Inferência Causal e Dados em Painel
- Aula 2-

Prof. Mestre. Omar Barroso Khodr

Instituto Brasileiro de Educação, Pesquisa e Desenvolvimento

MQO (Mínimos Quadrados Ordinários) agrupado(s)

- O método MQO agrupado, é uma técnica básica de regressão aplicada a dados em painel (ou dados longitudinais) onde observações são coletadas ao **longo do tempo** para as mesmas unidades (por exemplo, indivíduos, empresas, países).

TABLE 1.5 A Two-Year Panel Data Set on City Crime Statistics

obsno	city	year	murders	population	unem	police
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

Fonte: Wooldridge (2013)

TABLE 1.4 Pooled Cross Sections: Two Years of Housing Prices

obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85500	42	1600	3	2.0
2	1993	67300	36	1440	3	2.5
3	1993	134000	38	2000	4	2.5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
250	1993	243600	41	2600	4	3.0
251	1995	65000	16	1250	2	1.0
252	1995	182400	20	2200	4	2.0
253	1995	97500	15	1540	3	2.0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
520	1995	57200	16	1100	2	1.5

Fonte: Wooldridge (2013)

Definição MQO agrupado

- O MQO agrupado trata os dados em painel como uma única seção transversal, ignorando o tempo e as dimensões individuais.
- Ele estima uma **única equação de regressão** empilhando todas as observações ao longo do tempo.

Definição MQO agrupado

- Forma geral,
- $y_{it} = \beta_0 + \beta_1 x_{it1} + \beta_2 x_{it2} + \cdots + \beta_k x_{itk} + u_{it}$
- y_{it} : variável dependente para unidade ‘i’ no tempo ‘t’.
- $x_{it1} \ldots x_{itk}$: As variáveis explanatórias.
- u_{it} : Termo de erro.
- $\beta_0, \beta_1, \beta_2$: Os coeficientes a serem estimados.

Suposições (sob o modelo linear clássico)

- Para que MQO agrupados produzam estimativas imparciais e eficientes, os seguintes requisitos devem ser atendidos:

1. **Linearidade:** O modelo é linear em parâmetros.

2. **Amostragem aleatória:** As observações são i.i.d. (improváveis em dados em painel devido à correlação serial).

3. **Colinearidade inexistente:** Variáveis independentes não são perfeitamente correlacionadas.

Suposições (sob o modelo linear clássico)

4. Média Condicional Zero (Exogeneidade): $E(u_{it}|X) = 0$.

- Isso implica **ortogonalidade,** entre o termo de erro e o regressor. Ou seja, condicional a todos os regressores observados X, o valor esperado do termo de erro é zero [assim como a sua correlação].
- Nesse caso, nenhuma variável de viés é omitida e todas as variáveis y_{it} são inclusas em X, enquanto nenhuma é inclusa em u_{it}.
- Se esta suposição falhar (por exemplo, devido a variáveis omitidas, erro de medição ou simultaneidade), então $E(u_{it}|X) \neq 0$, leva a estimativas de MQO tendenciosas e inconsistentes.
- **Homoscedasticidade e correlação serial zero:** Ou seja, $\mathrm{Var}(u_{it}|X) = \sigma^2$ e $\mathrm{Cov}(u_{it}, u_{js}|X) = 0$; para $i \neq j$ ou $t \neq s$.

Exemplo

Exemplo (13.1) – Wooldridge

- Vamos considerar o exemplo 13.1 de Wooldridge (2013), utilizando dados da *General Social Survey* para explicar o número total de nascimento por mulheres (kids).
- Vamos supor que queremos utilizar 1972 como o ano-base para nosso estudo.
- Observe que para outros anos, a própria base de dados utilizou variáveis binárias para isolar o ano de 1972.
- Nesse contexto, nossa questão de interesse é saber: após controlados todos os outros fatores observados, o que aconteceu com as taxas de fertilidade ao longo do tempo?
- Os fatores **controlados** seriam educação, idade, etnia, região do país aonde as mulheres residiam quando tinham 16 anos e ambientes em que viviam quando tinham essa mesma idade.

