

Efecto del estatus socioeconómico y la edad sobre los resultados de Lectura en estudiantes de sexto grado

Por: Mario Raúl Moreno Grajeda¹

Julio de 2013

Resumen

Este artículo se enfoca en analizar si el estatus socioeconómico y la edad de los estudiantes tienen un efecto sobre los resultados de los alumnos de sexto grado en Guatemala, cuando se controla por la etnia y género del estudiante así como el área de ubicación de la escuela. Para tal efecto, se utilizan los datos de la evaluación del Nivel de Educación Primaria del año 2008. Los modelos de regresión utilizados sugieren que el estatus socioeconómico tiene un impacto positivo sobre el rendimiento de lectura de los estudiantes. Sin embargo, la edad de los estudiantes tiene un impacto negativo. En ambos casos, el impacto está sobre y va más allá de la etnia y género y de la ruralidad o urbanidad de las escuelas.

¹ Estudiante de Doctorado en el programa de Métodos Cuantitativos de la Educación en la Universidad de Minnesota. Exbecario Fulbright, con maestría en Economía Aplicada en Oklahoma State University. Colaborador del Ministerio de Educación. Para contactar al autor: mrmorenog@yahoo.com y mrmorenograjeda@hotmail.com.

I. Introducción

En Guatemala usualmente se utiliza las variables género, etnicidad, área y tiempo invertido en lectura, como variables predictoras del rendimiento escolar de lectura. Estas variables predictoras son importantes debido a que los registros oficiales indican que existe una brecha en la educación entre hombres y mujeres, en términos de años de escolaridad. En tal sentido, el promedio de años de escolaridad de las mujeres es de 5.26 años mientras que el de los hombres es de 6.34 años, según el Instituto Nacional de Estadística de Guatemala –INE– (2006). Además, el 13.85% de las mujeres entre 20 y 24 años de edad, ha finalizado la escuela primaria, el 11.9% el ciclo básico y el 9.8% el diversificado (Esquivel, 2006).

Respecto a la etnicidad, la encuesta de condiciones de vida del año 2006 (ENCOVI) reveló que 38.4% de la población pertenece a un grupo indígena. Considerando el área de residencia, la misma encuesta indicó que el 51% de la población vive en la área rural, en donde las condiciones de vida son menos favorables para los pobladores, comparados con el área urbana (INE, 2006). La variable tiempo invertido en lectura es importante debido a que se tiene la idea que cuanto más tiempo se invierta en leer, mayor será el rendimiento escolar.

Este artículo se enfoca en analizar a los estudiantes de sexto grado de Guatemala, utilizando los resultados de la evaluación del nivel primario del año 2008. Para tal efecto se utiliza diferentes modelos de regresión basados en dos importantes preguntas de investigación.

II. Preguntas de investigación

Los registros estadísticos oficiales de Guatemala dicen que casi el 51% de la población vive en niveles de pobreza. Por tal situación el estatus socioeconómico es una variable importante para determinar su impacto en los resultados de lectura. Por otra parte, los niños de sexto primaria deberían tener 12 años de edad cuando se encuentran en dicho grado, bajo el supuesto que la edad adecuada y reglamentaria para ingresar al nivel primario es de 7 años. Sin embargo, el Ministerio de Educación de Guatemala (2007) indica que el porcentaje promedio de estudiantes en sexto grado con extraedad es del 29.5%, y de estos, solo en el sector público es de 28.3%. Por ello, la edad de los estudiantes resulta ser una variable interesante de examinar y determinar su impacto sobre los resultados de lectura.

Entonces la primera pregunta de la investigación es:

- ¿El estatus socioeconómico –SES– y la edad de los estudiantes tienen algún efecto en los resultados de lectura, cuando el género del estudiante, el área de ubicación de la escuela y la etnia son controlados? ¿Cuál es la magnitud del efecto? ¿Existe un incremento significativo en la cantidad de la varianza explicada incluyendo el estatus socioeconómico de los estudiantes y su edad?

