# Tarea 6.

## Modelos de Supervivencia y de Series de Tiempo

### Ávila Argüello Carlos - 06

### Bonilla Cruz José Armando - 10

### Luna Gutiérrez Yanely - 51

### Reyes González Belén - 64

### Rivera Mata Dante Tristán - 67

Utiliza la base de datos de R llamada ovarian y realiza lo siguiente:

- 1. Realice un análisis descriptivo sobre el tiempo de supervivencia de los sujetos, ademńas de sus características generales y particulares. Obtenga la tabla de la estimación de Kaplan-Meier y grafique. En particular, grafique la función de supervivencia.

```
attach(ovarian)
summary(ovarian)
```

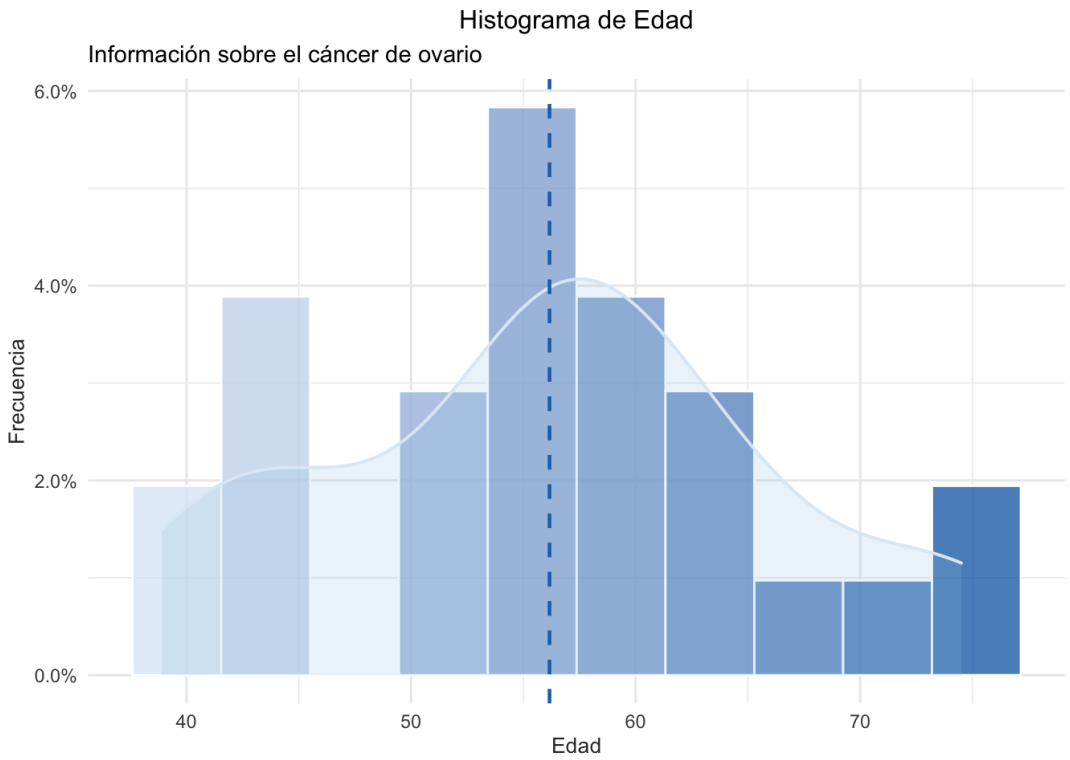
```
##      futime           fustat            age           resid.ds
##  Min.   :  59.0   Min.   :0.0000   Min.   :38.89   Min.   :1.000
##  1st Qu.: 368.0   1st Qu.:0.0000   1st Qu.:50.17   1st Qu.:1.000
##  Median : 476.0   Median :0.0000   Median :56.85   Median :2.000
##  Mean   : 599.5   Mean   :0.4615   Mean   :56.17   Mean   :1.577
##  3rd Qu.: 794.8   3rd Qu.:1.0000   3rd Qu.:62.38   3rd Qu.:2.000
##  Max.   :1227.0   Max.   :1.0000   Max.   :74.50   Max.   :2.000
##        rx         ecog.ps
##  Min.   :1.0   Min.   :1.000
##  1st Qu.:1.0   1st Qu.:1.000
##  Median :1.5   Median :1.000
##  Mean   :1.5   Mean   :1.462
##  3rd Qu.:2.0   3rd Qu.:2.000
##  Max.   :2.0   Max.   :2.000
```

La base de datos ‘ovarian’ nos dice el tiempo de supervivencia de una población que sufre cáncer de ovario. La descripción de la las componentes de la tabla es la que sigue:

La edad de los individuos se da en años. Columna: ‘age’.

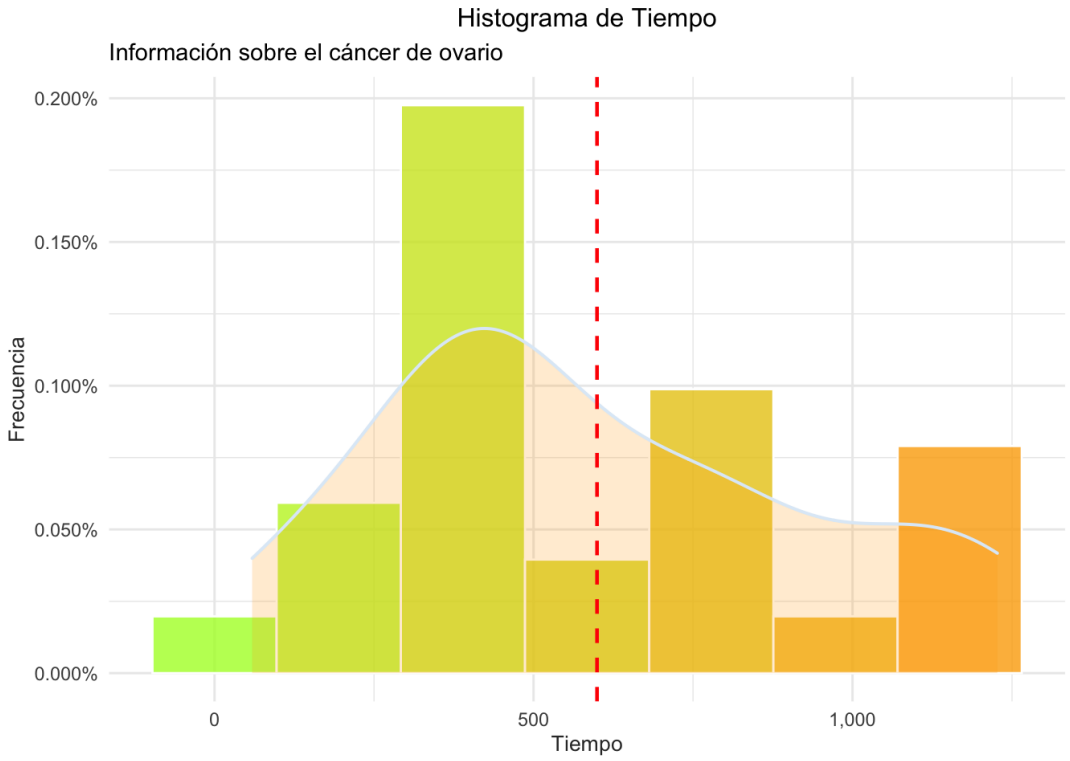
Si existieron secuelas a causa de la enfermedad se representa con uno (no) o dos (sí), esto se representa en la columna ‘resid.ds’. ECOG da la escala de mejoría (donde ‘1’ representa el mejor estado). ´rx´ dice el tipo de tratamiento usado. En la columna ‘fustat’ se dice si hubo falla o censura (1 ó 0 respectivamente). El tiempo se cuenta en días.

```
## Warning: Use of `cancer_ovario$age` is discouraged. Use `age` instead.
```

