Aplicación de la Teoría de Modelos Multinivel Lineal y No-Lineal utilizando el software especializado HLM7.

MSc. Welman Rosa Alvarado
FEDECRÉDITO, San Salvador
welman_16@hotmail.com

Resumen: Los modelos multinivel son básicamente un modelo de regresión de efectos mixtos, en donde se estudia una relación lineal entre dos o más variables en estudios realizados mediante un muestreo por agrupamiento, es decir, una técnica correlacional adecuada para analizar variaciones en las características de los sujetos que son miembros de un grupo que, a su vez, hace parte de otra agrupación, o sea, mediciones que forman una estructura agrupada y jerárquica. El modelo permite la descomposición de la variación de una variable criterio (como por ejemplo, rendimiento) en sus componentes "dentro del grupo" (dentro-escuela, dentro-departamentos) y "entre grupo" (entre-escuela, entre-departamento) y el análisis de la asociación entre variables en esos niveles de agregación.

Palabras clave: multinivel, regresión jerárquica, niveles de agregación.

Abstract: Multilevel models are basically a regression model mixed effect, where a linear relationship between two or more variables in studies is studied by sampling for clustering, that is, an appropriate correlation technique to analyze variations in the characteristics of the subjects that are members of a group which, in turn, is part of another group, that is, measurements and forming a nested hierarchical structure. The model allows the decomposition of the variation of a criterion variable (eg, yield) components "in-group" (within-school, within-departments) and "between group" (between-school, between-department) and analysis of the association between variables at these levels of aggregation.

Keywords*:* multilevel, hierarchical regression, levels of aggregation.

1. Teoría de modelos multinivel lineal.

1.1 Formulación del Modelo Multinivel Lineal.

Para entrar en materia, se presenta un ejemplo del estudio realizado por Goldstein (1999) a partir de los resultados obtenidos por alumnos en escuelas primarias (Junior School Proyect) en Londres, realizado por Mortimore et al (1988). Goldstein, utilizó una submuestra aleatoria

de la data de Mortimore, considerando 728 alumnos en 50 escuelas y como unidad de medida a los alumnos que están en cuarto año de aprendizaje, en el cual los alumnos cumplen sus ocho años de vida. Por otra parte, dentro de este estudio se utilizaron las puntuaciones de la prueba de matemática administrada en dos momentos junto con la información recogida del contexto social de los alumnos y de su género.

1.1.1 Regresión lineal simple

Para introducirnos a la teoría multinivel se hacen algunas conjeturaciones sobre qué tipo de relación seria de interés conocer a partir de la información de los gráficos.

En la figura 1.1 según el ejemplo tratado se muestra el diagrama de dispersión de las puntuaciones de la prueba de matemática en alumnos de 11 años de edad sobre las puntuaciones de la prueba de matemática en alumnos de 8 años de edad. En este diagrama no se hace ninguna distinción entre las escuelas a las cuales los alumnos pertenecen. Observamos que existe una dispersión estrecha de las puntuaciones de alumnos en edad de 11 años con el aumento de las puntuaciones de alumnos en edad de 8 años. Es importante recalcar que, al no haber distinción entre escuelas, no podemos ver si la escuela influye sobre las puntuaciones de los alumnos.

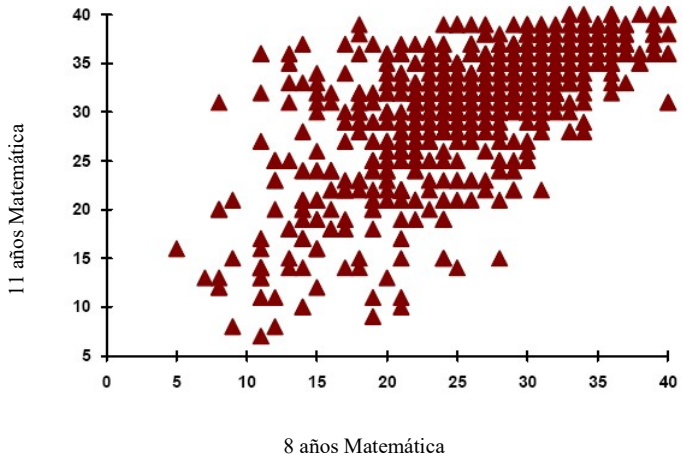


Figura 1.1: Diagrama de dispersión de las puntuaciones de la prueba de las matemáticas en alumnos de 11 años de edad sobre las puntuaciones de la prueba de matemática en alumnos de 8 años de edad.

En figura 1.2 se presenta para un caso particular de dos escuelas que han sido seleccionadas aleatoriamente, representadas por diversos símbolos. Observamos que conforme aumentan las puntuaciones de los alumnos de 8 años, las puntuaciones de alumnos de 11 años están entre 20 y 30 para la escuela 1 (símbolo círculo). Ahora bien, para esa misma escuela la puntuación de mayor edad se sobrepone a las puntuaciones de alumnos con menor edad. Sin embargo, si trazamos dos líneas de regresión para dichas escuelas, se tiene que las rectas no son paralelas, indicando que la escuela 2 (símbolo triángulo) tiene mejores puntuaciones en la prueba que la escuela 1. Además, hay un punto de intersección en las dos rectas o un balance de las puntuaciones obtenidas en alumnos de 8 y 11 años. Pero que el cambio surge después de ese punto de intersección, se observa que no solo la edad o variable explicativa del nivel alumno influye en su puntuación, sino que podemos pensar que existen otras características de la escuela, de tal modo que las características de la escuela estarían influyendo en las puntuaciones de los alumnos.

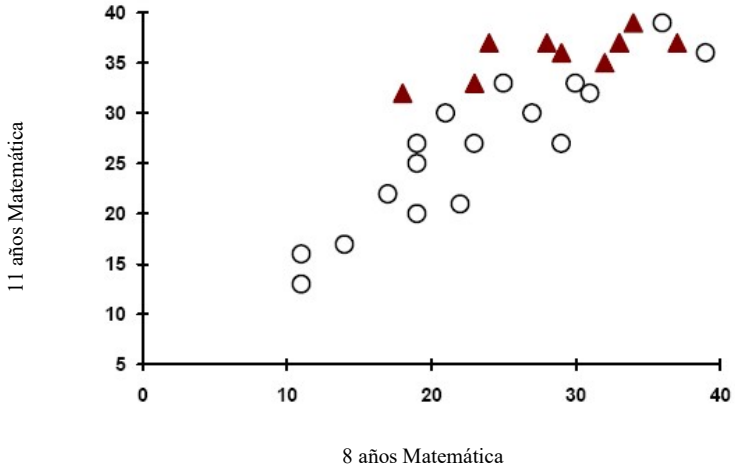


Figura 1.2: Diagrama de dispersión de las puntuaciones de la prueba de matemática para dos escuelas.

De las conjeturaciones hechas anteriormente, se escribe un modelo de regresión simple para una escuela, relacionando las puntuaciones de la prueba de matemática en alumnos de 11 años con las puntuaciones de 8 años, de la siguiente manera:

$$y_i = \alpha + \beta x_i + e_i \qquad (1.1)$$

Donde y_i es la puntuación del i-ésimo alumno, x_i la edad del alumno, $\beta = (\beta_1, \beta_2, ..., \beta_n)'$ es el efecto que tiene la edad sobre la puntuación del i-ésimo alumno y $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)'$ es el promedio de las puntuaciones eliminando el efecto de la edad de los alumnos.

Pero en la ecuación 1.1 solamente tenemos una regresión que permite conocer el efecto de la edad sobre la puntuación del estudiante en una escuela en particular con una muestra de tamaño k_1 estudiantes. Ahora bien, si queremos conocer el efecto de la edad sobre la puntuación de los estudiantes de más de una escuela, con muestras de tamaño k_j en cada una, entonces tendríamos n modelos de regresión lineal, de tal modo que podamos conocer que tanto influye la edad sobre la puntuación del i-ésimo alumno en la j-ésima escuela. Esto implica el ajuste de n modelos de regresión, mediante una forma parcial para cada escuela con tamaño k_j, tenemos

Para la escuela 1: $y_{i1} = \alpha_1 + \beta_1 x_{i1} + e_{i1}$; $i = 1, 2, ..., k_1$

Para la escuela 2: $y_{i2} = \alpha_2 + \beta_2 x_{i2} + e_{i2}$; $i = 1, 2, ..., k_2$

$\vdots$

Para la escuela n: $y_{in} = \alpha_n + \beta_n x_{in} + e_{in}$; $i = 1, 2, ..., k_n$

El sistema de ecuaciones de los n modelos anteriores se puede simplificar con el modelo siguiente:

$$y_{ij} = \alpha_j + \beta_j x_{ij} + e_{ij} \qquad i = 1, 2, ..., k_j \; ; \quad j = 1, 2, ..., n \qquad (1.2)$$

Donde

y_{ij} es la puntuación del i-ésimo alumno en la j-ésima escuela.

x_{ij} es la edad sobre la puntuación del i-ésimo alumno en la j-ésima escuela.

α representa el promedio de la puntuación muestral.

β_j representa los pesos asociados a la característica x_{ij} en la muestra completa.

e_{ij} es una variable aleatoria que representa el error de ajuste del modelo del i-ésimo alumno en la j-ésima escuela.

Los e_{ij} deben cumplir los siguientes supuestos:

a) La perturbación tiene esperanza nula, es decir:

$$E\left(e_{ij}\right)=0$$

b) La varianza de la perturbación es siempre constante, y no depende de x; conocido como homocedasticidad de la perturbación.

$$\text{var}\left(e_{ij}\right)=\sigma_{eo}^{2}$$

c) La perturbación tiene una distribución normal. Este supuesto es consecuencia del teorema central del límite.

d) Las perturbaciones son independientes entre sí.

Estas cuatro condiciones pueden expresarse igualmente respecto a la variable respuesta, o dependiente, como sigue:

e) La esperanza de la respuesta depende linealmente de x. Tomando esperanzas en (1.2), se tiene:

$$\begin{aligned} E\left(y_{ij}\right) &= E\left(\alpha_j+\beta_j x_{ij}+e_{ij}\right) \\ &= \alpha_j+\beta_j x_{ij} \end{aligned}$$

f) La varianza de la distribución de y_{ij} es constante.

$$\text{var}\left(y_{ij}\right)=\sigma_{eo}^{2}$$

g) La distribución de y para cada x es normal.

h) Las observaciones y_{ij} son independientes entre sí.

Ahora, si utilizamos el modelo de regresión (1.2) se tendrían que estimar α , β y σ_{eo}^{2} que representa 2n+1 parámetros, suponiendo que α_{j} , β_{j} y σ_{eo}^{2} es fijo para cada escuela $j=1,2,...,n$

1.1.2 El modelo de dos niveles.

Con el fin de conocer otras variables aleatorias que no han sido medidas en el alumno se debe considerar los parámetros como variables aleatorias. Es por ello que para hacer la ecuación (1.2) más auténtica de dos niveles, dejamos α_{j} y β_{j} como variables aleatorias convertidas. Para la consistencia de la notación sustituiremos α_{j} por β_{0j} y β_{j} por β_{1j} y asumiremos que

$$\beta_{0j}=\beta_{0}+u_{0j},\qquad \beta_{1j}=\beta_{1}+u_{1j},\quad j=1,2,...,n \tag{1.3}$$

Donde

β_{0j} es el logro promedio por escuela, está representado como una función de la gran media β_{0} o media de todas las escuelas, más una variable aleatoria u_{0j} que captura la variación en la puntuación promedia a través de escuelas. Dicho de otra forma, el u_{0j} contiene las variables no observables que hacen que la nota promedio de cada escuela no sean iguales. Por ejemplo, si todas las escuelas tuviesen media constante, entonces $u_{0j}=0$.

β_{1j} es el efecto de la variable edad, está representado como una función de la estimación de la media de las pendientes relativas al efecto de la variable edad más una variable aleatoria u_{1j} que captura la variación de los pesos asociados a la característica edad a través de escuelas. La variable u_{1j} recoge todas las variables no observables en la escuela que influyen en la edad de los estudiantes y hace que el peso de la variable edad difiera entre escuelas. Si el peso de la variable edad de los estudiantes fuese el mismo en todas las escuelas, entonces $u_{1j}=0$.

