Tarea 3: Análisis de Supervivencia

Miguel Ángel Beltrán Sánchez e Irene Extremera Serrano *

Especialización. Máster en Bioestadística.
Facultat de Ciències Matemàtiques. Universitat de València.

10 de noviembre de 2020

Este trabajo consiste en analizar datos sobre linfoma del sistema nervioso central procedentes de un ensayo clínico publicado en [1]. En efecto, se dispone de datos sobre 58 pacientes en los cuales se ha medido el tiempo de supervivencia, en años, desde el primer procedimiento BBBD[1] hasta el fallecimiento. Asimismo, también se dispone de información sobre si el paciente ha recibido radiación previa, el sexo, si la edad es superior a 60 años en el instante inicial y el llamado *Karnofsky score*.

El objetivo es realizar un análisis de supervivencia. Para ello, en primer lugar, se realizará una descriptiva numérica y gráfica del banco de datos. A continuación, se procederá a estimar las curvas de supervivencia y riesgo acumulado mediante curvas de Kaplan-Meier, diferenciando curvas según los diferentes factores medidos en el estudio. Posteriormente, se ajustará y encontrará el mejor modelo de regresión de Cox, sobre el que se realizará un diagnóstico junto al análisis de los residuos: hipótesis de riesgos proporcionales, bondad del ajuste y búsqueda de valores influyentes y anómalos. Finalmente, se interpretará el modelo y se extraerán las principales conclusiones.

1. Lectura y descriptiva

El banco de datos `linfoma.dat` está compuesto por variables cuantitativas y categóricas. Comentar que ha sido necesario convertir en factor a las distintas variables categóricas, para cualquier duda ver Anexo I: Código en `R`.

Con respecto a las variables categóricas, la variable sexo (`SEX`) está compuesta por dos categorías, la categoría de referencia que es hombre, codificado como "0" y compuesta por 38 sujetos, y la segunda que es mujer, codificado como "1" y compuesta por 20. Por otro lado, la variable grupo (`GROUP`) está compuesta por dos categorías, la de referencia que es el grupo sin tratamiento de radiación previo ("0") compuesta por 39 individuos y la segunda que sí ha tenido un tratamiento previo ("1"), que está compuesta solamente por 19 sujetos. En cuanto a las dos restantes, edad (`AGE60`) tiene como categoría de referencia las personas menores o iguales de 60 años ("0") y las mayores de 60 como segunda categoría ("1"), con 21 y 37

* *E-mail*: mianbel@alumni.uv.es y iexe@alumni.uv.es

[1] *Blood brain barrier disruption.*

individuos, respectivamente. Y, finalmente, el estado (`STATUS`) que como referencia tiene a los datos censurados del estudio ("0") y como segunda categoría las que han fallecido ("1"), o datos completos, con 22 y 36 pacientes, respectivamente.

A continuación, se presentan los diagramas de cajas pertenecientes a las variables continuas, esto es la Figura 1. A la izquierda puede apreciarse el diagrama de la variable *Karnofsky score* (`KPS.PRE.`). Este presenta el 75 % de sus valores entre 70 y 95, apreciándose una larga cola que llega hasta el valor 40 pero sin presencia de valores atípicos. Cabe mencionar que en el diagrama de la derecha perteneciente a la variable tiempo (`B3TODEATH`) desde el primer BBBD hasta muerte presenta una gran cantidad de valores atípicos (hasta 6) que superan los 5 años. La mayor concetración de valores se encuentra entre 1 y 3 años.

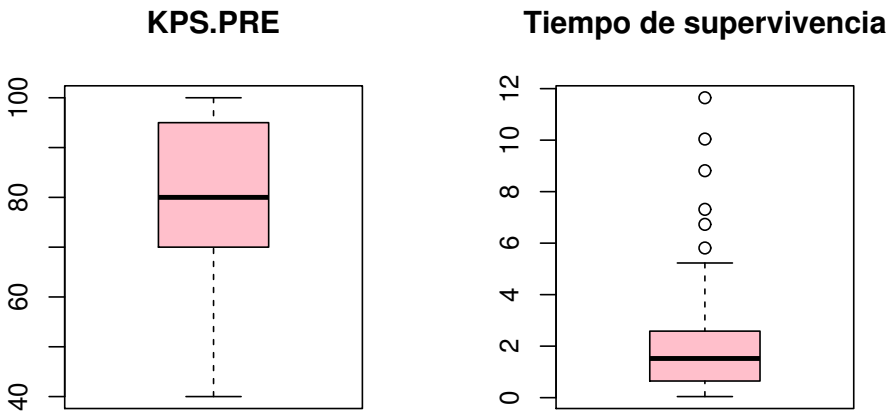

Figura 1: Diagrama de cajas de la variable `KPS.PRE.` (izquierda) y la variable tiempo `B3TODEATH` (derecha).

Continuando con la descriptiva, se puede apreciar en la Figura 2 un gráfico de nadadores, el cual es el más relevante a la hora de analizar la descriptiva de datos longitudinales y de supervivencia. En este gráfico se ha representado cada individuo y su tiempo desde el primer BBBD hasta la muerte. Cabe destacar que al final de cada línea aparece un punto coloreado o vacío que representa si el dato está censurado o no. Los puntos vacíos indican que es un dato censurado, es decir, ha llegado vivo al final del estudio o se ha perdido seguimiento. Por contra, los puntos rellenos indican que es un dato completo, esto significa que, efectivamente, ha fallecido en el estudio. Que haya datos censurados quiere decir que el sujeto en cuestión ha abandonado el experimento por causas ajenas a él, ya sea porque el estudio haya terminado o por exclusión del paciente del estudio.

Asimismo, se ha representado en colores según el grupo de tratamiento. En el grupo de radiados previamente, el individuo 5 y 16 son datos censurados, mientras que el resto son individuos que han fallecido (17 sujetos). En cuanto al grupo que no ha sido radiado con anterioridad, se aprecia que el número de datos censurados es mayor, concretamente de 20, mientras que la cantidad de fallecidos es de 19. Estas grandes diferencias en censurados de un grupo y otro puede ser determinante a la hora de extraer las conclusiones pertinentes.

