**Introducción**

La educación en Colombia no es ajena a los problemas de desigualdad estructurales que existen en el país. Éstos forman parte importante de los factores que reproducen la desigualdad social. Existe una conexión entre el nivel educativo y el ingreso, al igual que entre el nivel educativo de un individuo con el máximo nivel educativo alcanzado por sus padres. Existe transmisibilidad intergeneracional entre el nivel educativo de una persona con sus padres y en consecuencia en su nivel de ingreso (Esteban & Santiago, 2000). En la literatura sobre desigualdades en la educación es bien sabido que el desempeño de los estudiantes en ámbito académico está influenciado por factores externos socioeconómicos y culturales (Ángela, Andres, Carolina, & Jaime, 2017) y este factor se ve reflejado en los resultados de las pruebas estandarizadas colombianas de las pruebas Saber 11.

Desde la década de los 80, en Colombia se reglamentó el acceso a las instituciones de educación superior y desde eso los resultados de las pruebas saber 11 sirven a los estudiantes para acceder a distintas becas como la beca Andrés Bello para los 50 mejores del país o la Generación E, anteriormente conocida como Ser Pilo Paga. Los resultados de esta prueba no solo son diagnósticos del nivel educativo de un estudiante, sino que son trascendentes para el desarrollo posterior y el acceso a oportunidades académicas.

La desigualdad influye en dos direcciones en los estudiantes que presentan estas pruebas estandarizadas. Al presentarlas, aquellos quienes tienen condiciones socioeconómicas o culturales adversas suelen tener menores competencias para el examen y, en consecuencia, afectar, como se explicó anteriormente, las posibilidades de seguir con sus estudios universitarios. El propósito de este trabajo es dilucidar la influencia que tienen ciertas características socioeconómicas y culturales en el desempeño de los estudiantes en las pruebas saber ICFES 11.

**Variables a considerar**

En este trabajo se presentan tres regresiones usando tres formulaciones distintas de las variables. Los datos de resultados individuales para las Pruebas Saber de 2018 fueron promediados por colegio. A continuación detallamos las definiciones:

**ESTU\_TIENEETNIA:** Variable Dummy que indica la pertenencia de un individuo a una etnia. La pertinencia de esta variable en la consideración de la regresión es relativa a las diferencias socioeconómicas y culturales que existen en la población colombiana.

**FAMI\_ESTRATO\_VIVIENDA** Estrato al que pertenece el estudiante. Esta variable toma diferentes valores en función del estrato al que pertenece la residencia donde está radicado el estudiante.La función toma el valor de su estrato respectivo.

**FAMI\_PERSONASHOGAR:** Número de personas que conviven con el estudiante. Toma los valores 1, 3, 5, 7 o 9, correspondiente a las opciones 1 a 2, 3 a 4, 5 a 6, 7 a 8 o 9 o más personas.

**Conectividad:** De manera inclusiva, si el estudiante tiene computador o si cuenta con internet, toma el valor de 1.

**FAMI\_NUNLIBROS**: Variable de escala que muestra el número de libros en la casa. 10 si tiene de 0 a 10 libros. 25 si tiene de 11 a 25 libros, 50 si es la opción “26 a 100 libros” y por último 100 si es la opción “más de 100 libros”.

**ESTU\_DEDICACIONLECTURADIARIA** Esta variable se toma como una variable de escala. Se toma como grupos aproximados de 10 minutos (se selecciona un valor perteneciente al intervalo correspondiente): 0 si el estudiante no lee por entretenimiento, 3 si lo hace 0 a 30 minutos diarios, 6 si lee de 30 minutos a 1 hora, 12 si es entre una y dos horas y 18 si es más de dos horas.

**ESTU\_DEDICACIONINTERNET:** Variable de escala, de manera similar: 0 si no navega internet, 3 si navega 30 minutos o menos, 6 si lo hace entre 30 y 60 minutos, 12 si es entre 1 y 3 horas y 18 si es más de 3 horas.

**Alimentación:** Tomará el valor de 1 si se cumple que, por lo menos tres veces a la semana, el estudiante consume por lo menos dos de los tres grupos alimenticios considerados en la encuesta: leche y sus derivados, carnes y por último cereales frutas o legumbres.

**ESTU\_HORASASEMANATRABAJA:** Esta variable toma el valor de 1 si el individuo trabaja.

**COLE\_NATURALEZA:** Naturaleza del colegio.Toma valores de 1 cuando el colegio del estudiante es de carácter oficial o público y como 0 si el colegio es de carácter privado.

**COLE\_AREA\_UBICACION:** Ubicación del colegio. Toma valores de 1 cuando el colegio está ubicado en un entorno rural y 0 cuando está ubicado en el casco urbano.

**FAMI\_EDUCACIONPADRE, FAMI\_EDUCACIONMADRE y Educacion\_padres**

* Primera regresión: Se consideran aparte la educación máxima alcanzada por el padre y por la madre. En particular FAMI\_EDUCACIONPADRE toma el valor de 1 si el padre terminó el bachillerato y de 0 en caso contrario.
* Segunda regresión: Se considera una única variable Educacion\_padres que agrupa las dos variables anteriores con un operador lógico “o” inclusivo. Así, tomará el valor de 1 en caso de que alguno de los dos padres haya terminado del bachillerato.
* Tercera regresión: La variable Educacion\_padres ahora tomará valores enetros en el rango de 0 a 9, indicando el máximo nivel educativo alcanzado por alguno de los padres. Esta variable considera las opciones contempladas en la encuesta: ninguno, primaria incompleta, primaria completa, análogamente con bachillerato, educación técnica y educación profesional y por último posgrado.

