EVIDENCIAS EN PEDIATRÍA
Toma de decisiones clínicas basadas en las mejores pruebas científicas
www.evidenciasenpediatria.es

Fundamentos de medicina basada en la evidencia

Estudios de supervivencia. Modelo de riesgos proporcionales. Regresión de Cox

Ortega Páez E[1], Ochoa Sangrador C[2], Molina Arias M[3]
[1]Pediatra. Unidad de Gestión Clínica Góngora. Distrito Granada-Metropolitano. Granada España.
[2]Servicio de Pediatría. Hospital Virgen de la Concha. Complejo Asistencial de Zamora. Zamora. España.
[3]Servicio de Gastroenterología. Hospital Universitario La Paz. Madrid. España.

Correspondencia: Manuel Molina Arias: mma1961@gmail.com

Palabras clave en español: estadística; modelo de riesgos proporcionales; regresión de Cox; supervivencia.
Palabras clave en inglés: statistics; proportional hazards model; Cox regression; survival.

Fecha de recepción: 21 de septiembre de 2023 • **Fecha de aceptación:** 1 de octubre de 2023
Fecha de publicación del artículo: 11 de octubre de 2023

Evid Pediatr. 2023;19:48.

CÓMO CITAR ESTE ARTÍCULO

Ortega Páez E, Ochoa Sangrador C, Molina Arias M. Estudios de supervivencia. Modelo de riesgos proporcionales. Regresión de Cox. Evid Pediatr. 2023;19:48.

Este artículo está disponible en: http://www.evidenciasenpediatria.es/EnlaceArticulo?ref=2023;19:48.
 • ISSN: 1885-7388

Estudios de supervivencia. Modelo de riesgos proporcionales. Regresión de Cox

Ortega Páez E[1], Ochoa Sangrador C[2], Molina Arias M[3]
[1]*Pediatra. Unidad de Gestión Clínica Góngora. Distrito Granada-Metropolitano. Granada España.*
[2]*Servicio de Pediatría. Hospital Virgen de la Concha. Complejo Asistencial de Zamora. Zamora. España.*
[3]*Servicio de Gastroenterología. Hospital Universitario La Paz. Madrid. España.*

Correspondencia: Manuel Molina Arias: mma1961@gmail.com

En un capítulo anterior de esta serie, nos referimos al estudio de curvas de supervivencia según el método de Kaplan-Meier, válido para estimar las probabilidades acumuladas de supervivencia de una serie de sujetos en función del tiempo transcurrido de seguimiento y el evento de estudio, mediante comparaciones entre grupos diferentes definidos por un factor simple o variable categórica.

El modelo de riesgos proporcionales o regresión de Cox (RC), nombre que toma en honor al estadístico que la describió (Dave Cox, 1972), es hoy día una herramienta estadística ampliamente utilizada en medicina, esencial para el análisis de datos de supervivencia. Como cualquier modelo de regresión, es una función que analiza la relación entre una o más variables independientes o explicativas y la tasa de incidencia de un evento de interés, siendo capaz de predecir las probabilidades de supervivencia de un determinado sujeto a partir de los valores que tomen las variables predictivas. Comparte con los demás métodos de supervivencia (Kaplan-Meier) que para el cálculo del modelo computan todos los sujetos, ya que se tiene en cuenta el tiempo de todos los participantes que están en riesgo de presentar el evento en cada instante, tanto si el tiempo de seguimiento es completo hasta presentar el evento de interés o incompleto (censurados). A diferencia de otros modelos de regresión (lineal, logística) es un modelo **dinámico**, no solo nos dice el número de eventos que ocurren, sino a la velocidad a la que ocurren.

La RC es especialmente útil en:

- Ensayos clínicos donde la supervivencia o el tiempo hasta un evento de interés son resultados críticos.
- En estudios para determinar qué variables están asociadas con un mayor riesgo o probabilidad de experimentar el evento, puede ayudar a identificar poblaciones de alto riesgo y diseñar intervenciones preventivas o terapéuticas adecuadas.
- Para estimar la probabilidad de que un individuo o un grupo experimente un evento específico en el futuro. Para realizar un seguimiento adecuado de los pacientes y evaluar la supervivencia en periodos prolongados.

MODELO DE REGRESIÓN DE COX DE RIESGOS PROPORCIONALES

El modelo es el producto de dos funciones. La primera dependiente del tiempo [$h_0(t)$], y la segunda dependiente de las variables predictoras X ($e^{\beta x}$). Se considera un modelo semiparamétrico, ya que la primera función, que se obtiene a partir de los datos, está libre de cualquier distribución, mientras que la segunda función depende de la distribución de las variables predictoras.

La ecuación de regresión de Cox se puede expresar como:

$h(t;X) = h_0(t) * \beta' x = h_0(t) * e^{\beta_1 x_1+\beta_2 x_2+..\beta_i x_i}$. Tomando logaritmos neperianos se puede expresar como modelo log-lineal.

$$\ln h(t;X) = \ln h_0(t) + b_i x_i$$

***h*(t;X).** Función de riesgo. Representa la tasa de riesgo de un sujeto con valores = X (X_1,X_2… X_n) de la/s variable/s explicativas en un instante de tiempo determinado (t). Es la variable dependiente, codificada como 0: si, 1: no.

$h_0(t)$. Tasa basal de riesgo, que solo depende del tiempo. Representa la tasa instantánea de riesgo cuando el valor de las variables predictoras es igual a 0. Esta variable está unida en el análisis a la variable dependiente.

$e^{\beta x}$. Función exponencial. Es la combinación lineal de las variables predictoras o explicativas (X_n), que pueden ser cualitativas o cuantitativas.

Conceptos de *hazard*, *hazard ratio*, función acumulada de riesgo y función de supervivencia

- ***Hazard* (riesgo)** o riesgo instantáneo. Es el riesgo de experimentar un evento en un instante del estudio. Determina el riesgo de un paciente en un momento determinado, se podría calcular por medio del cociente de probabilidad de que el paciente sobreviva en un intervalo o incremento temporal, conociendo que ha sobrevivido hasta ese momento, dividido por el incremento de tiempo. El *hazard*

representa una probabilidad condicionada de presentar un evento en el siguiente instante, con la condición de que no se haya presentado antes del inicio de ese instante.

