Tarea 2
IN5244-1 Ciencia de datos

Cristopher Urbina [1]

Abstract

Este reporte evalúa el impacto de la política de reportes del MINEDUC, implementada el 28 de junio de 2022, en la revinculación y graduación de estudiantes desvinculados del sistema escolar. La variable de interés, ListedInDropoutReport, indica si un estudiante fue listado en el reporte enviado a las escuelas, y se estudia su relación con la probabilidad de reingreso a la educación y graduación al final del año. Se utiliza un enfoque de regresión discontinua (RD) para analizar los efectos de esta política, con SchoolDaysFromMay302022 como variable continua que permite evaluar discontinuidades en la probabilidad de inclusión en el reporte.

1. Pregunta 1:

Al dividir el conjunto de valores en torno al cutoff=0, se evidencia un cambio de tendencia claro, junto con una menor dispersión en los valores (ver figura 1).

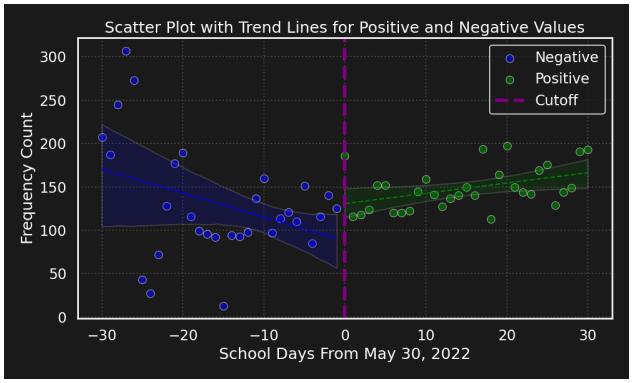

Figure 1. Gráfico de frecuencias antes y despues de la fecha de corte de asignación

Los resultados del test sugieren una manipulación significativa de la variable ‘running‘ alrededor del punto de corte, ya que se observó que los individuos a la derecha del corte presentaron una frecuencia promedio superior en aproximadamente 43.44 unidades en comparación con aquellos a la izquierda, con un valor de p menor a 0.001.

Table 1. Resultados de la regresión OLS para testear manipulación en variable running

VARIABLE	COEF.	STD	T	$P > \|t\|$
INTERCEPT	134.9460	1.260	107.113	0.000
IS_RIGHT	43.4399	2.171	20.013	0.000
SCHOOLDAYS	-1.6515	0.058	-28.556	0.000

2. Pregunta 2:

A) El gráfico de primera etapa es esencial en el análisis de regresión discontinua, ya que permite visualizar la relación entre la variable dependiente (ListedInDropoutReport) y la variable de corte (SchoolDaysFromMay302022). Al superponer líneas de proyección lineal a ambos lados del punto de corte, se pueden identificar discontinuidades que indican cómo la intervención afecta la variable de interés.

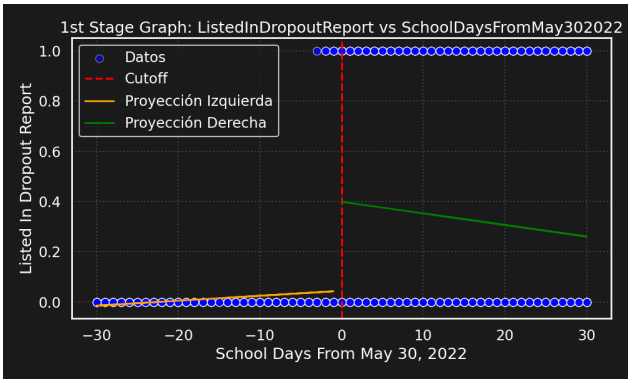

Figure 2. Gráfico de primera etapa, ListedInReport vs SchoolsDaysFromMay302022

B) Los resultados de la regresión indican que la variable ‘DroppedOutMay30Before‘ tiene un coeficiente de 0.3395 (p ¡ 0.001), lo que sugiere que ser listado como ”dropped out” antes del 30 de mayo aumenta significativamente la

probabilidad de estar en el reporte. Por otro lado, la variable 'SchoolDaysFromMay302022' muestra un coeficiente de 0.0027 (p ¡ 0.001), indicando que cada día escolar adicional incrementa la probabilidad de estar listado en el reporte en 0.27%. Sin embargo, la interacción entre estas dos variables presenta un coeficiente de -0.0077 (p ¡ 0.001), sugiriendo que el efecto positivo de ser "dropped out" disminuye ligeramente con cada día escolar adicional. Estos resultados sugieren que la probabilidad de estar listado en el reporte de deserción es estadísticamente distinta de cero en torno al punto de corte, destacando la importancia de las intervenciones en este contexto.