Como se mencionó, los datos son agrupados y promediados por colegio. Esto provoca que las variables consideradas como dummy representen un porcentaje de la población estudiantil de un colegio particular que cumple la condición especificada. Para que el efecto de incrementar un 1% sea registrado en la regresión. Las variables porcentuales, originalmente con valores entre 0 y 1, se transforman al intervalo [0,100].

Aunque las variables consideradas son directamente derivadas de las consideradas por la encuesta hecha en las Pruebas Saber y su relevancia puede ser asumida a priori, veremos en los análisis de regresión la relevancia de estas variables en el contexto de un modelo lineal.

**Resultados**

El objetivo de nuestro a análisis entender el efecto cuantitativo de estas variables tanto socioeconómicas como culturales sobre los resultados de los estudiantes en las pruebas estandarizadas (Saber11). El ejercicio de las tres regresiones se hizo con intención de dejar manifiesto el hecho y el efecto que tiene la arbitrariedad de la definición de las variables al momento de plantear un modelo. El efecto más notorio fue notar que, como cabía prever, las variables de educación padre y educación madre eran altamente correlacionadas. Adicionalmente, el factor de inflación de varianza de la variable Educacion\_padres en la última regresión (donde está en escala de 0 a 9), al poder oscilar en un rango de valores, es mayor y manifiesta una mayor colinealidad que en las regresiones donde se define como dummy. Se muestran en el presente documento, por brevedad, los resultados únicamente de la tercera regresión. Como visualización inicial, la variable categórica de la naturaleza del colegio muestra una gran separación entre los datos. En el siguiente boxplot se muestran los boxplots de colegios públicos (0) y privados (1), indicando la distribución de puntaje promedio en ambas categorías. Esto es una primera muestra del foco de una posible política pública a desarrollar:

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Antes de correr la regresión, mostramos también una visualización de la matriz de correlaciones entre las variables consideradas. Las altas correlaciones resultan obvias en su mayoría, como la dedicación a internet y la conectividad. Vemos que la variable de Educacion\_padres muestra una alta correlación con varias variables, al igual que estrato de vivienda. Más allá de la causalidad entre éstas, veremos más adelante que en sí misma, la educación de los padres es explicable en términos de las demás con un ajuste de regresión alto.

Imagen que contiene Gráfico

Descripción generada automáticamente

Es de mencionar la extrañeza de los resultados. Al contrario de la visualización y la separación que ofrece tanto la ubicación del colegio (urbana o rural) como la naturaleza (público o privado), donde ambas variables se consideran dummies y se dan positivas las opciones de colegios urbano para la primera y si la institución es privada para la segunda. La regresión arroja coeficientes negativos para éstas. Adicionalmente, en la tercera regresión, el coeficiente asociado a la variable de estrato es también negativo. A continuación, mostramos los resultados de la regresión.

Texto

Descripción generada automáticamente

Ignorando los resultados relativos a naturaleza, ubicación y estrato, los coeficientes porcentuales de etnia, conectividad, alimentación y trabajo del estudiante se corresponden con las creencias a priori. De igual forma, las variables no porcentuales como dedicación a lectura e internet y número de personas en el hogar tienen efectos relevantes. Todas las variables son significativas en la regresión, pero como vemos en el plot de residuales, al igual que en el test de Breusch Pagan, la heteroscedasticidad es evidente.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente Texto

Descripción generada automáticamente

A su vez observamos la tabla con los factores de inflación de varianza correspondientes a cada variable.

Texto

Descripción generada automáticamente

Los resultados no pueden ser ignorados, pero antes de dar nuestras conclusiones al respecto de la regresión, veremos el ajuste que tiene la variable educación padres como variable explicada en función de las otras. Mostraremos solamente la tabla de resultados.

Imagen que contiene exterior, texto, monitor, tabla

Descripción generada automáticamente

**Conclusiones:**

Aunque podemos concluir que, en efecto, las variables socioeconómicas y culturales repercuten significativamente en el desempeño de los estudiantes en las pruebas Saber11, y la relación explícita entre la educación de los padres y todas las demás variables promedio de una institución educativa, los resultados no son los esperados, además de ser claramente contradictorios en los datos de naturaleza y ubicación de la institución.

Nuestro análisis de la falla de los coeficientes incluye la posibilidad alta de distorsión en la manipulación de los datos, pues estos están agrupados como valores promedio para cada institución, adicionalmente, la transformación de los promedios en valores porcentuales, la multicolinealidad presente, las variables dummy que rompen los supuestos del modelo lineal al no ser variables continuas y por último, la presencia de valores atípicos o outliers que distorsionen el cálculo de los coeficientes adecuados. Esto presenta la necesidad de hacer cálculos con estimadores robustos de la matriz de covarianzas y explorar mejores aproximaciones a la manipulación de los datos para agruparlos por institución educativa y así esclarecer posibilidades de planeación de política pública.

# Bibliografía

E. N., & S. G. (2000). Educaicón, movilidad social y "trampa de pobreza". *Coyuntura Social*.

Á. L., A. V., C. S., & J. S. (2017). Desigualdad de oportunidades en el sistema de educación pública en Bogotá, Colombia. *Lecturas de economía*.