Hachuel, Leticia
Boggio, Gabriela
Wojdyla, Daniel
Borra, Virginia
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística

ESTUDIO DEL COMPORTAMIENTO DEL ESTIMADOR DEL EFECTO TRATAMIENTO EN ESTUDIOS CON ALEATORIZACIÓN A NIVEL INDIVIDUAL*

1. INTRODUCCIÓN

En muchas aplicaciones los individuos bajo estudio presentan algún tipo de agrupamiento que provoca que las observaciones provenientes de individuos de un mismo grupo tiendan a estar correlacionadas. En el análisis estadístico de este tipo de datos frecuentemente se busca modelar respuestas como función de covariables, ajustando los resultados por la correlación potencial de las respuestas de individuos dentro de un mismo grupo. Un enfoque posible es el que incluye en el modelo efectos no observados a nivel grupo que varían aleatoriamente y reciben el nombre de efectos aleatorios (Goldstein, 2003; Fitzmaurice et al., 2004).

Estos modelos constituyen la clase de los denominados modelos lineales generalizados mixtos (MLGM), los cuales admiten variables respuestas no normales y permiten modelar una función de la media a través de efectos fijos, asociados a variables medidas tanto a nivel individual como grupal, y de efectos aleatorios en el predictor lineal.

Este enfoque de modelización se caracteriza por especificar una función de probabilidad paramétrica completa, por lo que es posible obtener estimaciones máximo-verosímiles en las que basar la inferencia. Para ello se debe integrar la función de distribución conjunta de las respuestas con respecto a las distribuciones de los efectos aleatorios a fin de obtener la función de verosimilitud, que resulta ser una función de los parámetros de regresión y de otros parámetros de esas distribuciones. Resolver esa integral puede ser una tarea ardua, razón por la cual se han presentado distintas alternativas. Uno de los métodos más usados consiste en aproximar la función de verosimilitud usando técnicas de integración numérica. Otro enfoque, el de cuasiverosimilitud penalizada, propone aproximar la función de verosimilitud utilizando métodos de linearización (Agresti, 2002; Song, 2007). El paquete estadístico SAS implementa ambos enfoques mediante los procedimientos NLMIXED y GLIMMIX respectivamente.

El primer procedimiento de estimación provee aproximaciones a los estimadores máximo-verosímiles utilizando el método de cuadratura de Gauss-Hermite, el cual necesita mayor número de cuadraturas en la medida que el número de efectos aleatorios sea mayor. Por otro lado, varios autores han notado que las estimaciones cuasiverosímiles pueden ser asintóticamente sesgadas en situaciones donde se estudian un gran número de grupos con pocos individuos en cada uno de ellos, principalmente en el caso de res-

* En este trabajo participaron en calidad de auxiliares de investigación las alumnas de la carrera Licenciatura en Estadística Ivana Barbona y Virginia Pezza.

puesta binaria. Estos resultados parecen atenuarse cuando se observa un gran número de individuos reunidos en pocos grupos (Moerbeek et al, 2003; Bellamy et al, 2005).

Este trabajo tiene como objetivo estudiar el comportamiento de los estimadores del efecto de una covariable a nivel individual obtenidos por ambos métodos de estimación mediante un estudio por simulación. Para ello, se generan observaciones binarias agrupadas específicamente para un modelo logístico mixto simple bajo diferentes escenarios.

En la sección siguiente se presentan los MLGM y en particular el utilizado en este estudio. Luego se describen los métodos de estimación a comparar y seguidamente se presentan el algoritmo de generación de datos y los escenarios elegidos para realizar las simulaciones. Finalmente se muestran los resultados alcanzados y la discusión de los mismos.

2. METODOLOGÍA

2.1. Modelo lineal generalizado mixto

La premisa básica de los modelos lineales generalizados mixtos es que la correlación entre las unidades de un mismo grupo puede pensarse que surge por el hecho de compartir un conjunto de efectos aleatorios.

Condicional sobre los efectos aleatorios, las observaciones de diferentes grupos se suponen independientes y con una distribución de probabilidad perteneciente a la familia exponencial.

Sea Y_{ij} la respuesta para el j-ésimo individuo del i-ésimo grupo, pudiendo ser continua, binaria o de conteo. Asociado con cada Y_{ij} hay un vector (fila) X_{ij} de covariables de dimensión 1xp, las cuales pueden variar de grupo a grupo o bien de individuo a individuo dentro de cada grupo.

