

Datos en panel en R

Econometría Aplicada y Ciencia de Datos - CIDE Fall 2026

Carlos Brito

Trabajaremos con los datos *comportamiento_wide.csv*, que contienen información individual de niñas y niños, incluyendo su género, edad, raza e información de sus madres. Además, se incluye una medida auto reportada de autoestima (**self**) y una evaluación de comportamiento antisocial (**anti**). Se quiere conocer cómo influye la autoestima en el comportamiento antisocial. Para cada niño o niña hay tres observaciones en el tiempo. Se busca explicar el comportamiento antisocial en función de la autoestima y la condición de pobreza (**pov**):

$$anti_{it} = \alpha_i + \beta_1 self_{it} + \beta_2 pov_{it} + \varepsilon_{it}$$

Datos en panel

La base se encuentra en formato *wide*. Ponemos la base en formato *long*, donde haya una columna para cada variable y donde las filas representen a un individuo en un periodo.

```
data.comp <- read_csv("../files/comportamiento_wide.csv",  
                      locale = locale(encoding = "latin1"))  
colnames(data.comp)
```

La base transformada:

```
data.comp <- data.comp %>%  
  pivot_longer(c(anti90:anti94,self90:self94,pov90:pov94),  
              names_to = c("measure", "year"),  
              names_pattern = "(.*)(..)" %>%  
  pivot_wider(names_from = measure,  
              values_from = value)  
  
colnames(data.comp)
```

Estimadores para modelos de datos en panel

Estimador de MCO

Estimamos la ecuación de comportamiento antisocial empleando MCO *pooled*.

```
m.mco <- plm( anti ~ self + pov,  
              data=data.comp,  
              model="pooling",  
              index = c("id", "year"))
```

$$\hat{\beta}_{ols} = (X'X)^{-1} (X'y)$$

$$V(\hat{\beta}_{ols}) = (X'X)^{-1} X' \text{Var}(\epsilon|X) X (X'X)^{-1}$$

homoscedasticity

$$\text{Var}(\hat{\beta}_{ols}) = \sigma^2 (X'X)^{-1}$$

$$\widehat{\text{Var}}(\hat{\beta}_{ols}) = \hat{\sigma}^2 (X'X)^{-1}$$

$$\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{n-k}$$

↓
N x T

	MCO
self	-0.06510*** (0.01108)
pov	0.51581*** (0.07873)
Num.Obs.	1743

heteroscedasticity

$$\widehat{V}\hat{\beta}_{ols} = (X'X)^{-1} \left(\sum_i \sum_t x_{it} x_{it}' \hat{\epsilon}_{it}^2 \right) (X'X)^{-1}$$

La variable *self* tiene un efecto negativo y estadísticamente significativo sobre *anti*. El estimador de MCO será consistente solo si *self* y *pov* no están correlacionadas con el error, asumiendo que α_i puede escribirse simplemente como α .

Estimador de MCO y errores agrupados

El modelo pooled ignora la naturaleza en panel de los datos. Como tenemos a los mismos individuos en varios puntos del tiempo, los errores están agrupados.

```
vcov=list(NULL, clubSandwich::vcovCR(m.mco, type='CR1',  
  cluster=data.comp$id))
```

$$\hat{V}_{cl}(\hat{\beta}_{ols}) = (X'X)^{-1} \left(\sum_i \sum_t \sum_n x_{itk} x_{itn}' \hat{u}_{itk} \hat{u}_{itn} \right) (X'X)^{-1}$$

$$\hat{\beta}_{ols} = (X'X)^{-1} (X'y)$$

	MCO		MCO, errores agrupados
self	-0.06510*** (0.01108)	= ≠	-0.06510*** (0.01369)
pov	0.51581*** (0.07873)	= ≠	0.51581*** (0.10496)
Num.Obs.	1743		1743

Al tomar en cuenta la correlación entre individuos, los errores estándar son más grandes, pero los resultados siguen siendo significativos.

$$\hat{V}(\hat{\beta}_{ols}) \geq \hat{V}_{cluster}(\hat{\beta}_{ols}) \geq \hat{V}_{cl}(\hat{\beta}_{ols})$$

Estimador de efectos fijos

Estimamos la ecuación de comportamiento antisocial empleando el estimador *within*.

```
m.fe <- plm( anti ~ self + pov,  
            data=data.comp,  
            model="within",  
            index = c("id", "year"))
```

o adding α_i

$$anti_{it} = \alpha_i + \beta_1 self_{it} + \beta_2 pov_{it} + \varepsilon_{it}$$

	MCO	MCO, err. agrup.	Efectos fijos
self	-0.06510*** (0.01108)	-0.06510*** (0.01369)	-0.05150*** (0.01131)
pov	0.51581*** (0.07873)	0.51581*** (0.10496)	0.10490 (0.09922)
Num.Obs.	1743	1743	1743

α_i

ε_{it}

Si la heterogeneidad no observada y el error están potencialmente correlacionados, el estimador de efectos fijos elimina la heterogeneidad no observada y estima consistentemente los parámetros.