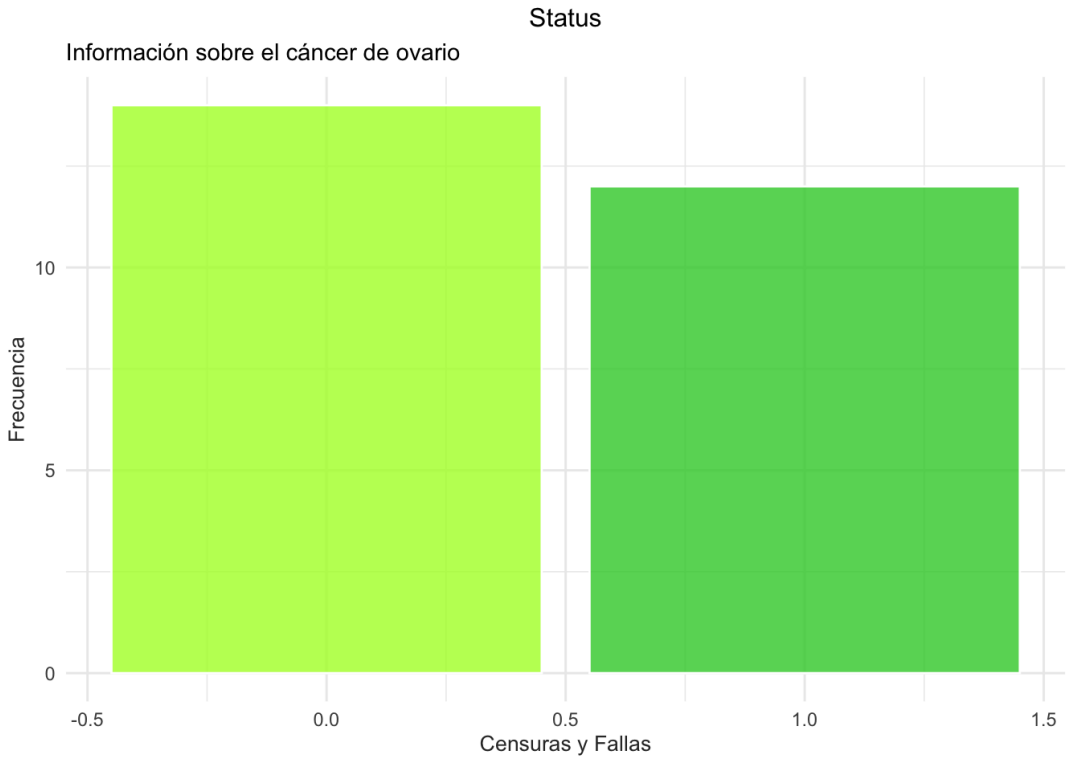



Notemos que el promedio de la edad es de 56 años aproximadamente, tiene tres modas y colas pesadas. La mayoría de la población oscila entre los 50 y 65 años de edad.

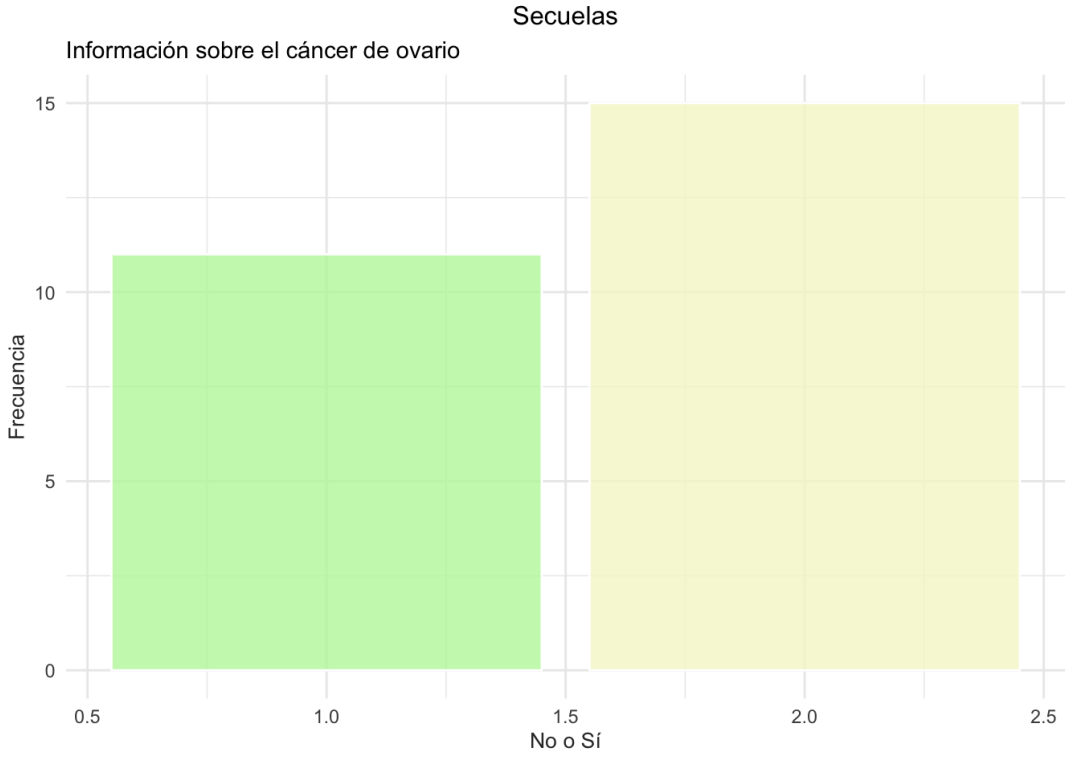
```
## Warning: Use of `cancer_ovario$futime` is discouraged. Use `futime` instead.
```