Table 2. Resultados de la Regresión de 'ListedInDropoutReport'

VARIABLE	COEF.	ERR. EST.	$P > \|t\|$
DROPOUT30	0.3395	0.018	0.000
SCHOOLDAYS	0.0027	0.001	0.000
INTERACTION	-0.0077	0.001	0.000

Al agregar tambien un efecto fijo por dia de la semana, ya que se observa una clara tendencia (ver figura):

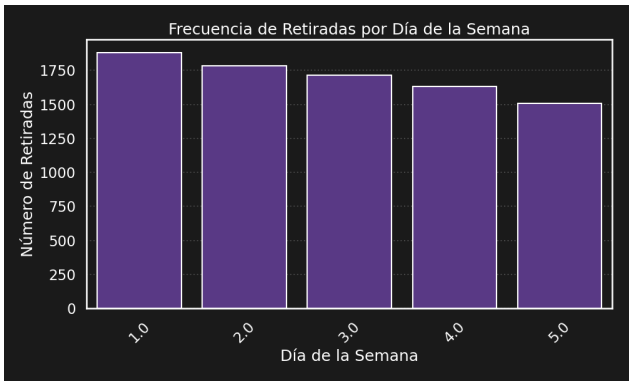

Figure 3. Gráfico de frecuencias de deserción según dia de la semana, 1 - 5 (Lunes - Viernes).

Table 3. Resultados de la regresión

VARIABLE	COEF.	ERR. EST.	$P > \|t\|$
DROPOUT30	0.3419	0.018	0.000
SCHOOLDAYS	0.0026	0.001	0.000
INTERACTION	-0.0077	0.001	0.000

El análisis de los resultados revela que el coeficiente de **DROPOUT30** ha aumentado ligeramente a 0.3419, lo que sigue indicando que la probabilidad de estar listado como "dropped out" es alta, manteniendo una significancia estadística (p ¡ 0.001). Por otro lado, el coeficiente de **SCHOOLDAYS** ha disminuido a 0.0026, lo que implica que cada día escolar adicional incrementa la probabilidad de estar listado en el reporte en 0.26%. A pesar de esta ligera reducción, el coeficiente sigue siendo estadísticamente significativo. En cuanto a la **interacción** entre las variables, su coeficiente se mantiene constante en -0.0077, indicando que el efecto negativo sobre la probabilidad de estar listado sigue siendo consistente en ambos modelos.

Al comparar los efectos, se observa que la inclusión de efectos fijos por día de la semana ha llevado a un ligero aumento en el coeficiente de **DROPOUT30**, sugiriendo que considerar el día de la semana puede revelar factores adicionales que influyen en la probabilidad de deserción. Aunque el coeficiente de **SCHOOLDAYS** ha disminuido ligeramente al agregar estos efectos fijos, continúa indicando que cada día escolar adicional tiene un efecto positivo en la probabilidad de estar listado en el reporte. Esto sugiere que el día de la semana puede tener un efecto moderador relevante en la interpretación de los días escolares. Finalmente, el coeficiente de interacción se mantiene constante, lo que implica que la relación entre **DROPOUT30** y **SCHOOLDAYS** es robusta a pesar de la inclusión de efectos fijos adicionales.

3. Pregunta 3:

Los resultados de la regresión que estima el impacto de haberse retirado de la escuela antes del cutoff sobre la probabilidad de estar revinculado en la escuela al 31 de agosto muestran lo siguiente:

DROPOUT30: El coeficiente es de 0.0485 con un error estándar de 0.021, lo que indica que ser listado como "dropped out" antes del 30 de mayo incrementa la probabilidad de estar revinculado en la escuela en aproximadamente 4.85%. Este resultado es estadísticamente significativo ($p < 0.05$), lo que sugiere un efecto positivo de la deserción sobre la revinculación.

