Taller 1. Econometría Avanzada

Daniel Ricardo Amaya Alba

Table of contents

1 Primer Ejercicio

Desde la perspectiva del capital humano, estudios como el de Cunha & Heckma (2007) encuentran que las intervenciones durante los primeros años de escolaridad son fundamentales para el desarrollo temprano de habilidades y, con ello, para la trayectoria de aprendizaje posterior de los individuos. Una forma de evaluar el éxito de estas intervenciones es mediante el nivel de habilidades cognitivas que desarrollan los estudiantes, dado que estas capacidades constituyen la base de cualquier aprendizaje futuro.

En este contexto, los gobiernos suelen mostrar interés por los programas de digitalización, ya que prometen acelerar el desarrollo del razonamiento y facilitar la adaptación al mundo moderno. Con base en lo anterior, la pregunta de investigación que guiará este ejercicio, inspirada en Cristia et al. (2012), es: ¿cuál es el efecto de participar en un programa de tecnología educativa sobre el desarrollo de habilidades cognitivas de razonamiento abstracto?

Para responder esta pregunta, usted analizará una política implementada por la Secretaría de Educación de Macondo en conjunto con el colegio Sonrisas, ubicado en una zona rural del país. La iniciativa buscaba transformar el modelo educativo tradicional mediante la entrega de laptops diseñadas específicamente para el aprendizaje autodidacta en entornos con limitaciones severas de infraestructura. El programa se enfocó exclusivamente en estudiantes de tercer grado de primaria, nivel que en este colegio cuenta con dos salones. El programa seleccionaba un salón para convertirlo en Aula Tecnológica.

Los estudiantes asignados al Aula Tecnológica realizaban actividades guiadas tres veces por semana utilizando las laptops. Estas actividades se desarrollaban únicamente durante el horario de clases, las cuales fueron adaptadas para incluir ejercicios lógicos y juegos de asociación visual orientados a fortalecer el razonamiento abstracto. Cada actividad estaba vinculada con el tema correspondiente a la clase del día. Adicionalmente, los estudiantes no podían llevar las laptops a sus hogares. Los estudiantes del otro salón, por su parte, continuaron con sus clases de manera habitual.

Considere una muestra de N estudiantes observados en un diseño de corte transversal. Defina Y_i como el puntaje del estudiante i en un test de habilidades cognitivas de razonamiento abstracto. Este test se aplica anualmente a nivel nacional a estudiantes de tercer grado, está definido por el Ministerio de Educación y sus puntajes oscilan entre 0 y 100. Por otro lado, Di es una variable dicotómica que toma el valor de 1 si el estudiante i fue asignado al Aula Tecnológica y 0 en caso contrario.

Para comenzar, suponga que el mecanismo de asignación al tratamiento es desconocido; es decir, no se conocen los criterios que determinaron la elección de un salón sobre el otro ni la forma en que los estudiantes son asignados a cada salón.

1.1 Punto1

Describa el problema de estimación utilizando el lenguaje de resultados potenciales

a) Teniendo en cuenta el contexto presentado, defina formalmente los resultados potenciales $Y_i(1)$ y $Y_i(0)$. Luego, describa el problema de inferencia causal en este contexto.**

R/: En este caso el resultado potencial $Y_i(1)$ se define como el puntaje que obtendría el estudiante i dado que fue asigando al aula tecnologica. Por su parte el resultado potencial $Y_i(0)$ es el resultado que obtendría el estudiante i dado que no fue asignado al aula tecnologica.

Para recuperar el efecto causal de la participación en las aulas tecnologicas en el puntaje obtenido por el estudiante i, nos gustaria ver la dferencia en sus resultados potenciales, es decir de sus puntajes dado que participio en el aula y que no participo en el aula:

$$\tau_i = Y_i(1) - Y_i(0)$$

Sin embargo, acá nos encontramos con el problema de la inferencia causal. En nuestra muestra N para el estudiante i solo se puede observar uno de los dos caminos, o el estudiante fue asignado ($D_i = 1$) o no fue asignado ($D_i = 0$). Es decir, lo que realmente tenemos en nuestra base de datos es:

$$Y_i = Y_i(1)D_i + (1 - D_i)Y_i(0)$$

De tal manera que el resultado observado Y_i para el estudiante con $D_i = 1$ sera su resultado potencial dado que participó $Y_i(1)$ y para el estudiante con $D_i = 0$ observaremos su resultado potencial dado que no participó $Y_i(0)$. Pero nunca seremos capaces de ver para el mismo estudiante i su resultado dado que participo $D_i = 1$ y su contrafactual $D_i = 0$.

Es por lo anterior que para trabajar con nuestra muestra utilizaremos no el resultado del inddividuo i sino que trabajaremos con el promedio de los resultados potenciales, es decir buscaremos estumar los efectos promedio, bien sea el efecto promedio del tramaiento ATE, el efecto promedio del tratamiento en los tratados ATT y el efecto promedio del tratamiento en los no tratados ATU.

b) Define en lenguaje matemático el ATE, el ATT y el ATU. Interprete cada uno de estos conceptos de acuerdo al contexto del caso

R/:

El ATE se define como el efectro promedio del tratamiento:

$$\tau_{ATE} = E[Y_i(1) - Y_i(0)]$$
$$\tau_{ATE} = E[Y_i(1)] - E[Y_i(0)]$$

Que para el caso de nuestra investigación representa la diferencia entre el puntaje promedio en el test de habilidad cognitivas si toda la población de estudiantes de tercer grado hubieran asistido a las aulas tecnológicas, frente al promedio si ninguno hubiera asisitido a dichas aulas.

El ATT se define como el efectro promedio del tratamiento en los tratados:

$$\tau_{ATT} = E[Y_i(1) - Y_i(0)|D_i = 1]$$
$$\tau_{ATT} = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1]$$

En este caso representa la diferencia en el puntaje promedio del test de habilidades cognitivas del salon que fue asignado a las aulas tecnológicas y el resultado que esos mismos estudiantes

hubieran obtenido si no hubieran sido asignados. Es decir, en este caso estamos restringiendo el efecto promedio al grupo que fue tratado.

El *ATU* se define como el efectro promedio del tratamiento en los no tratados:

$$\tau_{ATU} = E[Y_i(1) - Y_i(0)|D_i = 0]$$
$$\tau_{ATU} = E[Y_i(1)|D_i = 0] - E[Y_i(0)|D_i = 0]$$

En este caso representa la diferencia en el puntaje promedio del test de habilidades cognitivas del salon que no fue asignado a las aulas tecnológicas y el resultado que esos mismos estudiantes hubieran obtenido si hubieran sido asignados. Es decir, en este caso estamos restringiendo el efecto promedio al grupo que no fue tratado.

1.2 Punto2

Usted se encuentra interesada en estimar el siguiente efecto causal:

$$\tau = E[Y_i(1)] - E[Y_i(0)]$$

Para esto plantea una diferencia de medias ingenua

$$\tau_{naive} = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

a). Muestre formalmente que:

$$\tau_{naive} = ATT + E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]$$

R/:

$$\tau_{naive} = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$
$$\tau_{naive} = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] + E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 1]$$
$$\tau_{naive} = \underbrace{E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1]}_{ATT} + E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]$$
$$\tau_{naive} = ATT + E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]$$

b). Muestre formalmente que:

$$\tau_{naive} = ATU + E[Y_i(1)|D_i = 1] - E[Y_i(1)|D_i = 0]$$

R/:

$$\begin{aligned}
\tau_{naive} &= E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \\
\tau_{naive} &= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] + E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 1] \\
\tau_{naive} &= \underbrace{E[Y_i(1)|D_i = 0] - E[Y_i(0)|D_i = 0]}_{ATU} + E[Y_i(1)|D_i = 1] - E[Y_i(1)|D_i = 0] \\
\tau_{naive} &= ATU + E[Y_i(1)|D_i = 1] - E[Y_i(1)|D_i = 0]
\end{aligned}$$

c). Muestre formalmente que:

$$ATE = \pi ATT + (1 - \pi)ATU$$

R/: Esta demostración es diferente a las anteriores, es menos intuitiva. Sin embargo, podemos aplicar la ley de esperanzas iteradas a la definición de ATE:

$$\begin{aligned}
\tau_{ATE} &= E[Y_i(1)] - E[Y_i(0)] \\
\tau_{ATE} &= E[E[Y_i(1) - Y_i(0)|D_i]] \\
\tau_{ATE} &= E[Y_i(1) - Y_i(0)|D_i = 1]P(D_i = 1) + E[Y_i(1) - Y_i(0)|D_i = 0]P(D_i = 0)
\end{aligned}$$