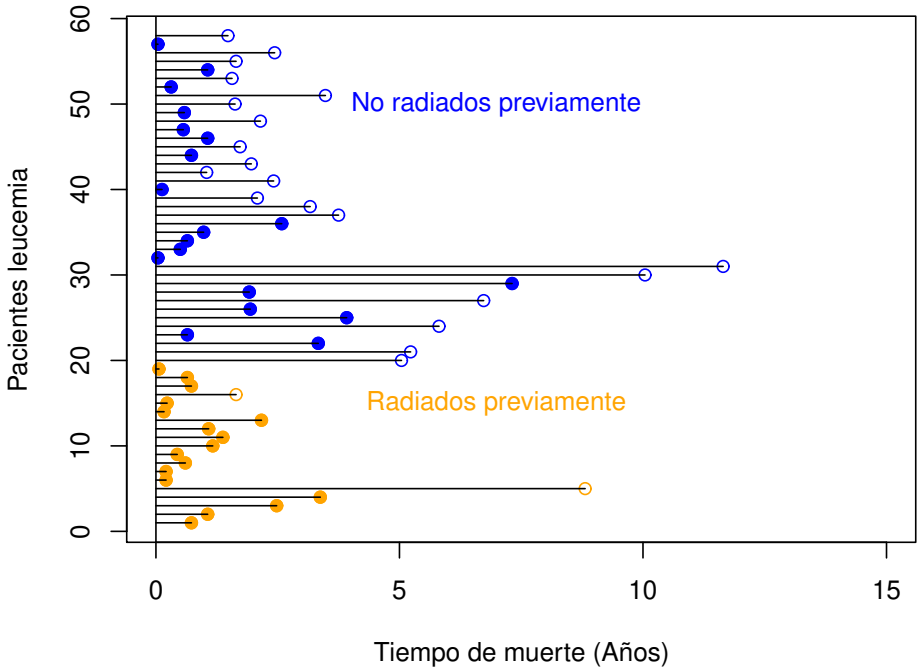

Figura 2: Gráfico de nadadores en donde aparece el tiempo de supervivencia de cada sujeto por grupos indicado por colores. Además se incluye si el dato es censurado o no mediante un punto relleno, dato completo, o vacío, dato censurado.

A continuación, se van a representar curvas de Kaplan-Meier, estimaciones de las funciones de supervivencia y riesgo acumulado. Esto permitirá, junto a diferentes test de comparativa de curvas, determinar si hay evidencia estadística suficiente para afirmar si hay diferencias en supervivencia entre los diferentes factores.

2. Curvas de Kaplan-Meier

Una vez realizada la descriptiva exhaustiva anterior se presenta uno de los principales objetivos que debe incluir un buen análisis de supervivencia, la estimación e interpretación de las funciones de supervivencia y riesgo acumulado.

Se define la *función de supervivencia* de una variable aleatoria T, que mide el tiempo de supervivencia desde un suceso inicial a uno final, como:

$$S(t) = 1 - F(t) = P(T > t),$$

es decir, la probabilidad de sobrevivir más tiempo en cada valor concreto del tiempo. Por otro lado, esta se relaciona con la *función de riesgo acumulado* de la siguiente forma:

$$H(t) = -\log S(t).$$

En la Figura 3 se han representado las estimaciones de la función de supervivencia y riesgo acumulado mediante las curvas de Kaplan-Meier. En efecto, en la izquierda se tiene la curva de

supervivencia de todos los pacientes con linfoma y a la derecha la curva de riesgo acumulado de todos estos mismos pacientes, es decir, sin distinguir en los factores del estudio. Se ha hecho uso de las funciones `survfit` y `Surv` del paquete `survival` para su obtención.

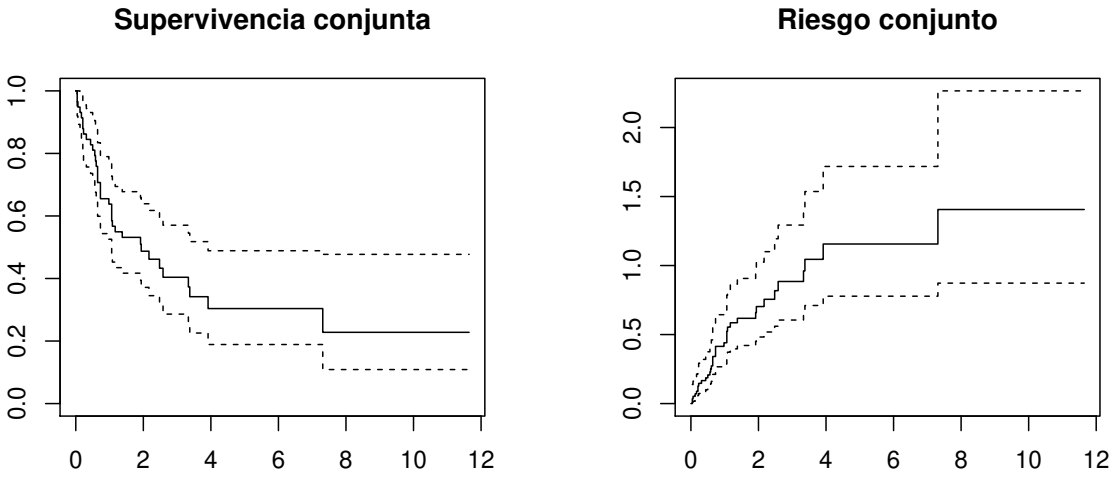

Figura 3: Curvas de Kaplan-Meier de supervivencia y riesgo acumulado de todos los pacientes con linfoma.

En la curva de supervivencia (izquierda) se observa, evidentemente, que a medida que pasan los años la probabilidad de seguir viviendo de las personas con linfoma se reduce considerablemente: se aprecia que la probabilidad en los primeros años decrece vertiginosamente de 0 hasta 4 años, pasando de una probabilidad de prácticamente 1 a 0.3. Además, parece que la probabilidad de fallecer toma un valor de 0.23 a partir de los 7 años mateniéndose estable, lo que indica que la probabilidad de supervivencia a partir de los 4 años es relativamente baja.

En cuanto a la curva de riesgo acumulado (derecha), se puede apreciar que el gráfico está mostrando cómo el riesgo de fallecer se acumula a medida que aumenta el periodo de tiempo. Se aprecia que crece muy rápidamente en los primeros años, tal y como se observaba en la función de supervivencia, debido a que el riesgo de fallecer es mucho mayor durante ese periodo. Sin embargo, a partir de aproximadamente los 4 años el riesgo permanece constante hasta pasados los 7 que se incrementa un poco.

Otro detalle a comentar y común en ambas curvas es el aumento del intervalo de confianza a medida que el tiempo transcurre. Esto es debido a que el número de individuos vivos en tiempos grandes es mucho menor que en periodos de tiempo más cortos.

Una vez visto cómo el es comportamiento de estas curvas con los datos conjuntos, a continuación se muestran las gráficas de Kaplan-Meier (Figura 4) correspondientes a las curvas de supervivencia con respecto a los factores `SEX`, `AGE60` y `GROUP`.

