

Introdução ao Software R e à Análise Econométrica

Agosto de 2018

Alexandre Xavier Ywata Carvalho
Geraldo Sandoval Góes

Introdução à Regressão Linear com Dados de Painel

Regressão com Dados de Paineis

- Considere o modelo de regressão tradicional:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \epsilon_i$$

- Nesse modelo, em geral, esse tipo de modelo se aplica a situações nas quais temos dados do tipo *cross-section*, ou dados de séries temporais
- Com o avanço nos métodos de coleta e armazenamento de informações, cada vez mais estão disponíveis bases de dados de painéis
 - Para cada unidade observacional, temos observações para diferentes unidades de tempo
- A possibilidade de observar os dados em diferentes instantes do tempo fornece a possibilidade de diferentes tipos de análise
- As técnicas para tratamento e análise de dados de painéis estão amplamente disponíveis
- No R, vamos utilizar nos exemplos bases de dados disponibilizadas pelos próprios pacotes. O principal pacote para dados de painéis é o “plm”
- Painéis podem ser: **balanceado** (mesmo número de períodos para cada unidade observacional) ou **não-balanceado**

Regressão com Dados de Painel

- Como exemplo, vamos carregar quatro tabelas de dados:

#---- carregando os dados

```
data("EmplUK", package = "plm")  
data("Wages", package = "plm")  
data("Grunfeld", package = "plm")  
data("Produc", package = "plm")
```

#--- descrição dos dados

```
?EmplUK  
?Wages  
?Grunfeld  
?Produc
```

#---- sumário dos dados

```
summary(EmplUK)  
summary(Wages)  
summary(Produc)  
summary(Grunfeld)
```

	firm	year	sector	emp	wage	capital	output
1	1	1977	7	5.041000	13.1516	0.5894	95.7072
2	1	1978	7	5.600000	12.3018	0.6318	97.3569
3	1	1979	7	5.015000	12.8395	0.6771	99.6083
4	1	1980	7	4.715000	13.8039	0.6171	100.5501
5	1	1981	7	4.093000	14.2897	0.5076	99.5581
6	1	1982	7	3.166000	14.8681	0.4229	98.6151
7	1	1983	7	2.936000	13.7784	0.3920	100.0301
8	2	1977	7	71.319000	14.7909	16.9363	95.7072
9	2	1978	7	70.642998	14.1036	17.2422	97.3569
10	2	1979	7	70.917999	14.9534	17.5413	99.6083
11	2	1980	7	72.030998	15.4910	17.6574	100.5501
12	2	1981	7	73.689003	16.1969	16.7133	99.5581

An **unbalanced** panel of 140 observations from 1976 to 1984
total number of observations : 1031
observation : firms, country : United Kingdom

Firm - firm index

Year - year

Sector - the sector of activity

Emp - employment

Wage - wages

Capital - capital

Output - output

Regressão com Dados de Painel

- As fórmulas para regressão com dados de painel são bastante flexíveis, permitindo a inclusão de lags (defasagens), leads (valores futuros) e diferenças (valor de uma variável menos um valor no período anterior)
- $\text{lag}(\log(\text{emp}), 1)$ indica o valor defasado (no período anterior) do logaritmo natural da variável “emp”
- $\text{lag}(\log(\text{wage}), 3)$ indica o valor defasado de três períodos anteriores, do logaritmo natural da variável “wage”
- $\text{diff}(\log(\text{capital}), 2)$ indica o valor de $\log(\text{capital})$ – o valor de $\log(\text{capital})$ dois períodos anteriores

#---- exemplo de regressão com dados de painel (fórmulas gerando o mesmo resultado)

```
formula1 <- log(emp) ~ lag(log(emp), 1) + lag(log(emp), 2) + lag(log(wage), 2) + lag(log(wage), 3) +  
                    diff(log(capital), 2) + diff(log(capital), 3)
```

```
Emp.mod1 <- plm(formul = formula1, data = EmplUK, model = "within")
```

```
summary(Emp.mod1)
```

```
formula2 <- log(emp) ~ lag(log(emp), 1) + lag(log(emp), 2) + lag(log(wage), 2) + lag(log(wage), 3) +  
                    I(log(capital) - lag(log(capital), 2)) + I(log(capital) - lag(log(capital), 3))
```

```
Emp.mod2 <- plm(formul = formula2, data = EmplUK, model = "within")
```

```
summary(Emp.mod2)
```

Regressão com Dados de Paineis

- Em geral, reescrevemos a equação linear para dados de painéis, indexando as unidades observacionais i ($i = 1, \dots, n$) e os períodos de tempo t ($t = 1, \dots, T$)

$$y_{i,t} = \beta_0 + \beta_1 x_{1,i,t} + \beta_2 x_{2,i,t} + \dots + \beta_k x_{k,i,t} + \epsilon_{i,t}$$

