Techniques for multilevel data: Application to the determinants of educational performance

Herrera Gómez, Marcos and Aráoz, M. Florencia and de Lafuente, Gisela and D'jorge, Lucrecia and Granado, M. José and Michel Rivero, Andrés and Paz Terán, Corina

Universidad Nacional de Tucumán (Argentina)

2005

Técnicas para Datos Multinivel:

Aplicación a los Determinantes del Rendimiento Educativo[1]

HERRERA GOMEZ, Marcos; ARÁOZ, Ma. Florencia; de LAFUENTE, Gisela; D´JORGE, Ma. Lucrecia; GRANADO, Ma. José; MICHEL RIVERO, Andrés; PAZ TERÁN, Corina

Magíster en Economía. Universidad Nacional de Tucumán
Año 2005

Resumen

Este trabajo consiste en la aplicación de la metodología multinivel, que contempla la interacción entre variables individuales y grupales. Se intenta medir los determinantes del rendimiento de los alumnos del último curso del polimodal en tres departamentos de la provincia de Tucumán (censo año 2000). Los datos tienen una estructura multinivel al pertenecer los alumnos a distintos colegios. El método GEE "Generalized Estimated Equation", adecuado a datos en conglomerados, permite modelar la correlación que existe entre los estudiantes dentro de un mismo colegio. Se observa un mejor ajuste respecto a Mínimos Cuadrados que supone independencia entre las observaciones.

Códigos JEL: [I2], [C4]

Abstract

This paper consists of the application of the Hierarchical lineal model (multilevel methodology) which takes in consideration the interaction between individual and aggregated variables. It is intented to measure determinants of student performance in their last year of school in three departments of Tucuman's province (educational census year 2000). Data has multilevel structure indeed students belong to different schools. The GEE method "Generalized Estimated Equation", suitable for conglomerate data, allows to model correlation between students within the same school. So multilevel modeling strategies are more likely to produce unbiased estimators than Least Squares, which suppose independence between observations.

Códigos JEL: [I2], [C4]

[1] Se agradece al Dr. Shrikant Bangdiwala y al Dr. Victor J. Elías por los comentarios realizados; al Dr. Juan José Llach por las sugerencias en cuanto al material especializado; y a las Sras Rosa María Humbert y María del Carmen Torino de la Dirección de Calidad Educativa de la Provincia de Tucumán por los datos suministrados.

Técnicas para Datos Multinivel:

Aplicación a los Determinantes del Rendimiento Educativo

1- Introducción

1.1- Objetivo

El objetivo del presente trabajo consiste en la comparación de métodos econométricos aplicados para datos multinivel. En particular nos proponemos estimar los determinantes del rendimiento escolar de los alumnos. Los métodos comparados son Mínimos Cuadrados (MC), Modelos de Promedios Poblacionales (GEE) y Modelos de Efectos Mixtos (Mixed).

1.2- Introducción

Cuando hablamos de datos que tienen una "estructura jerárquica" (o en forma de "cluster"[2]), nos referimos a unidades agrupadas en diferentes niveles. Frecuentemente, los modelos que trabajan ese tipo de datos se presentan en las áreas de salud y educación, donde la información se encuentra en forma anidada (por ejemplo, estudiantes dentro de escuelas).
El ignorar la importancia de los efectos de los grupos puede invalidar los resultados obtenidos del estudio de las relaciones entre datos con esas características (Goldstein; 1995). De modo que si empleáramos MC no podríamos diferenciar ni dimensionar apropiadamente los efectos de las características grupales e individuales, así como tampoco analizar sus interrelaciones. Este último aspecto es clave cuando hablamos de "Rendimiento Escolar" pues el mismo está influenciado simultáneamente por factores grupales (escuela, aula) e individuales (aptitud, nivel socioeconómico, sexo, etc).
El análisis multinivel nos permite estimar sin *sesgos* el efecto de variables contextuales e individuales. Además permite saber si la magnitud del efecto de los factores varía dentro de los diversos niveles de agregación y permite estimar también de forma *insesgada* las posibles interacciones entre los factores individuales y contextuales.
Cuando hablamos de educación, las observaciones individuales no son, en general, completamente independientes. En efecto, los alumnos de un mismo colegio tienden a parecerse entre ellos debido a un proceso de selección (por ejemplo, algunas escuelas atraerán principalmente alumnos de un nivel socioeconómico elevado) y a una historia común que los alumnos comparten por el hecho de concurrir a la misma escuela. De esta forma, la correlación promedio entre las variables de los alumnos de la misma escuela (conocida como la correlación intra-clase) será mayor que la correlación de las mismas variables medidas entre los alumnos de escuelas distintas. Los tests estadísticos tradicionales descansan en el supuesto de independencia de las observaciones, y como este supuesto no se cumple en esta clase de estructuras, los errores estándar estimados serán bastante reducidos, y esto conducirá a que la mayoría de los resultados sean espurios (Hox; 1995).
Este trabajo comienza con el análisis metodológico, a continuación se hace un recuento de los antecedentes del enfoque de la función de producción en educación. Luego se exponen el análisis descriptivo de los datos y los resultados comparando los distintos modelos de estimación. Por último, las conclusiones del trabajo.

[2] "Cluster" es un agrupamiento que contiene elementos de un menor nivel. Por ejemplo, alumnos en una escuela. El "nivel", por otra parte, es un componente de los datos jerárquicos. El nivel 1 es el menor nivel; por ejemplo, estudiantes dentro de las escuelas.

2- Análisis Metodológico

2.1. Datos con estructura multinivel

Los datos corresponden a una estructura multinivel cuando las variables son medidas en diferentes niveles, y estos niveles constituyen un diseño jerárquico o estructura anidada.
Existen dos estructuras principales de datos multinivel:

1. **Por conglomerado:** los datos presentan información acerca de sujetos o individuos que son parte de un grupo seleccionado (estudiantes que pertenecen a distintas escuelas: escuela>estudiante). Usualmente estos datos son recogidos en estudios observacionales con muestreo por conglomerado o en ensayos comunitarios.
2. **Longitudinales:** los datos presentan información acerca de qué ocurre en un grupo de unidades de investigación durante distintos momentos a través del tiempo (mediciones repetidas de un individuo a lo largo del tiempo: individuo>medición). Usualmente estos datos son recogidos en estudios de diseño longitudinal, como los estudios de cohortes.

El análisis estadístico tradicional se enfoca en un solo nivel. Cuando existen varios niveles se debe considerar que los datos están anidados y por lo tanto no debe ignorarse que existe una correlación entre las mediciones de la variable dependiente.
Esta correlación existente entre los individuos que están dentro de un mismo conglomerado, o entre las mediciones en el tiempo para un mismo sujeto, se denomina correlación intraclase, y representa el grado de similitud entre las unidades que pertenecen a un mismo grupo. Por ejemplo, cuanto más parecidos son los rendimientos académicos de los alumnos de un mismo colegio, más probable es que las causas del rendimiento tengan que ver con el colegio. La ausencia de dependencia en este caso, implica ausencia de efectos del colegio en el rendimiento individual.
Existen varias definiciones de este coeficiente de correlación (dependiendo del supuesto sobre el diseño de la muestra), pero básicamente puede representarse como:

$$\bullet\ \rho_I = \frac{\tau^2}{\tau^2 + \sigma^2} \qquad \text{E-1}$$

donde, τ^2 es la varianza **entre** grupos (between-group), σ^2 es la varianza **dentro** de los grupos (within-group), y donde $\tau^2 + \sigma^2$ es la varianza total de la variable dependiente.
En una estructura de datos multinivel, las variables se clasifican en variables a nivel macro (el nivel más alto) y variables a nivel micro (nivel más bajo). Por ejemplo, con datos por conglomerados las variables propias del estudiante son variables a nivel micro (sexo, edad, nivel económico-social), mientras que las variables que caracterizan a una escuela son variables a nivel macro (pública o privada). Con datos longitudinales, las variables a nivel micro corresponden a las mediciones en el tiempo de un individuo, mientras que las variables a nivel macro son las que caracterizan al individuo.

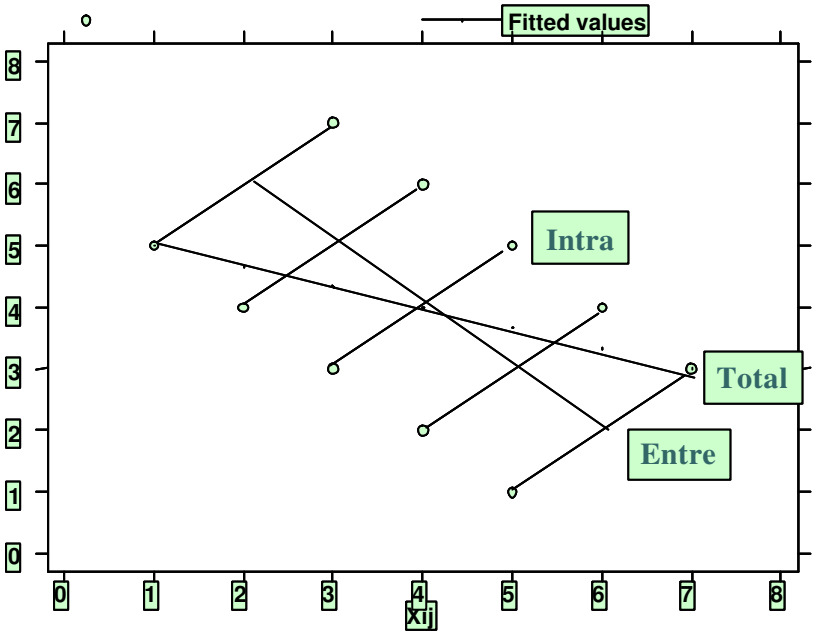
2.2. Análisis de los datos multinivel[3]

A la hora de analizar los datos se debe tener en cuenta cuál es la pregunta de investigación. Con datos por conglomerado si la pregunta de investigación se refiere al nivel macro (nivel colegio) basta con utilizar la metodología tradicional, con la correspondiente pérdida de información y potencia estadística, ya que se estarán promediando las observaciones de todos los estudiantes de una misma escuela. Con datos longitudinales, si la pregunta de investigación se refiere a un solo momento del tiempo, basta con el análisis tradicional y se pierde, igualmente, información y potencia estadística.