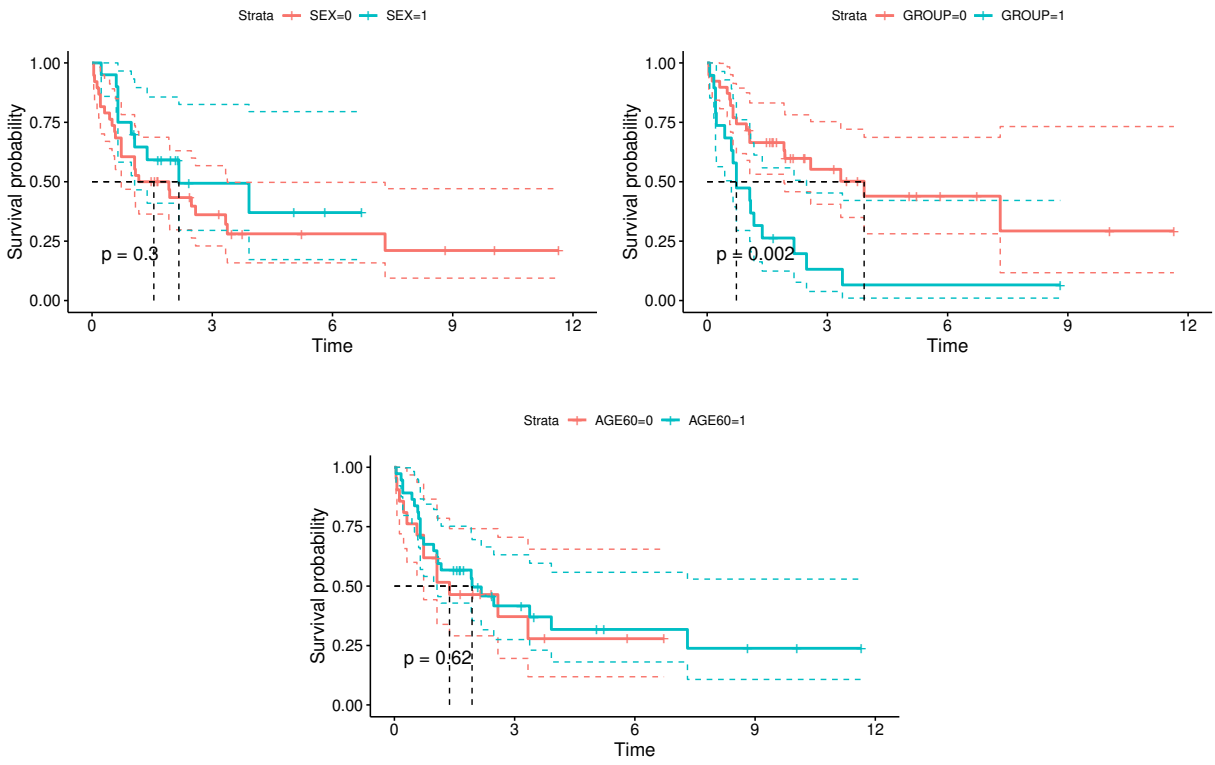

Figura 4: Curvas de supervivencia separadas por grupos de sexo, grupo de tratamiento y edad.

En la gráfica de arriba a la izquierda se aprecia la función de supervivencia correspondiente a los dos sexos. En rojo se representa la categoría de referencia hombres y en azul las mujeres. Parece que en mujeres la supervivencia es mayor que la de los hombres, aunque ligeramente.

En cuanto a la gráfica de arriba a la derecha, correspondiente al grupo de tratamiento, sí que se puede observar un espaciamiento más pronunciado entre el grupo radiado con anterioridad y el que no. Se observa que el grupo no radiado con anterioridad tiene una mayor supervivencia.

Finalmente, en la gráfica de supervivencia de abajo, correspondiente a los grupos de edad, parece que hay una muy ligera separación a partir de los 3 años, en donde los mayores de 60 tienen una mayor probabilidad de supervivencia.

Cabe destacar que al igual que en la gráfica de supervivencia conjunta, los intervalos de confianza son mayores a medida que se avanza en el tiempo. Dicho esto y con el objetivo de clarificar si efectivamente existen diferencias en la probabilidad de supervivencia en los grupos anteriores, se optó por realizar el test Log-Rank (Mantel-Haenszel) y el test de Breslow. Según ambos test, no existen diferencias significativas entre los grupos correspondientes a `AGE60` y `SEX` con unos p-valores resultantes del test Log-Rank y Breslow de 0.3 y 0.2 para `SEX` y de 0.6 y 0.5 para `AGE60`. Sin embargo, sí que se han encontrado diferencias significativas entre los grupos que han sido radiados o no con anterioridad, en donde los p-valores resultantes son de 0.002 en el test Log-Rank y de 0.005 en el de Breslow. Los p-valores del test Log-Rank aparecen también en las gráficas de la Figura 4.

3. Modelo de Cox

Hasta ahora se ha realizado una estimación e interpretación de la función de supervivencia y de riesgo acumulado mediante Kaplan-Meier. También se han comparado las funciones de supervivencia o de riesgo acumulado según el grupo de tratamiento, el sexo y el grupo de edad. En este apartado se pretende establecer la posible relación de las diferentes covariables con los tiempos de supervivencia.

En efecto, se modeliza la función de riesgo a partir de una estructura semiparamétrica:

$$h(t|x,\theta) = h_0(t)\exp\{\theta_1 x_1 + \theta_2 x_2 + \ldots, \theta_p x_p\},$$

donde t denota el tiempo, x el vector de predictores de y θ el vector de parámetros. Se asume que la función del tiempo $h_0(t)$, función de riesgo base, es la misma para cada individuo, por lo que se tiene la propiedad de riesgos proporcionales:

$$\frac{h(t|x,\theta)}{h(t|y,\theta)} = \exp\{\theta'(x-y)\}.$$

Mediante el uso de la función `coxph` del paquete `survival` se ajusta el modelo de Cox. Además, cada individuo tiene su propia estimación de la función de supervivencia, riesgo o riesgo acumulado. El procedimiento a seguir para la obtención del modelo ha sido el siguiente:

En primer lugar, se ha construido un modelo donde se incluyen todas las variables predictoras, esto es el grupo de tratamiento, el sexo, el grupo de edad y el *Karnofsky score*, incluyéndolas simplemente como efectos fijos, sin interacciones. Se ha obtenido que todas ellas son significativas a nivel $\alpha = 0{,}05$ salvo el grupo de edad, donde se ha obtenido un p-valor de 0.42 en el contraste de si el coeficiente asociado es nulo o no. El valor AIC para este modelo es 238.16.

Tras este primer ajuste, se ha valorado la posibilidad de otra relación funcional entre el tiempo de supervivencia y el *Karnofsky score*, única covariable numérica. En caso de no determinarse ninguna relación especial, todos los modelos a partir de este momento considerarán una relación lineal. En efecto, se ha determinado, y tal y como muestra el gráfico que enfrenta el *Karnofsky score* con los residuos martingala del modelo que excluye esta covariable numérica, que no existe una relación especial, esto puede verse en la Figura 5.

