Análisis de tiempos hasta que se produce un error de concordancia en cuatro estudiantes italianos de ELE. Material suplementario.

1. Nociones de selección de modelos según la teoría de la información

Se pueden usar las siguientes medidas para seleccionar modelos (Stroup, 2013; p. 193; Burnham & Anderson, 2010, caps. *2* y *4*) [$\theta = \hat{\sigma}, \hat{\beta}$ es el vector de los coeficientes fijos y aleatorios estimados]:

(a) *Schwarz*: $BIC = -2L(\theta) + (p \times log(s))$ [s= número de grupos; y $p = p_\sigma + (p_\beta = rank[X])$; o sea el número de parámetros fijos más los aleatorios]. Menos es mejor.

(b) *Akaike*: $AIC = -2L(\theta) + 2p$. Menos es mejor.

(c) *Akaike corregido*: $AICc = -2L(\theta) + 2p(n^*/(n^* - p - 1))$ [$n^* = N$, tamaño muestral]; Corrige por muestra pequeña. Menos es mejor. Como heurística, se debería usar cuando[1]: $\frac{n}{p} < 40$.

(d) Delta de Akaike: $\Delta AIC = \Delta = AIC_i - AIC_{min}$. Indican la distancia del modelo al mejor de todos (el de menor AIC). $\Delta_i \leq 2$ indica evidencia substancial para el modelo i.

(e) Pesos de Akaike (ω_i): indican el peso de la evidencia en favor de que modelo sea el mejor de entre todos los modelos candidatos. Es decir, responde a la pregunta: ¿Cómo soportan los datos al modelo i con respecto al resto de los modelos? Se define como:

$$\omega_{modelo_i} = \frac{exp(-\Delta_i/2)}{\sum_{i=1}^{R} exp(-\Delta_i/2)}$$

donde $i = 1, \ldots, R$ son los modelos considerados; y $\sum_{i=1}^{R} w_i = 1$.

(g) Ratio de evidencia ("Evidence Ratio", [ER]): Ratio entre el peso de Akaike del modelo i-ésimo y el peso de Akaike del j-*ésimo* modelo: $\frac{W(i)}{W(j)}$. Muchas veces resulta de interés establecer i como el índice del mejor modelo: $\frac{W(1)}{W(j)}$. Los *ER* son invariantes a los demás modelos, a parte de i y j. Responden a la pregunta: ¿Cuántas más veces apoyan los datos al (mejor) modelo i respecto del modelo j?

Una vez ordenados los modelos según alguno de los criterios, se puede reducir dicho conjunto por medio de un "conjunto de confianza" [*confidence set*] para el mejor modelo hallado. Burnham & Anderson (2010, p. 169) plantean tres alternativas: (i) sumar los pesos de Akaike de los modelos hasta alcanzar ≥ 0.95 (recuérdese que los pesos de Akaike suman *1*); (ii) tomar los modelos tal que

[1] Las medidas AIC y AIC_c convergen para n grande (manteniendo p constante). Es decir que cuando dicho ratio es suficientemente grande, tienden a seleccionar el mismo modelo. Entonces, en la práctica conviene usar siempre AIC_c.

$\Delta_i \leq 2$, ya que indican evidencia sustancial para el modelo i; (iii) establecer un corte usando ratios de evidencia (poniendo ahora el mejor modelo en el denominador), tal que[2]: $\frac{W(i)}{W(1)} > \frac{1}{8}$ ($\Delta_i = 2$). Los autores prefieren el tercer criterio debido a su invariancia por adición o borrado de modelos del conjunto de confianza.

Resulta imperativo tener en cuenta la incerteza debida al proceso de selección de modelos. De R modelos considerados se selecciona el mejor modelo i. Sin embargo, ¿Si hubieran cambiado los datos, se elegiría igualmente el modelo i como el mejor o habría variabilidad de entre las muestras de datos en cuanto al modelo elegido? Una forma de tener en cuenta dicha incerteza es estimar la probabilidad de que un determinado predictor x_j esté en el mejor modelo si se pudiera recoger una nueva muestra de datos. Se trata de una medida de importancia relativa de los predictores. Se lleva a cabo sumando los pesos de Akaike de los modelos en los cuales el predictor x_j está presente: $W_+ = w_i I_j(g_i)$; donde $I_j(g_i)$ es la función indicadora que es "1" si x_j está en el modelo g_i o cero, si no. Entonces, la importancia relativa es la proporción de modelos en los cuales la predictora está presente.

Si se diera el caso de que, por ejemplo, $w(i) > 0.9$, entonces el modelo i es un claro ganador. En dicho caso es válido hacer inferencia mediante la estimación de los coeficientes β_i y sus errores típicos serán condicionales al modelo seleccionado. Si embargo, muchas veces, especialmente si el conjunto de modelos a considerar es grande, los modelos con $\Delta_i \leq 2$ poseen pesos de Akaike similares o bien deltas de Akaike cercanos al cero. En este caso, β_i puede diferir en los modelos del conjunto considerado. Una solución es usar la información de todos los modelos involucrados mediante un promedio pesado de los coeficientes. En este caso, los errores típicos de los coeficientes estimados no son condicionales al modelo (ganador) en cuestión sino a todo el conjunto de modelos. Por lo tanto, dichos errores típicos "incondicionales" tienen en cuenta la varianza que proviene del proceso de selección de modelos. Para promediar los coeficientes se utilizó:

$$\bar{\beta}_j = \sum_{i=1}^{R} w_i \, I_j(g_i) \;\; \hat{\beta}_{j,i} = W_+ \, \hat{\beta}_{j,i}$$

donde:

$$I_j(g_i) = \begin{cases} 1 & x_j \in g_i \\ 0 & x_j \notin g_i \end{cases}$$

y la suma es sobre todos los modelos del conjunto: $i = 1, \ldots, R$. En este estimador se usan todos los modelos ("full average"), y cuando la predictora x_j no estuviera presente en un determinado modelo entonces $\beta_j = 0$. Tiene la ventaja de "correr

[2] También podrían usarse: *0.135* ($\Delta_i = 4$); *0.082* ($\Delta_i = 5$); *0.05* ($\Delta_i = 6$)

hacia cero" [*Shrinkage*] las estimaciones de parámetros presentes en "modelos malos". La varianza del estimador resulta:

$$\widehat{var}(\bar{\beta}_j) = \left[\sum_{i=1}^{R} w_i \sqrt{\widehat{var}(\bar{\beta}_j|g_i) + (\hat{\beta}_j - \bar{\beta}_j)^2}\right]^2$$

y su error típico: $\sqrt{\widehat{var}(\bar{\beta}_j)}$.

2. Nociones de análisis de supervivencia.

2.1. Modelo proporcional de Cox

La función de sobrevivencia indica la probabilidad de que un individuo "sobreviva" más allá de un cierto tiempo t (o sea que no experimente el evento hasta t): $S(t) = P(T > t) = 1 - F(t), t \in \mathbb{R} \geq 0$; donde $F(t) = P(T \leq t)$ es la función acumulada, es decir, la probabilidad de que el individuo viva menos o igual que t. La función de sobrevivencia es una curva decreciente que vale 1 en $S(t = 0)$ ["todos sobreviven"] y 0 en $S(t = \infty)$ ["nadie sobrevive"]. Por otro lado, la función de riesgo (o tasa de riesgo instantánea) indica la probabilidad de sobrevivir un intervalo Δ corto de tiempo adicional sabiendo que el individuo sobrevivió hasta el tiempo t. Muestra el riesgo de experimentar el evento en cada instante y puede tener cualquier forma funcional: $h(t) = \lim_{\Delta \to 0} \frac{P(t<T<t+\Delta|T>t)}{\Delta}$.