[3] Esta sección fue elaborada siguiendo a Twisk (2003) y Snijders & Bosker (1999) y apuntes de clases del Dr. Shrikant Bangdiwala.

Pero cuando la pregunta de investigación es de nivel micro o involucra datos de diferentes niveles, el análisis multinivel es el adecuado, ya que permite estudiar el efecto simultáneo de variables individuales y grupales, y sus respectivas interacciones, sobre una variable dependiente a nivel individual.
La estructura de nuestros datos es multinivel por conglomerado, ya que los alumnos están "anidados" en diferentes escuelas. Nuestro interés es conocer los efectos de variables propias del alumno (variables micro), como así también de variables a nivel del establecimiento al que pertenece (nivel macro), y sus interacciones, sobre el rendimiento del alumno (variable dependiente a nivel micro).
Al trabajar con datos multinivel, las regresiones a un nivel macro, es decir, agregando o promediando las variables por grupo, pueden resultar en coeficientes totalmente diferentes a los que se obtienen en regresiones a un nivel micro, es decir, una regresión por cada grupo. Los resultados para una base de datos hipotética pueden observarse en el siguiente gráfico.

Gráfico 1: Relaciones entre, intra y total. Adaptado de Snijders y Bosker (1999, p.28)



Las cinco rectas paralelas con pendiente positiva representan la relación dentro de cada grupo entre **Y** e **X** (relación intra); son las regresiones a nivel micro. La recta descendente más empinada representa la relación a nivel agregado, es decir, entre las medias de los grupos y corresponde a la regresión a nivel macro. La recta descendente más horizontal representa la relación total, y equivale a una regresión a nivel micro pero que ignora la estructura jerárquica o la dependencia entre las dos observaciones que hay en cada grupo.
Este ejemplo ilustra que los efectos de una variable explicativa sobre la variable dependiente dentro de un grupo, pueden tener incluso el signo contrario al del efecto que surge al evaluarse esta relación entre grupos. La relación verdadera entre **Y** y **X** se revela sólo cuando las relaciones between y within son consideradas conjuntamente, y esto se logra con un análisis de regresión multinivel.
Un modelo de regresión múltiple tradicional puede expresarse:

$$\bullet\ Y_{ij} = \beta_0 + \sum_{k=1}^{K} \beta_{1k} X_{ijk} + \sum_{m=1}^{M} \beta_{2m} Z_{jm} + \varepsilon_{ij} \qquad \text{E-2}$$

donde **i** representa a cada individuo (en nuestro caso, a cada alumno), **j** representa cada grupo o nivel macro (colegios), **k** representa cada variable a nivel micro (sexo del alumno, su nivel económico social, si repitió o no alguna vez un año, etc.), **m** representa cada variable a nivel macro (si se trata de un colegio público o privado, si tiene orientación técnica u otra, etc.).

Cuando se utiliza un modelo de este tipo para analizar datos multinivel, el supuesto latente es que la estructura multinivel está totalmente explicada por las variables grupales **Z** y las variables individuales **X**. Esto significa que si se consideran dos individuos y sus valores de **X** y **Z** están dados, entonces para su valor **Y** no interesa si ellos pertenecen al mismo o a diferentes grupos. Este tipo de modelos aún es utilizado en investigaciones con datos multinivel.

La idea básica del modelado multinivel es que la variable dependiente **Y**, que es una variable a nivel micro, tiene un aspecto individual como grupal. Lo mismo sucede con las variables independientes a nivel micro (**X's**): la media de $\mathbf{X_k}$ en un grupo puede ser diferente de la media en otro grupo, es decir, $\mathbf{X_k}$ tendrá una varianza entre grupos positiva. Esto significa que las variables explicativas a nivel individual frecuentemente también contienen alguna información sobre los grupos. Es por esta razón que incorporar a una regresión variables a nivel macro (**Z's**), no es suficiente para tomar en cuenta la estructura anidada de los datos.

El análisis de coeficientes aleatorios (Mixed), fue desarrollado inicialmente en las ciencias sociales, justamente para investigaciones en educación. Los investigadores, al observar que los resultados académicos de los alumnos dentro de una misma clase no eran independientes, y que los resultados de las clases dentro de un mismo colegio estaban correlacionados, concluyeron que debía realizarse alguna corrección por esta dependencia. La manera en que este tipo de análisis lleva a cabo esta corrección es permitiendo que los coeficientes de la regresión varíen de grupo a grupo.

Por lo tanto, en los modelos de efectos mixtos los coeficientes de la regresión pueden variar entre los miembros del nivel macro o grupos, y puede considerarse sólo el intercepto variable o el intercepto y la pendiente variables. Los parámetros variables usualmente no son de interés y se consideran aleatorios: la variación en el intercepto y la variación en las pendientes se distribuyen normalmente con media cero y una cierta varianza.

Un modelo de efectos mixtos con intercepto aleatorio sería:

$$\bullet\ y_{ij} = \beta_{0j} + \sum_{k=1}^{K} \beta_{1k} X_{ijk} + \sum_{m=1}^{M} \beta_{2m} Z_{jm} + \varepsilon_{ij} \qquad \text{E-3}$$

Se observa que β_0 varía para cada grupo *j*, y puede expresarse:

$$\beta_0 = \gamma_0 + U_{0j} \qquad \text{E-4}$$

Un modelo de efectos mixtos con intercepto y pendiente aleatoria sería:

$$\bullet\ y_{ij} = \beta_{0j} + \sum_{k=1}^{K} \beta_{1kj} X_{ijk} + \sum_{m=1}^{M} \beta_{2m} Z_{jm} + \varepsilon_{ij} \qquad \text{E-5}$$

Dado que las variables **Z** son a nivel grupal no tendría mucho sentido conceptualmente permitir que sus coeficientes dependan del grupo, es por eso que los coeficientes β_2 no cambian en la última expresión. Los que cambian son el intercepto y los coeficientes de pendiente β_1. Estos pueden expresarse:

$$\bullet\ \beta_{1k} = \gamma_{1k} + U_{1kj}\ , \quad \text{para k=1,...,K} \qquad \text{E-6}$$

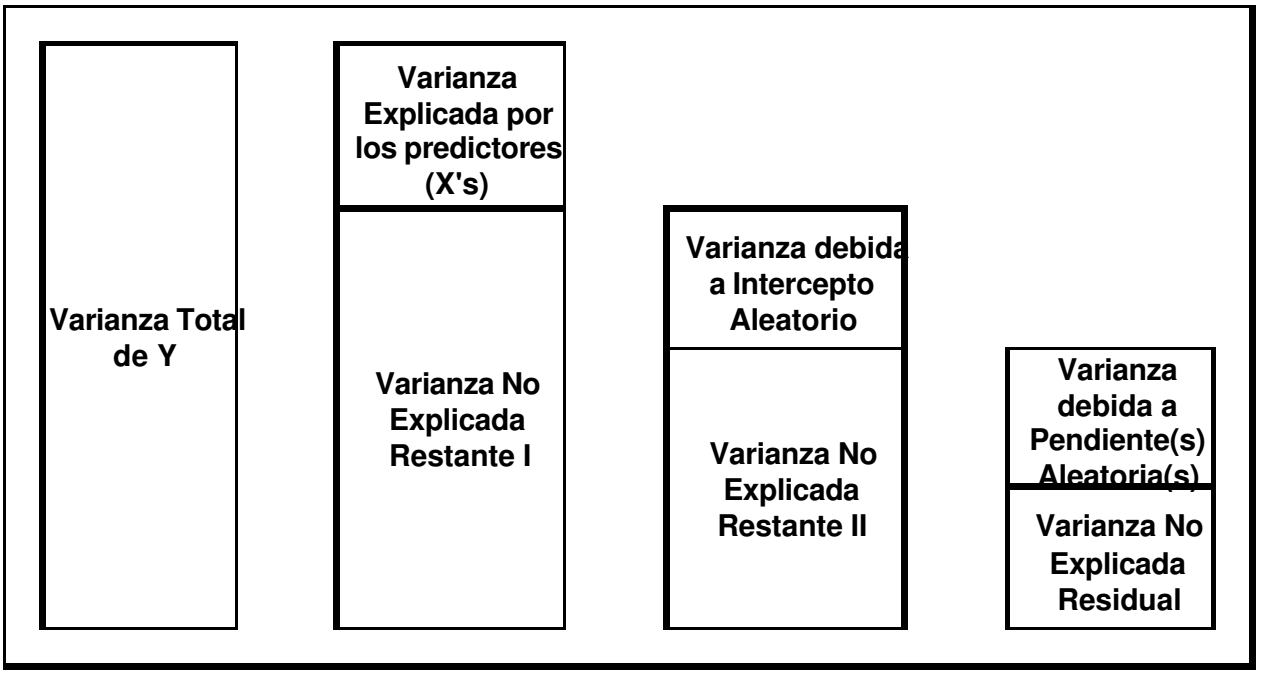
En general el modelo de efectos mixtos puede describirse como uno con efectos fijos y aleatorios (de ahí su nombre):

$$\underset{--}{Y} = \underset{--}{X}\underset{--}{\beta} + \underset{--}{Z}\underset{--}{\nu} \qquad \text{E- 7}$$

El vector $\underset{--}{\beta}$ captura los efectos fijos y el vector $\underset{--}{\nu}$, los efectos aleatorios. Los residuos ε están incluidos en la parte aleatoria de la regresión. En estos modelos nos interesa conocer los valores de los efectos fijos, $\underset{--}{\beta}$, y la contribución a la varianza total de establecer los coeficientes como aleatorios.[4]

La idea general del análisis con coeficientes aleatorios es que la varianza no explicada en la variable dependiente **Y** se divide en componentes diferentes. Uno de los componentes está relacionado con el intercepto aleatorio y otro componente está relacionado con las pendientes aleatorias. Estos componentes, constituyen la varianza a nivel macro (between groups). La varianza no explicada restante corresponde al nivel micro (within group), y es la varianza de los residuos o varianza no explicada residual. La suma de estas varianzas between y within, constituye la variabilidad total. Esto puede expresarse en el siguiente esquema:

Ilustración 1: Descomposición de la Varianza. Adaptado de Twisk (2003, p.81)



Por lo tanto, los modelos de efectos mixtos contienen variabilidad no explicada a dos niveles de anidamiento, y la participación de la variabilidad no explicada a través de los diversos niveles es la esencia de estos modelos.