Sustituyendo (1.3) en (1.2) tenemos que,

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \left(u_{0j} + u_{1j} x_{ij} + e_{0ij}\right) \tag{1.4}$$

Donde y_{ij} se ha expresado como la suma de una parte fija del modelo $\beta_0 + \beta_1 x_{ij}$ y una parte aleatoria $u_{0j} + u_{1j} x_{ij} + e_{0ij}$ dentro de las escuelas. Que al final la expresión (1.4) resulta ser un modelo lineal de efectos mixtos.

Para las perturbaciones se establecen los siguientes supuestos:

a) Las variables aleatorias u_{0j} y u_{1j} tiene esperanza nula, es decir:

$$E\left(u_{0j}\right) = E\left(u_{1j}\right) = 0 \tag{1.5}$$

b) La varianza de cada variable aleatoria u_{0j} y u_{1j} es siempre constante, y no depende de x.

$$\text{var}\left(u_{0j}\right) = \sigma^2_{u0}, \quad \text{var}\left(u_{1j}\right) = \sigma^2_{u1}, \quad \text{cov}\left(u_{0j}, u_{1j}\right) = \sigma_{u01} \tag{1.6}$$

c) Las perturbaciones son independientes entre si.

d) Para la variable respuesta o dependiente, se tiene que:

e) La esperanza de la respuesta depende linealmente de x. Tomando esperanzas en (1.4), se tiene que:

$$E\left(y_{ij}\middle|\beta_0,\beta_1,x_{ij}\right)=E\left(\beta_0+\beta_1x_{ij}+\left(u_{0j}+u_{1j}x_{ij}+e_{0ij}\right)\right)$$
$$=E\left(\beta_0\right)+E\left(\beta_1x_{ij}\right)+E\left(u_{0j}\right)+E\left(u_{1j}x_{ij}\right)+E\left(e_{0ij}\right)$$

f) Y según el supuesto (e) la esperanza matemática de la variable respuesta resulta ser,

$$\hat{y}_j\cong E\left(y_{ij}\middle|\beta_0,\beta_1,x_{ij}\right)=\beta_0+\beta_1x_{ij} \tag{1.7}$$

g) La distribución de y para cada x es normal.

h) La varianza de las perturbaciones del i-ésimo alumno en la j-ésima escuela e_{0ij} es constante.

$$\text{var}\left(e_{0ij}\right)=\sigma_{eo}^2 \tag{1.8}$$

Hay que observar que la expresión (1.3) no existe una medición de una variable en la j-ésima escuela. Solamente hemos considerado el caso simple de variables del alumno (nivel 1).

1.2 Asunción del modelo multinivel.

Snijjders y Brosker plantean las siguientes preguntas, que nos ayudan a comprobar supuestos:

a) La parte del modelo ¿contiene las variables adecuadas?
b) La parte aleatoria del modelo ¿contiene las variables adecuadas?
c) Los residuos del primer nivel, ¿están normalmente distribuidos?
d) Los coeficientes aleatorios del segundo nivel, ¿están normalmente distribuidos?
e) Los coeficientes aleatorios del segundo nivel, ¿tienen una matriz de varianzas-covarianzas constante?

1.3 Significación y ajuste de los modelos.

Como en cualquier otro modelo de regresión, la interpretación de los modelos multinivel depende de:

- La significación de los coeficientes de regresión.
- Como de bien el modelo ajusta los datos.

La teoría estadística que hay detrás del modelo de regresión multinivel es complicada. A partir de los datos observados, se quieren estimar los parámetros del modelo multinivel:

- Los coeficientes de regresión.
- Los componentes de la varianza.

Los estimadores más utilizados en el análisis de regresión multinivel son los estimadores de máxima verosimilitud. El objetivo de la estimación de máxima verosimilitud es encontrar un estimador del parámetro, dependiendo de los datos conocidos, más cercano al verdadero valor del parámetro. Es decir, dado un conjunto de datos y el modelo probabilístico subyacente, la estimación de máxima verosimilitud, toma el valor del parámetro que da lugar a la distribución con la que los datos son más probables.

Los estimadores de ML, nos van ayudar a contestar a las siguientes preguntas:

a) ¿Este predictor es estadísticamente significativo?

El procedimiento de máxima verosimilitud, produce, errores estándar, para la mayoría de las estimaciones. La significación de un predictor viene dada por la ratio entre el estimador del parámetro y su error típico. Este test es conocido como el Test de Wald. Esta distribución del estadístico de Wald sirve para aceptar o rechazar la hipótesis nula establecida sobre el estimador del parámetro β (H_0: $\beta = 0$).

Se verifica que:

$$\frac{\hat{\beta}}{\hat{s}_\beta} \sim N(0,1) \text{ o lo que es equivalente } \left(\frac{\hat{\beta}}{\hat{s}_\beta}\right)^2 \sim x^2{}_{1g.l.}$$

En la práctica, una regla general para determinar la significación de un predictor es que si,

$$z = \frac{parámetro}{error\ estandar} > 2 \rightarrow p < 0.05$$

b) ¿Aporta este modelo (con x predictores) información significativa comparado con el modelo nulo o con otro modelo alternativo?

Los procedimientos de máxima verosimilitud también producen un estadístico llamado ***Deviance***. Este estadístico indica como de bien ajusta el modelo a los datos. Si llamamos L_1, al valor del máximo de la función de verosimilitud (likelihood) en la estimación de los parámetros del modelo 1, entonces se define la deviance:

$$Dev = -2ln(L_1) \equiv -2log(likelihood)$$

Si en el modelo nulo no hay varianza estadísticamente distinta de 0 en los niveles contemplados, ningún modelo que se derive de este añadiendo variables explicativas, mejorará el ajuste, ya que, la varianza del intercepto no es significativamente distinta de 0. Si dos modelos están anidados, es decir, un modelo se obtiene a partir de otro más general eliminando parámetros de este último, entonces podemos compararlos. Para llevar a cabo la comparación hacemos uso de la deviance de cada modelo, ya que, la diferencia de las deviances se distribuye como una chi-cudrado con los grados de libertad iguales a la diferencia del número de parámetros estimados en los modelos que estamos comparando, bajo la hipótesis nula de que ambos modelos son iguales.

Ejemplo: supongamos que tenemos dos modelos M1 con m1 parámetros y M2 con m2 parámetros.

$$D_1 = -2ln(L_1) \;\; y \;\; D_2 = -2ln(L_2)$$

$$D = -2ln\left(\frac{L_2}{L_1}\right) \sim x^2{}_{m2-m1}$$

Si la diferencia es significativa, nos quedamos con el M2, y sino con M1, es decir, los parámetros que aparecen en el modelo 2 y no en el modelo 1 son significativamente distintos de 0 y, por tanto, las correspondientes variables de ajuste estarán asociadas de forma significativa con las variables respuesta.

c) ¿Cómo comparamos modelos no anidados?

Si los modelos que queremos comparar no están anidados, el principio de parsimonia nos indica que deberíamos escoger el modelo más simple. Pero también podemos usar el ***Criterio de información de Akaike***, conocido como AIC. Para un modelo de regresión multinivel de AIC se calcula a partir del valor de la deviance (Dev), y el número de parámetros estimados (q):

$$AIC = Dev + 2q$$

El AIC, es un índice de ajuste general, que asume que los modelos que se están comparando ajustan el mismo conjunto de datos, y usan un mismo método de ajuste. Un criterio de ajuste similar es el Criterio de información bayesiana de Schwarz, conocido como BIC y que viene dado por:

$$BIC = Dev + q.Ln(N)$$

Al igual que la deviance, cuanto menor es el valor de AIC y del BIC mejor es el ajuste. Tanto el AIC como el BIC, penalizan a los modelos con un elevado número de parámetros, pero el BIC impone una mayor penalización para la mayoría de tamaños muestrales, por ello, para los modelos multinivel con diferentes tamaños muestrales y varios niveles, y, por lo tanto, el criterio AIC es más recomendable que el BIC.

1.4 Métodos de estimación de los parámetros.

Hay dos tipos de parámetros: fijos y aleatorios. Los parámetros fijos corresponden a los efectos medios en la población, y son las pendientes y el intercepto. Los aleatorios corresponden a las varianzas y covarianzas de todos los niveles.

A la hora de estimar dichos parámetros, debemos distinguir entre métodos y algoritmos de estimación. Un método de estimación consiste en un conjunto de reglas y principios cuya aplicación da lugar a una ecuación o ecuaciones que ponen en relación los datos con el parámetro buscado. Los algoritmos son métodos matemáticos, que, por medio de iteraciones sucesivas, permiten obtener soluciones para dichas ecuaciones.

Métodos de Estimación		**Algoritmos de Estimación**	
Máxima Verosimilitud (ML)*	Tanto los coeficientes de regresión como los componentes de la varianza se incluyen en la función de verosimilitud.	Fisher Scoring	Se basa en la mejor aproximación de la función de verosimilitud y se puede usar con ML y REML.
Máxima verosimilitud restringida (REML)	Solo los componentes de la varianza se incluyen en la función de verosimilitud	Expectation-Maximization (EM)	Calcula estimadores máximo-verosímiles en casos en los que existen datos perdidos. Pero es muy lento. Se usa para REML.
		Iterative Generalized Least Squares (IGLS)	Refinamiento secuencial del procedimiento basado en

			Mínimos cuadrados. Produce estimadores sesgados de los parámetros aleatorios debido a que no tiene en cuenta la varianza muestral de la parte fija del modelo.
		Restricted IGLS (RIGLS)	Para estimar los parámetros usa el método REML, y se pueden conseguir estimadores insesgados. Aconsejable para muestras pequeñas.
Máxima verosimilitud (ML)*	Para modelos no lineales con variables Dicotómicas: utilizan expansiones de la serie de Taylos para linealizar el modelo multinivel.	Cuasi-verosimilitud Marginal (MQL)	Linealiza la parte fija del valor predecido de las variables dependientes.
Estimación Bayesiana	Combina la distribución a priori (integra conocimientos previos de los parámetros) con la función de verosimilitud, y produce una distribución a posteriori, que describe la incertidumbre de los parámetros después de observar los datos.	Método Monte carlo por Cadena de Markov	MCMC, Métodos de Monte Carlo por Cadenas de Markov. Simulación de cadenas de Markov convergentes hacia la distribución a posteriori de los parámetros.

1.5 Varianza.

En la regresión múltiple, la varianza explicada mide la proporción de la varianza total de la variable respuesta (Y) que es explicada por la relación lineal que existe entre Y y las variables explicativas del modelo (Xi). Para medir dicha proporción usamos el llamado coeficiente de determinación, más conocido por R^2, que toma valores entre 0 y 1. Un valor próximo a 1 se interpreta como un buen ajuste del modelo.

Pero ¿Cómo se mide en una regresión multinivel, la varianza explicada por el modelo?

Los modelos multinivel permiten dividir la varianza total en diferentes componentes de variación según los distintos niveles de agrupación de datos. Por ejemplo, en la investigación del asma en la infancia podemos considerar el estudio de los factores de riesgo de la proporción de asmáticos entre y variación dentro de las unidades del nivel superior (ciudad de residencia, por ejemplo).

a) Método Propuesto por Golstein.

Podemos resumir la importancia del segundo nivel (por ejemplo, si fuera hasta dos niveles), como la proporción de la varianza total explicada, que se conoce como "coeficiente de partición o división de la varianza" VPC y viene dado por la fórmula:

$$VCP = \frac{varianza\ residual\ del\ segundo\ nivel}{varianza\ residual\ del\ primer\ nivel + varianza\ residual\ del\ segudno\ nivel}$$

$$VCP = \frac{\sigma_{u_0}^2}{\sigma_{u_0}^2 + \sigma_e^2}$$

El VPC es útil en el caso que tengamos un modelo con una única fuente de variación en cada nivel, pero lo es menor, en un modelo de coeficientes aleatorios.