Defínase primero una función de riesgo "promedio" para el individuo típico denotada por $h_0(t)$. Luego, se puede especificar la función de riesgo para un individuo en particular, denotada por $h(t)$, y relacionarla con aquella promedio mediante un ratio de riesgo o *hazard ratio* (HR): $HR = \frac{h(t)}{h_0(t)} = \psi$. Obsérvese que se ha escrito ψ como una constante que *no depende* del tiempo, es decir que, no obstante $h(t)$ y $h_0(t)$ si lo hagan, su ratio se mantiene igual a lo largo del tiempo. Este es el supuesto de HR proporcional. También se puede escribir la expresión anterior como: $h(t) = h_0(t)\psi$, o sea que la función de riesgo de un individuo particular, $h(t)$, se define como el factor ψ que multiplica a la función de riesgo del individuo "promedio", $h_0(t)$. Si no hay covariables en el modelo, entonces $\psi = exp(0) = 1$; en cambio, si existieran, $\psi = exp(\mathbf{x^T\beta})$. El modelo de Cox proporcional para el i-ésimo individuo, dadas las covariables, se escribe pues como:

$$h_i(t \mid \mathbf{x}) = h_0(t)exp(\mathbf{x_i}^T\boldsymbol{\beta}) = h_0 exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right)$$

Por ejemplo, si tuviéramos una sola covariable binaria, $x \in \{0,1\}$, que compara un grupo control con otro tratado, el *hazard* del grupo control sería $h_C(t) = h_0(t)exp\big(\beta_1(x = 0)\big) = h_0(t)$ y el del grupo tratado: $h_T(t) = h_0(t)exp\big(\beta_1(x = 1)\big)$. El *hazard ratio* del grupo tratado relativo al control se escribe del modo siguiente:

$HR = \frac{h_T(t)}{h_C(t)} = \frac{h_0(t)exp(\beta_1)}{h_0(t)} = exp(\beta_1) = \psi$. Nótese que solamente la función $h_0(t)$ depende del tiempo, pero se cancela, por lo tanto el HR resulta proporcional. Entonces, el exponencial de los coeficientes del modelo indica el *HR* de la variable j-ésima, ajustada por las demás variables. Es el factor ψ por el cual se multiplica la función de riesgo si la j-ésima variable aumenta una unidad, manteniendo constantes los valores de las otras covariables. Se tienen las siguientes situaciones: (i) $\psi > 1$, entonces $h_T(t) > h_C(t)$, el riesgo basal *aumenta* en proporción de ψ cuando se usa el tratamiento en lugar del control; (ii) $\psi < 1$, entonces $h_T(t) < h_C(t)$, el riesgo basal *disminuye* en proporción de ψ cuando se usa el tratamiento en lugar del control; (iii) $\psi = 1$, entonces $h_T(t) = h_C(t)$, el riesgo basal no cambia.

El logaritmo del hazard es una función lineal de las variables explicativas:

$$ln[h_i(t \mid \mathbf{x})] = ln[h_0(t)]\mathbf{x_i}^T\boldsymbol{\beta} = ln[h_0(t)]\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right)$$

Donde: $\mathbf{x}$ es el vector de covariables para el individuo i-ésimo y $\boldsymbol{\beta}$, el vector de coeficientes. Cada predictor actúa de forma *multiplicativa* sobre el *hazard basal.* Además $h_0(t)$ puede asumir cualquier forma funcional. Los coeficientes betas se estiman por máxima (log)verosimilitud parcial. Para sacar la varianza de cada coeficiente es preciso calcular la información observada: $I(\hat{\beta}) = -log\mathcal{L}^{(2)}(\hat{\beta})$ (menos la segunda derivada de la log-verosimilitud parcial evaluada en $\hat{\beta}$), la cual es una medida de la curvatura de la función de verosimilitud en $\hat{\beta}$. A mayor curvatura, más información y por ende menos varianza. A menor curvatura, menos información y, por ende, más varianza. Como la relación es inversa, la varianza de $\hat{\beta}$ será $\widehat{Var}(\hat{\beta}) = I(\hat{\beta})^{-1}$y su error típico: $SE(\hat{\beta}) = \sqrt{I(\hat{\beta})^{-1}}$. Este último se usa para construir un test *z* como $Z = \hat{\beta}/SE(\hat{\beta})$ que rechazará $H_0: \beta = 0$ si $| Z | > z_{\alpha/2}$ o $Z^2 > \chi^2_{\alpha,df=1}$. Por último, cabe mencionar que el modelo puede estratificarse. En tal caso habrá un *hazard* basal (posiblemente de forma diferente) para *cada* estrato: $h_i(t \mid \mathbf{x}, V = j) = h_{0j}(t)exp(\mathbf{x_i}^T\boldsymbol{\beta}) = h_{0j}exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right)$.

Nótese que en la expresión anterior los coeficientes de una variable no dependen del estrato j. El *HR* de las variables es el mismo para todos los estratos. En cambio, se puede pensar que alguna variable (o todas) posee(n) *HR* diferente(s) según el estrato. Por ejemplo, en el modelo que sigue la variable X_2 interactúa con el nivel de estrato: $h_i(t \mid \mathbf{x}, V = j) = h_{0j}(t)exp(\mathbf{x_i}^T\boldsymbol{\beta}) = h_{0j}exp(\beta_1 x_1 + \beta_2 x_2 + \beta_{2:V} x_2 : V)$. Si el modelo tuviera todos los coeficientes diferentes para cada estrato, las estimaciones serían similares a tomar los individuos de cada estrato por separado, y ajustar un modelo por estrato.

2.2. Residuos y supuesto de proporcionalidad

Los residuos permiten evaluar: (i) el supuesto de proporcionalidad; (ii) posibles valores atípicos; (iii) posibles valores atípicos que son además influyentes. El primer tipo a considerar es el de los residuos *martingala*. Básicamente son, para

cada individuo, la diferencia entre el indicador de censura $\delta_i = I(\tau < \upsilon) = \{0: censura, 1: evento\}$ y el valor esperado de dicho indicador bajo el modelo de *hazards* proporcionales. Estos residuos suman 0 y tienen rango $[-\infty, 1]$ pero con distribución asimétrica. En cambio, los residuos de *devianza* poseen distribución simétrica con esperanza cero. La suma de estos residuos al cuadrado es el valor del test de cociente de verosimilitud entre el modelo con las covariables y el modelo reducido nulo. Residuales de devianza en valor absoluto > 2 indican valores atípicos, con una tolerancia del 5 %. Los residuos de martingala se usan para examinar la forma funcional de las variables continuas, es decir, el supuesto de linealidad. Los residuos "delta betas" (*dfbetas*) marcan cuáles individuos ejercen mayor influencia sobre la estimación de los coeficientes. Hay una serie de residuos para *cada* coeficiente estimado. Para el i-ésimo individuo se calcula la diferencia entre el valor estimado del coeficiente usando todas las observaciones y su valor estimado excluyendo al individuo en cuestión. Si la diferencia es mayor al 10 % se considera un individuo influyente. Los residuos se pueden estandarizar dividiéndolos por el error típico del coeficiente. En este caso, se consideran influyentes aquellos con desvíos superiores a ± 2. Por último es necesario mencionar a los residuos de *Schoenfeld. Cada* coeficiente tiene su serie de residuos. Los residuos de *Schoenfeld* son los términos individuales de la ecuación de score para cada coeficiente. Se definen como la diferencia entre el valor de la covariable x_i para el i-ésimo individuo y su valor esperado, $E[x_i]$. Los residuos escalados son: $r_j^* = r_j \times d \times var(\hat{\beta})$; $d = \sum(\delta = 1)$; los residuos multiplicados por el total de eventos, multiplicados por la varianza del coeficiente.

Si el supuesto de proporcionalidad se cumple, los residuos deberían ubicarse aleatoriamente en torno al cero. Ahora bien, decir que el $HR = exp(\hat{\beta})$ no es proporcional es lo mismo que decir que el coeficiente depende del tiempo: $HR = exp\left(\hat{\beta}(t)\right)$. Grambsch and Therneau (1994) consideraron definir el coeficiente dependiente del tiempo como: $\beta(t) = \beta + \theta g(t)$, donde $g(t)$ es una función del tiempo previamente definida para modelar la dependencia temporal. Si el HR resulta ser función del tiempo, entonces la esperanza de los residuos escalados es aproximadamente: $E[r_i^*] \approx \beta + \beta(t)$. Por lo tanto el coeficiente $\beta(t)$ se podría estimar como: $\hat{\beta}(t) = r_i^* + \beta$. Si se grafica $\hat{\beta}(t)$ versus los tiempos de falla ordenados, se revela la forma funcional de la dependencia temporal. Además es posible testear que $\theta = 0$, ajustando una recta al gráfico y evaluando la significatividad estadística de la pendiente. Si se rechaza, el coeficiente en cuestión depende del tiempo. Otro modo, más simple, de solucionar el problema de la no proporcionalidad consiste en estratificar según una partición de los datos en intervalos temporales disjuntos y estimar coeficientes específicos para cada estrato.