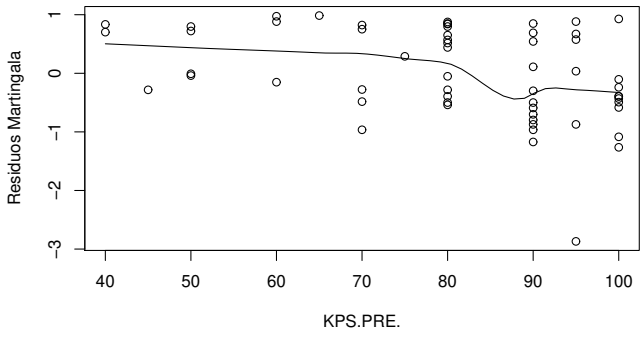

Figura 5: *Karnofsky score* vs. residuos martingala del modelo que excluye esta covariable.

A continuación, se ha valorado si se debería considerar alguno de los factores como estratos, es decir, una función no paramétrica del tiempo diferente para cada nivel del factor. Esta opción se ha rechazado pues el test de Therneau no rechaza los riesgos proporcionales. Gráficamente se observa en la Figura 6 suavizaciones aproximadamente horizontales aledeor de la media de los residuos escalados de Schoenfeld de cada variable predictora.

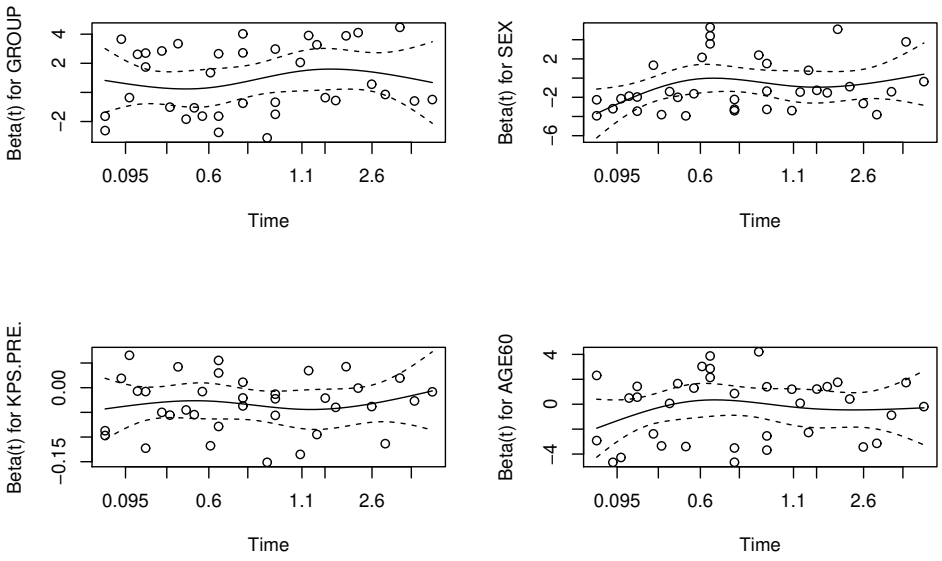

Figura 6: Representación gráfica del test de Therneau.

Además, con un test gráfico del logaritmo de la función de riesgo acumulado (`cloglog`), utilizando estratos para las variables categóricas, se observan estimaciones prácticamente paralelas (lo que significaría innecesario plantear estratos), esto puede apreciarse en la Figura 7. Sin embargo, considerando el estrato en grupo de tratamiento, las funciones del logaritmo de riesgo acumulado se cortan inicialmente. Haciendo un balance con el test de Therneau, se considera que no hay necesidad de incluir estratos.

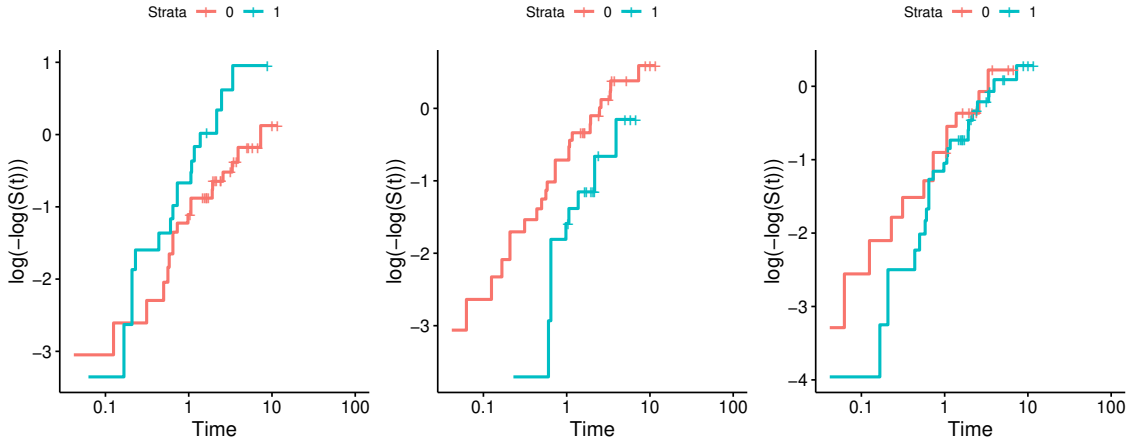

Figura 7: Logaritmo de la función de riesgo acumulado, `cloglog`, utilizando estratos, de izquierda a derecha: `GROUP`, `SEX` y `AGE60`.

Posteriormente, se ha utilizado la función `step` del paquete `stats` con la metodología *stepwise*

híbrida (`both`) para determinar si existe un mejor modelo que reduzca el AIC. Este modelo es aquel que elimina la variable del grupo de edad como predictor, la cual no era significativa. El AIC resultante es 236.81. Sin embargo, se ha planteado esta misma metodología pero incluyendo todas las interacciones dos a dos con tal de poder determinar si existe algún efecto potenciador entre variables y, efectivamente, el modelo resultante incluye todas las variables predictoras como efectos fijos y la interacción del sexo con el grupo de edad, siendo cada uno de ellos significativos a nivel $\alpha = 0{,}05$. El valor AIC resultante es 233.50, siendo este último modelo el que se ha determinado como el mejor. Por tanto, será este modelo sobre el que se realizará todo el diagnóstico: hipótesis de riesgos proporcionales, bondad del ajuste y análisis de los residuos. Además, en todas las selecciones de modelos anidados se ha realizado el contraste Ji-cuadrado de modelos anidades mediante la función `anova` del paquete `stats`.