En el caso de un modelo de **intercepto aleatorio** el VPC, por ejemplo, mide la correlación residual entre las respuestas de dos niños en la misma ciudad de residencia, y en este caso es conocido también como "Coeficiente de Correlación Intraclase, ρ", y se usa comúnmente como una medida de la importancia de considerar que los datos están agrupados o estructurados jerárquicamente.

El VPC es útil en el caso que tengamos un modelo con una única fuente de variación en cada nivel, pero lo es menos, en un **modelo de coeficientes aleatorios**.

b) Método propuesto por Snijders.

Snijders y Bosker proponen otra definición de la proporción de varianza explicada, y la llaman ***Reducción de la proporción de la predicción***.

En el marco de los modelos multinivel de 2 niveles, se puede elegir entre predecir el valor de Y para un individuo en un grupo, o predecir el valor medio de Y para un grupo. Lo que da lugar a dos formas de medir la proporción de varianza explicada:

i) Reducción de la proporción del error de predicción de un valor residual.

Cuando desconocemos el valor de x para un individuo, la mejor predicción es E(Yi) y la varianza del error de predicción es $var(Y_i)$. En cambio, cuando conocemos el valor de x la mejor predicción es:

$$\sum_{h=0}^{q} \beta_h x_{hij} \text{ y el error de predicción es } Y_{ij} - \sum_{h=0}^{q} \beta_h x_{hij} = u_{oj} + e_{ij}$$

Por tanto, la varianza del error de predicción es:

$$var\left(Y_{ij} - \sum_{h=0}^{q} \beta_h x_{hij}\right) = var(u_{oj} + e_{ij}) = \sigma_{u_0}^2 + \sigma_e^2$$

Y la reducción de la proporción de la varianza del error de predicción para el nivel 1 será:

$$R_1^2 = 1 - \frac{var\left(Y_{ij} - \sum_{h=0}^{q} \beta_h x_{hij}\right)}{var(Y_i)} = 1 - \frac{\sigma_{u_0}^2 + \sigma_e^2}{var(Y_i)}$$

Lo normal es desconocer el valor de estos parámetros, por ello la mejor forma de estimar R_1^2 es usar las estimaciones de dichos parámetros.

$$\hat{R}_1^2 = 1 - \frac{var\left(Y_{ij} - \sum_{h=0}^{q} \beta_h x_{hij}\right)}{var(Y_i)} = 1 - \frac{\sigma_{u_0}^2 + \sigma_e^2}{var(Y_i)}$$

Lo normal es desconocer el valor de estos parámetros, por ello la mejor forma para estimar R_1^2 es usar las estimaciones de dichos parámetros.

$$\hat{R}_1^2 = 1 - \frac{\left(\hat{\sigma}_{u_0}^2 + \hat{\sigma}_e^2\right)_N}{\left(\hat{\sigma}_{u_0}^2 + \hat{\sigma}_e^2\right)_A}$$

Donde,

N=varianza del modelo nulo.

A=varianza del modelo alternativo (con al menos un predictor)

ii) Reducción de la proporción del error de predicción de la medida de grupo.

La proporción de varianza explicada en el nivel 2 se puede definir como, la reducción en la proporción de la varianza del error de predicción de la media $\bar{Y}_{.j}$, de una unidad de nivel 2 elegida al azar.

Si conocemos los valores de todos los predictores x_{nij} para todos los i del grupo j, entonces la mejor predicción de $\bar{Y}_{.j}$ es el valor de la regresión $\sum_{h=0}^{q} \beta_h x_{hij}$ y la varianza del error de la predicción:

$$var\left(\bar{Y}_{.j} - \sum_{h=0}^{q} \beta_h x_{h.j}\right) = \hat{\sigma}_{u_0}^2 + \frac{\hat{\sigma}_e^2}{n_j}$$

Donde,

n_j es el número de unidades del nivel 1 en el grupo j.

Entonces a partir de estos datos definimos la reducción en la proporción de la varianza del error de predicción de $\bar{Y}_{.j}$ como:

$$R_2^2 = 1 - \frac{var\left(\bar{Y}_{.j} - \sum_{h=0}^{q} \beta_h x_{h.j}\right)}{var\left(\bar{Y}_{.j}\right)}$$

Y su estimación:

$$\hat{R}_2^2 = 1 - \frac{\left(\hat{\sigma}_{u_0}^2 + \frac{\hat{\sigma}_e^2}{n}\right)_N}{\left(\hat{\sigma}_{u_0}^2 + \frac{\hat{\sigma}_e^2}{n}\right)_A}$$

Donde,

N= varianza del modelo nulo.

A= varianza del modelo alternativo (con al menos un predictor).

La cantidad de varianza explicada en un segundo nivel es un único valor. Pero es posible que cada grupo j tenga un n_j distinto, entonces ¿Qué valor se debe usar? Se puede usar cualquier valor que sea considerado a priori, representativo de las unidades de nivel 2. Si los valores de n_j varían mucho en la población, se puede usar la media armónica $\frac{N}{\sum_J \frac{1}{n_j}}$

Respecto a R_1^2 y R_2^2, sus valores poblacionales no pueden ser menores de cero. En cambio, sus estimaciones pueden aumentar su valor al eliminar un predictor o disminuir al incluir un nuevo predictor, esto puede ser debido al azar o por una mala especificación de parte fija del modelo.

Estos cambios en los valores de $\hat{R}_1^2$ y $\hat{R}_2^2$ en una dirección equivocada sirven de diagnóstico para el investigador, para detectar posibles errores de especificación de la parte fija.

2. Muestreo

El muestreo es una herramienta que se usa en la investigación científica. Todo estudio ya sea observacional o experimental, lleva implícito en la fase de diseño la determinación del tamaño muestral necesario para la ejecución del mismo.

Existe bastante literatura acerca del cálculo del tamaño muestral en estudios multinivel. En los estudios más sencillos, de dos niveles, se deben estimar dos tamaños muestrales distintos:

- El tamaño de la muestra de las unidades del primer nivel (nj)
- El tamaño de la muestra de las unidades del segundo nivel (J)

El tamaño total de la muestra viene dado por: $\sum_{j=1}^{J} n_j$

En los estudios multinivel, generalmente el principal problema es determinar el tamaño muestral de las unidades del nivel grupo o segundo nivel, ya que este suele ser más pequeño que el tamaño de muestra del nivel individual.

En general, para calcular el número de unidades, individuos o pacientes necesarios en un estudio multinivel, lo primero es calcular un tamaño muestral para un muestreo aleatorio simple. Si queremos comparar las medias del grupo intervención y del grupo control podemos usar la siguiente fórmula:

$$N_1 = \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 * \sigma^2 * (r+1)}{d^2 * r}$$

Donde:

N_1: sujetos necesarios en el grupo intervención.

Z_t: Valor Z de la distribución normal que deja a su izquierda un área de t / $t = 1 - \alpha/2, 1 - \beta$, siendo α la probabilidad de cometer el error de tipo I y β la probabilidad de cometer el error de tipo II.

σ^2: Varianza de la variable respuesta.

d: Valor mínimo de la diferencia en media que, si existiera, se desea detectar con una probabilidad $1 - \beta$.

$r = \frac{N_0}{N_1}$: Razón del número de sujetos entre los comparados. (N_0 tamaño de la muestra del grupo control)

Si la variable respuesta es dicotómica, se puede usar la siguiente ecuación:

$$N_1 = \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 * \bar{p}(1-\bar{p}) * (r+1)}{(p_1 - p_0)^2 * r}$$

$$\bar{p} = \frac{p_1 + (r * p_0)}{1 + r}$$

Donde:

N_1: sujetos necesarios en el grupo intervención.

Z_t: Valor Z de la distribución normal que deja a su izquierda un área de t / $t = 1 - {}^{\alpha}/_{2}, 1 - \beta$, siendo α la probabilidad de cometer el error de tipo I y β la probabilidad de cometer el error de tipo II.

$\bar{p}$: media ponderada de p_0 y p_1

$r = \frac{N_0}{N_1}$: Razón del número de sujetos entre los comparados. (N_0 tamaño de la muestra del grupo control)

p_1: proporción de casos en el grupo intervención.

p_0: proporción de casos en el grupo control.

Después de calcular el tamaño muestral aleatorio simple es necesario multiplicar por un factor de corrección el cual tiene en cuenta el muestreo en dos etapas. En un muestreo aleatorio simple el error estándar de la media viene dado por la siguiente fórmula:

$$error\ est \coloneqq \frac{Desviación\ estandar}{\sqrt{tamaño\ muestral}}$$

Supongamos que tenemos N macro-unidades cuyo tamaño es n, entonces el tamaño total de la muestra es Nn. El efecto del diseño es el factor de corrección y es un número que indica cuando debemos ajustar el denominador de la formula anterior para tener en cuenta el cambio en el diseño (pasar de una etapa a dos); se define como el cociente entre la varianza obtenida con el nuevo diseño muestral y la varianza obtenida con el muestreo aleatorio simple para la misma población.

El efecto del diseño para un muestreo de dos etapas con igualdad de tamaño en las macro-unidades o unidades de segundo nivel es:

$$Efecto\ del\ diseño = 1 + (n - 1)\rho$$

Donde ρ es coeficiente de correlación intraclase (CCI).

Hay autores que proponen distintas reglas de oro, estas reglas son a menudo opiniones personales basadas en la experiencia:

Si se está interesado en los efectos fijos del modelo, 10 grupos en el segundo nivel serán suficientes: Si el interés está en los efectos contextuales como mínimo serán necesarios 30 grupos.

La regla del 30/30, los investigadores deben esforzarse para obtener una muestra de al menos 30 grupos con 30 individuos por grupo como mínimo.

3. Modelos multinivel no lineal: logistico.

Aunque los modelos multinivel fueron desarrollados originalmente para variables de respuesta con distribución normal y bajo los supuestos de una distribución normal de los errores en cada individuo, estos métodos han sido generalizados para situaciones en las que la variable respuesta es binomial, nominal y ordinal y para procesos donde la probabilidad del evento es pequeña y se puede modelas con una distribución de Poisson. Casos particulares son los siguientes:

3.1 Modelo Nulo

Se usa cuando nuestra variable dependiente toma dos valores. Es una extensión de los modelos multinomiales estándar. Siendo el modelo multinivel más simple, para una estructura jerárquica de 2 niveles, con una variable independiente, el intercepto aleatorio y link logit, es el siguiente:

$$log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 x_{ij} + u_{0j}$$

$$u_{0j} \sim N\left(0, \sigma_{u_0}^2\right) \; y \;\; \pi_{ij} = P\left(y_{ij} = 1\right)$$

El objetivo principal de una regresión logística es predecir la probabilidad π_i de que ocurra un evento Y, un individuo i, en función de un determinado número de variables.

Un modelo general para una variable respuesta dicotómica (Y_i) y una variable explicativa x_i es:

$$f\left(\pi_{ij}\right) = \beta_0 + \beta_1 x_{ij} \; / \; P\left(y_{ij} = 1\right) = 1$$

Siendo $f\left(\pi_{ij}\right)$ una transformación de π_i llamada link.

Las funciones link más conocidas son:

Link logit, donde $f(\pi_i) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right)$

Link probit, donde $f(\pi_i) = \varphi(\pi_i)$ es la función densidad acumulada de la distribución normal.

Link log-log, donde $f(\pi_i) = log\left(-log\left(1 - \pi_{ij}\right)\right)$

Si queremos extender nuestro modelo para tener en cuenta la estructura jerárquica de 2 niveles, iniciamos ajustando el modelo nulo con intercepto aleatorio. Nuestra variable respuesta es y_{ij} toma el valor 1 individuo i tiene la respuesta Y el grupo j y 0 si no.

Si usamos el link logit, nuestro

3.2 Modelo Nulo de 2 niveles.

Modelo nulo de dos niveles para una variable respuesta dicotómica queda de la siguiente manera:

$$logit(\pi_i) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_{0j}$$

$$\beta_{0j} = \beta_0 + u_{0j} \quad /u_{0j} \sim N\left(0, \sigma^2_{u_0}\right)$$

3.3 Modelo de Intercepto Aleatorio

Si queremos incluir en el modelo potenciales variables explicativas de la variable respuesta $X_1, \ldots, X_p$, obtenemos el siguiente modelo:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \sum_{P=1}^{p} \beta_p x_{(p)ij} + u_{0j}$$

$$/u_{0j} \sim N\left(0, \sigma^2_{u_0}\right)$$

El intercepto β_{0j} está formado por dos componentes: un efecto fijo β_0, igual para todos los grupos, y un efecto aleatorio u_{0j} específico para cada grupo (unidad de segundo nivel) j.