3. El modelo de eventos múltiples

Los modelos de esta sección tienen que ver con eventos que pueden repetirse. Las instancias se consideraron TOKENS de un TYPE, o sea que el [TYPE = "los profesores"] puede aparecer como [TOKEN = "lo profesores", "los profesor", ...]. Una concordancia / individuo (TYPE) puede registrar un error en el curso de

seguimiento del sujeto o bien varios errores. Para reflejar esto, los datos se dispusieron de manera de que cada fila representara un TYPE y la variable TIEMPO se reorganizó para expresar los intervalos entre los cuales se producían los errores. El cuadro *1* lo ejemplifica. El primer TYPE, "barrios amplios", tiene dos intervalos: *(0, 315]* y *(315, 312]*; el primero desde el cero hasta el tiempo *315*, donde se produce el error (STATUS = 1) y el segundo desde *316* a *312*, siendo este último el tiempo de censura (STATUS = 0). El TYPE "los profesores" registra dos TOKEN, el primero es un error en el tiempo *212* (intervalo *(0, 212]*) y el segundo con error en el tiempo *240* (intervalo *(240, 312]*); en el último periodo no se registra error, por tanto, va de *241* a *312*, el tiempo de censura. Por último, el TYPE "alemanes fieles" no registra error[3], por tanto, el intervalo va del cero hasta el tiempo de censura. Es necesario notar que cada TYPE tiene al menos un TOKEN. Por ejemplo, "los profesores" registra para SONIA ocho instancias, de las cuales solamente tres son errores, que es el tiempo que se registra. En cuanto a la relación entre TYPE y TOKEN, *247* TYPES de *1152* tienen entre *2* y *32* TOKENS, y 127 solamente *2*. Todas las covariables se asumen *fijas* en el tiempo. O sea que, por ejemplo, una covariable específica *repite* sus valores a lo largo de los eventos dentro de un mismo TYPE. En total hubo 1813 observaciones y 473 eventos de error.

Cuadro 1. Ejemplo de datos de sobrevivencia: modelo de eventos repetidos

ID	TYPE	TOKEN	STATUS	T_START	T_STOP	(otras variables)
1	barrios amplios	barrios amplio	1	0	115	...
1	barrios amplios	–	0	115	312	...
2	los profesores	los profesor	1	0	212	...
2	los profesores	los profesore	1	212	240	...
2	los profesores	–	0	240	312	...
3	alemanes fieles	alemanes fideles	0	0	312	...

Siguiendo a Hosmer, Lemeshow & May (2008, cap. 9.2), se definieron los siguientes modelos.

Andersen-Hill [AG]

El modelo estima una hazard basal común para todos los eventos y también coeficientes globales, que no dependen de los eventos. Se supone independencia entre los eventos dentro de cada sujeto. El modelo se escribe como sigue:

$$h_{ik}(t \mid \mathbf{x}) = Y_{ik}(t)h_0(t)exp\left(\mathbf{x_i^T}\boldsymbol{\beta}\right)$$

[3] "fideles" es un error por no conocer la base léxica, por eso no se consideró error de concordancia [*fedele* 'fiel'].

En donde $h_{ik}(t \mid \mathbf{x})$ representa la función de riesgo de error del k-ésimo evento dentro del i-ésimo individuo (TYPE) en el tiempo t; $h_0(t)$ es el *hazard* basal común para todos los eventos; $\mathbf{x_i}$ es el vector de covariables para el i-ésimo individuo; $\boldsymbol{\beta}$ es el vector de coeficientes y $Y_{ik} = \{0,1\}$ indica cuándo el i-ésimo individuo se halla bajo observación[4]. El modelo podría incluir estratos (que no sean eventos), y por ende, habría una *hazard* basal por estrato:

$$h_{ik}(t \mid \mathbf{x}, V = j) = Y_{ik}(t) h_{0j}(t) exp\left(\mathbf{x_i^T \boldsymbol{\beta}}\right)$$

Prentice, Williams and Peterson [PWP]

El modelo estratifica por evento; por lo tanto, hay una *hazard* basal diferente para cada estrato. Todos los individuos están en riesgo para el primer estrato pero solo aquellos con un evento en el estrato precedente permanecen en riesgo en el estrato siguiente. Por ejemplo, en el Cuadro *5*, el individuo *2* no estará en riesgo de sufrir el segundo evento (error) hasta que no haya sufrido el primer evento. Se pueden estimar coeficientes específicos para cada estrato / evento. En la práctica se deben limitar la cantidad de eventos para estratificar; ya que, de no hacerse, la cantidad de individuos en riesgo para los estratos posteriores se vuelve muy pequeña y las estimaciones no son confiables. El modelo se escribe como el anterior pero con un índice adicional en el *hazard* basal, indicando que es diferente para cada estrato k-ésimo.

$$h_{ik}(t \mid \mathbf{x}, V = k) = Y_{ik}(t) h_{0k}(t) exp\left(\mathbf{x_i^T \boldsymbol{\beta}}\right)$$

El supuesto de eventos (errores) independientes dentro de cada individuo / TYPE resulta irreal. Por lo tanto, para dar cuenta de la posible correlación dentro de cada TYPE, se utiliza un estimador "sándwich" para la varianza de los coeficientes, que ajusta por datos agrupados (varios eventos en individuos) [Lin & Wei, 1989][5].

Modelo de fragilidad compartida ("shared frailty") [Frailty]

El modelo asume que los sujetos (TOKEN) pueden estar expuestos a diferentes niveles de riesgo, ser unos más (menos) «frágiles» que otros, debido al efecto de covariables no observadas. Si los sujetos (concordancias TOKEN) que forman parte un grupo (concordancia TYPE) comparten el mismo nivel de fragilidad, el modelo se denomina de

[4] Bajo el modelo clásico de Cox $Y_{ik} = 0$ luego de que el evento se produce y el individuo deja el conjunto de riesgo. En cambio aquí $Y_{ik} = 1$, haciendo que los individuos permanezcan en el conjunto de riesgo.

[5] Dicho estimador es: $\hat{R}(\boldsymbol{\beta}) = \widehat{Var}\left(\hat{\boldsymbol{\beta}}\right)\left[\hat{\mathbf{L}}^T\hat{\mathbf{L}}\right]\widehat{Var}\left(\hat{\boldsymbol{\beta}}\right)$, donde $\widehat{Var}\left(\hat{\boldsymbol{\beta}}\right)$ es la varianza del coeficiente estimado basado en la inversa de la información observada y $\hat{\mathbf{L}}$ es una matriz $n \times p$ con n filas como individuos y p covariables, y cuyas filas contienen residuos de *score* (ver: Hosmer, Lemeshow & May, 2008; cap. 6).

«fragilidad compartida». Llámese u_i a la fragilidad del TYPE i-ésimo, que modela estos efectos de covariables no observadas. Dichos u_i son entonces similares a efectos aleatorios de un modelo mixto. El modelo incluye un coeficiente «fragilidad» que debería resultar significativo. La distribución de los efectos aleatorios se asume *Gamma* con esperanza $E[\mathbf{u}] = 1$ y varianza $Var(\mathbf{u}) = \theta$: $g(u_i, \theta) = \frac{u_i^{\frac{1}{\theta}-1} e^{-\frac{u_i}{\theta}}}{\Gamma\left(\frac{1}{\theta}\right)\theta^{\frac{1}{\theta}}}$.

Si un sujeto (TOKEN) de un grupo (TYPE) tiene $u_i > 1$, entonces es "frágil", es decir, con más riesgo de sufrir el evento "error". En cambio un sujeto con $u_i < 1$ será más "fuerte", con menos riesgo de sufrir dicho evento.