Puesto que el 70% de la población caracterizada como pobre en Guatemala vive en el área rural y 56% de estos pobres son personas indígenas, la interacción entre estatus socioeconómico y área y, la interacción entre estatus socioeconómico y etnia pueden considerarse como predictores de interés para el rendimiento de lectura, la segunda pregunta de investigación es:

- Una vez el estatus socioeconómico y la edad son incluidos en el modelo de regresión, ¿la interacción del estatus socioeconómico y la etnicidad y, la interacción de estatus socioeconómico y el área, tienen algún efecto sobre el rendimiento de lectura? Si lo tienen, ¿cuál es la magnitud de dicho efecto? Y ¿cuál es el impacto en el poder predictivo del modelo?

III. Limitación del estudio

Una limitación del estudio es que las inferencias son realizadas para los estudiantes, pero el muestreo fue realizado para escuelas y no para los alumnos. Otra limitación es que los datos mantienen una estructura anidada, pero el análisis se ha realizado con los procedimientos de regresión lineal estándar.

IV. Método

A fin de responder a las preguntas de investigación, se construyeron tres modelos de regresión con la técnica de mínimos cuadrados. El primer modelo se le ha llamado modelo 1 e incluye seis variables. El segundo modelo, llamado modelo 2 incluye ocho variables, mientras que el tercer modelo (modelo 3) incluye 10 variables. Todas las variables están descritas más adelante.

Dos test de “F incremental” fueron desarrollados para determinar el impacto de las variables sobre el poder predictivo del modelo. El primer test examinó el efecto de la magnitud del estatus socioeconómico y la edad de los estudiantes cuando estas variables fueron incluidas en el modelo, después de incluir las variables control; al hacer esto quedó determinado el modelo 2. El segundo test

se realizó para determinar el impacto de las interacciones SES y área y SES y etnia, cuando son incluidas en el modelo 3. La fórmula utilizada para el test “F incremental” fue:

$$F = \frac{R_{Full}^2 - R_{Reduce}^2}{1 - R_{Full}^2} \times \frac{df_{Residual F}}{df_{Change}} =$$

Donde

R_{Full}^2 = es el coefeciente de Determinacion del modelo Final.

R_{Reduce}^2 = es el coefeciente de Determinacion del modelo reducido.

$df_{Residual F}$ = grados de liberta del error del modelo Final.

df_{Change} = Cambio en los grados de libertad cuando se va del modelo reducido al Final.

El valor crítico para F fue calculado de acuerdo a $F_{dfchange,dfresidual}(0.95)$.

Los supuestos de normalidad, homocedasticidad y linearidad fueron chequeados para los modelos por medio de Q-Q plot, histogramas de los residuos, el plot de residuos estandarizados versus valores ajustados del modelo. También se chequearon los valores extremos y los valores de discrepancia. Se consideró como valor extremo a aquel cuyo valor “leverage” fue mayor a tres veces el valor de la media de los valores “leverage” (0.00102). También se consideró como valor de discrepancia a aquellos valores cuyo residuo estandarizado fue menor que -2 y mayor que 2.

4.1. Descripción de los datos

Los datos provienen de la evaluación nacional de lectura del nivel primario del año 2008. Todos los estudiantes evaluados pertenecen al sector público del Sistema Educativo Nacional de Guatemala. La muestra original incluyó a 32 459 estudiantes, distribuidos en 895 escuelas de los 22 departamentos del país. Sin embargo, cuando los modelos fueron construidos con el programa estadístico R, los datos faltantes fueron eliminados por el programa. Por tal razón el análisis final se llevó a cabo con 29 412 estudiantes de 872 escuelas de todo el país.

4.2. Descripción de las variables

Las variables independientes incluidas en los modelos de regresión se presentan en la tabla 1.