Matemáticamente, *hazard* (λ_t) en un tiempo t determinado se podría expresar como: número de eventos ocurridos (d_i) en un preciso instante dividido por el número total de personas en riesgo (n_i). Como es lógico, este riesgo instantáneo va cambiando durante el tiempo, dando lugar a una función de riesgo.

$$\lambda_t \frac{d_i}{n_i}$$

- ***Hazard ratio* (HR)** o cociente de riesgos instantáneos. Cuando tenemos un grupo con diferentes exposiciones en un tiempo determinado, se obtiene un *hazard* para cada grupo y al cociente entre los dos se denomina *hazard ratio* (HR), que se podría traducir como **cociente de riesgos instantáneos (CRI)**. La HR nos indica la relación entre una exposición y un evento en un tiempo t. Como vemos, la HR es una *odds*, representa cuánto más probable es que se produzca el suceso a que no se produzca en un grupo frente a otro. La interpretación es similar al riesgo relativo (RR), pero teniendo en cuenta el tiempo en el que se produce el suceso. Como cualquier cociente entre dos funciones, el valor nulo es igual a 1, lo que significa igualdad de probabilidad de producción del suceso en los dos grupos en el siguiente intervalo de tiempo. Una HR >1 indica más riesgo de producción en el grupo expuesto que en el control y una HR <1, menor riesgo en el grupo expuesto que en el control. Una HR = 3, significa que, si el sujeto no ha presentado todavía el suceso, tiene una probabilidad de tres veces más de presentarlo que un sujeto control durante el siguiente periodo de tiempo.

- **Función acumulada de riesgo.** Es el riesgo que un sujeto tiene (teniendo en cuenta los valores de las variables explicativas: X) de presentar el evento desde el inicio (t_0) hasta el tiempo que ocurre el suceso (t).

  $h(t;X) = h_0(t) * e^{\beta' x}$; tomando logaritmos neperianos, tenemos $h(t;X) = -\ln S(t;X)$

- **Función de supervivencia.** Es la probabilidad de supervivencia en el tiempo t de los sujetos con un determinado patrón de valores de las variables explicativas X. $S(t;X) = e^{-H(t;X)}$

INTERPRETACIÓN DE LOS PARÁMETROS

La estimación de los parámetros se realiza mediante la función de verosimilitud "parcial", porque la fórmula de probabilidad solo considera probabilidades para aquellos sujetos que mueren/fallan y no considera las probabilidades para aquellos sujetos que son censurados; dicho de otra manera, considera las verosimilitudes solo de los cambios ocurridos y no todos los sujetos de la muestra. Sin embargo, en el cálculo de las probabilidades de los tiempos de muerte sí tiene en cuenta todos los sujetos objeto de riesgo al inicio de los diferentes tiempos de muerte.

Veamos el caso más sencillo, donde la variable explicativa es cualitativa binaria X (0/1). Por ejemplo, un estudio que valore la muerte en grandes prematuros en función de presentar encefalopatía grave (X = 1) o no (X = 0).

$$h(t;X) = h_0(t) * e^{\beta x}$$

El *hazard* para cada grupo en el tiempo t sería:

$$h(t;X) = 1 = h_0(t) * e^{\beta}; = h(t; X = 0) = h_0(t) * e^{0}$$

El HR sería:

$$HR = \frac{h(t;X = 1)}{h(t;X = 0)} = \frac{h_0(t) * e^{\beta}}{h_0(t) * e^{0}} =$$
$$= e^{\beta} \rightarrow h(t;X = 1) = e^{\beta} * h(t;X = 0)$$

Esto quiere decir que la HR es equivalente a e^{β} y, como puede verse, es independiente del tiempo, solo depende de los datos de la variable explicativa. Se podría concluir que es el factor por el que se multiplica la tasa instantánea de riesgo cuando el valor de X aumenta en una unidad. En el ejemplo que nos ocupa con variable cualitativa, el valor se interpretaría como el cociente de riesgo instantáneo de muerte entre los grandes prematuros con encefalopatía grave respecto a los que no la tienen. En el caso de una variable cuantitativa (continua o discreta), el valor sería multiplicado por cada unidad de cambio de la variable X. Si β es positivo (+) nos indica un aumento de la tasa instantánea de riesgo cuando aumenta el valor de X (HR >1), y cuando tiene un signo negativo (-) significa una disminución de la tasa instantánea de riesgo al aumentar el valor de X (HR <1).

La significación estadística de la tasa de riesgo de la ocurrencia del evento y las variables explicativas se comprueban por la prueba de significación del coeficiente β y por el intervalo de confianza del coeficiente exponenciado (e^{β}).

La prueba de significación se estudia mediante la prueba de Wald (z), que compara el coeficiente β dividido por su error estándar (EE) con la ley normal estandarizada bajo la hipótesis nula de que z = 0.

$$Z = \frac{\beta}{EE}; \rightarrow (0,1);$$

Valores de $p < 0,05$ suponen que el coeficiente es significativo. Hay que decir que esta prueba es válida para muestras grandes. Cuando las muestras son pequeñas o la significación está cercana a 0,05 se recomienda realizar la prueba de verosimilitud que converge más rápido hacia la distribución normal.

El intervalo de confianza al 95% (IC 95) de e^{β} se calcula de la siguiente forma:

IC(95) = $e^{b} * e^{\pm 1,96EE}$. Donde el límite inferior (LI) = $e^{b} * e^{\pm 1,96EE}$; límite superior (LI) = $e^{b} * e^{\pm 1,96EE}$. Si el IC 95 contiene el valor nulo (la unidad) entonces el coeficiente no es significativo.

Veamos un ejemplo utilizando un programa de acceso libre, el *software* estadístico R (https://www.r-project.org/) con el plugin RCommander (https://www.rcommander.com/) y la base de datos EeP_Fund_SupervivenciaT12m.RData (https://evidenciasenpediatria.es/files/43-227-RUTA/EeP_Fund_SupervivenciaT12m.RData). Si necesita saber cómo instalar RCommander, puede consultar el siguiente tutorial en línea: http://sct.uab.cat/estadistica/sites/sct.uab.cat.estadistica/files/instalacion_r_commander_0.pdf

En la base de datos se recogen una serie de registros sobre la duración de la lactancia materna en un grupo de madres con sus hijos, además de otras variables, como el peso del recién nacido, la edad, el tipo de parto, el nivel de educación, etc. Tres variables tienen especial relevancia para el ejemplo:

1. La duración en meses de la lactancia materna durante el periodo de seguimiento de 1 año (LactDuracionHasta12m). Sería equivalente a la variable tiempo transcurrido (h_0t).

