Tarea 2 - Análisis de Supervivencia

Juan Esteban Sánchez Pulgarín, añadir los otros nombres

Librerías que se van a utilizar.

```
knitr::opts_chunk$set(fig.align = 'center', warning = F, message = F, comment = '')
```

```
library("survival")
library("tinytex")
library("tidyverse")
library("splines")
library("ggplot2")
library("gridExtra")
library("cowplot")
library("survminer")
```

Lectura de la base de datos y se convierten las variables categoricas a tipo factor.

```
base <- read.csv("ovarian.csv")

base$resid.ds <- as.factor(base$resid.ds)

base$ecog.ps <- as.factor(base$ecog.ps)

base$rx <- as.factor(base$rx)
```

Se tiene una base de datos que proviene de un estudio de supervivencia en mujeres con cáncer de ovario, sometidas a dos tratamientos distintos, en el cual se tienen en cuenta siete variables que son:

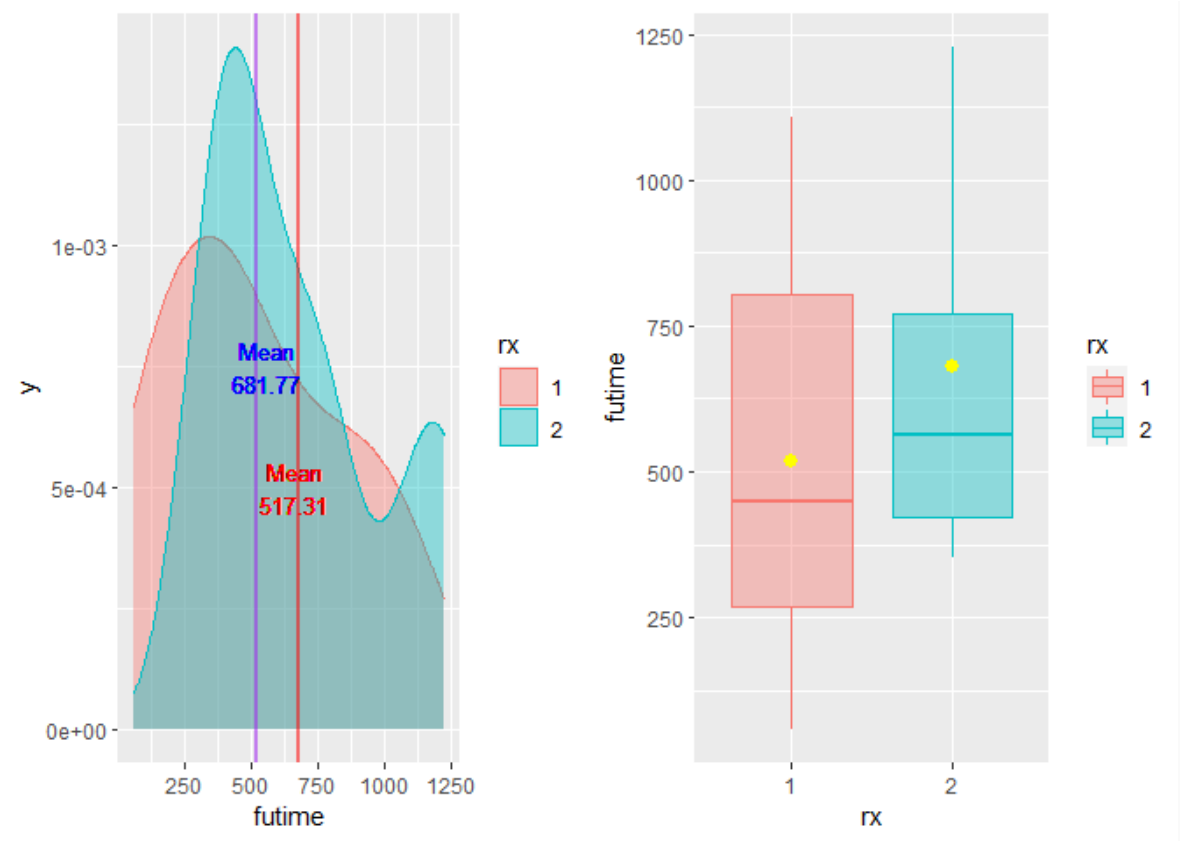
- futime (tiempo de supervivencia en días)
- fustat (estado de la censura donde fustat = 0 es censurado, y = 1 es defunción, representa la falla)
- age (edad de la paciente en años)
- resid.ds (persistencia de la enfermedad tras el tratamiento 1 = No, 2 = Si)
- rx (tratamiento a la que es sometida la paciente)

- ecog.ps (puntuación del test, ECOG = 1 es buen estado).

Primero se mirará unos gráficos descriptivos

Distribución del tiempo de supervivencia en función del tratamiento al que es sometido el paciente.

```
par(pty = 's')
a <- by(base$resid.ds, base$rx, summary)

means1 <-round(c(by(base$futime, base$rx, mean)), 2)


g1 <- ggplot(base, aes(x = futime, group = rx, color = rx, fill = rx))+
  geom_density(alpha = 0.4)+
  geom_vline(xintercept = means1[1], size = 0.9, color = "purple", alpha = 0.54)+
  geom_vline(xintercept = means1[2], size = 0.9, color = "red", alpha = 0.54)+
  geom_text(aes(x = means1[1]+150, label = paste0("Mean\n", means1[1]), y = 0.0005), size =
        3.5, col = "red")+
  geom_text(aes(x = means1[2] - 120, label = paste0("Mean\n", means1[2]), y=0.00075), size
        = 3.5, col = "blue")



g2 <- ggplot(base, aes(x = rx, y = futime, group = rx, color = rx, fill = rx))+
  geom_boxplot(alpha = 0.4)+
  stat_summary(aes(x = rx, y = futime, group = rx, color = rx), fun = mean, geom = "point",
        shape = 20, size = 4, color = "yellow", fill = "red", position = "identity")

plot_grid(g1, g2)
```