Para el caso particular en que Y_{ij} es una respuesta binaria correspondiente al j-ésimo individuo del i-ésimo grupo, un modelo logístico para Y_{ij} con interceptos aleatorios se especifica de la siguiente forma (Fitzmaurice et al., 2004):

1.- Condicional sobre un único efecto aleatorio, b_i, las Y_{ij} son independientes y tienen una distribución de probabilidad Bernoulli, con

$$Var(Y_{ij}/b_i)= E(Y_{ij}/b_i)\ \{1- E(Y_{ij}/b_i)\}\ \ (\phi=1). \qquad (1)$$

2.- La media condicional de Y_{ij} depende de efectos fijos y aleatorios a través de la siguiente expresón:

$$\ln\left\{\frac{Pr(Y_{ij}=1/b_i)}{1-Pr(Y_{ij}=1/b_i)}\right\}=\eta_{ij}=\mathbf{X}_{ij}\boldsymbol{\beta}+b_i\,. \qquad (2)$$

Es decir la media condicional de Y_{ij} se relaciona con el predictor lineal a través del enlace logit.

3.- El único efecto aleatorio b_i se supone que tiene una distribución Normal univariada con media cero y variancia σ_b^2.

2.2. Métodos de estimación

Una vez postulado un MLGM es de interés estimar los parámetros del modelo. Resulta natural elegir el método de estimación máximo-verosímil ya que para estos modelos es posible escribir la verosimilitud completa.

La función de verosimilitud se considera una función de los parámetros dadas las observaciones, esto es:

$$L(\beta,\sigma / y) = \int f(y / b;\beta).f(b;\sigma)db\,. \qquad (3)$$

Esta expresión generalmente es difícil de resolver y se complica aún más cuando el número de efectos aleatorios aumenta.

Cuando dicho número es pequeño se pueden utilizar métodos de integración numérica para aproximar la verosimilitud. El error inducido por reemplazar la integral por una suma finita, como lo hacen los métodos de cuadratura de Gauss-Hermite, se hace cada vez más difícil de controlar a medida que la dimensión de la integral aumenta. Las aproximaciones convergen a las estimaciones máximo-verosímiles (MV) a medida que el número de puntos de cuadratura se incrementa de una manera apropiada para la integración numérica (Song, 2007).

Si se está dispuesto a sacrificar exactitud por facilidad de implementación hay métodos que maximizan una aproximación analítica de la función de verosimilitud en lugar de maximizar a dicha función propiamente dicha. Este enfoque implica integrar una expansión por serie de Taylor de la función de verosimilitud. En particular, el algoritmo de Breslow y Clayton (1993) utiliza un argumento de cuasiverosimilitud penalizada ajustando iterativamente un modelo lineal mixto.

La principal ventaja de este enfoque es su relativa simplicidad evitando la integración numérica y siendo factible computacionalmente su uso para grandes conjuntos de datos. Sin embargo, este esquema iterativo no conduce a una estimación máximo-verosímil e incluso McCulloch (1997) usa argumentos analíticos y estudios por simulación para mostrar que el comportamiento del método puede ser bastante pobre en relación con el presentado anteriormente. Generalmente este enfoque se deteriora a medida que los datos se apartan de la distribución normal, como es el caso de datos binarios, y a medida que la magnitud de las componentes de varianza aumenta. En efecto, cuando la verdaderas componentes de varianza son grandes, el método de cuasiverosimilitud penalizada (CVP) tiende a producir estimaciones de ellas con sesgo negativo (Breslow y Lin, 1995). En forma general, entonces, los estimadores CVP son una buena aproximación a los máximo-verosímiles siempre que las varianzas de los efectos aleatorios sean relativamente pequeñas, es decir cuando los efectos fijos dominan el modelo. Agresti et al. (2000) recomiendan el uso de integración numérica en lugar de la aproximación CVP.

3. ESTUDIO DE SIMULACIÓN

Se diseña un estudio por simulación a fin de comparar el comportamiento de estimadores de un único coeficiente de regresión asociado a una covariable binaria a nivel individual en un modelo logístico con intercepto aleatorio. Este modelo se formaliza de la siguiente manera:

$$\ln\left\{\frac{Pr(Y_{ij}=1/b_i)}{1-Pr(Y_{ij}=1/b_i)}\right\} = \eta_{ij} = \beta X_{ij} + b_i\,. \qquad (4)$$

Los estimadores a evaluar son los obtenidos por el método de cuasiverosimilitud penalizada (CVP) implementado en el procedimiento GLIMMIX de SAS y el método de máxima verosimilitud (MV) proporcionado por el procedimiento NLMIXED del mismo programa computacional.