Tabla 1. Descripción de las variables

Correlativo	Variables Independientes	Tipo de Variable	Descripción de la Variable
1	Género	Nominal	Esta es una variable dummy codificada como 1 si el estudiante es niño y 0 si es niña.
2	Área	Nominal	Esta indica si el niño asiste a una escuela del área rural o del área urbana. Es una variable dummy codificada con 1 si el estudiante asiste a una escuela urbana y 0 si asiste a una escuela rural.
3	Etnia	Nominal	Esta es una variable dummy que indica la etnia de los estudiantes. Fue codificada como 1 si los estudiantes no son indígenas y 0 si son indígenas.
4	Tiempo invertido en lectura	Nominal/Ordinal	Tiempo 1: Variable dummy codificada 1 si el estudiante reportó leer una hora al día y 0 para el resto.
5			Tiempo 2: Variable dummy codificada 1 si el estudiante reportó leer dos horas al día y 0 para el resto.
6			Tiempo 3: Variable dummy codificada 1 si el estudiante reportó leer más de dos horas al día y 0 para el resto.
			El grupo de referencia los constituyen los estudiantes que reportaron leer menos de una hora al día.
7	Estatus Socioeconómico de los Estudiantes (SES)	Intervalo	Estatus socioeconómico de los estudiantes, es una variable continua construida por medio del análisis factorial.
8	Edad del estudiante	Intervalo	Aunque técnicamente hablando la variable es ordinal, ella puede ser tratada como variable de intervalo debido a que, por ejemplo, un estudiante con 14 años de edad tiene el doble de edad que un estudiante de 7 años.

Los puntos del rendimiento de lectura se constituyen como la variable dependiente del estudio. Es una variable continua con media de 500 puntos y desviación estándar de 100.

4.3. Diseño del estudio

El diseño del estudio se considera como correlacional puesto que los modelos que los datos han sido tomados de una base preexistente para construir los modelos de regresión. Esto significa que no se tiene control sobre el diseño original para la colección de los datos. Sin embargo, lo anterior no es una limitante en el uso de los datos, ya que al utilizar variables de control se genera un contrapeso generado por la no aleatoriedad.

V. Resultados

5.1. Supuestos del modelo

Al observar los Q-Q plot, las distribuciones de los residuales en los modelos tuvo un pequeño problema de la cola superior de la distribución. Sin embargo, los residuales de los histogramas siempre mostraron una aproximación a la distribución normal. De cualquier manera el teorema del límite central permitió mantener la distribución de normalidad debido al tamaño de la muestra. Respecto de la homoscedasticidad, se pudo determinar que el rango de los residuos se mantuvo constante a través de los niveles de los valores ajustados del modelo. Además, las distribuciones de los residuos se mantuvieron aproximadamente 50-50 arriba y debajo de las líneas de referencia. Todo esto se consideró como evidencia suficiente para mantener los supuestos de normalidad, homoscedasticidad y linealidad.

Al construir una matriz de correlaciones, la cual se observa en la tabla 3 del apéndice, se observó que la correlación entre las variables independientes tiempo 1 y tiempo 2 fue de -0.253, siendo el mayor valor de correlación. Por tal motivo, se consideró que no existió una colinearidad que causara problemas en los modelos.

5.2. Valores extremos y discrepantes

Un total de 425 valores “leverage” fueron considerados como extremos y 1 378 fueron clasificados con discrepantes. Es por ello que se examinó en las variables demográficas de los estudiantes para determinar la existencia de evidencia que permitiera verificar si dichos valores eran parte de otra población. Se encontró que 297 valores extremos y 850 discrepantes eran estudiantes con edades por arriba y por debajo de 12 años respectivamente. De esa cuenta, se desarrolló un análisis de sensibilidad entre los modelos con los datos sujetos al análisis y sin los datos extremos y discrepantes. El análisis mostró pequeños

cambios en las pendientes, sin embargo, al observar sus intervalos de confianza se comprobó se traslapaban. Además los valores de probabilidad (significancia estadística) también presentaron cambios irrelevantes manteniendo el estado original de la significancia; es decir, no se dio ningún cambio en las pendientes de significativa a no significativa y viceversa.

Los resultados anteriores son una evidencia que cualquiera de los dos modelos pudo brindar la misma información, aunque el impacto sobre el poder predictivo del modelo R^2 , sí cambio de 0.14 a 0.163. Con base en toda esta evidencia, se decidió utilizar los modelos sin valores extremos y discrepantes puesto que proveen la misma información; además ayudan a obtener un poder de predicción más alto en el modelo, lo cual es importante para la toma de decisiones en política educativa.

5.3. Modelos de Regresión

La tabla 2 muestra los tres modelos desarrollados en este estudio. Los coeficientes de regresión con dos asteriscos indican que la significancia es extrema y su valor de probabilidad es menor que 0.05, 0.01 y 0.001. Los coeficientes con un asterisco indican significancia menor que 0.05. La tabla también muestra los intervalos de confianza (solo para el modelo 3), los valores de R^2 y los valores críticos del test “F incremental”.