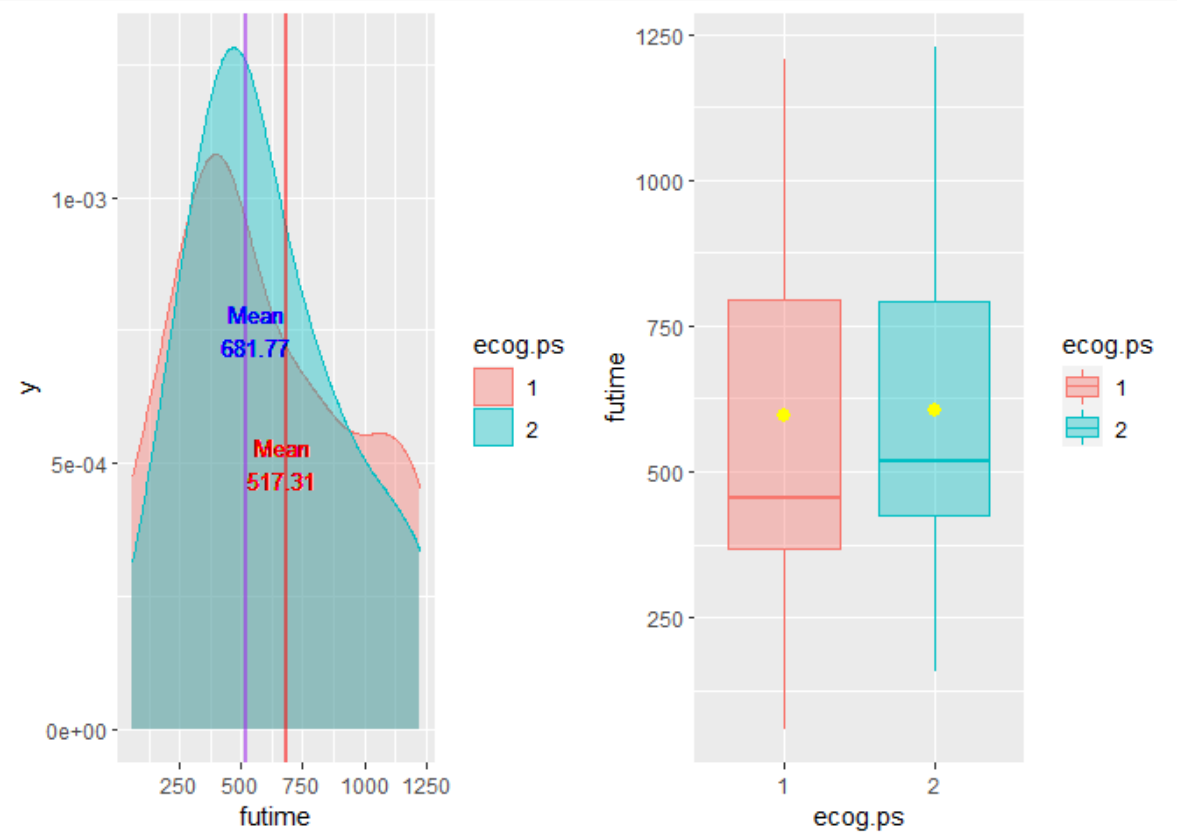


De los gráficos anteriores se observa que la distribución del tiempo de supervivencia frente al tratamiento al que es sometido el paciente presenta un leve cambio, pero las cajas de los boxplots no se traslapan demasiado, por lo tanto, se podría pensar que la influencia del tratamiento no es estadísticamente significativa.

Distribución del tiempo de supervivencia en función de la puntuación del test del paciente donde 1 significa buen estado y 2 mal estado.

```
par(pty = 's')
g3 <-ggplot(base, aes( x = futime, group = ecog.ps, color = ecog.ps, fill = ecog.ps))+
  geom_density(alpha = 0.4) +
  geom_vline(xintercept = means1[1], size = 0.9, color = "purple", alpha = 0.54)+
  geom_vline(xintercept = means1[2], size = 0.9, color = "red", alpha = 0.54)+
  geom_text(aes(x = means1[1] + 150, label = paste0("Mean\n", means1[1]), y = 0.0005), size
        = 3.5, col = "red")+
  geom_text(aes(x = means1[2] - 120, label = paste0("Mean\n", means1[2]), y = 0.00075),
        size = 3.5, col = "blue")



g4 <- ggplot(base, aes(x = ecog.ps, y = futime, group = ecog.ps, color = ecog.ps, fill =
        ecog.ps))+
  geom_boxplot(alpha = 0.4)+
  stat_summary(aes(x = ecog.ps, y = futime, group = ecog.ps, color = ecog.ps), fun = mean,
        geom = "point", shape = 20, size = 4, color = "yellow", fill = "red", position =
        "identity")
```

```
plot_grid(g3, g4)
```