La correlación intra-clase se define como: $IC = \theta/(2 + \theta)$. El modelo se escribe como riesgo del k-ésimo TOKEN del grupo (TYPE) i-ésimo *condicional* a las covariables y al efecto aleatorio u_i del i-ésimo grupo (TYPE):

$$h_{ik}(t \mid \mathbf{x}, u_i) = h_0(t) exp\left(\mathbf{x_{ik}^T}\boldsymbol{\beta} + u_i\right); u_i \sim Gamma(\theta)$$

Además, podría incluir estratos, en cuyo caso se agregaría un índice a la *hazard* basal para el j-ésimo estrato:

$$h_{ik}(t \mid \mathbf{x}, u_i, V = j) = h_{0j}(t) exp\left(\mathbf{x_{ik}^T}\boldsymbol{\beta} + u_i\right); u_i \sim Gamma(\theta)$$

3.1. Selección de modelos

Se ajustaron los siguientes modelos: (1) *AG*: modelo AG con varianza "sándwich" por TYPE; (2) *AG-STRATA*: modelo AG con varianza "sándwich" por TYPE y estratos por alumno: $V = 1,2,3,4$; (3) *PWP*: modelo PWP con estratos por eventos, reducidos a 3 eventos como máximo para evitar problemas de estimación ($V = 1,2,3$); y varianza "sándwich" por TYPE; (4) *Frailty*: modelo de "shared frailty" con distribución gamma de efectos aleatorios por TYPE; (5) *Frailty-STRATA*: modelo de "shared frailty" con distribución gamma de efectos aleatorios por TYPE; y estratos por alumno: $V = 1,2,3,4$.

El Cuadro *2* muestra la comparación de los modelos según las medidas de información AIC, delta y pesos de Akaike. Los dos mejores modelos son los que incluyen estratificación por alumno. El de "fragilidad compartida" es ligeramente mejor que el de AG. Sin embargo, en el primero el coeficiente de fragilidad no fue significativo ($p = 0.37$) y la varianza de los u_i fue muy baja ($\theta = 0.021$). Por lo tanto, se eligió el modelo AG estratificado por alumno, con varianza "sándwich" por individuo (TYPE).

Cuadro 2. Selección de modelos

	logLik[a]	AIC[b]	Delta[c]	Weight[d]
model.frailty.strata	-2727.688	5516.660	0.000	0.667
model.AG.strata	-2738.023	5518.046	1.387	0.333
model.PWP	-2993.258	6028.516	511.856	0.000
model.frailty	-3245.910	6556.913	1040.253	0.000
model.AG	-3258.391	6558.782	1042.123	0.000

[a]verosimilitud del modelo; [b]Akaike,
[c]delta de Akaike, [d]pesos de Akaike.

Se eligió el modelo AG estratificado por alumno, con varianza "sándwich" por individuo (TYPE). Se ajustó el modelo elegido con todas las predictoras mencionadas. Fueron $2^{17} = 131072$ modelos, jerarquizados mediante la medida de información AIC (como: $\frac{n}{p} = \frac{1813}{17} \approx 106 > 40$, no se usó la versión AICc corregida por tamaño muestral). Luego se examinó la frecuencia de las predictoras en el conjunto completo de modelos, que da un panorama de la incerteza por la selección. A continuación se redujo la cantidad de modelos al subconjunto "de confianza" con la regla $\frac{W(i)}{W(1)} > \frac{1}{8}$. Sobre dicho subconjunto se llevó a cabo un promedio de coeficientes con la varianza calculada con "full average". El Cuadro *3* muestra que las variables con porcentaje de elección arriba del *80* % son: ANIM, EST1, EST5, MORF.f. Las mismas variables son las que resultan significativas en los coeficientes promediados (Cuadro *4*).

Cuadro 3. Importancia Relativa de las predictoras

	Names	X
1	ANIM	1.00
2	EST1	0.99
3	EST5	0.98
4	MORF.f	0.89
5	ES	0.74
6	IMA.CONC.f	0.59
7	FAM.LEX.f	0.57
8	EST4	0.50
9	MOD	0.45
10	LDA	0.44
11	STEM.f	0.39
12	EST3	0.39
13	Fabs.SC.f	0.37
14	EST7	0.34
15	EST2	0.32
16	EST6	0.30
17	GRAMS	0.29

Cuadro 4. Promedio de los coeficientes con FULL AVERAGE

	Estimate	Std. Error	z value	Pr(> |z|)
ANIM1	0.412	0.115	3.596	0.000
ES1	0.244	0.250	0.976	0.329
ES2	-0.604	0.568	1.063	0.288
EST11	-0.495	0.118	4.198	0.000
EST41	-0.332	0.355	0.936	0.349
EST51	-0.598	0.185	3.224	0.001
FAM.LEX.f1	-0.107	0.108	0.989	0.323
IMA.CONC.f1	-0.114	0.117	0.978	0.328
MORF.f1	-0.519	0.174	2.987	0.003
MORF.f2	-0.491	0.279	1.760	0.078
STEM.f1	0.037	0.086	0.426	0.670
LDA1	-0.087	0.174	0.501	0.617
EST71	0.033	0.121	0.275	0.783
MOD1	-0.194	0.372	0.521	0.602
MOD2	0.046	0.096	0.478	0.633
MOD3	0.000	0.074	0.004	0.997
EST31	0.109	0.302	0.360	0.719
EST21	0.009	0.052	0.177	0.859
Fabs.SC.f1	-0.015	0.058	0.264	0.792
GRAMS1	0.003	0.032	0.097	0.923
EST61	0.007	0.063	0.104	0.917

3.2. Modelo ajustado

El modelo AG estratificado por alumno, con varianza “sándwich” por individuo (TYPE) con las predictoras ANIM, EST1, EST5, MORF.f, se escribe como:

$$h_{k(i)}(t \mid \mathbf{x}, V = j) = Y_{k(i)}(t) h_{0j}(t) exp(\boldsymbol{\beta}_1 ANIM + \boldsymbol{\beta}_2 EST1 + \boldsymbol{\beta}_3 EST5 + \boldsymbol{\beta}_4 MORF.f1 + \boldsymbol{\beta}_5 MORF.f2)$$
$$j = 1, \dots, 4; i = 1, \dots, 473; k = 1, \dots, 15; k(i) = 1, \dots, 1813; t = 1, \dots, 691$$

En donde: $h_{ik}(t \mid \mathbf{x})$ representa la función de riesgo de error del k-ésimo evento ($k = 1, \dots, 15$) anidado en el i-ésimo individuo (TYPE) [$i = 1, \dots, 473$] en el tiempo t ($t = 1, \dots, 691$); $h_{0j}(t)$ es el *hazard* basal específico para cada alumno ($j = 1, \dots, 4$); e $Y_{ik} = \{0,1\}$ indica cuándo el i(k)-ésimo individuo se halla bajo observación. Los efectos fijos (marginales) son: (i) β_1 es el efecto de ANIM; (ii) β_2 es el efecto de EST1; (iii) β_3 es el efecto de EST5; (v) β_4 es el efecto de MORF.f, nivel 1; (vi) β_5 es el efecto de MORF.f, nivel 2 [referencia de MORF.f es “0”].

La Figura *1*, a la izquierda, muestra los residuos “dfbetas” escalados. Se observa que ningún gráfico supera dos desvíos en valor absoluto. Por lo tanto, se

concluye que no hay observaciones influyentes. Por otra parte, en la derecha se muestran los residuos de devianza. El $5{,}6$ % supera en valor absoluto los dos desvíos (103 observaciones). Está en el borde del cinco por ciento esperado de valores atípicos.