- Vamos assumir, por enquanto, que o erro $\epsilon_{i,t}$ possui distribuição normal, com média zero, e variância σ_ϵ^2
- Além disso, vamos assumir por enquanto que $\epsilon_{i,t}$ são erros não correlacionados entre si
- O elemento $y_{i,t}$ corresponde ao valor da variável resposta da unidade i , no período t
- O item $x_{k,i,t}$ corresponde à k -ésima variável explicativa, para a unidade i , no período t
- Na versão da equação acima, os dados para cada unidade observacional estão “empilhados”
- Podemos estimar os parâmetros desconhecidos $\beta_0, \beta_1, \dots, \beta_k$ utilizando um estimador de mínimos quadrados ordinários para os dados empilhados
- A estimação usando os dados empilhados e aplicando um estimador de MQO é conhecida como estimação ou regressão do tipo “**pooled**”

Regressão com Dados de Painel

- Exemplo:

```
Emp.pooled1 <- plm(formula = formula1, data = EmplUK, model = "pooling")  
summary(Emp.pooled1)
```

- Output:

```
> summary(Emp.pooled1)  
Pooling Model
```

Call:

```
plm(formula = formula1, data = EmplUK, model = "pooling")
```

Unbalanced Panel: n=140, T=4-6, N=611

Residuals :

Min.	1st Qu.	Median	3rd Qu.	Max.
-0.72500	-0.05210	0.00401	0.05520	0.91900

Regressão com Dados de Paineis

- Exemplo (continuação):

Coefficients :

	Estimate	Std. Error	t-value	Pr(> t)
(Intercept)	0.0609696	0.0599305	1.0173	0.3094
lag(log(emp), 1)	0.9545302	0.0426280	22.3921	< 2.2e-16 ***
lag(log(emp), 2)	0.0337330	0.0424782	0.7941	0.4274
lag(log(wage), 2)	-0.0037779	0.0656215	-0.0576	0.9541
lag(log(wage), 3)	-0.0234908	0.0623861	-0.3765	0.7066
diff(log(capital), 2)	0.3174782	0.0409002	7.7623	3.581e-14 ***
diff(log(capital), 3)	-0.0111866	0.0325766	-0.3434	0.7314

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares: 1099.2

Residual Sum of Squares: 9.1562

R-Squared: 0.99167

Adj. R-Squared: 0.99159

F-statistic: 11984.9 on 6 and 604 DF, p-value: < 2.22e-16

Regressão com Dados de Paineis

- Exercício prático. Na regressão abaixo,
 - Adicione um lag de ordem 3 para o nível de emprego (emp), e um lag de ordem 1 para a variável de salários (wage)
 - Rode um modelo de painel com estimador do tipo “pooled”
 - Verifique as variáveis lag adicionadas são estatisticamente significantes.

```
formula1 <- log(emp) ~ lag(log(emp), 1) + lag(log(emp), 2) + lag(log(wage), 2)  
                + lag(log(wage), 3) + diff(log(capital), 2) + diff(log(capital), 3)
```

```
Emp.pooled1 <- plm(formula = formula1, data = EmplUK, model = "pooling")  
summary(Emp.pooled1)
```

Regressão com Dados de Paineis

- O problema com regressão do tipo pooled é que perdemos a oportunidade de tentar identificar diferenças intrínsecas entre as unidades observacionais
- Essas diferenças não necessariamente estão contabilizadas nas variáveis explicativas $x_{1,i,t}$, $x_{2,i,t}$, ..., $x_{k,i,t}$ (variáveis observáveis)
- Para contornar isso, nós trabalhamos então com a inclusão de variáveis dummy específicas para cada unidade observacional
- Por exemplo, em um painel de municípios, observados em diferentes anos, nós estamos tentando identificar os efeitos específicos das características de cada município individualmente
- Nesse caso, a regressão para dados de painéis é reescrita na forma:

$$y_{i,t} = \alpha_i + \beta_1 x_{1,i,t} + \beta_2 x_{2,i,t} + \dots + \beta_k x_{k,i,t} + \epsilon_{i,t}$$

- O parâmetro α_i corresponde ao efeito idiossincrático para a unidade observacional i
 - A ideia é que o termo contabilize por características da unidade i (municípios, por exemplo), que não se alteram ao longo dos anos, e que não sejam contabilizadas pelas variáveis observáveis $x_{1,i,t}$, $x_{2,i,t}$, ..., $x_{k,i,t}$

Resultado dos efeitos do FNE sobre o crescimento médio anual do PIB *per capita* no nível municipal – método painel de efeitos fixos

