*Taller 2 econometria 2.*

*Profesora: Juliana Helo*

*2025-1*

*Integrantes*

*Jose Ricardo Ricardo Hernandez - 202113889*

*Andres Serrano – 202116783*

*Link del repositorio con código y resultados:* [*https://github.com/jrrHernandez/talleres-econometria-2-en-r.git*](https://github.com/jrrHernandez/talleres-econometria-2-en-r.git)

**Análisis de la Regresión Lineal sobre el Ingreso y los Años de Educación (punto 1)**

En el marco del análisis de datos del **NLS Young Men Cohort** de la Encuesta Nacional Longitudinal (*BaseP1.dta*), nuestro grupo de investigación tiene como objetivo estimar la siguiente regresión lineal:

Yi=α+δSi+γXi+ϵiY\_i = \alpha + \delta S\_i + \gamma X\_i + \epsilon\_iYi​=α+δSi​+γXi​+ϵi​

Donde:

* YiY\_iYi​ representa el logaritmo del ingreso,
* SiS\_iSi​ corresponde a los años de educación,
* XiX\_iXi​ es un vector de características observables utilizadas como variables de control, que incluyen: la experiencia (\textit{exper}), la raza (\textit{black}), el estado civil (\textit{married}) y los años de educación del padre (\textit{fatheduc}) y de la madre (\textit{motheduc}).

A continuación, se abordan los siguientes puntos de análisis:

1. **Obtenga las estadısticas descriptivas de las variables contenidas en la matriz Xi y haga un analisis de estas. Comente para cada una de las variables ¿por que tiene sentido incluirlas como controles? ¿que sucederia si se omiten?**



el mínimo de experiencia son 0 años, mientras que el máximo es 22. En promedio se obtiene 8 años.

Para la educación del padre y la madre, su media es de 10 años y 12 años respectivamente. El padre con mayor experiencia cuenta con 18 años, mientras que el padre que cuenta con menor experiencia de cero años.

Igualmente, tanto la madre con mayor años de experiencia como la madre con menos años de experiencia es igual a la del padre.



El número de personas que no son de raza negra conforman el 84.14%, mientras que el 15.8% sí son de esta raza.

El estado civil de casado conforma el 72.07% de la muestra, mientras que el 27.9% no está casado.

¿Qué pasa si se omiten datos? Podemos tener un problema de variable omitida, donde el error captura el efecto de la variable omitida. Considerando lo anterior, si omitimos las variables, podemos tener un problema de endogeneidad.

1. **Considere que en el error se encuentra contenida la habilidad innata de la persona, la cual es una variable no observable. Explique con detalle qu´e implicaciones tiene lo anterior para la consistencia del estimador ˆ δ. Utilice todas las herramientas te´ oricas y gr´ aficas que le resulten ´ útiles**

****

Dado que hay otra(s) variables que expliquen el logaritmo del ingreso (habilidad innata), hay algo en el error que tiene relación con los años de educación.

Por lo tanto, hay un problema de endogeneidad. Al existir un problema de endogeneidad, la estimación de los años de educación está sesgada y es inconsistente.

¿Cómo probarlo?

Forma teórica:

1. Evaluamos el modelo incorrecto en el estimador de β₁.
2. Utilizamos propiedades y llegamos a sesgamiento e inconsistencia.
3. Si el efecto de la habilidad innata es positivo respecto a los años de educación, entonces estamos sobreestimando. Subestimamos cuando sucede lo contrario.

Forma empírica:

1. Sabemos que al existir endogeneidad sucede que E[u∗∣s]≠0, donde u∗=u+β2⋅Habilidad innata.
2. Utilizamos el test de correlación en R y obtenemos que existe una relación positiva.
3. ***Teniendo en cuenta que los años de educación son endógenos en la regresión (1), su grupo investigador decide implementar el método de variables instrumentales. Explique con detalle en qué consiste este método y cómo puede ayudar a solucionar la endogeneidad de la variable de años de educación y capturar su efecto.***

El método de variables instrumentales nos permite eliminar el sesgo generado por una variable endógena.

Cuando utilizamos una variable instrumental, buscamos eliminar la endogeneidad mediante una variable instrumental Zi​. Esta variable debe cumplir dos características fundamentales al mismo tiempo: relevancia y exogeneidad.

* Relevancia: Existe una relación entre la variable endógena y el instrumento.
* Exogeneidad: El instrumento Zi​ no tiene relación con el error.