Tabla 2. Modelos reducidos y finales

Variable	MODELO 1	MODELO 2	MODELO 3	
	B	B	B	95% CI
Constante	450.72	608.53	607.89	[596.33 619.45]
Género	-3.74**	-2.58*	-2.88**	[-4.82 -0.93]
Área	36.91**	27.82**	27.88**	[25.86 29.90]
Etnia	51.10**	41.54**	39.34**	[37.06 41.61]
Tiempo 1	4.14**	4.30**	4.49**	[2.40 6.58]
Tiempo 2	-2.62	-1.61	-2.42	[-6.12 1.28]
Tiempo 3	-9.57**	-8.37**	-9.49**	[-13.77 -5.21]
SES		12.28**	19.65**	[17.48 21.82]
Edad		-11.39**	-10.99**	[-11.85 -10.12]
SES*Área			-6.60**	[-8.92 -4.28]
SES*Etnia			-4.79**	[-6.81 -2.77]
R Cuadrado	0.100	0.139		0.163
Cambio de R Cuadrado		0.04		0.02
F incremental		665.92		1.14
F Crítica		2.99		1.07

* significance at 5%. ** significance at 5% and 1% and 0.1%.

VI. Discusión

El poder de predicción de los modelos aumentó en el modelo 2 cuando SES y la edad de los estudiantes fueron introducidas como variables predictoras. El valor de R^2 incrementó de 0.10 a 0.14 y el test “F incremental” tuvo significancia estadística lo cual indicó que al incluir esas variables, hay un incremento en la varianza explicada mucho más allá que las variables incluidas en el modelo 1. Esto quiere decir que estas dos variables ayudan a explicar la varianza del rendimiento de lectura en 4%. Y en consecuencia, las variables de interés más las variables control, explican un total de 14% en la variación del rendimiento de lectura.

Respecto a la interpretación de las pendientes del modelo 2, se puede decir que la variable tiempo 2 no fue significativa, pero el resto de las variables fueron extremadamente significativas estadísticamente. Por tal razón, el SES y la edad de los estudiantes presentaron un impacto en el rendimiento de lectura. Esto significa que, cuando SES aumenta una unidad, el rendimiento promedio de los estudiantes aumenta en 12.28 puntos en promedio, manteniendo el resto de variables constantes. Por otro lado, cuando la edad aumenta un año, el rendimiento promedio de lectura disminuye en 11.39 puntos, manteniendo las otras variables constantes. El resto de variables del modelo 2 son variables control por lo cual no son interpretadas aquí. Sin embargo, el lector puede interpretarlas de forma similar a SES y edad.

En el modelo 3 se incluyeron las interacciones entre SES y área, así como la interacción entre SES y etnia. La inclusión de dichas interacciones aumentó el valor de R^2 de 0.14 a 0.16, lo cual incrementó el poder predictivo del modelo en 2%. El test “F incremental” fue significativo al 5%. De esa forma se puede concluir que las dos interacciones explican 2% de la varianza del rendimiento de lectura, más allá de la explicación atribuible a los efectos principales de género, área, etnia, tiempo invertido en leer, SES y edad.

En el modelo tres la pendiente de la variable tiempo 2 tampoco no presentó significancia estadística, mientras que el resto de las pendientes fueron extremadamente significativas. Con respecto a las interacciones de SES y área, la relación entre dicha interacción y el rendimiento de lectura fue diferente entre los estudiantes que asisten a escuelas del área rural y estudiantes que asisten a escuelas del área urbana. Por otro lado, la interacción entre SES y etnia indicó que la relación entre SES y el rendimiento de lectura es diferente entre estudiantes que no son indígenas y quienes sí lo son.

