Métodos Quantitativos

Aula 10. Regressão linear, parte 2

Pedro H. G. Ferreira de Souza

pedro.ferreira@ipea.gov.br

Mestrado Profissional em Políticas Públicas e Desenvolvimento

Instituto de Pesquisa Econômica Aplicada (Ipea)

05 dez. 2022

Recapitulação

Introdução à regressão multivariada

Preparação para a aula

Estimativas de ponto

Inferência

Próxima aula

Regressão com uma variável independente

Estimação por MQO

$$y = \alpha + \beta x + \epsilon \quad \text{ou seja,} \quad E(y \mid x) = \alpha + \beta x$$

Pressupostos

Necessários para estimativas não enviesadas e eficientes.

Interpretação de coeficientes, testes de hipóteses e ICs

- Associação vs. causalidade
- Quantificando a incerteza

Recapitulação

Introdução à regressão multivariada

Preparação para a aula

Estimativas de ponto

Inferência

Próxima aula

O que é o erro?

O que é ϵ?

O erro ϵ contém tudo que causa Y mas não foi incluído no modelo.

- Outras variáveis que observamos ou poderíamos observar
- Choques aleatórios

Exogeneidade estrita

Para estimarmos o parâmetro β sem viés, X e ϵ têm que ser **não correlacionados**

$$E(\epsilon) = E(\epsilon \mid X) = 0 \quad \text{ou seja} \quad cov(X, \epsilon) = 0 \quad \text{e} \quad cor(X, \epsilon) = 0$$

Exogeneidade estrita

Aleatorização do tratamento

Forma ideal de garantir exogeneidade e isolar o efeito causal, contra viés de variável omitida. Mas obviamente nem sempre é possível fazer um experimento...

Regressão múltipla

Em estudos observacionais, tentamos **atenuar** o problema acrescentando variáveis ao modelo de regressão, isto é, “controles”.

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

$$E(y \mid x_1, x_2, \ldots, x_n) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Huntington-Klein 2022, fig 8.4

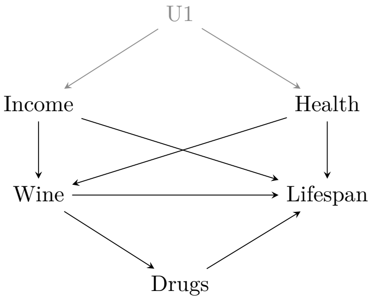
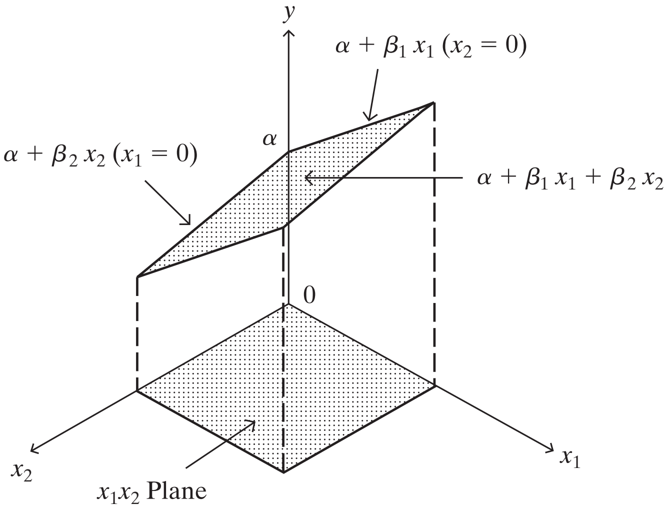


Figure 8.4: The Effect of Wine on Lifespan

Agresti 2018, fig 11.1



Recapitulação

Introdução à regressão multivariada

Preparação para a aula

Estimativas de ponto

Inferência

Próxima aula

Objetivos de hoje

1. Estimação de regressão múltipla por MQO
   - Pressupostos
   - Como estimar no R
   - Viés de variável omitida
2. Interpretação de resultados
   - Coeficientes
   - Erros padrão
   - Testes de hipóteses e ICs
3. Ajuste e seleção de modelos
   - Erro padrão da regressão, r^2 e r^2 ajustado
   - Teste F