En el caso más simple en que tan sólo hay una variable explicativa la formulación del modelo es:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 x_{(1)ij} + u_{0j}$$

$$/u_{0j} \sim N\left(0, \sigma^2_{u_0}\right)$$

$$\hat{\pi}_{ij} = \frac{\exp\left(\beta_0 + \beta_1 x_{(1)ij} + u_{0j}\right)}{1 + \exp\left(\beta_0 + \beta_1 x_{(1)ij} + u_{0j}\right)}$$

β_0 se interpreta como el intercepto de conjunto en la relación entre el log-odds y x. El intercepto para una determinada unidad del nivel 2 j es $\beta_0 + u_{0j}$ que será mayor o menor que el intercepto de conjunto dependiendo de si u_{0j} es mayor o menor de cero. Como en el caso

de modelo de respuesta continua, u_{0j} es conocido como el efecto aleatorio de grupo, el residuo de grupo o el residuo de segundo nivel. La varianza del intercepto entre grupos es $var(u_{0j}) = \sigma^2_{u_0}$, se conoce como la varianza residual entre grupos o varianza residual de segundo nivel. Y en el modelo con variables explicativas siempre la varianza no explicada de nivel 2. Las varianzas por definición son no negativas, por ellos cuando realizamos la prueba de la hipótesis nula $H_0: \sigma^2_{u_0} = 0$ la hipótesis alternativa debe ser unilateral $H_0: \sigma^2_{u_0} > 0$, por lo tanto, la probabilidad de que la estadística Z sea mayor o igual que una variable chi-cuadrado con tantos grados de libertad como parámetro haya en el modelo, hay que dividirla por 2.

El modelo multinivel para respuestas binarias se puede derivar también a través de una variable latente de contextualización. Asumimos que existe una variable continua $y*_{ij}$ subyacente a y_{ij} y así podemos formular el llamado modelo umbral, que permite la representación:

$$y_{ij} = \begin{cases} 1 & si \;\; y*_{ij} \geq 0 \\ 0 & si \;\; y*_{ij} < 0 \end{cases}$$

Teniendo en cuenta esta representación podemos escribir el siguiente modelo de 2 niveles de intercepto aleatorio para la variable inobservada $y*_{ij} = \beta_0 + \beta_1 x_{ij} + (u_{0j} + \varepsilon*_{ij})$

Para que represente un modelo de regresión logística, los residuos de primer nivel de la variable subyacente $y*$, deben tener una distribución logística. Lo que significa que:

- $P(\varepsilon*_{ij} < x) = logistic\,(x) \quad \forall\, x$
- La media de los residuos del primer nivel es 0
- La varianza es $\frac{\pi^2}{3} = 3.29$

Cuando se asume que $\varepsilon*_{ij}$ tiene esta distribución, el modelo logístico visto en el apartado anterior, es equivalente al modelo umbral definido aquí.

4. Aplicación de modelo multinivel lineal y logistico - utilizando el software hlm7.

4.1 Descripción Software HLM

El software HLM fue creado por Anthony Bryk en 1992, se trata de un programa diseñado específicamente para el desarrollo de Modelos Multinivel este caso en Estados Unidos. Fue a través del texto "Hierarchical Linear Models for Social and Behavioral Reserch: Applications and Data Analysis methods" (Bryk y Raudenbush, 1992) que se ha convertido en uno de los softwares más utilizados. HLM se distribuye a través de SSI-Scientific Software International – (www.ssicentreal.com).

El HLM ofrece al usuario una amplia gama de opciones de estimación, la imputación de ficheros desde diferentes softwares (SPSSS, Stata…), diferentes pruebas de hipótesis de razón de verosimilitud, la creación de gráficos, y la capacidad de manejar fácilmente modelos lineales jerárquicos a través del visor de operaciones en las que ofrece las ecuaciones de cada nivel, o su visor de modelo mixto donde se integran las ecuaciones de cada nivel en una única ecuación.

Las últimas versiones de HLM marcan claramente la diferencia frente a los paquetes estadísticos generales siendo mucho más intuitivas y cuidando el diseño la interfaz de trabajo. El HLM 7, la versión más reciente del programa puesta a la venta en marzo de 2013, incluye novedades como el cálculo de modelos de hasta cuatro niveles de análisis (anteriores versiones sólo dejaban calcular 2 y 3 niveles), o la imputación de los datos desde un fichero de datos (versiones anteriores requerían crear tantos ficheros de datos como niveles tuviera su estudio).

Para hacerse con el HLM el usuario tan sólo tendrá que acceder a la página de la distribuidora y solicitarlo vía e-mail. La buena noticia es que para todos aquellos que ya dispongan de una versión previa del software tendrán que pagar un precio considerablemente menor por hacerse con la última versión del mismo. La página ofrece precios diferentes para los distintos tipos de licencia (430$ desde la web de HLM) e incluso ofrece una versión estudiante de forma totalmente gratuita (http://www.ssicentral.com/hlm/student.html). Sin embargo, la

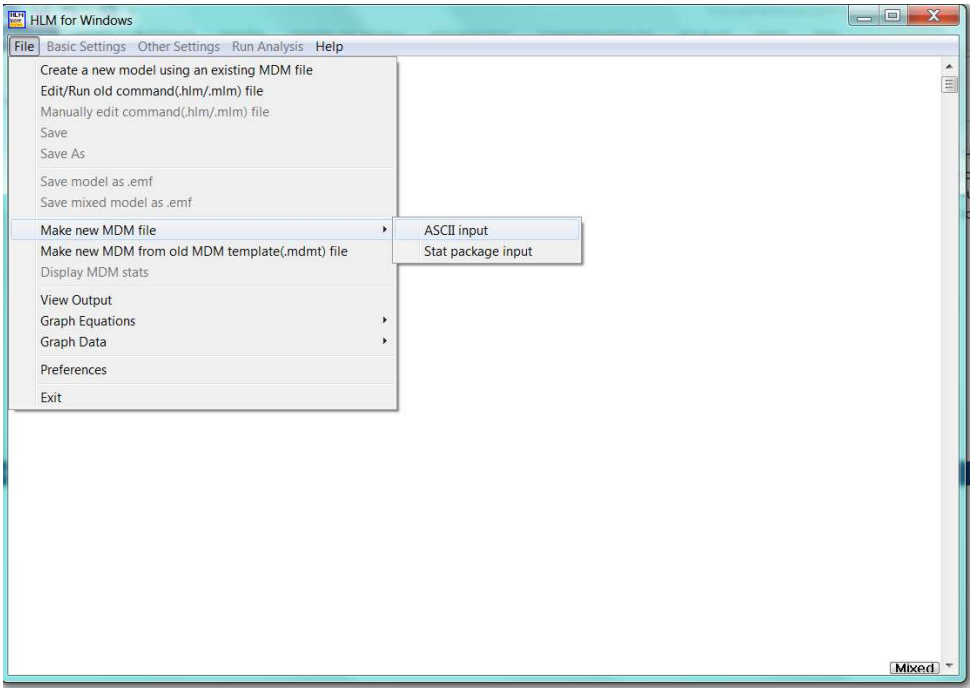
versión para el estudiante cuenta con importantes restricciones: no incluye la herramienta Stat/Transfer para la importación de datos y limita el número de observaciones en la creación de los modelos. Por ejemplo, para un modelo de tres niveles, el número máximo de observaciones que se puede utilizar en los niveles 1, 2 y 3 es de aproximadamente 8,000, 1,700 y 60, respectivamente. Para un modelo de dos niveles el número máximo de observaciones en los dos niveles es 8,000 en el nivel 1 y de 350 en el nivel 2. Además, no podrán ser incluidos más de 5 efectos en las ecuaciones de cualquier modelo, y el total de los efectos no podrán ser más de 25.

4.2 Preparando los Archivos en HLM.

- **Método 1 de entrada de datos:** Los archivos separados para cada nivel

Este método da lugar a un procesamiento más rápido, pero requiere más tiempo para establecer los datos. Requiere que los archivos separados deben crearse fuera de HLM 7 para cada nivel de análisis en el software HLM. Para los archivos SPSS, se trata de un formato con extensión .sav.

Por ejemplo, el software HLM 7 viene con ejemplos de archivos del estudio "High School and Beyond" Singer (1998). Los archivos SPSS para este ejemplo incluyen HSB1.SAV, que contiene el campo enlace del nivel 2 (ID es la identificación de la escuela o centro escolar) y las variables a nivel estudiantil. Hay varias filas por escuela, una fila por cada estudiante. Es fundamental que el archivo del nivel 1 este organizado de manera que todos los alumnos de una identificación de la escuela dado sean adyacentes (una relación unívoca).



Del mismo modo, el archivo de nivel escolar (nivel 2), HSB2.SAV, contiene el mismo campo de enlace para el nivel 2 las variables a nivel de la escuela.

- **Método de entrada 2:** Usando un archivo único.

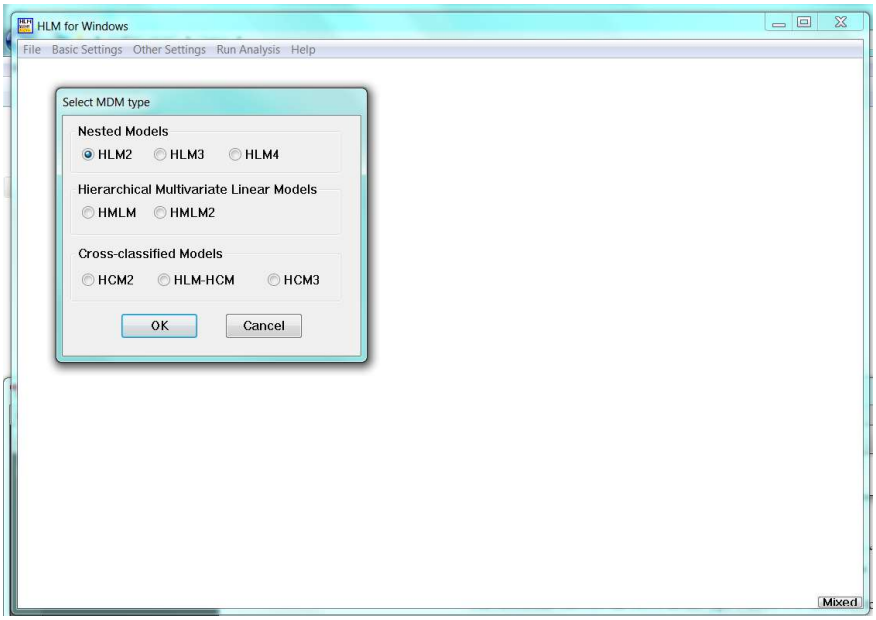
Este método 2 es más fácil en términos de gestión de datos y es el que se ilustra en este capítulo. Los mismos formatos de archivo paquete estadístico como para el Método 1 pueden estar utilizando. Para el ejemplo, el archivo de datos único debe ser ordenado de tal manera que todos los estudiantes para una identificación de la escuela dado tengan una relación unívoca.