Por lo tanto, la interacción entre Zi​ solo afecta a Yi a través de la variable endógena, garantizando así la exogeneidad.

1. ***Usted y su equipo investigador deciden utilizar como instrumento una variable dicotómica que señale si hay universidades en el condado (nearc4). Intuitivamente, comente por qué esta variable podría servirle de instrumento. Relacione su respuesta directamente con los supuestos que debe cumplir el instrumento.***

Cuando utilizamos la variable dicotómica *nearc4*, podemos concluir que el instrumento cumple las dos propiedades fundamentales.

**Relevancia:** Existe una relación entre los años de educación y la presencia de una universidad cerca del condado. En R se realiza lo siguiente:

cov(base1df$nearc4,base1df$educ)=0.1507828>0

**Exogeneidad:** En el caso anterior, concluimos que hay endogeneidad cuando E[u∗∣s]≠0, donde

u∗=u+β2⋅Habilidad innata

Por lo tanto, se puede argumentar que si un condado cuenta con una universidad, esto no necesariamente afecta la habilidad innata.

1. **Obtenga la regresión de la ecuación (1) y obtenga la regresión por medio del método de variables instrumentales. En una misma tabla, reporte los resultados de ambas regresiones, compárelos y coméntelos detalladamente. En específico:**

**¿Es coherente el resultado con el sentido del sesgo por variable omitida? Es decir, ¿se cree que la omisión de la habilidad innata sobreestimaría o subestimaría el rendimiento de la educación sobre los ingresos?**

****

Al comparar los estimadores de las 2 regresiones podemos notar que el estimador de la educación en OLS es mayor al de variables instrumentales. Por lo tanto, estaríamos sobreestimando con variable omitida.

1. **Reporte los resultados de la primera etapa del método MC2E (2SLS). Comente sus resultados a la luz del supuesto de relevancia del instrumento.**

****

En primera etapa queremos saber qué tan relevante es el instrumento respecto a la variable endógena. De acuerdo con la tabla, podemos ver un F-statistic de 343.619 y un p-valor menor a 0.01, lo cual quiere decir que nuestro instrumento es estadísticamente significativo

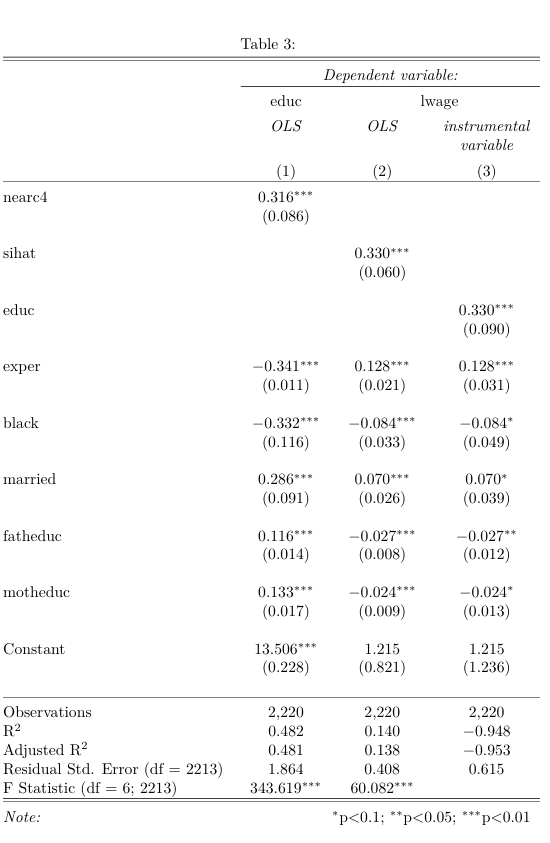
1. **Escriba y corra la forma reducida para este caso. Muestre que el resultado de la metodología de variables instrumentales es equivalente a dividir el estimador de la forma reducida entre el de la primera etapa. Explique de forma intuitiva este resultado.**

****

* De acuerdo con los resultados anteriores:
* Estimador por medio de IV (Instrumental Variables - ivreg): 0.330
* Estimador por medio de la primera etapa: 0.316
* Estimador de la forma reducida: 0.104

Al estimar LATE: 0.104/0.316 == 0.3291139

1. **Calcule de nuevo la regresión por medio de MC2E, pero de forma manual, y compare los resultados tanto de los estimadores como de los errores estándar.**

****

Todos los coefiecientes son estadísticamente signicativos puesto que sus errores estándar correspondientes, multiplicado por 2 son menores al estimador.