Exemplo (13.1) – Wooldridge

```
Call:
lm(formula = kids ~ educ + age + agesq + black + east + northcen +
    west + farm + othrural + town + smcity + y74 + y76 + y78 +
    y80 + y82 + y84, data = fert)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9878 -1.0086 -0.0767  0.9331  4.6548

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.742457   3.051767  -2.537 0.011315 *
educ        -0.128427   0.018349  -6.999 4.44e-12 ***
age          0.532135   0.138386   3.845 0.000127 ***
agesq       -0.005804   0.001564  -3.710 0.000217 ***
black        1.075658   0.173536   6.198 8.02e-10 ***
east         0.217324   0.132788   1.637 0.101992
northcen     0.363114   0.120897   3.004 0.002729 **
west         0.197603   0.166913   1.184 0.236719
farm        -0.052557   0.147190  -0.357 0.721105
othrural    -0.162854   0.175442  -0.928 0.353481
town         0.084353   0.124531   0.677 0.498314
smcity       0.211879   0.160296   1.322 0.186507
y74          0.268183   0.172716   1.553 0.120771
y76         -0.097379   0.179046  -0.544 0.586633
y78         -0.068666   0.181684  -0.378 0.705544
y80         -0.071305   0.182771  -0.390 0.696511
y82         -0.522484   0.172436  -3.030 0.002502 **
y84         -0.545166   0.174516  -3.124 0.001831 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.555 on 1111 degrees of freedom
Multiple R-squared:  0.1295,    Adjusted R-squared:  0.1162
F-statistic: 9.723 on 17 and 1111 DF,  p-value: < 2.2e-16
```

Exemplo (13.1) – Wooldridge

- Análise de resultados...
- Os coeficientes das variáveis dummies mostram uma queda da fertilidade a partir de 1976.
- Por exemplo, o coeficiente y84 indica que, mantendo fixos: educação, idade, e outros fatores. Uma mulher teve, em média, (-0.545) menos filhos em 1984 do que em 1972.

Efeitos Fixos (FE – Fixed Effects)

- O modelo EF controla características individuais específicas não observadas (c_i) que são constantes ao longo do tempo, mas podem ser correlacionadas com os regressores.
- Isso ajuda a eliminar o viés de variável omitida decorrente de fatores invariantes ao longo do tempo.
- Nota variável omitida: Quando uma variável que potencialmente afeta x e y não é incluída no modelo. A variável excluída pode ser correlacionada com as variáveis independentes incluídas, levando a estimativas enviesadas desses coeficientes. Isso significa que o modelo não captura com precisão a verdadeira relação entre as variáveis.

Efeitos Fixos (FE – Fixed Effects)

- E.g.,
- $y_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_k x_{it} + c_i + u_i$
- y_{it}: a variável dependente para uma ‘entidade’ i sob o tempo t.
- x_{it1}: regressores variáveis no tempo (e.g., salários, índices e mudanças de políticas).
- c_i : Efeito fixo não observado (constante ao longo do tempo para cada i). Por exemplo, gênero e etnia.
- u_i : Erro idiossincrático que ocorre entre i e t.

Efeitos Fixos (FE – Fixed Effects)

- E.g.,
- $y_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_k x_{it} + c_i + u_i$
- Exogeneidade Estrita:
- $E(u_i | x_{i1}, x_{i2}, \ldots, x_{it}, c_i) = 0$
- Os erros u_i não são correlacionados com todos os regressores passados, presentes e futuros, condicionados a c_i.
- Violações ocorrem se houver feedback (por exemplo, o futuro x_{it+1} depende do passado u_i).

FE Método de Estimação

- Subtraindo as médias temporais específicas da entidade ("médias dentro do grupo") de cada variável:
- $y_{it} - \overline{y_{it}} = \beta_1(x_{it} - \bar{x}_{it}) + (u_{it} - \bar{u}_{it})$
- Assim, o método MQO é aplicado na equação transformada eliminando c_i.
- Com isso, inclui uma variável dummy para cada entidade i.
- Equivalente ao estimador interno, mas computacionalmente intensivo para grandes espaços amostrais N.

FE Método de Estimação

- Com a operação anterior, obtemos:

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it1} + \beta_2 \ddot{x}_{it2} + ... + \beta_k x_{itk} + \ddot{u}_{it}, \quad \ddot{t} = 1, 2, ..., T,$$

- Que podemos estimar pelo MQO agrupado...