Pacotes

```
# Pacotes de uso geral
library(tidyverse)
library(broom)
# Pacote para estatisticas descritivas
library(summarytools)
# Pacote para visualizacao de resultados de modelos
library(modelsummary)
# Pacotes com bases de dados
library(HistData)
library(causaldata)
library(AER)
```

Bases de dados: exemplo #1

```
ex1_dados <-
  GaltonFamilies %>%
    mutate(filhos = 2.54 * childHeight,
           mae = 2.54 * mother,
           pai = 2.54 * father,
           num_irmaos = children - 1) %>%
      select(filhos, mae, pai, num_irmaos)
```

```
glimpse(ex1_dados)
```

```
## Rows: 934
## Columns: 4
## $ filhos     <dbl> 185.928, 175.768, 175.260, 175.260, 186.690, 184.1
## $ mae        <dbl> 170.18, 170.18, 170.18, 170.18, 168.91, 168.91, 16
## $ pai        <dbl> 199.39, 199.39, 199.39, 199.39, 191.77, 191.77, 19
## $ num_irmaos <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 1, 1, 4, 4, 4, 4, 4, 5, 5,
```

Bases de dados: exemplo #2

```
ex2_dados <-
  restaurant_inspections %>%
    rename(nota_inspecao = inspection_score,
           ano = Year,
           num_estab = NumberofLocations) %>%
      select(nota_inspecao, ano, num_estab)
```

```
glimpse(ex2_dados)
```

```
## Rows: 27,178
## Columns: 3
## $ nota_inspecao <dbl> 94, 86, 80, 96, 83, 95, 94, 100, 92, 91, 98, 94
## $ ano           <dbl> 2017, 2015, 2016, 2003, 2017, 2008, 2017, 2005,
## $ num_estab     <dbl> 9, 66, 79, 86, 53, 89, 28, 37, 109, 59, 45, 32,
```

Bases de dados: exemplo #3

```
data(CASchools)
ex3_dados <-
  CASchools %>%
    mutate(pc_aluno = computer / students,
           gasto_aluno = expenditure / 10^3) %>%
      rename(nota_ler = read, pct_out_idioma = english,
             pct_ajuda_alm = lunch,renda_dist = income) %>%
        select(nota_ler, starts_with('pct'),
               ends_with('aluno'), renda_dist)
```

```
glimpse(ex3_dados)
```

```
## Rows: 420
## Columns: 6
## $ nota_ler       <dbl> 691.6, 660.5, 636.3, 651.9, 641.8, 605.7, 604.
## $ pct_ajuda_alm  <dbl> 2.0408, 47.9167, 76.3226, 77.0492, 78.4270, 86
## $ pct_out_idioma <dbl> 0.000000, 4.583333, 30.000002, 0.000000, 13.85
```

Estimação

O modelo

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$$

Estimação

O modelo

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$$

Pressupostos

1. Linearidade
2. Exogeneidade estrita: $E(\epsilon \mid x_1, \ldots, x_k) = 0$
3. Sem colinearidade perfeita
4. Amostragem aleatória: erros não correlacionados, $cov(\epsilon_i, \epsilon_j) = 0$
5. Homoscedasticidade: $var(\epsilon \mid x_1, \ldots, x_k) = \sigma^2_\epsilon$

Estimação

O modelo

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$$

Pressupostos

1. Linearidade
2. Exogeneidade estrita: $E(\epsilon \mid x_1, \ldots, x_k) = 0$
3. Sem colinearidade perfeita
4. Amostragem aleatória: erros não correlacionados, $cov(\epsilon_i, \epsilon_j) = 0$
5. Homoscedasticidade: $var(\epsilon \mid x_1, \ldots, x_k) = \sigma_\epsilon^2$

Estimação

Minimizar **soma dos quadrados dos erros**: $\sum \epsilon_i^2 = \sum (y_i - \alpha - \beta_1 x_1 - \ldots - \beta_k x_k)$

Interpretação das estimativas de ponto

Regressão bivariada

Coeficiente de X equivale à inclinação da reta $a + bx$; efeito linear de x sobre y sem levar em conta outras variáveis.