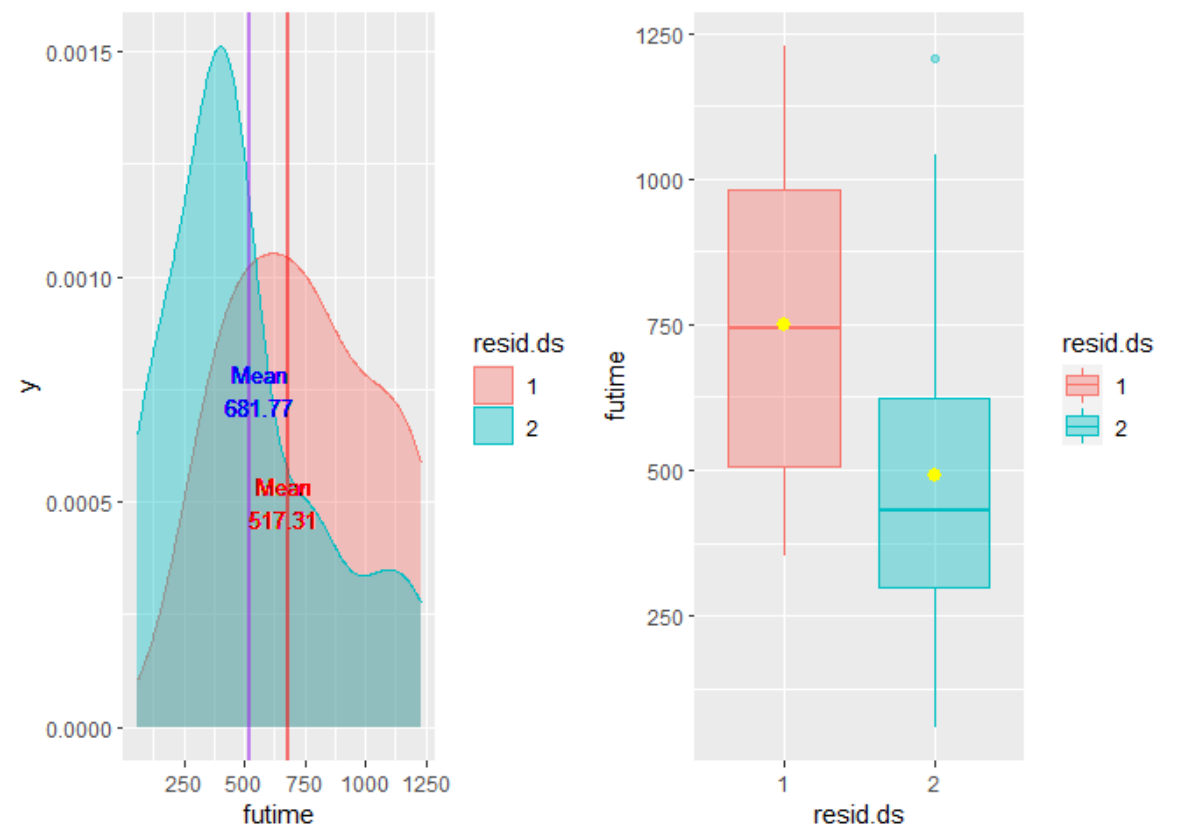


De los gráficos anteriores se observa que la distribución del tiempo de supervivencia frente a la puntación del test del paciente presenta un leve cambio, pero las cajas de los boxplots no se traslapan demasiado, por lo tanto, se podría pensar que la influencia de la puntación del test no es estadísticamente significativa.

Distribución del tiempo de supervivencia en función de la persistencia de la enfermedad tras el tratamiento, 1 = No y 2 = Si.

```
par(pty = 's')
g5 <-ggplot(base, aes(x = futime, group = resid.ds, color = resid.ds, fill = resid.ds))+
  geom_density(alpha = 0.4)+
  geom_vline(xintercept = means1[1], size = 0.9, color = "purple", alpha = 0.54)+
  geom_vline(xintercept = means1[2], size = 0.9, color = "red", alpha = 0.54)+
  geom_text(aes(x = means1[1] + 150, label = paste0("Mean\n", means1[1]), y = 0.0005), size
        = 3.5, col = "red")+
  geom_text(aes(x = means1[2] - 120, label = paste0("Mean\n", means1[2]), y = 0.00075),
        size = 3.5, col = "blue")



g6 <- ggplot(base, aes(x = resid.ds, y = futime, group = resid.ds, color = resid.ds, fill =
        resid.ds))+
```

```
  geom_boxplot(alpha = 0.4)+
  stat_summary(aes(x = resid.ds, y = futime, group = resid.ds, color = resid.ds), fun =
         mean, geom = "point", shape = 20, size = 4, color = "yellow", fill = "red",
         position = "identity")


plot_grid(g5, g6)
```



De los gráficos anteriores se observa que la distribución del tiempo de supervivencia frente a la persistencia de la enfermedad presenta un cambio en la distribución, similarmente las cajas de los boxplots se traslapan, por lo tanto, se podría pensar que la persistencia de la enfermedad es estadísticamente significativa.

1. Ajuste al menos tres modelos parámetricos de la familia AFT y seleccione uno de ellos de acuerdo a algún criterio o estadístico de los vistos en clase. Inicialmente, debe usar todas las covariables, y secuencialmente y con cada distribución, seleccione las variables estadísticamente importantes de acuerdo a su valor-p. Una vez seleccione un modelo, evalúe los posibles efectos confusores de las variables que no fueron

incluidas en el modelo (si es que aplica) e interacciones. INTERPRETE.

a. modelo exponencial

Se ajusta un primer modelo de regresión con distribución exponencial y con todas las covariables

```
#Modelo exponencial con todas las variables
result1 <- survreg(Surv(futime, fustat) ~ age + resid.ds + rx + ecog.ps, dist =
         "exponential", data = base)

summary(result1)
```

```
Call:
survreg(formula = Surv(futime, fustat) ~ age + resid.ds + rx +
    ecog.ps, data = base, dist = "exponential")
             Value Std. Error     z       p
(Intercept) 12.3913     2.0130  6.16 7.5e-10
age         -0.0875     0.0338 -2.59  0.0096
resid.ds2   -0.7659     0.7411 -1.03  0.3014
rx2          0.6269     0.6162  1.02  0.3090
ecog.ps2    -0.2523     0.6061 -0.42  0.6772

Scale fixed at 1

Exponential distribution
Loglik(model)= -90.6   Loglik(intercept only)= -98
    Chisq= 14.78 on 4 degrees of freedom, p= 0.0052
Number of Newton-Raphson Iterations: 5
n= 26
```

Del resumen anterior se observa que el modelo es estadísticamente significativo, pero la única variable significativa es la edad, por lo tanto, se puede pensar en comenzar a eliminar variables que no son estadísticamente significativas.