Método de estimação	Variável dependente – taxa de crescimento anual média do PIB <i>per capita</i>				
	Painel efeitos fixos	Painel efeitos fixos	Painel efeitos fixos	Painel efeitos fixos	
	(1)	(2)	(3)	(4)	
Alta renda_Proporção do FNE início do período (1º ano) em relação ao PIB do início de cada período	0.9982** (0.0157)	0.8501** (0.0208)	Alta renda_Proporção do FNE início do período (1º + 2º ano) em relação ao PIB do início de cada período	-0.0122 (0.5977)	-0.0380* (0.0665)
Dinâmica_Proporção do FNE início do período (1º ano) em relação ao PIB do início de cada período	0.1407*** (0.0006)	0.1225*** (0.0010)	Dinâmica_Proporção do FNE início do período (1º + 2º ano) em relação ao PIB do início de cada período	0.1282*** (0.0000)	0.1066*** (0.0000)
Baixa renda_Proporção do FNE início do período (1º ano) em relação ao PIB do início de cada período	0.4528*** (0.0000)	0.2129*** (0.001)	Baixa renda_Proporção do FNE início do período (1º + 2º ano) em relação ao PIB do início de cada período	0.0934*** (0.0002)	0.0273 (0.2259)
Estagnada_Proporção do FNE início do período (1º ano) em relação ao PIB do início de cada período	0.1508** (0.0411)	-0.0191 0.7733	Estagnada_Proporção do FNE início do período (1º + 2º ano) em relação ao PIB do início de cada período	0.1322*** (0.0000)	0.0639*** (0.0099)
Ln (PIB <i>per capita</i> no início de cada período)	-0.1693*** (0.0000)	-0.2944*** (0.0000)	Ln (PIB <i>per capita</i> no início de cada período)	-0.1681*** (0.0000)	-0.2936*** (0.0000)
Ln (anos médios de escolaridade no início de cada período, Rais)	0.0670*** (0.0000)	-0.0103** (0.0138)	Ln (anos médios de escolaridade no início de cada período, Rais)	0.0653*** (0.0000)	-0.01090*** (0.0091)
Ln (densidade populacional no início de cada período)	0.0926*** (0.0000)	-0.1280*** (0.0000)	Ln (densidade populacional no início de cada período)	0.0886*** (0.0000)	-0.1280*** (0.0000)
Efeitos fixos	Sim	Sim	Efeitos fixos	Sim	Sim
<i>Dummy</i> de tempo	Não	Sim	<i>Dummy</i> de tempo	Não	Sim
Número de observações (municípios)	5.946	5.946		5.946	5.946
R2 ajustado	0.1739	0.3368		0.1779	0.3403

Regressão com Dados de Paineis

- Além dos efeitos individuais de cada unidade observacional, podemos também incluir efeitos específicos δ_t dos períodos de tempo:

$$y_{i,t} = \alpha_i + \delta_t + \beta_1 x_{1,i,t} + \beta_2 x_{2,i,t} + \dots + \beta_k x_{k,i,t} + \epsilon_{i,t}$$

- Quando o número de unidades observacionais n não é muito grande, podemos estimar os efeitos α_i simplesmente adicionando dummies à regressão, da mesma forma que fizemos nas aulas anteriores
- No entanto, em geral, o número n é da ordem de milhares (exemplo, $n = 5564$ municípios), e teríamos que incluir 5564 dummies (ou 5563) na regressão
- Em estudos longitudinais de trabalhadores, por exemplo, o valor n pode chegar a dezenas ou centenas de milhares
- Computacionalmente, temos então um problema prático de estimar os coeficientes α_i , quando n é alto
- Uma parcela considerável dos avanços na análise de regressão com dados de painéis corresponde justamente a técnicas para estimarmos os coeficientes α_i
- Iremos agora discutir uma boa parte desses procedimentos

Regressão com Dados de Paineis

- Vamos considerar o caso mais geral, conforme regressão abaixo (os efeitos fixos de período δ_t podem estar representados por algumas das variáveis explicativas):

$$y_{i,t} = \alpha_i + \beta_1 x_{1,i,t} + \beta_2 x_{2,i,t} + \dots + \beta_k x_{k,i,t} + \epsilon_{i,t}$$

- As variáveis explicativas podem conter também defasagens, diferenças, vários tipos de variáveis dummy, etc.
- A literatura divide a estimação dos coeficientes α_i , de acordo com duas situações:
 - **Estimadores de efeitos aleatórios** para α_i - nesse caso, assume-se que os coeficientes α_i são termos aleatórios, com variância σ_α^2 , e esses termos não são correlacionados com os erros $\epsilon_{i,t}$
 - **Estimadores de efeitos fixos** para α_i - utilizamos esses estimadores quando a hipótese de que os coeficientes α_i são não correlacionados com os erros $\epsilon_{i,t}$ não é uma hipótese válida
- Dentro de cada um desses dois grandes grupos de estimadores, há uma série de variações
- Na prática, os estimadores de efeitos fixos são mais comuns, por conta da tendência de rejeitarmos a hipótese nula de que os coeficientes α_i são não correlacionados com os erros $\epsilon_{i,t}$

Estimadores de Efeitos Fixos

- No caso dos estimadores de efeitos fixos, temos então que estimar diretamente os coeficientes α_i na equação abaixo:

$$y_{i,t} = \alpha_i + \beta_1 x_{1,i,t} + \beta_2 x_{2,i,t} + \dots + \beta_k x_{k,i,t} + \epsilon_{i,t} \quad (\text{A})$$

- Para isso, vamos inicialmente aplicar a soma para todas as observações em cada unidade i , e dividir por n

$$\frac{1}{n} \sum_{t=1}^T y_{i,t} = \frac{1}{n} \sum_{t=1}^n [\alpha_i + \beta_1 x_{1,i,t} + \beta_2 x_{2,i,t} + \dots + \beta_k x_{k,i,t} + \epsilon_{i,t}]$$

$$\bar{y}_{i,\cdot} = \alpha_i + \beta_1 \bar{x}_{1,i,\cdot} + \beta_2 \bar{x}_{2,i,\cdot} + \dots + \beta_k \bar{x}_{k,i,\cdot} + \bar{\epsilon}_{i,\cdot} \quad (\text{B})$$

