

UNIVERSIDADE FEDERAL DE PELOTAS - UFPEL
PROGRAMA DE PÓS GRADUAÇÃO EM ORGANIZAÇÕES E MERCADO - PPGOM

Aplicação de Econometria Básica no R

Sérgio Alves Daneris

sergiodianerisalves@gmail.com

ÍNDICE

- 1.Pacote MASS (Base de Dados)
- 2.Analise Descritiva
- 3.Modelo OLS
- 4.Pacote STATS (Modelo OLS)
- 5.Testes
- 6.Resultados Modelo OLS
- 7.Modelo Efeitos Fixos
- 8.Pacote Wooldridge (Base de Dados)
- 9.Pacote PLM (Modelo de Efeitos Fixos)
- Referências

PACOTE MASS (BASE DE DADOS)

Pacote: MASS (Modern Applied Statistics with S).

- Desenvolvido por **William N. Venables e Brian D. Ripley (2002)**.
- O pacote contém mais de **80 conjuntos de dados reais**.
- Base de dados: **birthwt**.

Sobre: A Base de dados traz informações sobre mães e seus recém-nascidos no Baystate Medical Center, em Springfield (Massachusetts, EUA), coletados em 1986. O objetivo principal é **estudar os fatores de risco associados ao baixo peso ao nascer (definido como peso < 2.500 g)**.

low ← Indicador de baixo peso ao nascer (1 = < 2500g, 0 = ≥ 2500g)

age ← Idade da mãe (em anos)

lwt ← Peso da mãe antes da gravidez (em libras)

race ← Raça da mãe (1 = branca, 2 = negra, 3 = outra)

smoke ← 1 = mãe fumou durante a gravidez; 0 = não

ptl ← Número de partos prematuros anteriores

ht ← 1 = mãe tem hipertensão; 0 = não

ui ← 1 = presença de irritabilidade uterina; 0 = não

ftv ← Número de visitas de pré-natal no primeiro trimestre

bwt ← Peso do bebê ao nascer (em gramas)

ANÁLISE DESCRIPTIVA

```
summary(birthwt)
```

```
##      low       age      lwt      race
## Min. :0.0000  Min. :14.00  Min. : 80.0  Min. :1.000
## 1st Qu.:0.0000 1st Qu.:19.00 1st Qu.:110.0 1st Qu.:1.000
## Median :0.0000 Median :23.00 Median :121.0 Median :1.000
## Mean   :0.3122 Mean   :23.24 Mean   :129.8 Mean   :1.847
## 3rd Qu.:1.0000 3rd Qu.:26.00 3rd Qu.:140.0 3rd Qu.:3.000
## Max.   :1.0000 Max.   :45.00 Max.   :250.0 Max.   :3.000
## 
##      smoke      ptl      ht      ui
## Min. :0.0000  Min. :0.0000  Min. :0.00000  Min. :0.0000
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.0000
## Median :0.0000 Median :0.0000  Median :0.00000  Median :0.0000
## Mean   :0.3915 Mean   :0.1958  Mean   :0.06349  Mean   :0.1481
## 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:0.0000
## Max.   :1.0000 Max.   :3.0000  Max.   :1.00000  Max.   :1.0000
## 
##      ftv      bwt
## Min. :0.0000  Min. : 709
## 1st Qu.:0.0000 1st Qu.:2414
## Median :0.0000 Median :2977
## Mean   :0.7937 Mean   :2945
## 3rd Qu.:1.0000 3rd Qu.:3487
## Max.   :6.0000 Max.   :4990
```

```
str(birthwt)
```

```
## 'data.frame': 189 obs. of 10 variables:
## $ low : int 0 0 0 0 0 0 0 0 0 ...
## $ age : int 19 33 20 21 18 21 22 17 29 ...
## $ lwt : int 182 155 105 108 107 124 118 103 123 ...
## $ race : int 2 3 1 1 1 3 1 3 1 1 ...
## $ smoke: int 0 0 1 1 1 0 0 0 1 1 ...
## $ ptl : int 0 0 0 0 0 0 0 0 0 0 ...
## $ ht : int 0 0 0 0 0 0 0 0 0 0 ...
## $ ui : int 1 0 0 1 1 0 0 0 0 0 ...
## $ ftv : int 0 3 1 2 0 0 1 1 1 0 ...
## $ bwt : int 2523 2551 2557 2594 2600 2622 2637 ...
```

```
head(birthwt)
```

