INSTITUTO DE SOCIOLOGÍA
FACULTAD DE CIENCIAS SOCIALES

Tarea N°1

Inferencia Causal - SOL3063

Estudiante Andreas Laffert
Profesor Luis Maldonado
Ayudante Gustavo Ahumada

viernes 31, mayo 2024

Pregunta 1

Para estimar el efecto de la distancia en la intolerancia, Pepinsky et al. (2023) sostienen que es necesario controlar por efectos fijos para Estados. Al respecto y en base a su lectura del texto:

a) Señale qué tipo de características son controladas por estos efectos fijos.

Pepinsky et al. (2023) sostienen que los resultados obtenidos por Homola et al. (2020) son producto de una heterogeneidad espacial no observada, asociada a diferencias entre los Estados-Länder alemanes. Al controlar por estas características no observadas propias de los Estados, el efecto de la distancia al campo de concentración en la intolerancia desaparece. Los autores afirman que las variaciones en actitudes políticas entre los Estados en Alemania son bien conocidas en la investigación sobre ciencia política del país. Sin embargo, Homola et al. (2020) no consideran las posibles diferencias en actitudes entre Estados para analizar la relación entre la distancia a los campos y la intolerancia.

Estas diferencias entre Estados se refieren a características económicas y a la historia política e institucional, que pueden influir en la educación cívica y el currículum escolar, y, en consecuencia, en la opinión pública de los individuos de esos Estados (Pepinsky et al., 2023, p. 2). Así, estos factores tanto observados como no observados pueden afectar las actitudes políticas. Además, los autores identifican factores como la cultura política, las organizaciones de la sociedad civil y las tradiciones político-religiosas como covariables que utilizan en sus estimaciones. Si estos factores están correlacionados con la distancia al campo o si el mecanismo de proximidad al campo en la intolerancia varía entre Estados, no considerarlos produciría estimaciones sesgadas.

Para abordar este desafío, los autores proponen emplear efectos fijos para los Estados, ajustando por cualquier factor constante en el tiempo que varía entre Estados y que podría influir en la intolerancia. Siguiendo a Wooldridge (2009, p. 456), el estimador de efectos fijos incluye un término que captura las características observables e inobservables constantes en el tiempo dentro de las unidades, pero que varían entre ellas y que pueden actuar como factores de confusión en la relación entre las variables independientes y la variable dependiente. Este estimador aísla lo constante para centrarse en los factores que cambian en el tiempo (varianza *within*), eliminando así la heterogeneidad no observada entre unidades.

b) Señale y explique las condiciones bajos las cuales estos efectos fijos serían buenos controles.

De acuerdo con Pepinsky et al. (2023), los efectos fijos para Estados serían buenos controles si se cumple que este factor de confusión no sea descendiente causal del tratamiento, en este caso, de la distancia al campo de concentración. En detalle, para que los EF sean buenos controles para aislar la heterogeneidad no observada de características de los Estados es necesario que estas no se encuentren en el camino causal que conecta a la variable causal de interés con la variable de resultado (intolerancia). En el texto esto aparece bajo el supuesto 1, que establece que el confundidor (F) no es descendiente del tratamiento (T). En otras palabras, el confun-

didor no es afectado por el tratamiento directa o indirectamente. Si esta condición se cumple, controlar el efecto fijo no generaría sesgo de post-tratamiento. Gráficamente, esto implica que el DAG que representa la estimación se asemeja al DAG 1(a) de la Figura 1. Cumplir con esta condición ayuda a evitar el sesgo M-bias y asegura que la inclusión de efectos fijos en el análisis no distorsione los resultados (por ejemplo, DAG 2(b) de la Figura 2), permitiendo una estimación precisa del efecto causal entre el tratamiento y el resultado.

