Modelo de riesgos proporcionales: Validación

Giancarlo Sal y Rosas

Departmento de Ciencias
Pontificia Universidad Católica del Perú

May 26, 2017

Outline

Estudio: Prevención de abuso de drogas

```
> model1    <- coxph(Surv(time, censor)~ treat + age+beck+ndrugtx+age*site + race*site,data=u
> summary(model1)
Call:
coxph(formula = Surv(time, censor) ~ treat + age + beck + ndrugtx +
    age * site + race * site, data = uis, method = "breslow")

  n= 575, number of events= 464

              coef exp(coef)  se(coef)      z Pr(>|z|)
treat     -0.279626  0.756066  0.094292 -2.966 0.003021 **
age       -0.034082  0.966492  0.009677 -3.522 0.000428 ***
beck       0.009004  1.009045  0.004893  1.840 0.065724 .
ndrugtx    0.034932  1.035549  0.007996  4.369 1.25e-05 ***
site      -1.270718  0.280630  0.527412 -2.409 0.015981 *
race      -0.495251  0.609418  0.131416 -3.769 0.000164 ***
age:site   0.029814  1.030263  0.016068  1.855 0.063528 .
site:race  0.850385  2.340548  0.246461  3.450 0.000560 ***

          exp(coef) exp(-coef) lower .95 upper .95
treat        0.7561     1.3226   0.62849    0.9095
age          0.9665     1.0347   0.94833    0.9850
beck         1.0090     0.9910   0.99941    1.0188
ndrugtx      1.0355     0.9657   1.01945    1.0519
site         0.2806     3.5634   0.09982    0.7890
race         0.6094     1.6409   0.47104    0.7885
age:site     1.0303     0.9706   0.99832    1.0632
site:race    2.3405     0.4273   1.44387    3.7941
```

Estudio: Proceso de validación

- Este modelo no es el modelo final estrictamente hablando. Debemos realizar un diagnóstico para verificar si
  - El modelo ajusta adecuadamente los datos ?
  - Se cumple el supuesto de proporcionalidad ?
  - Existen observaciones que no son descritas adecuadamente por el modelo ?
  - Existen observaciones que son mas influyentes que otras ?

Técnicas de diagnóstico

Las preguntas planteadas pueden ser evaluadas graficamente mediante el calculo de residuos:

- Residuos de Cox-Snell
- Residuos de martingalas
- Residuos de score & delta-beta
- Residuos schoenfeld

Outline

Residuos de Martingalas

- En base a la teoria de conteo, el modelo de riesgos proporcionales se expresa como un modelo lineal

$$N(t) = \Lambda(t \mid Z) + M(t)$$

donde

  - $N(\cdot)$ representa la parte observada del modelo: El número de eventos
  - $\Lambda(\cdot)$ es el componente sistemático: Nuestro modelo
  - $M(\cdot)$ es el componente del error (Martingala): Juega un papel similar que en los modelos que conocen:
    - Su valor esperado es 0
    - Las observaciónes no estan correlacionadas

Residuos de Martingalas

Supongamos que para un sujeto el evento ocurre a los 200 dias y el máximo tiempo de seguimiento es un año:

- Conteo

$$N(t) = \begin{cases} 0, & t < 200 \\ 1, & t \geq 200 \end{cases}$$

- Componente sistemático

$$\Lambda(t) = \begin{cases} \Lambda_0(t)e^{Z\beta}, & t < 200 \\ \Lambda_0(200)e^{Z\beta}, & t \geq 200 \end{cases}$$

- En ambos casos el máximo valor que pueden tomar ambos componentes ocurre al final del seguimiento.