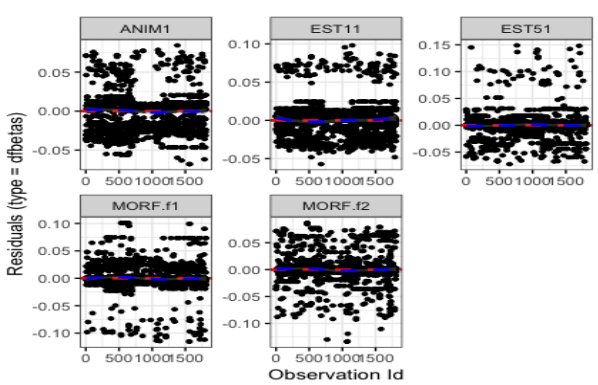
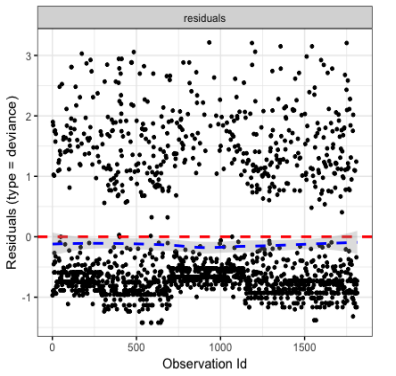


Figura 1. DFBETAS escalados



Resid. de devianza

La Figura 2 muestra los residuos de *Schoenfeld* versus el tiempo. La recta es $\beta(t)$. Se observa que el coeficiente ANIM tiene una leve pendiente negativa a medida que pasa el tiempo.

El Cuadro *5* muestra los p-valores del test de $\theta = 0$ para $\beta(t) = \beta + \theta g(t)$, donde $g(t)$ es una función del tiempo definida según: (i) $g(t) = rank(t)$; (ii) $g(t) = \hat{S}_{km}(t)$; (iii) $g(t) = t$; (iv) $g(t) = log(t)$. Según todas las funciones del tiempo, ANIM resulta significativo. Por lo tanto, no se cumple el supuesto de *hazards* proporcionales para dicha variable.

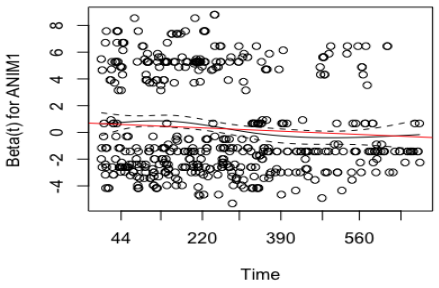


Figura 2. Residuos de Schoenfeld versus el tiempo. La recta roja es $\beta(t)$

Cuadro 5. Test para theta = 0, según diferentes funciones del tiempo: estadístico de chi cuadrado y p valor

	df	rank.chi2	rank.p	km.chi2	km.p	time.chi2	time.p	log.time.chi2	log.time.p
MORF.f1	1.0000	0.1430	0.7054	0.0087	0.9257	0.0002	0.9881	0.2011	0.6538
MORF.f2	1.0000	0.7136	0.3983	0.5088	0.4757	0.5284	0.4673	1.1555	0.2824
ANIM1	1.0000	11.4895	0.0007	11.0026	0.0009	11.0548	0.0009	7.2430	0.0071
EST11	1.0000	0.4291	0.5125	0.2645	0.6070	0.3078	0.5790	1.1808	0.2772
EST51	1.0000	1.5289	0.2163	1.3443	0.2463	1.3744	0.2411	1.4356	0.2309
GLOBAL	5.0000	11.8179	0.0374	11.2691	0.0463	11.2809	0.0461	7.7530	0.1704

Para acomodar esto, el modelo final incluyó el coeficiente de ANIM escrito como $\beta(t) = \beta + \theta g(t)$, donde $g(t) = t$, es decir dependiendo linealmente del tiempo. Se escribió como:

$$h_{k(i)}(t \mid \mathbf{x}, V = j) = Y_{k(i)}(t)h_{0j}(t)exp([\beta_0 + \beta_1(g(t) = t)]ANIM + \boldsymbol{\beta}_2 EST1$$
$$\boldsymbol{\beta}_3 EST5 + \boldsymbol{\beta}_4 MORF.f1 + \boldsymbol{\beta}_5 MORF.f2)$$
$$j = 1, \dots, 4; i = 1, \dots, 473; k = 1, \dots, 15; k(i) = 1, \dots, 1813; t = 1, \dots, 691$$

4. El modelo de riesgos competitivos

En este modelo se considera que hay k eventos (tipos de errores) posibles que compiten entre sí, de los cuales solo uno es observado y los demás están censurados. En este esquema se cuadruplicó *cada* individuo / instancia (sin considerar su TYPE) como se muestra en el Cuadro *6*:

Cuadro 6. Ejemplo de datos de sobrevivencia: modelo de riesgos competitivos

ID	INSTANCIA	ESPAÑOL	DE	A	TRANS	TIEMPO	STATUS	(otras variables)
1	mucho pensatores	muchos pensadores	1	2	1	3	0	...
1	mucho pensatores	muchos pensadores	1	3	2	3	0	...
1	mucho pensatores	muchos pensadores	1	4	3	3	1	...
1	mucho pensatores	muchos pensadores	1	5	4	3	0	...

El modelo de riesgos competitivos se entiende como si fuera una cadena de estados que comienza en el estado *1* y de allí se puede avanzar a cualquiera de los otros k estados, aquí los $k = 2,\ldots,5$ representan los estados de cada tipo de error. El cuadro *8* ilustra la organización de los datos para la instancia "mucho pensatores". Se establece que se parte del estado *1* (columna "DE") y se va hacia cualquiera de los cuatro estados posibles (columna "A"). Cada posible transición se numera en la columna "TRANS". La columna "TIEMPO" establece el tiempo del evento o de la censura, repetido para cada tipo de transición. Luego, la columna "STATUS" tiene un "1" en la transición donde se produce el evento y "0" en las demás (si no hubiera evento, hay "0" en todas las filas). O sea que, si hay evento, las otras causas posibles del error se hallan "censuradas", indicando que en ese instante de tiempo podrían ocurrir en lugar de la observada, aunque no se observen.

Para el caso de los riesgos competitivos, es preciso definir la función de *incidencia* acumulada para una causa específica (Putter et. al., 2007; Moore, 2016): $I_k(t) = P(T \leq t, C = k)$. Especifica la probabilidad acumulada para un individuo / concordancia de "morir" ("sufrir error") por la causa k-ésima ("tipo de error"). Es una función creciente pero, en el límite, alcanza la probabilidad de "muerte" por esa causa en particular; o sea que siempre está debajo de 1 ($I_k(\infty) = P(C = k)$). La función de riesgo instantánea (*hazard*) para una causa específica se define como la probabilidad de sobrevivir (no morir de la k-ésima causa) en un intervalo Δ corto de tiempo adicional sabiendo que el individuo sobrevivió hasta el tiempo t: $h_k(t) = \lim_{\Delta \to 0} \frac{P(t<T<t+\Delta, C=k|T>t)}{\Delta}$. Si se suman los *hazards* puntuales de todas las causas de "muerte" se obtiene la *hazard* total: $h(t) = \sum_{k=1}^{K} h_k(t)$. El modelo proporcional de Cox para riesgos competitivos se escribe como:

$$h_i(t \mid \mathbf{x}, C = k) = h_{0k}(t) exp(\mathbf{x_k}^T \boldsymbol{\beta_k}) = h_{0k} exp\left(\beta_{1k} x_{1k} + \beta_{2k} x_{2k} + \cdots + \beta_{pk} x_{pk}\right)$$

En donde: h_{0k} es la *hazard* basal para la causa k-ésima, $\mathbf{x_k}$ es el vector de covariables para la k-ésima causa y $\boldsymbol{\beta_k}$, su respectivo vector de coeficientes. O

sea que el modelo admite que HR de las covariables pueda ser específico para cada causa de “muerte” (error).

4.1. Selección de modelos

Ya que las concordancias están anidadas en sesiones (SESIÓN) y estas en los alumnos (ID), se consideraron datos agrupados en *52 clusters* dados por la cruza entre ID:SESIÓN. Se presume que las concordancias dentro de cada uno de dichos grupos se hallan más correlacionadas entre sí que con aquellas de otros grupos. Para dar cuenta de esta correlación se ajustó un modelo con varianza “sándwich” agrupando por ID:SESIÓN y otro de “fragilidad compartida” con efectos aleatorios de ID:SESIÓN. La medida de información de *Akaike* arrojó $AIC = 6628.656$ para el segundo modelo y $AIC = 6703.988$ para el primero; con lo cual se eligió el modelo de “fragilidad compartida”. Además, para el modelo elegido la varianza (de la distribución *Gamma*) de los efectos aleatorios fue $\theta = 0.29$ y la correlación intra grupo $IC = \theta/(2+\theta) = 0.29/(2+0.29) \approx 0.12$.