SCHOOLDAYS: El coeficiente de esta variable es de -0.0017 con un error estándar de 0.001, lo que implica que cada día escolar adicional disminuye la probabilidad de revinculación en aproximadamente 0.17%. Este resultado también es estadísticamente significativo ($p < 0.05$), sugiriendo que la asistencia a la escuela tiene un efecto negativo sobre la revinculación.

INTERACCIÓN: El coeficiente de la interacción es 0.0007 con un error estándar de 0.001 y un valor p de 0.536. Esto indica que no hay un efecto significativo de la interacción entre las variables, lo que sugiere que el impacto de ser "dropped out" no se ve afectado por el número de días escolares adicionales.

Los resultados aquí muestran que, aunque la probabilidad de revinculación aumenta para aquellos que se retiraron antes

del cutoff, la influencia de la cantidad de días desde el cutoff es negativa lo que indica que aumenta la probabilidad para aquellos que se retiraron antes del 30 de Mayo.

Table 4. Resultados de la Regresión de Revinculación

VARIABLE	COEF.	ERR. EST.	$P > \|t\|$
DROPOUT30	0.0485	0.021	0.019
SCHOOLDAYS	-0.0017	0.001	0.039
INTERACCIÓN	0.0007	0.001	0.536

Por otro lado, los resultados de la regresión que estima el impacto de haberse retirado de la escuela antes del cutoff sobre la probabilidad de graduarse en 2022 presentan los siguientes hallazgos:

DROPOUT30: El coeficiente es de 0.0395 con un error estándar de 0.023, lo que indica que ser listado como "dropped out" antes del 30 de mayo incrementa la probabilidad de graduarse en aproximadamente 3.95%. Aunque el resultado no es estadísticamente significativo al nivel del 5% ($p = 0.086$), se sugiere una tendencia positiva que podría ser relevante en un análisis más amplio.

SCHOOLDAYS : El coeficiente de esta variable es de -0.0027 con un error estándar de 0.001, indicando que cada día escolar adicional disminuye la probabilidad de graduación en aproximadamente 0.27%. Este resultado es estadísticamente significativo ($p < 0.01$), sugiriendo que la asistencia a la escuela tiene un efecto negativo sobre la graduación.

INTERACCIÓN : El coeficiente de la interacción es 0.0027 con un error estándar de 0.001 y un valor p de 0.024. Esto indica que el efecto positivo de haberse retirado antes del cutoff se incrementa con cada día escolar adicional, sugiriendo una interacción significativa entre las variables.

Al comparar estos resultados con los de la regresión anterior que evaluó la revinculación, observamos que la variable de deserción tiene un coeficiente positivo y un efecto significativo, aunque menor en comparación con el impacto sobre la revinculación. La variable de días escolares, aunque negativa, muestra un impacto más fuerte en la graduación para aquellos que se retiran antes del 30 de Mayo. Por último, la interacción entre la deserción y los días escolares se vuelve significativa en este contexto, indicando que el impacto de ser "dropped out" puede ser más positivo para aquellos se retiran mucho antes del 30 de Mayo.

Table 5. Resultados de la Regresión de Graduación en 2022

VARIABLE	COEF.	ERR. EST.	$P > \|t\|$
DROPOUT30	0.0395	0.023	0.086
SCHOOLDAYS	-0.0027	0.001	0.002
INTERACCIÓN	0.0027	0.001	0.024

4. Pregunta 4:

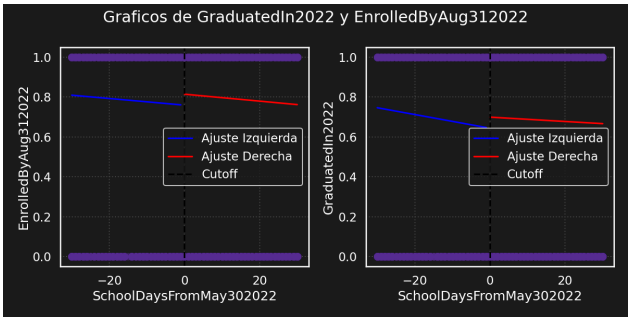

Figure 4. Graficos de GraduatedIn2022 y EnrolledByAug312022 en función de SchoolsDaysFromMay302022.