Exemplo – Efeitos Fixos

- Pergunta de Pesquisa:
- *Como a educação afeta os salários, após considerar características individuais não observadas e invariantes ao longo do tempo (por exemplo, capacidade, motivação)?*

Exemplo – Efeitos Fixos

- Vamos estimar o seguinte modelo de acordo com Wooldridge (2013).
- $\log(wage_{it}) = \beta_0 + \beta_1 edu_{it} + \beta_2 xp_{it} + \beta_3 xp_{it}^2 + c_i + u_i$
- edu_{it} : Educação
- xp_{it}: experiência
- c_i: Efeitos Fixos (não mudam ao longo do tempo)

Exemplo – Efeitos Fixos

- Vamos estimar o seguinte modelo de acordo com Wooldridge (2013).
- $\log(wage_{it}) = \beta_0 + \beta_1 edu_{it} + \beta_2 xp_{it} + \beta_3 xp_{it}^2 + c_i + u_i$
- Como é possível ver no conjunto de dados, algumas variáveis, como experiência, estado civil e filiação sindical **mudam ao longo do tempo**. Outras variáveis, como etnia e educação não mudam.

Exemplo – Efeitos Fixos

```
> summary(fe_model, vcov = vcovHC, type = "HC1")
Oneway (individual) effect Within Model

Note: Coefficient variance-covariance matrix supplied: vcovHC

Call:
plm(formula = lwage ~ educ + exper + I(exper^2), data = wagepan_
    model = "within")

Balanced Panel: n = 545, T = 8, N = 4360

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-4.1752156 -0.1221201  0.0079743  0.1566831  1.4875328

Coefficients:
             Estimate  Std. Error t-value  Pr(>|t|)
exper       0.12225704  0.01057300 11.5631 < 2.2e-16 ***
I(exper^2) -0.00452280  0.00068718 -6.5817 5.283e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    572.05
Residual Sum of Squares: 473.26
R-Squared:      0.1727
Adj. R-Squared: 0.054231
F-statistic: 201.917 on 2 and 544 DF, p-value: < 2.22e-16
```

Exemplo – Efeitos Fixos

- Estrutura do Painel: Painel Balanceado: n = 545, T = 8, N = 4360
- n = 545: Número de indivíduos (entidades).
- T = 8: Cada indivíduo é observado 8 vezes (por exemplo, 8 anos).
- N = 4360: Total de observações (545 * 8).

Exemplo – Efeitos Fixos

- Somente regressores que variam no tempo aparecem na saída (variáveis invariantes no tempo como educ são absorvidas em c_i).
- Nesse contexto, educação é invariante no tempo neste conjunto de dados, é absorvido pelos efeitos fixos (c_i) e não pode ser estimado

```
> summary(fe_model, vcov = vcovHC, type = "HC1")
Oneway (individual) effect Within Model

Note: Coefficient variance-covariance matrix supplied: vcovHC

Call:
plm(formula = lwage ~ educ + exper + I(exper^2), data = wagepan_
    model = "within")

Balanced Panel: n = 545, T = 8, N = 4360

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-4.1752156 -0.1221201  0.0079743  0.1566831  1.4875328

Coefficients:
             Estimate  Std. Error t-value  Pr(>|t|)
exper       0.12225704  0.01057300 11.5631 < 2.2e-16 ***
I(exper^2) -0.00452280  0.00068718 -6.5817 5.283e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    572.05
Residual Sum of Squares: 473.26
R-Squared:      0.1727
Adj. R-Squared: 0.054231
F-statistic: 201.917 on 2 and 544 DF, p-value: < 2.22e-16
```