***Punto2***

***Usted y su grupo de investigación se encuentran analizando la forma en que la representación política puede tener efectos sobre la violencia. Para este fin, usted cuenta con una base de datos con los resultados de las elecciones a alcaldía de varios municipios en Colombia entre 1997 y 2011. Particularmente, cuenta con la proporción de votos que capturó el partido de izquierda en ese municipio (vote share). Además, cuenta con un indicador del número de ataques de grupos paramilitares en los 4 años siguientes a la elección (parvio ip4 pc).***

***En este caso, la pregunta clave es si la elección de alcaldes pertenecientes a partidos de izquierda, es decir, aquellos que militan en grupos que han sido históricamente excluidos del poder local en Colombia, conduce a un incremento en los ataques violentos de paramilitares.***

* 1. ***Por medio de una regresión lineal simple, muestre cuánto cambian los ataques paramilitares cuando los alcaldes electos son de partidos asociados a la izquierda (left winner). ¿Este resultado permite concluir algo frente a la pregunta clave antes planteada?  
     Pista: Piense en los municipios con alcalde de izquierda electo como un grupo de tratamiento y en los municipios con alcalde de otra vertiente política como un grupo de control en torno a la construcción de un contrafactual.***

Tabla

El contenido generado por IA puede ser incorrecto.

La relación es negativa. Cuando un candidato de izquierda gana las elecciones, los incidentes violentos por los paramilitares disminuyen, en promedio, 0,7 puntos (aparentemente)

* 1. ***Ensutrabajo, Fergusson et al. (2021) utilizan como running variable la proporción de votos que capturó el partido de izquierda en cada municipio. Con esto, los autores aprovechan la variación exógena en las victorias de izquierda para verificar el efecto sobre la violencia. Muestre gráficamente que, siempre que la proporción de votantes de izquierda sea mayor a 0, entonces el/la alcalde electo será de izquierda y, en el caso contrario, será de otra vertiente política. Explique por qué esto permite utilizar una regresión discontinua aguda.***

******

Vemos que hay perfect compliance. Por lo tanto, podemos utilizar una regresión discontinua aguda***.***

* 1. ***Para toda la muestra:***
     + ***Grafique la relación existente entre la proporción de votantes y el número de ataques paramilitares.***
     + ***Corra un RDD utilizando una función lineal y una función cuadrática para la running variable sin interacciones.***
     + ***Interaccione la running variable con la variable de interés, tanto para la forma funcional lineal como para la cuadrática.***
     + ***Muestre todos los resultados en una sola tabla y coméntelos, indicando qué puede concluir de ellos.***

|  | Modelo RDD lineal | Modelo RDD cuadrático | Modelo RDD con interacción | Modelo RDD con interacción y cuadrático |
| --- | --- | --- | --- | --- |
| (Intercept) | 0.649 | 0.796 | 0.649 | 1.278 |
|  | (0.952) | (1.268) | (1.157) | (1.177) |
| D | 1.622 | 2.884 |  |  |
|  | (1.386) | (1.864) |  |  |
| vote\_share | -18.430 | -15.547 |  |  |
|  | (6.236) | (17.625) |  |  |
| vote\_share | 13.264 | -25.045 |  |  |
| Right | (10.993) | (30.521) |  |  |
| vote\_share\_sq |  | 6.866 |  |  |
|  |  | (39.271) |  |  |
| vote\_share\_sq |  | 113.432 |  |  |
| Right |  | (88.080) |  |  |
| vote\_share |  |  | -18.430 | -6.081 |
|  |  |  | (14.811) | (17.199) |
| left\_winner |  |  | 1.622 | 1.338 |
|  |  |  | (1.453) | (1.386) |
| vote\_share × left\_winner |  |  | 13.264 | -7.747 |
|  |  |  | (16.195) | (25.486) |
| vote\_share\_sq |  |  |  | 29.415 |
|  |  |  |  | (34.206) |
| Num.Obs. | 254 | 254 | 254 | 254 |
| R2 | 0.037 | 0.046 | 0.037 | 0.040 |
| R2 Adj. | 0.026 | 0.027 | 0.026 | 0.025 |
| AIC | 1763.2 | 1764.8 | 1763.2 | 1764.5 |
| BIC | 1780.9 | 1789.5 | 1780.9 | 1785.7 |
| Log.Lik. | -876.589 | -875.387 |  |  |
| RMSE | 7.63 | 7.59 | 7.63 | 7.62 |

Los coeficientes asociados a la variable dummy left\_winner y sus interacciones no son estadísticamente significativos. Adicionalmente, las desviaciones estándar son considerablemente altas. Por lo cual, no hay una relación clara y consistente entre la relación de un alcalde de izquierda y el numero de ataques paramilitares. En paralelo, el signo de los coeficientes cambia entre modelos, lo cual refuerza la idea de una carencia de tendencia clara.

