AULA ESTADÍSTICA

Serie Monográfica: introducción a la técnica de regresión discontinua difusa

Autor: Jorge Lorenzo

Año 2024

Regresión discontinua difusa (*fuzzy regression discontinuity design*, FRDD)

En esta monografía presentaremos una técnica de regresión que está siendo ampliamente utilizada en los modelos clasificatorios basados en aprendizaje automático. El objetivo es mostrar el modelo general y comentar sus aplicaciones en el análisis de datos educativos.

La técnica de regresión discontinua difusa (*fuzzy regression discontinuity design*, FRDD) es una metodología utilizada en econometría y estadística para estimar efectos causales de una intervención o tratamiento cuando los valores de corte establecidos por un umbral no son consistentes. En este sentido, la técnica se presenta como una variación de la regresión discontinua estándar (RDD), que se utiliza cuando la regla de asignación basada en un umbral es clara. Si bien la técnica es ampliamente utilizada en econometría, recientemente se han propuesto aplicaciones para datos educativos. Comenzaremos por algunas definiciones generales.

Una regresión discontinua consiste en evaluar la probabilidad de recibir un tratamiento, según un valor de umbral específico en una variable de asignación. Por lo general, esta variable de asignación se dicotomiza a partir de una variable continua. Por ejemplo, supongamos que deseamos asignar alumnos a un programa de promoción de las ciencias, para lo cual se toma como criterio las puntuaciones en una prueba cuyo umbral para asignar los individuos es de 70 puntos. Tal criterio determina la condición de asignación de manera clara: si un estudiante obtiene 70 puntos o más entra en el programa, caso contrario no ingresa. Una vez admitido un estudiante, recibe una beca para completar el programa. En estas circunstancias, es factible aplicar un modelo de regresión discontinua estándar. Pero, podría ser el caso que exista otro criterio complementario, como podría ser la valoración de un comité de profesores, sobre las aptitudes de los alumnos aspirantes. Al aplicar ambos criterios, puede ocurrir que se genere una frontera difusa en torno a la asignación al programa de los aspirantes.

La regresión discontinua difusa se aplica cuando el cumplimiento de la regla en el umbral no es perfecto. En otras palabras, no todos los individuos que superan el umbral reciben el tratamiento y algunos que no lo superan, sí lo reciben. Esto crea una situación en la que la probabilidad de recibir el tratamiento cambia en el umbral, pero no de manera perfecta. Entonces, en el programa de becas de promoción de las ciencias que se otorga a estudiantes, la puntuación de 70 en el examen, es un umbral perfecto para una regresión discreta estándar: cualquier estudiante con una puntuación de 70 o más recibiría la beca, y aquellos con menos de 70 no la recibirían. Sin embargo, el criterio de la valoración del comité de expertos puede producir decisiones discrecionales, y algunos estudiantes con puntuaciones inferiores a 70 ingresan en el programa, y otros con puntuaciones iguales o mayores a 70 no lo hacen. La aplicación de una regresión discontinua estándar, tendrá dificultades para llevar a un parámetro preciso las decisiones que se tomen al emplear el criterio de valoración del comité de profesores. No obstante, este problema se puede abordar mediante la aplicación de la regresión discontinua difusa. La cuestión que surge de lo expuesto sería: ¿por qué agregar un criterio más a la valoración obtenida por el resultado del examen? Pues bien, puede

ocurrir (y de hecho muchas veces ocurre), que el criterio del examen es incompleto, y puede producir un sesgo de selección, tal el caso de alumnos que por alguna circunstancia particular no alcanzan la puntuación umbral necesaria para entrar al programa y recibir la beca, aunque poseen aptitudes. Por otro lado, podría considerarse que algunos estudiantes que rinden adecuadamente, en el largo plazo no obtendrían ventajas del programa. Es por ello que se necesitarían de ambos criterios para la toma de decisión adecuada sobre cuáles individuos formarán parte del grupo que ingrese al programa de promoción de la ciencia mediante la beca. A partir de estas consideraciones veremos cómo la aplicación de la técnica de regresión discontinua difusa, nos ayuda a tomar la mejor decisión posible.

