Determinantes del rendimiento académico: un caso de estudio

Felipe Del Valle

09/10/2019

I) Introducción

Las diferencias de rendimiento académico entre géneros (Voyer and Voyer Susan D. 2014) ha sido un tema discutido extensivamente en las ciencias sociales. Estudios recientes (Marc Jackman and Morrain-Webb 2019) han mostrado, en condiciones particulares, que las mujeres tienden ligeramente a un mejor desempeño académico. Sin embargo, el resultado en pruebas parametrizadas no solo puede ser explorado ni explicado únicamente desde esta diferencia biológica. Entre los factores más relevantes al momento de predecir el rendimiento académico de una población están los socioeconómicos, como el nivel promedio de ingresos (Thomson 2018) o la formación educacional de los padres.

En este ejercicio se buscará aplicar las herramientas aprendidas en el curso BIO4022 para explorar una base de datos y obtener de ella información relevante para responder a una hipótesis. En este caso, se buscará determinar si existe o no un modelo lineal que permita explicar en este conjunto de observaciones, el rendimiento académico en pruebas de escritura, lectura y matemáticas. Lo anterior basado en posibles determinantes como el género, nivel de preparación de antes de las pruebas y el nivel académico de los padres.

II) Metodología y resultados

Origen de los datos

Para este ejercicio se trabajó con una base de datos disponible en https://www.kaggle.com/spscientist/students-performance-in-exams. Esta es de uso público y contiene datos que son relevantes para la pregunta de investigación.

Exploración de los datos

En primer lugar, ocupando las librerías de tidyverse (Wickham 2016b) se exploraron los datos para verificar cuáles eran los descriptores que se presentaban en las columnas obtenidas desde la base. A continuación se muestra el Chunk para estas instrucciones.

```
StudentsPerformance <- read_csv("StudentsPerformance.csv")
colnames(StudentsPerformance)
```

```
## [1] "gender"                       "race/ethnicity"
## [3] "parental level of education" "lunch"
## [5] "test preparation course"     "math score"
## [7] "reading score"                "writing score"
```

Para el interés de esta investigación, se ocuparán los siguientes: Género, scores de reading, writing y matemáticas; además de la preparación y el nivel de educación de los padres.

