

CAPÍTULO 2: ESTUDIO DE SIMULACIÓN

En la sección inmediatamente anterior se ha podido observar cómo los errores estándar asociados a las estimaciones de los parámetros mejoran cuando se introduce una matriz de correlación de trabajo para las observaciones de un mismo individuo (correlación intrasujeto). Sin embargo, nos parece obligatorio llevar a cabo un estudio de simulación para comprobar las mejoras que introducen las ecuaciones de estimación generalizadas.

Presentamos dos estudios de simulación, uno de variables respuesta binaria y otro de variables respuesta Poisson.

A continuación, se detallan los pasos seguidos en el estudio de simulación:

- 1- A partir del modelo marginal se calculan las probabilidades de las variables respuesta.
- 2- Generamos variables correlacionadas con respuesta binaria o Poisson.
- 3- Se calcula la estimación inicial del parámetro sin tener en cuenta la correlación entre las variables.
- 4- Se calcula la estimación de los parámetros teniendo en cuenta alguna correlación entre los parámetros.
- 5- Repetimos 2000 veces los Pasos 2 > 4 para muestras de 25, 50 y 100 individuos, con tres estructuras diferentes de correlación. Para comparar las estimaciones de los diferentes procedimientos se calcularán el valor medio de todas ellas, el error cuadrático medio y la probabilidad de cubrimiento al 95%.

2.1. MODELO PARA VARIABLES BINARIAS CORRELACIONADAS

Consideramos medidas repetidas tomadas en tres tiempos diferentes ($t = 3$) obtenidas de dos grupos donde las variables respuesta son variables binarias correlacionadas. Se supone que el modelo marginal viene dado por:

$$\text{logit}(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{1 - \mu_{ij}}\right) = \beta_0 + \beta_1 x_i + \beta_2 j + \beta_3 x_{ij}$$

donde

$$x_i = \begin{cases} 0 & \text{grupo 1} \\ 1 & \text{grupo 2} \end{cases} \quad \text{y } j = 1, 2, 3$$

quedando

$$\begin{cases} \text{logit}(\mu_{ij}) = \beta_0 + \beta_2 j & \text{grupo 1} \\ \text{logit}(\mu_{ij}) = \beta_0 + \beta_1 + (\beta_2 + \beta_3)j & \text{grupo 2} \end{cases}$$

siendo

$\beta_0 \rightarrow$ el intercepto del grupo 1

$\beta_1 \rightarrow$ el incremento del intercepto del grupo 2

$\beta_2 \rightarrow$ el efecto del tiempo en el grupo 1

$\beta_3 \rightarrow$ el incremento del efecto del tiempo en el grupo 2

Los verdaderos valores de los parámetros son elegidos como

$$\boldsymbol{\beta}_T = (\beta_0, \beta_1, \beta_2, \beta_3) = (0.1, 0.2, 0.2, 0.1)$$

Para simplificar asumimos que no hay sobredispersión, o lo que es lo mismo, que $\gamma = 1$. Las estructuras de \mathbf{R} que vamos a considerar para estudiar los efectos de las diferentes estructuras de correlación en los dos grupos son:

- Intercambiable con Alpha = 0.1
- Auto-regresiva(1) con Alpha = 0.1
- Independiente

A partir de aquí, una vez que hemos calculado la probabilidad asociada para cada grupo y fijada una estructura de correlación, vamos a utilizar el método propuesto por Park et al. [9] para generar variables correlacionadas con respuesta binaria que detallamos a continuación.

Se propone un algoritmo para generar un vector aleatorio $(Z_1, \dots, Z_k)^T$ de variables binarias tales que $E[Z_i] = p_i$, $q_i = 1 - p_i$ y $\text{corr}(Z_i, Z_j) = \rho_{ij} \geq 0, i \neq j$. Sea $\mathbf{X}(\alpha)$ una variable aleatorio Poisson

teniendo media $\alpha \geq 0$. Por convenio $X(0) = 0$. De aquí en adelante asumimos que las $X(\cdot)$ son independientes unas de otras si ellas aparecen con distintos subíndices.

Algoritmo:

Paso 0: Calculamos

$$\alpha_{ij} = \log \left[1 + \rho_{ij} \{q_i p_i^{-1} q_j p_j^{-1}\}^{\frac{1}{2}} \right]$$

para $1 \leq i, j \leq k$.

Sea $l = 0$ y $\alpha_{ij}^1 = \alpha_{ij}$, $1 \leq i, j \leq k$.

Paso 1: Sea $l = l + 1$.

Determinamos β_l y (r, s) .

Definimos

$$T_l = \{\alpha_{ij}^1 : \alpha_{ij}^1 > 0, 1 \leq i, j \leq k\}.$$

Sea $\beta_l = \alpha_{rs}^l$ el menos elemento del conjunto T_l . Si $\alpha_{rr} = 0$ ó $\alpha_{ss} = 0$, entonces parar.