Para aplicar la regresión discontinua difusa, se sigue un enfoque de dos etapas:

Primera etapa (Intensidad del tratamiento): se estima la probabilidad de recibir el tratamiento (en este caso, la beca) en función de la variable de asignación (puntuación en el examen) y se analiza la discontinuidad en esta probabilidad en el umbral.

$$Recibir\ la\ beca = \ \alpha + \beta * (puntuación \geq 70) + \varepsilon$$

Segunda etapa (Efecto del tratamiento): Se utiliza la probabilidad estimada de recibir el tratamiento para estimar el efecto del tratamiento en el resultado de interés (por ejemplo, rendimiento académico futuro).

$$Rendimento\ academico = \ \gamma + \delta * (p\ de\ recibir\ la\ beca\ \{estimada\}) + \nu$$

La interpretación de la regresión discontinua difusa es que, aunque no todos los estudiantes se adhieren perfectamente a la regla del umbral, la discontinuidad en la probabilidad de recibir el tratamiento en el umbral puede usarse para identificar el efecto causal del tratamiento. Si existe una discontinuidad en el rendimiento académico en el umbral, esta discontinuidad puede ser atribuida al efecto de la beca, siempre y cuando se cumplan ciertos supuestos, que son:

Continuidad: Las otras variables explicativas deben ser continuas en el umbral.

Manipulación del umbral: No debe haber manipulación de la puntuación en el examen cerca del umbral.

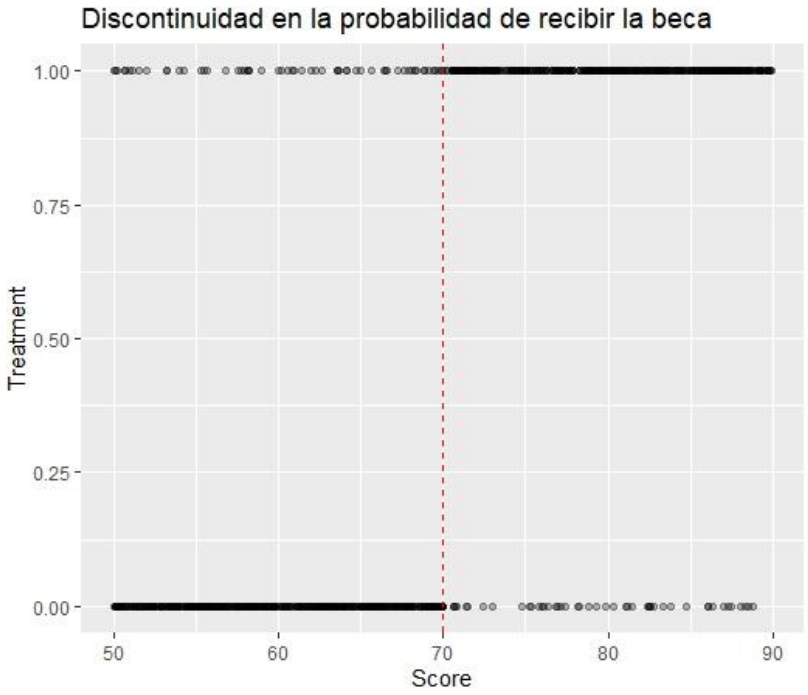
Validez del umbral: El umbral debe ser relevante para la asignación del tratamiento.

La regresión discontinua difusa es una técnica de análisis causal que permite a los investigadores lidiar con situaciones en las que el cumplimiento con la regla de asignación no es perfecto. Al estudiar la discontinuidad en la probabilidad de recibir el tratamiento en el umbral, podemos identificar efectos causales incluso en contextos de cumplimiento imperfecto.

Ejemplo a partir de R

En este apartado, se reproducirá el ejemplo usando una simulación mediante el software RStudio. Al final de la monografía se ofrece el código para reproducir el ejemplo. A continuación, se muestran las distintas etapas para analizar el ejemplo propuesto.