Residuos de Martingalas

- Es natural pensar en una forma de residuo (residuo de martingala) via

$$M(t) = N(t) - \Lambda(t \mid Z, \beta)$$

- Note que en teoria se puede calcular este residuo para cada tiempo. Sin embargo, su mayor utilidad se da al calcularlo al final del tiempo de seguimiento:

$$\begin{aligned} \hat{r}_{M,i} &= \delta_i - \hat{\Lambda}(t_i \mid Z_i) \\ &= \delta_i - \hat{\Lambda}_0(t_i) \exp(Z\hat{\beta}) \end{aligned}$$

donde
  - δ_i es el indicador de censura
  - $\hat{\Lambda}_0$ es la función acumulada de riesgo basal estimada en el tiempo máximo del individuo t_i

Residuos de Martingalas

- $r_{M,i}$ puede entenderse como la diferencia entre el número esperado de eventos (0 o 1) para el sujeto *ith* entre el tiempo $[0, t_i]$ y el número esperado de eventos de acuerdo al modelo
- Propiedades:
  - El valor esperado es 0
  - El rango de $r_{M,i}$ es $(-\infty, 0]$
  - Son aproximadamente no correlacionados (para muestras grandes)
- Se puede graficar $\hat{r}_{M,i}$ vs. una variable Z_j $(j < p)$ para analizar si esta ajusta adecuadamente los datos.

Residuos de Martingalas

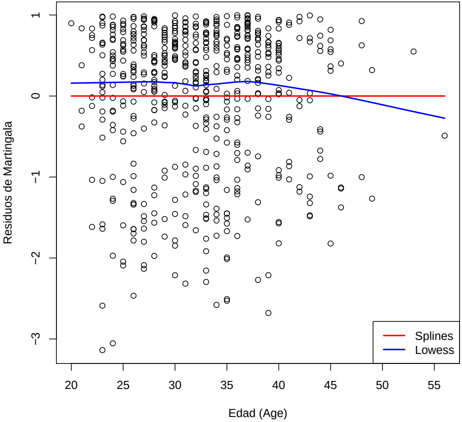

Figure : Residuos de Martingalas para edad

Residuos de Martingalas

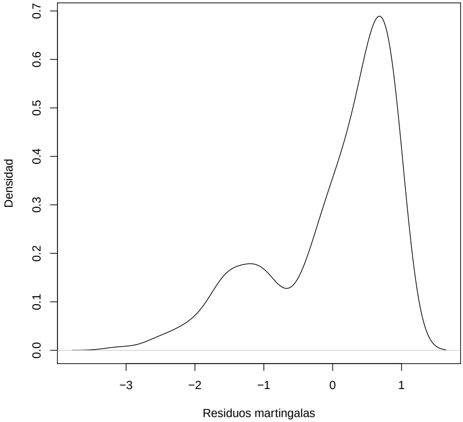

Figure : Residuos de Martingalas

Residuos de Martingalas

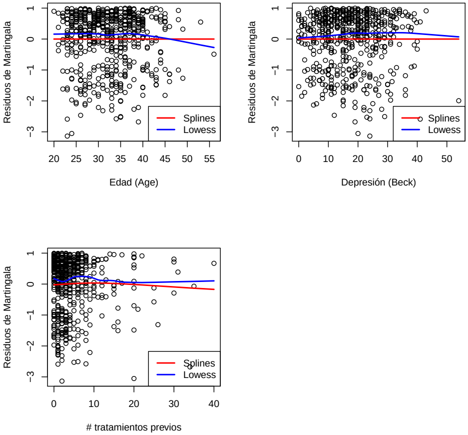

Figure : Residuos de Martingalas para las variables continuas edad, nivel de depresión y número de tratamientos previos