	low	age	lwt	race	smoke	ptl	ht	ui	ftv	bwt
## 85	0	19	182	2	0	0	0	1	0	2523
## 86	0	33	155	3	0	0	0	0	3	2551
## 87	0	20	105	1	1	0	0	0	1	2557
## 88	0	21	108	1	1	0	0	1	2	2594
## 89	0	18	107	1	1	0	0	1	0	2600
## 91	0	21	124	3	0	0	0	0	0	2622

MODELO OLS:

OLS (Mínimos Quadrados Ordinários)

É o procedimento de estimação mais fundamental em econometria. O princípio central do OLS é encontrar os parâmetros (coeficientes) de um modelo que minimizem a Soma dos Quadrados dos Resíduos. **(Hayashi,2000)**

Pressupostos:

- 1) Linearidade ← A relação entre a variável dependente e os regressores é linear nos parâmetros.
- 2) Exogeneidade Estrita ← Hayashi afirma que a média esperada do termo de erro , condicional a todos os regressores em todas as observações (passadas, presentes e futuras), é zero.
- 3) Não Multicolineariedade ← Isso significa que nenhum dos regressores pode ser escrito como uma combinação linear perfeita dos outros regressores. Se isso ocorrer, o modelo sofre de multicolinearidade perfeita e o OLS não pode ser calculado.
- 4) Variância de Erro Esférica: Este pressuposto tem duas partes e se refere à natureza da variância dos erros
 - 4.1) Homoscedasticidade: A variância do termo de erro, condicional aos regressores, é constante para todas as observações.
 - 4.2) Ausência de Correlação: Os termos de erro não são correlacionados entre observações distintas. Em modelos de séries temporais, isso significa que não há correlação serial

MODELO OLS:

Fórmula do modelo de Regressão Linear Múltipla

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \epsilon_i$$

- 1) **Y_i ← Variável Dependente** = É a variável que você está tentando entender ou prever
- 2) **X_i ← Variável Independente** = São as variáveis que você está usando para explicar o Y, ou seja, são os "preditores" ou "fatores".
- 2) **β ← Coeficientes** = São os parâmetros do modelo a serem estimados.
- 3) **ui ← Termo de Erro** = Também chamado de "resíduo". Ele representa tudo o que afeta o Y mas que não foi incluído nas suas variáveis X

PACOTE STATS

Este é um pacote R básico que inclui funções para operações estatísticas básicas, modelos lineares e testes de hipóteses.

Comandos:

Modelos de Regressão:

lm() → regressão linear OLS (Ordinary Least Squares)

glm() → modelos lineares generalizados (como regressão logística, Poisson, binomial)

nls() → regressão não linear

Testes de Estátisticos:

t.test() → teste t para média

anova() → análise de variância

chisq.test() → teste qui-quadrado

cor.test() → teste de correlação

Funções de resumo e agregação

summary(), aggregate(), fitted(), residuals(), predict()

TESTES:

1) Normalidade dos resíduos: Shapiro-Wilk

p-value > 0.05 → resíduos seguem distribuição normal.
p-value < 0.05 → (violação da normalidade.)

Ex: `shapiro.test(residuals(modelo1))`

2) Homocedasticidade: teste de Breusch-Pagan (pacote lmtest)

p-value > 0.05 → variância constante (OK).
p-value < 0.05 → heterocedasticidade (problema).