A pesar de que los modelos antes descriptos son los más difundidos para analizar datos multinivel, existe otra técnica que incluso es más conveniente para datos por conglomerados. Esta técnica es conocida como el enfoque "Generalizad Estimating Equations" (GEE) o modelos de promedios poblacionales.

El GEE es un procedimiento iterativo que usa cuasi-probabilidad para estimar los coeficientes de regresión (Liang y Zeger, 1986). El método GEE corrige la no-independencia entre los individuos de un mismo grupo (o entre las observaciones de un mismo sujeto) asumiendo a priori una cierta estructura de correlación para la variable dependiente medida

[4] En esta última expresión matricial, las variables **X** y **Z** no se corresponden con las variables a nivel micro y macro como en las expresiones anteriores. Aquí la **X** corresponden a lo que queda en la parte fija de la regresión, incluyendo variables macro y micro, y la **Z** corresponde a la parte aleatoria, es decir que incluye los $\mathbf{U_0}$, los $\mathbf{U_{1j}}$ y los $\mathbf{e_{ij}}$.

en un mismo grupo, es decir, se especifica una matriz de correlación entre los individuos dentro de un nivel macro:

$$\bullet \underset{--}{Y} = \underset{--}{X}\underset{--1}{\beta} + Z\underset{--2}{\beta} + \underset{--}{\varepsilon} \qquad \text{E- 8}$$

$$\bullet \varepsilon = \sigma_e^2 \begin{bmatrix} \Sigma & 0 & . & . & 0 \\ 0 & \Sigma & 0 & . & 0 \\ . & 0 & . & . & . \\ . & . & . & . & . \\ 0 & 0 & . & . & \Sigma \end{bmatrix} \qquad \text{E- 9}$$

donde Σ es la matriz de correlación entre los individuos de un mismo conglomerado. Existen tantas matrices Σ como conglomerados bajo análisis y el orden de estas matrices cuadradas puede diferir y corresponde al número de individuos por conglomerado.
En este caso se supone que las observaciones de los individuos de un mismo conglomerado están correlacionadas, esta correlación es la misma para cada conglomerado (Σ)[5], pero no existe correlación entre los conglomerados (**0**).
En estos modelos nos interesa conocer los valores de los parámetros β**,** y debemos imponer una estructura para las correlaciones. Existen varias estructuras posibles; una de ellas es la estructura de correlación "exchangeable" (intercambiable) o "compound symmetric", que es la usada en nuestra aplicación al ser la más apropiada para datos en conglomerado. Esta puede expresarse de la siguiente manera:

$$\bullet \Sigma = \begin{bmatrix} 1 & \rho & . & . & \rho \\ \rho & 1 & \rho & . & \rho \\ . & \rho & . & . & . \\ . & . & . & . & . \\ \rho & \rho & . & . & 1 \end{bmatrix} \qquad \text{E- 10}$$

que indica que al tomar el alumno A y el alumno B del mismo colegio, existe una relación entre ellos igual a la que existe entre A y C o a la de los alumnos C y D, es decir no hay un orden entre los alumnos. Esta estructura requiere la estimación de un solo parámetro: ρ**.**
Otras estructuras son más adecuadas para datos longitudinales, por ejemplo, las estructuras **m**-dependiente o la autoregresiva. Otras posibles estructuras son la no-estructurada, en la que todas las correlaciones son distintas, y la independiente, que va contra la intuición, porque asume independencia entre las observaciones y lo que se busca corregir es justamente la presencia de una dependencia entre las observaciones. La potencia del análisis estadístico está influenciada por la elección de la estructura, dado que cada una de ellas supone una cierta cantidad de parámetros a estimar. Por lo tanto la mejor estructura de correlación será la más simple de la que se ajusten a los datos.
El procedimiento de estimación en el análisis GEE, puede verse de la siguiente manera: Primero, se lleva a cabo un análisis de regresión lineal ingenuo, que asume que las observaciones dentro de un grupo son independientes. Luego, basado en los residuos de este análisis, se calculan los parámetros de la matriz de correlación. Por último, se re-estiman los coeficientes de regresión, corrigiendo por la dependencia de las observaciones. Aunque el proceso completo es un poco más complicado (el proceso alterna entre los pasos dos y tres, hasta que los coeficientes de la regresión y los desvíos estándar se estabilizan), básicamente consiste en estos tres pasos.

[5] El Programa Stata 9.0 permite especificar Σ diferente para cada conglomerado.

Esencialmente GEE trata la estructura de correlación “intraclase” como una variable de molestia y corrige la dependencia entre las observaciones de un mismo grupo según la siguiente ecuación:

$$\bullet\ y_{ij} = \beta_0 + \sum_{k=1}^{K} \beta_{1k} * X_{ijk} + \sum_{m=1}^{M} \beta_{2m} * Z_{jm} + corr_{ij} + \varepsilon_{ij}^{*} \qquad \text{E- 11}$$

El análisis GEE combina una relación dentro del grupo con una relación entre grupos, resultando en un coeficiente de regresión para cada variable explicativa, y esto tiene una implicancia para la interpretación de tales coeficientes. Supongamos que el coeficiente de una variable explicativa $\mathbf{X_1}$ es igual a 0.5. La interpretación de la magnitud del coeficiente tiene dos caras: (1) la interpretación entre grupos, indica que una diferencia entre dos grupos de una unidad de la variable $\mathbf{X_1}$, está asociada con una diferencia de 0.5 unidades en la variable dependiente **Y**; (2) la interpretación dentro del grupo indica que un cambio dentro de un grupo de una unidad en la variable $\mathbf{X_1}$ está asociado con un cambio 0.5 unidades en la variable **Y.** La interpretación “real” del coeficiente de regresión es una combinación de ambas relaciones, pero no es posible determinar la contribución de cada parte.

3- Función de producción educativa: antecedentes empíricos

3.1- Aspectos conceptuales

En la literatura económica se hace referencia a la importancia de la educación en la formación del capital humano como factor de crecimiento; y se observa que a nivel internacional los países más ricos son también "los más educados". El nivel de desarrollo de un país está fuertemente asociado a su stock de capital humano medido, fundamentalmente, por el nivel y la calidad educativa de los individuos[6].

Conocer cuáles son los factores que determinan esa calidad ha sido objeto de numerosos estudios. Los mismos consideran que el rendimiento educativo es una buena proxy de calidad [7] y concluyen que sus determinantes pueden agruparse en las siguientes categorías[8]:

1. Factores propios de cada **persona**, tales como el sexo o las habilidades innatas;
2. Factores propios de la **familia**, como el nivel socioeconómico (NES)[9], su tamaño o la presencia de uno o ambos padres;
3. Factores propios del **lugar de residencia**, ya sea el país, la provincia o la ciudad, según cual sea el nivel de agregación relevante al estudio;
4. Factores propios de las **escuelas** y los **maestros**.

Los modelos empleados para relacionar el rendimiento académico del alumno con cada uno de estos factores se basan en la teoría microeconómica de la firma, utilizando una función de producción educativa en la que intervienen variables tanto escolares como ambientales. A través de ella se trata de representar simplificadamente el proceso educativo y de determinar la significancia estadística de los factores que influyen en dicho proceso. Este tipo de estudio se presenta como una herramienta útil para la determinación de la importancia relativa de cada uno de los factores en el logro del producto.

El modelo que se plantea generalmente es el siguiente:

$$L_{i,t} = f(F_{i,t}, P_{i,t}, S_{i,t}, A_{i,t}) + \varepsilon, \qquad i = 1, ...N \qquad \text{E- 12}$$

donde:

$L_{i,t}$ es el rendimiento escolar del estudiante i en el período t

F_i son factores acumulativos de la familia del estudiante i

P_i son características de los pares (ambiente) del estudiante i

S_i son características de la escuela y los profesores del estudiante i

A_i son características del estudiante i

e es el error aleatorio

Es importante notar que el rendimiento de un alumno en el instante t es el producto de los numerosos factores y estos tienen entrada en distintos momentos del tiempo. Algunos provienen de las condiciones innatas del estudiante, mientras que otros dependen del ambiente en el que se encuentra interactuando. Además los efectos que provocan estos factores pueden ser duraderos, o disminuir en el tiempo. En este sentido debemos reconocer que la educación es un proceso acumulativo.