Se ajustó el modelo elegido con todas las predictoras. Fueron $2^{17} = 131072$ modelos, jerarquizados mediante la medida de información AIC (como: $\frac{n}{p} = \frac{7428}{17} \approx 436 > 40$, no se usó la versión AICc corregida por tamaño muestral). Luego se examinó la frecuencia de las predictoras en el conjunto completo de modelos, que da un panorama de la incerteza por la selección. A continuación se redujo la cantidad de modelos al subconjunto “de confianza” con la regla $\frac{W(i)}{W(1)} > \frac{1}{8}$. Sobre dicho subconjunto se llevó a cabo un promedio de coeficientes con la varianza calculada con “full average”. El Cuadro *7* muestra que las variables con porcentaje de elección arriba del 80 % son: ANIM, EST1, EST5, MORF.f, Fabs.SC.f, FAM.LEX.f, MOD, ES. A excepción de ES, las demás predictoras resultan significativas en los coeficientes promediados (Cuadro *8*).

Cuadro 7. Importancia Relativa de las predictoras

	Names	X
1	frailty(ID.SESION)	1.00
2	strata(trans)	1.00
3	MOD	1.00
4	EST5	1.00
5	MORF.f	0.99
6	FAM.LEX.f	0.99
7	EST1	0.97
8	Fabs.SC.f	0.93
9	ANIM	0.93
10	ES	0.83
11	EST4	0.70
12	STEM.f	0.56
13	EST3	0.52

14	EST7	0.44
15	EST6	0.43
16	IMA.CONC.f	0.36
17	EST2	0.36
18	GRAMS	0.35
19	LDA	0.28

Cuadro 8. Promedio de los coeficientes con FULL AVERAGE

	Estimate	Std. Error	z value	Pr(>|z|)
ANIM1	0.352	0.117	3.009	0.003
ES1	0.316	0.262	1.207	0.227
ES2	-0.694	0.571	1.216	0.224
EST11	-0.433	0.127	3.408	0.001
EST41	-0.534	0.369	1.446	0.148
EST51	-0.786	0.200	3.921	0.000
FAM.LEX.f1	-0.347	0.100	3.466	0.001
Fabs.SC.f1	-0.332	0.129	2.568	0.010
MOD1	-0.419	0.429	0.977	0.329
MOD2	0.518	0.126	4.128	0.000
MOD3	0.370	0.143	2.589	0.010
MORF.f1	-0.719	0.183	3.931	0.000
MORF.f2	-0.667	0.293	2.275	0.023
STEM.f1	0.108	0.129	0.837	0.402
EST31	0.312	0.475	0.655	0.512
EST71	0.118	0.214	0.550	0.582
EST61	0.071	0.159	0.445	0.656
GRAMS1	-0.027	0.076	0.351	0.726
IMA.CONC.f1	-0.022	0.069	0.312	0.755
EST21	0.017	0.074	0.230	0.818
LDA1	0.000	0.084	0.006	0.995

4.2. Modelo ajustado

Se compararon mediante AIC modelos de fragilidad compartida con y sin coeficientes específicos para cada causa de error; resultando en: $AIC(con) = 6595.025$; $AIC(sin) = 6626.958$. Conforme a que el AIC del modelo con coeficientes específicos fue menor, el modelo final incluyó los coeficientes específicos que resultaron significativos para cada tipo de error (datos completos en el material suplementario). El modelo de “fragilidad compartida” se escribió como:

$$h_{ik}(t \mid \mathbf{x}, u_i, C = j) = h_{0j}(t)exp(\boldsymbol{\beta}_1 Fabs.SC.f(1).1 + \boldsymbol{\beta}_2 Fabs.SC.f(1).2 + \boldsymbol{\beta}_3 Fabs.SC.f(1).3 + \boldsymbol{\beta}_4 MORF.f(1).2 + \boldsymbol{\beta}_5 MORF.f(1).3 + \boldsymbol{\beta}_6 MORF.f(2).3 + \boldsymbol{\beta}_7 MOD(2).3 + \boldsymbol{\beta}_8 MOD(2).4 + \boldsymbol{\beta}_9 MOD(3).3 + \boldsymbol{\beta}_{10} MOD(3).4 + \boldsymbol{\beta}_{11} ANIM(1).2 + \boldsymbol{\beta}_{12} ANIM(1).4 + \boldsymbol{\beta}_{13} FAM.LEX.f(1).2 + \boldsymbol{\beta}_{14} FAM.LEX.f(1).3 + \boldsymbol{\beta}_{15} FAM.LEX.f(1).4 + \boldsymbol{\beta}_{16} EST1(1).1 + \boldsymbol{\beta}_{17} EST1(1).4 + \boldsymbol{\beta}_{18} EST5(1).4 + u_i)$$
$$j = 1, \dots, 4; i = 1, \dots, 52; k = 1, \dots, 7428; t = 1, \dots, 691; u_i \sim Gamma(\theta)$$

En donde: $h_{ik}(t \mid \mathbf{x}, u_i, C = j)$ representa la función de riesgo de error de la k-ésima concordancia ($k = 1857 \times 4 = 7428$) dentro del i-ésimo grupo estratificado según el j-ésimo tipo de error ($j = 1, \dots, 4$) en el tiempo $t(t = 1, \dots, 691)$; $h_{0j}(t)$ es el *hazard* basal específico para cada tipo de error ($j = 1, \dots, 4$); y u_i, el i-ésimo factor aleatorio del grupo ID:SESIÓN ($i = 1, \dots, 52$). Los efectos fijos específicos para una determinada causa de error se detallan a continuación:

Cuadro 9. Efectos fijos del modelo de riesgos competitivos

Coeficiente	Descripción	Tipo de Error	Coeficiente	Descripción	Tipo de Error
$\boldsymbol{\beta}_1$	Efecto de Fabs.SC.f, nivel 1 (ref=0)	Género	$\boldsymbol{\beta}_{10}$	Efecto de MOD, nivel 3 (ref=0)	Mixto
$\boldsymbol{\beta}_2$	Efecto de Fabs.SC.f, nivel 1 (ref=0)	-e-epentética	$\boldsymbol{\beta}_{11}$	Efecto de ANIM, nivel 1 (ref=0)	-e-epentética
$\boldsymbol{\beta}_3$	Efecto de Fabs.SC.f, nivel 1 (ref=0)	Plural	$\boldsymbol{\beta}_{12}$	Efecto de ANIM, nivel 1 (ref=0)	Mixto
$\boldsymbol{\beta}_4$	Efecto de MORF.f, nivel 1	-e-epentética	$\boldsymbol{\beta}_{13}$	Efecto de FAM.LEX.f, nivel 1 (ref=0)	-e-epentética
$\boldsymbol{\beta}_5$	Efecto de MORF.f, nivel 1 (ref=0)	Plural	$\boldsymbol{\beta}_{14}$	Efecto de FAM.LEX.f, nivel 1 (ref=0)	Plural
$\boldsymbol{\beta}_6$	Efecto de MORF.f, nivel 2 (ref=0)	Plural	$\boldsymbol{\beta}_{15}$	Efecto de FAM.LEX.f, nivel 1 (ref=0)	Mixto
$\boldsymbol{\beta}_7$	Efecto de MOD, nivel 2 (ref=0)	Plural	$\boldsymbol{\beta}_{16}$	Efecto de EST1, nivel 1 (ref=0)	Género
$\boldsymbol{\beta}_8$	Efecto de MOD, nivel 2 (ref=0)	Mixto	$\boldsymbol{\beta}_{17}$	Efecto de EST1, nivel 1 (ref=0)	Mixto
$\boldsymbol{\beta}_9$	Efecto de MOD, nivel 3 (ref=0)	Plural	$\boldsymbol{\beta}_{17}$	Efecto de EST5, nivel 1 (ref=0)	Mixto

En la Figura *3* se muestran los residuos de devianza. El 3,7 % supera en valor absoluto los dos desvíos (513 observaciones de 13648). Está dentro del cinco por ciento esperado de valores atípicos. La Figura *4* muestra los residuos "dfbetas" escalados. Se observa que ningún gráfico supera dos desvíos en valor absoluto. Por lo tanto, se concluye que no hay observaciones influyentes.