En ambos gráficos se puede observar la probabilidad de estar inscrito al 31 de agosto de 2022 y la probabilidad de graduarse en 2022, respectivamente. A ambos lados del cutoff, la tendencia general es negativa, lo que indica que la probabilidad de mantenerse inscrito o de graduarse disminuye a medida que se aleja del punto de corte (lado derecho) y se acerca al punto de corte (lado izquierdo). Sin embargo, en ambos casos, se observa un aumento en la probabilidad en la discontinuidad en el cutoff (0), lo que sugiere un cambio abrupto en este punto.

En el gráfico de la izquierda (inscripción), a medida que nos acercamos al cutoff desde la izquierda, la probabilidad de estar inscrito disminuye. Después del cutoff, aunque hay un aumento en la probabilidad justo en el punto de corte, esta vuelve a disminuir a medida que nos alejamos hacia la derecha. De manera similar, en el gráfico de la derecha (graduación), la probabilidad de graduarse también muestra un comportamiento decreciente antes y después del cutoff, aunque el salto en la discontinuidad es más notable.

En ambos casos, el ajuste lineal muestra que la probabilidad se mantiene más alta justo después del cutoff, lo que podría sugerir que el evento ocurrido en esta fecha influye positivamente en los resultados, pero su efecto se atenúa a medida que el tiempo pasa.

5. Pregunta 5:

Table 6. Resultados de la Regresión de Revinculación

VARIABLE	COEF.	ERR. EST.	$P > \|t\|$
DROPOUT30	0.0448	0.031	0.152
SCHOOLDAYS	-0.0002	0.004	0.958
INTERACTION	-0.0017	0.005	0.713
SCHOOLDAYS^2	4.782E-05	0.000	0.665
INTERACTION^2	-2.014E-05	0.000	0.889

Los resultados de la regresión de revinculación se presentan en la Tabla 6. El coeficiente para la variable DroppedOutMay30Before es de 0.0448 con un error estándar de 0.031 y un valor p de 0.152, lo que sugiere que la relación entre haberse retirado antes del cutoff y la revinculación no es estadísticamente significativa. Por otro lado, la variable SchoolDaysFromMay302022 muestra un coeficiente de -0.0002 con un error estándar de 0.004 y un valor p de 0.958, indicando que no hay una relación significativa entre los días escolares y la revinculación. La interacción entre DroppedOutMay30Before y SchoolDaysFromMay302022 tiene un coeficiente de -0.0017 (error estándar: 0.005, valor p: 0.713), lo que también sugiere que no hay una influencia significativa de esta interacción en la revinculación. Finalmente, el término cuadrático SchoolDaysFromMay302022² presenta un coeficiente de 4.782e-05 con un error estándar de 0.000 y un valor p de 0.665, y la interacción cuadrática tiene un coeficiente de -2.014e-05 (error estándar: 0.000, valor p: 0.889), lo que confirma que la forma funcional cuadrática no aporta evidencia significativa para explicar la revinculación.

Table 7. Resultados de la Regresión de Graduación en 2022

VARIABLE	COEF.	ERR. EST.	$P > \|t\|$
DROPOUT30	0.0457	0.035	0.190
SCHOOLDAYS	-0.0013	0.004	0.755
INTERACTION	-0.0021	0.005	0.690
SCHOOLDAYS^2	4.214E-05	0.000	0.732
INTERACTION²	6.827E-05	0.000	0.672

Los resultados de la regresión para la variable GraduatedIn2022 se presentan en la Tabla 8. El coeficiente para la variable DroppedOutMay30Before es de 0.0457, con un error estándar de 0.035 y un valor p de 0.190, lo que sugiere que la relación entre haberse retirado antes del cutoff y la graduación no es estadísticamente significativa. La variable SchoolDaysFromMay302022 muestra un coeficiente de -0.0013 (error estándar: 0.004, valor p: 0.755), indicando que no hay una relación significativa entre los días escolares y la graduación. La interacción entre DroppedOutMay30Before y SchoolDaysFromMay302022 presenta un coeficiente de -0.0021 (error estándar: 0.005, valor p: 0.690), lo que también sugiere que esta interacción no tiene un impacto significativo en la graduación. Además, el término cuadrático SchoolDaysFromMay302022² tiene un coeficiente de 4.214e-05 con un error estándar de 0.000 y un valor p de 0.732, y la interacción cuadrática muestra un coeficiente de 6.827e-05 (error estándar: 0.000, valor p: 0.672), confirmando que la forma funcional cuadrática no proporciona evidencia significativa para explicar la graduación en 2022.