Para llevar adelante el estudio por simulación se generan datos a partir del algoritmo empleado por Evans y Hosmer (2004), que consideran el siguiente modelo:

$$\text{logit}\,\pi_{ij} = ca_i + \beta x_{ij}\,, \qquad (5)$$

siendo $\pi_{ij} = P(Y_{ij} = 1)$ la probabilidad de una respuesta positiva para el j-ésimo individuo del i-ésimo grupo.

El coeficiente aleatorio del modelo está conformado por a_i, el valor de una variable aleatoria normal estandarizada asumido por el grupo i-ésimo, multiplicado por una constante c, cuya magnitud produce diferentes niveles de correlación intra-grupo. X_{ij} es el valor asumido por una variable tipo Bernoulli con probabilidad igual a 0.5, y β es el coeficiente asociado.

Para valores predeterminados de β y c y valores aleatorios para a_i y 0 y 1 para X_{ij}, se determinan π_{ij}, a partir de los cuales se obtiene valores binarios 0 ó 1 comparando dicha probabilidad con un valor elegido al azar de una distribución Uniforme definida en el intervalo [0;1].

Cabe aclarar que se ha comprobado empíricamente que hay una relación directa entre la magnitud del coeficiente c y el grado de correlación entre las respuestas de los individuos de un mismo grupo.

3.1. Estudio de Montecarlo

A los valores binarios generados por el algoritmo recién descrito se les ajusta el modelo logístico mixto (4), con un coeficiente aleatorio, mediante los procedimientos GLIMMIX y NLMIXED de SAS, obteniendo las respectivas estimaciones del coeficiente de regresión, $\hat{\beta}$, y su desvío estándar.

Se repite este proceso de generación de datos y ajuste de modelos 1000 veces. Se calcula el promedio de las estimaciones del coeficiente de regresión, el promedio de los desvíos estándares de dichas estimaciones y el desvío estándar empírico del coeficiente estimado a través de las 1000 muestras de acuerdo a los dos métodos de estimación presentados. Los escenarios de análisis se definieron teniendo en cuenta diferentes valores y combinaciones de:

n: número de grupos en la muestra.

k: número de individuos dentro de cada grupo.

c: constante que interviene en el coeficiente aleatorio.

En este trabajo se presentan los resultados considerando en el modelo $\beta = 0.8$, distinto grado de correlación a través de los valores de c iguales 1, 2 y 4 y combinaciones entre valores de k iguales a 5, 10, 20 y 50 y valores de n iguales a 10, 30, 50 y 100.

4. RESULTADOS

Las Tablas 1, 2, 3 y 4 muestran los resultados hallados a través de la generación de 1000 muestras, del promedio de las estimaciones del coeficiente de regresión, el promedio de los desvíos estándares de dichas estimaciones y el desvío estándar empírico para los escenarios elegidos obtenidos mediante los dos métodos de estimación, CVP y MV.

Tabla 1: Promedio del coeficiente de regresión estimado por ambos métodos, su desvío estándar estimado y desvío estándar empírico según número de individuos por grupo (k) y valores de c para n=10 grupos.

Método	k	Baja correlación intragrupo c=1			Mediana correlación intragrupo c=2			Alta correlación intragrupo c=4		
		Promedio $\hat{\beta}$	Promedio DS($\hat{\beta}$)	DS empírico	Promedio $\hat{\beta}$	Promedio DS($\hat{\beta}$)	DS empírico	Promedio $\hat{\beta}$	Promedio DS($\hat{\beta}$)	DS empírico
CVP	5	0.801	0.677	0.718	0.722	0.779	0.769	0.595	0.959	0.818
	10	0.784	0.464	0.459	0.763	0.544	0.522	0.746	0.711	0.683
	20	0.780	0.325	0.335	0.772	0.382	0.379	0.785	0.506	0.489
	50	0.795	0.205	0.207	0.785	0.240	0.232	0.768	0.318	0.303
MV	5	0.905	7.795	1.056	0.960	7.493	1.378	1.256	1.496	4.470
	10	0.826	0.483	0.488	0.840	0.581	0.582	0.891	0.835	1.100
	20	0.806	0.333	0.345	0.808	0.394	0.398	0.853	0.536	0.555
	50	0.807	0.207	0.210	0.799	0.243	0.236	0.789	0.324	0.312