Visualización de la discontinuidad: Observamos una discontinuidad en la probabilidad de recibir el tratamiento alrededor del umbral de 70 puntos, pero no es perfecta debido a la variabilidad introducida por el criterio adicional. Específicamente, vemos que los puntos en el umbral se hallan dispersos tanto entre quienes no reciben el tratamiento, como entre quienes lo reciben. Si el umbral fuera perfecto, no debería observarse ningún caso en el valor 0 por encima del umbral 70, asimismo, no debería existir ningún caso por debajo de 70 en el valor 1. El hecho que aparezcan casos dispersos por encima y por debajo del umbral, se debe a la aplicación conjunta de los dos criterios propuestos para el otorgamiento de la beca.



Primera etapa: Estimamos la probabilidad de recibir el tratamiento en función del score y la discontinuidad en el umbral.

Residuales

Min	1Q	Median	3Q	Max
-0.91796	-0.11041	-0.09174	0.09840	0.90935

Coeficientes

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.002904	0.099356	0.029	0.977
score	0.001755	0.001641	1.069	0.285
I(score>= thres.)	0.759200	0.038305	19.820	<2e-16 ***

Los residuales representan las diferencias entre los valores observados y los valores predichos por el modelo. Son una medida de ajuste del modelo a los datos. Aquí, los valores de columna representan:

Min: El residuo mínimo observado.

1Q: El primer cuartil de los residuos, lo que significa que el 25% de los residuos son menores o iguales a este valor.

Median: El valor mediano de los residuos, indicando que el 50% de los residuos son menores o iguales a este valor.

3Q: El tercer cuartil de los residuos, indicando que el 75% de los residuos son menores o iguales a este valor.

Max: El residuo máximo observado.

Un resumen de los residuos da una idea de la distribución de los errores del modelo. Idealmente, se espera que los residuos estén distribuidos de manera simétrica alrededor de cero. En este ejemplo los residuos varían entre -0.91796 y 0.90935. La mediana de los residuos es -0.09174, indicando que la mayoría de los residuos están próximos a cero, lo cual muestra que el modelo tiene un buen ajuste.

La tabla coeficientes proporciona las estimaciones de los coeficientes del modelo, sus errores estándar, valores t, y valores p.

Estimate: La estimación del coeficiente para cada variable. Indica el cambio esperado en la variable dependiente por una unidad de cambio en la variable independiente, manteniendo otras variables constantes.

Intercept: El valor esperado de tratamiento cuando todas las variables independientes son cero.

score: La pendiente de la variable score, indicando el cambio en tratamiento por cada punto adicional en score.

I(score >= threshold): El cambio en tratamiento cuando score es mayor o igual a 70 (indicador binario).

Std. Error: El error estándar de la estimación del coeficiente, que mide la precisión de la estimación.

t value: El valor t para la prueba de hipótesis de que el coeficiente es igual a cero. Se calcula como Estimate / Std. Error.

Pr(>|t|): El valor p asociado con el valor t, que indica la probabilidad de observar un valor tan extremo como el valor t bajo la hipótesis nula de que el coeficiente es cero. Un valor p pequeño (usualmente < 0.05) indica que el coeficiente es significativamente diferente de cero.

La estimación del intercepto es principalmente un ajuste del modelo. El score de 0.0017 indica que, por cada punto adicional en score, la probabilidad de recibir la beca aumenta en 0.0017, manteniendo constante la variable indicadora de umbral. El valor p 0.285, sugiere que esta relación no es estadísticamente significativa al nivel del 5%.

I(score >= thresh), indica que el coeficiente es 0.7592, significa que cuando el score es mayor o igual a 70, la probabilidad de recibir la beca aumenta en 0.7592. El valor p (<2e-16) indica que esta relación es altamente significativa.

Segunda etapa: Utilizamos la probabilidad estimada de recibir el tratamiento para estimar el efecto del tratamiento en el resultado. La técnica de FRD permite aprovechar esta discontinuidad imperfecta para identificar el efecto causal del tratamiento. En este ejemplo, el efecto estimado del tratamiento (beca) en el resultado (rendimiento académico) se muestra al final del análisis.