Los insumos familiares tienden a ser medidos con las características sociodemográficas, como la educación de los padres, ingresos y tamaño familiar. Las características de los pares tienen que ver con esas mismas variables pero medidas en sus compañeros. Los insumos de la escuela incluyen características de los profesores; y en cuanto a la organización de la misma se toman variables como el tamaño de la clase.

---

[6] Otros componentes comúnmente citados para cuantificar el Capital Humano son el "Learning by Doing" y la Experiencia.

[7] En este punto reconocemos que tomar al rendimiento educativo como sinónimo de calidad es algo limitado, pues quedan excluidas las capacidades no cognoscitivas (ética, motivación, iniciativa, etc.)

[8] Llach, Juan Jose et. al; Educación para todos; 1999.

[9] El NES es una variable adoptada en los estudios del área. Tiene por finalidad aproximar el nivel socioeconómico del grupo familiar.

La limitación de este tipo de estimación tiene dos grandes aristas. La primera se relaciona con que no existe una adecuada medida de habilidad de los estudiantes. Muchas veces resulta imposible medir determinadas variables que hacen al rendimiento educativo del joven ya sea porque se carece de los recursos para hacerlo (coeficiente intelectual) o porque simplemente es imposible o sumamente complicado cuantificar ciertos fenómenos (capacidad del individuo para insertarse en el mercado laboral). Por otro lado, mientras que la educación es un proceso esencialmente dinámico y acumulativo, por lo que los insumos utilizados en el pasado afectarían el desempeño presente de los alumnos, en general sólo existen medidas contemporáneas de tales inputs. Cada uno de estos problemas conduce a un sesgo en los efectos estimados de los insumos educativos.

3.2- Literatura Empírica

Uno de los trabajos más conocidos y controversiales, aunque no el primero en el área de Economía de la Educación, fue el de Coleman et. al. (1966) pues concluyeron que los insumos escolares tenían poco o ningún efecto sobre las diferencias en el desempeño escolar, es decir que las diferencias en los resultados escolares se debían principalmente a la variación en el origen social del estudiante y que los antecedentes familiares explicaban una proporción altamente significativa de las diferencias en los logros. Posteriormente Jencks (1972), confirmaba los hallazgos de Coleman, concluyendo que, en la explicación del rendimiento escolar, lo más importante eran las características de los propios estudiantes, todo lo demás - recursos financieros de la escuela, sus políticas, las características de los maestros - tenía poca o ninguna relevancia.

Tabla 1: Efectos de algunos determinantes del rendimiento académico: Revisión Empírica – Economías Desarrolladas

VARIABLE	Estadísticamente SIGNIFICATIVA		Estadísticamente NO SIGNIFICATIVA
	Autores	Signo de la Relación	
Educación de los Padres	Coleman et. al. (1966) Jenks (1972) Summers & Wolfe (1977) Hanushek y Taylor (1990) Deller y Rudnicki (1993) Berger y Toma (1994)	Positiva	
Ingreso Familiar	Coleman et. al. (1966) Jenks (1972) Summers & Wolfe (1977) Hanushek y Taylor (1990) Deller y Rudnicki (1993) Berger y Toma (1994)	Positiva	
Tamaño de la Flia.			Resultados Ambiguos
Sexo	Summers y Wolfe (1977)	Los varones tienen, en promedio, rendimiento más bajo.	
Motivación del Estudiante[10]	Summers y Wolfe (1977)	Positiva	
Tamaño del curso	Krueger (1997)	Negativa	Summers y Wolfe (1977)
Tamaño de la Escuela	Summers y Wolfe (1977) Deller y Rudnicki (1993)	Negativa	
Nº de libros por alumno	Summers y Wolfe (1977)	Negativa	
Características Físicas de la Escuela			Summers y Wolfe (1977)

Fuente : Elaboración propia en base a Mizala y Romaguera (2000), Cervini (2002) y otros

[10] Medida a través de la asistencia a clases: Mayor ausentismo indicaría una menor motivación.

Hacia finales de los años ochenta, comenzó a emplearse una nueva técnica para abordar este problema, y revisiones realizadas con el uso de la técnica de meta-análisis (Cervini, 2002) concluyeron que no hay un factor que se sobreponga o que haga desaparecer totalmente los efectos de los otros y por lo tanto, la variancia total del rendimiento es explicable a través de una compleja gama de factores que interactúan entre sí y participan con una pequeña pero importante proporción. Los estudios parecen coincidir en que el nivel económico familiar es menos relevante que otras variables familiares, como por ejemplo, su capital cultural o sus actitudes y comportamiento relacionados a la escuela (Ej. el grado de injerencia en los estudios de los hijos). El balance final de las investigaciones en los países desarrollados es que los antecedentes familiares del estudiante se asocian significativamente con el rendimiento y si bien su peso relativo puede igualar al de otros factores, difícilmente será inferior.

Al revisar los estudios empíricos sobre la función de producción educativa en Economías Desarrolladas (**tabla 1**), encontramos que en varios de ellos se concluye que algunos de los insumos tradicionales no tienen un efecto estadísticamente significativo sobre el logro de los estudiantes. Sin embargo, esto no ocurre necesariamente en los estudios para las Economías en Desarrollo (**tabla 2**), posiblemente porque en estos países se observa una mayor dispersión en las características de los establecimientos educativos y los estudiantes, lo que permite estimar mejor la influencia de estos factores sobre el logro educativo.

Tabla 2: Efectos de algunos determinantes del rendimiento: Revisión Empírica Economías En Desarrollo

VARIABLE	Estadísticamente SIGNIFICATIVA		Estadísticamente NO SIGNIFICATIVA
	Autores	Signo de la Relación	
Variables Socioeconómicas	Mizala y Romaguera (1999) – Bolivia Mizala y Romaguera (2000) - Chile	Positiva	
Variables Asociadas con la Escuela y los Profesores[11]	Mizala y Romaguera (1999) – Bolivia Mizala y Romaguera (2000) - Chile	Positiva	
Establecimientos Públicos vs. Privados	Mizala y Romaguera (1999) – Bolivia Mizala y Romaguera (2000) - Chile	Los públicos tienen rendimientos menores.	
Sexo	Mizala y Romaguera (2000) - Chile	*Varones*: Mejor rendimiento en matemáticas *Mujeres*: Mejor rendimiento en lengua	
Tamaño del curso	Mizala y Romaguera (2000) - Chile	Negativa	Hanushek (1995)
Número de libros por alumno	Hanushek (1995)	Positiva	
Materiales de Aprendizaje de la Escuela	Harbison y Hanushek (1992) - Brasil	Positiva	
Características Físicas de la Escuela	Harbison y Hanushek (1992) – Brasil Hanushek (1995)	Positiva	

Fuente: Elaboración propia en base a Mizala y Romaguera (2000), Cervini (2002) y otros.

[11] Experiencia de los profesores, tareas diarias para la casa, tamaño de la escuela.

4- Datos y Análisis Descriptivo

En esta sección se describen los datos, se observan relaciones y se presentan ejemplos vinculados a los mismos.
En la provincia de Tucumán, en el año 2000, había 333.187 alumnos (aproximadamente un cuarto de la población total) distribuidos entre los diferentes niveles de la educación común. Para el último año del Nivel Medio o Polimodal había un total 12.245 alumnos repartidos en 192 unidades educativas. Nuestra población objetivo es la fracción de estos alumnos que en ese año residían en San Miguel de Tucumán, Tafí Viejo y Yerba Buena distribuidos en 97 colegios.

Tabla 3: El Nivel Medio en Números. Tucumán. Año 2000.

Jurisdicción	Población 2001	Unidades educativas Nivel Medio/Polimodal			Alumnos Inscriptos Nivel Medio/Polimodal		
		Total	Estatal	Privado	Total	Estatal	Privado
Total Provincial	1.338.523	192	79	113	61.355	37.122	24.233
Capital	527.607	90	22	68	30.048	15.643	14.405
Tafí Viejo	108.017	12	6	6	3.985	2.801	1.184
Yerba Buena	63.707	12	2	10	2.939	1.146	1.793
Total Localidades	699.331	114	30	84	36972	19.590	17.382
% Local./Total	**52,22**	**59,37**	**37,97**	**74,33**	**60,25**	**52,77**	**71,72**
Resto	639.192	78	49	19	24.383	17.532	6.851

Fuente: INDEC: Censo Nacional de Población y Vivienda 2001 - Dirección de Coordinación del SEN en base a datos de la Red Federal Educativa.

4.1- Fuente de datos

Los datos fueron obtenidos del censo realizado por el DINICE (Dirección Nacional de Información y Evaluación de la Calidad Educativa), organismo dependiente del Ministerio de Cultura y Educación de la Nación, en el marco del Operativo Nacional de Evaluación de Calidad Educativa, correspondiente al año 2000. El objetivo del mismo era medir los conocimientos de los estudiantes en las disciplinas de Lengua y Matemática, además de relevar información complementaria como la conformación de los hogares de los alumnos y su nivel económico social (NES), y la educación de los padres, entre otras.
Es importante destacar que el censo educativo realizado en el año 2000 provee una fuente de datos valiosa y de cobertura total para las escuelas y colegios del nivel medio o polimodal de la provincia de Tucumán.
Una fuente adicional de datos fue la Dirección de Calidad Educativa de la Secretaría de Educación de la Provincia de Tucumán que proporcionó valiosa información para el reconocimiento de los establecimientos y su ubicación geográfica.