- Subtraindo (A) – (B), obtemos a equação:

$$(y_{i,t} - \bar{y}_{i,\cdot}) = \beta_1 (x_{1,i,t} - \bar{x}_{1,i,\cdot}) + \dots + \beta_k (x_{k,i,t} - \bar{x}_{k,i,\cdot}) + (\epsilon_{i,t} - \bar{\epsilon}_{i,\cdot})$$

- Ou alternativamente

$$(y_{i,t} - \bar{y}_{i,\cdot}) = \beta_1 (x_{1,i,t} - \bar{x}_{1,i,\cdot}) + \dots + \beta_k (x_{k,i,t} - \bar{x}_{k,i,\cdot}) + \epsilon_{i,t}^* \quad (\text{C})$$

- Com $\epsilon_{i,t}^* = (\epsilon_{i,t} - \bar{\epsilon}_{i,\cdot})$

Estimadores de Efeitos Fixos

- Considere então a fórmula

$$(y_{i,t} - \bar{y}_{i,\cdot}) = \beta_1(x_{1,i,t} - \bar{x}_{1,i,\cdot}) + \dots + \beta_k(x_{k,i,t} - \bar{x}_{k,i,\cdot}) + \epsilon_{i,t}^*$$

- Podemos estimar os parâmetros β_1, \dots, β_k através da regressão da variável $(y_{i,t} - \bar{y}_{i,\cdot})$, versus as variáveis explicativas $(x_{1,i,t} - \bar{x}_{1,i,\cdot}), (x_{2,i,t} - \bar{x}_{2,i,\cdot}), \dots, (x_{k,i,t} - \bar{x}_{k,i,\cdot})$
- Um cuidado adicional deve ser tomado para o fato de o erro $\epsilon_{i,t}^* = (\epsilon_{i,t} - \bar{\epsilon}_{i,\cdot})$ não ser mais não-correlacionado. Devido ao termo $\bar{\epsilon}_{i,\cdot}$, aparecendo em todos os períodos para cada unidade i , o termo $\epsilon_{i,t}^*$ apresenta uma correlação com $\epsilon_{i,s}^*$, com $t \neq s$
- Vamos nos preocupar com os resíduos correlacionados mais adiante
- Um primeiro fato importante da equação na fórmula (C) é que, no estimador de efeitos fixos, não podemos incluir entre as variáveis explicativas uma variável que seja constante para todos os períodos de tempo, para todas as unidades i
- De fato, se tivermos $x_{1,i,t} = x_{1,i}$, para toda unidade i , então $\bar{x}_{1,i,\cdot} = x_{1,i}$; portanto, $(x_{1,i,t} - \bar{x}_{1,i,\cdot}) = x_{1,i} - x_{1,i} = 0$, para toda unidade i . Na nossa matriz X de variáveis explicativas vamos ter uma coluna somente com zeros, e não será possível estimar os coeficientes da regressão
- Portanto, quando estivermos usando estimadores de efeitos fixos, NÃO podemos incluir variáveis explicativas que não variem no tempo, para pelo menos algumas das unidades i

Estimadores de Efeitos Fixos

- Considere novamente a equação com os efeitos fixos α_i :

$$y_{i,t} = \alpha_i + \beta_1 x_{1,i,t} + \beta_2 x_{2,i,t} + \dots + \beta_k x_{k,i,t} + \epsilon_{i,t}$$

- Em princípio, o efeito fixo α_i serve justamente para capturar, em um único coeficiente, todas as especificidades da unidade (por exemplo, município) i
- Se uma determinada variável $x_{1,i,t}$ é constante ao longo de todos os períodos, não precisamos incluí-la na regressão. O seu efeito em princípio está sendo capturado pela constante α_i
- Considere então uma regressão na qual as unidades observacionais i sejam municípios brasileiros, e os períodos t sejam anos consecutivos
 - Nesse caso, é tentador incluir na regressão, variáveis explicativas com base no Censo 2010, e repetir esses valores para os demais anos
 - O problema dessa estratégia é justamente o fato de que, quando repetimos os valores do Censo 2010 para os demais anos, as variáveis resultantes não variam ao longo dos períodos
 - Portanto, não podemos utilizar estimadores de efeitos fixos nesses casos
 - Alternativas: usar fontes de dados que possuem informações anuais: Censo Escolar, DataSus, transferências de renda, RAIS etc.

