Introducción al Análisis de Supervivencia

Instructor: Juan Carlos Salazar Uribe*

Escuela de Estadística

Universidad Nacional de Colombia, Sede Medellín

jcsalaza@unal.edu.co

*'How long a minute is, depends on which side of the bathroom door you're on.' (Zall's Second Law)

Ejemplo: Obteniendo un P-P Plot

Considere los siguientes datos sobre tiempo de supervivencia en semanas de 42 pacientes con leucemia quienes participaron en un ensayo clínico para comparar un tratamiento basado en esteroides con un placebo. Se obtendrá un Percentile-Percentile plot de acuerdo a Collett. Para esto se usará el siguiente programa:

```
data exp;
input Time status gender logWBC Grupo @@;
cards;
35 0 1  1.45 0 34 0 1 1.47 0 32 0 1 2.20 0 32 0 1 2.53 0
25 0 1 1.78 0 23 1 1 2.57 0 22 1 1 2.32 0 20 0 1 2.01 0
19 0 0 2.05 0 17 0 0 2.16 0 16 1 1 3.60 0 13 1 0 2.88 0
11 0 0 2.60 0 10 0 0 2.70 0 10 1 0 2.96 0 9 0 0 2.80 0
7 1 0 4.43 0 6 0 0 3.20  0 6 1 0 2.31  0 6 1 1 4.06  0
6 1 0 3.28  0 23 1 1 1.97  1 22 1 0 2.73  1 17 1 0 2.95  1
15 1 0 2.30  1 12 1 0 1.50  1 12 1 0 3.06  1 11 1 0 3.49  1
11 1 0 2.12  1 8 1 0 3.52  1 8 1 0 3.05  1 8 1 0 2.32  1 8 1 1 3.26  1
5 1 1 3.49  1 5 1 0 3.97  1 4 1 1 4.36  1 4 1 1 2.42  1 3 1 1 4.01  1
2 1 1 4.91  1 2 1 1 4.48  1 1 1 1 2.80  1 1 1 1 5.00  1
;
run;
```

Ejemplo: Obteniendo un P-P Plot

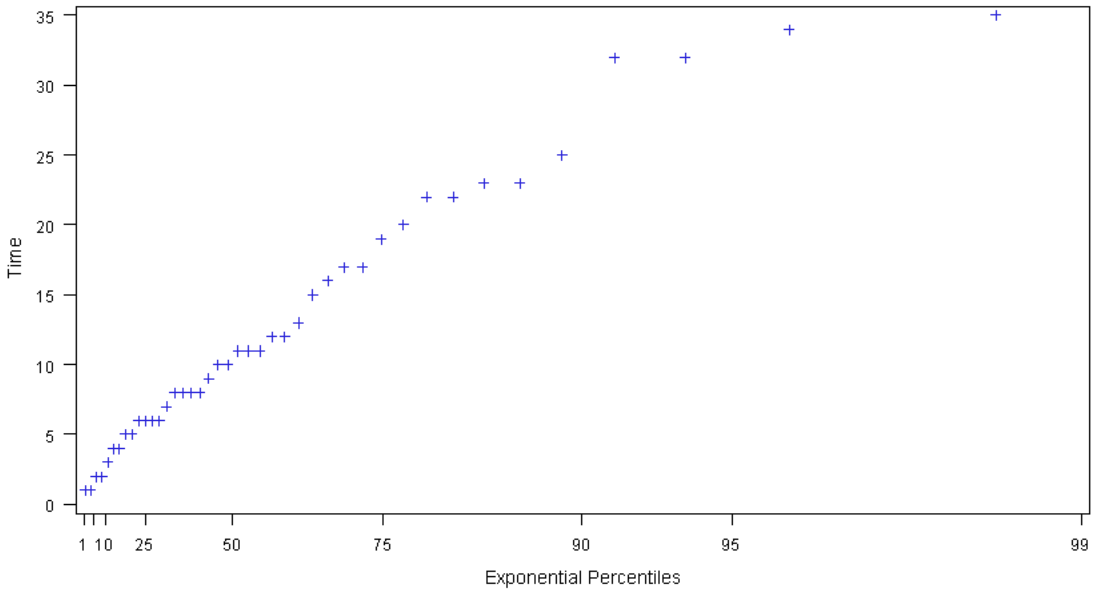
El SAS Proc Univariate, si bien no permite obtener el mismo gráfico que Collett explica en su libro, si permite obtener un percentile-percentile plot que permite evaluar el ajuste de un modelo, por ejemplo de un modelo exponencial*:

```
proc univariate data=exp;
var time;
probplot time /exponential;
run;
```