2. La variable lactancia materna al final del periodo (LactanMenos12mes). La ocurrencia del evento es la interrupción de la lactancia materna. La codificamos como 0 en el caso de persistir la lactancia materna (tiempo completo sin suceso) y como 1 en el caso de haberse interrumpido antes del final del estudio (ocurrencia del suceso). Sería equivalente a la variable dependiente, la tasa de instantánea de riesgo.

3. Edad de la madre en años (EdadMadre30) codificada como categórica: Mayor/igual 30 y Menor 30. Sería equivalente a la variable explicativa.

La instalación básica de RCommander no incluye las rutinas para realizar las técnicas de análisis de supervivencia, por lo que, antes empezar, debemos cargar un *plugin* (una extensión) denominado RcmdrPlugin.survival.

Una vez que hemos abierto el programa R, debemos instalar el paquete correspondiente al *plugin* (si no se ha instalado y usado antes), para lo cual, el método más rápido es teclear el comando install.packages(RcmdrPlugin.survival). Si este método falla, puede seleccionarse en R el menú Paquetes → Instalar paquetes(s)… y seleccionar un CRAN de la lista que ofrece la ventana emergente. Una vez hecho, veremos la lista de todos los paquetes de R ordenados alfabéticamente. Buscamos y marcamos RcmdrPlugin.survival y pulsamos OK.

Nuestra intención es estudiar si la duración de la lactancia materna está relacionada con la edad de madre mayor o menor de 30 años.

Para construir el modelo de Cox nos vamos a la pestaña de Rcommader de *Statistics* (estadística) → *Fif models* (ajuste de modelos) → *Cox regression model* (modelos de regresión de Cox). Se nos despliega un cuadro de dialogo donde debemos introducir las variables. En la pestaña *Data* (datos) en la ventana *Time or start/and times* (tiempos de inicio y final) introducimos la variable tiempo transcurrido (LactDuracionHasta12m) y en la ventana *Event indicator* (indicador del evento) la variable LactanMenos12mes, el resto de las ventanas las dejamos libres, ya que no vamos a aplicar ninguna estratificación, a continuación, seleccionamos todos los casos (*Subset expression: all vallid cases*). En la pestaña *Model* (modelo) seleccionamos el método de realización del modelo *Method for ties* (método) Efron que viene por defecto, los errores estándar robustos (*Robust Standard Errors* → *Default*) y le damos nombre al modelo (CoxModel.1). En la ventana Variables clicamos la variable explicativa Edad30 que está codificada como factor y se nos coloca en el cajón inferior (*Model Formula*- Fórmula del modelo). El resto lo dejamos como está y clicamos en OK (**Figura 1**).

En la **Figura 2** ofrecemos la ventana de salida con los resultados.

En primer lugar, nos aparece el nombre del modelo (CoxModel.1), la función en R coxph(Surv), las variables (LactDuracionHasta12m, LactanMenos12mes) ~ EdadMadre30, el método con el que se ha realizado el modelo (method="efron" y, por último, el nombre de la base de datos (data=EeP_Fund_SupervivenciaT12m). Se ha elegido el método de efron porque es el más válido en caso de empates y se acerva mucho al método exacto. Obsérvese que en la fórmula del modelo de regresión la variable dependiente está "unida" en el análisis a la variable tiempo trascurrido (primera parte de la ecuación).

Seguidamente, debemos fijarnos en el número total de registros (2195) y el número de eventos (las madres que han dejado la lactancia antes del año, 1821). En el modelo hay que destacar que la comparación se establece en la categoría de <30 años, el coeficiente β (coef: -0,2099), el coeficiente exponenciado (exp(coef) e^{β}: 0,81065, que corresponde con la HR, el error estándar del coeficiente (se (coef); 0,05), la prueba z de Wald para la significación el mismo (z: -3,53) y la probabilidad de la prueba (Pr)>z: 0,000403, que nos indica que es significativa p <0,0001(***), y los límites inferior y superior del IC 95 de la HR: 0,72 a 0,91.

Podríamos concluir que las madres lactantes de <30 años tienen un 18,94% (HR: 0,8106) menor riesgo de perder la lactancia al año respecto a las ≥30 años. El IC 95 no incluye el valor nulo (1), lo que significa que la relación es significativa