Donde:

$$\tau_{ATE} = \underbrace{E[Y_i(1) - Y_i(0)|D_i = 1]}_{ATT} \cdot \underbrace{P(D_i = 1)}_{\pi} + \underbrace{E[Y_i(1) - Y_i(0)|D_i = 0]}_{ATU} \cdot \underbrace{P(D_i = 0)}_{(1-\pi)}$$

Interpretando cada expresión:

a).

$$\tau_{naive} = ATT + \underbrace{E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]}_{\text{Selection Bias}}$$

Esta expresión indica que el τ_{naive} recupera el efecto causal sobre los tratados (ATT) si y solo si el sesgo de selección es igual a cero. Esto ocurre cuando el resultado promedio contrafactual del grupo tratado $E[Y_i(0)|D_i = 1]$ es idéntico al resultado promedio observado del grupo de control $E[Y_i(0)|D_i = 0]$. Para que esto se cumpla la composición de los salones debería ser homogénea o el proceso de asignación debería haber sido aleatorio, garantizando que no existan diferencias pre-existentes entre ambos grupos.

b).

$$\tau_{naive} = ATU + \underbrace{E[Y_i(1)|D_i = 1] - E[Y_i(1)|D_i = 0]}_{\text{Heterogeneity Bias}}$$

Por tro lado esta expresión nos indica que el τ_{naive} recupera el efecto promedio causal sobre los no tratado (ATU) si y solo si el sesgo de heterogeneidad en resultados del tratamiento es igual a cero. Esto ocurre cuando el resultado promedio de los estudiantes tratados $E[Y_i(1)|D_i = 1]$ es igual al resultado promedio que hubieran obtenido los estudiantes del grupo de control si se les hubiera asignado el tratamiento $E[Y_i(1)|D_i = 0]$. Esto quiere decir que el efecto de las computadoras sería el mismo, independientemente dle salon que hubiera sido seleccionado.

c).

$$ATE = \pi \cdot ATT + (1 - \pi) \cdot ATU$$

Por úlimo esta expresión nos dice que el efecto promedio del tratamiento es un promedio ponderado del efecto promedio en los tratado ATT y en los no tratados ATU.

¿Es cierto que si $\tau_{naive} = ATT$, entonces $\tau_{naive} = ATU$?

Esta afirmación no es cierta. Como vimos, el $\tau_{naive} = ATT$ sí y solo sí el resultado promedio contrafactual del grupo tratado $E[Y_i(0)|D_i = 1]$ es idéntico al resultado promedio observado del grupo de control $E[Y_i(0)|D_i = 0]$, lo cual no nos esta brindando ninguna información al respecto de algún posible sesgo de heterogeneidad del tratamiento, es decir no nos garantiza que el efecto del tratamiento sea el mismo para ambos grupos. Por su parte, si el $\tau_{naive} = ATU$ solo sabemos que hay ausencia de sesgo de heterogeneidad, pero no tenemos certeza al respecto de la comparabilidad de los grupos (sesgo de selección).

Teniendo como base la descomposicón que se hace en libro Causal Inference: The Mixtape de Scott Cunningham podemos demostrar la relación entre estos parametros:

$$\begin{aligned}
\tau_{ATE} &= \pi.ATT + (1 - \pi).ATU \\
\tau_{ATE} &= \pi.ATT + (1 - \pi).ATU + (1 - \pi).ATT - (1 - \pi).ATT \\
\tau_{ATE} &= ATT.\underbrace{(\pi + 1 - \pi)}_{1} - (1 - \pi).ATT + (1 - \pi).ATU \\
\tau_{ATE} &= ATT - (1 - \pi).ATT + (1 - \pi).ATU
\end{aligned}$$

Despejando ATT:

$$\tau_{ATT} = ATE + (1 - \pi).(ATT - ATU)$$

Remplazando la expresión que acabamos obtener de τ_{ATT} en nuestso esmitador τ_{naive}, obtenemos la descomposición completa del sesgo:

$$\tau_{naive} = ATT + \underbrace{E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]}_{\text{Selection Bias}}$$

$$\tau_{naive} = ATE + \underbrace{(1-\pi).(ATT - ATU)}_{\text{Heterogenity Bias}} + \underbrace{E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]}_{\text{Selection Bias}}$$

Por lo tanto en nuestro caso , es posible que el salón elegido fuera comparable inicialmente al de control (sin sesgo de selección), pero que las laptops tengan un retorno mayor en el salón tratado debido a factores específicos del grupo o del profesor. Por lo tanto para que $ATE = ATT = ATU$ necesitamos tener ausencia del sesgo de hterogeneidad y de selección.

1.3 Punto3

Para encontrar el efecto causal promedio de participar en el programa de tecnología educativa sobre el resultado en el test, considere el siguiente modelo lineal:

$$Y_i = \beta_0 + \tau D_i + U_i$$

Demuestre que el estimador $\hat{\tau}_{MCO}$ coincide con la diferencia de medias entre el grupo de tratados y el grupo de no tratados. Es decir, pruebe que $\hat{\tau}_{MCO} = \hat{\tau}_{DM}$, donde:

$$\hat{\tau}_{DM} := \frac{1}{N_1} \sum_{\{i:D_i=1\}} Y_i - \frac{1}{N_0} \sum_{\{i:D_i=0\}} Y_i,$$

y donde $N_1 = \sum_i D_i$; $N_0 = \sum_i (1 - D_i)$.

R/:

Para demostrar que $\hat{\tau}_{MCO} = \hat{\tau}_{DM}$, podemos estimar$\hat{\tau}_{MCO}$ haciendo la minimización de cuadrados a los residuos de nuestro modelo de regresión:

$$\min_{\hat{\beta}_0, \hat{\tau}} \sum_{i=1}^{N} (Y_i - \hat{\beta}_0 - \hat{\tau} D_i)^2$$

FOC respecto de $\hat{\beta}_0$:

$$-2 \sum_{i=1}^{N} (Y_i - \hat{\beta}_0 - \hat{\tau} D_i) = 0$$

$$\sum_{i=1}^{N} Y_i - N\hat{\beta}_0 - \hat{\tau}\sum_{i=1}^{N} D_i = 0$$

$$N\bar{Y} - N\hat{\beta}_0 - \hat{\tau}N_1 = 0$$

Despejando β_0:

$$\hat{\beta}_0 = \bar{Y} - \hat{\tau}\frac{N1}{N}$$

FOC respecto de $\hat{\tau}$:

$$-2\sum_{i=1}^{N}(Y_i - \hat{\beta}_0 - \hat{\tau}D_i)D_i = 0$$

Para que la sumatoria exista $D_i = 1$:

$$\sum_{i:D_i=1} (Y_i - \hat{\beta}_0 - \hat{\tau}) = 0$$

$$\sum_{i:D_i=1} Y_i - N_1\hat{\beta}_0 - N_1\hat{\tau} = 0$$

Despejando $\hat{\tau}$:

$$\hat{\tau} = \frac{1}{N_1}\sum_{i:D_i=1} Y_i - \hat{\beta}_0$$

$$\hat{\tau} = \bar{Y_1} - \hat{\beta}_0$$

Reemplazamos el β_0 que obtuvimos:

$$\hat{\tau} = \bar{Y_1} - \left(\bar{Y} - \hat{\tau}\,\frac{N_1}{N}\right)$$

$$\hat{\tau} - \hat{\tau}\frac{N1}{N} = \bar{Y_1} - \bar{Y}$$

$$\hat{\tau}\left(1 - \frac{N_1}{N}\right) = \bar{Y_1} - \bar{Y}$$

Si $\frac{N_1}{N}$ es la proporción de tratados, $1 - \frac{N_1}{N}$ es la proporción de controles:

$$\hat{\tau}\left(\frac{N_0}{N}\right) = \bar{Y_1} - \bar{Y}$$

El promedio es la suma del promedio de los grupos ponderado:

$$\hat{\tau}\left(\frac{N_0}{N}\right) = \bar{Y_1} - \left(\bar{Y_1}\frac{N_1}{N} + \bar{Y_0}\frac{N_0}{N}\right)$$

$$\hat{\tau}\left(\frac{N_0}{N}\right) = \bar{Y_1}\left(1 - \frac{N_1}{N}\right) - \bar{Y_0}\frac{N_0}{N}$$

$$\hat{\tau}\left(\frac{N_0}{N}\right) = \bar{Y_1}\left(\frac{N_0}{N}\right) - \bar{Y_0}\left(\frac{N_0}{N}\right)$$

$$\hat{\tau}_{MCO} = \bar{Y_1} - \bar{Y_0}$$

$$\hat{\tau}_{MCO} = \frac{1}{N_1}\sum_{\{i:D_i=1\}} Y_i - \frac{1}{N_0}\sum_{\{i:D_i=0\}} Y_i$$