Estimadores de Efeitos Fixos

- O estimador de efeitos fixos baseado na fórmula abaixo é conhecido como estimador “**within**” ou estimador “**demeaned**”

$$(y_{i,t} - \bar{y}_{i,\cdot}) = \beta_1(x_{1,i,t} - \bar{x}_{1,i,\cdot}) + \dots + \beta_k(x_{k,i,t} - \bar{x}_{k,i,\cdot}) + \epsilon_{i,t}^* \quad (C)$$

- Esse é um estimador muito utilizado na prática
- Com base nas estimativas $\hat{\beta}_1, \dots, \hat{\beta}_k$ para os parâmetros β_1, \dots, β_k , podemos empregar a equação (B), para encontrar estimativas para os efeitos fixos α_i

$$\hat{\alpha}_i = \bar{y}_{i,\cdot} - [\beta_1 \bar{x}_{1,i,\cdot} + \beta_2 \bar{x}_{2,i,\cdot} + \dots + \beta_k \bar{x}_{k,i,\cdot}]$$

- Em muitos casos, é útil analisar esses efeitos fixos, utilizando, por exemplo, análises gráficas

Estimadores de Efeitos Fixos

- Abaixo sintaxe em R para estimar uma regressão de painel com efeitos fixos, via estimador within

#---- regressão com estimador de efeitos fixos, do tipo within

```
Emp.within1 <- plm(formul = formula1, data = EmplUK, model = "within")  
summary(Emp.within1)
```

```
fixef(Emp.within1) #--- extraindo os efeitos fixos de cada unidade
```

#---- incluindo efeitos dos períodos

```
Emp.within1 <- plm(log(emp) ~ lag(log(emp), 1) + lag(log(emp), 2) + lag(log(wage), 2) + lag(log(wage), 3) +  
diff(log(capital), 2) + diff(log(capital), 3) + as.factor(year), data = EmplUK, model = "within")  
summary(Emp.within1)
```

Oneway (individual) effect Within Model

Call:

```
plm(formula = formula1, data = EmplUK, model = "within")
```

Unbalanced Panel: n=140, T=4-6, N=611

Residuals :

Min.	1st Qu.	Median	3rd Qu.	Max.
-0.5870	-0.0462	0.0035	0.0463	0.8170

Coefficients :

	Estimate	Std. Error	t-value	Pr(> t)
lag(log(emp), 1)	0.661436	0.045003	14.6977	< 2.2e-16 ***
lag(log(emp), 2)	0.014835	0.054911	0.2702	0.7871501
lag(log(wage), 2)	-0.024417	0.084921	-0.2875	0.7738394
lag(log(wage), 3)	0.092591	0.080758	1.1465	0.2521638
diff(log(capital), 2)	0.192682	0.041313	4.6640	4.059e-06 ***
diff(log(capital), 3)	0.124067	0.036670	3.3833	0.0007767 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares: 18.394

Residual Sum of Squares: 6.0765

R-Squared: 0.66965

Adj. R-Squared: 0.56663

F-statistic: 157.097 on 6 and 465 DF, p-value: < 2.22e-16

R^2 da regressão (C)

Regressão com Dados de Painel

- Exercício prático. Na regressão abaixo,
 - Adicione um lag de ordem 3 para o nível de emprego (emp), e um lag de ordem 1 para a variável de salários (wage)
 - Rode um modelo de painel com estimador do tipo “within”
 - Verifique as variáveis lag adicionadas são estatisticamente significantes.

```
formula1 <- log(emp) ~ lag(log(emp), 1) + lag(log(emp), 2) + lag(log(wage), 2)  
+ lag(log(wage), 3) + diff(log(capital), 2) + diff(log(capital), 3)
```

Estimadores de Efeitos Fixos

- Considere novamente o problema de estimar os coeficientes β_1, \dots, β_k na equação abaixo:

$$y_{i,t} = \alpha_i + \beta_1 x_{1,i,t} + \beta_2 x_{2,i,t} + \dots + \beta_k x_{k,i,t} + \epsilon_{i,t} \quad (\text{A})$$

- Considere a primeira defasagem da equação (A) acima

$$y_{i,t-1} = \alpha_i + \beta_1 x_{1,i,t-1} + \beta_2 x_{2,i,t-1} + \dots + \beta_k x_{k,i,t-1} + \epsilon_{i,t-1} \quad (\text{A1})$$

- Subtraindo (A) – (A1), obtemos a equação:

$$(y_{i,t} - y_{i,t-1}) = (\alpha_i - \alpha_i) + \beta_1 (x_{1,i,t} - x_{1,i,t-1}) + \dots + \beta_k (x_{k,i,t} - x_{k,i,t-1}) + (\epsilon_{i,t} - \epsilon_{i,t-1})$$

- Resultando:

$$(y_{i,t} - y_{i,t-1}) = \beta_1 (x_{1,i,t} - x_{1,i,t-1}) + \dots + \beta_k (x_{k,i,t} - x_{k,i,t-1}) + \tilde{\epsilon}_{i,t}$$

- Com o novo termo de erro $\epsilon_{i,t}^* = (\epsilon_{i,t} - \epsilon_{i,t-1})$. Esse termo também apresenta correlação, mas nós não iremos detalhar isso agora
- Note que a regressão acima está em primeiras diferenças tanto para as variáveis explicativas como para a variável resposta

$$\Delta y_{i,t} = \beta_1 \Delta x_{1,i,t} + \beta_2 \Delta x_{2,i,t} + \dots + \beta_k \Delta x_{k,i,t} + \tilde{\epsilon}_{i,t} \quad (\text{D})$$