Figura 1. Construcción de un modelo de Regresión Cox en R

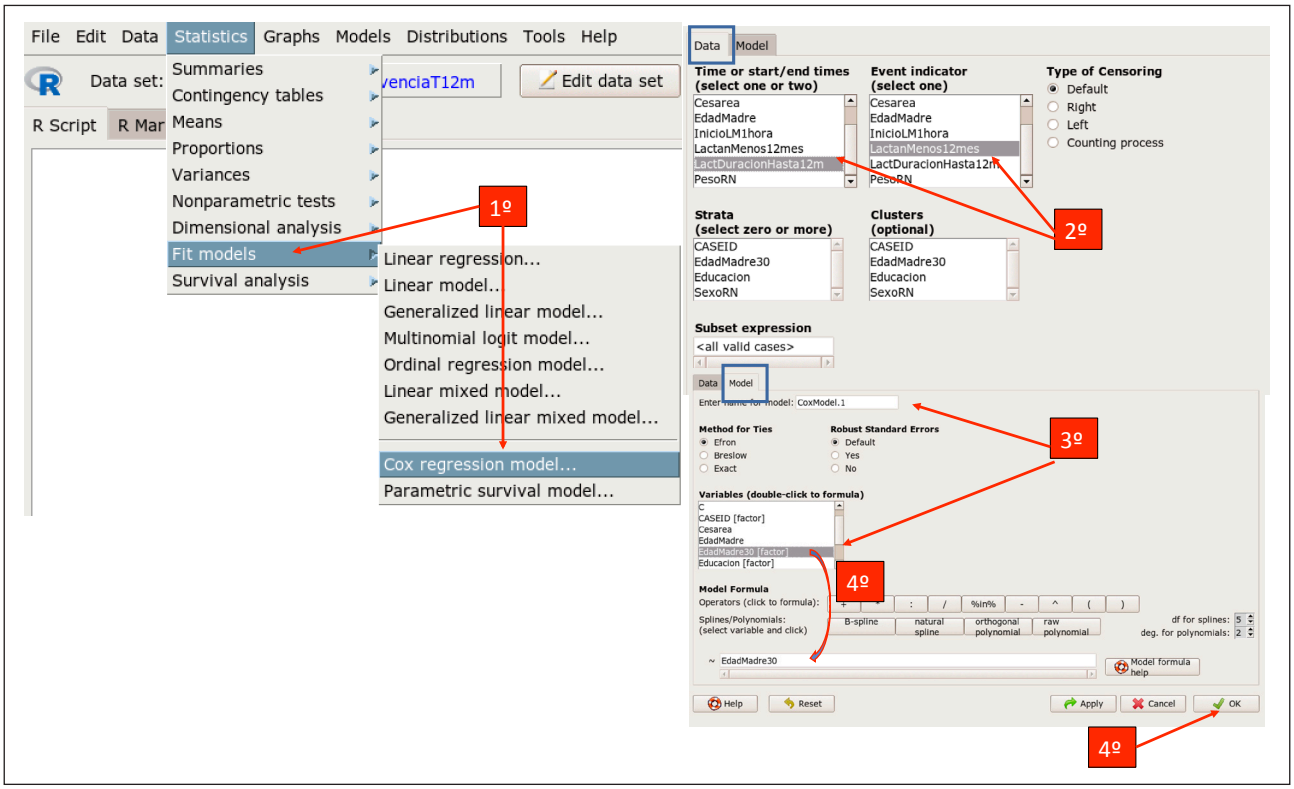


Figura 2. Resultados Regresión de Cox en R

```
Rcmdr>  CoxModel.1 <- coxph(Surv(LactDuracionHasta12m, LactanMenos12mes) ~
Rcmdr+    EdadMadre30, method="efron", data=EeP_Fund_SupervivenciaT12m)

Rcmdr>  summary(CoxModel.1)
Call:
coxph(formula = Surv(LactDuracionHasta12m, LactanMenos12mes) ~
   EdadMadre30, data = EeP_Fund_SupervivenciaT12m, method = "efron")

 n= 2195, number of events= 1821

                                coef     exp(coef)  se(coef)     z      Pr(>|z|)
EdadMadre30[T.Menor 30] -0.20992    0.81065  0.05933 -3.538 0.000403 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                                exp(coef) exp(-coef) lower .95 upper .95
EdadMadre30[T.Menor 30]    0.8107     1.234    0.7217    0.9106

Concordance= 0.52  (se = 0.006 )
Likelihood ratio test= 13.05  on 1 df,   p=0.0003
Wald test            = 12.52  on 1 df,   p=0.0004
Score (logrank) test = 12.56  on 1 df,   p=0.0004
```

HR

p valor de la prueba de significación

Inverso de la HR cuando el factor de riesgo es protector

Criterios de Bondad de ajuste del modelo

estadísticamente. R nos ofrece el valor del coeficiente exponenciado negativo (exp(-coef: 1,23), que en el caso de variables explicativas protectoras facilita la interpretación, al invertir la categoría de referencia, lo que se interpretaría como que las mujeres de ≥30 años tienen un 23% más riesgo de perder la lactancia materna que las de <30 años. R no nos ofrece el IC 95 (1,095 a 1,38), hay que calcularlo manualmente con la fórmula vista anteriormente:

$$\text{Límite inferior (LI)} = e^{b} * e^{-1,96EE} = e^{b} * e^{-1,96 * 0,0593} = 1,095$$
$$\text{Límite superior (LS)} = e^{b} * e^{+1,96EE} = e^{b} * e^{+1,96 * 0,0593} = 1,38$$

SUPUESTOS DE APLICACIÓN DEL MODELO

Para la aplicación de una RC es necesario que se cumplan ciertos supuestos:

1. **Supuesto de riesgos proporcionales**. Es el supuesto más importante, significa que el cociente de riesgos instantáneos (HR) a lo largo de todo el seguimiento debe ser constante. Ya vimos anteriormente que HR es independiente del tiempo transcurrido representado por el término no paramétrico del modelo (h_0t), que no presupone ningún supuesto; esto quiere decir que depende solamente de la distribución de los datos de las variables explicativas. En realidad, la HR final o global es consecuencia de un promedio de todas las HR parciales durante todo el seguimiento que deben ser constantes. En el caso de una variable cualitativa, donde existen dos categorías A y B, los valores de la función de riesgo que toman los sujetos de la categoría A deben ser proporcionales a los valores que toman los mismos sujetos con la categoría B [ht;X=A/ht;X=B=cte].

   La violación de este supuesto se puede corregir de dos formas. Estratificando por la variable que no cumple este supuesto o introduciendo una variable que represente el tiempo de seguimiento de cada variable que no sigue el supuesto y que se comporta como un término de interacción que multiplica su valor por el de la variable explicativa en cuestión. A este modelo extendido de la regresión de Cox se le denomina regresión de Cox dependiente de tiempo.

2. El **suceso** debe ser **irreversible**, ha de ocurrir una sola vez y la censura debe ser no informativa. El incumplimiento de este supuesto puede corregirse con la regresión de Cox de riesgos competitivos, donde el suceso en la variable de estudio puede repetirse más de una vez.

3. El **modelo** debe incluir una **variable dependiente**, tasa de riesgo del evento, que recoja el suceso del evento (1) o la ausencia del mismo (0), que va unida a la **variable tiempo trascurrido** y una o más **variables explicativas** que pueden ser cualitativas o cuantitativas.

4. **Supuesto log-lineal**. La relación entre el logaritmo neperiano (ln) de la tasa instantánea de riesgo de las variables explicativas debe ser lineal. $\ln(t;X) = \ln h_0(t) + b_i x_i$.

5. Los **tiempos de supervivencia** de los individuos son **independientes** entre sí, dados los valores de las variables predictoras. Esto significa que el tiempo de supervivencia de un individuo no debe influir en el tiempo de supervivencia de otro individuo.