Tabla 2: Promedio del coeficiente de regresión estimado por ambos métodos, su desvío estándar estimado y desvío estándar empírico según número de individuos por grupo (k) y valores de c para n=30 grupos

Método	k	Baja correlación intragrupo c=1			Mediana correlación intragrupo c=2			Alta correlación intragrupo c=4		
		Promedio $\hat{\beta}$	Promedio DS($\hat{\beta}$)	DS empírico	Promedio $\hat{\beta}$	Promedio DS($\hat{\beta}$)	DS empírico	Promedio $\hat{\beta}$	Promedio DS($\hat{\beta}$)	DS empírico
CVP	5	0.721	0.371	0.369	0.648	0.422	0.406	0.547	0.510	0.431
	10	0.755	0.262	0.261	0.730	0.303	0.295	0.689	0.382	0.361
	20	0.764	0.186	0.186	0.758	0.216	0.215	0.748	0.277	0.269
	50	0.785	0.117	0.117	0.780	0.137	0.133	0.769	0.177	0.173
MV	5	0.805	0.403	0.418	0.811	0.489	0.520	0.831	0.664	0.712
	10	0.811	0.275	0.283	0.816	0.325	0.333	0.831	0.428	0.447
	20	0.795	0.191	0.196	0.799	0.223	0.228	0.813	0.291	0.294
	50	0.800	0.119	0.119	0.796	0.139	0.137	0.791	0.180	0.178

Tabla 3: Promedio del coeficiente de regresión estimado por ambos métodos, su desvío estándar estimado y desvío estándar empírico según número de individuos por grupo (k) y valores de c para n=50 grupos

Método	k	Baja correlación intragrupo c=1			Mediana correlación intragrupo c=2			Alta correlación intragrupo c=4		
		Promedio $\hat{\beta}$	Promedio DS($\hat{\beta}$)	DS empírico	Promedio $\hat{\beta}$	Promedio DS($\hat{\beta}$)	DS empírico	Promedio $\hat{\beta}$	Promedio DS($\hat{\beta}$)	DS empírico
CVP	5	0.714	0.285	0.278	0.644	0.323	0.309	0.543	0.389	0.326
	10	0.743	0.202	0.200	0.719	0.233	0.229	0.676	0.293	0.273
	20	0.758	0.143	0.137	0.752	0.167	0.164	0.738	0.213	0.207
	50	0.786	0.091	0.092	0.784	0.106	0.103	0.775	0.136	0.132

MV	5	0.803	0.309	0.315	0.806	0.372	0.394	0.809	0.494	0.498
	10	0.804	0.213	0.220	0.807	0.250	0.258	0.814	0.327	0.334
	20	0.792	0.147	0.145	0.794	0.172	0.174	0.802	0.223	0.226
	50	0.801	0.092	0.094	0.801	0.107	0.105	0.799	0.138	0.136

Tabla 4: Promedio del coeficiente de regresión estimado por ambos métodos, su desvío estándar estimado y desvío estándar empírico según número de individuos por grupo (k) y valores de c para n=100 grupos

Método	k	Baja correlación intragrupo c=1			Mediana correlación intragrupo c=2			Alta correlación intragrupo c=4		
		Promedio $\hat{\beta}$	Promedio DS($\hat{\beta}$)	DS empírico	Promedio $\hat{\beta}$	Promedio DS($\hat{\beta}$)	DS empírico	Promedio $\hat{\beta}$	Promedio DS($\hat{\beta}$)	DS empírico
CVP	5	0.714	0.200	0.195	0.643	0.226	0.211	0.540	0.272	0.232
	10	0.740	0.143	0.139	0.711	0.164	0.158	0.665	0.205	0.194
	20	0.765	0.101	0.099	0.756	0.118	0.118	0.741	0.149	0.146
	50	0.788	0.064	0.065	0.786	0.075	0.075	0.781	0.098	0.093
MV	5	0.803	0.217	0.223	0.805	0.261	0.269	0.799	0.343	0.349
	10	0.799	0.150	0.152	0.798	0.176	0.179	0.800	0.229	0.236
	20	0.799	0.104	0.104	0.800	0.122	0.125	0.806	0.157	0.160
	50	0.802	0.065	0.067	0.804	0.076	0.076	0.804	0.096	0.096

Los resultados hallados se analizan teniendo en cuenta los dos procedimientos de estimación considerados.