Se ajusta un nuevo modelo, pero esta vez sin la variable **ecog.ps** que es la menos significativa

```
#Modelo exponencial sin ecog.ps
result2 <- survreg(Surv(futime, fustat) ~ age + resid.ds + rx, dist = "exponential", data =
         base)
summary(result2)
```

```
Call:
survreg(formula = Surv(futime, fustat) ~ age + resid.ds + rx,
    data = base, dist = "exponential")
              Value Std. Error     z       p
(Intercept) 12.3982     2.0401  6.08 1.2e-09
age         -0.0906     0.0335 -2.70  0.0069
resid.ds2   -0.6968     0.7296 -0.96  0.3395
rx2          0.6157     0.6189  0.99  0.3198

Scale fixed at 1

Exponential distribution
Loglik(model)= -90.7   Loglik(intercept only)= -98
    Chisq= 14.6 on 3 degrees of freedom, p= 0.0022
Number of Newton-Raphson Iterations: 5
n= 26
```

Del resumen anterior se observa que el modelo es estadísticamente significativo, pero la única variable significativa es la edad, por lo tanto, se puede pensar en seguir eliminando variables que no son estadísticamente significativas.

Se ajusta un nuevo modelo sin la variable **rx**, que no es la menos significativa, pero por las gráficas que se vieron previamente se puede pensar que tiene menos peso que **resid.ds**.

```
#Modelo exponencial sin rx
result3 <- survreg(Surv(futime, fustat) ~ age + resid.ds, dist = "exponential", data =
         base)
summary(result3)
```

```
Call:
survreg(formula = Surv(futime, fustat) ~ age + resid.ds, data = base,
    dist = "exponential")
             Value Std. Error     z       p
(Intercept) 13.442      2.042  6.58 4.6e-11
age         -0.102      0.035 -2.92  0.0035
resid.ds2   -0.733      0.715 -1.02  0.3056

Scale fixed at 1

Exponential distribution
```

```
Loglik(model)= -91.2   Loglik(intercept only)= -98
    Chisq= 13.62 on 2 degrees of freedom, p= 0.0011
Number of Newton-Raphson Iterations: 5
n= 26
```

Del resumen anterior se observa que el modelo es estadísticamente significativo, pero la única variable significativa es la edad, por lo tanto, se puede pensar en seguir eliminando variables que no son estadísticamente significativas.

Se ajusta un nuevo modelo sin la variable **resid.ds**

```
#Modelo sin rx
result4 <- survreg(Surv(futime, fustat) ~ age, dist = "exponential", data = base)
summary(result4)
```

```
Call:
survreg(formula = Surv(futime, fustat) ~ age, data = base, dist = "exponential")
             Value Std. Error     z       p
(Intercept) 13.9339     2.1030  6.63 3.5e-11
age         -0.1185     0.0339 -3.50 0.00047

Scale fixed at 1

Exponential distribution
Loglik(model)= -91.8   Loglik(intercept only)= -98
    Chisq= 12.51 on 1 degrees of freedom, p= 0.00041
Number of Newton-Raphson Iterations: 5
n= 26
```

Se observa que el modelo es estadísticamente significativo, sin embargo, vamos a probrar con otra distribución.

b. distribución log-normal

Se ajusta un primer modelo de regresión con distribución lognormal y con todas las covariables.

```
result1_logn<-survreg(Surv(futime,fustat)~resid.ds+ecog.ps+rx+age,dist="lognormal",data=base)

summary(result1_logn)
```

```
Call:
survreg(formula = Surv(futime, fustat) ~ resid.ds + ecog.ps +
    rx + age, data = base, dist = "lognormal")
              Value Std. Error     z       p
(Intercept) 10.3932     0.9833 10.57 < 2e-16
resid.ds2   -0.5331     0.3493 -1.53   0.127
ecog.ps2    -0.0205     0.3203 -0.06   0.949
rx2          0.6059     0.3027  2.00   0.045
age         -0.0676     0.0163 -4.15 3.3e-05
Log(scale)  -0.4590     0.2202 -2.08   0.037

Scale= 0.632

Log Normal distribution
Loglik(model)= -86.4   Loglik(intercept only)= -97.1
    Chisq= 21.43 on 4 degrees of freedom, p= 0.00026
Number of Newton-Raphson Iterations: 6
n= 26
```

Del resumen anterior se observa que el modelo es estadísticamente significativo, pero las únicas variables significativas son la edad y rx(tratamiento), por lo tanto, se puede pensar en comenzar a eliminar variables que no son estadísticamente significativas.

- Eliminando **ecog.ps** que es la menos significativa

```
result2_logn<-survreg(Surv(futime,fustat)~resid.ds+rx+age,dist="lognormal",data=base)
summary(result2_logn)
```

```
Call:
survreg(formula = Surv(futime, fustat) ~ resid.ds + rx + age,
    data = base, dist = "lognormal")
              Value Std. Error     z       p
(Intercept) 10.3867     0.9756 10.65 < 2e-16
resid.ds2   -0.5258     0.3288 -1.60   0.110
rx2          0.6030     0.2984  2.02   0.043
age         -0.0677     0.0161 -4.20 2.6e-05
Log(scale)  -0.4609     0.2178 -2.12   0.034

Scale= 0.631

Log Normal distribution
```

```
Loglik(model)= -86.4   Loglik(intercept only)= -97.1
    Chisq= 21.42 on 3 degrees of freedom, p= 8.6e-05
Number of Newton-Raphson Iterations: 6
n= 26
```

Del resumen anterior se observa que el modelo es estadísticamente significativo, pero las únicas variables significativas son la edad y rx(tratamiento), por lo tanto, se puede pensar en seguir eliminando variables que no son estadísticamente significativas.