Ex : `bptest(modelo1)`

Se houver heterocedasticidade, usa-se erros robustos: (pacote sandwich)

"HC0" White (clássico) ← básico
"HC1" White ajustado ← padrão em econometria
"HC2" ajusta pelo leverage ← amostras médias
"HC3" ← amostras pequenas
"HC4"- "HC5" versões mais robustas ← amostras muito pequenas

Ex: `coeftest(modelo1, vcov = vcovHC(modelo1, type = "HC1"))`

RESULTADOS MODELO OLS

```

modelo1 <- lm(low ~ smoke + ht, data = birthwt)

summary(modelo1)

##
## Call:
## lm(formula = low ~ smoke + ht, data = birthwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.6716 -0.2348 -0.2348  0.6139  0.7652
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.23480   0.04328  5.425 1.79e-07 ***
## smoke       0.15132   0.06790  2.228   0.027 *
## ht          0.28549   0.13592  2.100   0.037 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4556 on 186 degrees of freedom
## Multiple R-squared:  0.04862,    Adjusted R-squared:  0.03839
## F-statistic: 4.752 on 2 and 186 DF,  p-value: 0.009706

```

Interpretação:

Smoke (0.151): Fumar aumenta a probabilidade de baixo peso em 15.1 pontos percentuais

HT (0.285): Hipertensão aumenta a probabilidade de baixo peso em 28.5 pontos percentuais

Significância Estatística: Ambas variáveis são significativas ao nível de 5% ($p < 0.05$)

F-statistic = 4.752 (p = 0.0097) ← O modelo é significativo ($p\text{-value} = 0.0097 < 0.05$)

MODELO EFEITO FIXO:

O modelo de efeitos fixos (FE) é usado em dados em painel para controlar fatores não observados que:

São específicos de cada unidade (indivíduo, estado, empresa) e não variam ao longo do tempo.

$$y_{it} = \alpha_i + \beta X_{it} + u_{it}$$

y_{it} <- Variável dependente = É o valor do resultado de interesse, para o indivíduo i no tempo t

X_{it} <- Variável explicativa (ou conjunto de variáveis) = Representa os fatores que variam no tempo e entre indivíduos e que afetam Y_{it}

β <- Coeficiente de regressão = Mede o efeito médio das variáveis explicativas X_{it} sobre y_{it}

α_i <- Efeito fixo individual = Capta todas as características do indivíduo i que não mudam ao longo do tempo e que afetam Y_{it}

u_{it} <- Termo de erro = varia tanto entre indivíduos quanto ao longo do tempo, representando choques aleatórios ou fatores não observados que mudam no tempo

PACOTE WOOLDRIDGE (BASE DE DADOS)

Pacote: **Wooldridge**

O pacote wooldridge contém as **bases de dados reais e simuladas** utilizadas nos **exemplos e exercícios do livro do Wooldridge**. Ele **não é** um pacote de funções econômétricas, **e sim um repositório de dados para praticar regressões, modelos de painel, probit, logit, etc.**

Base de dados: **crime4** (C. Cornwell and W. Trumball (1994))

O objetivo principal é analisar como a taxa de criminalidade (crmrte) em 90 condados da Carolina do Norte (**630 observações / 7 anos = 90 condados**), durante os anos 1981-1987, é afetada por vários fatores de "custos" do crime.

- 1) install.packages('wooldridge')**
- 2) library(wooldridge)**
- 3) data(crime4)**

PACOTE WOOLDRIDGE (BASE DE DADOS)

county: identificador do condado

year: ano, de 81 a 87

crmrte: crimes cometidos por pessoa

prbarr: 'probabilidade' de prisão

prbconv: 'probabilidade' de condenação

prbpris: 'probabilidade' de sentença de prisão

avgsen: sentença média, em dias

polpc: polícia per capita

density: pessoas por milha quadrada

taxpc: receita de impostos per capita

west: =1 se no oeste da Carolina do Norte (N.C.)

central: =1 se no centro da Carolina do Norte (N.C.)

urban: =1 se em SMSA (Área Estatística Metropolitana)

pctmin80: percentual de minorias, 1980

wcon: salário semanal, construção

.....