Interpretação das estimativas de ponto

Regressão bivariada

Coeficiente de X equivale à inclinação da reta $a + bx$; efeito linear de x sobre y sem levar em conta outras variáveis.

Regressão múltipla

Coeficiente de x_k descreve o **efeito parcial** de x_k sobre y, **controlando** - isto é, mantendo constantes - as demais variáveis do modelo.

- Para cada x_k, b_k **indica a variação média em y associada a uma mudança de uma unidade em x_k**.

Efeito é causal? Depende do desenho de pesquisa, dos dados e dos pressupostos. Com dados observacionais, sempre há o risco de variável omitida.

O que acontece quando adicionamos variáveis?

Suponha que estimamos o modelo $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

Os coeficientes são estimados usando a variância "única" de cada variável. Podemos obter a mesma estimativa b_1 para β_1 de duas maneiras:

1. Estimando o modelo multivariado $y = a + b_1 x_1 + b_2 x_2 + e$

O que acontece quando adicionamos variáveis?

Suponha que estimamos o modelo $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

Os coeficientes são estimados usando a variância "única" de cada variável. Podemos obter a mesma estimativa b_1 para β_1 de duas maneiras:

1. Estimando o modelo multivariado $y = a + b_1 x_1 + b_2 x_2 + e$
2. Estimando o modelo $w = c + b_1 z + u$:
   2.1 Estimamos a regressão bivariada $y = c + dx_2 + u$ e criamos uma nova variável com o resíduo, isto é, $w = y - (c + dx_2)$
   2.2 Estimamos a regressão bivariada $x_1 = q + mx_2 + v$ e criamos mais uma variável com o resíduo $z = x_1 - (q + mx_2)$
   2.3 Finalmente, estimamos a regressão bivariada $w = h + b_1 z + l$

b_1 terá a mesma estimativa de ponto e o mesmo erro padrão, seja estimado por (1) ou por (2)

O que acontece quando adicionamos variáveis?

E se estimarmos os modelos...

$$y = \hat{a} + \hat{b}x_1 + \hat{e} \qquad\qquad y = \tilde{a} + \tilde{b_1}x_1 + \tilde{b_2}x_2 + \tilde{e}$$

Qual a relação entre $\hat{b}$ e $\tilde{b_1}$?

O que acontece quando adicionamos variáveis?

E se estimarmos os modelos...

$$y = \hat{a} + \hat{b}x_1 + \hat{e} \qquad\qquad y = \tilde{a} + \tilde{b_1}x_1 + \tilde{b_2}x_2 + \tilde{e}$$

Qual a relação entre $\hat{b}$ e $\tilde{b_1}$?

- Os dois serão idênticos **se e apenas se** o efeito parcial de x_2 sobre y for zero ($\hat{b_2} = 0$) **ou** não houver correlação entre x_1 e x_2 na amostra.

Caso contrário, $\hat{b} \neq \tilde{b_1}$, pois o modelo bivariado não controla pela influência de x_2 sobre y.

- O coeficiente $\hat{b}$ incorpora também o efeito da **variável omitida** x_2.

Viés de variável omitida

Suponha que o modelo populacional correto é $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$.

Se estimarmos somente $y = \hat{a} + \hat{b} x_1 + \hat{e}$, qual o efeito da omissão de x_2 sobre $\hat{b}$?

Viés de variável omitida

Suponha que o modelo populacional correto é $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$.

Se estimarmos somente $y = \hat{a} + \hat{b}x_1 + \hat{e}$, qual o efeito da omissão de x_2 sobre $\hat{b}$?