Outline

1. Motivación

2. Residuos
   - Residuos de Martingalas
   - Residuos de Cox-Snell
   - Residuos de desviación
   - Residuos de Score

3. Proporcionalidad
   - Residuos Schoenfeld

Motivación

- Recordemos que si $T \sim F$ donde F es continua, entonces

$$P(F(T) \leq y) = P(X \leq F^{-1}(y)) = F[F^{-1}(y)] = y$$

entonces $F(T) \sim U(0,1)$ y dado que esta distribución es simetrica

$$S(T) \sim U(0,1)$$

- Adicionalmente

$$\begin{aligned} P[\Lambda(T) \leq a] &= P[-\log(S(T)) \leq a] = P[S(T) \geq e^{-a}] \\ &= 1 - e^{-a} \end{aligned}$$

entonces

$$\Lambda(T) \sim Exp(1)$$

Residuos de Cox-Snell

- Supongamos que el modelo de Cox (que hemos construido) es correcto. Es decir

$$\lambda(t \mid Z_i) = \lambda_0(t) \exp(Z_i \beta)$$

o equivalentemente

$$\Lambda(t \mid Z_i) = \Lambda_0(t) \exp(Z_i \beta)$$

- El resultado anterior nos dice que

$$\{\hat{\Lambda}_i = \hat{\Lambda}(t \mid Z_i)\}_{i=1}^{n}$$

debe ser una muestra aleatoria de observaciones de una distribución exponencial éstandar sujeta a censura por la derecha.

Residuos de Cox-Snell

- Definimos los residuos de Cox-Snell como

$$r_{CS,i} = \hat{\Lambda}(t \mid Z_i) = \hat{\Lambda}_0(t) \exp(Z_i \hat{\beta})$$

y note que este se puede expresar en función de los residuos de Martingalas

$$r_{CS,i} = \delta_i - r_{M,i}$$

- Proceso
  - Sea $Y_i = \hat{\Lambda}_i(t \mid Z_i)$
  - Graficamos $-\log(S(Y_i))$ vs. Y_i y deberia ser una linea recta
  - Graficamos $\log(-\log(S(Y_i)))$ vs. $\log(Y_i)$ y deberia ser una linea recta con pendiente 1.

Residuos de Cox-Snell

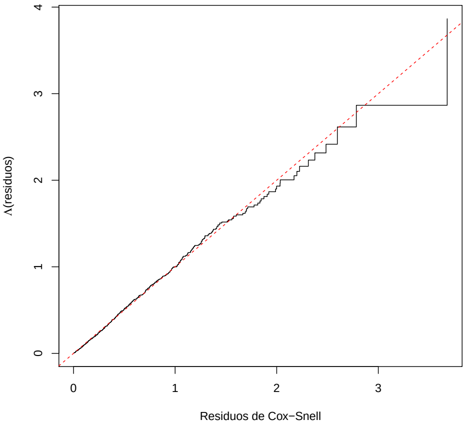

Figure : Residuos de Cox-Snells para el modelo

Residuos de Cox-Snell

Cuidado:

- En el caso de no linealidad, el gráfico no nos dice cual lejos del modelo estamos.
- Esto no necesariamente implica que los tiempos que estamos estudiando tienen una distribución exponencial
- Usamos $\hat{\beta}$ y $\hat{\Lambda}_0$ (pues no conocemos β ni Λ_0), entonces $\hat{\Lambda}(t)$ no necesariamente tiene una distribución exponencial.

Outline

Residuos de desviación

- Los residuos de martingalas tienden a ser asimétricos
- Los residuos de desviación estan definidos por:

$$r_{d,i} = signo(\hat{r}_{M,i})\sqrt{2}\sqrt{-\hat{r}_{M,i} - \delta_i \log\left(\delta_i - \hat{r}_{M,i}\right)}$$

- Estos residuos tienen una distribución mas simétrica y por ende pueden ser aproximados por la distribución normal estándar
- Se puede graficar los residuos vs. el predictor lineal o covariables individuales