- Eliminando **resid.ds**

```
result3_logn<-survreg(Surv(futime,fustat)~rx+age,dist="lognormal",data=base)
summary(result3_logn)
```

```
Call:
survreg(formula = Surv(futime, fustat) ~ rx + age, data = base,
    dist = "lognormal")
             Value Std. Error     z       p
(Intercept) 10.5449     1.0482 10.06 < 2e-16
rx2          0.6904     0.3169  2.18   0.029
age         -0.0765     0.0171 -4.48 7.4e-06
Log(scale)  -0.3813     0.2189 -1.74   0.082

Scale= 0.683

Log Normal distribution
Loglik(model)= -87.7   Loglik(intercept only)= -97.1
    Chisq= 18.93 on 2 degrees of freedom, p= 7.8e-05
Number of Newton-Raphson Iterations: 6
n= 26
```

- Las variables **rx** y **age** son significativas para ajustar el modelo de regresión usando la distribución lognormal, sin embargo, vamos a probar con otra distribución.

c. distribución log-logistic

- Usando todas las variables

```
result1_logl<-survreg(Surv(futime,fustat)~resid.ds+ecog.ps+rx+age,dist="loglogistic",data=base)

summary(result1_logl)
```

```
Call:
survreg(formula = Surv(futime, fustat) ~ resid.ds + ecog.ps +
    rx + age, data = base, dist = "loglogistic")
              Value Std. Error     z       p
(Intercept) 10.4756     1.0427 10.05 < 2e-16
resid.ds2   -0.5469     0.3530 -1.55    0.12
ecog.ps2    -0.0346     0.3147 -0.11    0.91
rx2          0.5957     0.3039  1.96    0.05
age         -0.0690     0.0174 -3.97 7.2e-05
Log(scale)  -0.9869     0.2412 -4.09 4.3e-05

Scale= 0.373

Log logistic distribution
Loglik(model)= -86.8   Loglik(intercept only)= -97.4
    Chisq= 21.06 on 4 degrees of freedom, p= 0.00031
Number of Newton-Raphson Iterations: 6
n= 26
```

Del resumen anterior se observa que el modelo es estadísticamente significativo, pero las únicas variables significativas son la edad y rx(tratamiento) que tiene una significancia de borde, por lo tanto, se puede pensar en comenzar a eliminar variables que no son estadísticamente significativas.

- Eliminando la variable **ecog.ps** que es la menos significativa.

```
result2_logl<-survreg(Surv(futime,fustat)~resid.ds+rx+age,dist="loglogistic",data=base)
summary(result2_logl)
```

```
Call:
survreg(formula = Surv(futime, fustat) ~ resid.ds + rx + age,
    data = base, dist = "loglogistic")
              Value Std. Error     z       p
(Intercept) 10.4561     1.0209 10.24 < 2e-16
resid.ds2   -0.5371     0.3404 -1.58    0.11
rx2          0.5941     0.3026  1.96    0.05
age         -0.0691     0.0173 -4.00 6.4e-05
Log(scale)  -0.9892     0.2398 -4.12 3.7e-05

Scale= 0.372

Log logistic distribution
```

```
Loglik(model)= -86.8   Loglik(intercept only)= -97.4
    Chisq= 21.05 on 3 degrees of freedom, p= 1e-04
Number of Newton-Raphson Iterations: 6
n= 26
```

Del resumen anterior se observa que el modelo es estadísticamente significativo, pero las únicas variables significativas son la edad y rx(tratamiento) que tiene una signifcancia de borde, por lo tanto, se puede pensar en seguir eliminando variables que no son estadísticamente significativas.

- Se elimina **resid.ds**

```
result3_logl<-survreg(Surv(futime,fustat)~rx+age,dist="loglogistic",data=base)
summary(result3_logl)
```

```
Call:
survreg(formula = Surv(futime, fustat) ~ rx + age, data = base,
    dist = "loglogistic")
             Value Std. Error     z       p
(Intercept) 10.8071     1.1538  9.37 < 2e-16
rx2          0.5911     0.3298  1.79 0.07309
age         -0.0805     0.0187 -4.31 1.6e-05
Log(scale)  -0.9050     0.2385 -3.79 0.00015

Scale= 0.405

Log logistic distribution
Loglik(model)= -88   Loglik(intercept only)= -97.4
    Chisq= 18.62 on 2 degrees of freedom, p= 9e-05
Number of Newton-Raphson Iterations: 5
n= 26
```

- Se elimina **rx**, aunque se podria seguir considerando como significancia de borde.

```
result3_logl<-survreg(Surv(futime,fustat)~age,dist="loglogistic",data=base)
summary(result3_logl)
```

```
Call:
survreg(formula = Surv(futime, fustat) ~ age, data = base, dist = "loglogistic")
             Value Std. Error     z       p
```

```
(Intercept) 11.6495     1.3767  8.46 < 2e-16
age         -0.0887     0.0225 -3.95 7.9e-05
Log(scale)  -0.8003     0.2428 -3.30 0.00098

Scale= 0.449

Log logistic distribution
Loglik(model)= -89.6   Loglik(intercept only)= -97.4
    Chisq= 15.61 on 1 degrees of freedom, p= 7.8e-05
Number of Newton-Raphson Iterations: 5
n= 26
```

- Solo la variable **age** es significativa para ajustar el modelo de regresión usando la distribución loglogistic.

Ahora se van a evaluar los modelos con el criterio AIC y BIC

```
criterioAIC <- AIC(result1, result1_logl, result1_logn, result2, result2_logl,
         result2_logn, result3, result3_logl, result3_logn, result4)

criterioBIC <- BIC(result1, result1_logl, result1_logn, result2, result2_logl,
         result2_logn, result3, result3_logl, result3_logn, result4)
```

```
library(kableExtra)

tabla <- cbind(criterioAIC, criterioBIC)

kable(tabla[ ,c(1, 2, 4)]) %>%
  kable_styling(full_width = F)
```

	df	AIC	BIC
result1	5	191.2886	197.5791
result1_logl	6	185.6466	193.1952
result1_logn	6	184.8166	192.3652
result2	4	189.4634	194.4958
result2_logl	5	183.6588	189.9492
result2_logn	5	182.8207	189.1112

	df	AIC	BIC
result3	3	188.4435	192.2178
result3_logl	3	185.1017	188.8760
result3_logn	4	183.3164	188.3488
result4	2	187.5586	190.0748

De la tabla se observa que el modelo tres con distribución lognormal es el que cuenta con menor BIC, pero en el otro criterio es el segundo modelo con menor AIC, entonces, por el criterio de parsimonia se escoge el modelo tres con distribución lognormal ya que cuenta con pocos parámetros y las diferencias de los criterios entre los modelos no es muy grande.