VII. Conclusiones

Al incluir las variables SES y edad en el modelo 1, formando así el modelo 2, la varianza explicada aumentó 4% lo cual aumenta el poder predictivo del modelo de regresión. El test “F incremental” validó dicha afirmación. El efecto promedio de SES sobre el rendimiento de lectura fue de 12.28 y el de edad fue de -11.39 en el modelo 2. Al incluir las interacciones entre SES y área, y SES y etnia en el modelo 2, lo cual formó el modelo 3, incrementó el poder de predicción del modelo en 2%. Además, el efecto de dichas interacciones fueron de -6.60 y -4.79 respectivamente. Esto significa que el efecto de SES sobre el área no es el mismo en el área rural que en el área urbana; de igual manera, el efecto de SES sobre el rendimiento de lectura tampoco es el mismo para los estudiantes no indígenas que para los indígenas. Toda esta evidencia ayuda a concluir que SES, edad y las interacciones, son variables importantes que aumentan el poder predictivo del modelo. Adicionalmente, se puede concluir que esas variables tienen un efecto estadísticamente significativo, sobre el rendimiento de lectura de los estudiantes de sexto grado en Guatemala.

VIII. Implicación de política

Puesto que SES fue significativo, es importante que no solo el Ministerio de Educación sino el Gobierno de Guatemala puedan implementar políticas educativas que ayuden a paliar la situación económica de los estudiantes. En tal sentido Willms (2006) explica de forma completa cinco políticas de intervención que involucran al SES y el aprendizaje de los estudiantes, así pues el Ministerio de Educación y el Gobierno de Guatemala podrán analizar y seleccionar las políticas que mejor se adapten a las condiciones del país. Los programas sociales, como el de transferencias condicionadas, podrían ayudar a mejorar la situación socioeconómica de los estudiantes.

La edad demostró ser una variable que disminuye el rendimiento escolar de lectura. Esto incluye estudiantes con sobreedad y estudiantes con extraedad, variables que son definidas en cierto sentido por la repitencia y por el acceso tarde a la escuela. En tal sentido, las autoridades educativas deben crear programas remediales para repitentes y estrategias docentes que ayuden a disminuir la repitencia, así como crear estrategias que incentiven a los padres a enviar a los niños a la escuela en la edad indicada.

Puesto que la interacción del SES y el área de residencia indica que el SES tiene un efecto diferente en el área rural que en el área urbana (el efecto es mucho mayor el área rural), el Ministerio de Educación deberá considerar aplicar políticas educativas diferenciadas en cada una estas áreas.

El caso de la interacción de SES y etnia también indica que el efecto que tiene SES sobre los no indígenas es diferente al efecto que tiene sobre los indígenas: el impacto es mayor en los estudiantes indígenas. Por tal razón, las políticas educativas que deberá implementar el Gobierno de Guatemala a través del Ministerio de Educación, deben ser focalizadas a la población indígena pobre; estas políticas, sean las que sean, deben ser diferenciadas a la políticas utilizadas con los estudiantes ladinos pobres.

IX. Referencias

Esquivel, F. (2006). *Situación del Sistema Educativo Guatemalteco*. Recuperado el 07 de noviembre de

<http://siteresources.worldbank.org/GUATEMALAINSPANISHEXT/Resources/500796-1166830633691/Situacion2004Inf.completo.pdf>

Instituto Nacional de Estadísticas de Guatemala. (2006). *Encuesta de Condiciones de Vida de Guatemala*. Recuperado el 06 de diciembre de

http://www.ine.gob.gt/np/encovi/ENCOVI2006/Resultados_Nacionales.pdf

Ministerio de Educación de Guatemala (2007). *Plataforma de Información Social Integrada*. Dirección de Planificación Educativa, Recuperado el 01 de diciembre de <http://200.6.193.201/businessobjects/enterprise115/desktoplaunch/InfoView/logon/logoff.do>

Ministerio de Educación de Guatemala. (2008). *Anuario Estadístico*. Dirección de Planificación Educativa. Recuperado el 01 de diciembre, de

<http://www.mineduc.gob.gt/estadistica/2007/anuario/>

Willms, J.D. (2006). *Learning Divides: Ten Policy Questions About the Performance and Equity of Schools and Schooling Systems*. UNESCO, Institute for Statistics, Montreal. UIS WorkingPaper No. 5.