Residuos de Desviación

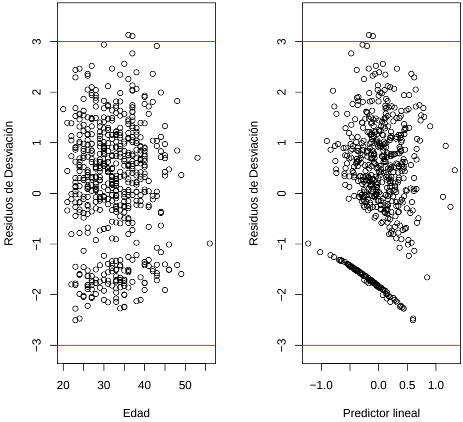

Figure : Identificación de valores extremos

Outline

Residuos de Score

Recordemos que la verosimilitud parcial es

$$L_p(\beta) = \prod_{i=1}^{n} \left[\frac{e^{z_i^t \beta}}{\sum_{j \in R_i} e^{z_j^t \beta}} \right]^{\delta_i}$$

Por ende la función de score tiene la forma

$$U_{\beta_k} = \frac{\partial \log(L_p)}{\partial \beta_k} = \sum_{i=1}^{n} \delta_i \left[z_{ik} - \frac{z_{ik} e^{z_i^t \beta}}{\sum_{j \in R_i} e^{z_j^t \beta}} \right] = \sum_{i=1}^{n} \delta_i \left[z_{ik} - \bar{z}_{w_i k} \right]$$

donde $\bar{z}_{w_i k}$ es el estimador del valor esperado del factor en el conjunto de riesgo $R(i) = R(t_i)$

Residuos de Score

Notemos que

$$\hat{\hat{z}}_{w_i k} = \frac{z_{ik} e^{z_i^t \hat{\beta}}}{\sum_{j \in R_i} e^{z_j^t \hat{\beta}}}$$

entonces

- La diferencia

  $$(z_{ik} - \hat{\hat{z}}_{w_i k})$$

  mide el grado de diferencia del individuo i con respecto al grupo en riesgo.
- Se calcula para cada covariable
- No esta definido para datos censurados
- La suma de $r_{ik} = (z_{ik} - \hat{\hat{z}}_{w_i k})$ es igual a zero.

Residuos de Score

- El residuo del proceso de score para el individuo *ith* en la covariable *kth* es

$$L_{ik} = \sum_{j=1}^{n} (z_{ik} - \bar{z}_{w_j}) dM_i(t_j)$$

  donde $dM_i(t_j)$ es el cambio en el residuo de martingala para el sujeto *ith* en el tiempo t_j

- El residuo de score es el estimador de L_{ik} ($r_{SC,ik}$)

Residuos de Score

- El sujeto *ith* tiene k residuos de score asociados

$$r_{SC,i} = (\hat{r}_{SC,i1}, \ldots, \hat{r}_{SC,ik})$$

donde

$$r_{SC,il} = \sum_{t_{(k)} \leq t_{(i)}}^{n} (z_{il} - \hat{\bar{z}}_l) r_{M,k}$$

- Largos valores de $r_{SC,il}$ implica importante influencia del la observación i-ésima tanto en el tiempo como en la covariable.

Residuos de Score

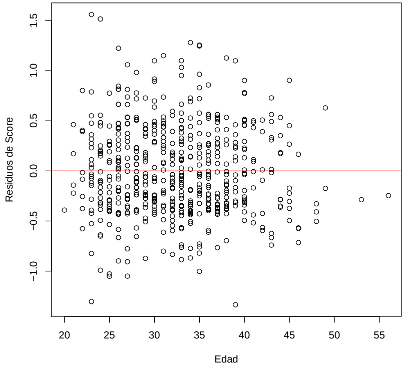

Figure : Residuos de Score para la variable edad

Residuos de Score

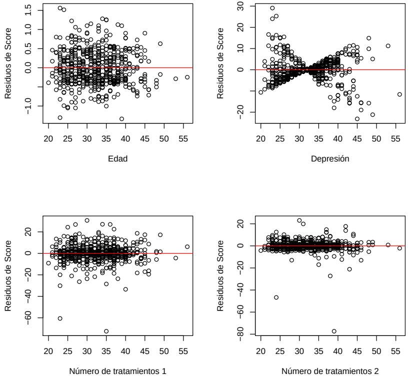

Figure : Residuos de Score pra las cuatro primeras covariables del modelo

Puntos de influencia

- Valores extremos son aquellos tiempos extremos observados (W_i, Δ_i) dados las variables Z_i
  - Valores de los residuos de martigalas o desviación largos.
- Existen tambien observaciones que son valores atípicos en el plano de las covariables Z.
- **Alta influencia**: Una observación que es la conbinación de ambos casos.
  - Esto nos indica que tiene alta influencia en la estimación del vector de regresión $\hat{\beta}$.