Estimadores de Efeitos Fixos

- O estimador com base na regressão abaixo é conhecido como estimador de efeitos fixos, do tipo **primeiras diferenças (“first differences”)**

$$\Delta y_{i,t} = \beta_1 \Delta x_{1,i,t} + \beta_2 \Delta x_{2,i,t} + \dots + \beta_k \Delta x_{k,i,t} + \tilde{\epsilon}_{i,t} \quad (D)$$

- Diferentemente do estimador do tipo “within”, no caso do estimador de primeiras diferenças não é possível obter os coeficientes $\hat{\alpha}_i$
- Abaixo a sintaxe para estimar no R

#---- regressão com estimador de efeitos fixos, do tipo first differences

```
Emp.fd1 <- plm(formul = formula1, data = EmplUK, model = "fd")  
summary(Emp.fd1)
```

```
fixef(Emp.fd1) #--- extraindo os efeitos fixos de cada unidade (vai dar erro!)
```

Oneway (individual) effect First-Difference Model

Call:

```
plm(formula = formula1, data = EmplUK, model = "fd")
```

Unbalanced Panel: n=140, T=4-6, N=611

Observations used in estimation: 471

Número de observações utilizadas

Residuals :

```
Min. 1st Qu. Median 3rd Qu. Max.
-0.92100 -0.05530 0.00815 0.05450 0.91800
```

Coefficients :

```
Estimate Std. Error t-value Pr(>|t|)
(intercept) -0.0475518 0.0081571 -5.8295 1.042e-08 ***
lag(log(emp), 1) 0.1329223 0.0449745 2.9555 0.003280 **
lag(log(emp), 2) 0.1117792 0.0572729 1.9517 0.051576 .
lag(log(wage), 2) 0.0606044 0.0884320 0.6853 0.493483
lag(log(wage), 3) -0.0676549 0.0861997 -0.7849 0.432935
diff(log(capital), 2) 0.1158038 0.0364845 3.1741 0.001603 **
diff(log(capital), 3) 0.1714872 0.0370334 4.6306 4.740e-06 ***
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R² da regressão (D)

Total Sum of Squares: 10.637

Residual Sum of Squares: 9.0427

R-Squared: 0.14991

Adj. R Squared: 0.13892

F-statistic: 13.6375 on 6 and 464 DF, p-value: 2.7803e-14

Regressão com Dados de Painel

- Exercício prático. Na regressão abaixo,
 - Adicione um lag de ordem 3 para o nível de emprego (emp), e um lag de ordem 1 para a variável de salários (wage)
 - Rode um modelo de painel com estimador do tipo “first differences”
 - Verifique as variáveis lag adicionadas são estatisticamente significantes.

```
formula1 <- log(emp) ~ lag(log(emp), 1) + lag(log(emp), 2) + lag(log(wage), 2)  
+ lag(log(wage), 3) + diff(log(capital), 2) + diff(log(capital), 3)
```

Estimadores de Efeitos Fixos

- Voltemos novamente ao problema de estimar os coeficientes β_1, \dots, β_k na equação abaixo:

$$y_{i,t} = \alpha_i + \beta_1 x_{1,i,t} + \beta_2 x_{2,i,t} + \dots + \beta_k x_{k,i,t} + \epsilon_{i,t} \quad (\text{A})$$

- Vamos agora derivar o terceiro tipo de estimador de efeitos fixos, conhecido como estimador “**between**”
- Para isso, lembrando a equação (B) acima:

$$\bar{y}_{i,\cdot} = \alpha_i + \beta_1 \bar{x}_{1,i,\cdot} + \beta_2 \bar{x}_{2,i,\cdot} + \dots + \beta_k \bar{x}_{k,i,\cdot} + \bar{\epsilon}_{i,\cdot} \quad (\text{B})$$

- Podemos reagrupar os termos, inserindo um intercepto α , obtendo

$$\bar{y}_{i,\cdot} = \alpha + \beta_1 \bar{x}_{1,i,\cdot} + \beta_2 \bar{x}_{2,i,\cdot} + \dots + \beta_k \bar{x}_{k,i,\cdot} + [\alpha_i - \alpha + \bar{\epsilon}_{i,\cdot}]$$

- Ou, reescrevendo,

$$\bar{y}_{i,\cdot} = \alpha + \beta_1 \bar{x}_{1,i,\cdot} + \beta_2 \bar{x}_{2,i,\cdot} + \dots + \beta_k \bar{x}_{k,i,\cdot} + \bar{\epsilon}_{i,\cdot}^* \quad (\text{B1})$$

- Onde $\bar{\epsilon}_{i,\cdot}^* = [\alpha_i - \alpha + \bar{\epsilon}_{i,\cdot}]$ é um termo de erro nessa nova equação. Esse termo também é correlacionado entre si, mas pode ser ajustado devidamente na estimação
- O estimador “**between**” corresponde simplesmente a uma regressão *cross-section* das médias da variável predita versus as médias das variáveis explicativas

Estimadores de Efeitos Fixos

- O estimador “**between**” corresponde simplesmente a uma regressão *cross-section* das médias da variável predita versus as médias das variáveis explicativas
- Da mesma forma que no estimador de primeiras diferenças, não é possível obter diretamente estimativas para os efeitos fixos α_i
- O estimador do tipo “**between**” não apresenta vantagens em relação aos demais estimadores
- Sintaxe no R:

#---- regressão com estimador de efeitos fixos, do tipo between