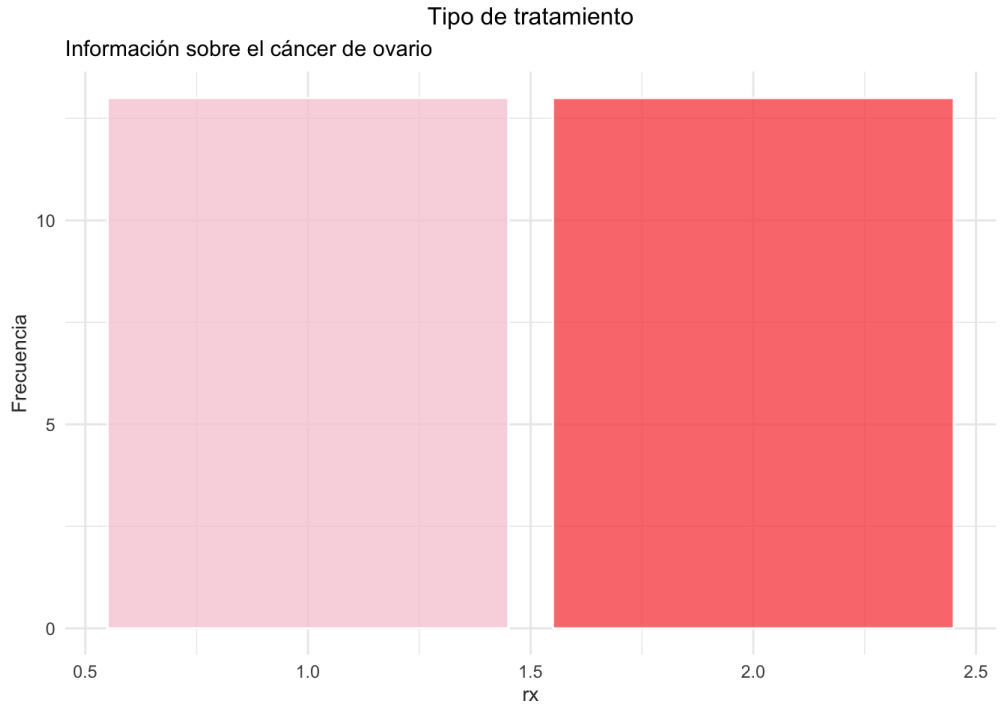

Hay un sesgo positivo dados los datos atípicos (que en su mayoría toman valores mayores al promedio) para las tiempos de supervivencia. Además parece que se presentan dos modas.



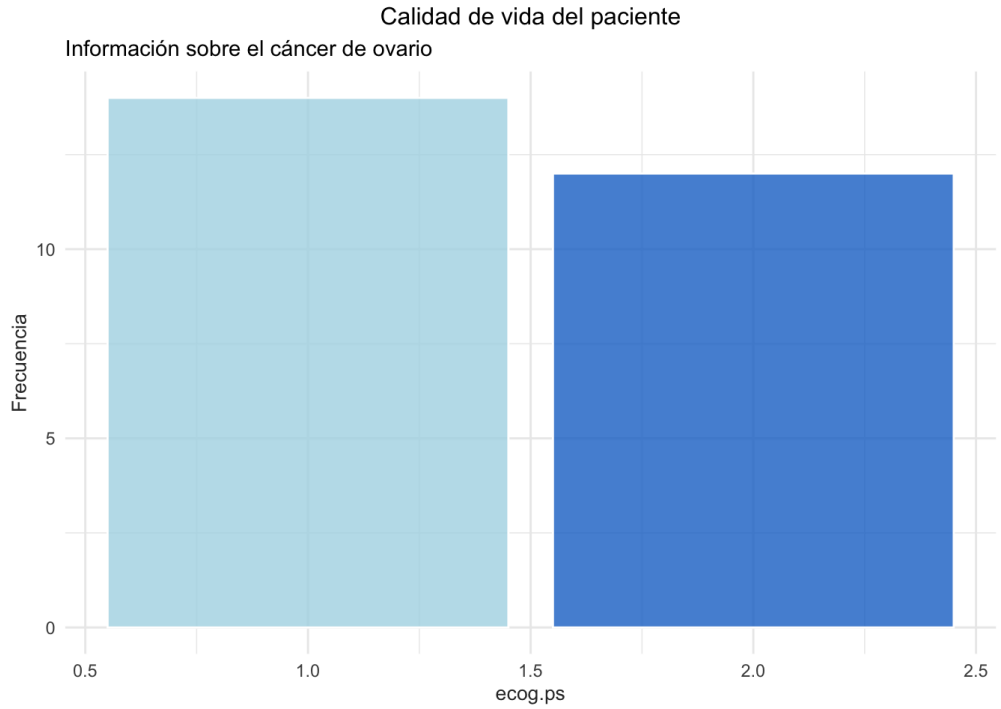

Notemos que hay mayor presencia de censuras que se pueden atribuir a los costos del tratamiento o del estudio, lo cual implica que la estimación sobre el tiempo de supervivencia puede verse afectada considerablemente.




Notemos que la mayoría de las personas en el estudio presentaron secuelas relacionadas con el tratamiento, esto puede dar pauta a estudios futuros y el análisis del tratamiento. Desde la perspetiva estadística, la función de supervivencia puede verse afectada.




Del gráfico anterior notemos que la muestra estudiada considera que el número de pacientes del tratamiento 1 y 2 es el mismo para cada uno.