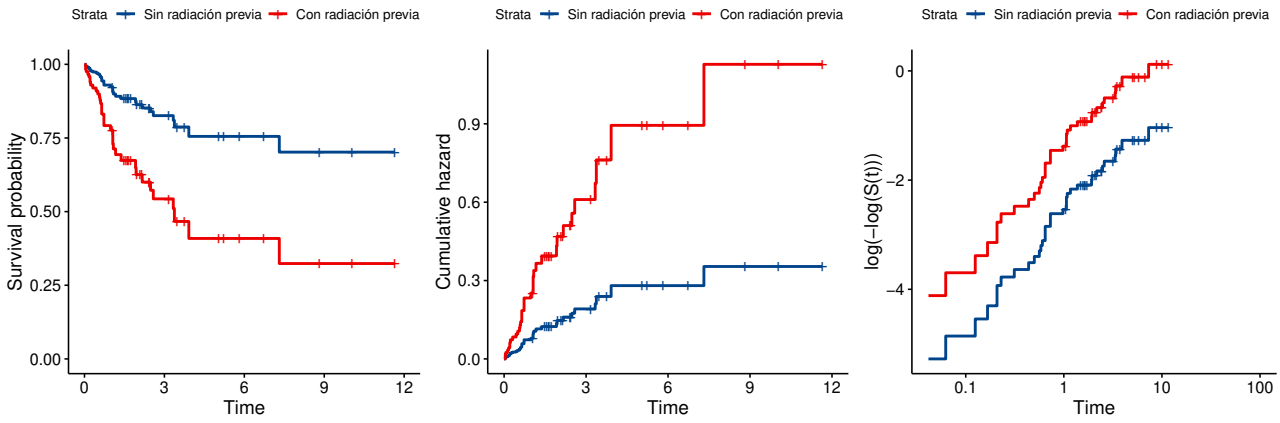

Figura 8: Ejemplo de estimaciones de la función de supervivencia, riesgo acumulado y su logaritmo.

Igual que en cualquier modelo de Cox, existe una curva de supervivencia ajustada para cada paciente. En la Figura 8 se han representado las estimaciones de las funciones de supervivencia, riesgo acumulado y su logaritmo para dos pacientes, uno sin radiación previa y otro con radiación previa, con sexo mujer, menor o igual a 60 años y con valor de *Karnofsky score* igual a su valor mediano, 80.

Es evidente que las estimaciones del logaritmo del riesgo acumulado son paralelas, pues se asume un modelo de Cox sin estratos, es decir riesgos proporcionales. A continuación, se realizará el diagnóstico del modelo y, por último, se procederá a su interpretación.

4. Diagnóstico del modelo

Una vez que se ha identificado cuál es el modelo definitivo de Cox, el considerado como el mejor, se procede al análisis de sus residuos con el objetivo de determinar si el modelo es o no adecuado: bondad del ajuste, valores influyentes y anómalos.

Antes de todo, se ha de comprobar de nuevo, mediante el test de Therneau, si se siguen teniendo riesgos proporcionales, condición de aplicabilidad del modelo de Cox. Efectivamente, se ha obtenido p-valores de 0.65 en `KPS.PRE.`, 0.30 en `GROUP`, 0.71 en `AGE60`, 0.15 en `SEX` y

0.20 en `AGE60:SEX`; todos ellos lo suficientemente grandes para no rechazar la hipótesis de riesgos proporcionales.

Además, se ha de valorar la bondad del juste mediante los residuos Cox-Snell y deviance. Primero se ha de comprobar que los residuos de Cox-Snell pueden aproximarse por una exponencial de parámetro 1 y, efectivamente, es lo que se observa en la Figura 9, al comprobar el gráfico QQ donde se comparan los cuantiles de los residuos Cox-Snell con los de una distribución exponencial de parámetro 1.

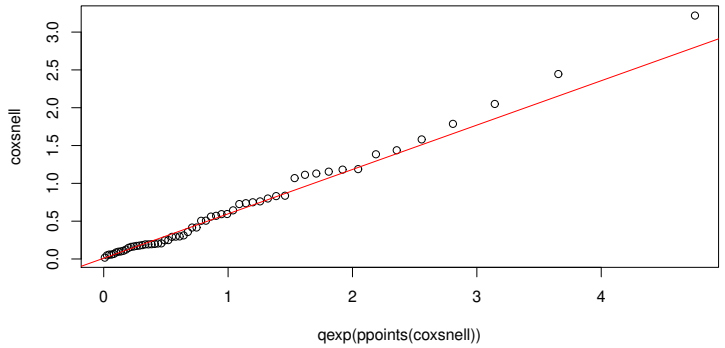

Figura 9: *QQ plot* para la bondad del ajuste de residuos Cox-Snell.

Asimismo, los residuos deviance parecen comportarse aproximadamente normales. Para ello se ha realizado el gráfico QQ presente en la Figura 10, donde se aprecia que el ajuste parece bueno. Además, para cerciorar esta condición de normalidad, también se ha realizado el test de normalidad de Kolmogorov-Smirnov con la función `ks.test` del paquete `stats`, dando un p valor de 0.46, por lo que no se rechaza la hipótesis nula de que los residuos deviance son aproximadamente normales.

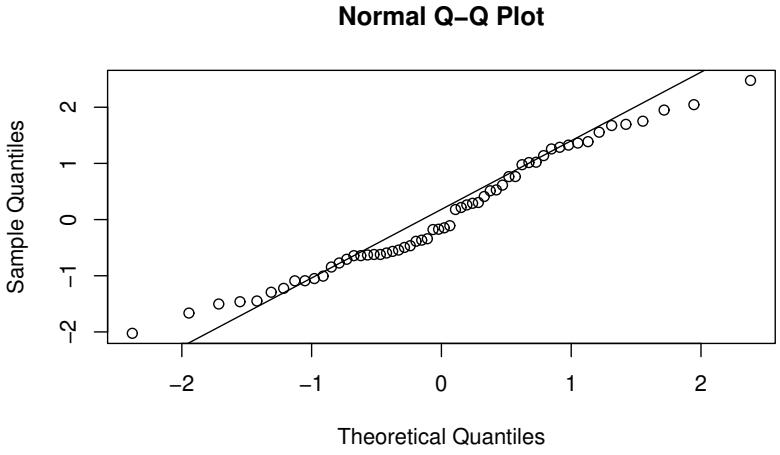

Figura 10: Gráfico QQ de residuos deviance para la condición de normalidad.

Asimismo, al representar la gráfica del tiempo frente a estos residuos deviance se observa una nube de puntos decreciente, por lo que las estimaciones de probabilidad de muerte en tiempos

grandes están sesgadas, sobreestimando la probabilidad de muerte, esto es la Figura 11. Sin embargo, no es tan problemático este hecho, lo sería si ocurriera en tiempos pequeños.