4.2- Niveles y variables de estudio

Nuestros niveles de estudio son dos: **Colegios** en los que se imparte educación de Nivel Medio, ubicados en las localidades de San Miguel de Tucumán, Yerba Buena y Tafí Viejo (nivel macro); y **alumnos** del último año del Nivel Medio de esos colegios (nivel micro).
Nuestra variable a explicar es el rendimiento promedio del alumno. La construcción de la misma se basa en la calificación media obtenida por cada alumno en un examen que evalúa sus nociones en Lengua y en Matemáticas. El rango de calificaciones oscila entre 0 (puntaje mínimo) y 100 (puntaje máximo). Cabe aclarar que una calificación de 100 no implica excelencia, pues lo que se mide con estas evaluaciones es el alcance de los conocimientos considerados como mínimos para ese nivel.
La siguiente tabla resume nuestras variables explicativas teniendo en cuenta el tipo y la definición empleada en las estimaciones.

Tabla 4: Variables explicativas a nivel Micro

	Variables	Tipo	Definición
MICRO	Sexo	Dummy	0= Mujer
			1= Hombre
	K_libro[12]	Dummy	0= No tiene en su casa revistas o enciclopedias para estudiar
			1= Tiene revistas o enciclopedias para estudiar
	Comp.[13]	Dummy	0= No tiene computadora en su casa
			1= Tiene computadora en su casa
	No repite[14]	Dummy	0= Repitió al menos una vez
			1= No repitió ninguna vez
	No trabaja[15]	Dummy	0= Trabaja
			1= No trabaja
	No violencia[16]	Dummy	0= Presenció algún acto de violencia en el último mes
			1= No presenció algún acto de violencia en el último mes
	Respons[17]	Categórica	1= No tiene los libros, fichas y apuntes que le pidieron este año
			2= Tiene algunos de los libros, fichas y apuntes que le pidieron año
			3= Tiene todos los libros, fichas y apuntes que le pidieron este año
	Nes[18]	Numérica	Nivel Económico Social: Posesión de Bienes + Nivel educativo de los padres. Suma ponderada de ambos elementos.
	Pares Nes	Numérica	Promedio del Nes de los compañeros del alumno dentro del aula

Las variables de mayor interés para nuestro estudio son *Nes* del alumno, *Pares Nes*, *Compu*, *K_libro* y *Sector*. Las demás variables desempeñan la función de controlar características de los alumnos y de los establecimientos.

Tabla 5: Variables Explicativas a nivel Macro

	Variables	Tipo	Definición
MACRO	Sector	Dummy	0= Colegio Público
			1= Colegio Privado
	Classize_col	Numérica	Cantidad promedio de alumnos dentro del aula por colegio
	Técnico[19]	Dummy	0= Colegio con otra orientación
			1= Colegio con orientación técnica

4.3- Construcción del NES (Nivel Económico Social)

En una primera etapa construimos dos variables:

A. **Bienes:** Esta variable esta construida como la suma de respuestas a las preguntas binarias acerca de la posesión o no de determinados bienes en el hogar donde habita el alumno. Los bienes son: calefón/termotanque – freezer/heladera con freezer – cocina a gas – ventilador – horno microondas – video casettera – lavarropa – secarropa – computadora – internet – equipo de música – auto propio – teléfono –

[12] Pregunta 14 del cuestionario al alumno de Media.
[13] Pregunta 7_10 del cuestionario al alumno de Media.
[14] Pregunta 19 del cuestionario al alumno de Media.
[15] Pregunta 26_1 del cuestionario al alumno de Media.
[16] Pregunta 24_5 del cuestionario al alumno de Media. Respuesta 2, 3 y 4: tabulado 0; Respuestas 1: tabulado 1.
[17] Correspondiente a la pregunta 20 del cuestionario al alumno de media.
[18] Estandarización de la suma de la pregunta 7 mas estandarización de la suma de la pregunta 8 y 9.
[19] Estrato 6 del cuestionario al alumno de Media.

TV color – TV por cable – video filmadora – aire acondicionado. El máximo valor alcanzable por un alumno es 18.

B. **Educa:** Esta variable es la suma de la educación del padre y de la madre, cada una de las cuales está, a su vez categorizada, de la siguiente forma:
   - (2) Primario incompleto
   - (3) Primario completo
   - (4) Secundario incompleto
   - (5) Secundario completo
   - (6) Universitario/Terciario incompleto
   - (7) Universitario/Terciario completo

El dato faltante a esta respuesta puede ser un caso de no educación, o un caso en que el alumno no responde. El máximo valor alcanzable es 14, y el mínimo 2.
En la segunda etapa estandarizamos cada una de estas variables por su máximo valor, y luego sumamos las dos variables a igual ponderación.

$$NES = BE \times 0{,}5 + EE \times 0{,}5 \qquad \text{E- 13}$$

donde:
BE = *Bienes* estandarizado
EE = *Educa* estandarizado
A continuación presentamos una descripción de la cantidad de colegios por ciudad de acuerdo a si son de gestión pública o privada. Cabe destacar que las ciudades son geográficamente aledañas por lo que se podría considerar un único conglomerado.

Tabla 6: Establecimientos y Alumnos por Ciudad según Sector

Sector		San Miguel	Tafí Viejo	Yerba Buena	Total
Público	Establecimientos	22	3	1	26
	Alumnos	1674	200	47	1921
Privado	Establecimientos	59	3	9	71
	Alumnos	2141	62	258	2461

Fuente: Dirección de Evaluación Educativa de la provincia de Tucumán.

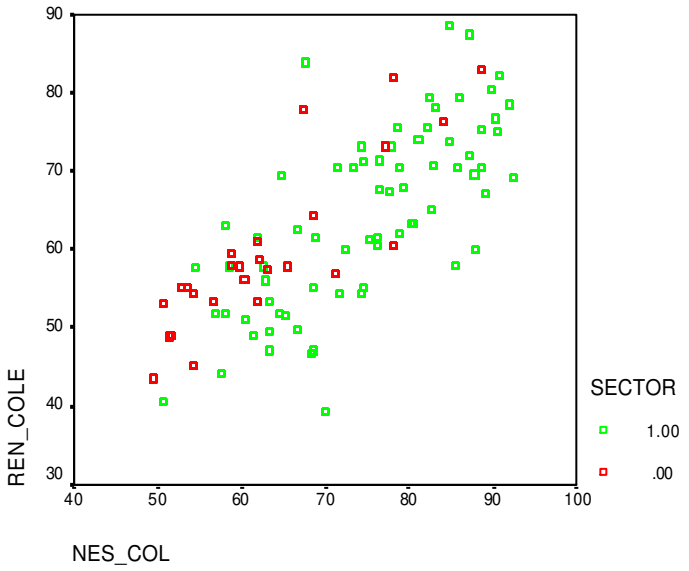
La siguiente tabla describe los estadísticos de las variables de interés. Encontramos que la media del rendimiento promedio entre las evaluaciones tomadas para medir los conocimientos mínimos de Lengua y Matemática es un valor cercano al 65%. Se puede destacar que el 46% de los alumnos poseen computadoras y que la cantidad de alumnos que han repetido al menos una vez es del 15%.

Tabla 7: Estadísticos Descriptivos

Variables	N	Mínimo	Máximo	Suma	Media	Desv. típ.
REN_PRO	4382	18.75	98.75		64.89	15.67
SEXO	4382	0	1	1858	.42	.49
K_LIBRO	4382	0	1	4235	.97	.18
COMPU	4382	0	1	2008	.46	.49
NO_REPITE	4382	0	1	3736	.85	.35
NO_TRABAJA	4382	0	1	3672	.84	.37
NO_VIOLENCIA	4382	0	1	3350	.76	.42
RESPONS	4382	1	3	9788	2.23	.57
NES	4382	14.29	100.00		70.96	16.85
PARES_NES	4382	41.67	93.78		70.96	11.86
SECTOR	4382	0	1	2461	.56	.49
TECNICO	4382	0	1	452	.10	.30
CLASSIZE_COL	4382	11	76.5		33.95	12.80

El siguiente gráfico muestra la presencia de una relación claramente positiva entre el rendimiento promedio y la variable NES:

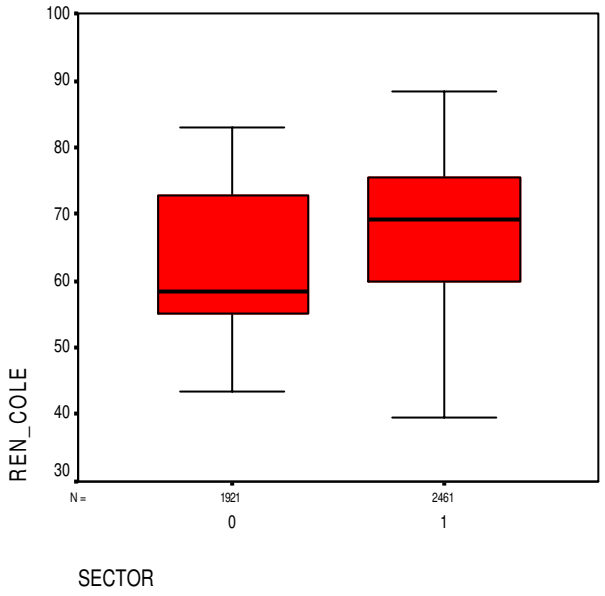
Gráfico 2: Rendimiento Promedio y NES por Establecimiento en cada sector



Esta relación nos indica que frente a un mayor nivel socioeconómico de los alumnos, su rendimiento promedio es mayor. Esto nos da un indicio de la relevancia de esta variable para nuestro análisis.