- **Montando y creando el archivo MDM.**

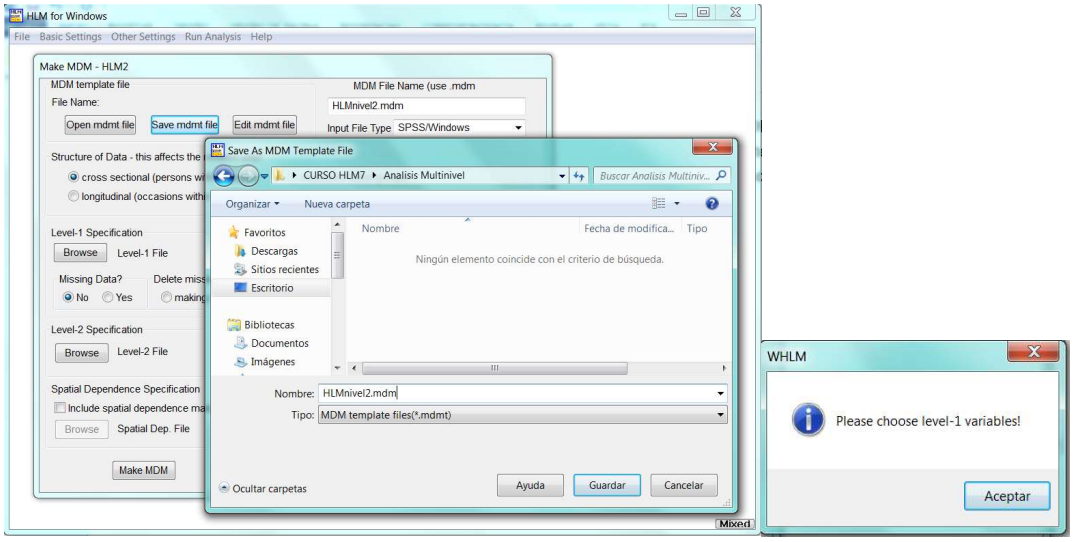
El siguiente paso es crear el archivo .MDM, que es nativa de datos de software de HLM formato.

- Hacer click en menú archivo – crear nuevo archivo MDM.

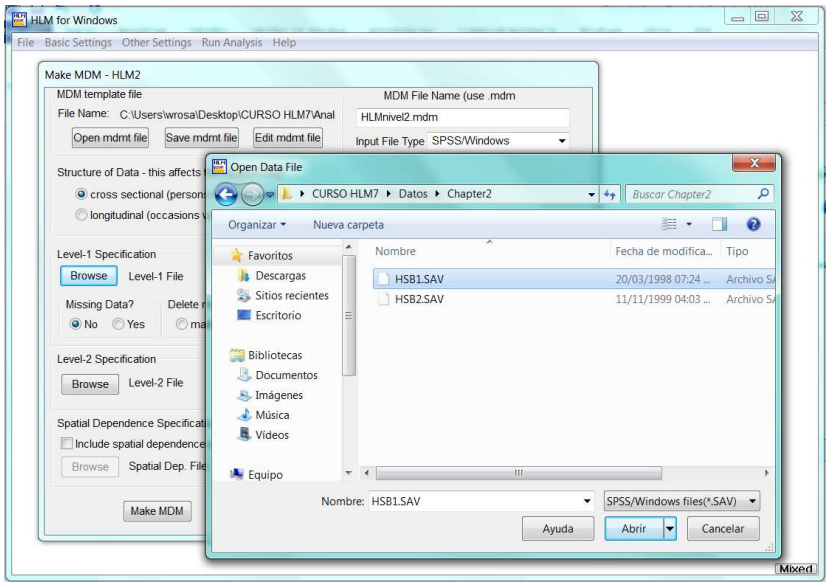
- Luego en la ventana que despliegue seleccionar de todas las opciones el método estadístico para 2 niveles (HLM2).



- El siguiente paso es darle un nombre al archivo MDM (para este ejemplo pondremos HLMnivel2.mdm) utilizando la extensión .mdm. Posteriormente en la opción Guardar archivo mdmt, buscamos el directorio de la carpeta donde guardaremos nuestra plataforma de trabajo. Ver imagen siguiente. Luego hacemos click en Guardar, aparecerá una notificación diciendo que seleccionemos la ubicación de los datos.

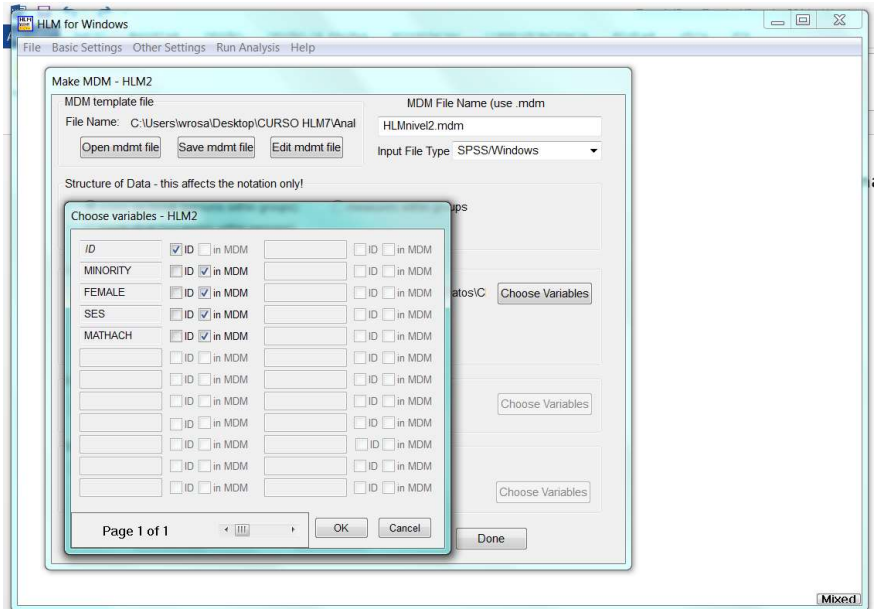


- Una vez guardado el archivo semilla para el análisis multinivel. Buscamos y especificamos todos los atributos del archivo perteneciente al nivel 1. Ver imagen siguiente.

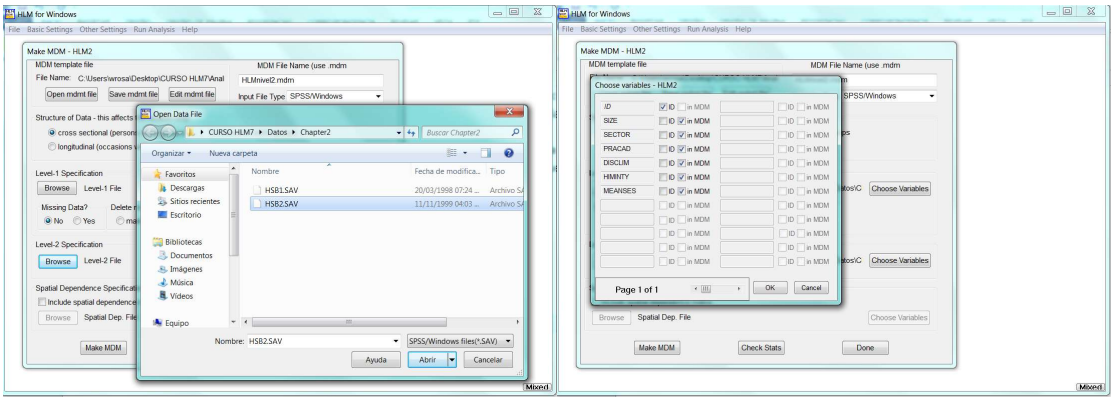


- En el archivo HSB1.SAV se muestra información relacionada al estudiante cómo el identificador del estudiante (ID), minoridad del estudiante (minority), genero femenino (female), nivel socioeconómico (ses) y resultado de matemática (mathach). Es de considerar, que para analizar este archivo, se hace unicamente para el genero femenino.

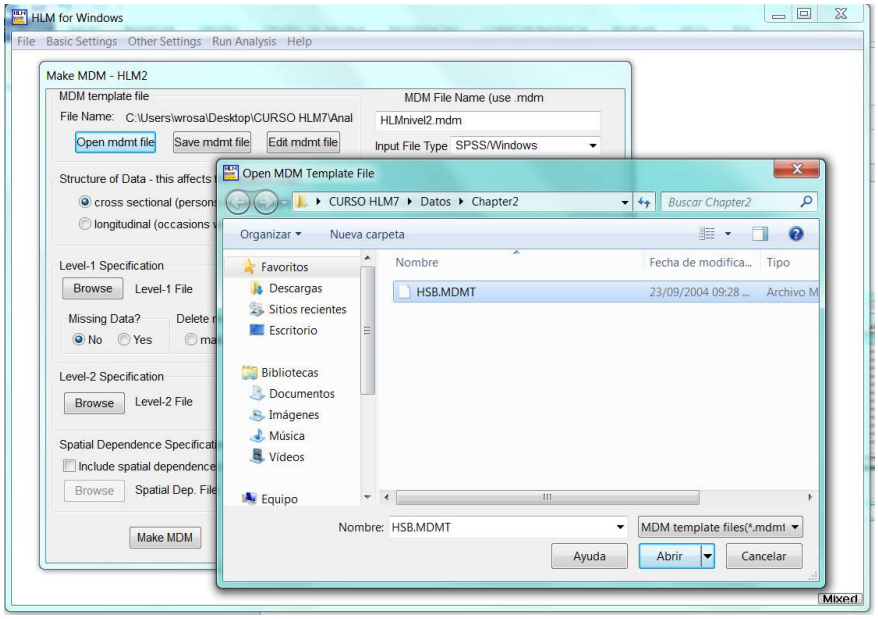
- Seleccionamos dichas variables dejando como identificador el campo ID. Ver imagen siguiente.



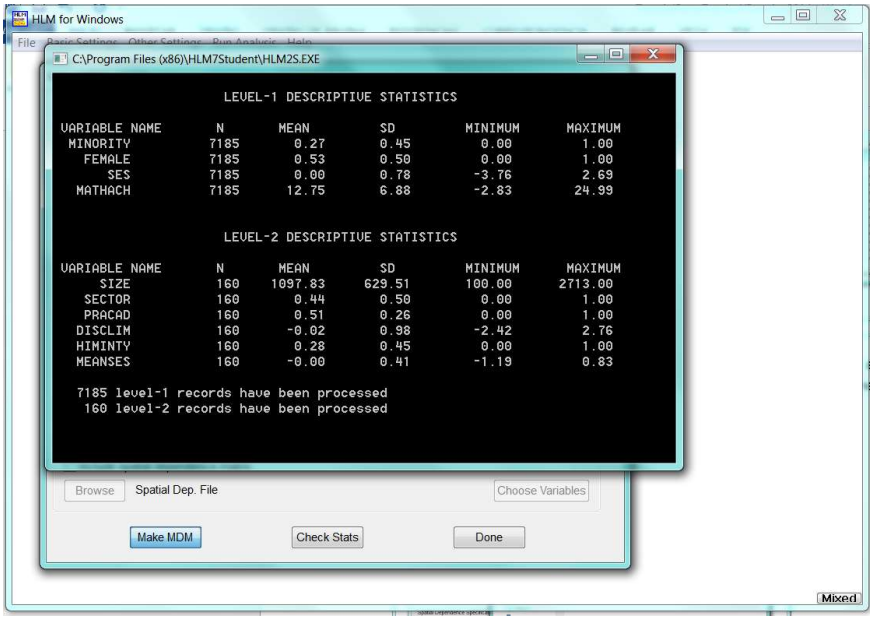
- Existen otras opciones luego de aceptar las variables identificadas para el analisis. Estas opciones son: a) si queremos trabajar con datos perdidos (missing), es decir, si seleccionamos que deseamos trabajar con datos perdidos hay que seleccionar si (yes), b) Si trabajaremos con datos perdidos entonces hay que decirle al software que borré los datos cuando trabaje con el archivo mdm, o si unicamente en la corrida de analisis. Para el ejemplo lo dejaremos por defecto.
- Posteriormente, seleccionamos la ruta del archivo para los datos pertenecientes al nivel 2 (centros escolares o escuela). Igual al procedimiento que realizamos para el archivo nivel 1. Ver imagen siguiente:



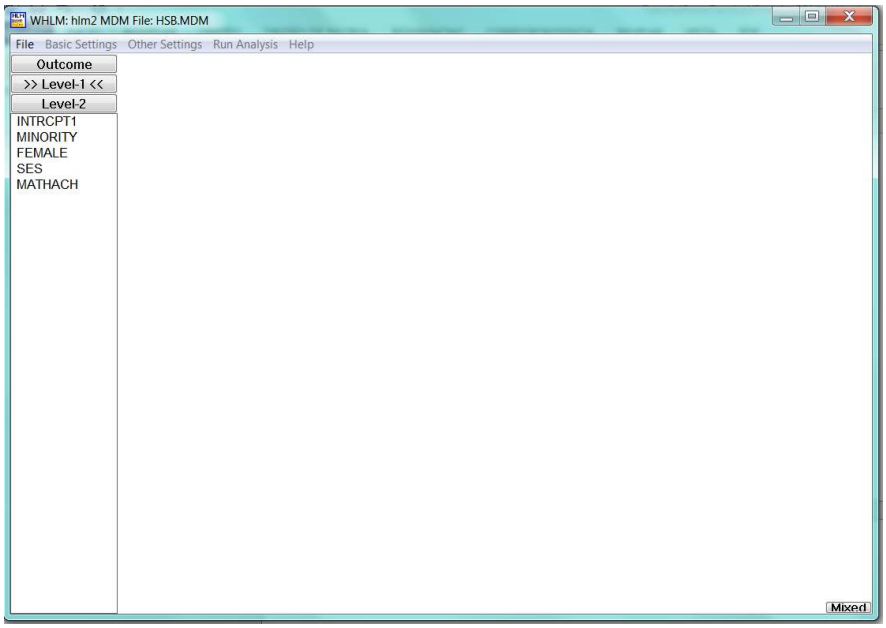
- Una vez terminado el procedimiento de montar los dos archivos ahora procedemos a seleccionar el archivo semilla creado por el software. Ver imagen siguiente:



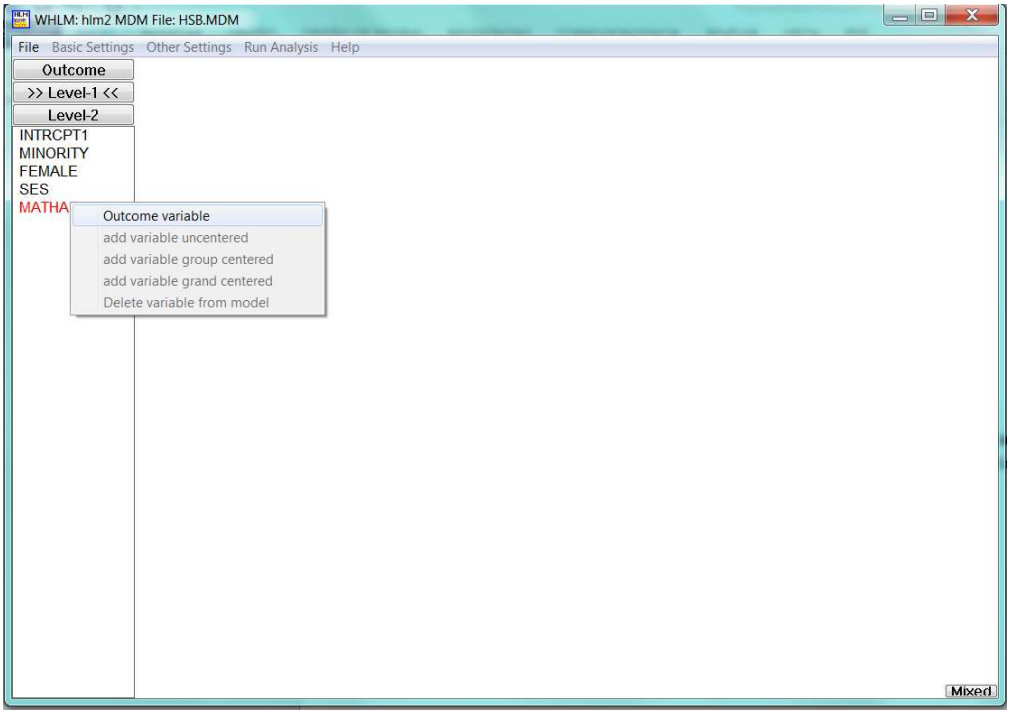
- Finalmente, si todo esta bien, y en dichos archivos existen las relaciones para el estudiante y el centro, hacemos click en Make MDM. Y el software HLM creará estadísticas para dichos niveles (estudiante y escuela). Ver imagen siguiente:



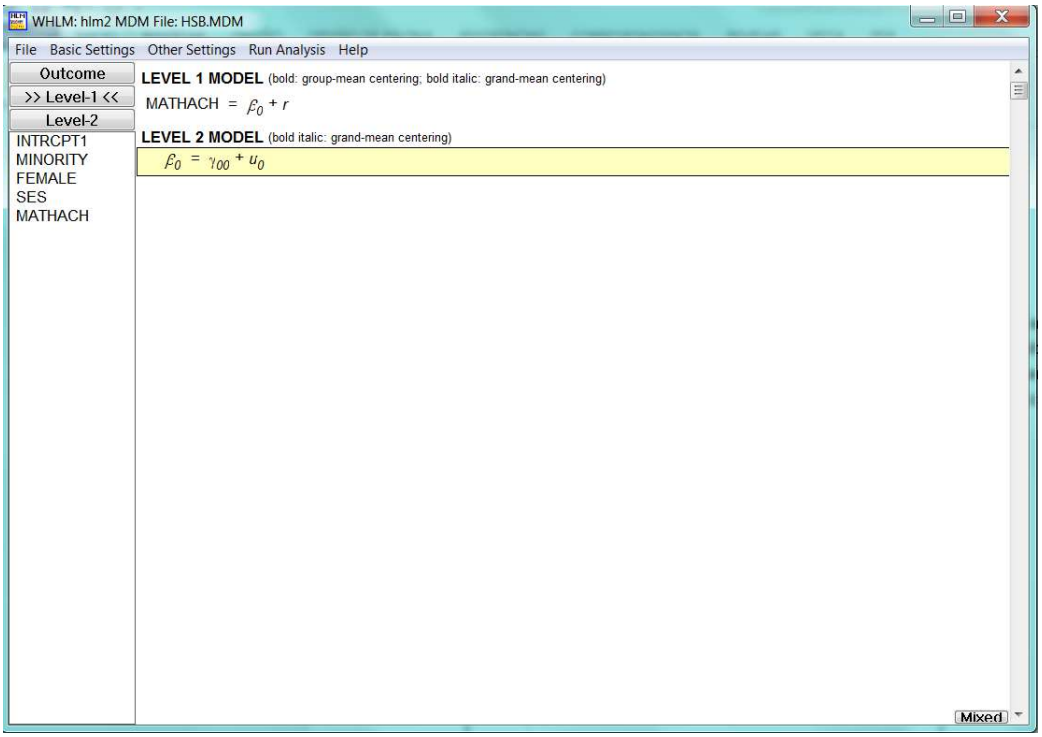
- Si existiera algun error el programa nos dijera que probablemente existen observaciones en el nivel 1 que no tiene relación con el nivel 2. La solución en este caso es borrar tales observaciones y trabajar sólo con aquellas donde exista relación con el ID.
- Finalmente, hacemos click en done y se desplagará la plataforma para realizar el análisis multinivel lineal, según este ejemplo. Asi:



- Para trabajar con dicho visor es necesario seleccionar la variable factor a analizar, para este ejemplo seleccionaremos la variable Resultado de Matemática (MATHACH). En la imagen siguiente se muestra al momento de seleccionar dicha variables una serie de opciones como: variable salida (Outcome variable), entre otras. Seleccionamos variable salida.



- En la imagen siguiente se puede apreciar la plataforma para poder realizar el procedimiento de análisis multinivel.

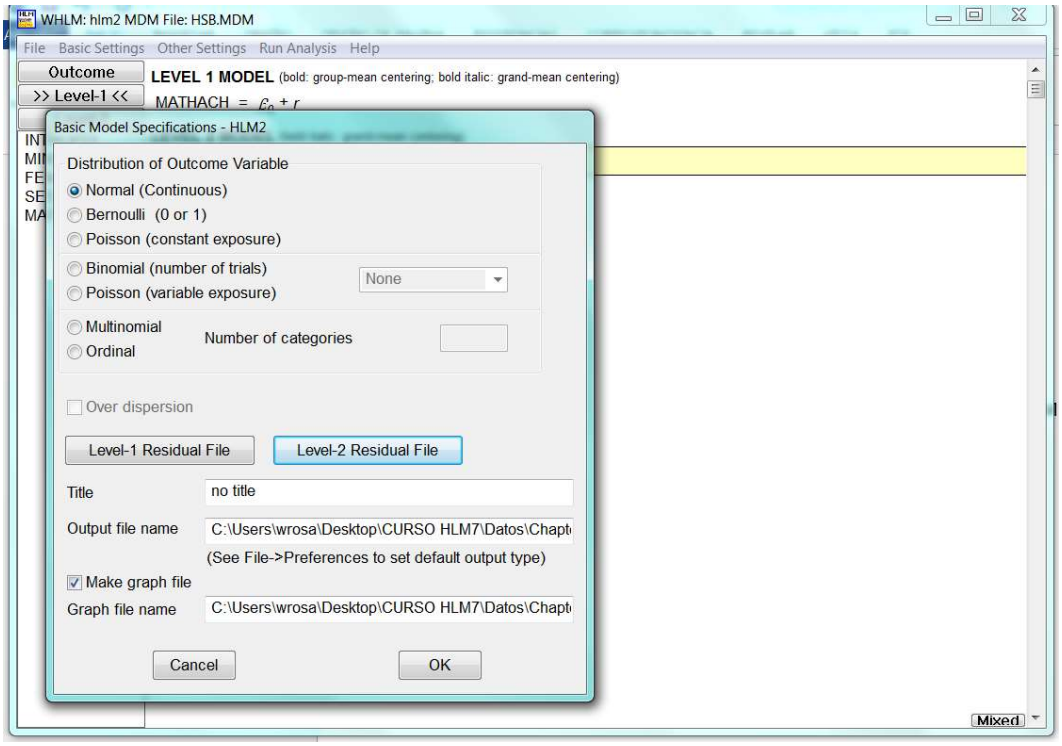


4.3 Plataforma del programa.

- **Descripción barra de menú Configuración Basica.**

En la parte superior del visor de análisis multinivel del software HLM se observa la barra de menu. Todo lo anterior lo hemos realizado en el menu archivo. En la imagen siguiente se observa el menu configuraciones basicas (basic settings). Se muestra un listado de aquellas distribuciones para la variables salida como:

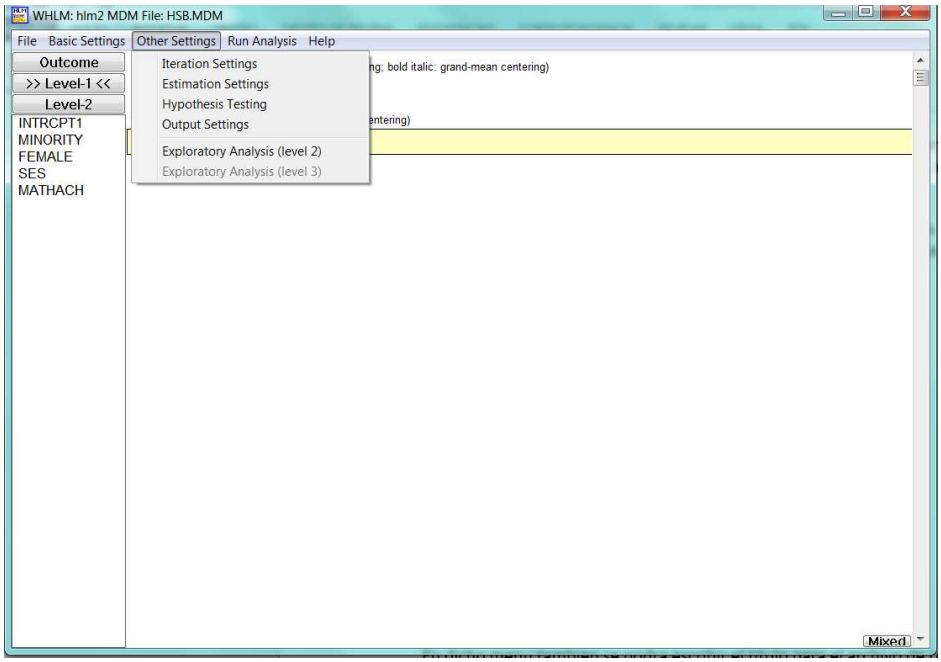
- **La distribución normal:** para una variables aletaria continua.
- **La distribución Bernoulli:** para una variable aleatoria discreta, donde la variable respuesta toma valores 0 y 1.
- Entre otras.