Estimador de efectos aleatorios

Estimamos la ecuación de comportamiento antisocial empleando efectos aleatorios.

```
m.re <- plm( anti ~ self + pov,  
             data=data.comp,  
             model="random",  
             index = c("id", "year"))
```

	MCO	MCO, err. agrup.	Ef. fijos	Ef. aleatorios
self	-0.06510*** (0.01108)	-0.06510*** (0.01369)	-0.05150*** (0.01131)	-0.05673*** (0.01022)
pov	0.51581*** (0.07873)	0.51581*** (0.10496)	0.10490 (0.09922)	0.29241*** (0.08196)
Num.Obs.	1743	1743	1743	1743

Si la α_i heterogeneidad no observada y el error ϵ son independientes, el estimador de efectos aleatorios es válido. MCO pooled también es consistente pero no eficiente.

Inclusión de características invariantes en el tiempo

Si deseamos incorporar el género (**gender**): no es posible estimarlo con efectos fijos, por lo que este modelo queda descartado. La alternativa es MCO pooled o un modelo de efectos aleatorios.

```
m.sex.f <- plm( anti ~ self + pov + gender,  
  data=data.comp, model="within",  
  index = c("id", "year"))
```

```
m.sex.a <- plm( anti ~ self + pov + gender,  
  data=data.comp, model="random",  
  index = c("id", "year"))
```

	Ef. fijos	Ef. aleatorios (género)
self	-0.05150*** (0.01131)	-0.05856*** (0.01022)
pov	0.10490 (0.09922)	0.30500*** (0.08146)
gender		-0.48047*** (0.10710)
Num.Obs.	1743	1743

Prueba de Hausman

Realizamos una prueba de Hausman para determinar si se prefiere un modelo de efectos fijos o uno de efectos aleatorios.

`phptest(m.fe, m.re)`

$$\begin{aligned} H_0: \hat{\beta}_w &\xrightarrow{P} \beta; \hat{\beta}_{RE} \xrightarrow{P} \beta && \text{qi ind. of } X_{it} \\ H_1: \hat{\beta}_w &\xrightarrow{P} \beta; \hat{\beta}_{RE} \not\xrightarrow{P} \beta && \text{qi is correlated with } X_{it} \end{aligned}$$

$\beta_1 \rightarrow$ subvector of coef of X_{it} not constant on time

Hausman Test

$$\sim \chi^2(\dim[\beta_1])$$

data: anti ~ self + pov

chisq = 13.578, df = 2, p-value = 0.001126

alternative hypothesis: one model is inconsistent

$\hookrightarrow$ Reject H_0

Se rechaza H_0 de que los coeficientes estimados son iguales. Hay evidencia de que se prefiere un modelo de efectos fijos, aunque no se podrá estimar el coeficiente de variables invariantes en el tiempo.

Equivalencia de estimadores

Equivalencia de estimadores

Comprobamos que el estimador de efectos fijos es equivalente a MCO con dummies de individuos.

```
m.fe <- plm( anti ~ self + pov,  
            data=data.comp, model="within",  
            index = c("id", "year"))  
  
m.dummy <- lm(anti ~ self + pov + factor(id),  
              data=data.comp)
```

	MCO	MCO con dummies
self	-0.05150*** (0.01131)	-0.05150*** (0.01131)
pov	0.10490 (0.09922)	0.10490 (0.09922)
Num.Obs.	1743	1743

Equivalencia de estimadores

Comprobamos que en un modelo de efectos fijos las características que no varían en el tiempo no pueden ser identificadas. Añadimos **black** para comprobarlo.

```
summary(plm( anti ~ self + pov + black,  
             data=data.comp, model="within",  
             index = c("id", "year")))
```

Coefficients:

	Estimate	Std. Error	t-value	Pr(> t)
self	-0.051495	0.010530	-4.8902	1.149e-06 ***
pov	0.104899	0.093880	1.1174	0.2641

(black omitido por ser invariante en el tiempo)

Equivalencia de estimadores

$$\hookrightarrow \hat{\beta}_{\text{W}} \text{ is the } \hat{\beta}_{\text{OLS}} \text{ for: } y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)' \beta + (e_{it} - \bar{e}_i)$$