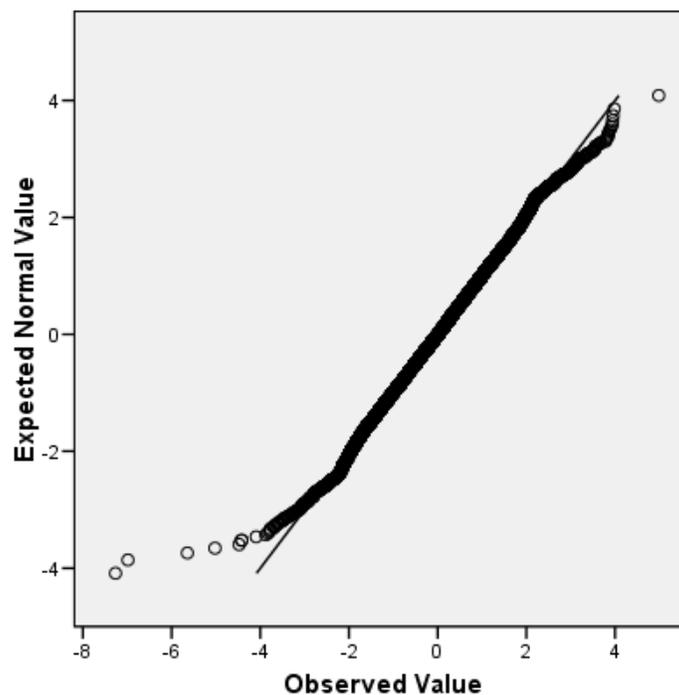
APÉNDICE A

Tabla 3. Matriz de correlaciones par a las variables explicativas

		STUDENTS GENDER	SCHOOL IS LOCATED IN URBAN OR RURAL AREA	ETHNICITY INDIGENOUS AND NON INDIGENOUS	TIME_READ 1	TIME_READ 2	TIME_READ 3
STUDENTS GENDER	Pearson Correlation	1	-.030(**)	-.047(**)	-.022(**)	-.018(**)	-.012(*)
	Sig. (2-tailed)		0	0	0	0.002	0.039
	N	29412	29412	29412	29412	29412	29412
SCHOOL IS LOCATED IN URBAN OR RURAL AREA	Pearson Correlation	-.030(**)	1	.076(**)	-0.005	-.016(**)	-.019(**)
	Sig. (2-tailed)	0		0	0.377	0.008	0.001
	N	29412	29412	29412	29412	29412	29412
ETHNICITY INDIGENOUS AND NON INDIGENOUS	Pearson Correlation	-.047(**)	.076(**)	1	0.007	.013(*)	.021(**)
	Sig. (2-tailed)	0	0		0.221	0.022	0
	N	29412	29412	29412	29412	29412	29412
TIME_READ1	Pearson Correlation	-.022(**)	-0.005	0.007	1	-.253(**)	-.217(**)
	Sig. (2-tailed)	0	0.377	0.221		0	0
	N	29412	29412	29412	29412	29412	29412
TIME_READ2	Pearson Correlation	-.018(**)	-.016(**)	.013(*)	-.253(**)	1	-.077(**)
	Sig. (2-tailed)	0.002	0.008	0.022	0		0
	N	29412	29412	29412	29412	29412	29412
TIME_READ3	Pearson Correlation	-.012(*)	-.019(**)	.021(**)	-.217(**)	-.077(**)	1
	Sig. (2-tailed)	0.039	0.001	0	0	0	
	N	29412	29412	29412	29412	29412	29412