```
Emp.between1 <- plm(formul = formula1, data = EmplUK, model = "between")  
summary(Emp.between1)
```

```
fixef(Emp.between1) #--- extraindo os efeitos fixos de cada unidade (vai dar erro!)
```

Oneway (individual) effect Between Model

Call:

```
plm(formula = formula1, data = EmplUK, model = "between")
```

Unbalanced Panel: n=140, T=4-6, N=611

Observations used in estimation: 140

Número de observações utilizadas
(cross-section)

Residuals :

Min.	1st Qu.	Median	3rd Qu.	Max.
-0.080100	-0.011600	0.000133	0.013700	0.072600

Coefficients :

	Estimate	Std. Error	t-value	Pr(> t)
(Intercept)	0.066345	0.029245	2.2686	0.0249 *
lag(log(emp), 1)	1.931319	0.044720	43.1867	< 2.2e-16 ***
lag(log(emp), 2)	-0.937184	0.044536	-21.0431	< 2.2e-16 ***
lag(log(wage), 2)	-0.061661	0.070879	-0.8699	0.3859
lag(log(wage), 3)	0.042590	0.067088	0.6348	0.5266
diff(log(capital), 2)	0.506499	0.058375	8.6767	1.259e-14 ***
diff(log(capital), 3)	-0.306741	0.037957	-8.0813	3.445e-13 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R² da regressão cross-section (B1)

Total Sum of Squares: 253.01

Residual Sum of Squares: 0.085322

R-Squared: 0.99966

Adj. R Squared: 0.99965

F-statistic: 65710.6 on 6 and 133 DF, p-value: < 2.22e-16

Estimadores de Efeitos Aleatórios

- Voltemos novamente ao problema de estimar os coeficientes β_1, \dots, β_k na equação abaixo:

$$y_{i,t} = \alpha_i + \beta_1 x_{1,i,t} + \beta_2 x_{2,i,t} + \dots + \beta_k x_{k,i,t} + \epsilon_{i,t} \quad (\text{A})$$

- Uma das grandes desvantagens de utilizarmos regressão de painel com efeitos fixos é que não podemos incluir no lado direito da equação variáveis explicativas que não variam no tempo
- Quando o termos α_i não são correlacionados com os erros $\epsilon_{i,t}$, nós podemos empregar estimadores de efeitos aleatórios
- Esses estimadores permitem a inclusão de variáveis explicativas que não variem no tempo, o que pode ser muito útil em várias situações
- Vamos reescrever a equação (A) na forma

$$y_{i,t} = \alpha + (\alpha_i - \alpha) + \beta_1 x_{1,i,t} + \beta_2 x_{2,i,t} + \dots + \beta_k x_{k,i,t} + \epsilon_{i,t}$$

- Simplificando, temos:

$$y_{i,t} = \alpha + \beta_1 x_{1,i,t} + \beta_2 x_{2,i,t} + \dots + \beta_k x_{k,i,t} + \alpha_i^* + \epsilon_{i,t} \quad (\text{E})$$

- Para o estimador de efeitos aleatórios, o termo $\alpha_i^* = (\alpha_i - \alpha)$ é considerado uma variável aleatória, com média 0 e variância σ_α^2

Estimadores de Efeitos Aleatórios

- Para o estimador de efeitos aleatórios, o termo α_i^* é considerado uma variável aleatória, com média 0 e variância σ_α^2

$$y_{i,t} = \alpha + \beta_1 x_{1,i,t} + \beta_2 x_{2,i,t} + \dots + \beta_k x_{k,i,t} + \alpha_i^* + \epsilon_{i,t}$$

- No caso dos estimadores de efeitos aleatórios, podemos considerar um erro composto $v_{i,t} = \alpha_i^* + \epsilon_{i,t}$, resultando

$$y_{i,t} = \alpha + \beta_1 x_{1,i,t} + \beta_2 x_{2,i,t} + \dots + \beta_k x_{k,i,t} + v_{i,t} \quad (\text{E1})$$

- O estimador de efeitos aleatórios nada mais é do que um estimador de mínimos quadrados, com base na equação (E1), levando em consideração a estrutura de variância dos erros, devido à composição $v_{i,t} = \alpha_i^* + \epsilon_{i,t}$
- Para isso, rodamos um estimador de mínimos quadrados ordinários na equação

$$(y_{i,t} - \hat{\lambda} \bar{y}_{i,\cdot}) = \alpha^* + \beta_1 (x_{1,i,t} - \hat{\lambda} \bar{x}_{1,i,\cdot}) + \dots + \beta_k (x_{k,i,t} - \hat{\lambda} \bar{x}_{k,i,\cdot}) + v_{i,t}^*$$

- Onde:

$$\hat{\lambda} = 1 - \frac{\sigma_\epsilon}{\sqrt{\sigma_\epsilon^2 + T\sigma_\alpha^2}}$$

Estimadores de Efeitos Aleatórios

- Para isso, rodamos um estimador de mínimos quadrados ordinários na equação

$$(y_{i,t} - \hat{\lambda} \bar{y}_{i,\cdot}) = \alpha^* + \beta_1(x_{1,i,t} - \hat{\lambda} \bar{x}_{1,i,\cdot}) + \dots + \beta_k(x_{k,i,t} - \hat{\lambda} \bar{x}_{k,i,\cdot}) + v_{i,t}^*$$

- Onde:

$$\hat{\lambda} = 1 - \frac{\sigma_\epsilon}{\sqrt{\sigma_\epsilon^2 + T\sigma_\alpha^2}}$$