* 1. ***El choque exógeno proviene de discontinuidades en la running variable. Por tanto, los autores argumentan que, justo alrededor del umbral (en el margen), la victoria de un alcalde de izquierda ocurre prácticamente de manera aleatoria. Repita los 3 puntos del literal c) pero estableciendo un ancho de banda de ±5 puntos porcentuales alrededor del umbral de proporción del electorado del 50% (0.5). Explique también cuál es el trade-off al que se enfrenta al momento de definir un ancho de banda.***

|  | Modelo RDD lineal | Modelo RDD cuadrático | Modelo RDD con interacción | Modelo RDD con interacción y cuadrático |
| --- | --- | --- | --- | --- |
| (Intercept) | 0.649 | 0.796 | 0.649 | 1.278 |
|  | (0.952) | (1.268) | (1.157) | (1.177) |
| D | 1.622 | 2.884 |  |  |
|  | (1.386) | (1.864) |  |  |
| vote\_share | -18.430 | -15.547 |  |  |
|  | (6.236) | (17.625) |  |  |
| vote\_share | 13.264 | -25.045 |  |  |
| Right | (10.993) | (30.521) |  |  |
| vote\_share\_sq |  | 6.866 |  |  |
|  |  | (39.271) |  |  |
| vote\_share \_sq\_right |  | 113.432 |  |  |
|  |  | (88.080) |  |  |
| vote\_share |  |  | -18.430 | -6.081 |
|  |  |  | (14.811) | (17.199) |
| left\_winner |  |  | 1.622 | 1.338 |
|  |  |  | (1.453) | (1.386) |
| vote\_share × left\_winner |  |  | 13.264 | -7.747 |
|  |  |  | (16.195) | (25.486) |
| vote\_share\_sq |  |  |  | 29.415 |
|  |  |  |  | (34.206) |
| Num.Obs. | 254 | 254 | 254 | 254 |
| R2 | 0.037 | 0.046 | 0.037 | 0.040 |
| R2 Adj. | 0.026 | 0.027 | 0.026 | 0.025 |
| AIC | 1763.2 | 1764.8 | 1763.2 | 1764.5 |
| BIC | 1780.9 | 1789.5 | 1780.9 | 1785.7 |
| Log.Lik. | -876.589 | -875.387 |  |  |
| RMSE | 7.63 | 7.59 | 7.63 | 7.62 |

Al utilizar un ancho de banda grande, se utilizan mas datos. Por lo cual, la varianza de los estimadores va a ser más pequeña. Sin embargo, al incluir mas observaciones alejadas del umbral podemos tener estimadores más sesgados.

Por otra parte, un ancho de banda grande, reduce la cantidad de datos. Por lo cual, la varianza de los estimadores va a aumentar. No obstante, la estimación local va a estar menos sesgada y será mas precisa.

* 1. ***Ahora, repita nuevamente el literal c) pero con un ancho de banda de ±10 puntos porcentuales. Compare los resultados obtenidos en los últimos 3 literales. ¿Qué puede concluir frente a la pregunta clave?***