Cómo vemos, $\hat{\tau}_{MCO} = \hat{\tau}_{DM}$.

b).¿que supuesto debe cumplirse para que $\hat{\tau}_{DM}$ sea un esitimador insesgadod el efecto causal promedio?

Para verificar que supuesto se debe cumplir vamos a evaluar la insesgadez tomando la esperanza matematica de nuestro estimador:

$$E[\hat{\tau}_{DM}] = E[\bar{Y_1} - \bar{Y_0}]$$
$$E[\hat{\tau}_{DM}] = E[\bar{Y_1}] - E[\bar{Y_0}]$$

Si usamos notación de resultados potenciales:

$$E[\hat{\tau}_{DM}] = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

Como en la primera demostración del taller, si sumamos y restamos el contrafactual tenemos que:

$$E[\hat{\tau}_{DM}] = \underbrace{E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1]}_{ATT} + \underbrace{E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]}_{\text{Selectión Bias}}$$

Por lo tanto, para que nuestro estimador $\hat{\tau}_{DM}$ se insesgado necesitamos evitar el sesgo de selección lo cual se logra por medio de la aleatorización garantizando que $(Y_i(1), Y_i(0)) \perp D_i$

Ahora veamos esto como lo podemos obtener tambien de la recta de regresión. Recordemos que nuestra linea de regresión es:

$$Y_i = \beta_0 + \tau D_i + U_i$$

Partamos del mismo punto que antes:

$$\begin{aligned} E[\hat{\tau}_{DM}] &= E[\bar{Y_1} - \bar{Y_0}] \\ E[\hat{\tau}_{DM}] &= E[\bar{Y_1}] - E[\bar{Y_0}] \\ E[\hat{\tau}_{DM}] &= E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \end{aligned}$$

Ahora si tomamos estos valores esperados de nuestra recta obtenemos:

$$\begin{aligned} E[Y_i|D_i = 1] &= \beta_0 + \tau + E[U_i|D_i = 1] \\ E[Y_i|D_i = 0] &= \beta_0 + E[U_i|D_i = 0] \end{aligned}$$

Por lo tanto tenemos:

$$E[\hat{\tau}_{DM}] = \beta_0 + \tau + E[U - i|D_i = 1] - (\beta_0 + E[U_i|D_i = 0])$$

tal que:

$$E[\hat{\tau}_{DM}] = \tau + \underbrace{E[U_i|D_i = 1] - E[U - i|D_i = 0]}_{\text{Selection Bias}}$$

Por lo tanto en terminos de la recta de regresión el supuesto de independencia $(Y_i(1), Y_i(0)) \perp D_i$ se traduce en el supuesto de exogenidad $E[U_i|D_i] = 0$. Esto nos señala que, para tener un estimador insesgado, no debemos tener factores no observables correlacionados con el tratamiento que afecten el resultado, haciendo a los grupos diferentes."

1.4 Punto4

Suponga ahora que la asignación de estudiantes a cada salón se realiza de manera aleatoria. De igual forma, la decisión sobre qué salón recibe el tratamiento también se tomó aleatoriamente. Además, todos los estudiantes pertenecientes al salón asignado al tratamiento efectivamente participaron en el programa. Demuestre que el estimador

$$\hat{\tau}_{DM} \xrightarrow{p} \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]$$

Puede suponer que el supuesto de SUTVA se cumple. Posteriormente, defina intuitivamente y explique la importancia de la propiedad de consistencia de un estimador.

R/:

Para demostrar que $\hat{\tau}_{DM} \xrightarrow{p} \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]$ empecemos por:

$$\text{plim}[\hat{\tau}_{DM}] = \text{plim}\left(\left(\frac{1}{N_1}\sum_{i:D_i=1} Y_i\right) - \left(\frac{1}{N_0}\sum_{i:D_i=0} Y_i\right)\right)$$

Como vimos en clase, por mapeo continuo (Slutsky):

$$\text{plim}[\hat{\tau}_{DM}] = \text{plim}\left(\frac{1}{N_1}\sum_{i:D_i=1} Y_i\right) - \text{plim}\left(\frac{1}{N_0}\sum_{i:D_i=0} Y_i\right)$$

Así mismo por WLLN:

$$\text{plim}[\hat{\tau}_{DM}] = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

Si escribimos en resultado potencial:

$$\text{plim}[\hat{\tau}_{DM}] = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0]$$

Aplicamos la suma y resta del contrafactual:

$$\text{plim}[\hat{\tau}_{DM}] = \underbrace{E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1]}_{ATT} + \underbrace{E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]}_{\text{Selectión Bias}}$$

Cómo sabemos por el enunciado que la asignación a los salones fue aleatoría entonces podemos estar tranquilo respecto del sesgo de selección, es decir sabemos que $((Y_i(1), Y_i(0)) \perp D_i)$:

$$\text{plim}[\hat{\tau}_{DM}] = \underbrace{E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1]}_{ATT}$$

Simplficando:

$$\text{plim}[\hat{\tau}_{DM}] = \underbrace{E[Y_i(1) - Y_i(0)|D_i = 1]}_{ATT}$$

que es lo mismo que $\hat{\tau}_{DM} \xrightarrow{p} \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]$

2 Segundo Ejercicio

Tal como señalan Abadie et al. (2022), los errores estándar clusterizados han adquirido un rol central en el trabajo empírico y esto no es casualidad. Si bien una parte sustancial de los esfuerzos de los investigadores se concentra en la obtención de estimadores consistentes e insesgados, resulta también fundamental poder realizar inferencia estadística válida sobre dichos estimadores. En este contexto, una de las prácticas más extendidas en la literatura empírica es el reporte de errores estándar clusterizados. No obstante, su uso plantea varias preguntas recurrentes en la práctica: ¿en qué situaciones es apropiado utilizar este tipo de errores estándar?, ¿cuándo tienen un impacto importante sobre la inferencia estadística? y ¿a qué nivel debería realizarse la clusterización?

Teniendo en cuenta lo anterior, suponga que usted desea estudiar **el impacto de que un estudiante de raza negra tenga un profesor de la misma raza**. Es bien sabido que existen brechas raciales en los resultados educacionales, lo cual resulta problemático por diversas razones: la movilidad social, los ingresos futuros y el bienestar social están estrechamente vinculados al nivel educativo. En este contexto, usted decide seguir la línea de investigación propuesta por Gershenson et al. (2022) y busca replicar algunos de sus principales resultados, en particular aquellos relacionados con el desempeño en exámenes y la entrada a la universidad.

Para realizar estos ejercicios, usted dispone de una base de datos de corte transversal denominada `race_teaching.dta` que proviene de un experimento realizado en EE.UU. que aleatorizaba la raza de los profesores en diferentes salones de kínder. Sus datos están a nivel individual, donde cada observación es una persona de raza negra que participó en el experimento.

Esta base contiene información sobre los resultados en el SAT (examen estandarizado para el ingreso a la universidad) y una variable indicadora que señala si el estudiante asistió a la universidad. A estas variables de interés se las denota por Y_i. Adicionalmente, la base incluye una variable indicadora D_i, que toma el valor de uno si el estudiante i tuvo un profesor de raza negra en kínder y cero en caso contrario. Asimismo, se incorporan características geográficas, tales como el salón, el colegio, el condado y el estado al que pertenece cada individuo.