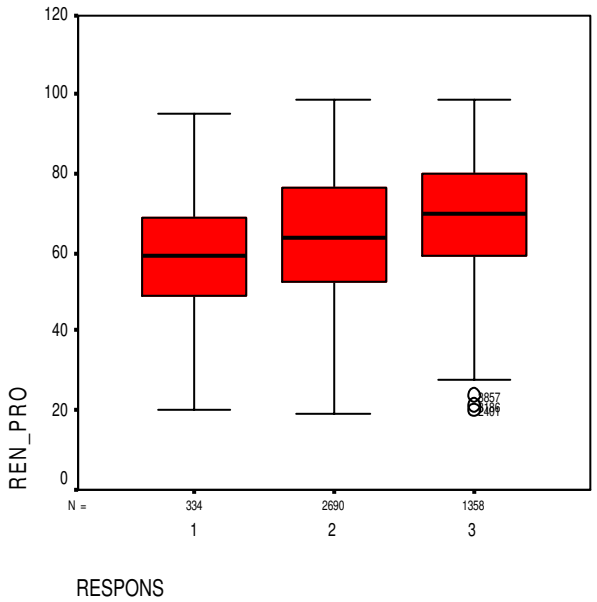
El siguiente gráfico muestra la diferencia entre los rendimientos de los colegios públicos y privados, observándose una diferencia en la mediana a favor de los últimos, pero con mayor dispersión.

Gráfico 3: Box Plot del Rendimiento Promedio por Establecimiento y en cada sector.



Para analizar el esfuerzo del alumno generamos la variable "Responsabilidad" (respons). El próximo gráfico presenta el Box Plot entre *ren*dimiento *pro*medio y esta variable e indica que los alumnos tienen un rendimiento superior (medido por la mediana) en la medida en que son más "responsables".

Gráfico 4: Box Plot del Rendimiento Promedio por Alumno y Responsabilidad

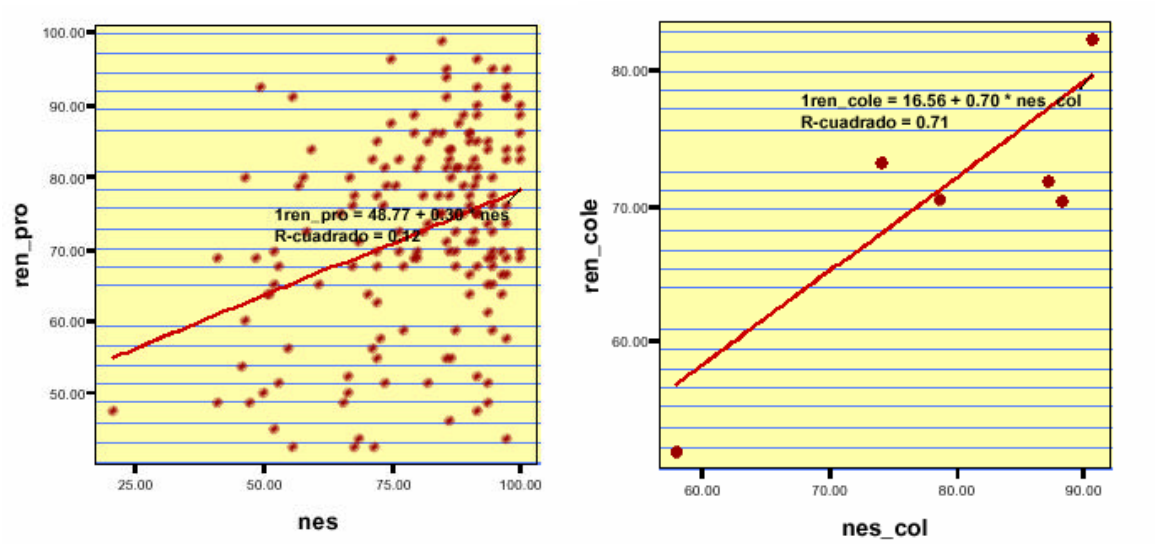


Para graficar la diferencia más notoria entre la metodología de Mínimos Cuadrados y la propuesta en este trabajo presentamos a continuación una muestra de 168 alumnos correspondientes a 6 establecimientos privados pertenecientes a la localidad de Yerba Buena.

El gráfico 4 muestra una estimación ingenua sin considerar que los alumnos están agrupados dentro de colegios. Esta estimación puede ser mejorada mediante la incorporación de variables dummies para capturar el efecto colegio pero la estimación con

MC continúa ignorando la correlación entre los individuos, o sea trabaja con $\Sigma_i = \mathbf{1}$ (matriz identidad de orden **i**).

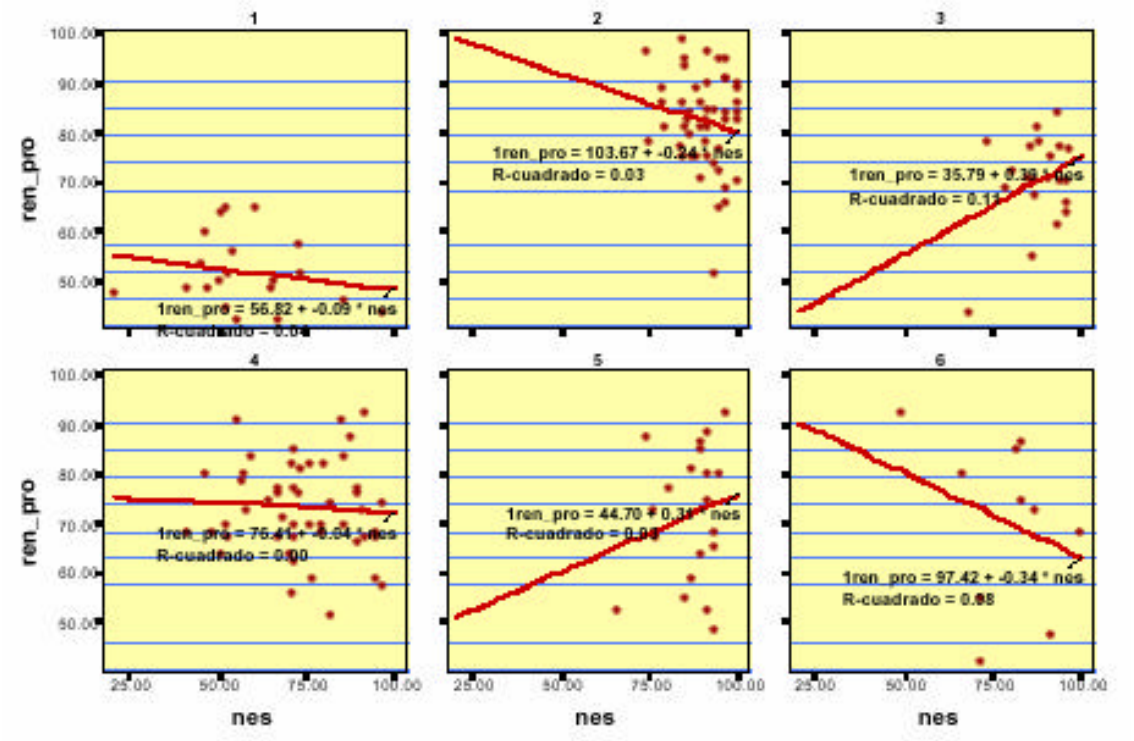
Gráfico 5: Regresión Total (MC) y Regresión *ENTRE*-establecimientos



Tradicionalmente cuando se conocía que los datos se encontraban anidados, los observaciones micro se resumían en un promedio a nivel macro realizándose una estimación a nivel macro como la observado en el gráfico 6. El problema de este procedimiento es que la estimación pierde potencia estadística debido a la disminución en la cantidad de observaciones.

El propósito del Modelo de Efectos Poblacionales (GEE) y del Modelo de Efectos Mixtos es capturar la relación tanto macro como micro. Los coeficientes estimados serán el resultado de una combinación de los coeficientes *intra* y *entre;* y su ponderación dependerá del grado de correlación intraclase. En nuestro ejemplo el GEE y el Mixed capturaran tanto la relación presentada en el gráfico 6 como en el 7, sin pérdida de potencia estadística.

Gráfico 6: Regresión *INTRA*-establecimientos



5- Estimaciones de los diferentes modelos

Esta sección contiene los resultados arrojados por las diferentes técnicas de estimación de los coeficientes buscados. El primer paso es calcular el grado de correlación intraclase (**?**), para justificar el empleo de una técnica que ajuste la dependencia de las observaciones; se muestra además las correlaciones entre las variables predictoras. En segundo lugar, presentamos una tabla resumen de los cuatro modelos considerados en la estimación. Por último, se realiza un análisis del residuo para cada modelo, para obtener una aproximación de la bondad del ajuste.

5.1-Estimación de la Correlación Intraclase. Empty Model

Tabla 8: Cálculo de la correlación Intraclase

ren_pro	Coeficente
sigma_u (variación *entre*)	11.14
sigma_e (variación *intra*)	11.28
Rho (?)	.49

El modelo vacío (empty) se calcula mediante una descomposición de varianza, comúnmente llamada ANOVA, de la variable dependiente (ren_pro). Vemos que la correlación intraclase es de un valor de 49%. Esto nos brinda la información sobre la conveniencia del análisis de tipo multinivel con datos anidados por colegios en el estudio de la relación entre el rendimiento promedio de los alumnos y las variables explicativas consideradas.

Tabla 9: Correlaciones entre variables predictoras

Variable	k_libr	pc	No Repite	No Trabaj	No Vcia.	resp.	nes	Par nes	cla col	sector	tecn.
k_libro	1										
compu	0.10	1									
no_repite	0.02	0.14	1								
no_trabaj	0.05	0.11	0.08	1							
no_viol.	0.01	0.02	0.02	0.07	1						
respons	0.10	0.14	0.09	0.03	0.07	1					
nes	0.18	**0.62**	0.18	0.14	0.02	0.20	1				
pares_nes	0.11	**0.43**	0.24	0.15	0.04	0.19	**0.66**	1			
class_col	-0.01	-0.11	0.01	-0.10	-0.05	-0.04	-0.18	-0.26	1		
Sector	0.06	0.24	0.08	0.15	0.05	0.08	**0.39**	**0.56**	**-0.44**	1	
tecnico	-0.02	-0.01	-0.06	-0.06	0.02	-0.03	-0.09	-0.12	-0.04	-0.23	1

Se resaltan en “negrita” las correlaciones (positivas y negativas) más elevadas.