Calculamos S_l :

Sea $S_l^0 = \{r, s\}$

Para $i = 1, \dots, k$; Sea

$$\begin{aligned} S_l^i &= S_l^{i-1} \cup \{i\} \text{ si } \alpha_{ij}^l > 0 \text{ para todo } j \in S_l^{i-1} \\ &= S_l^{i-1} \text{ en caso contrario} \end{aligned}$$

Sea $S_l = S_l^k$

Paso 2: Actualizamos los α_{ij} :

$$\begin{aligned} \alpha_{ij}^{l+1} &= \alpha_{ij}^l - \beta_l, \quad \forall i \in S_l \text{ y } j \in S_l \\ &= \alpha_{ij}^l \quad \forall i \notin S_l \text{ o } j \notin S_l. \end{aligned}$$

Si todos los $\alpha_{ij}^{l+1} = 0$ para $1 \leq i, j \leq k$, entonces vamos al Paso 3. En caso contrario vamos al Paso 1.

Paso 3: Sea $\tau = l$.

$$Y_i = \sum_{l=1}^{\tau} X_l(\beta_l) I_{S_l}(i); \text{ para } i = 1, \dots, k$$

y el conjunto

$$Z_i = I_{\{0\}}(Y_i)$$

donde I_A es la función indicatriz de un conjunto A tal que $I_A(y) = \begin{cases} 1 & y \in A \\ 0 & y \notin A \end{cases}$

Nótese que en el Paso 1 el mínimo β_l es elegido de tal manera que asegura que todas las variables aleatorias Poisson tienen media no negativa. Es posible que α_{rs} y S_l en el Paso 1 puedan no ser determinadas unívocamente. En este caso podemos elegir α_{rs} y S_l arbitrariamente.

Calculamos el vector aleatorio binario $\mathbf{Z} = (Z_1, Z_2, Z_3)$ a partir del algoritmo anterior generando tantas variables Poisson independientes de parámetro β_l como sean necesarias.

La comparación de los diferentes procesamientos se lleva a cabo utilizando las siguientes medidas:

El valor promedio de los parámetros estimados como

$$\sum_{k=1}^{2000} \frac{\hat{\beta}_{k,i}}{2000}$$

donde $\hat{\beta}_{k,i}$ es el estimador del parámetro β_i , $i = 0, \dots, 3$ de la réplica k-ésima.

El error cuadrático medio (ECM_i) de $\hat{\beta}_{k,i}$

$$ECM_i = \sqrt{\sum_{k=1}^{2000} \frac{(\hat{\beta}_{k,i} - \beta_T)^2}{2000}}$$

donde β_T el verdadero valor del parámetro β .

La probabilidad de cubrimiento al 95%, se calcula como la proporción de las réplicas en las que el Intervalo de Confianza al 95% (IC) contiene el verdadero valor del parámetro. Para cada réplica calculamos el IC de β_i de la forma

$$\hat{\beta}_{k,i} \pm 1,96s_i$$

Donde s_i el error estándar estimado de $\hat{\beta}_{k,i}$.

TABLA 9. RESULTADOS SIMULACIÓN PARA RESPUESTAS BINARIAS CON MATRIZ DE CORRELACIÓN DE TRABAJO INTERCAMBIABLE 0.1 EN AMBOS GRUPOS

		25	50	100
β_0	Media	0.0934	0.0840	0.1021
	ECM	0.6282	0.4439	0.3112
	Prob Cub	0.9660	0.9545	0.9565
β_1	Media	0.2021	0.2257	0.1862
	ECM	0.9436	0.6456	0.4490
	Prob Cub	0.9505	0.9505	0.9565
β_2	Media	0.2047	0.2105	0.2004
	ECM	0.2983	0.2123	0.1465
	Prob Cub	0.9635	0.9475	0.9520
β_3	Media	0.1119	0.0930	0.1080
	ECM	0.4544	0.3102	0.2145
	Prob Cub	0.9450	0.9305	0.9290

TABLA 10. RESULTADOS SIMULACIÓN PARA RESPUESTAS BINARIAS CON MATRIZ DE CORRELACIÓN DE TRABAJO Ar1 0.1 EN AMBOS GRUPOS

		25	50	100
β_0	Media	0.1146	0.1055	0.0993
	ECM	0.6190	0.4462	0.3082
	Prob Cub	0.9575	0.9545	0.9580
β_1	Media	0.1568	0.1958	0.2005
	ECM	0.9283	0.6619	0.4581
	Prob Cub	0.9515	0.9470	0.9460
β_2	Media	0.1988	0.2035	0.1997
	ECM	0.2930	0.2133	0.1447
	Prob Cub	0.9655	0.9460	0.9565
β_3	Media	0.1286	0.1031	0.1041
	ECM	0.4462	0.3128	0.2133
	Prob Cub	0.9500	0.9350	0.9255