Delta-beta

- Supongamos que $\hat{\beta}_k$ es el estimador de β_k usando el total de datos
- Supongamos que $\hat{\beta}_{k(-j)}$ es el estimador de β_k retirando la observación (W_j, Δ_j, X_j)
- Definimos los valores delta-beta ($\Delta\beta_{kj}$) como

$$\Delta\beta_{kj} = \hat{\beta}_k - \hat{\beta}_{k(-j)}$$

  que es una medida de la influencia de la observación j-ésima en la estimación de $\hat{\beta}_k$

Delta-beta

- En principio, esto implicaria construir $n+1$ modelos. Por suerte podemos aproximar $\Delta\beta_{kj}$ mediante

$$\Delta\beta_{kj} = \hat{\beta}_k - \hat{\beta}_{k(-j)} \approx \hat{V}_j \times r_{SC,j}$$

  donde
  - $\hat{V}_j$ es la j-ésima fila de la matriz de varianza de los estimadores $\hat{\beta}$
  - $r_{SC,j}$ son los residuos score del sujeto j
- Cada observación, tiene un valor de $\Delta\beta$ para cada covariable del modelo

Residuos de Score

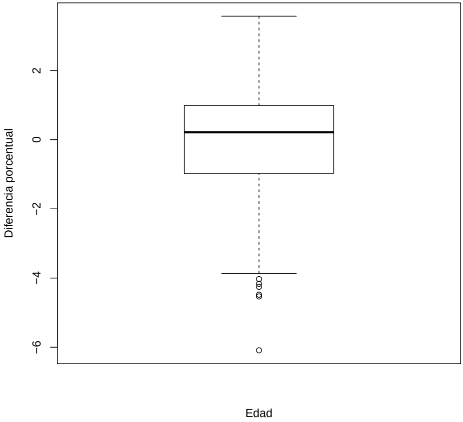

Figure : Distribución de $\Delta\beta_{edad}$ para la variable edad

Residuos de Score

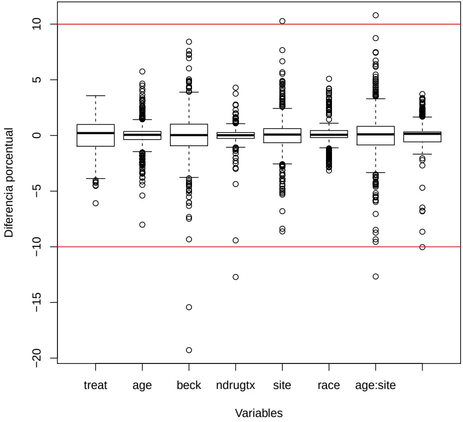

Figure : Distribución de $\Delta\beta$ porcentual para las covariables del modelo