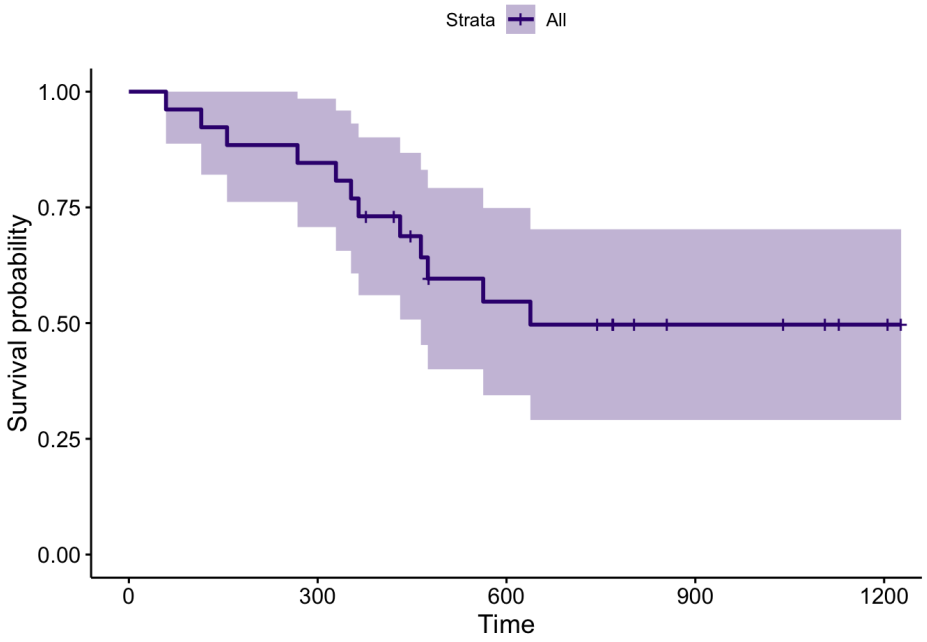



La mayoría de los pacienten tienen una escala de bienestar ‘1’, en esta escala las pacientes no tienen síntomas y pueden realizar todas sus actividades. Por otro lado, las pacientes pertenecientes a la escala ‘2’ no pueden realizar todas las actividades y presentan síntomas.

# Tabla Kaplan-Meier

```
fit_cancer_ovario<-survfit(Surv(ovarian$futime,ovarian$fustat)~1,type="kaplan-meier",
                conf.type='plain')
summary(fit_cancer_ovario)
```

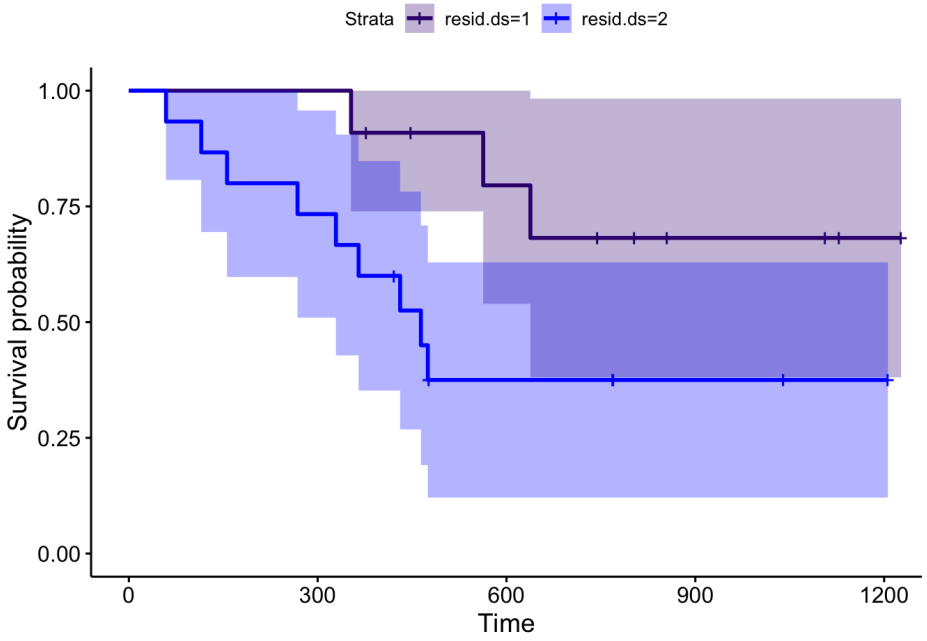
```
## Call: survfit(formula = Surv(ovarian$futime, ovarian$fustat) ~ 1, type = "kaplan-meier",
##     conf.type = "plain")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    59     26       1    0.962  0.0377        0.888        1.000
##   115     25       1    0.923  0.0523        0.821        1.000
##   156     24       1    0.885  0.0627        0.762        1.000
##   268     23       1    0.846  0.0708        0.707        0.985
##   329     22       1    0.808  0.0773        0.656        0.959
##   353     21       1    0.769  0.0826        0.607        0.931
##   365     20       1    0.731  0.0870        0.560        0.901
##   431     17       1    0.688  0.0919        0.508        0.868
##   464     15       1    0.642  0.0965        0.453        0.831
##   475     14       1    0.596  0.0999        0.400        0.792
##   563     12       1    0.546  0.1032        0.344        0.749
##   638     11       1    0.497  0.1051        0.291        0.703
```




La supervivencia tiene un decremento de tipo lineal hasta antes del día 600 del estudio y posteriormente tenemos una constante dado que la población remanente presenta censuras lo cual a su vez implica que la función de supervivencia estimada no toma el valor de cero.

- 2. Identifique las variables que afectan el tiempo de supervivencia, de manera exploratoria (estadísticas descriptivas y/o gráficas) y formal (usando intervalos de confianza y pruebas de hipótesis). En particular, pruebe si el tiempo de supervivencia es el mismo para ambos tratamientos. Use alguna prueba de hipótesis, con un nivel de confianza del 95%.

SECUELAS




De las gráficas de supervivencia podemos notar que las personas que presesentan secuelas (res.ds=2) tienen probabilidades mayores de presentar alguna falla. De los intervalos de confianza podemos observar que en los primeros días sí hay una diferencia significativa. Pero comprobemoslo con pruebas de hipótesis.

Procederemos con la prueba de long-rank

```
#H0:son iguales vs H1: son dsitintas
survdiff(Surv(cancer_ovario$futime, cancer_ovario$fustat)~cancer_ovario$resid.ds,rho=0)
```

```
## Call:
## survdiff(formula = Surv(cancer_ovario$futime, cancer_ovario$fustat) ~
##     cancer_ovario$resid.ds, rho = 0)
##
##                            N Observed Expected (O-E)^2/E (O-E)^2/V
## cancer_ovario$resid.ds=1 11        3     6.26      1.70      3.62
## cancer_ovario$resid.ds=2 15        9     5.74      1.85      3.62
##
##  Chisq= 3.6  on 1 degrees of freedom, p= 0.06
```