Evaluando interacción en el modelo

```
#Interaction

resultFinal <- survreg(Surv(futime, fustat) ~ rx*age, dist = "lognormal", data = base)
summary(resultFinal)
```

```
Call:
survreg(formula = Surv(futime, fustat) ~ rx * age, data = base,
    dist = "lognormal")
             Value Std. Error     z       p
(Intercept) 10.2591     1.0583  9.69 < 2e-16
rx2          3.9955     3.9076  1.02   0.307
age         -0.0717     0.0174 -4.12 3.8e-05
rx2:age     -0.0565     0.0662 -0.85   0.393
Log(scale)  -0.3844     0.2191 -1.75   0.079

Scale= 0.681

Log Normal distribution
Loglik(model)= -87.2   Loglik(intercept only)= -97.1
    Chisq= 19.8 on 3 degrees of freedom, p= 0.00019
Number of Newton-Raphson Iterations: 6
n= 26
```

Se observa que la interacción en el modelo no es significativa, por lo tanto, los efectos principales tanto de la edad como rx son estadísticamente validos.

Evaluando un posible efecto de confusión con la variable resid.ds

```
#Confounding

resultConf <- survreg(Surv(futime, fustat) ~ rx + age + resid.ds, dist = "lognormal", data
         = base)
summary(resultConf)
```

```
Call:
survreg(formula = Surv(futime, fustat) ~ rx + age + resid.ds,
    data = base, dist = "lognormal")
              Value Std. Error     z       p
(Intercept) 10.3867     0.9756 10.65 < 2e-16
rx2          0.6030     0.2984  2.02   0.043
age         -0.0677     0.0161 -4.20 2.6e-05
resid.ds2   -0.5258     0.3288 -1.60   0.110
Log(scale)  -0.4609     0.2178 -2.12   0.034

Scale= 0.631

Log Normal distribution
Loglik(model)= -86.4   Loglik(intercept only)= -97.1
    Chisq= 21.42 on 3 degrees of freedom, p= 8.6e-05
Number of Newton-Raphson Iterations: 6
n= 26
```

Resumen del modelo escogido

```
summary(result3_logn)
```

```
Call:
survreg(formula = Surv(futime, fustat) ~ rx + age, data = base,
    dist = "lognormal")
              Value Std. Error     z       p
(Intercept) 10.5449     1.0482 10.06 < 2e-16
rx2          0.6904     0.3169  2.18   0.029
age         -0.0765     0.0171 -4.48 7.4e-06
Log(scale)  -0.3813     0.2189 -1.74   0.082

Scale= 0.683
```

```
Log Normal distribution
Loglik(model)= -87.7   Loglik(intercept only)= -97.1
    Chisq= 18.93 on 2 degrees of freedom, p= 7.8e-05
Number of Newton-Raphson Iterations: 6
n= 26
```

Evaluando confusión.

```
#To evaluate confounding use abs((beta_unadjusted-beta_adjusted)/beta_unadjusted)

rxCon <- abs((0.6904 - 0.6030)/0.6904)*100
ageCon <- abs((-0.0765 - -0.0677)/-0.0765)*100

rxCon
```

```
[1] 12.65933
```

```
ageCon
```

```
[1] 11.50327
```

Puesto que el cambio porcentual en las estimaciones de ambos parámetros después de considerar el efecto de **resid.ds** es mayor al 10%, se puede argumentar que la variable resid.ds es una variable de confusión y por lo tanto, debe ser incluida en el modelo.

Interpretación del modelo seleccionado.

```
summary(resultConf)
```

```
Call:
survreg(formula = Surv(futime, fustat) ~ rx + age + resid.ds,
    data = base, dist = "lognormal")
             Value Std. Error     z       p
(Intercept) 10.3867     0.9756 10.65 < 2e-16
rx2          0.6030     0.2984  2.02   0.043
age         -0.0677     0.0161 -4.20 2.6e-05
```

```
resid.ds2   -0.5258     0.3288 -1.60   0.110
Log(scale)  -0.4609     0.2178 -2.12   0.034

Scale= 0.631

Log Normal distribution
Loglik(model)= -86.4   Loglik(intercept only)= -97.1
    Chisq= 21.42 on 3 degrees of freedom, p= 8.6e-05
Number of Newton-Raphson Iterations: 6
n= 26
```

Del resumen se observa que el modelo lognormal es estadísticamente significativo para explicar el tiempo de supervivencia de las mujeres con cáncer de ovario, el signo positivo de la estimación del coeficiente asociado a **rx** indica que pertenecer al grupo 2, o sea, el segundo tratamiento representa un aumento en el tiempo de supervivencia, similarmente, para la variable **edad**, la estimación del beta es negativo, lo que indica es que a mayor edad el tiempo de supervivencia disminuye, por ultimo, para la variable **resid.ds** el coeficiente es negativo, indicando así que pertenecer al segundo grupo disminuye la supervivencia, es decir, si la enfermedad persiste el tiempo de supervivencia disminuye.