6. No debe haber resultados influyentes o valores atípicos, lo que quiere decir que resultados muy apartados positivos o negativos pueden afectar a la validez del modelo.

1. DIAGNÓSTICO DEL MODELO

1. 1. Diagnóstico de bondad de ajuste del modelo

Se realiza por tres pruebas de hipótesis: prueba de razón de máxima verosimilitud (*Likelihood ratio test*), prueba de Wald y el *Score (longrank) test*. Son pruebas de hipótesis globales, tienen en cuenta todas las variables del modelo y siguen una distribución $\chi^2_{gl=n}$; donde n = n.º de variables del modelo.

- Prueba de razón de máxima verosimilitud. Es la prueba de hipótesis de la verosimilitud parcial del modelo. Es la más aconsejable, ya que se afecta poco por el tamaño muestral.

- Prueba de Wald. Prueba de hipótesis basada en que los coeficientes β ajustados se aproximan a una distribución normal. El contraste se realiza bajo la hipótesis nula de que todos los coeficientes son igual a cero $\beta = 0$. Esta prueba es muy sensible al tamaño muestral, por lo que no es aconsejable en muestras pequeñas.

- *Score (logrank) test*. Prueba de hipótesis que utiliza las derivadas del logaritmo de la verosimilitud parcial bajo la hipótesis nula de que el vector de los coeficientes es nulo.

Es frecuente en los modelos de regresión medir el grado de varianza explicada por el modelo ajustado por el valor de la R^2. En la RC la R^2 no es un buen parámetro, ya que su valor nunca puede llegar a 1 por el término no paramétrico del modelo. Por esta razón los paquetes informáticos proporcionan la *Concordance* (concordancia), que se interpretaría como la clasificación correcta de los pares de observaciones en términos de tiempo hasta el evento. Es útil para evaluar la capacidad predictiva del modelo, se considera como una generalización del área bajo la curva ROC en modelos de supervivencia. La concordancia toma valores entre 0 y 1, donde un valor de 0,5 indica que el modelo no tiene capacidad de discriminación (similar a una moneda lanzada al aire) y un valor de 1 indica una capacidad perfecta de discriminación (todos los individuos se clasifican correctamente según su riesgo).

2. COMPROBACIÓN DE LOS SUPUESTOS DEL MODELO

2.1. Supuesto de riesgos proporcionales

Se puede estudiar de dos formas, mediante gráficos o por pruebas de hipótesis.

Métodos gráficos

El más sencillo es la comprobación de las **curvas de supervivencia ajustadas** por el modelo para cada categoría de la variable explicativa, donde se relaciona el logaritmo de la probabilidad de supervivencia en el eje de la Y con el tiempo trascurrido en el eje de la X. Si las curvas de supervivencia son paralelas entre sí, podemos asumir que los cocientes de riesgos instantáneos a lo largo del tiempo son proporcionales.

Gráfico de los residuos de Schoenfield escalados. Se calculan por cada variable del modelo por separado. Son la diferencia de los valores obtenidos por el modelo de cada variable ajustados por las demás covariables y los valores esperados en un momento dado, calculado por el promedio ponderado (multiplicando cada valor por el inverso de la varianza) de los valores de las covariables entre todos los individuos que están en riesgo en ese momento dado. La gráfica muestra los valores de los residuales en el eje de la Y ordenados frente al tiempo de estudio en el eje de la X. El supuesto de proporcionalidad se cumple si los valores son independientes del tiempo (no existe ningún patrón de distribución), se distribuyen alrededor del valor 0 y comprendidos entre +/- 2 desviaciones estándar.

Métodos estadísticos mediante pruebas de hipótesis

Es más confiable que el anterior, ya que a veces la inspección gráfica puede ser engañosa. Se realiza mediante una correlación lineal entre los residuos escalados de Schoenfield y el tiempo de estudio. Bajo la hipótesis nula de que la ausencia de correlación ($H_0 = \rho = 0$) es sinónimo de riesgos proporcionales. Valores p <0,05 suponen rechazar la hipótesis nula de riesgos proporcionales.

2.2. Supuesto de no linealidad

Este supuesto solo es aplicable a las variables continuas. Se realiza mediante los residuos de Martingala, que son los valores calculados por el modelo menos los esperados. Tienen una distribución asimétrica. El gráfico se obtiene representando los residuos de Martingala del modelo nulo en el eje de la Y frente a los residuos estimados por el modelo. Valores alrededor del 0 suponen que no existe linealidad, valores cercanos a 1 suponen que los individuos tuvieron el evento demasiado pronto y valores cercanos a 0 suponen que lo tuvieron demasiado tarde.

2.3. Supuesto de los valores atípicos o influyentes

- Residuos de desviación o *residual deviance* (desviación), que consiste en la normalización de los residuos de Martingala. Deben distribuirse de forma simétrica alrededor de 0 con una desviación estándar de 1. Valores positivos suponen que los individuos tuvieron el suceso demasiado pronto, valores negativos es que lo tuvieron demasiado tarde.

- Residuos DFBETA. Se calculan para cada variable tomando la diferencia entre los coeficientes estimados por el modelo y los coeficientes estimados después de eliminar una observación individual del conjunto de datos. El resultado es un conjunto de DFBETA para cada observación y para cada variable del modelo. Refleja cuánto cambia el coeficiente estimado para cada variable cuando se excluye una observación en particular. Los valores deben ser simétricos alrededor de 0. Valores grandes de DFBETA indican que la eliminación de esa observación del análisis afectaría considerablemente los coeficientes estimados y sería un valor influyente. Los DFBETAS son los DFBETA estandarizados.

Siguiendo con nuestro ejemplo, tenemos:

Bondad de ajuste del modelo

En la última fila de la **Figura 2** representamos el criterio de bondad de ajuste del modelo. Vemos cómo la prueba de verosimilitud parcial (p = 0,0003), la prueba de Wald (p = 0,0004) y la prueba *Score* (p = 0,0004) son todas significativas, hecho que apoya que el modelo reproduce la probabilidad pronosticada. Es de destacar que la prueba de verosimilitud parcial es la más potente.

Por último, la concordancia del modelo es de 0,52. Lo que significa que el modelo es poco discriminativo.