visualización de los descriptores de interés

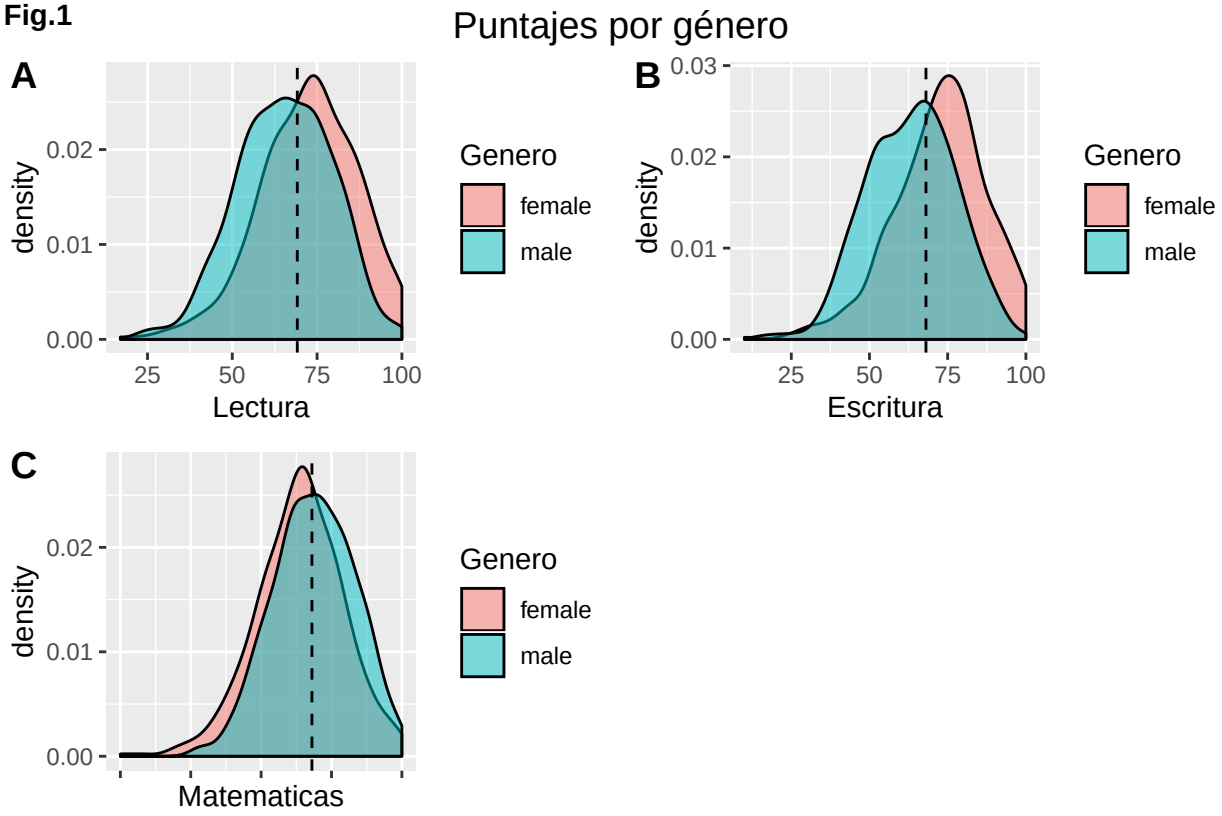
De manera posterior, se ocupó el paquete ggplot (Wickham 2016a) para ordenar los datos visualmente y entender el tipo de distribución que muestran para las pruebas de lectura, escritura y matemáticas. Esto se resume en los siguientes gráficos.

```
## Warning: Ignoring unknown parameters: binwidth
```

```
## Warning: Ignoring unknown parameters: binwidth

## Warning: Ignoring unknown parameters: binwidth
```

Fig.1



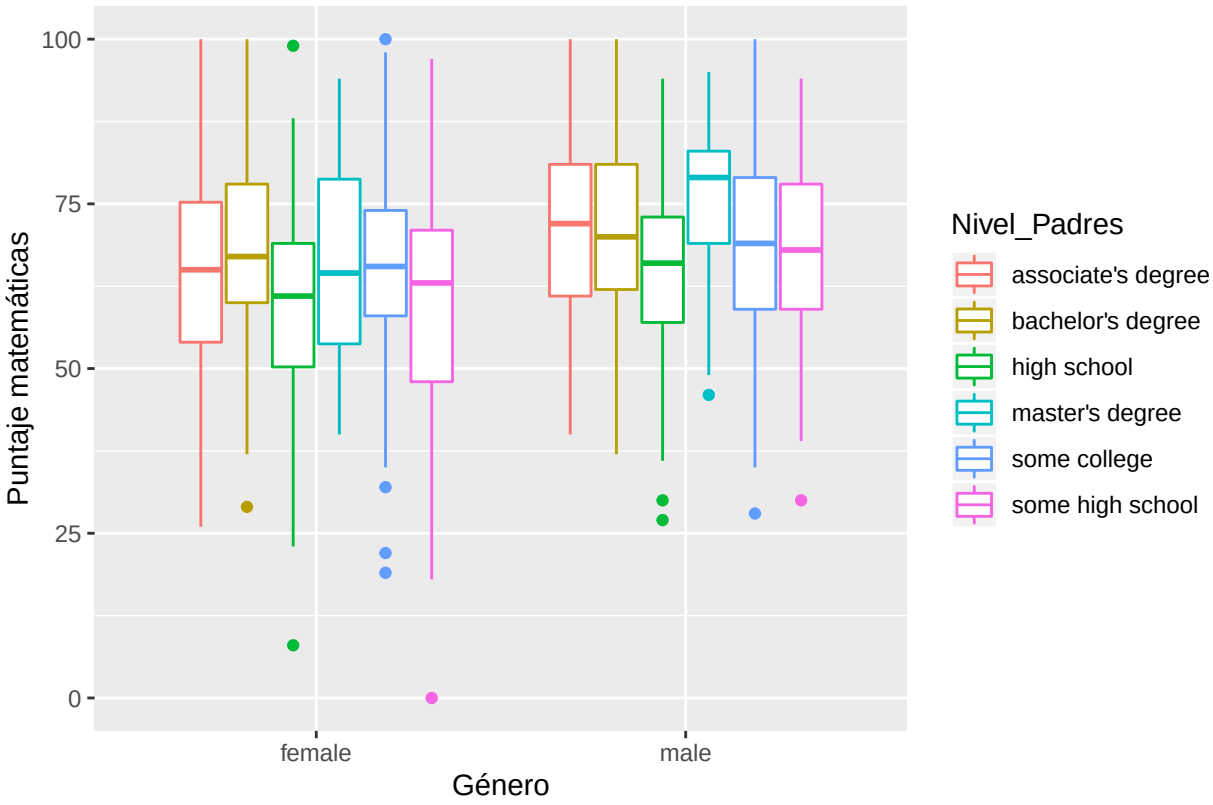
Distribución de los puntajes en las respectivas pruebas (Verde=Hombres, Sandía=Mujeres)