PACOTE PLM (MODELO DE EFEITOS FIXOS)

Pacotes para modelos econométricos

- **PLM (Panel Linear Models)** : Estimação de modelos econométricos de dados em painel (**dados com dimensão tempo e indivíduo**)

Principais Metodos de Regressão: **Pooled OLS, Fixed Effects (Efeitos Fixos), Random Effects (Efeitos Aleatórios), Between, First Differences.**

Tipos de dados usados: **Dados em painel.**

pdata.frame() <- Transforma a base comum em painel (define id e tempo)

Ex: pdata.frame(base, index = c("id", "ano"))

plm() <- Estima o modelo de regressão para dados em painel

Ex: plm(y ~ x1 + x2, data = painel, model = "within")

phtest() <- Teste de Hausman para escolher entre efeitos fixos e aleatórios

Ex: phtest(modelo_fixo, modelo_aleatorios)

PACOTE PLM (MODELO DE EFEITOS FIXOS)

```
summary(modelo_fx)

## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgson +
##       lpolpc + ldensity + lpctymle + lwcon + lwtrd + lwser, data = crime4
##       model = "within", index = c("county", "year"))
##
## Balanced Panel: n = 90, T = 7, N = 630
##
## Residuals:
##      Min.    1st Qu.     Median    3rd Qu.     Max.
## -0.65852798 -0.07651194 -0.00089633  0.07839218  0.59157023
##
## Coefficients:
##             Estimate Std. Error t-value Pr(>|t|)
## lprbarr   -0.3951986  0.0334312 -11.8213 < 2.2e-16 ***
## lprbconv  -0.3111145  0.0217581 -14.2988 < 2.2e-16 ***
## lprbpris  -0.2146370  0.0333242  -6.4409 2.668e-10 ***
## lavgson    0.0322195  0.0258941   1.2443  0.21395
## lpolpc    0.4230518  0.0274764  15.3969 < 2.2e-16 ***
## ldensity   0.1448029  0.2618459   0.5530  0.58049
## lpctymle   0.3954462  0.2217544   1.7833  0.07512 .
## lwcon     -0.0535321  0.0390972  -1.3692  0.17151
## lwtrd     -0.0574097  0.0415584  -1.3814  0.16773
## lwser     -0.0043194  0.0198641  -0.2174  0.82794
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:  17.991
## Residual Sum of Squares: 11.14
## R-Squared: 0.38079
## Adj. R-Squared: 0.26513
## F-statistic: 32.5929 on 10 and 530 DF, p-value: < 2.22e-16
```

Interpretação

Variável dependente: lcrmrte (log da taxa de crimes)

Variáveis Significativas ($p < 0.05$):

Iprbarr (-0.395) - Probabilidade de prisão

Aumento de 1% na probabilidade de prisão reduz taxa de crimes em $\approx 0.395\%$

Iprbpris (-0.215) - Probabilidade de prisão se condenado

Aumento de 1% reduz crimes em $\approx 0.215\%$

Ipolpc (0.423) - Policiais per capita

Aumento de 1% em policiais aumenta crimes em $\approx 0.423\%$

Variáveis Não Significativas:

lavgson (sentença média): efeito positivo não significativo

ldensity (densidade populacional): não significativo

lwcon, lwtrd, lwser (variáveis de salários): nenhuma significativa

REFERÊNCIAS

- HAYASHI, Fumio. *Econometrics*. Princeton: Princeton University Press, 2000.
- WOOLDRIDGE, Jeffrey M. *Econometric Analysis of Cross Section and Panel Data*. 2. ed. Cambridge, MA: MIT Press, 2010.
- Croissant Y, Millo G (2008). "Panel Data Econometrics in R: The *plm* Package." *Journal of Statistical Software*, 27(2), 1–43.
- R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*, Fourth edition. Springer, New York.