Comprobamos que el estimador de efectos fijos es equivalente a MCO sobre el modelo en diferencias con respecto a la media. Conservamos dos periodos consecutivos y observaciones completas.

```
data.comp.sub <- data.comp %>%  
  dplyr::select(id, year, anti, self, pov) %>%  
  filter(year==90 | year==92)  
  
data.comp.sub <- data.comp.sub[complete.cases(data.comp.sub), ]
```

Equivalencia de estimadores

Creamos las variables como diferencias respecto a la media y estimamos el modelo within y el modelo de MCO en las variables transformadas:

```
data.comp.sub <- data.comp.sub %>%  
  group_by(id) %>%  
  mutate(m.anti = mean(anti), m.self = mean(self),  
         m.pov = mean(pov)) %>%  
  mutate(dm.anti = anti - m.anti, dm.self = self - m.self,  
         dm.pov = pov - m.pov)
```

Handwritten annotations for the code above:

- $\bar{y}_i$ points to `mean(anti)`
- $\bar{x}_{1i}$ points to `mean(self)`
- $\bar{x}_{2i}$ points to `mean(pov)`
- $y_{it} - \bar{y}_i$ points to `anti - m.anti`
- $(x_{it} - \bar{x}_i)$ points to `self - m.self`

```
m.fe.sub <- plm(anti ~ self + pov, data=data.comp.sub,  
               model="within", index = c("id", "year"))  
  
m.demean <- lm(dm.anti ~ dm.self + dm.pov, data.comp.sub)
```

Handwritten annotations for the model above:

- $y_{it} - \bar{y}_i$ points to `dm.anti`
- $(x_{it} - \bar{x}_i)$ points to `dm.self + dm.pov`

	Efectos fijos	MCO demean
self / dm.self	-0.03835** (0.01487)	-0.03835** (0.01487)
pov / dm.pov	0.19469 (0.14506)	0.19469 (0.14506)
Num.Obs.	1162	1162

Equivalencia de estimadores

Comprobamos que el estimador de efectos fijos es equivalente a MCO sobre el modelo en primeras diferencias.

```
data.comp.sub <- data.comp.sub %>%
```

```
  group_by(id) %>%
```

```
  mutate(d.anti = anti-dplyr::lag(anti, order_by = year),
```

```
         d.self = self-dplyr::lag(self, order_by = year),
```

```
         d.pov = pov-dplyr::lag(pov, order_by = year)) %>%
```

```
  ungroup()
```

```
m.difs <- lm(d.anti ~ -1 + d.self + d.pov, data=data.comp.sub)
```

	Ef. fijos	MCO demean	MCO 1 ^{ras} dif.
self	-0.03835** (0.01487)		
pov	0.19469 (0.14506)		
dm.self		-0.03835** (0.01487)	
dm.pov		0.19469 (0.14506)	
d.self			-0.03835** (0.01487)
d.pov			0.19469 (0.14506)
Num.Obs.	1162	1162	581

$$\hat{\beta}_{\text{fixed diff}} = \hat{\beta}_{\text{OLS of:}}$$

$$(y_{it} - y_{it-1}) = (x_{it} - x_{it-1})'\beta + (e_{it} - e_{it-1})$$

Why? $T=2$

Why?

Errores estándar

Estructura de los datos

Tomamos los datos *mlbook1.csv* con información sobre 2287 estudiantes en 131 escuelas. Nos interesa la relación entre **iq_vert** y **langpost**. Las variables **schoolnr** y **pupilnr** identifican escuelas y estudiantes:

$$\underbrace{\text{language fit score}}_{\text{langpost}_i} = \alpha + \beta \underbrace{\text{Verbal IQ Score}}_{\text{iqvert}_i} + BX_i + \varepsilon_i$$

donde X_i incluye sexo, minoría y años repetidos.

Es probable que estemos ante un problema de errores agrupados: los estudiantes de una misma escuela comparten características no observadas correlacionadas, rompiendo el supuesto de independencia.

Estimamos la ecuación ignorando la agrupación de datos.

```
m.mco <- lm(langpost ~ iq_verb + sex + minority + repeatgr,  
            data=data.examen)
```

	MCO
iq_verb	2.48635*** (0.07233)
sex	2.42228*** (0.28871)
minority	-0.03701 (0.62762)
repeatgr	-4.40860*** (0.43222)
Num.Obs.	2287

Una hora más en la prueba de aptitud incrementa en 2.49 puntos la calificación. El error estándar es 0.072.