Pregunta 2

Considere el gráfico 2(b) de la figura 2 en Pepinsky et al. (2023). En este caso, explique por qué controlar por F genera sesgo:

De acuerdo con el supuesto 1, para que controlar por el confounder F no induzca sesgo de postratamiento, es esencial que esta variable confounder F no sea descendiente del tratamiento T. En otras palabras, este supuesto establece que el tratamiento no debe afectar al confounder para que controlar por él no sea problemático.

Sin embargo, incluso cuando se cumple el supuesto 1, controlar por F puede generar un sesgo de colisión. Como mencionan Pepinsky et al. (2023), "El supuesto 1 excluye la forma más simple de sesgo del colisionador en la que F es descendiente tanto de T como de Y" (p.3). Por lo tanto, los autores proponen un segundo supuesto para evitar el sesgo de colisión, el cual establece que F no debe ser descendiente ni de una variable U_1 de la cual T también sea descendiente, ni de una variable U_2 de la cual Y también sea descendiente.

En el escenario donde F es descendiente de U_1, del cual T también es descendiente, y F es descendiente de U_2, del cual Y también es descendiente, controlar por F introduciría un sesgo al condicionar por una variable colisionadora. Al hacerlo, se crea un camino no causal entre T y Y a través de U_1, generando una asociación espuria entre T y Y que no existiría sin controlar por F.

En resumen, en situaciones donde U_1 causa directamente tanto el tratamiento T como el confounder F, y U_2 causa directamente la variable de resultado Y y el confounder F, controlar por el efecto fijo introduciría sesgo M-bias al distorsionar la relación entre T y Y. Esto resalta la importancia de identificar y evitar el sesgo al controlar por variables que pueden actuar como colisionadores en el análisis causal.

Pregunta 3

Estime los modelos 1, 2, 3 y 4 de Table 1 en Pepinsky et al. (2023). Específicamente:

a) Reporte sus resultados en una tabla de calidad similar a la Table 1 del artículo bajo replicación. Use las covariables mencionadas arriba.[1]

[1]Note que no debe incluir como variable independiente la población en 1925 y que debe utilizar el estimador least-squares-dummy-variables-estimator (LSDV) para su modelo con efectos fijos.

Table 1: Replicación de modelos Pepinsky et al. (2023)

	Modelo 1	Modelo 2	Modelo 3	Modelo 4
Distancia al campo	-0.009^{**}	0.004	-0.011^{**}	0.002
	(0.003)	(0.004)	(0.003)	(0.004)
% Judíos (1925)			−1.274	−6.553
			(1.025)	(3.631)
% Desempleo (1933)			0.698	1.047
			(0.441)	(0.603)
Participación partido nazi (1933)			-0.377^{*}	-0.758^{**}
			(0.179)	(0.227)
Länder FE	No	Si	No	Si
Método	OLS	OLS	OLS	OLS
Adj. R^2	0.005	0.031	0.008	0.039
Num. obs.	2075	2075	2075	2075

Nota: Las celdas contienen coeficientes de regresión con errores estándar entre paréntesis. $^{**}p < 0.01$; $^{*}p < 0.05$

b) Señale a lo menos 2 diferencias conceptuales entre los modelos 3 y 4.

En la Tabla 1 se presentan los modelos replicados de Pepinsky et al. (2023). Por un lado, el Modelo 3 utiliza un estimador OLS combinado (pooled OLS), que combina todos los datos sin considerar los efectos específicos de cada unidad y asume independencia entre las observaciones. Esto puede resultar en estimaciones menos eficientes y sesgadas debido a la heterogeneidad no observada entre los Estados, subestima los errores estándar y entrega pruebas estadísticas muy elevadas con valores p muy bajos (2009). Por otro lado, el Modelo 4 incorpora efectos fijos de los Estados, lo que permite controlar y capturar la variabilidad específica de cada Estado en la estimación del efecto. Al incluir los efectos fijos de los Estados, se mejora la eficiencia de las estimaciones al tener en cuenta las características únicas de cada Estado y al controlar por la heterogeneidad no observada que podría influir en la relación entre la distancia a los campos de concentración y la intolerancia. Con esto, en el Modelo 4 el coeficiente de la distancia al campo no solo deja de ser estadísticamente significativo, sino que también da vuelta el signo a positivo.