TABLA 11. RESULTADOS SIMULACIÓN PARA RESPUESTAS BINARIAS CON MATRIZ DE CORRELACIÓN DE TRABAJO INDEPENDIENTE EN AMBOS GRUPOS

		25	50	100
β_0	Media	0.0943	0.1013	0.1076
	ECM	0.6469	0.4448	0.3142
	Prob Cub	0.9550	0.9425	0.9485
β_1	Media	0.2145	0.1771	0.2047
	ECM	0.9329	0.6386	0.4650
	Prob Cub	0.9485	0.9535	0.9425
β_2	Media	0.2095	0.1993	0.1982
	ECM	0.3060	0.2088	0.1471
	Prob Cub	0.9570	0.9480	0.9530
β_3	Media	0.1035	0.1178	0.0993
	ECM	0.4521	0.3114	0.2185
	Prob Cub	0.9350	0.9380	0.9255

En las tres tablas anteriores se han marcado en negrita los menores errores cuadrático medios y las mayores probabilidades de cubrimiento. Observamos que se obtienen mejores resultados cuando se introduce una matriz de correlación de trabajo. En este caso, la matriz de correlación de trabajo AR(1) es la que obtiene los mejores resultados. Además, el estudio de simulación también se probó para otros valores de alfa y para diferentes estructuras de la matriz de correlación de trabajo entre los grupos, pero los resultados eran muy parecidos y optó por poner estos a modo de ejemplo.

2.2. MODELO PARA VARIABLES POISSON CORRELACIONADAS

Consideramos medidas repetidas tomadas en tres tiempos diferentes ($t = 3$) obtenidas de dos grupos donde las variables respuesta son variables Poisson correlacionadas. Se supone que el modelo marginal viene dado por:

$$\log(\mu_{ij}) = \beta_0 + \beta_1 x_i + \beta_2 j + \beta_3 x_{ij}$$

donde

$$x_i = \begin{cases} 0 & \text{grupo 1} \\ 1 & \text{grupo 2} \end{cases} \quad \text{y } j = 1, 2, 3$$

quedando

$$\begin{cases} \log(\mu_{ij}) = \beta_0 + \beta_2 j & \text{grupo 1} \\ \log(\mu_{ij}) = \beta_0 + \beta_1 + (\beta_2 + \beta_3)j & \text{grupo 2} \end{cases}$$

siendo

$\beta_0 \rightarrow$ el intercepto del grupo 1

$\beta_1 \rightarrow$ el incremento del intercepto del grupo 2

$\beta_2 \rightarrow$ el efecto del tiempo en el grupo 1

$\beta_3 \rightarrow$ el incremento del efecto del tiempo en el grupo 2

Los verdaderos valores de los parámetros son elegidos como

$$\beta_T = (\beta_0, \beta_1, \beta_2, \beta_3) = (0.1, 0.2, 0.2, 0.1)$$

Para simplificar asumimos que no hay sobredispersión, o lo que es lo mismo, que $\gamma = 1$. Las estructuras de \mathbf{R} que vamos a considerar para estudiar los efectos de las diferentes estructuras de correlación en los dos grupos son:

- Intercambiable con Alpha = 0.1
- Auto-regresiva(1) con Alpha = 0.1
- Independiente

A partir de aquí, una vez que hemos calculado la probabilidad asociada para cada grupo y fijada una estructura de correlación, vamos a utilizar el método propuesto por Park y Shin [8] para generar variables correlacionadas con respuesta Poisson que detallamos a continuación.

Generamos variables Poisson correlacionadas $(\mathbf{Z}_1, \dots, \mathbf{Z}_k)'$ a partir de variables respuesta Poisson independientes.

Paso 1: Sea $l = 1$

Calculamos $\alpha_{ij}^l = \rho_{ij}(\mu_i \mu_j)^{\frac{1}{2}}$ para $1 \leq i, j \leq k$.

Construimos un grafo conectando los vértices i, j con $\alpha_{ij} > 0$ donde cada vértice representa una de las k variables aleatorias. El valor de α_{ij} se asigna a la arista (i, j) que une los vértices i, j .

Paso 2: A la arista con menor valor se denota por (r, s) .

Sea $\beta_l = \alpha_{rs}^l$.

Paso 3: Encontramos un subconjunto S_l de vértices unidos que contengan los vértices r, s .