Estimamos ahora los errores robustos a heterocedasticidad tipo HC1.

```
vcov=list(NULL, "HC1")
```

$$\hat{V}_{ols} = (X'X)^{-1} \left(\sum_i \sum_k x_{ik} x_{ik}' \hat{\epsilon}_{ik}^2 \right) (X'X)^{-1}$$

	MCO	MCO, err. robustos
iq_verb	2.48635*** (0.07233)	2.48635*** (0.07587)
sex	2.42228*** (0.28871)	2.42228*** (0.28852)
minority	-0.03701 (0.62762)	-0.03701 (0.61245)
repeatgr	-4.40860*** (0.43222)	-4.40860*** (0.44862)
Num.Obs.	2287	2287

El coeficiente es el mismo. El error estándar es aproximadamente 5% más grande, 0.076.

Estimador de efectos fijos

Estimamos usando MCO y efectos fijos de escuela.

```
m.mco.ef <- lm(langpost ~ iq_verb + sex + minority + repeatgr +  
  factor(schoolnr), data=data.examen)
```

	MCO	MCO, err. rob.	Ef. fijos
iq_verb	2.48635*** (0.07233)	2.48635*** (0.07587)	2.25997*** (0.07310)
sex	2.42228*** (0.28871)	2.42228*** (0.28852)	2.41900*** (0.26990)
minority	-0.03701 (0.62762)	-0.03701 (0.61245)	0.23262 (0.70803)
repeatgr	-4.40860*** (0.43222)	-4.40860*** (0.44862)	-4.43503*** (0.42731)
Num.Obs.	2287	2287	2287

Al incluir efectos fijos de escuela controlamos por características no observadas a nivel escuela, incorporadas como desplazamientos de la ordenada al origen.

Errores agrupados

$$\hat{V}(\hat{\beta})_{OLS_{CL=school}} = (X'X)^{-1} \left(\sum_{g=1}^G \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} x_{gi} x_{gj} \hat{u}_{gi} \hat{u}_{gj} \right) (X'X)^{-1}$$

Estimamos con errores **agrupados a nivel escuela** (sin efectos fijos).

```
clubSandwich::vcovCR(m.mco, type='CR1',  
  cluster=data.examen$schoolnr)
```

=

it is also rob.

	MCO	Err. rob.	Ef. fijos	Err. agrupados
iq_verb	2.48635*** (0.07233)	2.48635*** (0.07587)	2.25997*** (0.07310)	2.48635*** (0.08986)
sex	2.42228*** (0.28871)	2.42228*** (0.28852)	2.41900*** (0.26990)	2.42228*** (0.28436)
minority	-0.03701 (0.62762)	-0.03701 (0.61245)	0.23262 (0.70803)	-0.03701 (0.85396)
repeatgr	-4.40860*** (0.43222)	-4.40860*** (0.44862)	-4.43503*** (0.42731)	-4.40860*** (0.39547)
Num.Obs.	2287	2287	2287	2287

Los errores agrupados toman en cuenta la correlación de errores dentro de cada escuela.
El error de iq_verb es aprox. 20 % mayor respecto a MCO clásico y robusto.

Errores agrupados: factor de Moulton

Considere un modelo simple con datos agrupados:

$$y_{ig} = \alpha + \beta x_{ig} + u_{ig}$$

con tamaño promedio de grupo $\bar{N}_g$. Moulton (1990) muestra que el error estándar de MCO está sesgado hacia abajo por un factor igual a la raíz de

$$\tau \approx 1 + \rho_x \rho_u (\bar{N}_g - 1)$$

donde ρ_x es la correlación dentro de grupos de x y ρ_u la de los errores. Dependiendo del signo y magnitud de ρ_x y ρ_u , $\sqrt{\tau}$ puede ser menor que 1, por lo que el error agrupado puede ser menor que el de MCO. τ se conoce como el *factor de Moulton*.

refers barely
very so it is actually
a cluster level data

high $|\rho_u|$: more more
together within
cluster

Estimador de efectos fijos y errores agrupados

Combinamos efectos fijos de escuela con errores agrupados a nivel escuela.



	MCO	Err. rob.	Ef. fijos	Err. agrup.	Ef. fijos + agrup.
iq_verb	2.48635*** (0.07233)	2.48635*** (0.07587)	2.25997*** (0.07310)	2.48635*** (0.08986)	2.25997*** (0.08879)
sex	2.42228*** (0.28871)	2.42228*** (0.28852)	2.41900*** (0.26990)	2.42228*** (0.28436)	2.41900*** (0.27519)
minority	-0.03701 (0.62762)	-0.03701 (0.61245)	0.23262 (0.70803)	-0.03701 (0.85396)	0.23262 (0.69585)
repeatgr	-4.40860*** (0.43222)	-4.40860*** (0.44862)	-4.43503*** (0.42731)	-4.40860*** (0.39547)	-4.43503*** (0.39617)
Num.Obs.	2287	2287	2287	2287	2287

Obtenemos un coeficiente estimado de 2.26 para iq_verb, con un error estándar mayor, 0.08879.