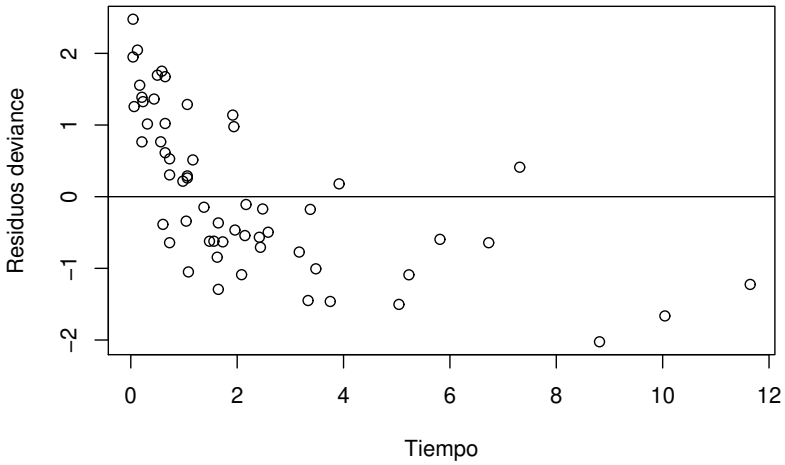

Figura 11: Residuos deviance frente al tiempo.

También se ha revisado la presencia de observaciones influyentes que pudiesen estar afectando al ajuste del modelo mediante los residuos dfbetas. A continuación, en la Figura 12 se aprecian las distintas observaciones y el valor de su residuo dfbeta correspondiente en cada variable.

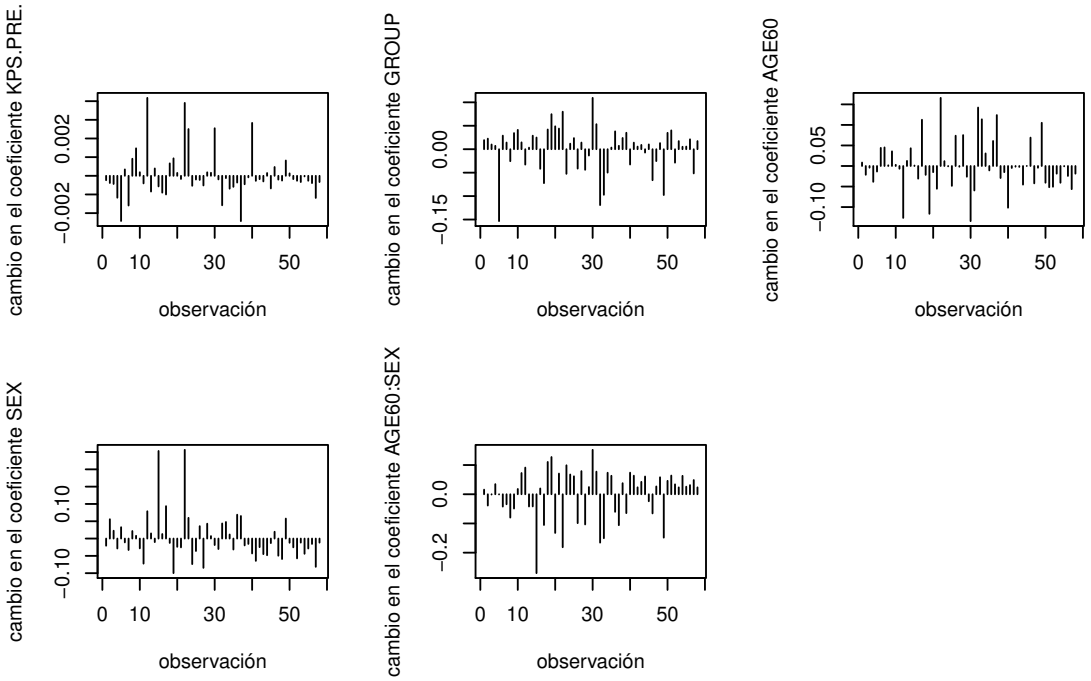

Figura 12: Residuos dfbetas correspondientes a las covariables incluidas en el modelo.

Se puede ver que los valores son bastante reducidos en los dfbetas a excepción de dos valores presentes en la gráfica de sexo, en donde hay dos valores que superan el 0.20, concretamente los sujetos 15, el cual también tiene un valor muy bajo del dfbeta en la edad, y 22. Asimismo,

el paciente 5 tiene un dfbeta muy bajo en grupo.

Una vez comprobado que hay observaciones que pueden ser influyentes y pueden estar afectando al ajuste del modelo, se ha comprobado si hay observaciones anómalas, las cuales no tienen por qué ser influyentes. Para ello se han mirado los residuos deviance y se ha comprobado si superan los valores de 2 y -2. El resultado fue la presencia de tres valores de -2.025, 2.48 y 2.046 correspondientes a los sujetos 5, 32 y 40, respectivamente. Cabe mencionar que dos de esas observaciones sobrepasan muy levemente el valor de 2 y -2, y solo es una la que toma un valor más o menos alto de 2.48 correspondiente al individuo 32.

Seguidamente se comprobó si había valores anómalos en las covariables, para ello se utilizaron los residuos parciales y se pudo comprobar que no parece haber ningún tipo de tendencia y se distribuyen de forma más o menos homogénea, como se aprecia claramente en la Figura 13.

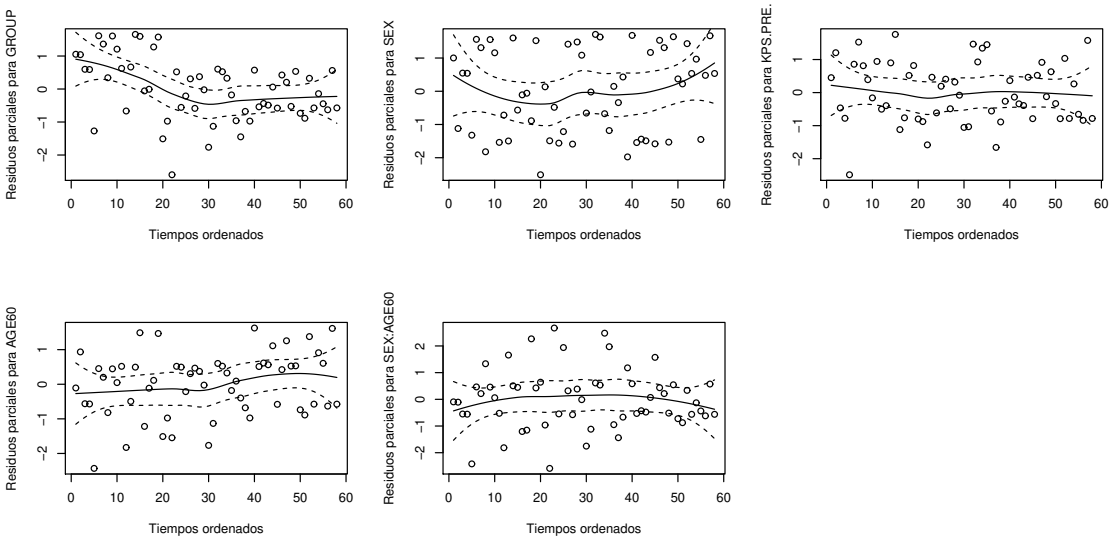

Figura 13: Residuos parciales correspondientes a los individuos y sus covariables.