En dicho menu tambien se podra escribir el titulo para el archivo de resultado o salida. Para nuestro ejemplo, escribamos en el combo de texto “Análisis Multinivel para el Aprendizaje”.

- **Descripción barra de menu Otras configuraciones.**

Aquí se especifica la configuración sobre las iteraciones, es decir, cuantas iteraciones hay que especificar para que al momento de la corrida de análisis pueda existir una convergencia o solución. También, se especifican configuraciones para la estimación, prueba de hipótesis, entre otras. Ver imagen siguiente.



4.4 Metodologia para Realizar una Análisis Multinivel.

- **Metodologia.**

Ejemplo practico: considerando una variable de salida Resultado de Prueba de Matemática.

Generalmente la técnica estadística que se utiliza para determinar si una variable es un factor asociado al resultado escolar de los estudiantes, es analizar el grado de asociación lineal, y para decidir si la asociación es estadísticamente significativa se usa el test de la razón de máxima verosimilitud. Trabajando con la hipótesis de nulidad de diferencia igual a cero, la diferencia entre los valores de máxima verosimilitud de dos modelos sigue la distribución de Chi-cuadrado, con grados de libertad igual al número de nuevos parámetros. Habitualmente

para indicar el nivel de significación de cada estimación se usan como referencia el límite de probabilidad propuestos o utilizados por Fisher (0.01, 0.05, 0.0025). En este trabajo se utilizará el límite de 0.05.

En esta etapa verificaremos la correspondencia entre la información contenida en el centro educativo y familia, con la información del estudiante perteneciente a ese centro educativo. El procedimiento a seguir se describe a continuación:

a) **Realizar un análisis de regresión múltiple** donde la variable dependiente será el factor o constructo a analizar sea rendimiento o repitencia del estudiante versus las variables contextuales, pertenecientes al entorno del centro escolar. El propósito de este análisis es conocer aquellas variables que afectan más al factor mencionado a través de un contraste estadístico de tal forma que aquellas variables que no sean estadísticamente significativas no serán consideradas en el modelo estudiantil.
b) **Realizar análisis One-way:** consiste en realizar un modelo multinivel donde se considera el promedio del rendimiento y repitencia del centro educativo más en error aleatorio para en n-ésimo centro escolar. Construyendo de esta manera un modelo multinivel con una parte fija y otra aleatoria. El resultado será tres modelos multinivel one-way.

La estimación de los modelos nulos representa el punto de partida de todo análisis multinivel y presenta las siguientes formas funcionales y supuestos:

- Los modelos nulos no contienen ningún tipo de predictor (variable exógena), bien individual o grupal, excluyendo aquellas variables que pertenecen al nivel inferior o superior.
- El coeficiente intercepto equivale a la media global conformada por parte de la parte fija del modelo.
- Los términos de error del nivel escuela y del estudiante siguen una distribución normal con media igual a cero y varianzas iguales.

- La varianza total es igual a la suma de las varianzas de u_{0j} y e_{0ij} y la importancia del grupo, la cual es la proporción de la varianza total atribuible a ese nivel (ρ), aunque sin ningún control respecto del efecto de las variables de contexto.

c) Modelo Multinivel Óptimo: aquí se incluye unicamente aquellas variables que resultaron ser estadísticamente significativas.

4.5 Aplicación modelo multinivel lineal.

Planteamiento y Resultado del Modelo Multinivel Nulo.

Summary of the model specified

Level-1 Model

MATHACHij = β0j + rij

Level-2 Model

β0j = γ00 + u0j

Mixed Model

MATHACHij = γ00 + u0j+ rij

Donde,

$MATHACH_{ij}$: es el Resultado de Prueba de Matemática del i-ésimo estudiante en la j-ésima escuela.

γ00 : es el promedio global de reprobación Resultado de Prueba de Matemática de los centros escolares.

u0j : es el error aleatorio para el nivel dos (centros escolares)

rij : es el error aleatorio para el nivel uno (estudiantes).

Final Results - Iteration 4

Iterations stopped due to small change in likelihood function

$\sigma^2 = 39.14831$

τ

INTRCPT1,β_0	8.61431

Random level-1 coefficient	Reliability estimate
INTRCPT1,β_0	0.901

The value of the log-likelihood function at iteration 4 = -2.355840E+004

Final estimation of fixed effects:

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	12.636972	0.244412	51.704	159	<0.001

Final estimation of fixed effects (with robust standard errors)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	12.636972	0.243628	51.870	159	<0.001

Final estimation of variance components

Random Effect	Standard Deviation	Variance Component	*d.f.*	χ^2	*p*-value
INTRCPT1, u_0	2.93501	8.61431	159	1660.23259	<0.001
level-1, *r*	6.25686	39.14831			

Statistics for current covariance components model

Deviance = 47116.793477

Number of estimated parameters = 2

4.5.1 Análisis Modelo Nulo.

El coeficiente de correlación intra-clase, representa en este modelo la proporción de varianza de la variable respuesta Resultado de Prueba de Matemática, el cual es estimado de la siguiente manera:

$$\rho = \frac{\hat{\sigma}_{u0}^2}{\hat{\sigma}_{u0}^2 + \hat{\sigma}_{e0}^2} = \frac{8.61431}{8.61431 + 39.14831} = \frac{8.61431}{47.7626} = 18.03\%$$

De la variación total de la nota de reprobación es el 18.03% y se debe a las diferencias entre los centros escolares. Representa el peso que tienen las características grupales o internas al sistema escolar en la explicación de las variaciones totales del resultado de prueba de matemática. Por otro lado, la estimación del nivel 1, alumno (81.97%) representa el peso de los factores externos al sistema escolar.

Un indicador global de la fiabilidad es el promedio de fiabilidad de los centros escolares, el cual está dado por:

$$\hat{\lambda}_j = \sum \frac{\hat{\lambda}_j}{j} = 0.901$$

Lo cual nos indica que la media muestral tiende a ser un buen estimador de la medida verdadera del centro escolar.

Planteamiento y Resultado del Modelo Óptimo.

Level-1 Model

MATHACHij = β0j + β1j*(MINORITYij) + β2j*(SESij) + rij

Level-2 Model

β0j = γ00 + γ01*(SIZEj) + γ02*(SECTORj) + u0j

β1j = γ10

β2j = γ20

MINORITY SES have been centered around the group mean. SIZE SECTOR have been centered around the grand mean.

Mixed Model

MATHACHij = γ00 + γ01*SIZEj + γ02*SECTORj

+ γ10*MINORITYij

+ γ20*SESij

+ u0j+ rij

Donde,

$MATHACH_{ij}$: es el Resultado de Prueba de Matemática del i-ésimo estudiante en la j-ésima escuela.

$MINORITY_{ij:}$ es la variable minoridad del i-ésimo estudiante en la j-ésima escuela.

SES_{ij} : es el nivel socioeconómico del i-ésimo estudiante en la j-ésima escuela.

SIZEj : es la variable Tamaño de la j-ésima escuela

SECTORj: es la variable Sector que pertenece la j-ésima escuela

β_{1j} : es el peso esperado de la variable minoridad en la j-ésima escuela.

B_{2j} : es el peso esperado de la variable del nivel socio-económico en la j-ésima escuela.

β_{0j} : es el promedio global de reprobación Resultado de Prueba de Matemática de los centros escolares.

u0j : es el error aleatorio para el nivel dos (centros escolares)

rij : es el error aleatorio para el nivel uno (estudiantes).

Final Results - Iteration 4

Iterations stopped due to small change in likelihood function

σ2 = 36.12688

τ

INTRCPT1,β0	6.70134

Random level-1 coefficient	Reliability estimate
INTRCPT1,β0	0.885

The value of the log-likelihood function at iteration 4 = -2.326450E+004

Final estimation of fixed effects:

Fixed Effect	Coefficient	Standard error	t-ratio	Approx. d.f.	p-value
For INTRCPT1, β0					
INTRCPT2, γ00	12.620818	0.217552	58.013	157	<0.001
SIZE, γ01	0.000616	0.000390	1.580	157	0.116
SECTOR, γ02	3.153818	0.490215	6.434	157	<0.001
For MINORITY slope, β1					
INTRCPT2, γ10	-2.895582	0.220143	-13.153	7023	<0.001
For SES slope, β2					
INTRCPT2, γ20	1.952476	0.108873	17.934	7023	<0.001

Final estimation of fixed effects (with robust standard errors)

Fixed Effect	Coefficient	Standard error	t-ratio	Approx. d.f.	p-value
For INTRCPT1, β0					
INTRCPT2, γ00	12.620818	0.215521	58.560	157	<0.001
SIZE, γ01	0.000616	0.000381	1.616	157	0.108
SECTOR, γ02	3.153818	0.458148	6.884	157	<0.001
For MINORITY slope, β1					

INTRCPT2, γ10	-2.895582	0.258532	-11.200	7023	<0.001
For SES slope, β2					
INTRCPT2, γ20	1.952476	0.120874	16.153	7023	<0.001

Final estimation of variance components

Random Effect	Standard Deviation	Variance Component	d.f.	χ2	p-value
INTRCPT1, u0	2.58869	6.70134	157	1391.46372	<0.001
level-1, r	6.01056	36.12688			

Statistics for current covariance components model

Deviance = 46528.996829

Number of estimated parameters = 2

4.5.2 Análisis Modelo Multinivel Lineal Óptimo

En la tabla de estimación de los parámetros del modelo multinivel óptimo para la nota de reprobación se observan las variables que resultaron ser estadísticamente significativas a un nivel de significancia del 5%, luego de realizar una calibración del modelo ajustado partiendo de la eliminación de aquellas variables que no resultaron ser estadísticamente significativas.

En dicha tabla se muestra que el intercepto o media global del rendimiento matemática de los estudiantes de los centros escolares es de 13.48. El promedio de nota de reprobación de la j-ésima escuela está determinada por la regresión de las variables externas pertenecientes al nivel 2 o centros escolares. Dichas variables con la estimación de sus coeficientes de regresión indica que a un nivel de significancia del 5% están explicando la nota de reprobación de un estudiante, no obstante, si consideramos la estimación de fiabilidad para la interpretación del parámetro de cada variable, hay que considerar que sólo el 65.3% refleja

una fiabilidad, únicamente en términos de relación con el rendimiento de matemática del estudiante.

a) Método propuesto por Golstein.

i) Variación explicada para el nivel 2.

La varianza explicada en este nivel se calcula comparándola con la del modelo nulo, así:

$$Var_{\exp} = \frac{\hat{\sigma}^2_{u0(modelo\ nulo)} - \hat{\sigma}^2_{u0(modelo\ optimo)}}{\hat{\sigma}^2_{u0(modelo\ nulo)}} = \frac{8.61431 - 6.70134}{8.61431} = 22.20\%$$

Por tanto la proporción de varianza explicada por el modelo del nivel 2 es igual a 22.2%, y significa que un 22.2% de la variabilidad debida al estudiante se explica por características internas al sistema escolar, el resto es debido a otras variables ya sea pertenecientes al estudiante y factores externos al sistema educativo como: delincuencia, pobreza, nivel socioeconómico, entre otros.

ii) Variación explicada en el nivel 1[1]

La varianza explicada en este nivel se calculará comparándola con la del modelo nulo así:

$$\text{var\_exp\_}nivel\_1 = \frac{\left(\hat{\sigma}^2_{e0}(Modelo\ nulo) - \hat{\sigma}^2_{e0}(Modelo\ ajustado)\right)}{\hat{\sigma}^2_{e0}(Modelo\ nulo)}$$

$$\boldsymbol{var_{exp_{nivel_1}}} = \frac{(39.14831 - 36.12688)}{39.14831} = \mathbf{7.7\%}$$

[1] Para mayor información sobre este concepto ver Anthony S. Bryk, Stephen W Raundenbush. “Hierarchical Linear Models: Applications and Data Analysis Methods”. Advanced Quantitative Techniques in the Sciences Series, 1992, pág. 70

Por tanto, la proporción de varianza explicada por el modelo del nivel 1 es igual al 7.7%, y significa que un 7.7% de la variabilidad debida al estudiante se explica por las variables minoridad y nivel socioeconómico.