Antes de abordar las decisiones empíricas que deberá tomar en su investigación, usted decide analizar primero, desde una perspectiva teórica, los errores estándar clusterizados. De este modo, podrá comprender su formulación, sus principales propiedades y sentar las bases para determinar en qué contextos resulta apropiado utilizarlos y cómo deben implementarse.

2.1 Punto1

Realice un análisis crítico de los supuestos de homocedasticidad y no-correlación del término de error. Para ello:

- Defina ambos supuestos, tanto desde un punto de vista matemático como intuitivo, en el contexto del caso de estudio.
- Discuta si resulta razonable asumir el cumplimiento de estos supuestos dada la pregunta de investigación. Justifique su respuesta utilizando un ejemplo intuitivo para cada supuesto.

R/:

El supuesto de homocedasticidad implica que $E[e_i^2|x] = \sigma^2$ que tambien podría ser visto como $E[e_i^2|D_i = 1] = \sigma^2$ y $E[e_i^2|D_i = 0] = \sigma^2$, lo que quiere decir que la varianza de los errores es constante para cada estudiante independientemente de sus caracteristicas (X). En este caso, quiere decir que la dispersión en los resultados de la prueba SAT para cada estudiante es la misma para el grupo de estudiantes tratados y no tratados.

Por otro lado, el supuesto de no-correlación implica que $E[e_i e_j] = 0$. Esto quiere decier que conocer el error del estudiante i no me dice nada al respecto del resultado del estudiante j. Es decir, si a un estudiante le va muy bien en la prueba, por algun factor diferente al tratamiento, eso no quiere decir que a otro estudiante le vaya a ir igual de bien. Es decir, si un estudiante el dia del examen se desperto con dolor de cabeza, eso no nos dice nada respecto de otros estudiantes.

El supuesto de homocedasticidad resulta poco plausible en este escenario. Tal como lo señalan Angrist y Pischke (2009) en la vida real, aunque no conocemos la CEF, es viable pensar que esta no es lineal necesarimente, por lo que el hecho que no sea lineal ya va a generar por si mismo que tengamos variabilidad en la forma en que la línea de regresión se aproxima a la CEF. Así mismo, hay que tener en cuenta que una de nuestras variables dependientes es binaria (asístencia a la universidad), de acuerdo con los autores mencionados, esto anula el supusto de homocedasticidad porque por simple construcción va a haber variación en los errores.

Por otro lado, Abadie et al.(2023) señalan que el ajuste por cluster de errores estandar se debe hacer por nivel de muestreo y de asignación de tratamiento. En este caso el tratamiento fue asignado por salones, por lo que no solo no deberíamos asumir errores homocedasticos sino aplicar cluster. Esto tambien implica que debemos ser concientes de que es muy probable que el supuesto de no-correlación $E[e_i e_j] = 0$ no se cumpla, pues en este caso pueden haber factores no observados en cada salon que si correlaciónen los resultados del estudiante i con el estudiante j que pertenecen al mismo salon, lo que vimos formalmente en clase como $E[e_{gi} e_{gh}] \neq 0$. Por ejemplo, si el profesor de un salón específico se enferma o, por el contrario, posee una motivación excepcional independientemente de su raza, estos factores afectarán simultáneamente el término de error de todos los estudiantes de ese curso, generando una correlación intra-clúster que debe ser corregida.

2.2 Punto2

Por el momento, suponga que usted desea estimar el modelo sin intercepto:

$$Y_i = \beta X_i + \varepsilon_i$$

> **Nota:** Considere que es un modelo donde las variables Y y X se han re-centrado con respecto a sus medias. En general, note que eliminar el intercepto tiene consecuencias importantes si no se realiza dicho proceso.

Donde X_i es una variable determinística. Determine la varianza del estimador de β bajo los siguientes escenarios:

- **Caso 1:** Se cumplen los supuestos de homocedasticidad y no-correlación del término de error.
- **Caso 2:** Se cumple el supuesto de no-correlación, pero existe heterocedasticidad.
- **Caso 3:** Existe heterocedasticidad y correlación entre las características no-observadas de los individuos que asisten al mismo colegio.

Para cada caso, identifique el estimando poblacional correspondiente y proponga un estimador adecuado de la varianza. Justifique brevemente su respuesta.

R/:

En primer lugar partamos de la defínción de nuestro esmitado $\hat{\beta}$:

Nuestro modelo es: $Y_i = \beta X_i + \varepsilon_i$

El problema de minimización es:

$$min \sum_{i=1}^{N} (Y_i - \beta X_i)^2$$

Si obtenemos la FOC:

$$\frac{\partial SSR}{\partial \beta} = 2 \sum_{i=1}^{N} (Y_i - \beta X_i)(-X_i)$$

$$\sum_{i=1}^{N} (Y_i X_i - \hat{\beta} X_i^2) = 0$$

Despejando $\hat{\beta}$:

$$\sum_{i=1}^{N} (Y_i X_i) - \hat{\beta} \sum_{i=1}^{N} X_i^2 = 0$$

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(Y_i X_i)}{\sum_{i=1}^{N} X_i^2}$$

Ahora veamos la varianza en cada uno de los casos:

- **Caso 1:** Se cumplen los supuestos de homocedasticidad y no-correlación del término de error.

el caso nos plantea que $Var(e_i|x_i) = \sigma^2$ y que $E[e_i e_j] = 0, \forall i \neq j$.

Para encontrar la varianza peimro organicemos remplazando Y_i:

$$\hat{\beta} = \frac{\sum_{i=1}^{N}((\beta X_i + e_i)X_i)}{\sum_{i=1}^{N} X_i^2}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(\beta X_i^2 + e_i X_i)}{\sum_{i=1}^{N} X_i^2}$$

$$\hat{\beta} = \frac{\beta \sum_{i=1}^{N} X_i^2 + \sum_{i=1}^{N} e_i X_i}{\sum_{i=1}^{N} X_i^2}$$

$$\hat{\beta} = \beta + \frac{\sum_{i=1}^{N} e_i X_i}{\sum_{i=1}^{N} X_i^2}$$

Aplicamos el operador de la varianza:

$$\text{Var}(\hat{\beta}|x_i) = \underbrace{\text{Var}(\beta)}_{\text{Var de conste} = 0} + \text{Var}\left(\frac{\sum_{i=1}^{N} X_i e_i}{\sum_{i=1}^{N} X_i^2}\right)$$

Aplicamos la propiedad de la varianza $Var(aX + b) = a^2 Var(X)$:

$$\text{Var}(\hat{\beta}|x_i) = \text{Var}\left(\frac{\sum_{i=1}^{N} X_i e_i}{\sum_{i=1}^{N} X_i^2}\right)$$

$$\text{Var}(\hat{\beta}|x_i) = \frac{1}{\left(\sum_{i=1}^{N} X_i^2\right)^2} \sum_{i=1}^{N} X_i^2 \underbrace{\text{Var}(e_i|x_i)}_{\sigma^2}$$

$$\text{Var}(\hat{\beta}|x_i) = \frac{\sigma^2}{\sum_{i=1}^{N} X_i^2}$$

Una vez obtenido nuestro estimando, como no conocemos σ^2 usamos su estimador $\hat{\sigma}^2$. Por lo tanto nuestro estimador de la varianza es:

$$\widehat{\text{Var}(\hat{\beta})} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{N} X_i^2}$$

Donde:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} e_i^2}{N - \underbrace{K}_{\text{en este caso} = 1}}$$

- **Caso 2:** Se cumple el supuesto de no-correlación, pero existe heterocedasticidad.