Según lo obtenido, se puede apreciar que la distribución en frecuencia de los puntajes se asemeja a una curva normal. A esto se suma que las medias para cada genero están prácticamente superpuestas. Esto sugiere que no existirían mayores diferencias entre los géneros para cualquier evaluación en este caso.

Diferencias entre géneros por prueba y nivel de los padres

Luego, se evaluaron visualmente los puntajes por género para comprender cuál es su posible relación con el nivel educacional de los padres. En primer lugar se ilustran los datos obtenidos para los puntajes en matemáticas.

Fig.2: Puntaje en matemáticas según género y nivel de los padres



Se aprecia que al igual que en lo que respecta a las diferencias por género, el nivel educacional de los padres no estaría afectando a este parámetro.

Resultados para escritura y lectura

Siguiendo la misma lógica anterior se analizaron los resultados para las pruebas de lectura y escritura, obteniéndose los siguientes gráficos.

Fig.3: Puntaje en lectura según género y nivel de los padres

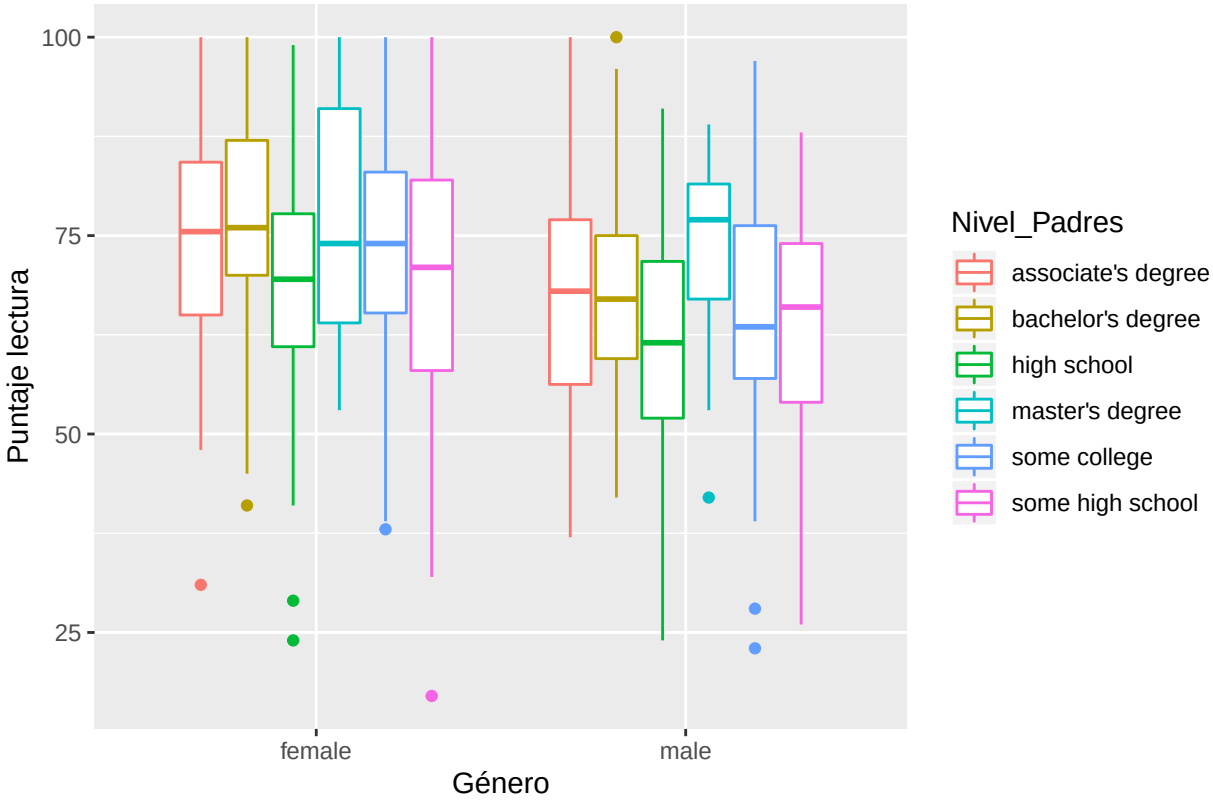
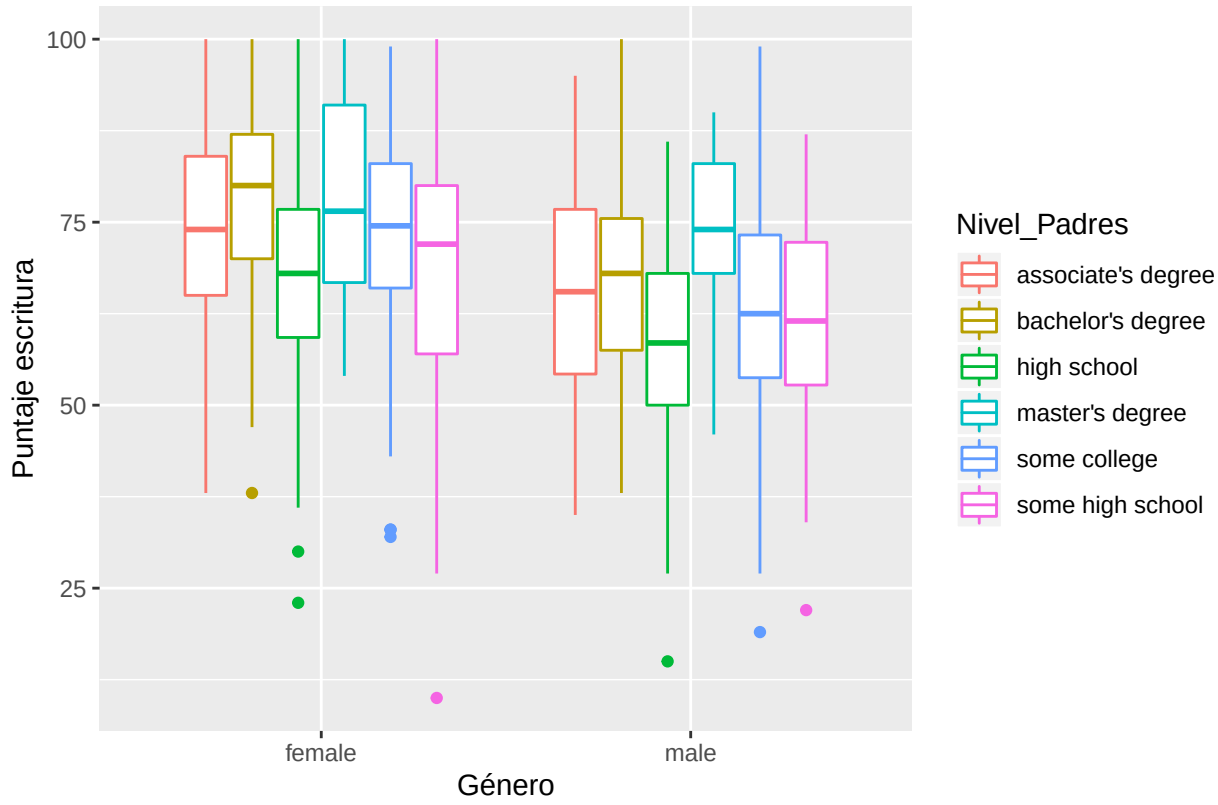


