5.4      7.6        4.9      2.5      6.7
## 5          f    white bus         3.0      4.1        8.0      7.1      3.7
## 6          m    white bus         6.2      6.2        4.1      5.6      6.7
```

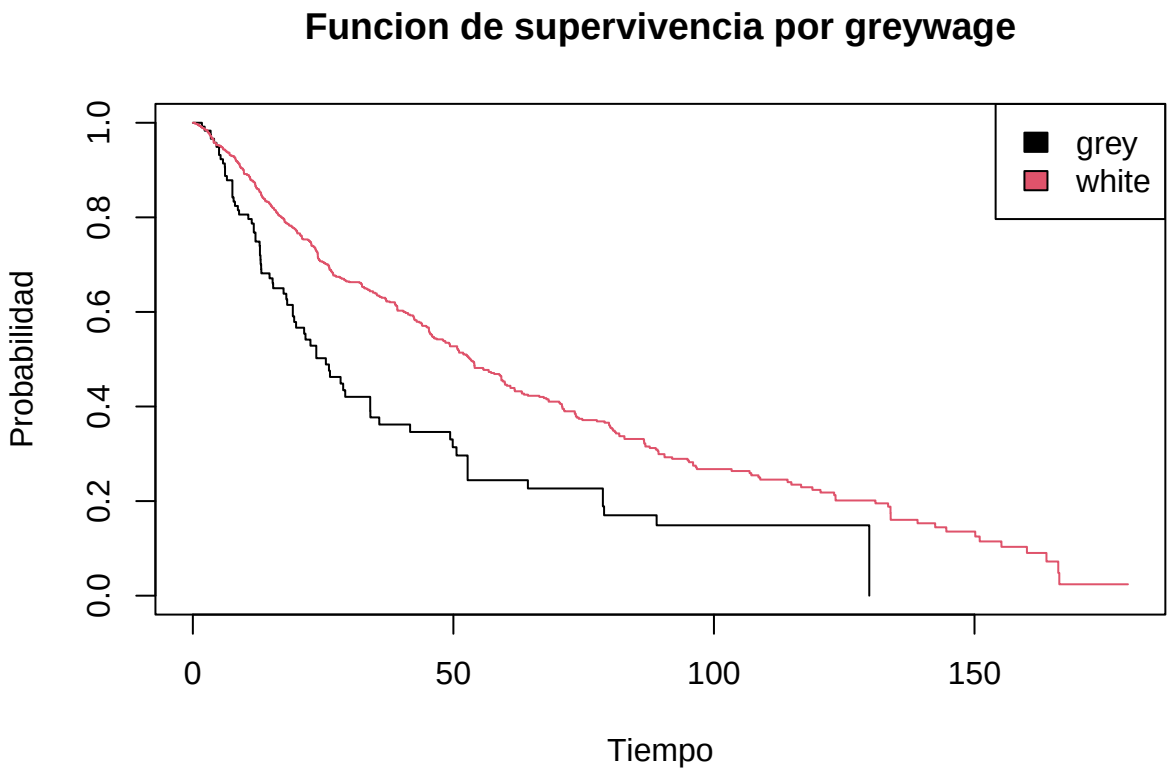
En particular hay una variable llamada greywage de particular interes. Para explicarla damos la siguiente definicion. Sea A un trabajador. Decimos que A tiene whitewage si y solo si A recibe exactamente el pago que dice su contrato. Decimos que A tiene greywage si y solo si A no tiene whitewage. Asi, la variable greywage indica si el trabajador tiene salario whitewage o si tiene greywage.

Asi pues, surge la siguiente pregunta: ¿Un empleado tiene mas riesgo de renunciar dado que su salario es greywage? Procedemos a responder a esta pregunta. Para eso haremos la estimacion de la funcion de superviviencia de Kaplan - Meier.

```
#Inicializamos el objeto de supervivencia
surv_obj = Surv(data$stag, data$event)

#Hacemos el ajuste Kaplan-Maier
fit= survfit(surv_obj ~ data$greywage, conf.type = 'plain', type = 'kaplan-meier', conf.int = 0.95)

#Hacemos la grafica.
plot(fit, col = c(1,2), xlab = 'Tiempo', ylab = 'Probabilidad')
title(main = 'Funcion de supervivencia por greywage')
legend(x = 'topright', legend = c('grey', 'white'), fill = c(1,2))
```



Vemos que la curva de supervivencia asociada a la curva de whitewage es significativamente mayor que la curva de superviviencia asociada al greywage. De tal manera que sospechamos que las personas con greywage tienen mayor riesgo de renuncia. Para comprobarlo formalmente aplicamos la prueba de Log-Rank.

Recordemos que la prueba Log Rank tiene la siguiente prueba de hipotesis: $H_0 : S_1(t) = S_2(t)$ y $H_a : S_1(t) \neq S_2(t)$. Es decir, la hipotesis nula es que las dos funciones de superviviencia son iguales.

```
#Aplicamos la prueba de Log-Rank
survdiff(surv_obj ~ data$greywage , rho = 0, data = data)
```

```
## Call:
## survdiff(formula = surv_obj ~ data$greywage, data = data, rho = 0)
##
##                         N Observed Expected (O-E)^2/E (O-E)^2/V
## data$greywage=grey    127       73     43.2     20.47      22.3
## data$greywage=white  1002      498    527.8      1.68      22.3
##
##  Chisq= 22.3  on 1 degrees of freedom, p= 2e-06
```

Vemos que el p-value es menor que 0.05 (de hecho es muy pequeño). De tal suerte que rechazamos la hipotesis nula y concluimos formalmente que las funciones de superviviencia son distintas.

Asi pues, podemos concluir el hecho de que un trabajador tenga greywage si aumenta su riesgo de renuncia.