Este caso nos plantea que que $\text{Var}(e_i|x_i) = \sigma_i^2$ y que $E[e_i e_j] = 0, \forall i \neq j$.

$$\text{Var}(\hat{\beta}|x_i) = \frac{1}{\left(\sum_{i=1}^{N} X_i^2\right)^2} \sum_{i=1}^{N} X_i^2 \underbrace{\text{Var}(e_i|x_i)}_{\sigma^2}$$

$$\text{Var}(\hat{\beta}|x_i) = \frac{1}{\left(\sum_{i=1}^{N} X_i^2\right)^2} \sum_{i=1}^{N} X_i^2 \sigma_i^2$$

$$\text{Var}(\hat{\beta}|x_i) = \frac{\sum_{i=1}^{N} X_i^2 \sigma_i^2}{\left(\sum_{i=1}^{N} X_i^2\right)^2}$$

Por lo tanto el estimador de la varianza siguiendo la formulación de errores robustos de White es:

$$\widehat{\text{Var}(\hat{\beta})} = \frac{\sum_{i=1}^{N} X_i^2 \hat{e}_i^2}{\left(\sum_{i=1}^{N} X_i^2\right)^2}$$

Donde $\hat{e}_i = (Y_i - \hat{\beta} X_i)$. Es importante tener en cuenta que no podemos conocer la varianza de cada persona i por lo tanto hacemos una aproximación a traves del residual estimado.

- **Caso 3:** Existe heterocedasticidad y correlación entre las características no-observadas de los individuos que asisten al mismo colegio.

Este caso nos plantea que que $\text{Var}(e_{gi}|X) = \sigma^2_{gi}$, que existe correlación dentro de los grupostal que $E[e_{gi}e_{gj}] \neq 0$ para estudiantes del mismo colegio, pero hay independencia entre coelgios $E[e_{gi}e_{hj}] = 0$.

entonces:

$$\hat{\beta} - \beta = \frac{\sum_{i=1}^{N} X_i e_i}{\sum_{i=1}^{N} X_i^2}$$

Como tenemos correlación al interior de los colegios, agrupamos la suma del numerador por grupos (colegios) g, en lugar de individuos i:

$$\hat{\beta} - \beta = \frac{\sum_{g=1}^{G} \left(\sum_{i \in g} X_{gi} e_{gi}\right)}{\sum_{i=1}^{N} X_i^2}$$

Aplicamos la varianza recordando que $E[e_{gi}e_{hj}] = 0$ y que la varianza de la suma es la vianza de cada grupo:

$$\text{Var}(\hat{\beta}|x_i) = \frac{\sum_{g=1}^{G} \text{Var}\left(\sum_{i \in g} X_{gi} e_{gi}\right)}{\left(\sum_{i=1}^{N} X_i^2\right)^2}$$

Que es lo mismo que:

$$\text{Var}(\hat{\beta} \mid x_i) = \frac{\sum_{g=1}^{G} E\left[\left(\sum_{i \in g} X_{gi} e_{gi}\right)^2\right]}{\left(\sum_{i=1}^{N} X_i^2\right)^2}$$

En este caso nuestro estimador sería:

$$\widehat{\text{Var}(\hat{\beta} \mid x_i)} = \frac{\sum_{g=1}^{G} \left(\sum_{i \in g} X_{gi} \hat{e}_{gi}\right)^2}{\left(\sum_{i=1}^{N} X_i^2\right)^2}$$

Este estimador es el adecuado porque captura la estructura de correlación intra-clúster. Al elevar al cuadrado la suma a nivel de grupo, el estimador incorpora no solo la varianza individual (heterocedasticidad), sino también todas las covarianzas entre estudiantes del mismo colegio.

Es importante notar que, aunque el tratamiento se asignó a nivel de salon, el supuesto del problema indica correlación de errores a nivel de colegio. Por lo tanto, para evitar subestimar la varianza, la variable de agrupación G corresponde a los colegios.

2.3 Punto3

Uno de sus colegas le indica que no es necesario tener en cuenta el cluster a nivel de colegio y le propone usar la siguiente expresión como estimador de la varianza en el caso 3:

$$\frac{\sum_i \sum_j x_i x_j \hat{\varepsilon}_i \hat{\varepsilon}_j}{(\sum_i x_i)^2}$$

- **Utilizando sus conocimientos de teoría de regresión lineal, explique brevemente por qué este estimador es inválido.**
- **Compare el estimador que usted propuso para el caso 3 con el que le propuso su colega. ¿Cómo soluciona su estimador el problema que identificó?**

R/:

El estimador es inválido por dos motivos.

1). En primer lugar, el numerador colapsa matemáticamente a cero. Esto se puede notar al recordar que una propiedad fundamental de MCO es que los residuos son ortogonales a los regresores. Tengamos en cuenta que esa ortogonalidad viene del mismo proceso de minimización de cuadrados:

$$min \sum_{i=1}^{N} (Y_i - \beta X_i)^2$$

FOC:

$$\frac{\partial SSR}{\partial \beta} = 2 \sum_{i=1}^{N} (Y_i - \beta X_i)(-X_i)$$

Igualando a cero para encontrar el mínimo:

$$\sum_{i=1}^{N} (X_i) \underbrace{(Y_i - \hat{\beta} X_i)}_{\hat{\varepsilon}_i} = 0$$

Lo que quiere decir que:

$$\sum_{i=1}^{N} X_i \hat{\varepsilon}_i = 0$$

Si tomamos el numerador propuesto

$$\sum_i \sum_j x_i x_j \hat{\varepsilon}_i \hat{\varepsilon}_j$$

y agrupamos las sumatorias:

$$\underbrace{\sum_i x_i \hat{\varepsilon}_i}_{0} \underbrace{\sum_j x_j \hat{\varepsilon}_j}_{0}$$

Por ortogonalidad, el numerador propuesto colapsaría a cero, resultando en una varianza estimada de cero.

2). En términos conceptuales, asume que existe correlación entre todos los individuos de la muestra, ignorando la independencia entre colegios $E[e_{gi}e_{hj}] = 0$.

Por su lado, nuestro estimador propuesto para el caso 3:

$$\widehat{\text{Var}(\hat{\beta})} = \frac{\sum_{g=1}^{G} \left(\sum_{i \in g} X_{gi} \hat{e}_{gi} \right)^2}{\left(\sum_{i=1}^{N} X_i^2 \right)^2}$$

soluciona el problema porque ordena y agrupa las operaciones de manera correcta.

Para poder entender esto, centrémonos de nuevo el numerador $\sum_{g=1}^{G} \left(\sum_{i \in g} X_{gi} \hat{e}_{gi} \right)^2$.

En primer lugar, calcula la suma de los productos cruzados para cada colegio g. En este caso, al ser la suma de cada colegio (local), se evita que la condición de ortogonalidad tenga efecto, pues no es la suma total de los productos; un colegio puede tener contribuciones positivas y otro negativas a los errores. Luego, al elevar eso al cuadrado, convertimos esos balances positivos y negativos en estrictamente positivos. Finalmente, aplicamos la sumatoria de todos los colegios, logrando así capturar la verdadera variabilidad intra-clúster y respetando la independencia entre colegios.

2.4 Punto 4

¿Compare los estimadores robustos a heterocedasticidad (caso 2) y robustos a clusters (caso 3) que usted propuso. Para ello:

- **Explique por qué el estimador robusto a clusters también es robusto a heterocedasticidad.**
- **¿Es razonable esperar que la varianza estimada mediante el estimador robusto a clusters sea mayor que la obtenida con el estimador robusto a heterocedasticidad? Justifique su respuesta.**

R/:

Para responder a estas preguntas veamos los dos estimadores. El de errores robustos a heterocedasticidad:

$$\widehat{\text{Var}(\hat{\beta})}_{White} = \frac{\sum_{i=1}^{N} X_i^2 \hat{e}_i^2}{\left(\sum_{i=1}^{N} X_i^2\right)^2}$$

Frente al de errores robustos por cluster:

$$\widehat{\text{Var}(\hat{\beta})}_{Cluster} = \frac{\sum_{g=1}^{G} \left(\sum_{i \in g} X_{gi} \hat{e}_{gi}\right)^2}{\left(\sum_{i=1}^{N} X_i^2\right)^2}$$

Como se puede notar, ambos estimadores estiman la varianza de cada observación i por medio de $\hat{e}_{gi}$ y $\hat{e}_i$. Es decir, ambos tienen en cuenta la heterocedasticidad, que significa una varianza diferente en los errores para cada valor de X.