Fig.4: Puntaje en escritura según género y nivel de los padres



Sorprendentemente, hasta ahora, los determinantes de género y de preparación educacional de los padres parecen no tener un efecto notorio sobre la distribución de los puntajes para todas las pruebas realizadas.

Elección de un modelo

Para comprender si estos resultados se ajustan a algún modelo que pueda predecir el rendimiento académico según sus determinantes, se probarán los siguientes. Se verificará si el rendimiento está explicado por el nivel de los padres y el género; considerando un modelo lineal. Se resume en tablas el valor de r cuadrado y AIC obtenido para este caso.

$$Rendimiento = NivelPadres + Género$$

Table 1: Modelo 1 Lectura

r.squared	p.value	df	AIC
0.1	0	7	8110.42

Table 2: Modelo 1 Escritura

r.squared	p.value	df	AIC
0.15	0	7	8131.37

Table 3: Modelo 1 Matemáticas

r.squared	p.value	df	AIC
0.06	0	7	8225.58

El modelo de lectura tiene un r cuadrado de 0.1, el de escritura uno de 0.15 y el de matemáticas uno de 0.06

Elección de un segundo modelo

Luego, se probó un segundo modelo lineal que busca explicar el rendimiento mediante la preparación antes de las pruebas y el género de los estudiantes.

$$Rendimiento = Preparación + Género$$

Table 4: Modelo 2 Lectura

r.squared	p.value	df	AIC
0.1	0	7	8108.97

Table 5: Modelo 2 Escritura

r.squared	p.value	df	AIC
0.15	0	7	8131.37

Table 6: Modelo 2 Matemáticas

r.squared	p.value	df	AIC
0.06	0	7	8226.47

El modelo de lectura tiene un r cuadrado de 0.1, el de escritura uno de 0.16 y el de matemáticas uno de 0.06

III) Conclusiones y discusión

Considerando lo analizado en este trabajo, se puede concluir que un modelo lineal para explorar las diferencias en puntajes de estas pruebas parametrizadas no es suficiente y debería considerar un análisis más robusto y probablemente con otros indicadores. Esto queda demostrado por los altos indicadores de AIC y bajos de r cuadrado en los casos probados. De la misma manera, es posible también discutir que la transparencia de esta base de datos no permite saber de dónde se obtuvieron y si son o no ficticios. Finalmente, debería tratrse estos datos con otro tipo de modelo que se adecue a estudios anteriores para encontrar un buen predictor.

Referencias

Marc Jackman, W., and Judith Morrain-Webb. 2019. "Exploring gender differences in achievement through student voice: critical insights and analyses." *Cogent Education.* doi:10.1080/2331186x.2019.1567895.

Thomson, Sue. 2018. "Achievement at school and socioeconomic background—an educational perspective." *Npj Science of Learning.* doi:10.1038/s41539-018-0022-0.

Voyer, Daniel, and D. Voyer Susan D. 2014. "Gender differences in scholastic achievement: A meta-analysis." *Psychological Bulletin.* doi:10.1037/a0036620.

Wickham, Hadley. 2016a. *ggplot 2: Elagant graphics for data analysis.* doi:10.1007/978-0-387-98141-3.

———. 2016b. "tidyverse: Easily Install and Load 'Tidyverse' Packages."