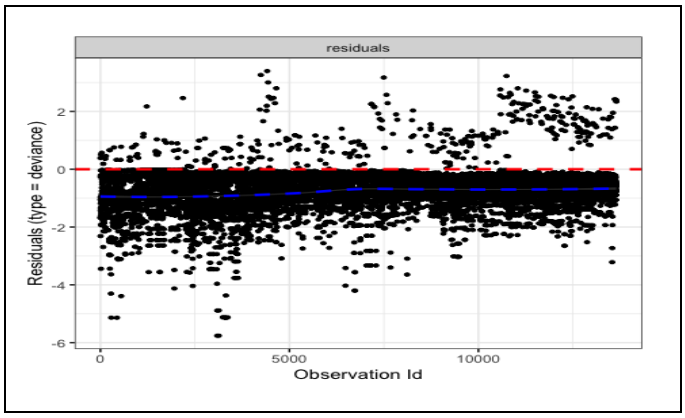


Figura 3. Residuos de devianza

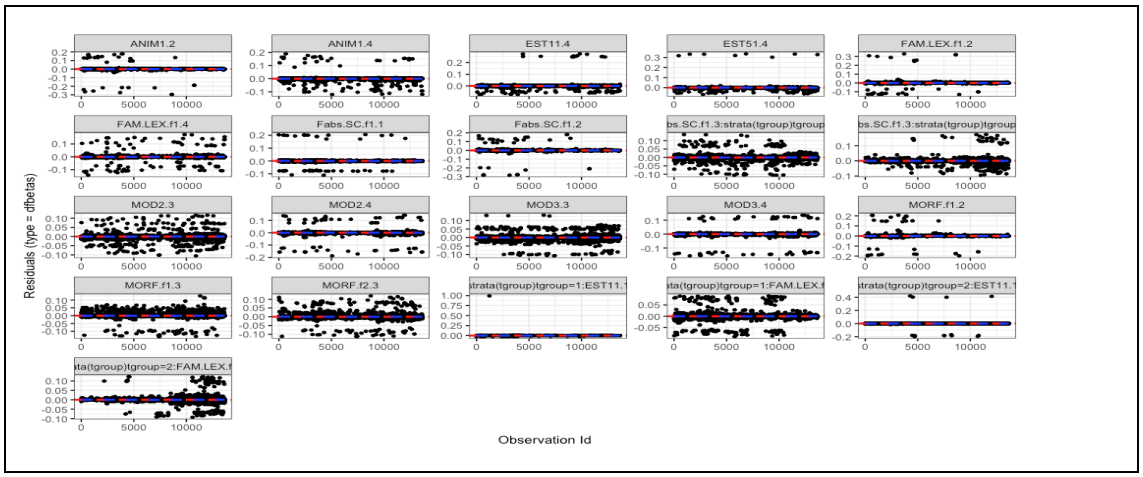


Figura 4. Residuos DFBETA escalados, para cada coeficiente

El Cuadro *10* muestra los p-valores del test de $\theta = 0$ para $\beta(t) = \beta + \theta g(t)$, donde $g(t)$ es una función del tiempo definida según: (i) $g(t) = rank(t)$; (ii) $g(t) = \hat{S}_{km}(t)$; (iii) $g(t) = t$; (iv) $g(t) = log(t)$. Se observa que Fabs.SC.f(1).3, FAM.LEX.f(1).3 y EST1(1).1 resultan significativas. Por lo tanto, no se cumple el supuesto de *hazards* proporcionales para dichas variables.

Cuadro 10. Test para theta = 0, según diferentes funciones del tiempo: estadístico de chi cuadrado y p valor

	Df	rank.chi2	rank.p	km.chi2	km.p	t.chi2	t.p	log.t.chi2	log.t.p
Fabs.SC.f1.1	1.0000	0.1424	0.7059	0.1719	0.6785	0.2095	0.6471	0.5122	0.4742
Fabs.SC.f1.2	1.0000	0.0874	0.7675	0.0487	0.8254	0.0590	0.8081	0.5978	0.4394
Fabs.SC.f1.3	1.0000	3.4814	0.0621	4.7143	0.0299	4.9946	0.0254	1.8575	0.1729
MORF.f1.2	1.0000	0.1431	0.7052	0.0605	0.8057	0.0558	0.8132	0.7727	0.3794
MORF.f1.3	1.0000	0.0212	0.8842	0.2167	0.6416	0.2568	0.6124	0.0219	0.8823
MORF.f2.3	1.0000	0.2631	0.6080	0.3644	0.5461	0.3385	0.5607	0.1632	0.6862
MOD2.3	1.0000	0.9175	0.3381	0.9863	0.3206	1.2790	0.2581	0.7758	0.3784
MOD2.4	1.0000	0.1414	0.7069	0.0188	0.8910	0.0436	0.8345	0.4572	0.4989
MOD3.3	1.0000	1.2978	0.2546	1.0419	0.3074	0.9624	0.3266	0.9200	0.3375
MOD3.4	1.0000	1.1326	0.2872	0.8331	0.3614	0.8269	0.3632	2.0319	0.1540
ANIM1.2	1.0000	0.0004	0.9840	0.0000	0.9960	0.0002	0.9888	0.4288	0.5126

ANIM1.4	1.0000	1.0977	0.2948	1.4705	0.2253	1.3641	0.2428	1.0714	0.3006
FAM.LEX.f1.2	1.0000	0.1446	0.7037	0.0515	0.8204	0.0587	0.8086	1.2557	0.2625
FAM.LEX.f1.3	1.0000	6.5752	0.0103	4.8466	0.0277	4.2661	0.0389	7.1557	0.0075
FAM.LEX.f1.4	1.0000	0.1442	0.7041	0.0051	0.9432	0.0000	0.9976	0.8679	0.3515
EST11.1	1.0000	5.4959	0.0191	4.2131	0.0401	4.0634	0.0438	4.4186	0.0355
EST11.4	1.0000	0.0042	0.9483	0.1776	0.6734	0.1269	0.7217	0.2670	0.6054
EST51.4	1.0000	0.6369	0.4248	0.5995	0.4388	0.5602	0.4542	0.0697	0.7917
GLOBAL	18.0000	21.5100	0.2545	19.5436	0.3591	19.3426	0.3710	22.9307	0.1933

Para acomodar esto, se ajustó el modelo estratificando para estas variables según $V = \{1: t \leq 300; 2: t > 300\}$. Es decir que para Fabs.SC.f(1).3, FAM.LEX.f(1).3 y EST1(1).1 hubo un coeficiente específico para cada intervalo de datos. El modelo final fue el siguiente:

$$h_{ik}(t \mid \mathbf{x}, u_i, C = j, V = w) = h_{0j}(t)exp(\boldsymbol{\beta}_1 Fabs.SC.f(1).1 + \boldsymbol{\beta}_2 Fabs.SC.f(1).2 + \boldsymbol{\beta}_{3,w=1} Fabs.SC.f(1).3 + \boldsymbol{\beta}_{4,w=2} Fabs.SC.f(1).3 + \boldsymbol{\beta}_5 MORF.f(1).2 + \boldsymbol{\beta}_6 MORF.f(1).3 + \boldsymbol{\beta}_7 MORF.f(2).3 + \boldsymbol{\beta}_8 MOD(2).3 + \boldsymbol{\beta}_9 MOD(2).4 + \boldsymbol{\beta}_{10} MOD(3).3 + \boldsymbol{\beta}_{11} MOD(3).4 + \beta_{12} ANIM(1).2 + \boldsymbol{\beta}_{13} ANIM(1).4 + \boldsymbol{\beta}_{14} FAM.LEX.f(1).2 + \boldsymbol{\beta}_{15,w=1} FAM.LEX.f(1).3 + \boldsymbol{\beta}_{16,w=2} FAM.LEX.f(1).3 + \boldsymbol{\beta}_{17} FAM.LEX.f(1).4 + \boldsymbol{\beta}_{18,w=1} EST1(1).1 + \boldsymbol{\beta}_{19,w=2} EST1(1).1 + \boldsymbol{\beta}_{20} EST1(1).4 + \boldsymbol{\beta}_{21} EST5(1).4 + u_i)$$
$$j = 1, \dots, 4; i = 1, \dots, 52; k = 1, \dots, 13648; t = 1, \dots, 691; w = 1,2; u_i \sim Gamma(\theta)$$

A continuación se presenta el ajuste completo del modelo de riesgos competitivos de "fragilidad compartida", con un coeficiente para cada tipo de error.