Veámoslo más de cerca. Para ello, expandamos el numerador de nuestro estimador por cluster:

$$\sum_{g=1}^{G} \left(\sum_{i \in g} X_{gi} \hat{e}_{gi}\right)^2$$

Esto quedaría:

$$\sum_{g=1}^{G} \left[\left(\sum_{i \in g} X_{gi}^2 \hat{e}_{gi}^2\right) + \left(\sum_{i \in g} \sum_{j \in g, j \neq i} X_{gi} X_{gj} \hat{e}_{gi} \hat{e}_{gj}\right) \right]$$

Como se puede notar, el término de la izquierda es la expresión del estimador dos, que reconoce la heterocedasticidad.

Por otro lado, al verlo, se nota que es la parte del estimador de White **más** otra cosa, que representa la correlación intracluster. Por lo tanto, al permitir la correlación intracluster (generalmente positiva), solo viendo la ecuación podemos notar que la varianza va a ser mayor con el estimador de cluster.

Habiendo adquirido una mayor claridad sobre la teoría detrás de los errores estándar clusterizados, usted procede a intentar replicar algunos de los resultados del trabajo de Gershenson et al. (2022), integrando los conocimientos sobre errores estándar clusterizados desarrollados en los incisos anteriores.

En particular, recuerde que el experimento se llevó a cabo sobre estudiantes de kínder de raza negra y que la exposición a profesores de distinta raza fue aleatorizada *dentro de cada colegio*, asignando estudiantes y docentes a diferentes salones. Una ilustración intuitiva de este contexto puede verse en la serie *Abbott Elementary*, que retrata un colegio con una población estudiantil predominantemente negra cuyos docentes pertenecen a distintas razas. El experimento se implementó en múltiples condados y estados de E.E.U.U.

Bajo este diseño, los estudiantes de kínder de raza negra fueron expuestos de manera aleatoria a profesores de la misma raza o de razas distintas *condicional en el colegio*. En consecuencia, haciendo uso de sus conocimientos sobre diseños experimentales, usted estima la siguiente ecuación:

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i.$$

2.5 Punto 5

Usando el caso estudio, identifique qué factores pueden generar una mayor varianza estimada por parte de los errores estándar clusterizados a nivel de colegio. Para ello:

- **Usando el estimador que usted propuso en el caso 3 inciso 3, enuncie qué mecanismos matemáticos incrementan la varianza estimada.**
- **Por cada uno de los mecanismos hallados, dé un ejemplo concreto usando el contexto del caso estudio que lo respalde.**
- **Estime el modelo anterior utilizando como variable dependiente: i) el puntaje en matemáticas del SAT y ii) la probabilidad de ingresar a la universidad. Reporte los resultados en una tabla, incluyendo errores estándar robustos a heterocedasticidad y errores estándar clusterizados a nivel de colegio.**

R/:

El estimador de cluster que obtuvimos fue

$$\widehat{\text{Var}(\hat{\beta})}_{Cluster} = \frac{\sum_{g=1}^{G} \left(\sum_{i \in g} X_{gi} \hat{e}_{gi} \right)^2}{\left(\sum_{i=1}^{N} X_i^2 \right)^2}$$

Matemáticamente existen dos mecanismos fundamentales que incrementan la varianza estimada: un aumento en el numerador o una disminución en el denominador. Veamos de qué consta el numerador:

$$\sum_{g=1}^{G} \left[\left(\sum_{i \in g} X_{gi}^2 \hat{e}_{gi}^2 \right) + \left(\sum_{i \in g} \sum_{j \in g, j \neq i} X_{gi} X_{gj} \hat{e}_{gi} \hat{e}_{gj} \right) \right]$$

Como vimos anteriormente, el primer término representa la varianza de las observaciones individuales. Dicha varianza está compuesta de X_{gi} y de $\hat{e}_{gi}$, que representan respectivamente el valor de la variable independiente para el individuo i en el grupo o *cluster* g y su residuo respectivo. Por lo anterior, la varianza se puede inflar matemáticamente si tenemos datos muy atípicos en X_{gi} o residuos muy grandes $\hat{e}_{gi}$.

En cuanto al segundo término, este representa la correlación intra-cluster. Si existe una alta correlación entre los individuos del mismo grupo, los productos cruzados en nuestra sumatoria serán grandes y positivos, incrementando significativamente la varianza.

Ahora veamos el denominador:

$$\left(\sum_{i=1}^{N} X_i^2 \right)^2$$

Representa el cuadrado de la variabilidad total de X. Funciona como un estabilizador, a mayor información, es decir, más datos o más dispersión en X, más grande es este denominador y más pequeña es la varianza y viceversa.

Aterricemos estos mecanismos al caso. En nuestro modelo con intercepto, la variable relevante es la versión centrada $(D_i - \bar{D})$

Para derivar el estimador exacto en este contexto, partimos del problema de minimización:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{N} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i)^2$$

FOC respecto de $\hat{\beta}_0$:

$$-2\sum_{i=1}^{N}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i) = 0$$

$$\sum_{i=1}^{N} Y_i - N\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{N} D_i = 0$$

$$N\hat{\beta}_0 = \sum_{i=1}^{N} Y_i - \hat{\beta}_1 \sum_{i=1}^{N} D_i$$

Despejando β_0:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{D}$$

FOC respecto de $\hat{\beta}_1$:

$$-2\sum_{i=1}^{N}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i)D_i = 0$$

$$\sum_{i=1}^{N}(Y_i - (\bar{Y} - \hat{\beta}_1 \bar{D}) - \hat{\beta}_1 D_i)D_i = 0$$

$$\sum_{i=1}^{N}(Y_i - \bar{Y} + \hat{\beta}_1 \bar{D} - \hat{\beta}_1 D_i)D_i = 0$$

$$\sum_{i=1}^{N}[(Y_i - \bar{Y}) - (D_i - \bar{D})hat\beta_1]D_i = 0$$

$$\sum_{i=1}^{N} D_i(Y_i - \bar{Y}) - hat\beta_1 \sum_{i=1}^{N} D_i(D_i - \bar{D}) = 0$$

Que es lo mismo que:

$$\sum_{i=1}^{N}(D_i - \bar{D})(Y_i - \bar{Y}) - \hat{\beta}_1 \sum_{i=1}^{N}(D_i - \bar{D})(D_i - \bar{D}) = 0$$

Por lo tanto depejando $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(D_i - \bar{D})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(D_i - \bar{D})^2}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(D_i - \bar{D})Y_i}{\sum_{i=1}^{N}(D_i - \bar{D})^2}$$

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{N}(D_i - \bar{D})\varepsilon_i}{\sum_{i=1}^{N}(D_i - \bar{D})^2}$$

Si aplicamos la varianza condicional y permitimos la correlación dentro de los clusters:

$$\text{Var}(\hat{\beta}_1) = \frac{1}{\left(\sum_{i=1}^{N}(D_i - \bar{D})^2\right)^2} \text{Var}\left(\sum_{i=1}^{N}(D_i - \bar{D})\varepsilon_i\right)$$

Entonces:

$$\widehat{\text{Var}(\hat{\beta})}_{Cluster} = \frac{\sum_{g=1}^{G}\left(\sum_{i\in g}(D_{gi} - \bar{D})\hat{e}_{gi}\right)^2}{\sum_{i=1}^{N}(D_i - \bar{D})^2}$$

Por lo tanto en el caso de nuestra regresión el numerador expandido, que captura los mecanismos de inflación de varianza, es:

$$\sum_{g=1}^{G}\left[\left(\sum_{i\in g}(D_{gi} - \bar{D})^2\hat{e}_{gi}^2\right) + \left(\sum_{i\in g}\sum_{j\in g, j\neq i}(D_{gi} - \bar{D})(D_{gj} - \bar{D})\hat{e}_{gi}\hat{e}_{gj}\right)\right]$$

Veamos, el primer mecanismo asociado a la varianza individual. Como ya mencionamos Se observa un incremento si hay factores residualesgrandes no explicados por el tratamiento. Por ejemplo, un shock exógeno, como una ola de desempleo o divorcios en las familias de un grupo específico de estudiantes, aumentaría sus residuos individuales inflando la varianza total.

Por otro lado, en cuanto a la correlación intra-cluster, sabemos que a mayor correlación entre compañeros de colegio, mayor varianza. Supongamos que el colegio $g = 1$ sufrió un fallo eléctrico y se quedó sin aire acondicionado durante los exámenes. Este factor ambiental afecta negativamente a todos los estudiantes de ese colegio simultáneamente. Sus errores estarán correlacionados positivamente, incrementando drásticamente el segundo término del numerador.