Exemplo – Efeitos Fixos

- R-quadrado (0,1727): Apenas 17,27% da variação intra-individual nos salários é explicada por $exper$ e $exper^2$.
- Lembrando que o R-quadrado indica que o modelo a variabilidade na variável dependente.
- O R-quadrado baixo é comum em modelos FE porque a maior parte da variação ocorre entre indivíduos (absorvida por c_i).
- R-quadrado ajustado (0,054): Penaliza o número de preditores. Ainda menor, sugerindo poder explicativo limitado.

```
> summary(fe_model, vcov = vcovHC, type = "HC1")
Oneway (individual) effect Within Model

Note: Coefficient variance-covariance matrix supplied: vcovHC

Call:
plm(formula = lwage ~ educ + exper + I(exper^2), data = wagepan_
    model = "within")

Balanced Panel: n = 545, T = 8, N = 4360

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-4.1752156 -0.1221201  0.0079743  0.1566831  1.4875328

Coefficients:
             Estimate  Std. Error t-value  Pr(>|t|)
exper       0.12225704  0.01057300 11.5631 < 2.2e-16 ***
I(exper^2) -0.00452280  0.00068718 -6.5817 5.283e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    572.05
Residual Sum of Squares: 473.26
R-Squared:      0.1727
Adj. R-Squared: 0.054231
F-statistic: 201.917 on 2 and 544 DF, p-value: < 2.22e-16
```

Exemplo – Efeitos Fixos

- O pequeno valor de p (< 2,2e-16) rejeita a hipótese nula, confirmando que $exper$ e $exper^2$ são importantes coletivamente.
- **A Experiência Importa:** Tanto os termos lineares quanto os quadráticos são estatisticamente significativos, mostrando que os salários aumentam com a experiência, mas a uma taxa decrescente.
- **Efeitos Fixos Trabalhados:** O modelo controlou a heterogeneidade não observada e invariante no tempo (por exemplo, habilidade, gênero).

```
> summary(fe_model, vcov = vcovHC, type = "HC1")
Oneway (individual) effect Within Model

Note: Coefficient variance-covariance matrix supplied: vcovHC

Call:
plm(formula = lwage ~ educ + exper + I(exper^2), data = wagepan_
    model = "within")

Balanced Panel: n = 545, T = 8, N = 4360

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-4.1752156 -0.1221201  0.0079743  0.1566831  1.4875328

Coefficients:
             Estimate  Std. Error t-value  Pr(>|t|)
exper       0.12225704  0.01057300 11.5631 < 2.2e-16 ***
I(exper^2) -0.00452280  0.00068718 -6.5817 5.283e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    572.05
Residual Sum of Squares: 473.26
R-Squared:      0.1727
Adj. R-Squared: 0.054231
F-statistic: 201.917 on 2 and 544 DF, p-value: < 2.22e-16
```

Exemplo – Efeitos Fixos

- **R-quadrado baixo:** O modelo explica pouco da variação salarial intra-individual, sugerindo que outros fatores que variam com o tempo (por exemplo, mudanças de emprego, treinamento) podem ser necessários.
- **Erros Padrão Robustos:** O uso de vcovHC garante uma inferência válida, apesar da potencial heterocedasticidade/correlação serial.

```
> summary(fe_model, vcov = vcovHC, type = "HC1")
Oneway (individual) effect Within Model

Note: Coefficient variance-covariance matrix supplied: vcovHC

Call:
plm(formula = lwage ~ educ + exper + I(exper^2), data = wagepan_
    model = "within")

Balanced Panel: n = 545, T = 8, N = 4360

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-4.1752156 -0.1221201  0.0079743  0.1566831  1.4875328

Coefficients:
             Estimate  Std. Error t-value  Pr(>|t|)
exper       0.12225704  0.01057300 11.5631 < 2.2e-16 ***
I(exper^2) -0.00452280  0.00068718 -6.5817 5.283e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    572.05
Residual Sum of Squares: 473.26
R-Squared:      0.1727
Adj. R-Squared: 0.054231
F-statistic: 201.917 on 2 and 544 DF, p-value: < 2.22e-16
```

Por que não encontramos Educ?

- Perceba que Educ não têm variação temporal. Nesse contexto, essa coluna é invariante. O que nos dá a missão em colocar como um efeito fixo e eliminar esse efeito.

Próximos passos

- Percebemos que pelo nosso R-Quadrado que nosso poder explicativo dos salários não foi muito forte.
- Pensando em causalidade, como poderíamos melhorar nosso modelo e quais outros regressores poderíamos utilizar?

Bibliografia

- Angrist, J.D. and Pischke, J.-S., 2009. Mostly Harmless Econometrics. 1st ed. Princeton University Press.
- Cunningham, S., 2021. Causal Inference: The Mixtape. Yale University Press. https://doi.org/10.2307/j.ctv1c29t27.
- Wooldridge, J.M. (2013) Introductory econometrics: a modern approach. 5th ed. Michigan State University.