Estimación MV

Cuando la correlación intragrupo es baja (c=1) y el número de individuos por grupo (k) chico, el estimador del coeficiente de regresión, en el caso de muestras chicas (n=10) presenta sesgos pequeños que disminuyen con el aumento del tamaño del grupo. Cuando el tamaño de la muestra aumenta (n ≥30), cualquiera sea el número de individuos por grupo, el sesgo es casi nulo.

A medida que aumenta la correlación intragrupo (c=2 o c=4) los sesgos para tamaños de grupo y muestra pequeños son mayores. Sin embargo a partir de muestras de tamaño 30 estos sesgos tienden a anularse.

Respecto a la variancia de los estimadores, sólo se aprecian estimaciones poco confiables de la misma cuando tanto el tamaño de la muestra como el número de individuos por grupo es muy chico.

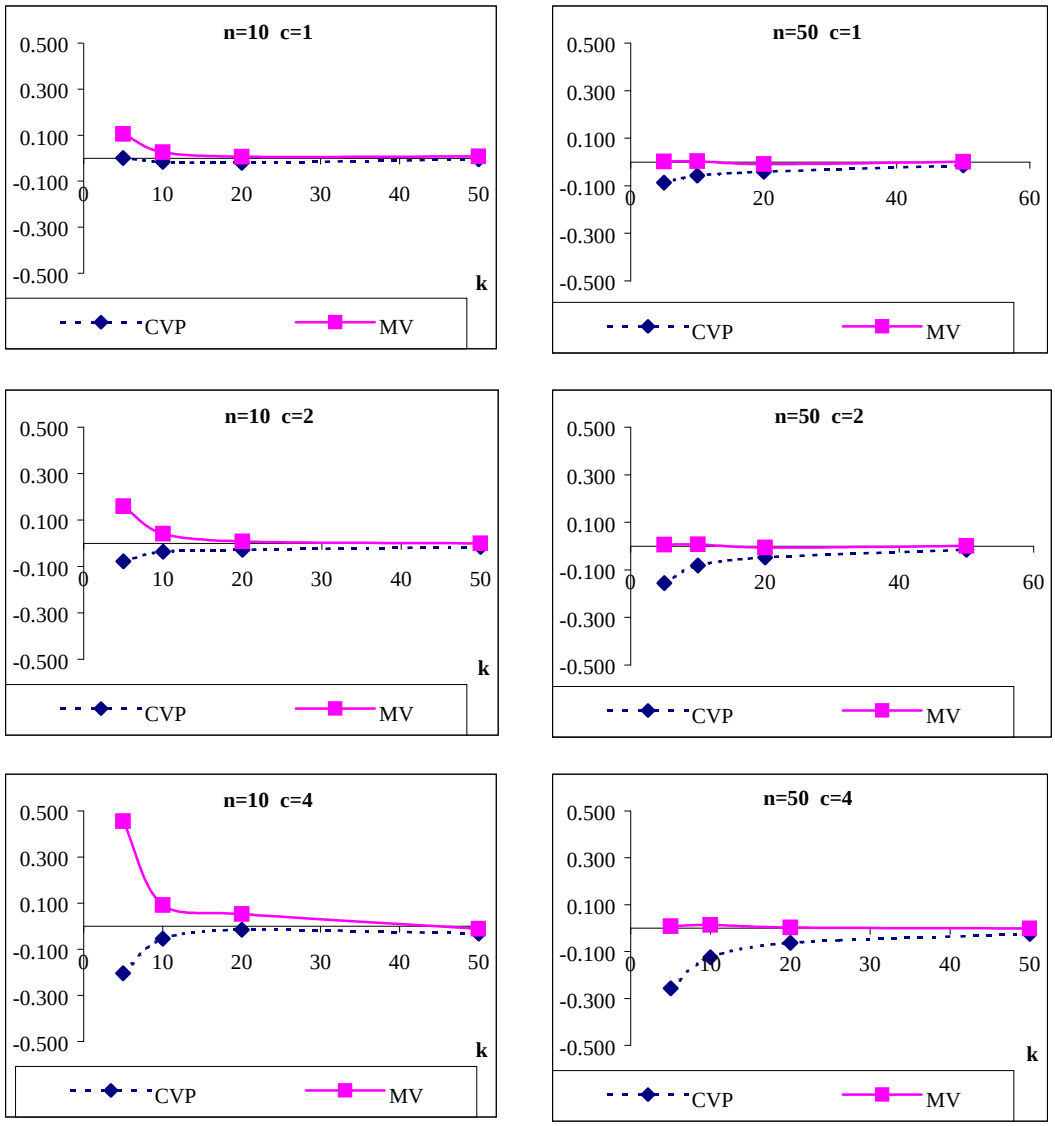
Estimación CVP

Cuando la correlación intragrupo es pequeña (c=1), para muestras pequeñas el estimador CVP del coeficiente de regresión casi no presenta sesgo, cualquiera sea el tamaño del grupo. Al aumentar la correlación intragrupo, es notable el aumento del sesgo para grupos de tamaño pequeño (k=5 y 10).