Residuo de Score

- ¿ Del grafico podemos ver que algunas observaciones afectan en un 10% la estimación del coeficiente ?

```
> round(per_rdb[rowSums(abs(per_rdb) > 10)>0,],3)
     treat    age    beck ndrugtx   site   race age:site site:race
7    3.569  0.783  -1.140 -12.710  2.207 -0.744    3.683    -1.421
338  3.476  2.568 -19.281   1.976  1.971 -2.053    3.112    -0.671
372  3.304 -5.397 -15.426   0.796 -5.079 -2.842   -5.841    -1.303
519 -4.253 -3.317  -3.221  -9.421 10.276  0.752   10.798     1.873
573 -4.481 -0.722  -3.446  -1.087 -2.099 -0.019   -2.715   -10.030
585 -2.180  0.055   4.427   1.143 -8.620 -0.084  -12.668     2.163

> uis[rowSums(abs(per_rdb) > 10)>0,]
     id age beck hercoc ivhx ndrugtx race treat site los time censor   ok nivhx
7     7  39   19      4    3      34    0     1    0 179  459      1 TRUE     1
338 338  35   54      4    2       1    0     1    0  29  621      0 TRUE     0
372 372  23   41      3    1       1    0     1    0 144  546      0 TRUE     0
519 519  24   20      3    2      20    0     0    1 108  540      0 TRUE     0
573 573  35   23      3    1       5    1     0    1 183  540      0 TRUE     0
585 585  49    4      4    2       2    0     0    1 177  547      0 TRUE     0
```

Variable de depresión

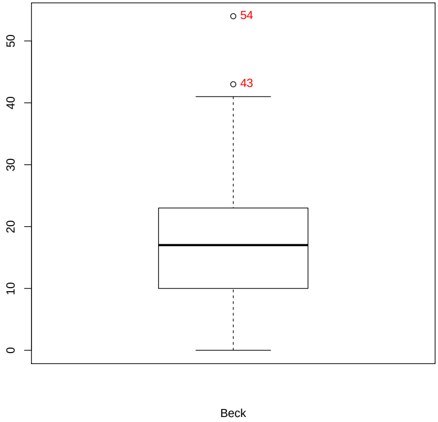

Figure : Distribución de la variable de depresión

Outline

Residuos Schoenfeld

- Sea $D(t_{(k)})$ es el conjunto de fallas en el tiempo $t_{(k)}$
- Consideremos la diferencia entre el valor de la covariable del individuo z_{il} donde $i \in D(t_{(k)})$ y el promedio ponderado de la variable en el conjunto

$$z_{il} - \bar{z}_l(\beta, t_{(k)})$$

- El residuo Schoenfeld esta definido por

$$r_{S,lk} = r_{S,lk}(\beta) = \sum_{i \in D(t_{(k)})} \delta_{ik}[z_{il} - \bar{z}_l(\beta, t_{(k)})]$$

donde δ_{ik} es igual a 1 si el sujeto i falla en $t_{(k)}$ y 0 en otro caso.

Residuos Schoenfeld

- Bajo el supuesto de que el modelo de proporcionalidad es el adecuado, se tiene
  - Los residuos tienen valor esperado 0
  - Los residuos no tienen correlación
- En la práctica los residuos son estimados por

$$\hat{r}_{S,lk} = r_{S,lk}(\hat{\beta})$$

- Estos residuos se pueden estandarizar

$$r^*_{S,k} = r^*_{S,k}(\beta) = Var(\hat{\beta})^{-1} r_{S,k}(\beta)$$

Residuos Schoenfeld

- Si una variable en específico tiene un efecto que cambia en el tiempo
$$\beta_l(t) = \beta_l + \rho g(t)$$
- Las propuestas mas comunes para $g(t)$ son $g(t) = t$, $g(t) = \log(t)$
- Entonces, se puede comprobar que
$$E[\hat{r}_{S,lk}] \approx \rho g(t_{(k)})$$

Residuos Schoenfeld

- Este resultado sugiere graficar $\hat{r}_{S,lk}$ vs. $g(t_{(k)})$ para examinar violaciones al supuesto de proporcionalidad.
- Adicionalmente podrias estudiar la hipótesis

$$H_0 : \rho = 0 \quad , \quad H_a : \rho \neq 0$$

- En R toda esta implementación se hace con la función **cox.zph**

Función cox.phz

```
> ph <- cox.zph(m11)
> ph
               rho    chisq      p
age       0.041178 7.63e-01 0.3823
beck     -0.084863 3.14e+00 0.0763
ndrugfp1  0.000593 1.62e-04 0.9898
ndrugfp2  0.009611 4.26e-02 0.8364
nivhx    -0.003690 6.43e-03 0.9361
race      0.052684 1.30e+00 0.2537
treat     0.091425 3.90e+00 0.0483
site      0.047981 1.05e+00 0.3059
agesite  -0.043967 8.62e-01 0.3532
racesite -0.017842 1.48e-01 0.7005
GLOBAL          NA 1.12e+01 0.3450
```

- La variable que mide depresión y la variable que mide el efecto del tratamiento no cumplen la proporcionalidad
- La prueba de proporcionalidad global no es significativa.