Cuadro 11. Modelo de riesgos competitivos

	Coef	se.coef.	Chisq	DF	p	exp.coef.	lower95	upper95
Fabs.SC.f1.1	-0.972	0.303	10.248	1.000	0.001	0.378	0.209	0.686
Fabs.SC.f1.2	1.056	0.524	4.066	1.000	0.044	2.876	1.030	8.031
Fabs.SC.f1.3	-0.315	0.152	4.325	1.000	0.038	0.729	0.542	0.982
Fabs.SC.f1.4	-0.258	0.269	0.924	1.000	0.336	0.772	0.456	1.308
MORF.f1.1	0.278	0.480	0.335	1.000	0.563	1.320	0.515	3.384
MORF.f1.2	-1.379	0.612	5.075	1.000	0.024	0.252	0.076	0.836
MORF.f1.3	-0.837	0.173	23.510	1.000	0.000	0.433	0.309	0.608
MORF.f1.4	-0.440	0.355	1.529	1.000	0.216	0.644	0.321	1.293
MORF.f2.1	0.245	0.566	0.187	1.000	0.665	1.278	0.421	3.873
MORF.f2.2	-0.502	0.634	0.626	1.000	0.429	0.605	0.175	2.098
MORF.f2.3	-0.907	0.218	17.247	1.000	0.000	0.404	0.263	0.619
MORF.f2.4	0.370	0.389	0.903	1.000	0.342	1.447	0.675	3.100
MOD1.1	-0.505	1.036	0.237	1.000	0.626	0.604	0.079	4.601
MOD1.3	-0.322	0.523	0.380	1.000	0.538	0.725	0.260	2.018

MOD1.4	-0.294	1.039	0.080	1.000	0.777	0.745	0.097	5.709
MOD2.1	0.551	0.294	3.514	1.000	0.061	1.734	0.975	3.085
MOD2.2	0.373	0.507	0.540	1.000	0.463	1.451	0.537	3.921
MOD2.3	0.409	0.160	6.568	1.000	0.010	1.505	1.101	2.058
MOD2.4	1.062	0.281	14.304	1.000	0.000	2.892	1.668	5.013
MOD3.1	-0.691	0.404	2.921	1.000	0.087	0.501	0.227	1.107
MOD3.2	0.481	0.647	0.553	1.000	0.457	1.618	0.455	5.751
MOD3.3	0.487	0.171	8.108	1.000	0.004	1.627	1.164	2.275
MOD3.4	0.730	0.324	5.064	1.000	0.024	2.076	1.099	3.921
ANIM1.1	0.470	0.277	2.889	1.000	0.089	1.600	0.931	2.753
ANIM1.2	1.654	0.392	17.844	1.000	0.000	5.229	2.427	11.266
ANIM1.3	0.089	0.140	0.403	1.000	0.526	1.093	0.830	1.439
ANIM1.4	0.491	0.242	4.100	1.000	0.043	1.633	1.016	2.626
FAM.LEX.f1.1	-0.110	0.263	0.175	1.000	0.676	0.896	0.535	1.501
FAM.LEX.f1.2	-1.295	0.425	9.278	1.000	0.002	0.274	0.119	0.630
FAM.LEX.f1.3	-0.271	0.125	4.705	1.000	0.030	0.762	0.597	0.974
FAM.LEX.f1.4	-0.510	0.222	5.274	1.000	0.022	0.601	0.389	0.928
EST11.1	-1.721	0.474	13.174	1.000	0.000	0.179	0.071	0.453
EST11.2	-0.118	0.454	0.068	1.000	0.795	0.888	0.365	2.165
EST11.3	-0.097	0.138	0.489	1.000	0.484	0.908	0.692	1.191
EST11.4	-0.996	0.310	10.329	1.000	0.001	0.369	0.201	0.678
EST51.1	-0.792	0.507	2.438	1.000	0.118	0.453	0.168	1.224
EST51.3	-0.155	0.217	0.516	1.000	0.473	0.856	0.560	1.309
EST51.4	-1.388	0.409	11.495	1.000	0.001	0.250	0.112	0.557

coef: betas estimados, exp(coef): hazard ratios, se(coef): error típico de betas estimados, robust se: error típico de beta con estimador sandwhich, z: coef / robust se, Pr(z): p-valor, lower 95: extremo izquierdo de intervalo de confianza de 95 por ciento para hazard ratio, upper 95: extremo derecho de intervalo de confianza de 95 por ciento para hazard ratio.

Ejemplos del corpus con MORF.f (bajo) para errores de plural y -e-epentética

Cuadro 12. Instancias de distancia baja de MORF.f.

	INSTANCIA	TIPO_ERROR	ID	SESION	LINEA
43	vacacione agreables	3	1	4	10
52	calles grandas (1)	2	1	4	60
322	tu pies	3	2	1	78
335	su padres	3	2	1	218
350	mi compañeros	3	2	2	142
356	mis amigo	3	2	2	225
442	su ministros	3	2	6	187
459	la curas (1)	3	2	7	16
526	su ojos	3	2	8	263
534	las investigacione	3	2	9	16
565	el desfiles	3	2	10	65
626	mi compañeros	3	2	12	6

670	las mujeras	2	2	13	129
672	[mujeras] famosas <de los políticos>	2	2	13	129
698	las mujeras	2	2	14	236
750	diferente personas	3	3	2	59
780	tu manos	3	3	2	172
817	sus acuerdo	3	3	4	26
827	su productos	3	3	4	81
914	su condiciones	3	3	7	40
925	su pasiones (1)	3	3	7	150
934	su pensamientos	3	3	7	189
957	diferente postaciones	3	3	8	48
986	tu parámetros (2)	3	3	9	30
1034	cuatrociento kilómetros	3	3	10	71
1039	la [motos] (1)	3	3	10	179
1040	[motos] más cómoda (2)	3	3	10	179
1046	las moto (1)	3	3	10	216
1047	únicas moto (2)	3	3	10	216
1097	cuatrociento habitants	3	3	12	15
1151	su aspectos (1)	3	3	13	125
1213	mi amigos (2)	3	4	1	250
1243	parte diferentes	3	4	1	425
1264	tu amigos	3	4	2	133
1267	tu amigas	3	4	2	138
1282	otras radio	3	4	2	249
1283	radio especializadas	3	4	2	255
1298	su tareas	3	4	3	50
1333	su libros	3	4	3	217
1358	costumbre diferentes	3	4	4	20
1383	[costumbres] diferente	3	4	4	183
1459	su instintos (1)	3	4	6	70
1475	su principios	3	4	6	197
1515	su razones	3	4	7	79
1516	las foto	3	4	7	95
1630	su juegos	3	4	9	109
1642	mis amigo (2)	3	4	9	168
1645	mucha motivaciones	3	4	9	185
1651	mi amigos	3	4	9	214
1660	su padres	3	4	9	248
1710	su textos	3	4	10	110
1734	su hijos	3	4	10	207
1764	los video	3	4	11	161

Código de R

Con el objetivo de fomentar la replicación de los resultados de este trabajo, se brinda en adjunto el script de R usado para el análisis, junto a las bases de datos.

Referencias

Burnham, K. P., & Anderson, D. R. (2010). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer.

Grambsch, P. M., & Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3), 515–526.

Hosmer, D. W., Lemeshow, S., & May, S. (2008). *Applied survival analysis: regression modeling of time to event data.* Wiley.

Lin, D. Y., & Wei, L. J. (1989). The robust inference for the cox proportional hazards model. *Journal of the American Statistical Association*, 84, 1074–1078.

Moore, D. F. (2016). *Applied Survival Analysis Using R.* Springer.

Putter, H.; Fiocco, M., & Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11), 2389–2430.

Stroup, W. W. (2013). *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. CRC Press, Chapman Hall.