5.2- Comparación de los modelos

El primer modelo que presentamos corresponde a una regresión simple estimada mediante Mínimos Cuadrados Ordinarios. Esta es una regresión ingenua ya que ignora la forma en que fueron recogidos los datos, y considera que las observaciones a nivel micro (alumnos) son independientes entre si.

La segunda regresión utiliza la corrección robusta de los errores y tiene en cuenta que las observaciones se encuentran anidadas respecto a la variable CUE (Colegio). [20]Los coeficientes estimados no difieren del modelo anterior, pero se corrigen los desvíos estándar de los estimados aumentándolos. Se mantienen las 4382 observaciones, ahora agrupadas en 97 colegios.

La siguiente estimación corresponde al modelo GEE robusto, apropiado para la estructura de nuestros datos. En este caso se especifica una estructura de correlación particular entre las observaciones de los alumnos dentro de un mismo colegio. La estructura de correlación elegida es Intercambiable (Exchangable o "compound symmetric"). Supone que la relación entre cualesquiera dos alumnos de un mismo colegio es la misma e implica estimar un único parámetro.

Finalmente estimamos el modelo Mixed con intercepto aleatorio. Este modelo trata de capturar la misma variabilidad que el modelo GEE, al permitir que los interceptos varíen de acuerdo a cada colegio, pero los coeficientes no son los mismos, dado que no logra capturar totalmente la correlación existente dentro de un colegio. El modelo de efectos mixtos es más adecuado para datos longitudinales, sin embargo se puede adaptar para datos en conglomerado.

Es notoria la diferencia entre los coeficientes de MC (1 y 2) con los otros modelos 3 y 4. Es importante observar que los coeficientes del GEE y Mixed presentan coeficientes bastante similares y las variables que presentan significancia estadística son las mismas. Los modelos MC, en cambio, difieren tanto en la magnitud y signo de los coeficientes como en la significancia de las variables.

El coeficiente de la variable sexo pasa de ser positivo (en **1** y **2**) a negativo (en **3** y **4**), aunque el MC robusto obtiene la no significancia del coeficiente igual que el GEE y el Mixed. Para la variable *no_trabaja*, las estimaciones de MC difieren sustancialmente de las de los otros modelos. Es de notar que el MC robusto trata de corregir el desvío pero deja inalterado el coeficiente provocando que la variable sea no significativa; por el contrario, los modelos **3** y **4** ajustan además el coeficiente capturando parte de la variabilidad debida a los clusters y esto hace que se mantenga la significancia.

Un coeficiente importante es el de Sector en donde todos los modelos, inclusive el ingenuo, estiman que no es significativo. El GEE y el Mixed, al capturar parte de la correlación entre los individuos, aumentan los valores del coeficiente pero también aumentan los errores estimados. Los coeficientes de las demás variables son similares en los dos últimos modelos.

Respecto a las variables de mayor importancia en la determinación del rendimiento se destaca Pares Nes, cuyo coeficiente es casi 7 veces superior al del Nes propio del individuo. Esto indicaría que el nivel socioeconómico de los pares es más relevante que el propio. Este hallazgo es consistente con los resultados obtenidos por otros autores. Otra variable importante es no_repite, ya que captura la historia de desempeño del individuo, y vale notar que su coeficiente es significativo en todas las estimaciones. Además se encuentran coeficientes significativos para K_libros, compu, respons,y no_violencia.

La variable classize_col, variable macro, es significativa y positiva para todas las estimaciones, lo cual va en contra de la intuición. Quizás esto pueda ser consecuencia del modo en que fue construida la variable: cantidad de alumnos del colegio en su último año dividido la cantidad de aulas del último año. Así puede que en realidad esta variable no capture el efecto aula (negativo) sino más bien el efecto escala (positivo), por el cual el

[20]En modelos con robust clusters, se tiene en cuenta la existencia de esta correlación, por lo tanto se corrigen los desvíos estándar de los coeficientes estimados pero no se modifican los estimados. En el primer modelo se sobreestima la significancia del efecto de cada variable; al corregir por la estructura anidada de los datos, los desvíos estándar aumentan y por esto los estadísticos t disminuyen. La corrección que realiza considera el efecto diseño: al existir correlación entre los individuos de una misma escuela la varianza del error se encuentra "inflada" por el siguiente factor que es el efecto diseño: $\sigma^2_{corr} = \sigma^2_{ind}\left[1+(n-1)\rho\right]$; donde σ^2_{corr} representa la varianza del error cuando existe una correlación dentro de grupos, y σ^2_{ind} representa esta varianza cuando las observaciones son independientes unas de otras.

tamaño del establecimiento se considera una Proxy del prestigio del colegio. Este problema debe ser analizado aún con mayor profundidad.
Otro posible problema en las estimaciones reside en los datos recogido ya que no se encuentran encuestados todos los alumnos de los cursos considerados, existen datos faltantes, que pueden estar generando un sesgo de selección.

Tabla 10: Comparación entre los modelos

Variables	MODELOS 1 MC INGENUO	2 ROBUSTO	3 GEE ROBUSTO	4 MIXED I. ALEAT.
ren_pro	Coef.y (s.d.)	Coef.y (s.d.)	Coef.y (s.d.)	Coef.y (s.d.)
sexo	0.85 (0.42)	ns 0.85 (1.09)	ns - 0.34 (0.49)	ns - 0.28 (0.45)
k_libro	2.34 (1.09)	2.34 (1.20)	2.29 (1.04)	2.27 (1.04)
compu	1.35 (0.50)	1.35 (0.43)	1.00 (0.37)	1.02 (0.44)
no_repite	4.74 (0.57)	4.74 (0.79)	2.86 (0.61)	2.94 (0.53)
no_trabaja	0.88 (0.53)	ns 0.88 (1.54)	1.70 (0.82)	1.72 (0.50)
no_violencia	1.90 (0.46)	1.90 (0.62)	1.41 (0.51)	1.46 (0.42)
Respons	1.85 (0.35)	1.85 (0.50)	1.03 (0.36)	1.08 (0.31)
Nes	0.09 (0.02)	0.09 (0.03)	0.07 (0.03)	0.08 (0.02)
nes_priv	- 0.07 (0.03)	ns - 0.07 (0.05)	ns - 0.03 (0.04)	ns - 0.03 (0.03)
pares_nes	0.69 (0.02)	0.69 (0.06)	0.46 (0.10)	0.52 (0.04)
Sector	ns 1.59 (1.83)	ns 1.59 (3.69)	ns 2.55 (3.32)	ns 1.84 (2.44)
Tecnico	ns 0.91 (0.69)	ns 0.91 (1.98)	ns 2.11 (2.41)	ns 2.13 (2.06)
Classize_col	0.18 (0.02)	0.18 (0.07)	0.16 (0.06)	0.16 (0.05)
_cons	- 8.41 (2.11)	ns - 8.41 (5.19)	ns 10.64 (7.19)	ns 6.24 (3.78)
Varianza **ren_pro**	245.42	245.42	245.42	245.42
R-squared	0.34	0.34	0.30	0.33
Rho			0.36	0.20
Scale parameter:			170.82	

Errores estándares entre paréntesis
ns No significativo al 90% de confianza.

Al no variar demasiado los resultados con GEE y Mixed (modelos que tienen en cuenta la estructura multinivel), consideramos que el primero es el que mejor se ajusta a la estructura por conglomerado de nuestros datos y por eso lo consideramos como el modelo definitivo.

A partir del modelo elegido (GEE robust) podemos concluir que:

- Un aumento del nivel económico social promedio de los pares de 1 punto porcentual incrementa el rendimiento del alumno en 0.46 puntos. Recordar que el rendimiento se mide de 0 a 100.
- Un aumento del nivel económico social del alumno de 1 punto porcentual incrementa su rendimiento en 0,07 puntos.
- Si se trata de un alumno que no ha repetido un curso, este tiene un rendimiento superior en 2.86 puntos respecto a un alumno que repitió alguna vez.
- Un alumno que tiene algún libro o enciclopedia en su casa tendrá un rendimiento superior en 2,29 puntos en relación a alguno que no tiene.
- Si el alumno no trabaja tendrá un rendimiento superior en 1,70 puntos en relación al alumno que trabaja.
- Un alumno que manifiesta que su colegio es un lugar seguro (entiéndase que no ha presenciado ningún acto de violencia en el último mes) tendrá un rendimiento superior en 1,41 puntos en relación al alumno que señala inseguridad.
- El alumno que tiene computadora en su casa tendrá un rendimiento superior en 1 punto en relación al alumno que no tiene.
- El aumento de un punto porcentual del *classize_col*, aumenta en promedio el rendimiento de cada alumno de dicho colegio en 0,16 puntos.

En resumen, de las 8 variables independientes consideradas para explicar el rendimiento del alumno, el promedio del nivel económico-social del colegio es la que cobra mayor relevancia por la magnitud del impacto estimado, mientras que el efecto de las demás es prácticamente despreciable al considerarse sobre un total de 100 puntos de rendimiento.[21]

Por otro lado, se debe tener en cuenta que el costo de política gubernamental que implica aumentar 0.10 el nivel económico-social promedio del colegio es muy alto, ya que comprende mejorar la educación de los padres como el nivel de ingresos de todo un conjunto de alumnos.