Por último, en ambos casos el efecto es sensible a la forma funcional utilizada.

6. Pregunta 6:

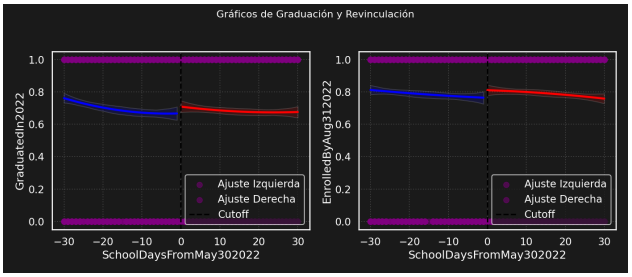

Figure 5. Graficos de GraduatedIn2022 y EnrolledByAug312022 en función de SchoolsDaysFromMay302022 al cuadrado.

Con la proyección cuadrática, la probabilidad de inscribirse en agosto, al lado derecho del punto de corte, muestra un cambio en su tendencia: aunque sigue siendo a la baja, se vuelve más constante en los puntos cercanos al corte. En contraste, la probabilidad de graduarse disminuye de manera más pronunciada al acercarse al punto de corte desde el lado izquierdo. En el lado derecho, se observa un salto inicial en la probabilidad de graduación, que luego se mantiene casi constante, disminuyendo su pendiente de forma gradual.

Al agregar una forma funcional cúbica, las curvas de probabilidad de graduación se aproximan en la discontinuidad (cutoff), lo que indica que la probabilidad de graduarse en torno al punto de corte se vuelve similar para los grupos a ambos lados del mismo. Sin embargo, al analizar la probabilidad de revinculación, se observa que, aunque también tienden a acercarse, persiste una brecha significativa entre ambas proyecciones. Esto sugiere que, a pesar de las similitudes en las probabilidades de graduación, las condiciones que afectan la revinculación son diferentes y requieren un análisis más profundo para comprender los factores subyacentes.

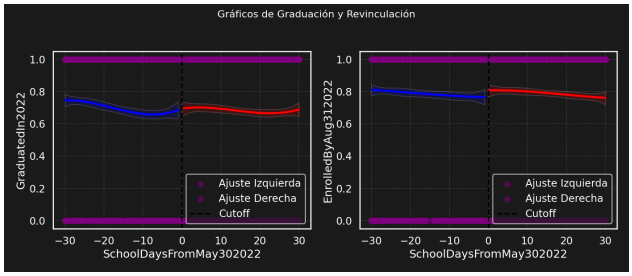

Figure 6. Graficos de GraduatedIn2022 y EnrolledByAug312022 en función de SchoolsDaysFromMay302022 al cubo.

7. Pregunta 7:

Este ejercicio es importante porque permite visualizar cómo la variable de corte (cutoff) impacta en las probabilidades de graduación. Al observar los resultados a cada lado del cutoff, podemos identificar si existe un cambio abrupto en la probabilidad de graduación en el punto de corte. Esto es fundamental en análisis de regresión discontinua, donde el interés radica en entender el efecto causal de una intervención o tratamiento.

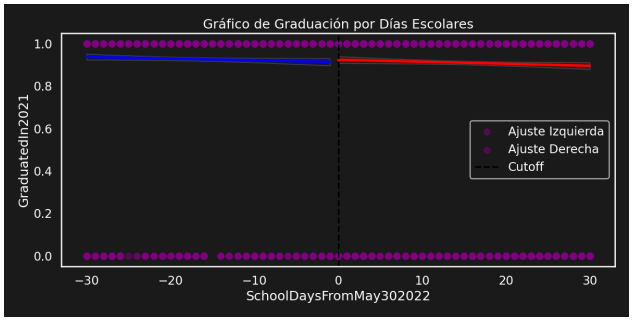

Figure 7. Grafico de GraduatedIn2022 en función de SchoolsDaysFromMay302022.

En el gráfico de graduación se observa la probabilidad de graduarse en 2021 en función de la variable de corte. A ambos lados del cutoff, la tendencia general es negativa, lo que indica que la probabilidad de graduarse disminuye a medida que se aleja del punto de corte por el lado derecho y se acerca al punto de corte por el lado izquierdo. Sin embargo, se evidencia un aumento en la probabilidad en la discontinuidad del cutoff (0), lo que sugiere un cambio abrupto en este punto.