Punto 2

Los siguientes datos se refieren a dos grupos de mujeres con cáncer de ovario.

a) Ajuste un modelo Weibull a estos datos considerando como variable independiente la variable 'grupo' que en este caso tiene dos niveles: 1 y 2. Usando las respectivas estimaciones Grafique la función hazard

```
base2 <- read.csv2("Tumor.csv")
base2$Grupo <- as.factor(base2$Grupo)

#Implementando un modelo de regresion Weibull

r2 <- survreg(Surv(Tiempo, Status) ~ Grupo, dist = "weibull", data = base2)
summary(r2)
```

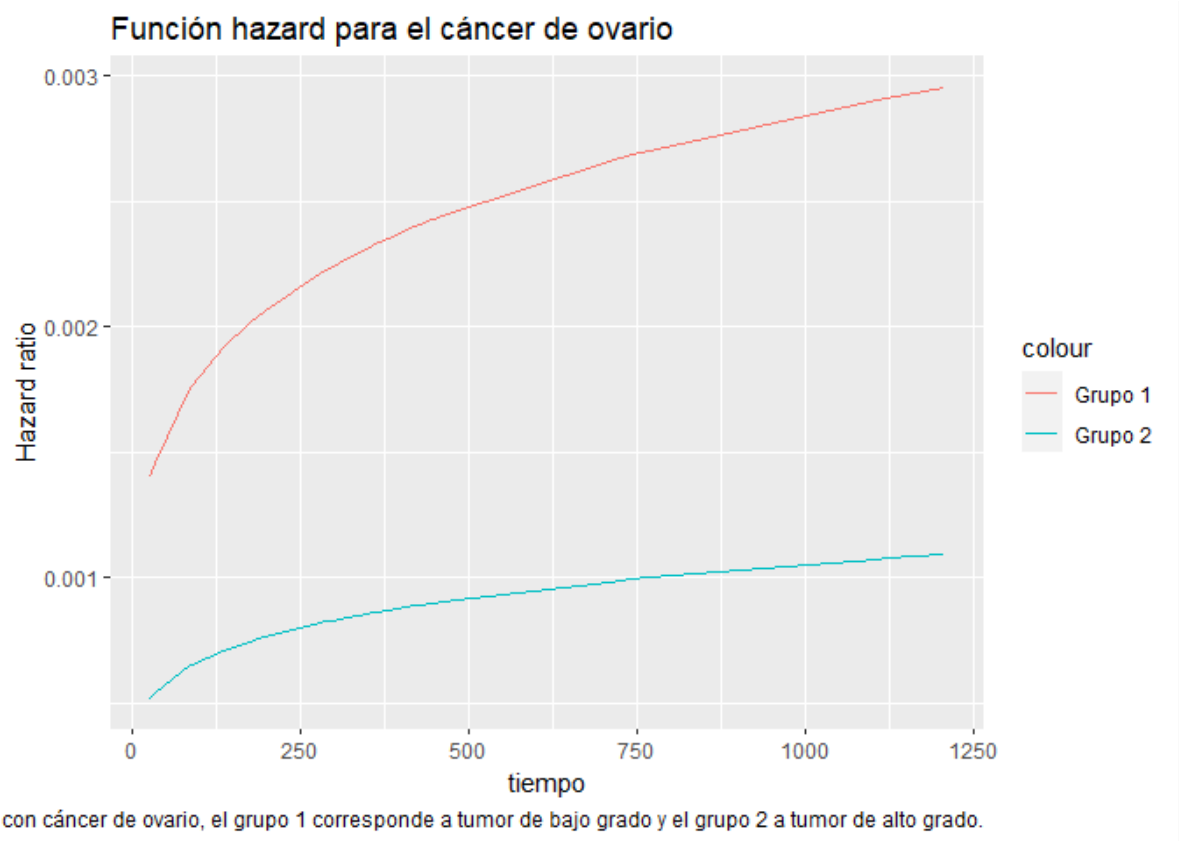
```
Call:
survreg(formula = Surv(Tiempo, Status) ~ Grupo, data = base2,
    dist = "weibull")
             Value Std. Error     z      p
(Intercept)  7.015      0.354 19.81 <2e-16
Grupo2      -0.829      0.418 -1.98  0.047
```

```
Log(scale)  -0.181      0.191 -0.95  0.343

Scale= 0.834

Weibull distribution
Loglik(model)= -142.1   Loglik(intercept only)= -144.2
    Chisq= 4.22 on 1 degrees of freedom, p= 0.04
Number of Newton-Raphson Iterations: 5
n= 34
```

Se observa que el modelo es estadísticamente significativo y que la variable grupo afecta el tiempo de supervivencia.

Grafico de la función hazard

```
require(flexsurv)

flexgg <- flexsurvreg(Surv(Tiempo,Status) ~ Grupo, dist="weibull", data=base2) %>%
    summary(type = "hazard") %>% data.frame

ggplot() + geom_line(aes(x = Grupo.1.time, y = Grupo.1.est, col = "red"), data = flexgg) +
  geom_line(aes(x = Grupo.2.time, y = Grupo.2.est, col = "blue"), data = flexgg) +
  labs(title='Función hazard para el cáncer de ovario',
       caption='Mujeres con cáncer de ovario, el grupo 1 corresponde a tumor de bajo grado
         y el grupo 2 a tumor de alto grado.', x = 'tiempo', y = 'Hazard ratio') +
  scale_color_hue(labels = c("Grupo 1", "Grupo 2"))
```



con cáncer de ovario, el grupo 1 corresponde a tumor de bajo grado y el grupo 2 a tumor de alto grado.

Del grafico anterior se observa que la función hazard estimada es creciente, se observa que crece más rapido en los primeros días, por lo tanto, la probabilidad de que falle aumenta más rápido en los primero días y luego sigue creciendo, pero de manera más suave, también se observa que el grupo 1 tiene mayor riesgo que el grupo 2.

b) Con estos datos y sin usar el log-rank test pruebe la hipótesis $H_0 : S_1(t) = S_2(t)$ para todo t. Use un $\alpha = 0,05$ ¿Qué se puede concluir? ¿Es plausible la conclusión?