Residuos:

Min	1Q	Median	3Q	Max
-10.1433	-2.6429	-0.0108	2.5354	11.1120

Coeficientes:

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	78.4583	0.1677	467.89	<2e-16 ***
predicted_treatment	17.9655	0.2679	67.06	<2e-16 ***

La tabla de resultados de la segunda etapa del modelo de regresión, indica la probabilidad estimada de recibir el tratamiento (predicted_treatment) para estimar su efecto sobre la variable de resultado. La tabla de residuos, representan las diferencias entre los valores observados y los valores predichos por el modelo y son una medida de ajuste del modelo a los datos, aquí los valores ofrecidos son los mismos que en la primera etapa. Dado que aquí se analiza la distribución de los errores del modelo, idealmente se espera que los residuos estén distribuidos de manera simétrica alrededor de cero, tal como aparecen los valores tabulados.

La tabla de coeficientes, es igual a la presentada en la primera etapa, excepto que en este caso las filas contienen:

Intercept: El valor esperado de resultado cuando todas las variables independientes son cero.

predicted_treatment: La pendiente de la variable predicted_treatment, indica el cambio en el resultado por cada unidad adicional en tratamiento.

Interpretación de los resultados: la estimación del intercepto es 78.45, lo que significa que, si predicted_treatment es 0, el valor esperado de resultado es 78.45, valor puede interpretarse como el rendimiento académico promedio de los estudiantes que no reciben la beca. En predicted_treatment, el coeficiente es 17.96, que indica que, por cada unidad adicional en la probabilidad estimada de recibir el tratamiento (beca), el rendimiento académico aumenta en 17.96 puntos. El valor p sugiere que esta relación es altamente significativa.

En concreto coeficiente positivo y significativo de predicted_treatment indica que recibir la beca tiene un efecto considerable en el rendimiento académico posterior de los estudiantes en el área de ciencias. Esta interpretación es crucial para la técnica de regresión discontinua difusa, ya que permite identificar el efecto causal del tratamiento a pesar de la discontinuidad imperfecta en la asignación del tratamiento.

Vamos a interpretar otros valores adicionales obtenidos del modelo de regresión discontinua difusa (FRD) que hemos utilizado para estimar el efecto del tratamiento sobre el rendimiento académico:

Error Estándar Residual (Residual Standard Error - RSE)

Residual standard error: 3.362 con 998 grados de libertad

El error estándar residual (RSE) es una medida de la dispersión de los residuos del modelo. Es esencialmente una estimación de la desviación estándar de los mismos. El valor 3.362 indica que, en promedio, los residuos del modelo (las diferencias entre los valores observados y los valores predichos) tienen una desviación estándar aproximada de 3. Los grados de libertad para el RSE se calcularon sobre una muestra de 1000 casos, que al restarse los parámetros estimados quedan en 998 (en este ejemplo los parámetros son el intercepto y predicted_treatment).

R^2 múltiple

El valor de R^2 múltiple es de 0.8184, representa una medida de la proporción de la varianza en la variable dependiente que es explicada por el modelo. Para este ejemplo, R^2= 0.8184 indica que el 81.84% de la variabilidad en el rendimiento académico es explicada por la probabilidad estimada de recibir el tratamiento (beca para ingresar al programa). R^2 ajustado, es una medida que ajusta el valor según el número de predictores en el modelo. Es particularmente útil cuando se compara el ajuste entre modelos con diferentes números de predictores. Para este ejemplo, no se encuentran diferencias apreciables entre estas medidas ya que R^2 ajustado= 0.8182.

El contraste del modelo se realiza con ayuda de la distribución F, para lo cual se informa que el valor de F= 4497; 1; DF 998, p-value: < 2.2e-16. Este estadístico evalúa la hipótesis nula de que todos los coeficientes del modelo son iguales a cero (es decir, que el modelo no tiene ningún poder explicativo). Los valores obtenidos indican que al menos uno de los coeficientes del modelo es significativamente diferente de cero.

Conclusión:

RSE (3.362): Indica que el modelo tiene una precisión moderadamente alta, con residuos que varían en promedio alrededor de 3.362 unidades de los valores observados. R^2 múltiple de 0.8184 indica que el modelo explica aproximadamente el 81.84% de la variabilidad en el rendimiento académico, lo que sugiere un fuerte poder explicativo del modelo. El contraste F muestra que el modelo es altamente significativo y que la probabilidad estimada de recibir el tratamiento tiene un efecto considerable y significativo en el rendimiento académico.