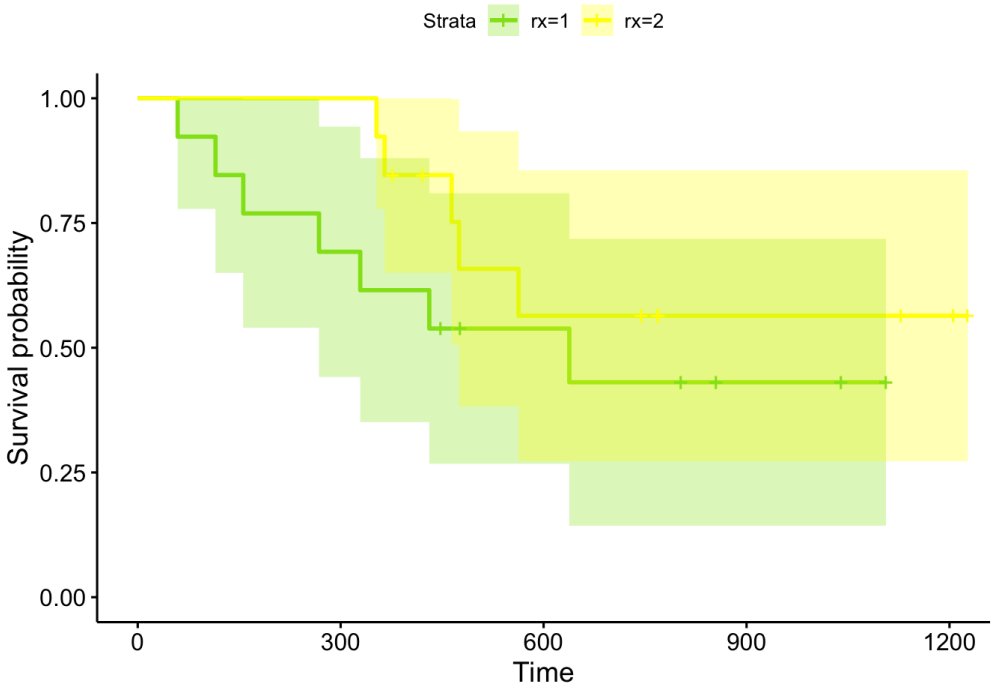
Ahora con la prueba de peto-peto

```
#H0:son iguales vs H1: son dsitintas
survdiff(Surv(cancer_ovario$futime, cancer_ovario$fustat)~cancer_ovario$resid.ds,rho=1)
```

```
## Call:
## survdiff(formula = Surv(cancer_ovario$futime, cancer_ovario$fustat) ~
##     cancer_ovario$resid.ds, rho = 1)
##
##                            N Observed Expected (O-E)^2/E (O-E)^2/V
## cancer_ovario$resid.ds=1 11     1.95     4.79      1.68      4.31
## cancer_ovario$resid.ds=2 15     7.45     4.61      1.75      4.31
##
##  Chisq= 4.3  on 1 degrees of freedom, p= 0.04
```

Al utilizar las dos pruebas estadísticas podemos notar que no son consistentes, ya que una rechaza con un nivel de significacia del $5\%$ y la otra lo acepta. Pero nos quedaremos con la prueba de peto-peto ya que gráficamente sí podiamos observar una diferencia significativa y esta prueba nos dice que $p - value < 0.05$ por lo que hay pruebas estadístias para rechazar H0 y aceptamos que la supervivencia es distitanta dado que sí hay secuelas o no.

TRATAMIENTO




De las gráficas de supervivencia podemos suponer que las personas con distintos tratamientos tienen la misma supervivencia, pero procederemos a hacer pruebas formales

Procederemos con la prueba de long-rank

```
#H0:son iguales vs H1: son dsitintas
survdiff(Surv(cancer_ovario$futime, cancer_ovario$fustat)~cancer_ovario$rx,rho=0)
```

```
## Call:
## survdiff(formula = Surv(cancer_ovario$futime, cancer_ovario$fustat) ~
##     cancer_ovario$rx, rho = 0)
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## cancer_ovario$rx=1 13        7     5.23     0.596      1.06
## cancer_ovario$rx=2 13        5     6.77     0.461      1.06
##
##  Chisq= 1.1  on 1 degrees of freedom, p= 0.3
```

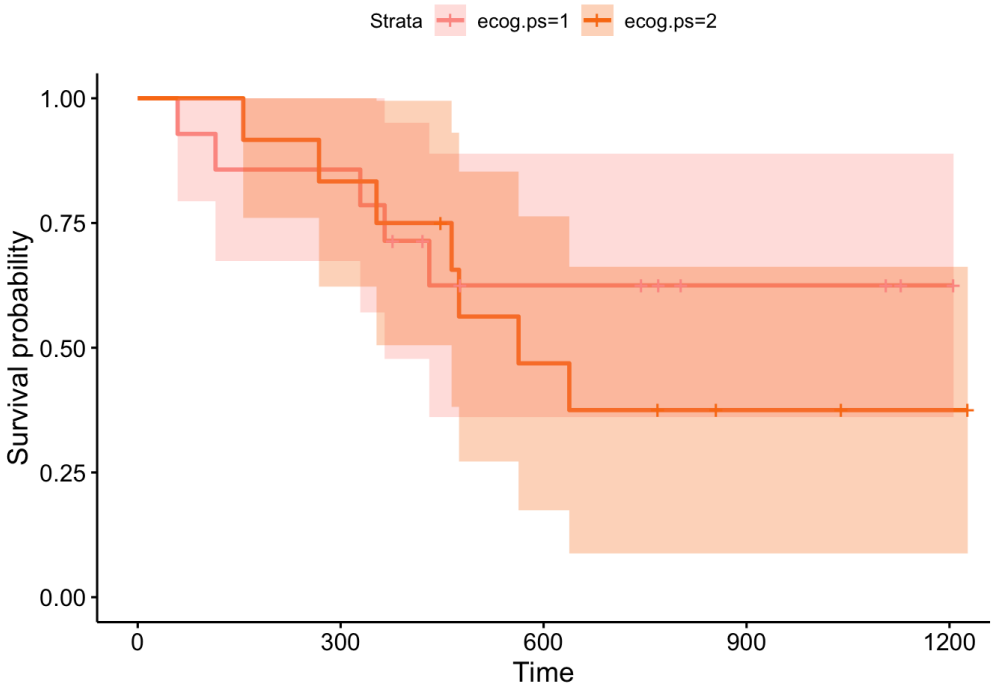
Ahora con la prueba de peto-peto

```
#H0:son iguales vs H1: son dsitintas
survdiff(Surv(cancer_ovario$futime, cancer_ovario$fustat)~cancer_ovario$rx,rho=1)
```

```
## Call:
## survdiff(formula = Surv(cancer_ovario$futime, cancer_ovario$fustat) ~
##     cancer_ovario$rx, rho = 1)
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## cancer_ovario$rx=1 13     5.89     4.12     0.761      1.68
## cancer_ovario$rx=2 13     3.50     5.27     0.595      1.68
##
##  Chisq= 1.7  on 1 degrees of freedom, p= 0.2
```

Ambas pruebas estadísticas son consistentes ya que se acepta con un nivel de significacia del $5\%$ i.e. $p-value > 0.05$ por lo que podemos decir que, no hay pruebas estadísticas suficientes para afirmar que hay una diferencia en la supervivecia respecto a los tratamientos