Por úlitmo, veamos como funcionaría el mecanismo del denominador, que en este caso sería:

$$\left(\sum_{i=1}^{N}(D_i - \bar{D})^2\right)^2$$

Supongamos que en nuestra muestra solo tenemos 10 estudiantes tratados (con profesor de su misma raza) frente a miles de controles. En este caso, la variación $(D_i - \bar{D})^2$ es mínima, lo que hace que el denominador sea muy pequeño. Al dividir por un número pequeño, la varianza resultante será muy grande, indicando que nos falta información para estimar el efecto con precisión.

Veamos la estimación de los modelos con errores robustos para heterocedasticidad y con cluster de errores estandar a nivel colegio:

```
#Cargando paquetes necesarios
library(haven)
library(fixest)
library(modelsummary)
library(dplyr)
library(kableExtra)

#Importando los datos para la estimación del modelo

race_teaching_1_ <- read_dta("race_teaching (1).dta")

#Haciendo las regresiones con errores heterocedasticos
Reg_Hete_SAT<-feols(sat_math ~ D, race_teaching_1_, vcov = "hetero")
Reg_Hete_Col<-feols(college ~ D, race_teaching_1_, vcov = "hetero")

#Haciendo las regresiones con cluster por colegio
Reg_Clus_SAT<-feols(sat_math ~ D, race_teaching_1_, vcov = ~ school )
Reg_Clus_Col<-feols(college ~ D, race_teaching_1_, vcov = ~ school)

#Haciendo tabla autocontendia

#Definimos el mapa de coeficientes y la información visible
mapa_coeficientes <- c("D" = "Profesor Misma Raza",
                       "(Intercept)" = "Media sin Tratamiento")
mapa_gof <- list(list("raw" = "nobs",
                      "clean" = "Observaciones", "fmt" = 0),
                 list("raw" = "r.squared",
                      "clean" = "R²", "fmt" = 3))

#Generamos la tabla
```

```r
modelsummary(
  list(
    "(Robusto)" = Reg_Hete_SAT,
    "(Cluster)" = Reg_Clus_SAT,
    "(Robusto)" = Reg_Hete_Col,
    "(Cluster)" = Reg_Clus_Col
  ),
  coef_map = mapa_coeficientes,
  gof_map = mapa_gof,
  fmt = 3,
  stars = c('*' = .1, '**' = .05, '***' = .01),
  title = "Tabla 1. Efecto de la asignación racial docente",
  notes = list("Nota: Errores estándar entre paréntesis."),

  output = "kableExtra"
) %>%
  add_header_above(c(" " = 1, "Puntaje SAT" = 2,
                     "Ingreso Universidad" = 2)) %>%

  kable_styling(
    latex_options = c("hold_position",
                      "scale_down", "striped"),
    bootstrap_options = c("striped", "hover", "condensed"),
    full_width = FALSE,
    position = "center"
  )
```

Table 1: Tabla 1. Efecto de la asignación racial docente

	Puntaje SAT		Ingreso Universidad	
	(Robusto)	(Cluster)	(Robusto)	(Cluster)
Profesor Misma Raza	4.607***	4.607***	0.198***	0.198***
	(0.064)	(0.065)	(0.003)	(0.004)
Media sin Tratamiento	499.616***	499.616***	0.513***	0.513***
	(0.045)	(0.319)	(0.002)	(0.010)
Observaciones	99 204	99 204	99 204	99 204
R^2	0.050	0.050	0.041	0.041

* p < 0.1, ** p < 0.05, *** p < 0.01

Nota: Errores estándar entre paréntesis.

2.6 Punto 6

¿Qué posible impacto puede tener un nivel de agregación más grande al clusterizar sobre la varianza estimada? Use como guía los siguientes puntos:

- **Use el estimador de la varianza que usted propuso en el inciso 3, caso 3 y compare la varianza teórica que se obtiene al clusterizar por colegio vs por condado**.
- **Repita las estimaciones del inciso 5, concentrándose únicamente en la variable Y_i correspondiente al puntaje de matemáticas en la prueba SAT. Reporte los coeficientes estimados junto con errores estándar robustos a heterocedasticidad y errores estándar clusterizados a nivel de: i) salón, ii) colegio, iii) condado y iv) estado**
- **interprete los resultados obtenidos. ¿Las estimaciones son consistentes con lo que usted encuentra teóricamente?**

(Pendiente demostración matematica la hago después no se como hacerla)

Siguiendo a Angrist y Pischke y los apuntes de Hansen, y considerando que la unidad de colegio se encuentra anidada dentro de condado y este a su vez en estado, la teoría predice el comportamiento de la varianza. Siempre que exista una correlación intra-cluster positiva a un nivel superior (es decir, que los colegios dentro del mismo condado o estado compartan choques comunes), la varianza estimada clusterizada a estos niveles debería ser mayor que la varianza robusta o clusterizada a niveles inferiores

Respecto a la validez de esta inferencia, Angrist menciona que la corrección por cluster funciona adecuadamente cuando el número de grupos es suficientemente grande $G \approx 42$. Hansen advierte que con un G pequeño, los errores estándar clusterizados pueden volverse imprecisos y sufrir de un sesgo hacia abajo.

En la práctica econométrica, se sugiere clusterizar al nivel más agregado donde exista correlación de los errores o donde se haya asignado el tratamiento. Si al subir el nivel de clusterización los errores estándar aumentan, estamos corrigiendo un sesgo negativo y evitando falsos positivos. Por otro lado, si los errores estándar permanecen estables al subir de nivel, es indica que no hay una correlación espacia a ese nivel. Sin embargo, considero que debe ser una buena práctica reportar estos niveles superiores como prueba de robustez por transparencia con los resultados.

Ahora veamos la estimación del modelo tomando como variable Y_i el puntaje en matematicas en la prueba SAT usando errores robustos para heterocedasticidad y errores clusterizados a nivel de los modelos con errores robustos para heterocedasticidad y con cluster de errores estandar a nivel i) colegio, ii) condado y iii) estado:

```r
#Cargando paquetes necesarios
library(haven)
library(fixest)
library(modelsummary)
library(dplyr)
library(kableExtra)

#Importando los datos para la estimación del modelo

race_teaching_1_ <- read_dta("race_teaching (1).dta")

#Haciendo las regresion con errores heterocedasticos
Reg_Hete<-feols(sat_math ~ D, race_teaching_1_, vcov = "hetero")

#Haciendo las regresiones con cluster por colegio
Reg_Clus_School<-feols(sat_math ~ D, race_teaching_1_, vcov = ~ school )

#Haciendo las regresiones con cluster por condado
Reg_Clus_County<-feols(sat_math ~ D, race_teaching_1_, vcov = ~ county )

#Haciendo las regresiones con cluster por estado
Reg_Clus_State<-feols(sat_math ~ D, race_teaching_1_, vcov = ~ state )

#Haciendo tabla autocontendia

#Definimos el mapa de coeficientes y la información visible
mapa_coeficientes <- c("D" = "Profesor Misma Raza",
                       "(Intercept)" = "Media sin Tratamiento")
mapa_gof <- list(list("raw" = "nobs",
                      "clean" = "Observaciones", "fmt" = 0),
                 list("raw" = "r.squared",
                      "clean" = "R²", "fmt" = 3))

#Generamos la tabla
modelsummary(
  list(
    "Robustos" = Reg_Hete,
    "Cluster Coelgio" = Reg_Clus_School,
    "Cluster Condado" = Reg_Clus_County ,
    "Cluster Estado" = Reg_Clus_State
  ),
  coef_map = mapa_coeficientes,
```

```
  gof_map = mapa_gof,
  fmt = 3,
  stars = c('*' = .1, '**' = .05, '***' = .01),
  title = "Tabla 1. Efecto de la asignación
  racial docente en diferentes niveles cluster",
  notes = list("Nota: Errores estándar entre paréntesis."),

  output = "kableExtra"
) %>%
  add_header_above(c(" " = 1, "Puntaje SAT" = 4)) %>%

  kable_styling(
    latex_options = c("hold_position",
                      "scale_down", "striped"),
    bootstrap_options = c("striped", "hover", "condensed"),
    full_width = FALSE,
    position = "center"
  )
```

Table 2: Tabla 1. Efecto de la asignación racial docente en diferentes niveles cluster

	Puntaje SAT			
	Robustos	Cluster Coelgio	Cluster Condado	Cluster Estado
Profesor Misma Raza	4.607***	4.607***	4.607***	4.607***
	(0.064)	(0.065)	(0.067)	(0.060)
Media sin Tratamiento	499.616***	499.616***	499.616***	499.616***
	(0.045)	(0.319)	(0.683)	(1.204)
Observaciones	99 204	99 204	99 204	99 204
R^2	0.050	0.050	0.050	0.050

* p < 0.1, ** p < 0.05, *** p < 0.01

Nota: Errores estándar entre paréntesis.