Además, el Modelo 4 con efectos fijos de los Estados también permite controlar de manera más efectiva los posibles factores de confusión no observados a nivel estatal que podrían influir en la relación entre la distancia a los campos de concentración y la intolerancia. Al incluir los efectos fijos, se tiene en cuenta la influencia de todas las características específicas de cada Estado que podrían estar correlacionadas con la variable independiente (distancia a los campos de concentración). Esto ayuda a mitigar el riesgo de sesgo causado por variables omitidas a nivel estatal que podrían distorsionar la estimación del efecto causal. En definitiva, la inclusión de efectos fijos por Estado en el Modelo 4 no solo mejora la eficiencia de las estimaciones, sino que también contribuye a controlar de mejor manera los posibles factores de confusión no observados a nivel estatal, fortaleciendo la validez interna de la relación causal investigada al explotar la varianza *within* (Wooldridge, 2009).

Pregunta 4

Pepinsky et al. (2023) sostienen que controlar por efectos fijos para Estados sería problemático si la variable causal de interés varía principalmente entre los Estados. Para evaluar empíricamente este tema, los autores estiman una serie de tests de Hausman y reportan los resultados en la Table A2 del Appendix D. Al respecto y usando intolerancia como variable dependiente, replique los modelos 1, 2 y 3 del Panel A de la Table A2. Reporte sus resultados en una tabla como la Table A2 e interprete los resultados de los tests de Hausman, indicando diferencias sustantivas/conceptuales entre los modelos que están siendo comparados.[2]

En la Tabla 2 se muestran los resultados de la replicación de test de Hausmann de Pepinsky et al. (2023). En el primer test, que contrasta el estimador efectos aleatorios (EA) con OLS combinado, se obtuvo un valor p < 0.05, siendo estadísticamente significativo. Esto nos permite rechazar con un 95% de confianza la hipótesis nula (H_0) de que no hay diferencias entre ambos estimadores, sugiriendo que EA es preferible debido a que, por lo general, es más eficiente que OLS combinado, según Wooldridge (2009). En el segundo test, que contrasta el estimador de EF con OLS combinado, también se obtuvo un valor p estadísticamente significativo, permitiendo sostener con un 95% de confianza que EF es preferible debido a su capacidad para controlar la heterogeneidad no observada que OLS combinado no puede capturar. El último test es el verdadero test de Hausman que compara el estimador de EF con el estimador de EA. La hipótesis nula (H_0) es que no hay diferencias sistemáticas entre los estimadores, implicando que los efectos específicos de las unidades no están correlacionados con las variables independientes. En este test, se obtuvo un valor p < 0.05, por lo que se rechaza la H_0 con un 95% de confianza. Esto sugiere que el supuesto clave del modelo EA, de que los factores no observados no se correlacionan con los predictores, es falso, haciendo del estimador EF un estimador más consistente.

Table 2: Replicación de test de Hausman Pepinsky et al. (2023)

	Modelo 1	Modelo 2	Modelo 3
Distancia al campo	-0.009^{**}	-0.000	0.004
	(0.003)	(0.004)	(0.004)
Método	Pooled	RE	FE
RE v Pooled		0.002	
FE v Pooled		0	
FE v RE		0.044	
Adj. R^2	0.005	-0.000	-0.007
Num. obs.	2075	2075	2075

Nota: Las celdas contienen coeficientes de regresión con errores estándar entre paréntesis. $^{**}p < 0.01$; $^{*}p < 0.05$

Existen diferencias sustantivas entre los modelos que se están comparando. Por un lado, el estimador combinado con OLS agrupa todas las observaciones sin considerar los efectos específicos de cada unidad y asume independencia entre las observaciones. Esto puede resultar en estimaciones menos eficientes debido a la heterogeneidad no observada entre las unidades,

[2]Note que no debe incluir covariables.

subestima los errores estándar y proporciona pruebas estadísticas muy elevadas con valores p muy bajos (Wooldridge, 2009). Además, el modelo OLS combinado no controla por la posible correlación entre las observaciones dentro de las mismas unidades, lo que puede llevar a conclusiones incorrectas sobre la significancia de los predictores.