Del resumen del modelo visto previamente se sabe que el coeficiente estimado para la variable grupo es de -0.829 unidades en el logaritmo del tiempo de supervivencia, es decir estar en el grupo 2, disminuye en -0.829 unidades el logaritmo de la supervivencia, para probar que pertenecer a un grupo u otro afecta el tiempo de supervivencia se construirá un intervalo de confianza.

Entonces queremos probar que:

$$H_0 : S_1(t) = S_2(t)$$
$$H_1 : S_1(t) \neq S_2(t)$$

Para realizarlo se construira un intervalo de confianza

$$\hat{\beta}_j \pm z_{1-\alpha/2}\widehat{SE(\hat{B}_j)}$$
$$Luego:$$
$$-0.829 \pm z_{0.975}0.418$$
$$-0.829 \pm 1.959964 \times 0.418$$
$$-0.829 \pm 0.819265$$
$$(-1.648265,\ -0.009735)$$

Como el intervalo no contiene al cero, entonces, con una confianza del 95% se muestra que hay suficiente evidencia para rechazar H_0, es decir que la supervivencia del grupo 1 es diferente a la supervivencia del grupo 2, por lo tanto, pertenecer a un grupo u otro, o sea, tener un tumor de bajo grado presenta un tiempo de supervivencia diferente a uno de alto grado.

Esta conclusión es bastante lógica, ya que cuando una persona tiene un tumor de alto grado se espera que sus consecuencias medicas sean más severas que cuando el tumor es de bajo grado.

Punto 3

Suponga que una unidad puede fallar por dos causas. Por ejemplo, una persona puede morir ya sea por una enfermedad del corazón o por una enfermedad de riñón. Sean

Y_1 = Tiempo de supervivencia con enfermedad del corazón

Y_2 = Tiempo de supervivencia con enfermedad del riñón

La pregunta de interés es ¿Cuál de las dos enfermedades causarán primero la muerte a un paciente? Sea $T = mín(Y_1, Y_2)$. Suponga que $Y_1 \sim exp(\lambda_1)$ y $Y_2 \sim exp(\lambda_2)$ y que Y_1 y Y_2 son independientes. Bajo estas condiciones halle la f.d.p de la v.a. T. ¿Qué se puede decir del hazard de T?

Sea la función de densidad de probabilidad de una exponencial de una exponencial

$$f(t) = \lambda e^{-\lambda t}$$

Entonces la función conjunta de $T = mín(Y_1, Y_2)$ es

$$F(t) = P(min(Y1, Y2) < t) = 1 - P(min(Y1, Y2) > t)$$

$$F(t) = 1 - P(Y1 > t \wedge Y2 > t) = 1 - (P(Y1 > t) * P(Y2 > t))$$

$$F(t) = 1 - \left(\int_{-\infty}^{t} \lambda_1 e^{-\lambda_1 t} dt * \int_{-\infty}^{t} \lambda_2 e^{-\lambda_2 t} dt\right)$$

$$F(t) = 1 - \lambda_1\lambda_2 \left(\int_{-\infty}^{t} e^{-(\lambda_1+\lambda_2)t} dt\right)$$

$$F(t) = 1 - \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)t}$$

Entonces la f.d.p de la v.a. T es

$$f(t) = \frac{d}{dt}\left(1 - \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}e^{-(\lambda_1+\lambda_2)t}\right)$$

$$f(t) = \lambda_1\lambda_2 e^{-(\lambda_1+\lambda_2)t}$$

En términos de la función de Hazard la distribución conjunta encontrada se puede interpretar como que las enfermedades tienen un efecto aditivo, es decir, entre mas enfermedades tenga el paciente más probable es que fallezca temprano, las enfermedades entonces “compiten” para ver cual se manifiesta primero en el paciente.

Punto 4

Pruebe lo siguiente si log $T \sim normal(\mu, \sigma)$ entonces $T \sim lognormal(\mu, \sigma)$.

Pruebe lo siguiente si log $T \sim normal(\mu, \sigma)$ entonces $T \sim lognormal(\mu, \sigma)$.

Sea $Y = T$ entonces $x = ln(t)$ y $\frac{dx}{dy} = \frac{1}{x}$

Como $f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(t-\mu)^2}{2\sigma^2}}$ para $x > 0$, usando el método de cambio de variable

$$g(y) = f(ln(t))|(1/t)| = f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(ln(t)-\mu)^2}{2\sigma^2}}\frac{1}{t}$$

Entonces $T \sim lognormal(\mu, \sigma)$

Punto 5

Pruebe lo siguiente si $T \sim lognormal(\mu, \sigma)$ entonces log$T \sim lognormal(\mu, \sigma)$.

Tenemos que $y = log(t)$, implica que $t = e^y$ y por lo tanto, $\frac{dt}{dy} = e^y$. Como $f(t) = \frac{1}{\sqrt{2\pi}\sigma t}e^{-0.5\left(\frac{log(t-\mu)}{\sigma}\right)^2}$

Usando la técnica de cambio de variable

$$
\begin{aligned}
g(y) &= f(e^y)e^y \\
&= \frac{1}{\sqrt{2\pi}\sigma e^y} e^{-0.5\left(\frac{log(e^y-\mu)}{\sigma}\right)^2} e^y \\
&= \frac{1}{\sqrt{2\pi}\sigma} e^{-0.5\left(\frac{log(e^y-\mu)}{\sigma}\right)^2} \\
&= \underbrace{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{log(t-\mu)^2}{2b^2}}}_{\text{función de densidad de una normal}}
\end{aligned}
$$

Entonces $Y = log(T) \sim N(\mu, \sigma)$

Punto 6

Pruebe lo siguiente si $T \sim loglogis(\mu, \sigma)$ entonces log$T \sim logis(\mu, \sigma)$.

Punto 7

Pruebe lo siguiente si log$T \sim logis(\mu, \sigma)$ entonces $T \sim loglogis(\mu, \sigma)$.