Función cox.zph

```
> phl <- cox.zph(m11,transform="log")
> phl
                rho  chisq     p
age         0.03799 0.6497 0.420
beck       -0.06494 1.8402 0.175
ndrugfp1   -0.00154 0.0011 0.974
ndrugfp2    0.00451 0.0094 0.923
nivhx      -0.01127 0.0600 0.807
race        0.04349 0.8878 0.346
treat       0.06477 1.9564 0.162
site        0.05065 1.1682 0.280
agesite    -0.05269 1.2378 0.266
racesite   -0.00731 0.0248 0.875
GLOBAL           NA 6.8007 0.744
```

- Usando $g(t) = \log(t)$, resulta que no hay evidencia de falta de proporcionalidad.
- Podemos entender un poco mas el problema si lo analizamos graficamente.

Función cox.zph

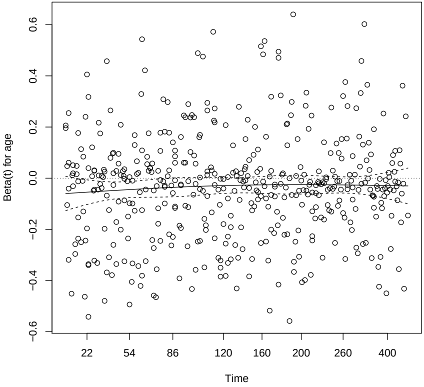

Figure : Estudio de proporcionalidad de la variable treat usando $g(t) = t$

Función cox.zph

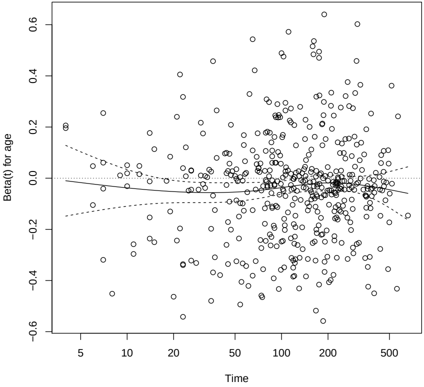

Figure : Estudio de proporcionalidad de la variable treat usando $g(t) = \log(t)$

Función cox.zph

- Al estudiar nuestro modelo, no encontramos evidencia para rechazar la hipótesis de proporcionalidad de las variables incluidas en este
- El hecho que no podamos rechazar $H_0 : \rho = 0$, no implica que sea cierto.
- En la practica, probablemente nunca se cumple la proporcionalidad
- El diagnostico grafico es muy provechoso en el proceso de decisión.

Referencias

Amber, G. and Royston, P. (2001). Fractional polynomial model selection procedures: Investigation of type i error rate. *Journal of Statistical Simulations and Computation*, 69(1):89–108.

Braga, A., Bressan, V., Colosimo, E., and Bressan, A. (2006). Investigating the solvency of Brazilian credit unions using proportional hazard model. *Annals of Public and Cooperative Economics*, 77(1):83–106.

Royston, P. and Altman, D. (1994). Regression using fractional polynomials of continuos covariates: parsimonious modelling. *Applied Statistics*, 43(3):419–467.

Sánchez, J., Sal Y Rosas, V., Hughes, J., Baeten, J., Fuchs, J., Buchbinder, S., Koblin, B., Casapia, M., Ortiz, A., and Celum, C. (2011). Male circumcision and risk of hiv acquisition among msm. *AIDS*, 25(4):519–23.