*Este método se debe usar con cuidado ya que no tiene en cuenta la censura. Sin embargo, si la tasa de censura no es muy alta se puede usar

Ejemplo: Obteniendo un P-P Plot

Se obtiene:



Ejemplo: Obteniendo un P-P Plot

Para obtener un gráfico similar al de Collett se puede usar el siguiente código.

Note que por ejemplo, para estimar el percentil 10[*] se toma el promedio de la estimación (es una estimación del logaritmo del tiempo obtenida con el modelo y por lo tanto incorpora la censura a derecha[**]) de este para cada una de las observaciones:

```
symbol1 v=square c=blue;
proc lifereg data=exp;
model Time*status(0)=logWBC Grupo / dist=exponential;
output out=Superv cdf=F p=predicted quantile=0.1 to 0.9 by 0.1;
title 'Ajuste del modelo exponencial';
run;
proc means data=superv mean noprint;
class grupo _PROB_;
var predicted;
output out=medias mean=mean;
run;
```

[*] Puesto que SAS transforma la exponencial a una distribucipon de valor extremo, la función cuantil es $u_p = \log(-\log(1-p))$

[**] El p-ésimo percentil estimado está dado por $\widehat{y}_p = \widehat{\log(t_p)} = \widehat{\beta}^T \mathbf{x} + \widehat{\sigma} u_p$

Ejemplo: Obteniendo un P-P Plot

```
data medias1;
set medias;
if grupo=. then delete;
if _PROB_=. then delete;
run;

data g0;
set medias1;
if grupo=0;
mean_grupo0=mean;
keep _PROB_ mean_grupo0;
run;
data g1;
set medias1;
if grupo=1;
mean_grupo1=mean;
keep _PROB_ mean_grupo1;
run;
```

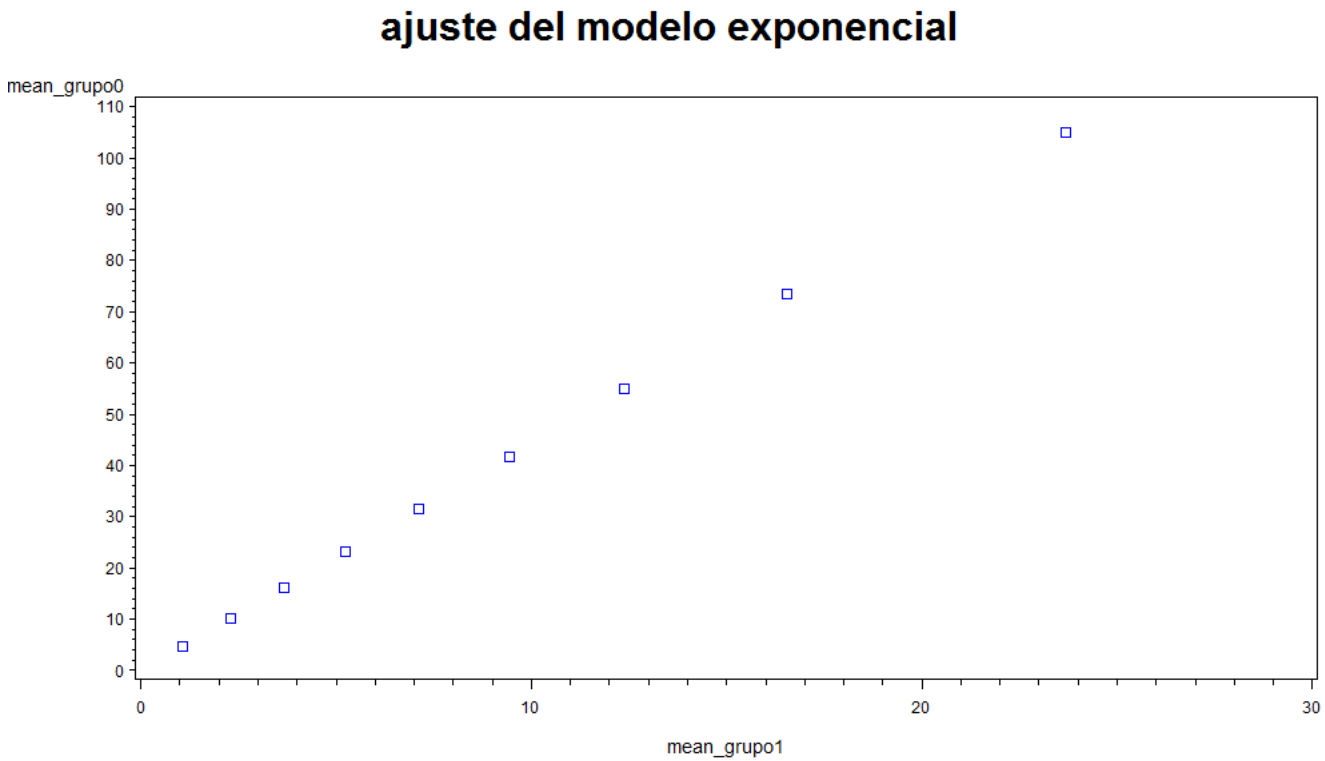
Ejemplo: Obteniendo un P-P Plot

```
data collett;
merge g1 g0;
run;

proc gplot data=collett;
plot mean_grupo0*mean_grupo1;
run;
title 'ajuste del modelo exponencial';
quit;
```