Comprobación de los supuestos del modelo

1. **El suceso** (pérdida de la lactancia materna) **es único, irreversible** (no se considera la recuperación de la lactancia) y la censura (los que no presentan el evento) no es informativa, y es independiente de la causa de la pérdida de la lactancia.

2. **Variables incluidas en el modelo.** Existe una variable dependiente: LactanMenos12mes con dos posibles estados (0 No, 1 Sí), una variable tiempo transcurrido: LactDuracionHasta12m y una variable explicativa: EdadMadre30 con dos categorías ≥30 años/<30 años.

3. **Los tiempos de supervivencia** de los individuos (los sucesos de los eventos) son **independientes** entre sí en cada sujeto.

La comprobación de los tres supuestos que vienen a continuación se realiza con el estudio de los residuos, que no son más que los valores observados menos los esperados por el modelo más o menos transformados.

4. **Supuesto de riesgos proporcionales**. Rcommander, en su rutina de análisis, ofrece la curva de supervivencia media de la variable, que en el caso de que sea cualitativa no diferencia la curva por categorías; es preciso realizarlo manualmente. Podemos emplear la función ggsurvplot de la librería survminer, como se detalla en la **Figura 3**. Podemos observar que las mujeres de <30 años tienen mayor supervivencia que las de ≥30 años y que ambas curvas son paralelas.

   Rcommander realiza el gráfico de los residuales de Schoenfield y la prueba de hipótesis mediante la función cox.zph() en la pestaña *Models* (modelos) → *Numerical diagnostics* (diagnósticos numéricos). Computa bien la prueba de hipótesis (**Figura 4**), pero da error en la gráfica de los residuos de Schoenfield, al no calcular bien los grados de libertad de la prueba (df), ya que por defecto utiliza df = 4.

   La prueba realiza dos contrastes: uno global, que incluye todas las variables del modelo, y otro por cada variable por separado, en este caso al existir una única variable ambos contrastes coinciden. Observamos que la prueba no es significativa (p = 0,68), lo que implica que no es posible rechazar la hipótesis nula de riesgos proporcionales.

   Para realizar el gráfico de Schoenfield es necesario hacerlo de forma manual limitando el análisis a df = 2 (**Figura 5**). Podemos observar que los puntos de ambas categorías están dispersos en una línea horizontal, indicando que el supuesto de proporcionalidad se cumple. También nos ofrece la prueba de hipótesis vista anteriormente (p = 0,68).

5. **Supuesto de no linealidad**. Rcommander en la pestaña *Models* (modelos) → *Graphs* (gráficos) → *Plot null Martingale residuals*. Como se dijo anteriormente, en variables cualitativas la no linealidad no es un problema, solo es aplicable a variables cuantitativas. De hecho, si se realiza el gráfico no es interpretable, ya que obtenemos dos líneas paralelas que corresponden a las observaciones de las dos categorías en los extremos sin que se pueda ver la distribución alrededor el 0 (**Figura 6**).

Figura 3. Comprobación supuesto de riesgos proporcionales. Curvas de logaritmo de la supervivencia ajustada y tiempo

Código de sintaxis

- Install.packages ("survminer")
- Library("suvminer")
- ggsurvplot(survfit(Surv(LactDuracionHasta12m, LactanMenos12mes) ~ EdadMadre30, data=EeP_Fund_SupevivenciaT12m),conf.int = TRUE,ggtheme = theme_bw(), fun = "log")

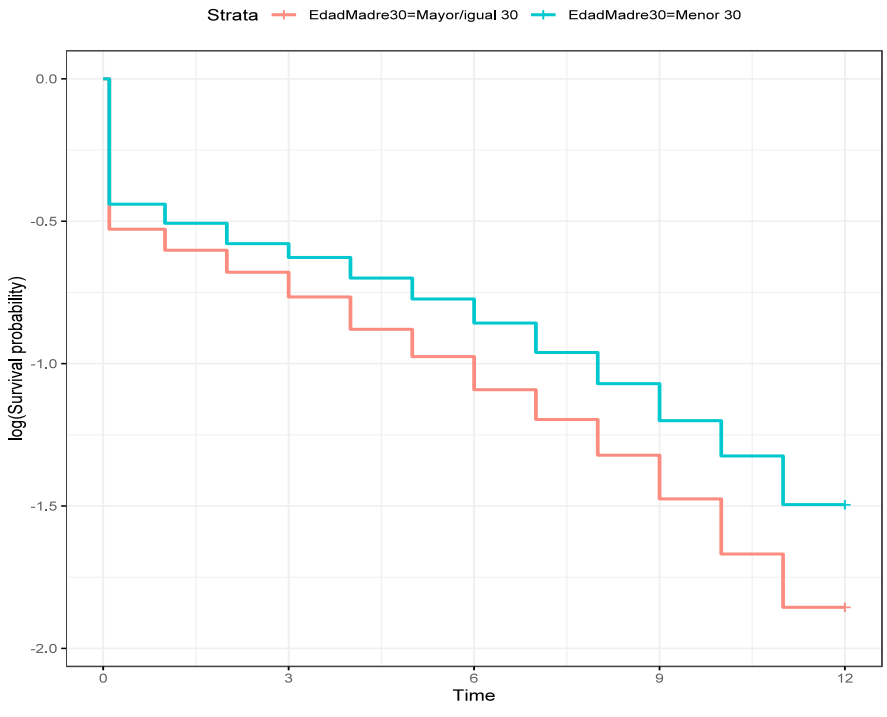
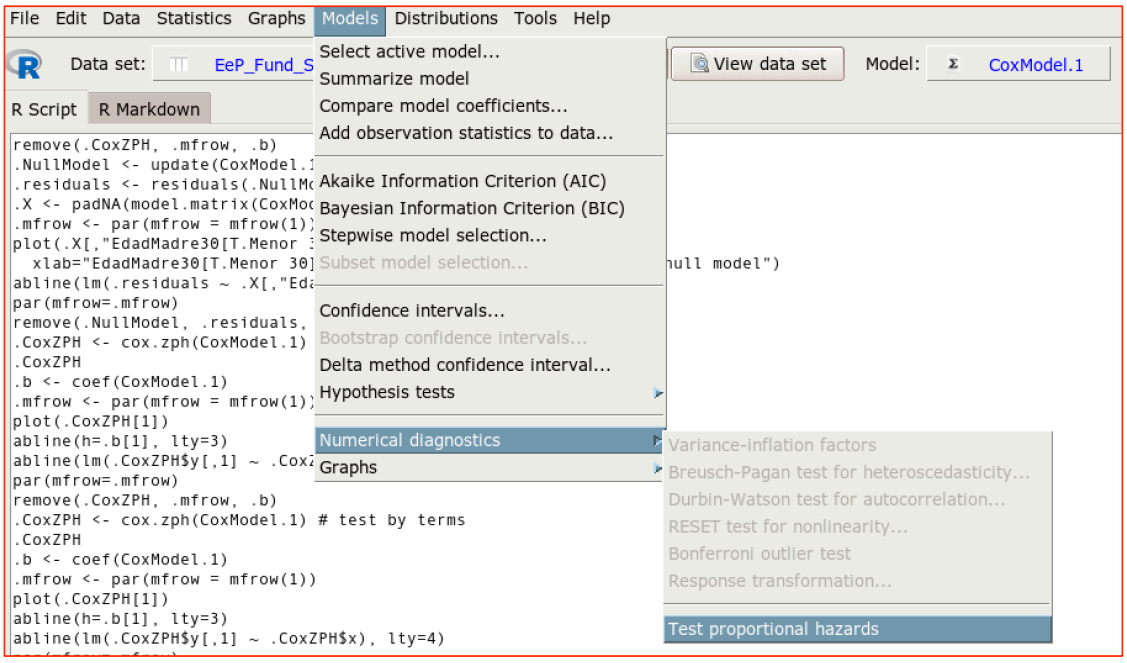


