Departamento de Ingeniería Civil Industrial
Facultad de Ciencias Físicas y Matemáticas
Universidad de Chile
IN4143-1 Análisis de Datos e Inferencia Causal

Laboratorio nº3

Integrantes: Adolfo Rojas
Profesor: Daniel Schwartz P.
Omar D. Perez
Auxiliar: Antonia Aceituno
Camila Galarce
Ayudantes: Bastián Medina
Guillermo Escobar Fuentes

Fecha de entrega: 22 de octubre de 2024
Santiago de Chile

1) *Tabla de regresiones*

	OLS (sin IV)	ITT	LATE	LATE 2SLS manual (sin control)	LATE 2SLS (sin control)	LATE 2SLS (con control)
Coeficiente	0.83602	−0.09691	0.5737852	−0.3268	−0.3268	−0.456373
Error Estándar	0.06682	0.07653	–	0.2581	0.2859	0.303232

Tabla 1: Tabla comparativa

2) A partir de los resultados obtenidos en el ítem anterior responda a las siguientes preguntas

   a) *¿Se observa evidencia de endogeneidad en la estimación de i.*
   Sí, se puede visualizar un sesgo que termina sobreestimando el efecto de X, basta con ver la diferencia entre el coeficiente estimado en i. y los coeficientes estimados de LATE a 2SLS (sin var. de controles)

   b) *En relación a los errores estándar asociados a la variable endógena de las regresiones iv., v. y vi., ¿son diferentes? ¿por qué?*
   Sí, entre las dos primeras que no tienen en cuenta variables de control, estas son distintas solo por la forma en que está implementada la doble regresión en el paquete vs hacer dos regresiones manuales, ahora la diferencia respecto a vi y el resto es que esta al incluir variables de control se esperaría una mejor aproximación del efecto causal y menor sigma (solo si las variables de control cumplen con el supuesto de independencia)

   c) *Compare los coeficientes asociados a la variable endógena de i., iii., iv., v. y vi. ¿Son diferentes? ¿por qué?*
   Sí, son diferentes por lo discutido en el ítem 2.a, respecto a iii esta solo da el efecto de los compliers, sobre iv y v estos deben ser los mismos puesto que se obtienen de la misma forma, por otro lado el LATE 2SLS con Q_1, Q_2, Q_3 al considerar las mismas se obtiene un coeficiente que refleja mejor el efecto causal de X y por eso la diferencia

3) *Compruebe si se cumplen los supuestos requeridos para que la estimación LATE sea válida (mediante IV)*

   Usando el ivreg para la regresión vi tenemos un p-value del test de Weak instrument extremadamente pequeño $= 3.95e-11$ lo que nos dice que en efecto el instrumento es relevante, ahora sobre el supuesto de independencia, no es verificable al 100 % pero haciendo una especie de tabla de balance donde si las covariables $Q_i, i \in \{1, 2, 3\}$ son en promedio iguales para ambos grupos podríamos decir que se tiene independencia. Finalmente para el supuesto de *exclusion restriction* es razonable asumir que las reparaciones en las calles aumentaron los tiempos de viaje, lo que provocó que muchos trabajadores llegaran tarde a sus trabajos y que la llegada tarde de un trabajador puede reducir su productividad en ese día debido a menos horas trabajadas. Ahora para argumentar la ausencia de un efecto directo de Z y X solo falta ponerse en el caso y ser realistas pues no hay razón alguna para pensar que el hecho de que haya reparaciones en el camino tenga un impacto directo en la productividad

4) *¿Cuál cree usted que es el histograma asociado a las estimaciones de 2SLS?*

El histograma de verde (el de más a la izquierda) se trata del OLS puesto que esta es más eficiente (menor desviación estándar) pero sesgada (para este caso en que hay endogeneidad) en comparación al LATE por 2SLS que es más dispersa pero sin sesgo

1. Anexo

Código 1: Código

```r
library("MASS")
library("dplyr")
library("ivreg")
library("readxl")

df <- read_excel("datos/df_lab3_2024.xlsx")

# Compliers vs No compliers
table(df[, c('Z', 'X')])

pi1 = sum(df$Z==1 & df$X==1)/sum(df$Y[df$Z==1]) # 235/(235+0): compliers del grupo de tratamiento
pi0 = sum(df$Z==0 & df$X==1)/sum(df$Y[df$Z==0]) # 91/(91+174): no compliers del grupo de control (z=0, x=1, o sea que fueron tratados pero no asignados) / grupo de control

# Estimaciones OLS, ITT y LATE
OLS = summary(lm(Y ~ X, data=df))
OLS
print(OLS$coefficients[2,1]) # recibir la estimación asociada a la variable endógena (fila 2 columna 1)

# LATE = ITT/ITTd
ITT = summary(lm(Y ~ Z, data=df))
print(ITT)
ITT = mean(df$Y[df$Z==1]) - mean(df$Y[df$Z==0])
ITTd = pi1 - pi0
LATE = ITT/ITTd        # Solo funciona para estimaciones sin variables de controles
print(LATE)

# LATE en 2 etapas
first_stage = lm(X ~ Z, data=df)
df$X_hat = predict(first_stage, newdata=df)

second_stage = summary(lm(Y ~ X_hat, data=df))
print(second_stage)

# LATE 2SLS sin controles
iv_reg1 = summary(ivreg(Y ~ X | Z, data=df))   # Aquí te da el test de relevancia (weak instrument) e independencia (Hausman)
print(iv_reg1)

```

```r
# LATE 2SLS con controles
iv_reg2 = summary(ivreg(Y ~ X + Q1 + Q2 + Q3 | Z + Q1 + Q2 + Q3, data=df))
print(iv_reg2)

### Chequear supuestos

# El instrumento es relevante (!= 0)
summary(lm(X ~ Z, data=df))
summary(lm(X ~ Z, data=df))coefficients[2,1]

# Supuesto de independencia (si se aplica asignación aleatoria), esto se puede seudo verificar con el test de Hausman o con un balance donde si las covariables son en promedio iguales para ambos grupos se tiene independencia
df %>% group_by(Z) %>% summarize("promedio Q" = mean(Q1))
df %>% group_by(Z) %>% summarize("promedio Q" = mean(Q2))
df %>% group_by(Z) %>% summarize("promedio Q" = mean(Q3))

# Restricciones de exclusión, argumentar que el instrumento no afecta directamente a la variable dependiente

###

# Cuando hay endogeneidad LATE a dos etapas es más ineficiente que OLS pero es insesgado
```