Por otro lado, un modelo con estimador de efectos fijos (EF) incluye un término constante que captura factores constantes observados y no observados dentro de las unidades que pueden estar correlacionados con las covariables, permitiendo controlar la heterogeneidad no observada que puede actuar como un factor de confusión en la relación causal. En consecuencia, el estimador EF asume una correlación arbitraria entre α_i y los predictores X_{it}. Este enfoque es útil cuando se sospecha que hay características invariables en el tiempo dentro de las unidades que podrían sesgar los resultados si no se controlan adecuadamente.

Por último, el modelo de efectos aleatorios (EA) es similar al de EF en cuanto a sus supuestos, pero agrega uno adicional: que no existe correlación entre los factores no observados y las covariables. Sustantivamente, en el caso de Pepinsky et al. (2023), emplear un estimador EA implica asumir que las características no observadas de los Estados no se correlacionan con ninguna covariable en ningún periodo. Esto permite en los modelos EA incluir variables constantes en el tiempo y estimar sus efectos, además de permitir correlación serial. Sin embargo, si la correlación entre los efectos no observados y las covariables es significativa, las estimaciones del modelo EA serán sesgadas e inconsistentes.

En definitva, si se tienen buenas razones para considerar que no hay correlación entre el término α_i y los predictores X_{it} en todos los periodos, es posible emplear un EA; de lo contrario, es preferible utilizar un EF. Además, el uso de efectos fijos es generalmente más robusto cuando se quiere controlar por la heterogeneidad no observada que es constante en el tiempo pero varía entre unidades. En contraste, el modelo OLS combinado es el más sencillo pero también el más propenso a proporcionar estimaciones sesgadas y menos eficientes.

Referencias

Homola, J., Pereira, M. M., & Tavits, M. (2020). Legacies of the Third Reich: Concentration Camps and Out-group Intolerance. *American Political Science Review*, *114*(2), 573–590. https://doi.org/10.1017/S0003055419000832

Pepinsky, T. B., Goodman, S. W., & Ziller, C. (2023). Modeling Spatial Heterogeneity and Historical Persistence: Nazi Concentration Camps and Contemporary Intolerance. *American Political Science Review*, *118*(1), 519–528. https://doi.org/10.1017/S0003055423000072

Wooldridge, J. M. (2009). *Introductory econometrics: a modern approach* (4th ed). Mason, OH: South Western, Cengage Learning.