Figura 4. Comprobación supuesto riesgos proporcionales. Prueba de hipótesis




```
Rcmdr>  .CoxZPH <- cox.zph(CoxModel.1) # test by terms

Rcmdr>  .CoxZPH
                  chisq df    p
EdadMadre30  0.166  1   0.68
GLOBAL          0.166  1   0.68

Rcmdr>  .b <- coef(CoxModel.1)
```

Figura 5. Comprobación supuesto riesgos proporcionales. Método gráfico y prueba de hipótesis

Código de sintaxis

- Install.packages ("survminer")
- Library("suvminer")
- test.ph<- cox.zph(CoxModel.1)
- ggcoxzph(test.ph, df=2)

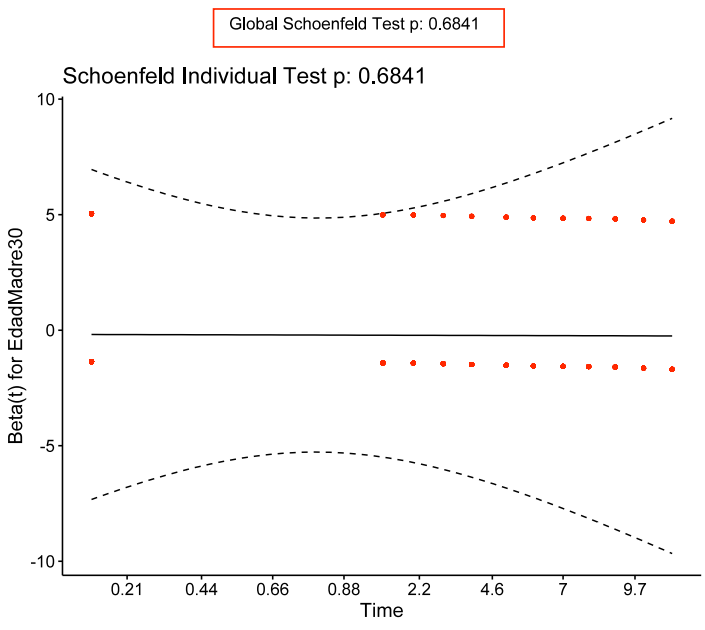
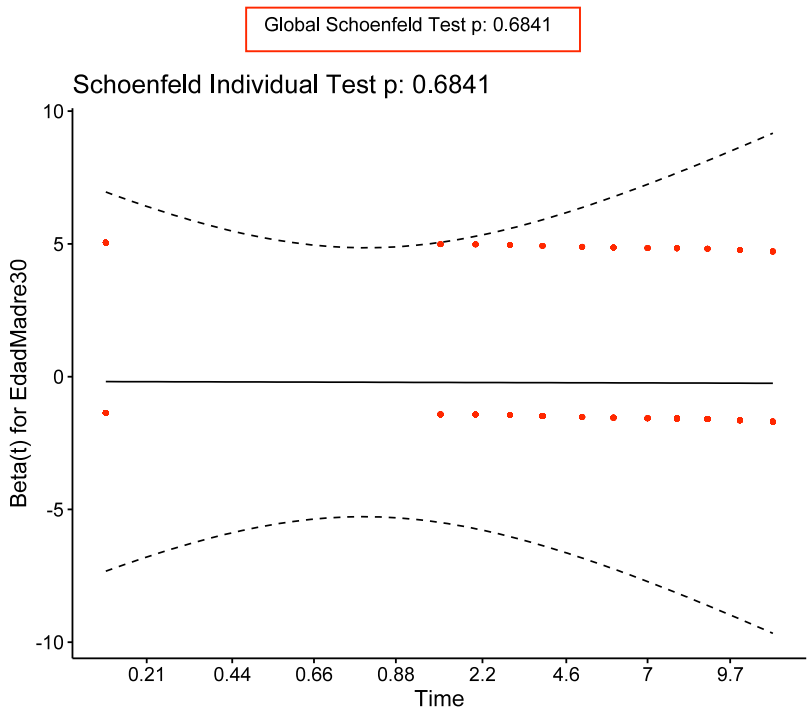