ECOG




De las gráficas de supervivencia podemos suponer que las personas con distintos tratamientos tienen la misma supervivencia, pero procederemos a hacer pruebas formales

Procederemos con la prueba de long-rank

```
#H0:son iguales vs H1: son dsitintas
survdiff(Surv(cancer_ovario$futime, cancer_ovario$fustat)~cancer_ovario$rx,rho=0)
```

```
## Call:
## survdiff(formula = Surv(cancer_ovario$futime, cancer_ovario$fustat) ~
##     cancer_ovario$rx, rho = 0)
##
##                       N Observed Expected (O-E)^2/E (O-E)^2/V
## cancer_ovario$rx=1 13         7     5.23     0.596      1.06
## cancer_ovario$rx=2 13         5     6.77     0.461      1.06
##
##  Chisq= 1.1  on 1 degrees of freedom, p= 0.3
```

Ahora con la prueba depeto-peto

```
#H0:son iguales vs H1: son dsitintas
survdiff(Surv(cancer_ovario$futime, cancer_ovario$fustat)~cancer_ovario$rx,rho=1)
```

```
## Call:
## survdiff(formula = Surv(cancer_ovario$futime, cancer_ovario$fustat) ~
##     cancer_ovario$rx, rho = 1)
##
##                       N Observed Expected (O-E)^2/E (O-E)^2/V
## cancer_ovario$rx=1 13     5.89     4.12     0.761      1.68
## cancer_ovario$rx=2 13     3.50     5.27     0.595      1.68
##
##  Chisq= 1.7  on 1 degrees of freedom, p= 0.2
```

Ambas pruebas estadísticas son consistentes ya que se acepta con un nivel de significacia del $5\%$ i.e. $p-value > 0.05$ por lo que podemos decir que, no hay pruebas estadísticas suficientes para afirmar que hay una diferencia en la supervivecia respecto a los tratamientos

- 3. Ajuste un modelo de riesgos proporcionales de Cox para definir la contribución de las variables al tiempo de supervivencia de las pacientes.

$$p-value < \alpha \iff \text{rechazar } H_0$$

```
#Primero observamos la significancia por cada covariable:
coxph(Surv(cancer_ovario$futime,cancer_ovario$fustat)~cancer_ovario$age)
```

```
## Call:
## coxph(formula = Surv(cancer_ovario$futime, cancer_ovario$fustat) ~
##     cancer_ovario$age)
##
##                       coef exp(coef) se(coef)     z       p
## cancer_ovario$age 0.16162   1.17541  0.04974 3.249 0.00116
##
## Likelihood ratio test=14.29  on 1 df, p=0.0001564
## n= 26, number of events= 12
```

```
#Como p-value es menor que alpha, entonces rechazamos la hipótesis nula que dice que "el coeficiente asociado al factor edad es cero (lo cual implicaría que no es segnificativo dicho facotr para el modelo de riesgos proporcio nales)" por lo que la edad sí es una covariable significativa para el modelo al nivel de significancia del 20%.
coxph(Surv(cancer_ovario$futime,cancer_ovario$fustat)~cancer_ovario$resid.ds)
```

```
## Call:
## coxph(formula = Surv(cancer_ovario$futime, cancer_ovario$fustat) ~
##     cancer_ovario$resid.ds)
##
##                            coef exp(coef) se(coef)     z      p
## cancer_ovario$resid.ds2 1.2092    3.3507   0.6724 1.798 0.0721
##
## Likelihood ratio test=3.76  on 1 df, p=0.05251
## n= 26, number of events= 12
```

```
#Sí es significativa la covariable "secuelas".

coxph(Surv(cancer_ovario$futime,cancer_ovario$fustat)~cancer_ovario$rx)
```

```
## Call:
## coxph(formula = Surv(cancer_ovario$futime, cancer_ovario$fustat) ~
##     cancer_ovario$rx)
##
##                       coef exp(coef) se(coef)      z    p
## cancer_ovario$rx2 -0.5964    0.5508   0.5870 -1.016 0.31
##
## Likelihood ratio test=1.05  on 1 df, p=0.3052
## n= 26, number of events= 12
```

```
#No es significativa la covariable "tipo de tratamiento".
coxph(Surv(cancer_ovario$futime,cancer_ovario$fustat)~cancer_ovario$ecog.ps)
```

```
## Call:
## coxph(formula = Surv(cancer_ovario$futime, cancer_ovario$fustat) ~
##     cancer_ovario$ecog.ps)
##
##                           coef exp(coef) se(coef)     z     p
## cancer_ovario$ecog.ps2 0.3984    1.4894   0.5864 0.679 0.497
##
## Likelihood ratio test=0.47  on 1 df, p=0.4935
## n= 26, number of events= 12
```

```
#No es significativa la covariable "calidad del paciente".
```

Por tanto podemos elegir el modelo de riesgios proporcionale que se compone de las covariables “edad” y “secuelas”. Descartamos las que no son significativas dado que la información que pueden proporcionar al modelo en érminos del riesgo es mínima en relación a el conjunto de las cuatro covariables.

```
riesgos_proporcionales<-coxph(Surv(cancer_ovario$futime,cancer_ovario$fustat)~ cancer_ovario$age + cancer_ovario
$resid.ds)
```

• ¿Cuál es la estimación puntual para los coeficientes de regresión? Interprete los coeficientes de regresión.

```
riesgos_proporcionales$coefficients
```

```
##       cancer_ovario$age cancer_ovario$resid.ds2
##               0.1445394               0.6141357
```

```
exp(riesgos_proporcionales$coefficients)
```