En relación a la estimación de la variancia de los estimadores, este procedimiento proporciona estimaciones confiables en todos los casos.

Los gráficos resultan más elocuentes para demostrar el comportamiento de los sesgos asociados a las estimaciones obtenidas por ambos procedimientos (Figura 1).

Figura 1: Promedio del sesgo en la estimación del efecto del coeficiente de regresión obtenido de acuerdo a ambos métodos de estimación según número de grupos (n) y niveles de correlación (c) para 10 y 50 individuos por grupo (k)



En ellos es posible observar la notable influencia del tamaño de la muestra (n) en el comportamiento MV del coeficiente de regresión, mostrando sesgos despreciables aún cuando el tamaño del grupo (k) es reducido si n es igual a 50. En cambio el estimador CPV reduce el sesgo al aumentar k siguiendo una trayectoria similar para diferentes tamaños de muestra.

5. CONSIDERACIONES FINALES

En este trabajo se estudia el comportamiento de los estimadores obtenidos por los métodos de estimación de cuasiverosimilitud penalizada y máxima-verosimilitud del efecto fijo asociado a una variable dicotómica medida a nivel individual en un modelo logístico simple con intercepto aleatorio.

El estudio por simulación pone en evidencia que el método CVP produce estimacio-

nes sesgadas con sesgos inversamente proporcionales al tamaño del grupo. Dicho sesgo también se ve afectado por el grado de correlación intra-grupo aumentando en forma directa con éste, pero, en cambio, poca es la influencia del tamaño muestral, es decir la cantidad de grupos en la muestra. Por el contrario, el estimador MV, si bien presenta sesgos apreciables para pocos grupos de tamaño reducido, dichos sesgos disminuyen notablemente si el número de grupos considerados aumenta.

Además de considerar estos resultados, es de destacar que otro elemento a tener en cuenta a la hora de elegir el método de estimación es el interés que se pueda tener en la estimación explícita de los efectos aleatorios. Si existiera este interés, el método de estimación apropiado es el de cuasiverosimilitud penalizada, ya que la estimación basada en el método numérico de cuadratura los considera parámetros de ruido e integra a través de ellos.

6. REFERENCIAS BIBLIOGRÁFICAS

Agresti, A. (2002). *Categorical Data Analysis*, 2nd ed. John Wiley & Sons.

Agresti, A; Booth, J.; Hobert, J.; Caffo, B. (2000) "Random-effects modeling of categorical response data". *Sociological Methodology*, 27-80

Breslow, N.; Clayton, D. (1993). "Approximate inference in generalized linear mixed models". *Journal of the American Statistical Association*, 88: 9-25.

Breslow, N.; Lin, X. (1995). "Bias correction in Generalized Linear Mixed Models with a Single Component of Dispersion". *Biometrika*, 82: 81-91.

Bellamy, L.; Li, Y.; Lin, X.; Ryan, L. (2005). "Quantifying PQL bias estimating cluster-level covariate effects in generalized linear mixed models for group-randomized trials". *Statistica Sinica*, 15: 1015 -1032.

Evans, S. R.; Hosmer, D. W. (2004). "Goodness of Fit Tests for Logistic GEE Models: Simulation Results". *Communications in Statistics: Simulation and Computation*, 33(1): 247-258.

Fitzmaurice, G.; Laird, N.; Ware, J. (2004). *Applied longitudinal analysis*. John Wiley & Sons.

Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd ed. Kendall's Library of Statistics. London.

McCulloch, C. (1997). "Maximum likelihood algorithms for generalized linear mixed models" *Journal of the American Statistical Association*, 92: 162-170.

Moerbeek, M.; Van Breukelen, G.; Berger, M. (2003). "A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies". *Journal of Clinical Epidemiology*, 56: 341-350.

SAS Institute, Inc. (2004). SAS/STAT User's guide, version 9.1.3. Cary, NC, USA.

Song, P. X. K. (2007) *Correlated data análisis: modeling, analytics and applications*. Springer, New York.