Ejemplo: Obteniendo un P-P Plot

Este programa produce:



¿Cómo obtener residuales en SAS

Considere los siguientes datos sobre tiempo de supervivencia en semanas de 42 pacientes con leucemia quienes participaron en un ensayo clínico para comparar un tratamiento basado en esteroides con un placebo. ¿Cómo obtener algunos de los residuales estudiados a fin de evaluar el ajuste de un modelo paramétrico exponencial? Para esto se puede usar el siguiente programa.

¿Cómo obtener residuales en SAS

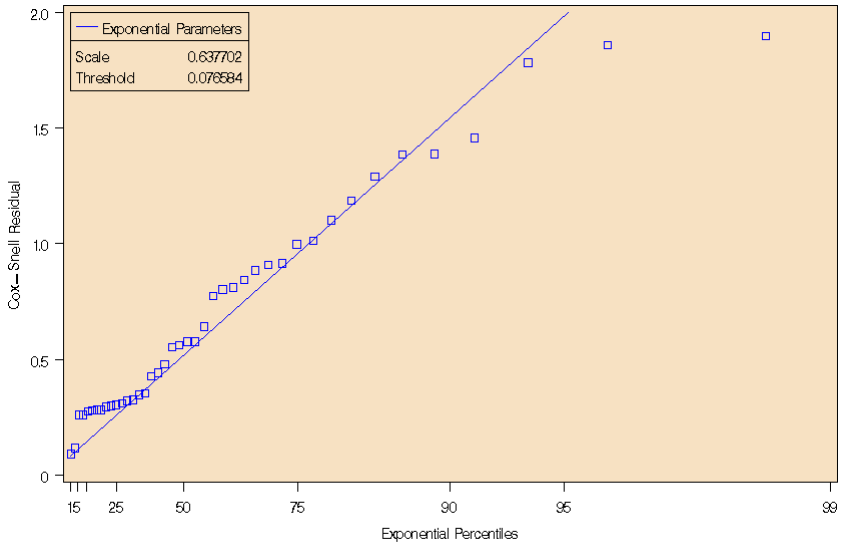
```
data exp;
input Time status gender logWBC Grupo @@;
cards;
35 0 1  1.45 0 34 0 1 1.47 0 32 0 1 2.20 0 32 0 1 2.53 0
25 0 1 1.78 0 23 1 1 2.57 0 22 1 1 2.32 0 20 0 1 2.01 0
19 0 0 2.05 0 17 0 0 2.16 0 16 1 1 3.60 0 13 1 0 2.88 0
11 0 0 2.60 0 10 0 0 2.70 0 10 1 0 2.96 0 9 0 0 2.80 0
7 1 0 4.43 0 6 0 0 3.20  0 6 1 0 2.31  0 6 1 1 4.06  0
6 1 0 3.28  0 23 1 1 1.97  1 22 1 0 2.73  1 17 1 0 2.95  1
15 1 0 2.30  1 12 1 0 1.50  1 12 1 0 3.06  1 11 1 0 3.49  1
11 1 0 2.12  1 8 1 0 3.52  1 8 1 0 3.05  1 8 1 0 2.32  1 8 1 1 3.26  1
5 1 1 3.49  1 5 1 0 3.97  1 4 1 1 4.36  1 4 1 1 2.42  1 3 1 1 4.01  1
2 1 1 4.91  1 2 1 1 4.48  1 1 1 1 2.80  1 1 1 1 5.00  1
;
run;
```