- Outros tipos de estimadores de efeitos aleatórios estão disponíveis, e todos visam a separar a variabilidade dos termos idiossincráticos (σ_α^2) da variabilidade dos resíduos (σ_ϵ^2)
- Em alguns casos, dependendo da base de dados, é possível que haja problemas numéricos, incorrendo em valores negativos para alguns dos dois termos (σ_α^2 ou σ_ϵ^2)
- Resta agora estudarmos como identificar se devemos utilizar estimador de efeitos fixos ou estimador de efeitos aleatórios
- O teste comumente empregado é o teste de Hausman

> `phtest(fixed, random)`

Efeitos Fixos versus Efeitos Aleatórios

- O teste comumente empregado é o teste de Hausman, para diferenciar entre efeitos fixos e efeitos aleatórios

> `phtest(fixed, random)`

- A diferença básica entre esses dois tipos de modelos é que, para os efeitos aleatórios, assumimos que não existe correlação entre os termos idiossincráticos α_i e os resíduos $\epsilon_{i,t}$; para o estimador de efeitos fixos, podemos ter ou não correlação entre α_i e $\epsilon_{i,t}$
- Podemos testar se existe ou não correlação de forma indireta
- Rodamos o estimador de efeitos fixos (por exemplo, within) e rodamos o estimador de efeitos aleatórios
- Testamos então a diferença, estatisticamente, entre os parâmetros estimados pelos dois estimadores
- Se tivéssemos apenas um coeficiente δ , a estatística de Hausman teria a forma:

$$T = \frac{\hat{\delta}_{FE} - \hat{\delta}_{RE}}{\sqrt{\hat{Var}(\hat{\delta}_{FE}) - \hat{Var}(\hat{\delta}_{RE})}}$$

Efeitos Fixos versus Efeitos Aleatórios

- Na prática, temos mais de coeficientes para testarmos (coeficientes do modelo de regressão de painel)
- Nesse caso, a estatística de Hausman tem uma expressão mais complexa, mas a ideia é a mesma
- A hipótese nula do teste é que os coeficientes são conjuntamente diferentes do modelo de efeitos fixos e de efeitos aleatórios
- Dado que o método de efeitos fixos é mais flexível, caso rejeitemos a hipótese nula, rejeitamos indiretamente o modelo de efeitos aleatórios
- Caso rejeitemos a hipótese nula, mantemos o modelo de efeitos fixos; caso contrário, podemos usar o modelo de efeitos aleatórios
- Além disso, é necessário testar se de fato precisamos de um modelo com termos idiossincráticos α_i
- Para isso, temos um teste específico, para o qual a hipótese nula é $H_0: \sigma_\alpha^2 = 0$; caso rejeitemos a hipótese nula, justifica-se o uso de modelos de efeitos fixos ou aleatórios; caso contrário, podemos usar o pooled OLS

`pFtest(fixed1, pooled1)`

Regressão com Dados de Paineis

- Diante das várias possibilidades de tipos de estimadores, é importante termos uma sequência de procedimentos para empregar na prática
- Procedimento geral para estimação de modelos de painéis:
 - Estime um modelo de efeitos fixos, utilizando estimador within ou de primeiras diferenças
 - Estime um modelo de efeitos aleatórios
 - Use o teste de Hausman; se as estimativas dos coeficientes β_1, \dots, β_k forem significativamente diferentes, o estimador de efeitos fixos será mais apropriado
 - Caso contrário, teste a hipótese nula de que a variância σ_α^2 entre os termos α_i é igual a zero ($H_0: \sigma_\alpha^2 = 0$)
 - Se rejeitamos a hipótese nula $H_0: \sigma_\alpha^2 = 0$, então o estimador de efeitos aleatórios serão o mais apropriado
 - Caso contrário, podemos usar o Pooled MQO

Regressão com Dados de Paineis

- Exercício – para entregar em duas semanas

Com base no arquivo “Analise_de_Regressao_com_Dados_Painel.R”, considere o modelo com dados de painéis, com a fórmula:

```
formula <- log(gsp) ~ log(water) + log(hwy) + log(util) + log(pc) + lag(log(gsp), 1)
              + lag(log(emp), 1) + log(pcap)
```

Questão 1: rode um modelo de efeitos fixos com estimador ‘within’

Questão 2: rode um modelo de efeitos aleatórios, usando o default do plm

Questão 3: compare o estimador de efeitos fixos ao estimador de efeitos aleatórios, usando um teste de Hausman

Questão 4: rode um modelo sem os termos idiossincráticos

Questão 5: teste a necessidade dos termos idiossincráticos, usando teste de hipótese

Questão 6: qual modelo você usaria ao final dos procedimentos?