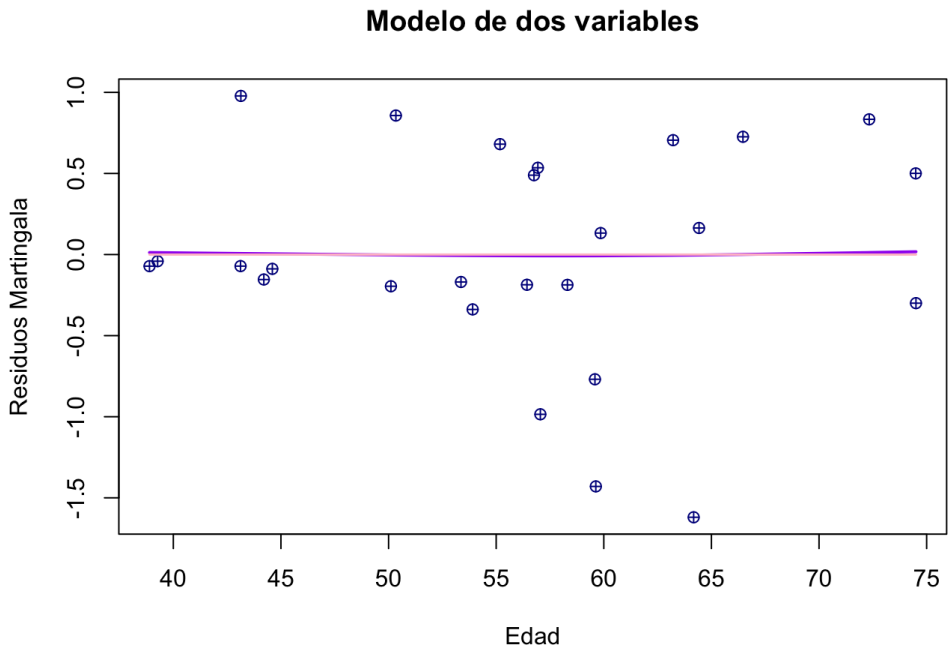
```
##       cancer_ovario$age cancer_ovario$resid.ds2
##                1.155507                1.848059
```

Si observamos el riesgo en un tiempo específico (dada $t_0 \geq 0$) para una persona de edad x, al aumentar la edad en un año (x+1), la función de riesgo (que corresponde al timepo de falla, muerte por cáncer de ovario) aumenta $15\%$ con respecto a la persona de edad x. Análogamente si se presentan secuelas en una individuo (en un tiempo fijo) aumenta la función de riesgo en un $84\%$ con respecto a la persona que no presenta secuelas.

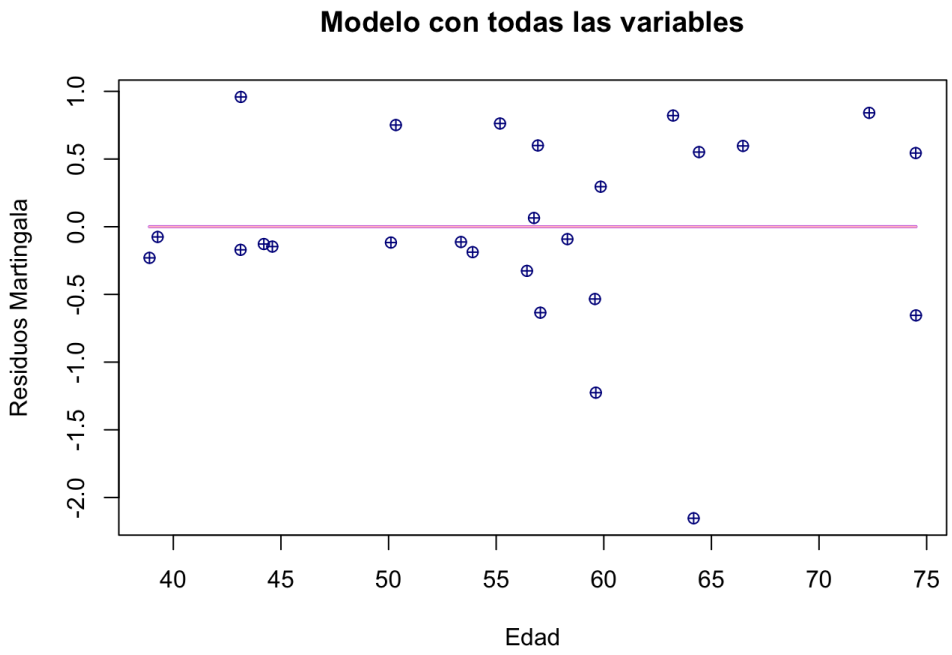
Lo anterior implica que, a grosso modo, a mayor edad se tiene mayor riesgo de falla y por ende menor tiempo de supervivencia. Y si se presentan secuelas (por la enfermedad) aumenta el riesgo de muerte por cáncer de ovario, lo que quiere decir que el tiempo de supervivencia para las pacientes que sufren secuelas disminuye.

• ¿Las variables explicativas tienen o no efecto en el modelo? Justifique.

```
## Call:
## coxph(formula = Surv(cancer_ovario$futime, cancer_ovario$fustat) ~
##     cancer_ovario$age + cancer_ovario$resid.ds)
##
##                            coef exp(coef) se(coef)     z       p
## cancer_ovario$age       0.14454   1.15551  0.05142 2.811 0.00494
## cancer_ovario$resid.ds2 0.61414   1.84806  0.73358 0.837 0.40249
##
## Likelihood ratio test=15.03  on 2 df, p=0.0005453
## n= 26, number of events= 12
```



```
## Call:
## coxph(formula = Surv(cancer_ovario$futime, cancer_ovario$fustat) ~
##     cancer_ovario$age + cancer_ovario$resid.ds + cancer_ovario$rx +
##         cancer_ovario$ecog.ps)
##
##                             coef exp(coef) se(coef)      z       p
## cancer_ovario$age         0.12481   1.13294  0.04689  2.662 0.00777
## cancer_ovario$resid.ds2   0.82619   2.28459  0.78961  1.046 0.29541
## cancer_ovario$rx2        -0.91450   0.40072  0.65332 -1.400 0.16158
## cancer_ovario$ecog.ps2    0.33621   1.39964  0.64392  0.522 0.60158
##
## Likelihood ratio test=17.04  on 4 df, p=0.001896
## n= 26, number of events= 12
```




Con el modelo de riesgos proporcionales usando dos variables (las que nosotros propusimos) podemos notar que la covariable que tiene una mayor significancia es la edad, lo cual sucede también en el modelo que consifera todas las variables de la base de datos original.

Por otro lado, podemos notar en las gráficas anteriores (de los residuos martinagalas) que para el modelo de dos covariables, los resiguos aproximan mejor a cero que en el modelo donde se consideran todas las covariables, lo cual implica que para la covariable edad se cumple mejor el supuesto de linealidad cuando solo hay dos covariables que cuando se consideran todas.

Por lo tanto las variables explicativas en el modelo que nosotros propusimos sí tienen efecto en el modelo.

• Obtenga un intervalo de confianza al 95% para la estimación de los coeficientes de regresión.