¿Cómo obtener residuales en SAS

```
proc lifereg data=exp;
model Time*status(0)=logWBC Grupo / dist=exponential;
output out=Superv cdf=F cres=coxsnell sres=stutres;
run;
goptions ftext=SWISS ctext=BLACK htext=1 cells;
symbol v=SQUARE c=BLUE h=1 cells;
proc univariate data=Work.Superv   noprint;
var COXSNELL;
probplot / caxes=BLACK cframe=CXF7E1C2 waxis= 1
hminor=0 vminor=0   name='PROB'
exponential( sigma=est theta=est color=BLUE l=1 w=1);inset exponential;
run;
```

¿Cómo obtener residuales en SAS

Se obtiene:

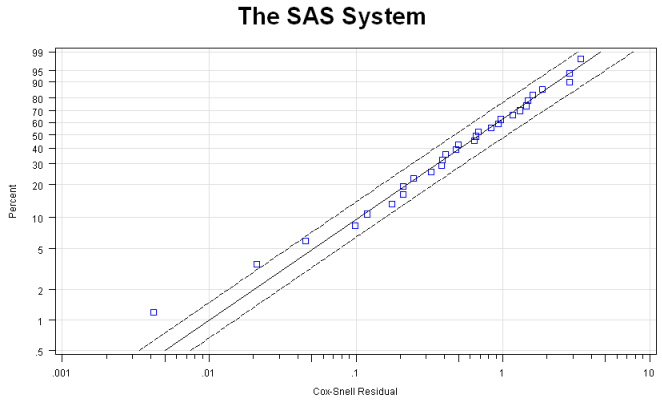


¿Cómo obtener residuales en SAS

Otra forma de obtener un probability plot con bandas simultáneas para los residuales

```
proc lifereg data=Superv;
model coxsnell*status(0)=logWBC Grupo / dist=exponential;
probplot nocenplot;
run;
```

Se obtiene:



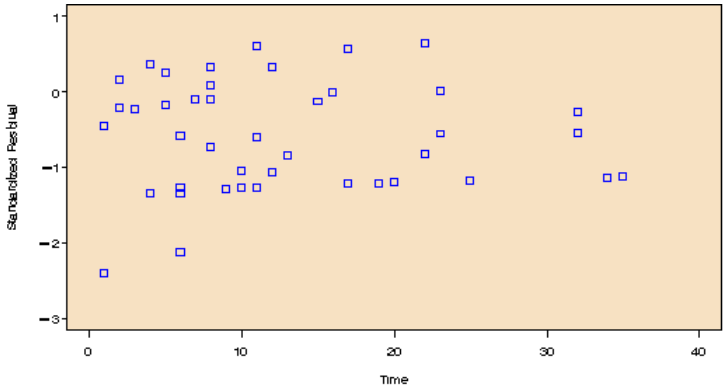
Este gráfico confirma que estos residuales se pueden ver como provenientes de una exponencial.

¿Cómo obtener residuales en SAS

Para los residuales estandarizados:

```
proc gplot data=superv;
plot stutres*time=1;
symbol1 v=circle;
run;
```

Se obtiene:



¿Cómo obtener residuales en SAS

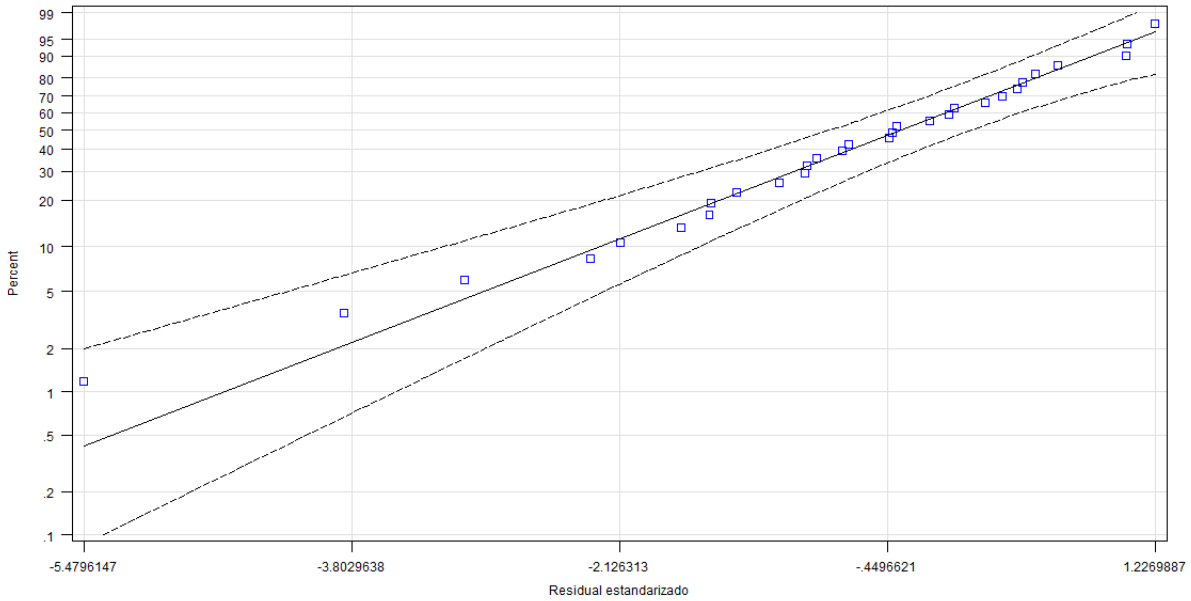
Ahora se ajustará un modelo Weibull:

```
proc lifereg data=exp;
model Time*status(0)=logWBC Grupo / dist=weibull;
output out=Superv cdf=F cres=coxsnell sres=stutres;
probplot nocenplot;
run;

symbol v=SQUARE c=BLUE h=1 cells;
proc lifereg data=superv;
model stutres*status(0)=logWBC Grupo / dist=weibull nolog;*Opcion nolog
especifica una DVES;
probplot nocenplot;
run;
```

¿Cómo obtener residuales en SAS

Se obtiene:



Este gráfico confirma que los errores se comportan como una DVES. Parece que el modelo Weibull ajusta mejor que el exponencial.

Modelo de riesgos proporcionales de Cox

Un modelo de regresión paramétrico trata de alcanzar simultáneamente dos objetivos:

1) Describir la distribución subyacente del tiempo de falla (componente de error)

2) Caracterizar cómo esa distribución cambia en función de las covariables (componente sistemática)

Modelo de riesgos proporcionales de Cox

En algunos campos de aplicación (por ejemplo, confiabilidad) es importante contar con modelos que satisfagan ambos objetivos, pero en otros campos (por ejemplo, epidemiología) se requiere que el modelo satisfaga el objetivo 2 ya que las inferencias se basan casi siempre en las estimaciones de los parámetros que conforman la componente sistemática.

Modelo de riesgos proporcionales de Cox

Los modelos que no requieren que se especifique una distribución específica para representar los tiempos de falla se conocen como Modelos Semi-paramétricos.

Cuando se introdujeron los modelos paramétricos se indicó cómo se podían especificar estos en términos de la función hazard.

Modelo de riesgos proporcionales de Cox

Por ejemplo, el modelo Weibull permite escribir su función hazard como:

$$\begin{aligned} h(t;\mathbf{x}) &= \left(\frac{1}{\sigma}t^{\frac{1}{\sigma}-1}\right)e^{-\frac{\beta_0}{\sigma}-\frac{1}{\sigma}\sum_{j=1}^{k}\beta_j x_j} \\ &= h_0(t)e^{-\frac{\beta_0}{\sigma}-\frac{1}{\sigma}\sum_{j=1}^{k}\beta_j x_j} \end{aligned}$$

El hazard baseline está completamente especificado $h_0(t) = \frac{1}{\sigma}t^{\frac{1}{\sigma}-1}$.