|  | Modelo RDD lineal | Modelo RDD cuadrático | Modelo RDD con interacción | Modelo RDD con interacción y cuadrático |
| --- | --- | --- | --- | --- |
| (Intercept) | 1.167 | 0.796 | 1.167 | 0.865 |
|  | (1.237) | (1.268) | (0.974) | (0.985) |
| D | 2.622 | 2.884 |  |  |
|  | (1.738) | (1.864) |  |  |
| vote\_share | -14.865 | -15.547 |  |  |
|  | (24.376) | (17.625) |  |  |
| vote\_share | -25.004 | -25.045 |  |  |
| Right | (33.723) | (30.521) |  |  |
| vote\_share\_sq |  | 6.866 |  |  |
|  |  | (39.271) |  |  |
| vote\_share\_sq  right |  | 113.432 |  |  |
|  |  | (88.080) |  |  |
| vote\_share |  |  | -14.865 | -36.014 |
|  |  |  | (16.474) | (55.085) |
| left\_winner |  |  | 2.622 | 2.563 |
|  |  |  | (1.788) | (1.873) |
| vote\_share × left\_winner |  |  | -25.004 | 18.107 |
|  |  |  | (28.335) | (120.232) |
| vote\_share\_sq |  |  |  | -223.159 |
|  |  |  |  | (537.085) |
| Num.Obs. | 167 | 254 | 167 | 167 |
| R2 | 0.020 | 0.046 | 0.020 | 0.021 |
| R2 Adj. | 0.002 | 0.027 | 0.002 | -0.003 |
| AIC | 1090.1 | 1764.8 | 1090.1 | 1092.0 |
| BIC | 1105.7 | 1789.5 | 1105.7 | 1110.7 |
| Log.Lik. | -540.051 | -875.387 |  |  |
| RMSE | 6.14 | 7.59 | 6.14 | 6.14 |

Al comparar los estimadores de cada ancho de banda se puede concluir que no existe una evidencia significativamente robusta que afirme que la elección de alcaldes de izquierda conduzca a un incremento en los ataques paramilitares. Esto se debe a la alteración del signo del estimador entre modelos.

***Punto 3***

***En un trabajo de Leite y Rizzi (2014) los autores buscan evaluar la hipótesis de la ignorancia racional (Downs, 1957). De acuerdo con esta hipótesis, en elecciones grandes los individuos reconocen que su voto tiene una probabilidad considerablemente baja de ser decisivo y, por lo tanto, en muchos casos no incurren en costos asociados a tener un voto informado.***

***Para lo anterior, Leite y Rizzi (2014) analizan el caso de Brasil. En este país, el voto es obligatorio para las personas que tienen 18 años o más y es voluntario para quienes son menores. Cargue la BaseP3.dta para el desarrollo de los siguientes literales.***

* 1. ***Ustedes deciden utilizar el puntaje en un quiz de conocimiento político como variable de resultado (politicalquiz). Estandaricen esta variable con la función std(). Recuerden que, para aplicar una función en la generación de una variable, debe utilizarse el comando egen.***

En código

* 1. ***Expliquen: ¿por qué tiene sentido utilizar la edad como running variable para un diseño de regresión discontinua en este caso?***

Dado que la edad es un factor fundamental a la hora de votar, puesto que al cumplir 18 años es obligatorio, la edad nos sirve como running variable para estimar el salto en la probabilidad de votar a medida que esta aumenta.

* 1. ***¿Qué inconvenientes tendría llevar a cabo una Regresión Discontinua Aguda? Explique intuitivamente y apóyese en un gráfico en el que se muestre el salto en la probabilidad de votar cuando se cumplen los 18 años. Puede utilizar herramientas estadísticas y gráficas adicionales que considere útiles.***

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

Podemos ver que no hay perfect compliance, pues a medida que aumenta la edad, la probabilidad de votar vota igualmente.

* 1. **Ahora, creen una variable bajo el nombre compulsoryvote que tome valor de 1 para aquellas personas cuyo voto es obligatorio y 0 para aquellas cuyo voto es voluntario. ¿Por qué tendría sentido utilizar esta variable como instrumento en una Regresión Discontinua Borrosa?**
* En código. Explicación: os individuos no pueden manipular su edad, es una variable continua y además existe una discontinuidad entre votar y si cumple con 18 años

***e) Para correr la regresión discontinua borrosa, compruebe el cumplimiento de los siguientes supuestos. Utilicen solo aquellas personas cuyo cumpleaños está a máximo 6 meses de distancia de las elecciones (sample6m), así como asuman una función lineal de la edad recentrada (diasalt1) y su interacción con vote.***

***En la vecindad del punto de corte, hay discontinuidad en la probabilidad de votar.***

***Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.***

***Hay continuidad local para las siguientes covariables: white, female, college mother, livewithparents y work. Utilice la variable categ´ orica del nivel escolar como control para todos los casos. [REALIZADO EN EL CODIGO]***

***No hay manipulación en la variable de focalización.***

***Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.***

La densidad es similar en el punto de corte, es decir, no hay manipulación del instrumento