Código de R

```
knitr::opts_chunk$set(echo = F,
                      warning = F,
                      error = F,
                      message = F)
if (! require("pacman")) install.packages("pacman")

pacman::p_load(tidyverse,
               rio,
               here,
               vtable,
               kableExtra,
               sjPlot,
               sjmisc,
               estimatr,
               panelr,
               plm,
               clubSandwich,
               texreg)

options(scipen=999)
rm(list = ls())
db_or <- rio::import(file = here("input", "data", "EVS_main.csv")) %>%
  as_tibble()


miles <- function(x) {
  format(round(as.numeric(x),0), big.mark = ".")
}

decimales <- function(x) {
  format(round(as.numeric(x), 2), decimal.mark = ",")
}

custom_extract <- function(model) {
  tr <- extract(model)

  # Identificar índices a conservar (excluyendo "R$^2$", "s_idios" y "s_i
  gof_indices <- which(!(tr@gof.names %in% c("R$^2$", "s_idios", "s_id"))

  # Actualizar gof, gof.names y gof.decimal simultáneamente
```

```
  tr@gof.names <- tr@gof.names[gof_indices]
  tr@gof <- tr@gof[gof_indices]
  tr@gof.decimal <- tr@gof.decimal[gof_indices]

  return(tr)
}
# set theme

theme_nice <- function() {
  theme_bw() +
    theme(text = element_text(family = "serif"))


}

theme_set(theme_nice())

# Seleccionar

db <- db_or %>%
  dplyr::select(intolerance, Distance, prop_jewish25,
                unemployment33, nazishare33, state) %>%
  janitor::clean_names()

# Filtrar: No

# Recodificar
sjmisc::frq(db$state)
db$state <- as.factor(db$state)

# Tratamiento casos perdidos
colSums(is.na(db)) # no NA

# Transformar y derivar: No

# FE con LSDV
m1 <- lm(formula = intolerance ~ distance,
         data = db)

m2 <- lm(formula = intolerance ~ distance + state,
         data = db)
```

```r
m3 <- lm(formula = intolerance ~ distance + prop_jewish25 +
         unemployment33 + nazishare33, data = db)

m4 <- lm(formula = intolerance ~ distance + prop_jewish25 +
         unemployment33 + nazishare33 + state, data = db)

models <- list(m1, m2, m3, m4)

models1 <- map(models, custom_extract)

#lm_robust(formula = intolerance ~ distance,
#          data = db,
#          clusters = state,
#         se_type = "CR2",
#          fixed_effects = ~ state)

#coef_test(m2, vcov = "CR2", cluster = db$state)

#plm(formula = intolerance ~ distance,
#    data = db,
#    model = "within",
#    index = "state")

ccoef <- list(
  distance = "Distancia al campo",
  prop_jewish25 = "% Judíos (1925)",
  unemployment33 = "% Desempleo (1933)",
  nazishare33 = "Participación partido nazi (1933)"
  )

texreg::texreg(l = models1,
               caption = paste("(\\#tab:table1)","Replicación de modelos
               stars = c(0.05, 0.01),
               custom.note = "Nota: Las celdas contienen coeficientes de
               custom.model.names = c("Modelo 1", "Modelo 2", "Modelo 3",
               custom.coef.map = ccoef,
               leading.zero = T,
               float.pos = "h!",
               use.packages = F,
               booktabs = TRUE,
               scalebox = 0.90,
```

```
                digits = 3,
                custom.gof.rows = list("Länder FE" = c("No", "Si", "No","S
                                       "Método" = rep("OLS", 4)))



m_pooled <- plm(formula = intolerance ~ distance, data = db,
                model = "pooling", index = "state")

m_re <- plm(formula = intolerance ~ distance, data = db,
            model = "random", index = "state")

m_fe <- plm(formula = intolerance ~ distance, data = db,
            model = "within", index = "state")

hausman_re_pooled <- plm::phtest(m_re, m_pooled)

hausman_fe_pooled <- plm::phtest(m_fe, m_pooled)

hausman_fe_re <- plm::phtest(m_fe, m_re)

models <- list(m_pooled, m_re, m_fe)

models2 <- map(models, custom_extract)

texreg::texreg(l = models2,
          stars = c(0.05, 0.01),
          leading.zero = T,
          float.pos = "h!",
          use.packages = F,
          booktabs = TRUE,
          scalebox = 0.90,
          digits = 3,
          caption = paste("(\\#tab:table2)","Replicación de test de Hausm
          custom.note = "Nota: Las celdas contienen coeficientes de regre
          custom.model.names = c("Modelo 1", "Modelo 2", "Modelo 3"),
          custom.coef.map = list(distance = "Distancia al campo"),
          custom.gof.rows = list("Método" = c("Pooled", "RE", "FE"),
                                 "RE v Pooled" = c("",  round(hausman_re_
                                 "FE v Pooled" = c("",  round(hausman_fe_
                                 "FE v RE" = c("",  round(hausman_fe_re$p
```

)