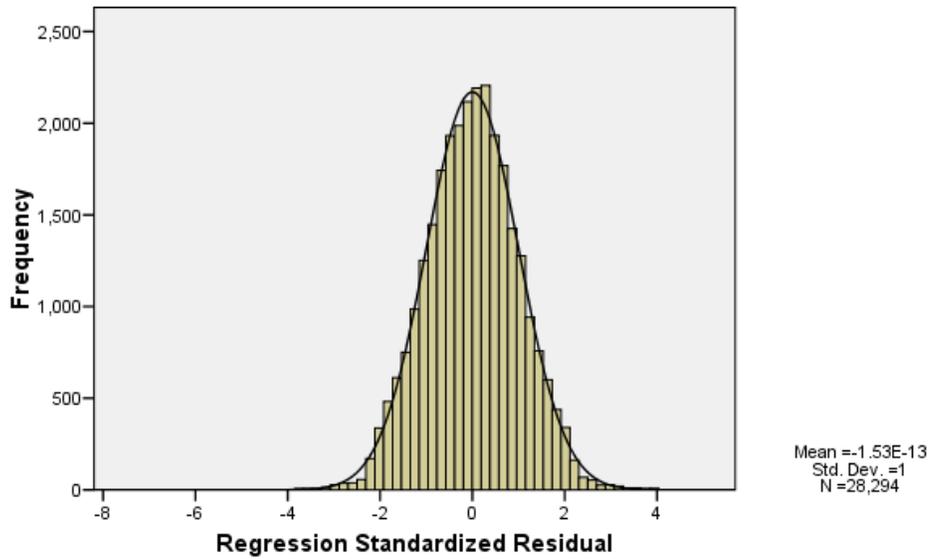
APÉNDICE B

Figura 1. Plot Q-Q Normal de Residuos Studentizados



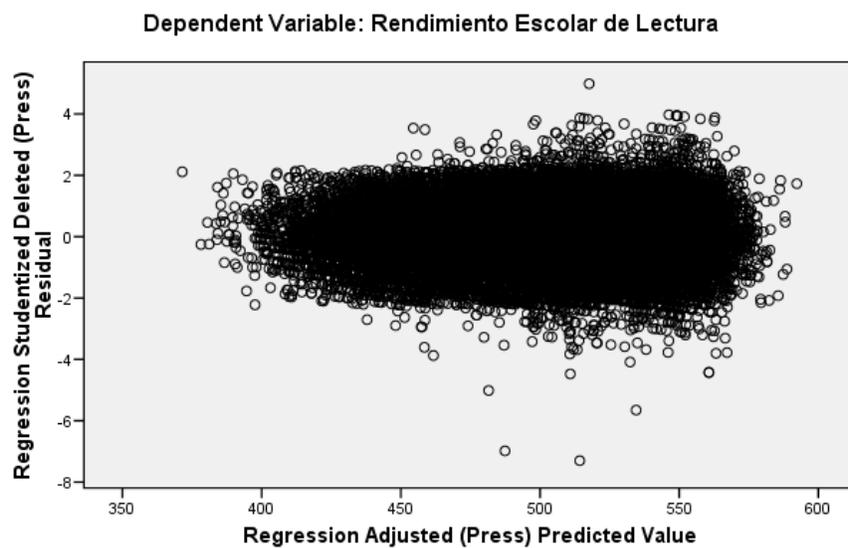
APÉNDICE C

Figura 2. Histogramas de Residuo



APÉNDICE D

Figura 3. Residuos Estandarizados vrs. Valores Ajustados



APÉNDICE E

Tabla 4. Modelo final con y sin valores extremos y discrepantes

Variable	Modelo final con valores extremos y discrepantes				Modelo final con valores extremos y discrepantes			
	B	p-value	95% CI		B	p-value	95% CI	
Constante	610.22	0.00	597.92	622.51	607.89	0.00	596.33	619.45
Género	-2.53*	0.02	-4.64	-0.43	-2.88**	0.00	-4.82	-0.93
Área	27.99**	0.00	25.81	30.17	27.88**	0.00	25.86	29.90
Etnia	40.16**	0.00	37.74	42.58	39.34**	0.00	37.06	41.61
Tiempo 1	4.31**	0.00	2.04	6.58	4.49**	0.00	2.40	6.58
Tiempo2	-1.67	0.41	-5.65	2.31	-2.42	0.20	-6.12	1.28
Tiempo3	-8.66**	0.00	-13.16	-4.15	-9.49**	0.00	-13.77	-5.21
SES	18.92**	0.00	16.68	21.16	19.65**	0.00	17.48	21.82
Edad	-11.35**	0.00	-12.27	-10.43	-10.99**	0.00	-11.85	-10.12
SES*Área	-5.81**	0.00	-8.21	-3.40	-6.60**	0.00	-8.92	-4.28
SES*Etnia	-4.95**	0.00	-7.10	-2.81	-4.79**	0.00	-6.81	-2.77
R Cuadrado		0.140				0.163		
F		480.38				552.21		
Cambio R Cuadrado						0.023		
Cambio en F						71.83		