Finalmente, se ha comprobado mediante el criterio VIF si existe algún tipo de problema de colinealidad entre las variables utilizadas. El valor VIF mayor correspondía a la interacción con un valor de 4, un valor bastante bajo, por lo que se ha concluido que no hay ningún problema de colinealidad. Por tanto, en estas circunstancias de que el modelo es correcto y ha pasado el diagnóstico, se procede a interpretar los resultados.

5. Interpretación y conclusiones

Una vez realizado el diagnóstico, es momento de interpretar cómo cambia el riesgo de supervivencia según el nivel de los factores y cuando se incrementa en una unidad la covariable *Karnofsky score*. En efecto, las estimaciones de los diferentes coeficientes asociados a las variables predictoras son el logaritmo del riesgo relativo de ese nivel del factor o, en el caso de *Karnofsky score*, el logaritmo del riesgo relativo de incrementar en unidad este puntaje, siempre manteniendo constantes el resto de variables predictoras. Se recalca que todas las interpretaciones que vienen a continuación sobre cambios en el riesgo a partir de una covariable

se tienen, en cualquier caso, *manteniendo el resto de covariables constantes*.

En el Cuadro 5 se tienen las exponenciales de los coeficientes, es decir, los riesgos relativos. En cuanto a la covariable de *Karnofsky score*, un incremento en una unidad de esta, manteniendo el resto constantes, implica que el riesgo se ve multiplicado por 0,97, lo que supone una ligera reducción de este, al ser menor que 1.

Covariable	GROUP1	SEX1	KPS.PRE.	AGE601	SEX1:AGE601
Riesgo	3.19	0.12	0.97	0.35	8.50

Cuadro 1: Riesgos relativos del mejor modelo.

Por otro lado, solamente por pertenecer al grupo de tratamiento con radiación previa, el riesgo se ve multiplicado por 3.19, esto es más del triple que el grupo sin radiación previa. En cuanto a cómo interviene el sexo y el grupo de edad, se han de distinguir varios casos.

En primer lugar, hombre (SEX=0) y edad menor o igual de 60 (AGE60=0). En este caso, el riesgo relativo, proporcionado exclusivamente por estas variables, es 1, pues todos los coeficientes asociados a efectos fijos e interacción son cero.

En segundo lugar, hombre (SEX=0) y edad mayor de 60 (AGE60=1). El riesgo relativo, proporcionado exclusivamente por estas variables, es 0.35, luego esta combinación de niveles de factores reduce el riesgo multiplicándolo por 0.35, esto es un 65 % menos de riesgo en hombres mayores de 60 años frente a hombres menores o iguales a 60.

En tercer lugar, mujer (SEX=1) y edad menor de 60 (AGE60=0). El riesgo relativo, proporcionado exclusivamente por estas variables, es 0.12, luego esta combinación de niveles de factores reduce el riesgo multiplicándolo por 0.12. Si comparamos el riesgo relativo de una mujer menor o igual a 60 años frente a un hombre de este grupo de edad, se tiene 0.12/1=0.12. Esto es un 88 % menos de riesgo en una mujer que en un hombre de edad menor o igual a 60.

En cuarto y último lugar, mujer (SEX=1) y edad mayor de 60 (AGE60=1). El riesgo relativo, proporcionado exclusivamente por estas variables, es 0.36 (=0.12·0.35·8.50), luego esta combinación de niveles de factores reduce el riesgo multiplicándolo por 0.36. Si comparamos el riesgo relativo de una mujer mayor de 60 años frente a una mujer menor o igual a 60, se tiene 0.36/0.12=3. Luego se estima que una mujer mayor de 60 años tiene un riesgo 3 veces mayor que una mujer menor o igual a 60, manteniendo el resto constante. Asimismo, una mujer de edad mayor de 60 tiene un riesgo relativo frente a un hombre de mayor de 60 de 0.36/0.35$\approx$ 1.03. Por tanto, el riesgo es ligeramente mayor en mujeres mayores de 60 que en hombres de este mismo rango de edad.

Cabe resaltar que las grandes diferencias en censurados de un nivel de un factor y otro, además de las diferencias en número de pacientes en cada nivel, puede haber afectado a la hora de extraer las conclusiones pertinentes.