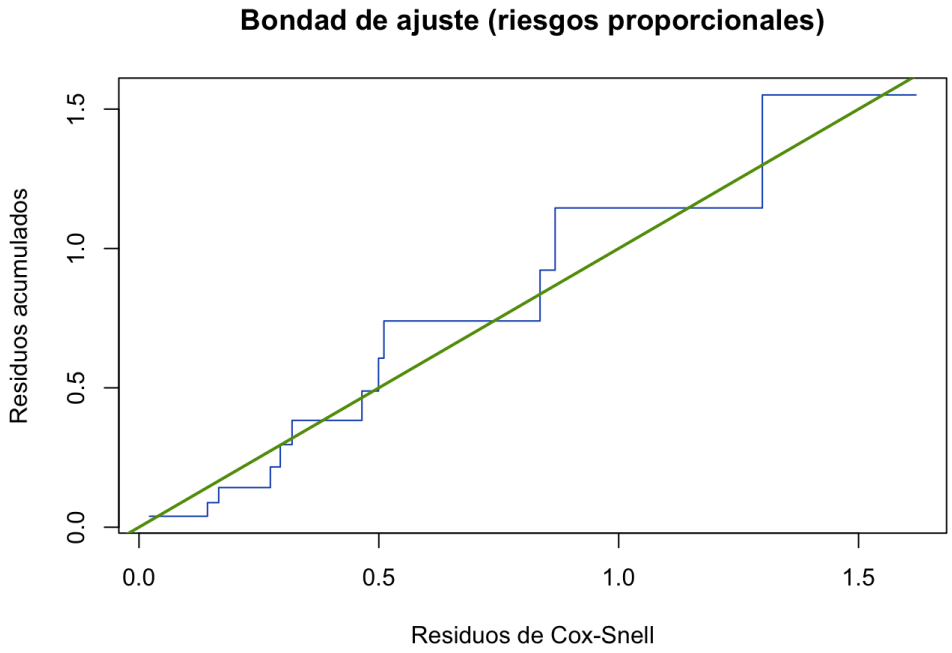
```
confint(riesgos_proporcionales)
```

```
##                               2.5 %    97.5 %
## cancer_ovario$age         0.04376539 0.2453135
## cancer_ovario$resid.ds2 -0.82365000 2.0519214
```

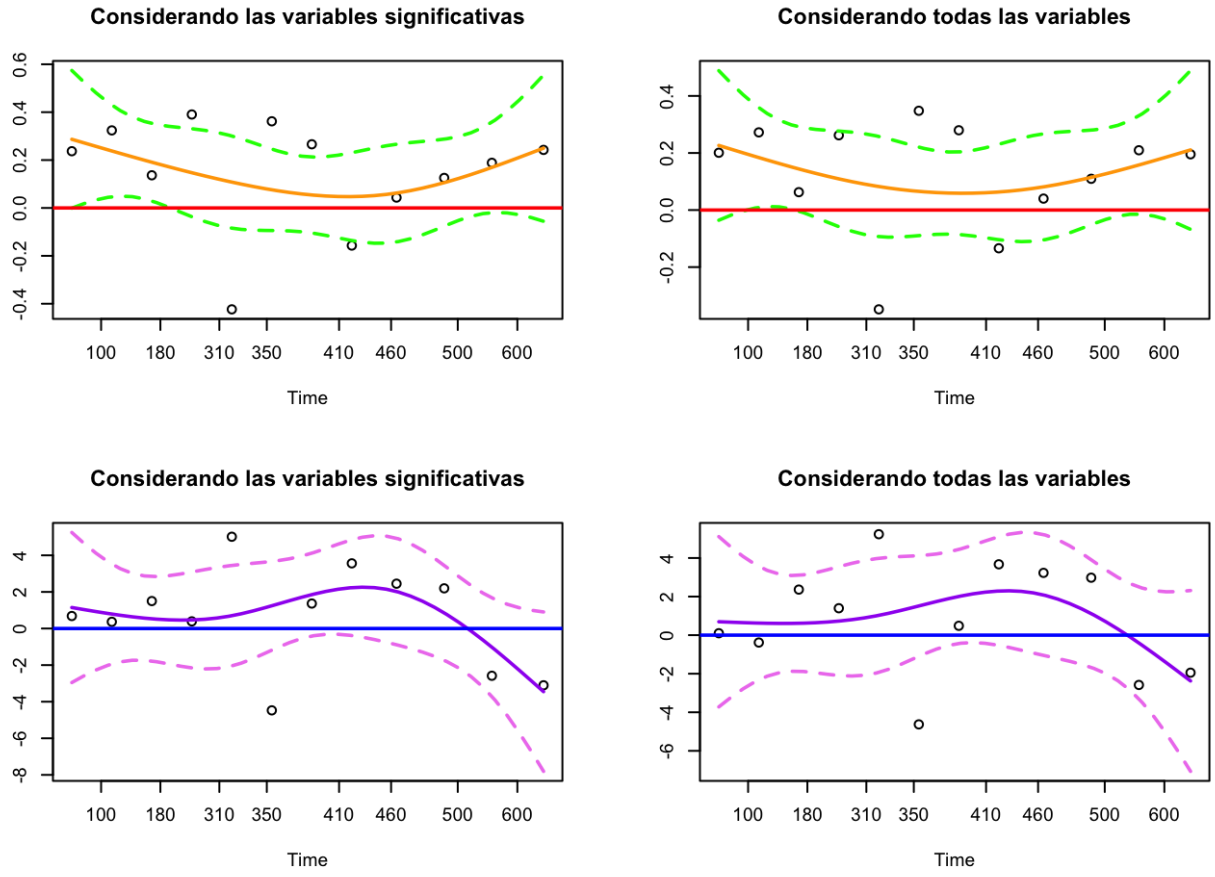
Notemos que la variable *secuelas* no es significativa ya que el cero pertenece al intervalo de confianza obtenido, mientras que para la *edad* sí hay significancia del parámetro.

• ¿Es válido tu modelo de acuerdo al supuesto de riesgos proporcionales? Realizas las pruebas y concluye.

Para contestar esta pregunta, consideremos lo siguiente:




Buscamos que la gráfica escalonada se asemeja lo más posible a la identidad, y como se puede observar al inicio de la gráfica la estimación se pega a la identidad en os primeros puntos, lo qe nos habla de un buen ajuste, sin embargo, como tenemos pocos datos, podemos observar que los residuos se alejan de forma considerable en las últimas observaciones.




Se observa que las variables significativas se “pegan”a las líneas horizontales, cumpliendo así con el supuesto de linealidad. Por otro lado, los puntos se dispersan de manera aleatoria lo que nos da a entender que se está cumpliendo el supuesto de riesgos proporcionales.

Para detectar datos atípicos necesitamos realizar los plots de los residuos de la devianza y de los residuos de dfbetas, sin embargo tenemos muy pocos datos, por lo que remover los datos atípicos podría afectar de forma significativa las estimaciones y el tamaño de la muestra.

Por las graficas anteriores podemos concluir que, a pesar de tener pocos datos, considerar las covariables *edad* y *secuelas* da un ajuste que cumple los supuestos de linealidad y de riesgos proporcionales.