Ejercicio.

De la estimación de Varianza propuesto por Golstein, Snijders. Calcular:

a) El coeficiente de partición o división de la varianza VPC para el modelo óptimo.
b) Reducción de la proporción del error de predicción de un valor residual.
c) Reducción de la proporción del error de predicción de la medida de grupo.
d) Analizar dichos resultados.

4.6 Aplicación Modelo Multinivel No Lineal.

4.6.1 Planteamiento del Modelo Nulo repitencia

Summary of the model specified

Level-1 Model

$\text{Prob}(REP1_{ij}=1|\beta_j) = \phi_{ij}$

$\log[\phi_{ij}/(1 - \phi_{ij})] = \eta_{ij}$

$\eta_{ij} = \beta_{0j}$

Level-2 Model

$\beta_{0j} = \gamma_{00} + u_{0j}$

Level-1 variance = $1/[\phi_{ij}(1-\phi_{ij})]$

Mixed Model

$\eta_{ij} = \gamma_{00} + u_{0j}$

The value of the log-likelihood function at iteration 6 = -2.444614E+003

Results for Non-linear Model with the Logit Link Function Unit-Specific Model, PQL Estimation - (macro iteration 7)

τ

INTRCPT1,β_0	1.31578

Random level-1 coefficient	Reliability estimate
INTRCPT1,β_0	0.688

The value of the log-likelihood function at iteration 2 = -9.980880E+003

Final estimation of fixed effects: (Unit-specific model)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	-2.033670	0.073305	-27.743	355	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	-2.033670	0.130854	(0.113,0.151)

Final estimation of fixed effects (Unit-specific model with robust standard errors)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					

INTRCPT2, γ_{00}	-2.033670	0.073198	-27.783	355	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	-2.033670	0.130854	(0.113,0.151)

Final estimation of variance components

Random Effect	Standard Deviation	Variance Component	*d.f.*	χ^2	*p*-value
INTRCPT1, u_0	1.14707	1.31578	355	1489.67665	<0.001

Results for Population-Average Model

The value of the log-likelihood function at iteration 2 = -1.011173E+004

Final estimation of fixed effects: (Population-average model)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	-1.730812	0.069381	-24.947	355	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	-1.730812	0.177141	(0.155,0.203)

Final estimation of fixed effects (Population-average model with robust standard errors)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	-1.730812	0.064992	-26.631	355	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	-1.730812	0.177141	(0.156,0.201)

4.6.2 Planteamiento Modelo Multinivel No Lineal Repitencia.

Summary of the model specified

Level-1 Model

$\text{Prob}(REP1_{ij}=1|\beta_j) = \phi_{ij}$

$\log[\phi_{ij}/(1 - \phi_{ij})] = \eta_{ij}$

$\eta_{ij} = \beta_{0j} + \beta_{1j}*(MALE_{ij}) + \beta_{2j}*(PPED_{ij})$

Level-2 Model

$\beta_{0j} = \gamma_{00} + \gamma_{01}*(MSESC_j) + u_{0j}$

$\beta_{1j} = \gamma_{10} + u_{1j}$

$\beta_{2j} = \gamma_{20} + u_{2j}$

Level-1 variance = $1/[\phi_{ij}(1-\phi_{ij})]$

Mixed Model

$\eta_{ij} = \gamma_{00} + \gamma_{01}*MSESC_j$

$+ \gamma_{10}*MALE_{ij}$

$+ \gamma_{20}*PPED_{ij}$

$+ u_{0j} + u_{1j}*MALE_{ij} + u_{2j}*PPED_{ij}$

Results for Non-linear Model with the Logit Link Function Unit-Specific Model, PQL Estimation - (macro iteration 626)

τ

INTRCPT1,β_0	1.32276	0.06223	-0.22261
MALE,β_1	0.06223	0.11094	0.05098
PPED,β_2	-0.22261	0.05098	0.09298

τ (as correlations)

INTRCPT1,β_0	1.000	0.162	-0.635
MALE,β_1	0.162	1.000	0.502
PPED,β_2	-0.635	0.502	1.000

Random level-1 coefficient	Reliability estimate
INTRCPT1,β_0	0.380
MALE,β_1	0.051
PPED,β_2	0.029

Note: The reliability estimates reported above are based on only 239 of 356 units that had sufficient data for computation. Fixed effects and variance components are based on all the data.
The value of the log-likelihood function at iteration 2 = -9.968402E+003

Final estimation of fixed effects: (Unit-specific model)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	-2.036507	0.095061	-21.423	354	<0.001
MSESC, γ_{01}	-0.300702	0.192141	-1.565	354	0.118
For MALE slope, β_1					
INTRCPT2, γ_{10}	0.454126	0.076809	5.912	355	<0.001
For PPED slope, β_2					
INTRCPT2, γ_{20}	-0.530918	0.097587	-5.440	355	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	-2.036507	0.130484	(0.108,0.157)
MSESC, γ_{01}	-0.300702	0.740299	(0.507,1.080)
For MALE slope, β_1			
INTRCPT2, γ_{10}	0.454126	1.574796	(1.354,1.832)
For PPED slope, β_2			
INTRCPT2, γ_{20}	-0.530918	0.588065	(0.485,0.713)

Final estimation of fixed effects (Unit-specific model with robust standard errors)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	-2.036507	0.094394	-21.575	354	<0.001
MSESC, γ_{01}	-0.300702	0.203403	-1.478	354	0.140
For MALE slope, β_1					
INTRCPT2, γ_{10}	0.454126	0.075885	5.984	355	<0.001
For PPED slope, β_2					
INTRCPT2, γ_{20}	-0.530918	0.095354	-5.568	355	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	-2.036507	0.130484	(0.108,0.157)
MSESC, γ_{01}	-0.300702	0.740299	(0.496,1.104)
For MALE slope, β_1			
INTRCPT2, γ_{10}	0.454126	1.574796	(1.356,1.828)
For PPED slope, β_2			
INTRCPT2, γ_{20}	-0.530918	0.588065	(0.487,0.709)

Final estimation of variance components

Random Effect	Standard Deviation	Variance Component	*d.f.*	χ^2	*p*-value
INTRCPT1, u_0	1.15011	1.32276	237	423.90971	<0.001
MALE slope, u_1	0.33308	0.11094	238	215.29381	>0.500
PPED slope, u_2	0.30492	0.09298	238	166.08884	>0.500

Note: The chi-square statistics reported above are based on only 239 of 356 units that had sufficient data for computation. Fixed effects and variance components are based on all the data.

Results for Population-Average Model

The value of the log-likelihood function at iteration 2 = -9.330431E+003

Final estimation of fixed effects: (Population-average model)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	-1.664338	0.082690	-20.127	354	<0.001
MSESC, γ_{01}	-0.283367	0.175996	-1.610	354	0.108
For MALE slope, β_1					
INTRCPT2, γ_{10}	0.418750	0.060204	6.956	355	<0.001
For PPED slope, β_2					
INTRCPT2, γ_{20}	-0.482296	0.078926	-6.111	355	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	-1.664338	0.189316	(0.161,0.223)
MSESC, γ_{01}	-0.283367	0.753243	(0.533,1.065)
For MALE slope, β_1			
INTRCPT2, γ_{10}	0.418750	1.520061	(1.350,1.711)

For PPED slope, β_2			
INTRCPT2, γ_{20}	-0.482296	0.617364	(0.529,0.721)

Final estimation of fixed effects (Population-average model with robust standard errors)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	-1.664338	0.059670	-27.892	354	<0.001
MSESC, γ_{01}	-0.283367	0.139462	-2.032	354	0.043
For MALE slope, β_1					
INTRCPT2, γ_{10}	0.418750	0.045914	9.120	355	<0.001
For PPED slope, β_2					
INTRCPT2, γ_{20}	-0.482296	0.059139	-8.155	355	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	-1.664338	0.189316	(0.168,0.213)
MSESC, γ_{01}	-0.283367	0.753243	(0.573,0.991)
For MALE slope, β_1			
INTRCPT2, γ_{10}	0.418750	1.520061	(1.389,1.664)
For PPED slope, β_2			
INTRCPT2, γ_{20}	-0.482296	0.617364	(0.550,0.694)

Análisis:

En cuento al factor de **Repitencia**, se tiene que, dado que en los modelos de regresión logística se supone que los residuos siguen una distribución de ese tipo, el valor de la varianza de $r_{ij}(\sigma^2)$ es igual a $\frac{\pi^2}{3} = 3.29$ y representa la varianza residual dentro de cada centro. El error aleatorio u_{0j} es la desviación no explicada de la escuela "j" respecto del promedio del

conjunto de las escuelas o centros escolares. Se supone que es una variable aleatoria independiente con media igual a cero y varianza igual a τ^2, la cual representa la varianza entre centros.

En este tipo de modelos, el efecto de las variables explicativas se evalúa a través de los denominados odd-ratios (razones de probabilidad). Estos se calculan como $\exp(\beta)$, siendo β un nombre genérico asignado a los coeficientes de la regresión. Los odds-ratios miden la probabilidad de que ocurra un suceso Y, condicionada al mismo evento X. Las razones de probabilidad asociadas a variables cuyos coeficientes son positivos son mayores a uno, mientras que las asociadas a coeficientes negativos son menores a la unidad.

En el modelo anterior "**Modelo Nulo**" resulta útil estimarlo, debido a que permite conocer qué proporción de la desigualdad en los resultados se debe a diferencias entre centros (coeficiente de correlación intraclase, ρ) y qué proporción se vincula con diferencias en su interior. En este caso como modelo de regresiones multinivel logística este coeficiente se calcula aplicando la ecuación $\rho = \tau^2/(\tau^2 + \sigma^2)$, donde σ^2 es constante igual a 3.29.

En el cuadro de estimación de los efectos fijos y aleatorios del modelo óptimo para el factor repitencia, se observa que los Odd-ratio mayores a la unidad, el signo del coeficiente estimado es negativo. Indicando un aumento en probabilidad de que un estudiante repita de grado en un centro educativo. Caso contrario, si el valor del ratio es menor a la unidad, entonces indica que existirá una menor incidencia en la probabilidad de que pueda repetir cierto estudiante en un centro educativo.

5. Bibliografia

Browne, W.J. And J. Rasbash, Multilevel Modelling (1999). Institute of Education, University of London.

Cervini, Rubén (2002). Desigualdades Socioculturales en el Aprendizaje de Matemática y Lengua de la Educación Secundaria en Argentina.

Ferrão. Leite. Beltrão (2001). Introdução à modelagem multinível em Avaliação Educacional. Ministério do Planejamento, Orçamento e Gestão Instituto Brasileiro de Geografia e Estatística – IBGE Escola Nacional de Ciências Estatísticas. Rio de Janeiro.

Goldstein, Harvey. Rasbash. Yang. Geoffrey, Woodhouse (1993). Sally Thomas. A multilevel analysis of school examination results. Oxford review of education, vol 19, No. 4. Goldstein. Tutorial in biostatistics Multilevel modelling of medical data. Institute of Education; University of London; London; U.K. Goldstein, H. (1997). Methods in school effectiveness research. School effectiveness and school improvement.

Gujarati, Damodar N. Econometría Básica. Tercera edición, United States Military Academy, West Point, 1997.

Joop J. Hox, 2010. Multilevel Analysis. Techniques and Applications Quantitative Methodology Series. Second Edition.

Montgomery. Peck. Vining (2002). Introducción al Análisis de Regresión Lineal. Primera edición Mexico.

Naderiand J. Mace (2002). Education and Earnings; A Multilevel Analysis A Case Study of the Manufacturing Sector in Iran. Management and Planning Organisation, Tehran, Iran. University of London, London, U.K.

Rosa, Welman (2005). Tesis: Teoría de Modelos Multinivel y sus Aplicaciones. Universidad de El Salvador, C.A.