La función hazard permite especificar modelos de regresión al dejarse escribir en función del tiempo y de las covariables.

Modelo de riesgos proporcionales de Cox

De manera análoga y con el fin de relajar algún supuesto distribucional, se puede pensar en formular un modelo semi-paramétrico donde el hazard baseline permanezca sin específicar, es decir,

$$h(t;\mathbf{x}) = h_0(t)e^{\sum_{j=1}^{k}\beta_j x_j}$$

que satisfaga el supuesto de hazards o riesgos proporcionales. En esta expresión $h_0(t)$ es el hazard baseline. Este modelo se conoce como Modelo de hazards proporcionales de Cox*.

*Cox, D.R. (1972) Regression models and life tables (with discussion). *J.R. Statist. Soc.* B **34**, 187-220

Modelo de riesgos proporcionales de Cox

$h_0(t)$ también puede verse como una 'versión inicial del hazard', previo a la consideración de cualquier covariable, ya que

$$\begin{aligned} h(t;\mathbf{x}) &= h_0(t)e^{\sum_{j=1}^{k}\beta_j x_j} \\ &= h_0(t)e^{0} \\ &= h_0(t)\times 1 \\ &= h_0(t) \end{aligned}$$

Conclusión: Si no hay covariables en el modelo $h(t;\mathbf{x}) = h_0(t)$, $(0 < h(t;\mathbf{x}) < +\infty)$.

Modelo de riesgos proporcionales de Cox

Notas:

1) Se considerarán covariables que no dependen del tiempo. Cuando las covariables dependen del tiempo el modelo de Cox ya no satisface la propiedad de hazard proporcionales (HP) y se denomina Modelo de Cox extendido.

2) El modelo de Cox es robusto en el sentido de que aunque el hazard baseline no está especificado, usualmente proporciona estimaciones razonables del vector de parámetros β. Este modelo produce hazards muy flexibles que incluyen el de forma de bañera.

Modelo de riesgos proporcionales de Cox

3) Los resultados que se obtiene con el modelo de Cox son muy similares a los que se obtienen cuando el modelo paramétrico está bien especificado. Por ejemplo, si el modelo paramétrico correcto es el Weibull, el modelo de Cox producirá resultados cercanos a aquellos producidos por el Weibull.

4) Si se tiene un buen grado de certeza del modelo a usar*, prefiera usar un modelo paramétrico en vez del modelo de Cox. En caso de duda el modelo de Cox es una elección frecuentemente segura y el usuario no tiene que preocuparse por la forma distribucional de los tiempos de falla.

*Aunque en general, nunca hay seguridad absoluta de que un modelo particular sea el correcto, solo se puede argumentar que es útil.

Modelo de riesgos proporcionales de Cox

Considere de nuevo la ecuación:

$$h(t;\mathbf{x}) = h_0(t)e^{\sum_{j=1}^{k}\beta_j x_j}$$

Este ecuación es el producto de dos factores: un hazard baseline $h_0(t)$ que no se especifica pero se asume positivo y una función exponencial elevada a un término lineal que es función de las covariables. $h_0(t)$ caracteriza la forma en que el hazard cambia como función del tiempo de supervivencia, mientras que $e^{\sum_{j=1}^{k}\beta_j x_j}$ caracteriza la forma en la que el hazard cambia como función de las covariables a la vez que garantiza que el hazard es positivo. Una característica del modelo de Cox es que aún si $h_0(t)$ no se especifica todavía es posible estimar el vector de parámetros β de la función $e^{\sum_{j=1}^{k}\beta_j x_j}$ (matricialmente $e^{\beta^T \mathbf{x}}$)