El resultado de nuestra estimación da cuenta de lo que dijimos anteriormente. El estimador puntual es idéntico en todos los casos 4.607 con le mismo nivel de significacia estadistica $p < 0.001$. Tambien podemos ver que el error estándar aumenta a 0.065 cuando aplicamos cluste de colegio y a 0.067 cuando aplicamos cluster de condado. No obstante al agrupar por estado los eerores estandar caen 0.060.

Entoces podemos decir que el incremento de colegio a condado da cuenta de la existencia de correlación positiva entre colegios del mismo condado, lo que nos indica que el cluster a nivel Condado sería la estimación más adecuada sobreto si tenemos en cuenta que en nuestra

muestra $G > 500$ para el nivel de condado. Por su parte la disminución cuando se aplica cluster a nivel de estado seguramente tiene que ver con un sesgo en el calculo de los errores asociado a un $G = 50$ que si bien es superior al aproximado mencionado por Angrist, como explica Hansen en sus notas, aún hay mucho por aprender sobre el correcto nivel para aplicar la correcta clusterización de los errores estandar.

2.7 Punto 7

Usted decide escribirle a su amiga cercana Susan Athey para pedirle su opinión sobre el comentario hecho por el referee #2. Ella le envía un codigo en **R**:

```
#-------------------------------------------------
# Housekeeping\\Codigo Susan
#-------------------------------------------------

# Semilla para replicación.
set.seed(31416)

# Cargar las librerías necesarias.
library(sandwich)
library(lmtest)

# Parámetros de la simulación
#-------------------------------------------------

# Número de individuos en la población.
n_population <- 1e7

# Número de colegios.
schools <- 100

# Número de estudiantes por colegio
stud_by_school <- n_population / schools

# Probabilidad de aparecer en la muestra.
p_sample <- 0.01

# Número de simulaciones.
n_sim <- 500


# Construir base de datos poblacional
```

```r
#-------------------------------------------------

# Base para la construcción.
population <- data.frame(id = 1:n_population)

# Determinar el tratamiento.
population$D <- rbinom(n_population,
                       size = 1, prob = 0.5)

# Determinar un colegio para cada individuo.
population$schools <- rep(1:schools,
                          each = stud_by_school)

# Determinar el efecto por colegio.
population$school_effect <- c(rep(1, n_population / 2),
                              rep(-1, n_population / 2))

# Construir la variable de resultado.
population$Y <- population$school_effect * population$D
+ runif(n_population)
```

Usando ese código, realice el siguiente procedimiento:

- Tome 1′000 muestras de la población de interés y estime el modelo usando errores estándar robustos a heterocedasticidad y clusterizados a nivel de colegio.
- Para cada una de las muestras, determine el intervalo de confianza al 95% usando ambos errores estándar.
- Para cada una de las muestras, determine si el intervalo de confianza "cubre" el cero (es decir, el intervalo de confianza incluye el cero). Note que, en la simulación, el efecto esperado es cero por construcción.
- Determine los errores estándar robustos a heterocedasticidad y clusterizados promedio.
- Determine el porcentaje de veces que los intervalos de confianza "cubren" el cero.
- Interprete sus resultados. ¿Es cierto que siempre es mejor clusterizar al nivel más agregado?

R/:

```r
#| echo: true
#| results: 'hide'

#Tomando 1000 muestras y estimando modelos
#_______________________________________________________________________________
library(fixest) #Usammos fixest por facilidad

# Semilla para replicación, ponemos nuestra propia semilla :3.
set.seed(2896)

# Número de simulaciones.
n_sim1 <- 1000

#Definimos tamaño muestra

Tamano_muestra <- n_population*p_sample

# Vectores para guardar resultados
se_robusto <- numeric(n_sim1)
se_cluster <- numeric(n_sim1)
betas      <- numeric(n_sim1)

#Construyendo bucle para hacer las estimaciones


for (i in 1:n_sim1) {
  indices <- sample(1:n_population, size = Tamano_muestra)
  MuestraTemporal<- population[indices, ]
  modelo<- feols(Y~D, data = MuestraTemporal)
  #guardamos los coeficientes
  betas[i] <- coef(modelo)["D"]
  #Guardamos los errores estandar
  se_robusto[i]<- se(modelo, se = "hetero")["D"]
  se_cluster[i]<- se(modelo, cluster= ~schools)["D"]
}
#(Para general el bucle se usó ayuda de la IA).

#_____________________________________________________
# Calculando Intervalos de Confianza
#_____________________________________________________

#Definimos valor crítico
```

```r
valor_critico <- 1.96

#Calculando Intervalo Para Errores Robustos
#Limite inferior = \hatBeta- 1.96*se_robusto
ci_robusto_inf <- betas - (valor_critico * se_robusto)
#Limite superior = \hatBeta + 1.96*se_robusto
ci_robusto_sup <- betas + (valor_critico * se_robusto)

#Calculando Intervalos Para Errores Cluster
#Limite inferior = \hatBeta- 1.96*se_cluster
ci_cluster_inf <- betas - (valor_critico * se_cluster)

#Limite superior = \hatBeta + 1.96*se_cluster
ci_cluster_sup <- betas + (valor_critico * se_cluster)

#_____________________________________________________________
# Calculando si los intervalos cubren cero
#_____________________________________________________________
cubre_cero_robusto <- (ci_robusto_inf <= 0) & (ci_robusto_sup >= 0)
cubre_cero_cluster <- (ci_cluster_inf <= 0) & (ci_cluster_sup >= 0)

#_____________________________________________________________
#Calculando promedio de errores para los dos casos
#_____________________________________________________________
Media_se_robusto<-mean(se_robusto)
Media_se_cluster<-mean(se_cluster)
```

Consolidando resultados en la tabla sugerida:

```r
##Creando tabla
library(kableExtra)
library(dplyr)

#Haciedno Data Frame
tabla_final <- data.frame(
  media_rob = Media_se_robusto,
  procent_cobertura_rob   = mean(cubre_cero_robusto) * 100,
  media_clus = Media_se_cluster,
  porcent_cobertura_clus   = mean(cubre_cero_cluster) * 100
)
```

```r
tabla_final %>%
  kbl(
    col.names = c("V Promedio", "% Cobertura", "V Promedio", "% Cobertura"),
    digits = 4,
    caption = "Tabla 1: Cobertura teórica de heterocedasticidad vs clusters",
    align = "c",
    booktabs = TRUE
  ) %>%

  add_header_above(c("Heterocedasticidad" = 2, "Cluster" = 2)) %>%

  kable_styling(
    latex_options = c("hold_position", "scale_down", "striped"),
    bootstrap_options = c("striped", "hover", "condensed"),
    full_width = FALSE,
    position = "center"
  )
```

Table 3: Tabla 1: Cobertura teórica de heterocedasticidad vs clusters

Heterocedasticidad		Cluster	
V Promedio	% Cobertura	V Promedio	% Cobertura
0.0045	94.5	0.1006	100

frente a la pregunta de si siempre es mejor clusterizar al nivel más agregado, la respuesta no es unívoca. Si bien clusterizar a un nivel agregado es necesario para reconocer la estructura de correlación y evitar falsos positivos, esto depende crucialmente del número de grupos (G) disponibles. Como sugiere la literatura, con un G muy pequeño los estimadores de varianza clusterizados se vuelven sesgados e inestables, haciendo que nuestros tests de hipótesis sean poco confiables.