A medida que nos acercamos al cutoff desde la izquierda, la probabilidad de graduarse disminuye. Después del cutoff, aunque se observa un aumento en la probabilidad justo en el punto de corte, esta tendencia a graduarse vuelve a disminuir a medida que nos alejamos hacia la derecha. El ajuste lineal muestra que la probabilidad de graduarse se mantiene más alta justo después del cutoff, lo que podría indicar que el evento relacionado con esta fecha influye positivamente en los resultados. No obstante, su efecto se atenúa con el tiempo.

8. Pregunta 8:

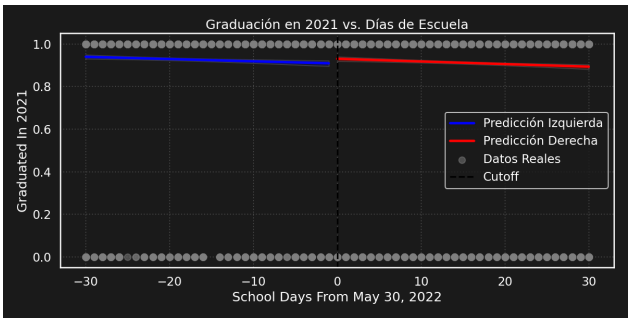

Figure 8. Grafico de GraduatedIn2021 en función de SchoolsDaysFromMay302022.

Al incorporar covariables en el modelo, se observa un ligero cambio en la pendiente de la relación entre la variable de interés y la variable de corte antes del cutoff. Este cambio sugiere que los factores adicionales, como el GPA en 2021, la asistencia, el género y el estatus de migrante, pueden influir en la probabilidad de graduación, afectando la manera en que esta se comporta al aproximarse al punto de corte.

9. Pregunta 9:

El estimador de regresión discontinua fuzzy (RD Fuzzy) se utiliza para estimar el efecto causal de una intervención o tratamiento en situaciones donde el tratamiento no se asigna de manera estricta en el punto de corte, sino que la probabilidad de recibir el tratamiento cambia alrededor de este punto. A continuación se presenta la formulación matemática del estimador RD Fuzzy:

Estimador RD Fuzzy

El estimador de RD Fuzzy se puede expresar matemáticamente como:

$$\hat{\tau}_{RD} = \frac{E[Y(1)|X = c] - E[Y(0)|X = c]}{E[D|X = c] - E[D|X < c]}$$

DONDE:

- $Y(1)$ es el resultado cuando se recibe el tratamiento (en este caso, estar listado en el reporte).

- $Y(0)$ es el resultado cuando no se recibe el tratamiento.
- X es la variable de ejecución (running variable).
- c es el punto de corte.
- D es una variable indicadora que toma el valor de 1 si un individuo recibe el tratamiento (es listado en el reporte) y 0 en caso contrario.

Numerador y Denominador

- **Numerador:**

$$E[Y(1)|X = c] - E[Y(0)|X = c]$$

Representa la diferencia en la expectativa del resultado Y (revinculación escolar) en el punto de corte c entre los que reciben el tratamiento y los que no lo reciben. Este es el efecto causal estimado del tratamiento en el punto de corte.

- **Denominador:**

$$E[D|X = c] - E[D|X < c]$$

Mide la diferencia en la probabilidad de recibir el tratamiento en el punto de corte c frente a aquellos que están justo por debajo del mismo. Este término ajusta la estimación del efecto causal por el hecho de que no todos los individuos en el lado derecho del corte reciben el tratamiento.

Supuestos de Identificación del Método RD Fuzzy

Para que el estimador de regresión discontinua fuzzy (RD Fuzzy) sea válido y produzca estimaciones causales, es necesario que se cumplan ciertos supuestos de identificación.