Anexo I: Código de R

```
Paquetes y lectura de datos
pacman::p_load(survival, ggplot2, gridExtra, survminer, rms, xtable,
    install = FALSE)
linfoma <- read.csv("linfoma.dat", sep="")

linfoma$GROUP<-as.factor(linfoma$GROUP)
linfoma$SEX<-as.factor(linfoma$SEX)
linfoma$AGE60<-as.factor(linfoma$AGE60)
attach(linfoma)

#Descriptiva

par(mfrow=c(1,2))
boxplot(linfoma$KPS.PRE,col='PINK', main= 'KPS.PRE')
boxplot(linfoma$B3TODEATH,col='PINK', main= 'Tiempo de supervivencia
   ')

# Swimmer plot
color <- c('BLUE','ORANGE')
plot(B3TODEATH,1:length(B3TODEATH),xlim=c(0,15), xlab='Tiempo de
   muerte (Anyos)', ylab='Pacientes leucemia',col=color[GROUP], main
   ='Swimmer Plot')
points(B3TODEATH[STATUS==1],(1:length(B3TODEATH))[STATUS==1], pch
   =19,col=color[GROUP[STATUS==1]])
for (i in 1:length(B3TODEATH)) lines(c(0,B3TODEATH[i]),c(i,i))
abline(v=c(0,38),lty=c(1,2))
text(7,50,'No radiados previamente',col=color[1])
text(7,15,'Radiados previamente',col=color[2])

#Kaplan-Meier

km.linfoma <- survfit(Surv(B3TODEATH,STATUS)~1, data=linfoma) #
   Curvas de Kaplan-Meier

par(mfrow=c(1,2))
plot(km.linfoma,conf.int=TRUE, main='Curva de Supervivencia conjunta
   ') #supervivencia
plot(km.linfoma,fun = "cumhaz", conf.int=TRUE, main='Curva de Riesgo
   conjunta') #riesgo
summary(km.linfoma)

par(mfrow=c(1,3))
km.linfoma.sexo <- survfit(Surv(B3TODEATH,STATUS)~SEX, data=linfoma)
    # Curvas de Kaplan-Meier
ggsurvplot(km.linfoma.sexo, data = linfoma,
           conf.int = TRUE,
           conf.int.style = "step",
           surv.median.line = "hv",
```

```
            pval = TRUE) #p valor log rank

km.linfoma.grupo <- survfit(Surv(B3TODEATH,STATUS)~GROUP, data=
   linfoma) # Curvas de Kaplan-Meier
ggsurvplot(km.linfoma.grupo, data = linfoma,
            conf.int = TRUE,
            conf.int.style = "step",
            surv.median.line = "hv",
            pval = TRUE)

km.linfoma.edad <- survfit(Surv(B3TODEATH,STATUS)~AGE60, data=
   linfoma) # Curvas de Kaplan-Meier
ggsurvplot(km.linfoma.edad, data = linfoma,
            conf.int = TRUE,
            conf.int.style = "step",
            surv.median.line = "hv",
            pval = TRUE)

# SEXO
survdiff(Surv(B3TODEATH,STATUS)~SEX,
data=linfoma) # test log-rank
survdiff(Surv(B3TODEATH,STATUS)~SEX,
data=linfoma, rho=1) # test Breslow
# GRUPOS
survdiff(Surv(B3TODEATH,STATUS)~GROUP, data=linfoma) # test log-rank
survdiff(Surv(B3TODEATH,STATUS)~GROUP, data=linfoma, rho=1) # test
   Breslow
# EDAD
survdiff(Surv(B3TODEATH,STATUS)~AGE60, data=linfoma) # test log-rank
survdiff(Surv(B3TODEATH,STATUS)~AGE60, data=linfoma, rho=1) # test
   Breslow

#Cox
ajuste <- coxph(Surv(B3TODEATH,STATUS) ~ GROUP + SEX + KPS.PRE. +
   AGE60, data = linfoma)
summary(ajuste); AIC(ajuste)

#Estudiamos relaciones funcionales
ajuste1 <- coxph(Surv(B3TODEATH,STATUS) ~ GROUP + SEX + AGE60, data
   = linfoma)
scatter.smooth(linfoma$KPS.PRE., ajuste1$resid, type="p", xlab="KPS.
   PRE.",ylab="Residuos␣Martingala")

# Consideramos estratos?
PH.test <- cox.zph(ajuste); PH.test #no hay que poner estratos
par(mfrow=c(2,2)); plot(PH.test)

ajuste.strata <- coxph(Surv(B3TODEATH,STATUS) ~ strata(GROUP) + SEX
   + KPS.PRE. + AGE60, data = linfoma)
```

```
p1<-ggsurvplot(survfit(ajuste.strata), data = linfoma, fun = "
   cloglog")
ajuste.strata <- coxph(Surv(B3TODEATH,STATUS) ~ GROUP + strata(SEX)
   + KPS.PRE. + AGE60, data = linfoma)
p2<-ggsurvplot(survfit(ajuste.strata), data = linfoma, fun = "
   cloglog")
ajuste.strata <- coxph(Surv(B3TODEATH,STATUS) ~ GROUP + SEX + KPS.
   PRE. + strata(AGE60), data = linfoma)
p3<-ggsurvplot(survfit(ajuste.strata), data = linfoma, fun = "
   cloglog")
grid.arrange(p1$plot,p2$plot,p3$plot, ncol=3)

#Seleccion de modelos
ajuste.AIC <-step(ajuste,direction='both',trace=0)
anova(ajuste.AIC,ajuste)
ajuste <- coxph(Surv(B3TODEATH,STATUS) ~ (GROUP + SEX + KPS.PRE. +
   AGE60)^2, data = linfoma)
ajuste.AIC <-step(ajuste,direction='both',trace=0)
anova(ajuste.AIC,ajuste)

#MEJOR MODELO
ajuste.final<-coxph(Surv(B3TODEATH,STATUS) ~ GROUP + SEX + KPS.PRE.
   + AGE60 + SEX:AGE60, data = linfoma)
AIC(ajuste.final); summary(ajuste.final)

datos.nuevos <- cbind.data.frame(GROUP=c("0","1"), SEX="1",KPS.PRE.=
   median(linfoma$KPS.PRE.), AGE60="0")

p1<-ggsurvplot(survfit(ajuste.final,newdata = datos.nuevos), data =
   linfoma, palette = "lancet", conf.int = FALSE, legend.labs=c("Sin
   ␣radiacion␣previa", "Con␣radiacion␣previa"))

p2<-ggsurvplot(survfit(ajuste.final,newdata = datos.nuevos), data =
   linfoma, palette = "lancet", conf.int = FALSE, fun = "cumhaz",
   legend.labs=c("Sin␣radiacion␣previa", "Con␣radiacion␣previa"))

p3<-ggsurvplot(survfit(ajuste.final,newdata = datos.nuevos), data =
   linfoma, palette = "lancet", conf.int = FALSE, fun = "cloglog",
   legend.labs=c("Sin␣radiacion␣previa", "Con␣radiacion␣previa"))
grid.arrange(p1$plot,p2$plot,p3$plot, ncol=3)

#Diagnostico

PH.test <- cox.zph(ajuste.final)

martingala <- ajuste.final$residuals
coxsnell <- linfoma$STATUS - martingala
qqplot(qexp(ppoints(coxsnell), rate=1),coxsnell)
qqline(coxsnell, distribution = function(p) qexp(p), col=2)
```

```
res.deviance<-resid(ajuste.final,type="deviance")
qqnorm(res.deviance); qqline(res.deviance)
ks.test(res.deviance, "pnorm")
plot(linfoma$B3TODEATH,res.deviance,type='p',ylab='Residuos␣deviance
   ',xlab='Tiempo')
abline(h=0)


res.dfbeta <- resid(ajuste.final,type='dfbeta')
indice = 1:nrow(res.dfbeta)

par(mfrow=c(2,3))
for (i in 1:5) plot(indice,res.dfbeta[,i],type='h', xlab='
   observacion', ylab=paste('cambio␣en␣el␣coeficiente',(names(ajuste
   .final$coef))[i], sep = "␣"))
ggcoxdiagnostics(ajuste.final, type = "dfbeta", linear.predictions =
    FALSE)

res.partial<-resid(ajuste.final,type='partial')
tiempos <- as.numeric(rownames(res.partial))
par(mfrow=c(2,3))
for (i in 1:5){sm.res <- loess(res.partial[,i] ~ tiempos)
destip <- predict(sm.res,se=T)$se
plot(tiempos,res.partial[,i],xlab='Tiempos␣ordenados',type='p', ylab
   =paste('Residuos␣parciales␣para',(names(ajuste.final$coef))[i],
   sep = "␣"))
lines(tiempos,sm.res$fit)
lines(tiempos,sm.res$fit+2*destip,lty=2)
lines(tiempos,sm.res$fit-2*destip,lty=2)}

vif(ajuste.final)
```

Bibliografía

[1] Dahlborg, et al. (1996) *The Cancer Journal for Scientific American.* 2, 166-174