***f) Corran una regresión por MCO, una Regresión Discontinua Aguda y una Regresión Discontinua Borrosa, cuya variable Y sea el puntaje del quiz político estandarizado y su variable X sea haber votado (vote). Para todos los casos, utilicen las variables escola, votedbefore2010, white, female y collegemother como controles. Además, para las regresiones discontinuas, utilicen solo aquellas personas cuyo cumpleaños está a máximo 6 meses de distancia de las elecciones (sample6m), así como asuman una función lineal de la edad recentrada (diasalt1) y su interacción con vote. Reporten los resultados de las 3 regresiones en una sola tabla.***

| Variable | MCO | Sharp\_RD | Fuzzy\_RD |
| --- | --- | --- | --- |
| (Intercept) | -0.62821666 | -0.7507365305 | -0.15550579 |
| vote | 0.46068725 | 0.4577369450 |  |
| escola | 0.03516437 | 0.0277533302 |  |
| votedbefore2010 | 0.45358570 | 0.5485922042 |  |
| female | -0.15062796 | -0.1396044876 |  |
| collegemother\_y | 0.33249296 | 0.3081859067 |  |
| diasalt\_1 |  | -0.0002034587 | 0.00600224 |
| white |  | 0.1898346504 |  |
| vote:diasalt\_1 |  | 0.0005738256 |  |
| diasalt\_1:vote |  |  |  |

***g) Interprete el resultado para la variable de interés (vote) para cada uno de los 3 casos anteriores y compárelos. En la comparación, lo más importante es la intuición. Por tanto, tenga presente cuáles son las poblaciones que se comparan en cada metodología y explique por qué la Regresión Discontinua Borrosa es la más confiable. En específico, recuerde que en una Regresión Discontinua Borrosa se obtiene un estimador LATE que es local en dos sentidos: 1) individuos en la vecindad del punto de corte; 2) para los compliers.***

Mínimos Cuadrados Ordinarios (MCO)

El coeficiente de vote en MCO es 0.4607.

Este método estima la relación promedio entre votar y la variable de interés en toda la muestra. Suponiendo que todas las variables relevantes están incluidas y no hay sesgo de selección, MCO podría dar una buena aproximación, pero esto rara vez ocurre en datos observacionales.

Regresión Discontinua Sharp (Sharp RD)

El coeficiente de vote en este modelo es 0.4577, muy cercano al de MCO.

Este método aprovecha el punto de corte de la variable running variable (e.g., una puntuación de elegibilidad) para estimar el efecto causal en los individuos justo en torno al umbral.

Su fortaleza es que, bajo supuestos de continuidad, las diferencias en vote pueden interpretarse como efectos causales locales.

Regresión Discontinua Borrosa (Fuzzy RD)

En este caso, no se reporta directamente un coeficiente para vote, lo que sugiere que el modelo utiliza vote como una variable instrumental en una estimación de variables instrumentales (IV).

A diferencia de Sharp RD, donde el cumplimiento de la regla de asignación es perfecto, aquí hay incumplimiento (algunos individuos por debajo del umbral reciben el tratamiento y algunos por encima no lo reciben).

La estimación de Fuzzy RD entrega un LATE (Local Average Treatment Effect), que es local en dos sentidos:

Se estima solo en una vecindad cercana al punto de corte, lo que limita la extrapolación.

Se aplica únicamente a los compliers, es decir, aquellos individuos cuya participación en el tratamiento (vote) depende del umbral de elegibilidad***.***

* 1. ***Con lo anterior, en respuesta a la pregunta: ¿El hecho de votar tiene un efecto sobre el conocimiento político? Explique cómo el estimador LATE le permite llegar a esta conclusión. ¿Qué podría decir frente a la hipótesis de la ignorancia racional?***

Con base en la estimación mediante Regresión Discontinua Borrosa (Fuzzy RD), podemos argumentar que votar sí tiene un efecto sobre el conocimiento político, pero esta conclusión aplica específicamente a los compliers en la vecindad del punto de corte.

**Efecto Local**:

* No estimamos un efecto promedio en toda la población, sino solo para los individuos que se encuentran cerca del punto de corte (e.g., elegibilidad para votar).
* Esto significa que el resultado no es necesariamente generalizable a toda la población, sino solo a aquellos cuyo comportamiento respecto a votar está determinado por el umbral.
* Solo observamos el impacto en los compliers, es decir, los individuos que votan únicamente porque fueron asignados al grupo de tratamiento