1. **Continuidad de la variable de resultado:** La variable de resultado Y debe ser continua alrededor del punto de corte c. Esto implica que no deben haber saltos abruptos en la distribución de Y a menos que sean provocados por la asignación al tratamiento.
2. **Asignación al tratamiento basada en la running variable:** La probabilidad de recibir el tratamiento debe depender de la variable de ejecución (running variable) X en torno al punto de corte c. Es decir, los individuos justo por encima y justo por debajo del corte deben tener probabilidades diferentes de recibir el tratamiento, reflejando así una discontinuidad en la asignación.
3. **Independencia condicional:** Los resultados potenciales $Y(0)$ y $Y(1)$ deben ser independientes de la asignación al tratamiento, condicionado a la running variable X. Esto significa que, una vez controlado por X, la asignación al tratamiento no debe estar correlacionada con el resultado.
4. **No manipulación de la running variable:** No debe haber manipulación en la variable de ejecución X alrededor del punto de corte. Es decir, los individuos no deben poder influir en su posición en relación con el cutoff para obtener el tratamiento. Esto asegura que cualquier discontinuidad observada en la probabilidad de tratamiento se deba a la asignación natural y no a intentos de manipulación.
5. **Estabilidad de los efectos:** Los efectos del tratamiento deben ser estables alrededor del punto de corte. Esto significa que el efecto causal del tratamiento sobre la variable de resultado no debería variar abruptamente en torno al cutoff.

Implementación del Método RD Fuzzy

La implementación del método de regresión discontinua fuzzy (RD Fuzzy) se realiza en varias etapas. A continuación, se describen los pasos operacionales clave para llevar a cabo este método:

1. **Definir la variable de ejecución (running variable):** Identificar la variable que determina la asignación del tratamiento. Esta variable debe ser continua y tener un punto de corte c que se utilizará para distinguir entre los grupos de tratamiento y control.
2. **Definir el tratamiento:** Establecer la variable de tratamiento T, que puede ser una variable binaria que indique si un individuo fue tratado o no (por ejemplo, $T = 1$ si fue listado en el reporte y $T = 0$ en caso contrario).
3. **Calcular la probabilidad de tratamiento:** Estimar la probabilidad de recibir el tratamiento $P(T = 1|X)$ en función de la variable de ejecución X. Esto generalmente se realiza utilizando un modelo de regresión logística, donde se modela la relación entre T y X alrededor del punto de corte.
4. **Ajustar modelos de regresión:** Ajustar un modelo de regresión lineal para la variable de resultado Y considerando la variable de ejecución X y la variable de tratamiento T:

$$Y = \alpha + \beta_1 T + \beta_2 X + \beta_3 (X - c) + \epsilon$$

Donde β_1 captura el efecto del tratamiento, y β_3 representa el efecto de la running variable alrededor del cutoff.

5. **Estimación del efecto causal:** Utilizar el estimador de variables instrumentales (IV) para obtener el efecto causal del tratamiento en la variable de resultado. El IV es la probabilidad estimada de recibir el tratamiento basada en la running variable.

6. **Evaluar la robustez de los resultados:** Realizar análisis de sensibilidad y pruebas de robustez para evaluar la validez de las conclusiones. Esto puede incluir la verificación de supuestos como la continuidad de la variable de resultado y la no manipulación de la running variable.

10. Pregunta 10:

A) Dado los siguientes estimadores:

$$\text{ITT\_estimate} = 0.0485$$

$$\text{first\_stage\_estimate} = 0.3419$$

Calculamos el estimador RD Fuzzy como sigue:

$$RD \text{ Fuzzy} = \frac{\text{ITT\_estimate}}{\text{first\_stage\_estimate}} = \frac{0.0485}{0.3419} \approx 0.141$$

La interpretación del estimador RD Fuzzy es la siguiente:

El estimador RD Fuzzy es aproximadamente 0.141*, lo que indica que por cada incremento en la probabilidad de ser listado, se espera un aumento del* 14.1% *en la probabilidad de revinculación escolar.*

B) A continuación, se presentan los resultados de la segunda regresión, donde se evalúa el efecto de la inclusión en el reporte de deserción sobre la probabilidad de revinculación escolar, usando el metodo de variables instrumentales. La tabla incluye las estimaciones de los coeficientes, los errores estándar y los valores P asociados.

Table 8. Resultados de la Regresión de Graduación en 2022

VARIABLE	COEF.	ERR. EST.	$P > \|t\|$
LISTEDREPORT	0.1420	0.060	0.019
SCHOOLDAYS	-0.0020	0.001	0.027
INTERACTION	0.0018	0.001	0.148

El estimador del impacto es 0.142, lo que indica que por cada unidad de incremento en la probabilidad de ser listado en el reporte, se espera un aumento de 14.2% en la probabilidad de revinculación escolar. Este resultado es estadísticamente significativo, dado que el valor P asociado es 0.019, lo que indica que la relación observada es poco probable que sea debida al azar.