Modelo de riesgos proporcionales de Cox

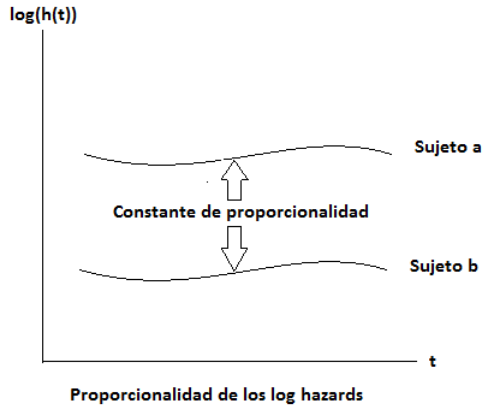
El modelo de Cox se denomina de 'hazards proporcionales' ya que la razón de hazards (HR) para dos sujetos a y b (asumiendo que las covariables asociadas a cada uno de ellos no dependen del tiempo) satisface:

$$\begin{aligned} HR &= \frac{h(t;\mathbf{x}^{(a)})}{h(t;\mathbf{x}^{(b)})} \\ &= \frac{h_0(t)e^{\beta^T\mathbf{x}^{(a)}}}{h_0(t)e^{\beta^T\mathbf{x}^{(b)}}} \\ &= e^{\beta^T\left(\mathbf{x}^{(a)}-\mathbf{x}^{(b)}\right)} \end{aligned}$$

Esta razón NO depende del tiempo (es decir, el HR es constante en el tiempo, solo depende de la diferencia en los valores de la covariables). El HR permite medir el tamaño del efecto de una covariable particular.

Modelo de riesgos proporcionales de Cox

Si se grafican los logaritmos de los hazards de 2 individuos cualquiera, la propiedad de proporcionalidad en los hazards implica que estas funciones deben ser paralelas:



Proporcionalidad de los log hazards

En este caso la constante de proporcionalidad está dada por $\beta^T \left(\mathbf{x}^{(a)} - \mathbf{x}^{(b)} \right)$.

Modelo de riesgos proporcionales de Cox

Formulación del modelo de Cox de HP en términos del hazard acumulado y la función de supervivencia. ¿Cuál es la función de supervivencia para un modelo con hazard $h(t;\mathbf{x}) = h_0(t)e^{\beta^T \mathbf{x}}$? Esta pregunta se puede responder notando que:

$$S(t;\mathbf{x}) = e^{-H(t;\mathbf{x})}$$

Donde $H(t;\mathbf{x})$ es el hazard acumulado para un sujeto con covariables $\mathbf{x} = (x_1, x_2, \ldots, x_k)$.

Modelo de riesgos proporcionales de Cox

Asumiendo que el tiempo de supervivencia es continuo,

$$\begin{aligned} H(t;\mathbf{x}) &= \int_0^t h(u;\mathbf{x})du \\ &= \int_0^t h_0(u)e^{\beta^T\mathbf{x}}du \\ &= e^{\beta^T\mathbf{x}}\int_0^t h_0(u)du \\ &= e^{\beta^T\mathbf{x}}H_0(t) \end{aligned}$$

Modelo de Cox en términos del hazard acumulado. En esta expresión $H_0(t)$ es el hazard baseline acumulado. Esta relación puede pensarse como una medida del riesgo acumulado baseline el cual se modifica de acuerdo a la función $e^{\beta^T\mathbf{x}}$.

Modelo de riesgos proporcionales de Cox

A partir de la relación anterior, se puede formular el modelo de Cox en términos de la supervivencia:

$$\begin{aligned} S(t;\mathbf{x}) &= e^{-H(t;\mathbf{x})} \\ &= e^{-e^{\beta^T \mathbf{x}} H_0(t)} \\ &= \left[e^{-H_0(t)}\right]^{e^{\beta^T \mathbf{x}}} \\ &= [S_0(t)]^{e^{\beta^T \mathbf{x}}} \end{aligned}$$

Modelo de Cox en términos de la supervivencia*. En esta expresión $S_0(t)$ es el supervivencia baseline.

*Permite obtener curvas de supevivencia ajustadas por covariables.

Modelo de riesgos proporcionales de Cox

De esta última relación, una estimación de la curva de supervivencia corregida por covariables está dada por:

$$\widehat{S(t;\mathbf{x})} = \left[\widehat{S_0(t)}\right]^{e^{\hat{\beta}^T \mathbf{x}}}$$

Donde $\widehat{\beta}$ es una estimación del vector β y $\widehat{S_0(t)}$ es una estimación de la supervivencia baseline $S_0(t)$. Pero ¿Cómo obtener $\widehat{S_0(t)}$? SAS permite obtener a $\widehat{\beta}$ y a $\widehat{S_0(t)}$. Se verán los detalles de la manera en que se estima la supervivencia baseline.