Estos resultados refuerzan la conclusión de que la técnica de regresión discontinua difusa (FRD) ha sido efectiva en identificar y estimar el efecto causal del tratamiento (beca) sobre el rendimiento académico en este ejemplo. Concretamente, si se emplean dos criterios conjuntos (una prueba y el juicio de un comité de expertos), para asignar a un grupo de estudiantes a un programa de promoción de las ciencias mediante el otorgamiento de una beca, se observa que en los casos que cumplen parcialmente los criterios de inclusión, mejoran su rendimiento en esta área. Por lo cual, este modelo de asignación compuesto es mejor que la utilización de un solo criterio.

Hacerlo en RStudio

```
install.packages("dplyr")
install.packages("ggplot2")
install.packages("ivreg")
library(dplyr)
library(ggplot2)
library(ivreg)
set.seed(42)
# Simular datos
n <- 1000
score <- runif(n, 50, 90)  # Puntuaciones entre 50 y 90
noise <- rnorm(n)  # Ruido
# Umbral para la beca
threshold <- 70
# Variable de tratamiento difuso
# Algunos estudiantes con score >= 70 no reciben la beca y algunos con score < 70 sí la reciben
treatment <- as.numeric(score >= threshold) + rbinom(n, 1, 0.1) - rbinom(n, 1, 0.1)
treatment <- as.numeric(treatment > 0)
# Variable de resultado
outcome <- 50 + 5 * treatment + 0.5 * score + noise
# Crear DataFrame
data <- data.frame(score = score, treatment = treatment, outcome = outcome)
ggplot(data, aes(x = score, y = treatment)) +
geom_point(alpha = 0.3) +
geom_vline(xintercept = threshold, color = 'red', linetype = 'dashed') +
labs(x = 'Score', y = 'Treatment', title = 'Discontinuidad en la probabilidad de recibir la beca')
# Primera etapa
first_stage <- lm(treatment ~ score + I(score >= threshold), data = data)
data$predicted_treatment <- predict(first_stage, data)
summary(first_stage)
# Segunda etapa
second_stage <- lm(outcome ~ predicted_treatment, data = data)
summary(second_stage)
# Efecto estimado del tratamiento
effect_estimate <- coef(second_stage)['predicted_treatment']
print(paste("Efecto estimado del tratamiento:", round(effect_estimate, 2)))
install.packages("dplyr")
install.packages("ggplot2")
install.packages("ivreg")
library(dplyr)
library(ggplot2)
library(ivreg)
set.seed(42)
# Simular datos
n <- 1000
score <- runif(n, 50, 90)  # Puntuaciones entre 50 y 90
noise <- rnorm(n)  # Ruido
# Umbral para la beca
threshold <- 70
# Variable de tratamiento difuso
# Algunos estudiantes con score >= 70 no reciben la beca y algunos con score < 70 sí la reciben
treatment <- as.numeric(score >= threshold) + rbinom(n, 1, 0.1) - rbinom(n, 1, 0.1)
treatment <- as.numeric(treatment > 0)
# Variable de resultado
outcome <- 50 + 5 * treatment + 0.5 * score + noise
# Crear DataFrame
data <- data.frame(score = score, treatment = treatment, outcome = outcome)
ggplot(data, aes(x = score, y = treatment)) +
geom_point(alpha = 0.3) +
geom_vline(xintercept = threshold, color = 'red', linetype = 'dashed') +
labs(x = 'Score', y = 'Treatment', title = 'Discontinuidad en la probabilidad de recibir la beca')
# Primera etapa
first_stage <- lm(treatment ~ score + I(score >= threshold), data = data)
data$predicted_treatment <- predict(first_stage, data)
summary(first_stage)
# Segunda etapa
second_stage <- lm(outcome ~ predicted_treatment, data = data)
summary(second_stage)
# Efecto estimado del tratamiento
effect_estimate <- coef(second_stage)['predicted_treatment']
print(paste("Efecto estimado del tratamiento:", round(effect_estimate, 2)))
```