Paso 4: Actualizamos los valores de las aristas de la siguiente forma

$$\begin{aligned}\alpha_{ij}^{l+1} &= \alpha_{ij}^l - \beta_l, \quad \forall i \in S_l \text{ y } j \in S_l \\ &= \alpha_{ij}^l \quad \forall i \notin S_l \text{ y } j \notin S_l.\end{aligned}$$

Se quitan las aristas (i, j) con $\alpha_{ij}^{l+1} = 0$. Si todas las nuevas aristas tienen valor ≤ 0 entonces vamos al Paso 5.

Si no, $l = l + 1$ y volvemos al Paso 2.

Paso 5: Si algún vértice tiene valor $\alpha_{ij}^{l+1} < 0$, entonces el algoritmo falla y se para.

En caso contrario, se construye una matriz $k \times \tau$, $\mathbf{T} = [T_{il}]$ con

$$T_{il} = \begin{cases} 1, & \text{si } i \in S_l, \\ 0, & \text{si } i \notin S_l \end{cases} \quad \text{para } i = 1, \dots, k, l = 1, \dots, \tau$$

Entonces, $\mathbf{Z} = \mathbf{T} \times \mathbf{Y}$, donde \mathbf{Y} está formado por variables independientes Poisson de parámetros β_l , son un subconjunto de variables Poisson correlacionadas.

TABLA 12. RESULTADOS SIMULACIÓN PARA RESPUESTAS POISSON CON MATRIZ DE CORRELACIÓN DE TRABAJO INTERCAMBIABLE 0.1 EN AMBOS GRUPOS

		25	50	100
β_0	Media	0.0839	0.0969	0.0945
	ECM	0.2539	0.1761	0.1234
	Prob Cub	0.9415	0.9405	0.9545
β_1	Media	0.1971	0.2006	0.2058
	ECM	0.3264	0.2256	0.1630
	Prob Cub	0.9490	0.9575	0.9535
β_2	Media	0.2041	0.2003	0.2017
	ECM	0.1074	0.0742	0.0519
	Prob Cub	0.9440	0.9560	0.9525
β_3	Media	0.1025	0.0994	0.0975
	ECM	0.1363	0.0952	0.0677
	Prob Cub	0.9000	0.8440	0.7065

TABLA 13. RESULTADOS SIMULACIÓN PARA RESPUESTAS POISSON CON MATRIZ DE CORRELACIÓN DE TRABAJO AR1 0.1 EN AMBOS GRUPOS

		25	50	100
β_0	Media	0.0846	0.0943	0.0974
	ECM	0.2478	0.1752	0.1224
	Prob Cub	0.9490	0.9485	0.9555
β_1	Media	0.2066	0.2029	0.1990
	ECM	0.3288	0.2305	0.1613
	Prob Cub	0.9480	0.9465	0.9560
β_2	Media	0.2038	0.2011	0.2001
	ECM	0.1047	0.0752	0.0519
	Prob Cub	0.9590	0.9530	0.9605
β_3	Media	0.0969	0.0989	0.1007
	ECM	0.1366	0.0971	0.0670
	Prob Cub	0.8940	0.8350	0.7240

TABLA 14. RESULTADOS SIMULACIÓN PARA RESPUESTAS POISSON CON MATRIZ DE CORRELACIÓN DE TRABAJO INDEPENDIENTE EN AMBOS GRUPOS

		25	50	100
β_0	Media	0.0946	0.0949	0.0958
	ECM	0.2485	0.1755	0.1225
	Prob Cub	0.9330	0.9365	0.9605
β_1	Media	0.2055	0.2058	0.2000
	ECM	0.3220	0.2335	0.1642
	Prob Cub	0.9565	0.9520	0.9495
β_2	Media	0.2002	0.1999	0.2009
	ECM	0.1049	0.0762	0.0530
	Prob Cub	0.9540	0.9420	0.9715
β_3	Media	0.0971	0.0985	0.0996
	ECM	0.1378	0.0971	0.0674
	Prob Cub	0.9035	0.8435	0.7290

En las tres tablas anteriores se han marcado en negrita los menores errores cuadrático medios y las mayores probabilidades de cubrimiento. Observamos que se obtienen mejores resultados cuando se introduce una matriz de correlación de trabajo. En este caso, la matriz de correlación de trabajo AR(1) es la que obtiene los mejores resultados. Además, el estudio de simulación también se probó para otros valores de alfa y para diferentes estructuras de la matriz de correlación de trabajo entre los grupos, pero los resultados eran muy parecidos y optó por poner estos a modo de ejemplo.

BIBLIOGRAFÍA

- [9-6] Park, C.G., Park, T. y Shin, D.W. (1996). A simple method for generating correlated binary variates. *The American Statistician*, 50, (4), 306-310.
- [8-5] Park, C.G. y Shin, D.W. (1998). An algorithm for generating correlated random variables in a class of infinitely divisible distributions. *Journal Statist. Comput. Simul.*, 61, 127-139.