5.3- Bondad del ajuste

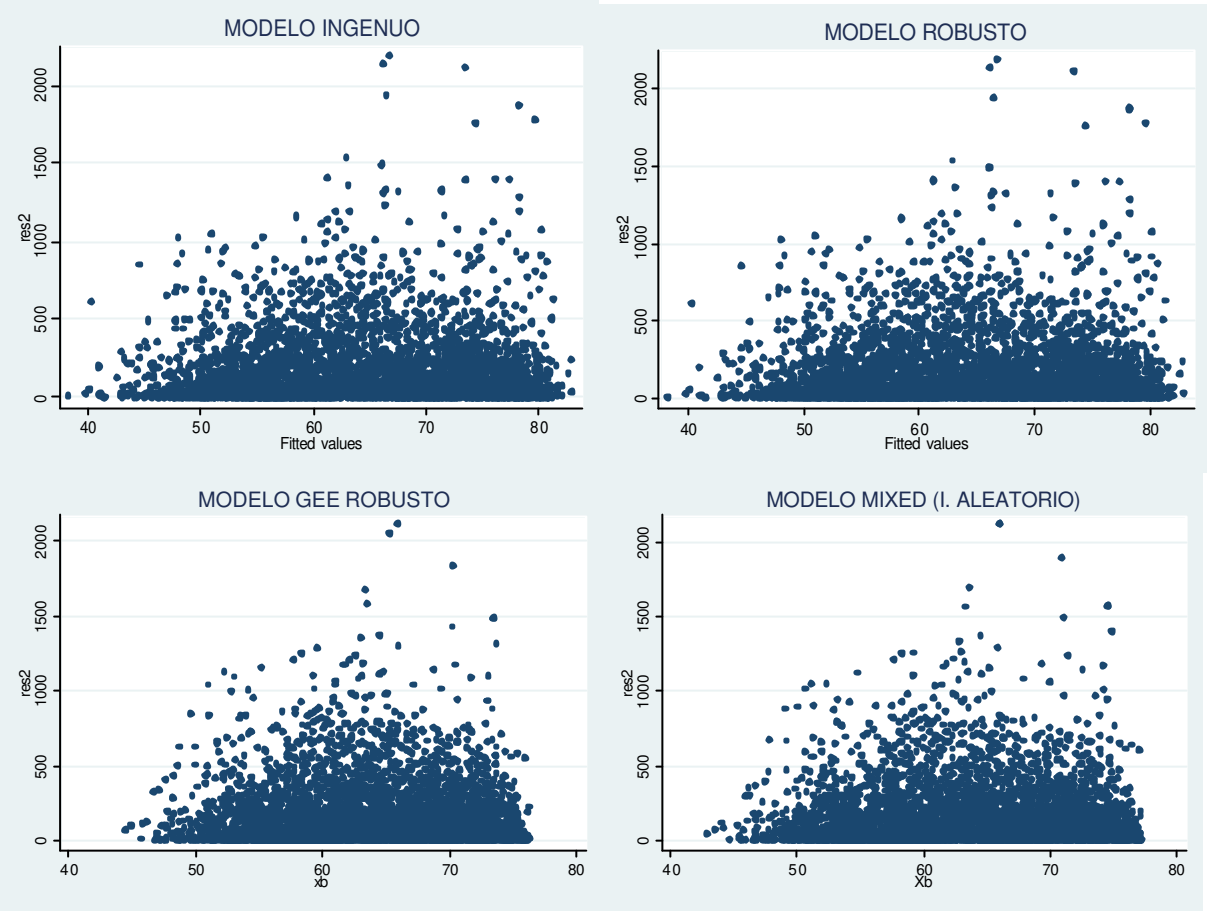
A partir de las regresiones se obtuvo la medida de bondad del ajuste[22] para los cuatro modelos. La mínima es del 30%, lo cual no es despreciable. Los R^2 reportados en trabajos de esta área de estudio, generalmente son menores que los estimados aquí, a pesar de que son más exhaustivos respecto a las variables explicativas que incluyen.

No se aprecia algún comportamiento o patrón anómalo en los residuos de los cuatros modelos. Se observa que los residuos de los modelos **1** y **2** son los mismos porque el **2** sólo corrige los desvíos de los estimadores. También es de destacar que en los modelos **3** y **4** existe mayor concentración en los residuos al cuadrado contrastado con la variable *ren_pro* estimada por el modelo. Esto es una señal que el ajuste realizado por el modelo está acorde con lo esperado.

[21] Se debe tener en cuenta que las variables dummies en nuestro modelo son variables de control.

[22] Para el caso del modelo **3** el R^2 es obtenido mediante la diferencia entre 1 y el cociente entre el parámetro de escala y la varianza de *ren_pro*.

Gráfico 7: Residuos de los diferentes modelos



6- Conclusiones

El trabajo muestra las ventajas de introducir el enfoque GEE y Multinivel a una estructura de datos anidada. Los resultados indican que estos enfoques mejoran las estimaciones debido a que corrigen los errores estándares y los coeficientes de las variables de interés; aspectos que el método más sofisticado de Mínimos Cuadrados no logra capturar.

Se utilizaron diferentes modelos para estimar el rendimiento educativo y de esta manera visualizar la contribución de cada método, seleccionando como el más adecuado al modelo GEE, dada la estructura por conglomerados de los datos.

El modelo Mixed generalmente es utilizado para datos longitudinales, lo que en economía se denomina Panel Data. En este trabajo se presenta su adaptación a datos de corte transversal y sus resultados son casi tan buenos como los del GEE.

Nuestro análisis se basa en la teoría de la función de producción aplicada a la educación.

De los resultados obtenidos en el análisis de regresión, se resalta la importancia del nivel socioeconómico de los compañeros dentro del colegio en relación al nivel socioeconómico individual a la hora de explicar el rendimiento del alumno.

Sin ignorar las limitaciones de la información obtenida y el sesgo de selección que pudiera existir, consideramos que la metodología propuesta es una buena alternativa a tener en cuenta para investigaciones que contemplen datos multinivel.

Bibliografía

- Cervini, Rubén (2000). Desigualdades socioculturales en el aprendizaje de Matemática y Lengua de la educación secundaria en Argentina: un modelo de tres niveles; RELIEVE, v.8, n. 2, p. 135-158; http://www.uv.es/RELIEVE/v8n1/RELIEVEv8n2_1.htm.

- Coleman, J., Campbell, E., Hobson C., McPartland, J., Mood, A., Weinfeld, F. y York, R. (1966) "Equality of Educational Opportunity"; U.S. Department of Health, Education and Welfare, Office of Education.Washington: Government Printing Office.

- Delprato, Marcos (1999). "Determinantes del rendimiento educativo del nivel primario aplicando la Técnica de Análisis Multinivel"; Documento de trabajo 27, IERAL; Córdoba.

- Delprato, Marcos (2000). "Determinantes del rendimiento educativo y de la repitencia en la Capital Federal"; Documento de trabajo 28, IERAL; Córdoba.

- Goldstein, Harvey (1995). Multilevel Statistical Models, London, Edward Arnold: New York,

- Halstead Press. http://www.arnoldpublishers.com/support/goldstein.htm

- Hanushek, E. y Taylor L.(1990),"Alternative Assesments of the Performance of Schools. Measurement of State variations in Achievement", The Journal of Human Resources, Vol. 25, Nº2.

- Hanushek, E.(1995), "Interpreting recent research on schooling in developing countries" The World Bank Research Observer, vol 10, Nº2, 227-246.

- Harbison R. y Hanushek, E (1992), Educational Performance of the Poor: Lesson from Rural Northeast Brazil, World Bank Oxford University Press.

- Heinrich, Carolyn J.; Lynn, Laurence E., Jr. (1999). Means and ends: a comparative study of empirical methods for investigating governance and performance; The University of Chicago.

- Hox, J.J. (1994) Applied Multilevel Análisis. Amsterdan: TT-Publikaties. Available in electronic form at http://www.ioe.ac.uk/multilevel/amoboek.pdf.

- Jencks, C. et.al. (1972). "Inequality: a reassessment of the effects of family and schooling in America; New York, Basic.

- Krueger, A. (1997), "Experimental estimates of education production functions", NBER Working Paper Nº 6051.

- Laird, N. M. and Ware, J. H. (1982). Random Effects models for longitudinal data. Biometrics, 38,963-74.

- Liang, K-Y and Zeger, S. L., (1986). Longitudinal data analysis using generalized linear models. Biometrica, 73, 45-51.

- Llach, Juan José; Montoya, Silvia; Roldán, Flavio (1999). Educación para todos; IERAL.

- MCyE, DINIECE, Base de datos Operativo Nacional de Evaluación 2000; en Internet: www.diniece.org.ar.

- Mizala, A, Romaguera, P. y T. Reinaga (1999), "Factores que determinan el desempeño escolar en Bolivia", Documento de Trabajo N°61, Centro de Economía Aplicada, Depto. De Ingeniería Industrial, Universidad de Chile.

- Mizala, A., Romaguera, P.(2000), "Determinación de Factores Explicativos de los resultados escolares en Educación Media en Chile", Serie Economía N°85, Centro de Economía Aplicada, Dpto. de Ingeniería Industrial, Universidad de Chile.

- Snijders, Tom; Bosker, Roel (1999). Multilevel analysis: An introduction to basic advanced multilevel modeling; SAGE Publications.

- Summers, A. y Wolfe, B. " Do school made a difference?" (1977), American Economic Review.

- Twisk, Jos W. R. (2003). Applied longitudinal data análisis for epidemiology; Cambridge University Press.