Figura 6. Supuesto de no linealidad. Residuos de Martingala

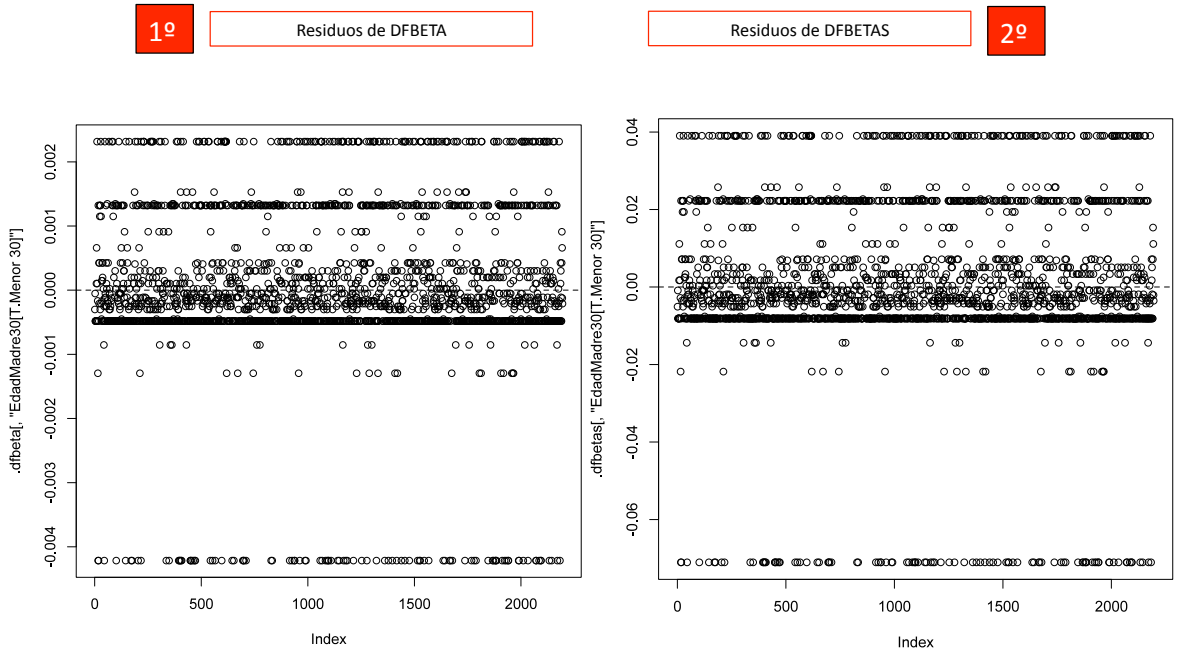
Código de sintaxis

- Install.packages ("survminer")
- Library("suvminer")
- test.ph<- cox.zph(CoxModel.1)
- ggcoxzph(test.ph, df=2)



6. **Supuesto de ausencia de valores atípicos o influyentes.** Rcommander en la pestaña *Models* (modelos) → *Graphs* (gráficos), se despliegan dos opciones: *Plot survival-regression dfbetas* (residuos DFBETA), *Plot survival-regression dfbeta* (DFBETA estandarizados). Los gráficos se exponen en la **Figura 7**. Podemos observar que los valores están alrededor del valor 0, no existen resultados muy alejados y que la distribución se realiza alrededor de dos líneas paralelas correspondientes a los dos grupos.

Figura 7. Residuales DFBETA y DFBETAS



RELACIÓN ENTRE *HAZARD RATIO*, RIESGO RELATIVO Y *ODDS RATIO*

El riesgo relativo (RR) y la *odds ratio* (OR) nos dan una fotografía en un momento dado, son medidas estáticas de riesgo. La HR es dinámica, ya que tiene en cuenta el tiempo en que se producen los sucesos, y nos ofrece una película a través del tiempo. Los tres estimadores ofrecen resultados diferentes. El RR siempre estará más cercano a la unidad, la OR será la más lejana y la HR estará en una posición intermedia. Cuanto más largo sea el seguimiento, mayor la incidencia de eventos y mayor el valor del RR, más divergencia habrá entre RR y HR. El motivo es que la RC tiene en cuenta el momento concreto en que se produjo el evento y los datos censurados, de ahí que los denominadores a lo largo del tiempo vayan cambiando, siendo cada vez menores; esto trae consigo un aumento de la HR. Debemos tener en cuenta que la equivalencia de los tres parámetros pierde sentido si los seguimientos de los pacientes no son homogéneos.

BIBLIOGRAFÍA

- Domenech Massons JM, Navarro Pastor JB. Análisis de la supervivencia y modelo de riesgos proporcionales de Cox. Editorial Signo. Barcelona. 2008.
- Martínez-González MA, Álvaro A, López Fidalgo J. ¿Qué es una hazard ratio? Nociones de análisis de supervivencia. Med Clin (Barc). 2008;131(2):65-72.
- Ruiz-Canela E, López Fidalgo J, Martínez-González MA. Aspectos avanzados de la Regresión de Cox. En: Martínez González MA, Sánchez-Villegas A, Toledo Atucha EA, Faulin Falardo J (Eds.). Bioestadística amigable, 4.ª Ed. Elsevier España SL. Barcelona, 2020;451-69.
- Therneau T, Atkinson E. Concordance. March 2, 2022. Vignettes from package 'survival'.
- Assessment of Discrimination in Survival Analysis (C-statistics, etc.). En: Rpubs [en línea] [consultado el 09/10/2023]. Disponible en: https://rpubs.com/kaz_yos/survival-auc
- Martinez J. Análisis de supervivencia en R. En: Rstudio [en línea] [consultado el 09/10/2023]. Disponible en: http://rstudio-pubs-static.s3.amazonaws.com/316989_83cbe556125645b698c9ff6cf88c4c1a.html#44_validaci%C3%B3n_de_supuestos
- Therneau TM, Lumley T, Atkinson E, Crowson C. Suvival Analysis. 14 de agosto de 2023 [en línea] [consultado el 09/10/2023]. Disponible en: https://cran.r-project.org/web/packages/survival/survival.pdf
- Fox J, Weisberg S. Cox Proportional-Hazards Regression for Survival Data in R An Appendix to An R Companion to Applied Regression. 3rd Ed. Última revisión: 28/09/2018 [en línea] [consultado el 09/10/2023]. Disponible en: https://socialsciences.mcmaster.ca/jfox/Books/Companion/appendices/Appendix-Cox-Regression.pdf
- Cox Proportional-Hazards Model. En: STDHA [en línea] [consultado el 09/10/2023]. Disponible en: www.sthda.com/english/wiki/wiki.php?title=cox-proportional-hazards-model
- Cox Model Assumptions. En. STDHA [en línea] [consultado el 09/10/2023]. Disponible en: www.sthda.com/english/wiki/wiki.php?title=cox-model-assumptions