Como mostra Wooldridge (p. 87-89), o valor esperado $y = \hat{a} + \hat{b}x_1 + \hat{e}$ será:

$$E(\hat{b}) = \beta_1 + \beta_2 \delta_1$$

… em que δ_1 é o coeficiente da regressão $x_2 = \tilde{\delta_0} + \tilde{\delta_1} x_1$

	Corr(x_1,x_2) > 0	**Corr(x_1,x_2) < 0**
$\beta_2 > 0$	positive bias	negative bias
$\beta_2 < 0$	negative bias	positive bias

Exemplo #1: altura de filhos e pais

Regressões com y = altura dos filhos

```
ex1_1 <- lm(filhos ~ mae,
            data = ex1_dados)
ex1_2 <- lm(filhos ~ mae + pai,
            data = ex1_dados)
ex1_3 <- lm(filhos ~ mae + pai + num_irmaos,
            data = ex1_dados)
```

Cada objeto `ex1_1`, `ex1_2`, `ex1_3` contém os resultados das regressões.

Tentem digitar `print(ex1_1)` e `summary(ex1_1)`, por exemplo.

Exemplo #1: altura de filhos e pais

```
ex1_1$coefficients
```

```
## (Intercept)         mae
## 118.3312440   0.3145428
```

```
ex1_2$coefficients
```

```
## (Intercept)         mae         pai
##  57.5139305   0.2905100   0.3682823
```

```
ex1_3$coefficients
```

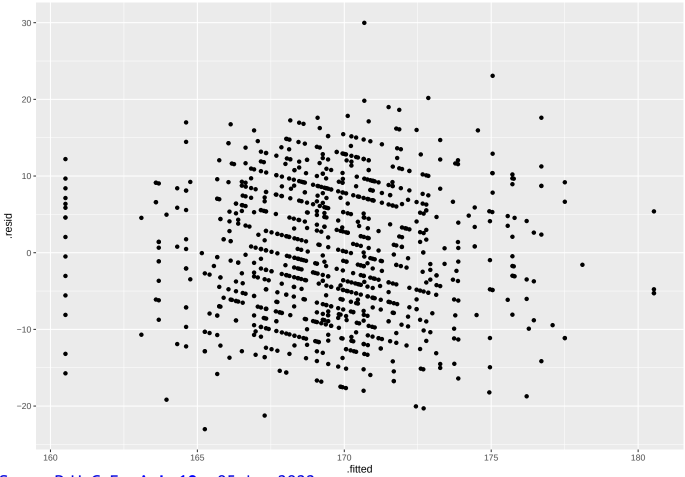
```
## (Intercept)         mae         pai  num_irmaos
##  62.6757224   0.2872942   0.3501145  -0.2794492
```

```
summary(ex1_3) # output no proximo slide
```

```
##
## Call:
## lm(formula = filhos ~ mae + pai + num_irmaos, data = ex1_dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.0144  -6.6828  -0.6204   7.0352  29.9783
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 62.67572   10.96016   5.719 1.45e-08 ***
## mae          0.28729    0.04838   5.938 4.06e-09 ***
## pai          0.35011    0.04525   7.737 2.64e-14 ***
## num_irmaos  -0.27945    0.10417  -2.683  0.00743 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.58 on 930 degrees of freedom
## Multiple R-squared:  0.1121, Adjusted R-squared:  0.1092
## F-statistic: 39.14 on 3 and 930 DF,  p-value: < 2.2e-16
```

Exemplo #1: altura de filhos e pais

```
ex1_3_yhat <- augment(ex1_3)
qplot(.fitted, .resid, data = ex1_3_yhat, geom = 'point')
```



Exemplo #2: restaurantes no Alasca

Estatísticas descritivas

```
descr(ex2_dados, stats = c('mean','sd','min','med','max','n.valid'))
```

```
## Descriptive Statistics
## ex2_dados
## N: 27178
##
##                      ano   nota_inspecao   num_estab
## ------------- ---------- --------------- -----------
##          Mean    2010.34           93.64       64.77
##       Std.Dev       5.95            6.26       84.27
##           Min    2000.00           66.00        1.00
##        Median    2009.00           95.00       41.00
##           Max    2019.00          100.00      646.00
##       N.Valid   27178.00        27178.00    27178.00
```

Exemplo #2: restaurantes no Alasca

Estimando as regressões

```
ex2_1 <- lm(nota_inspecao ~ num_estab,
            data = ex2_dados)
ex2_2 <- lm(nota_inspecao ~ num_estab + ano,
            data = ex2_dados)
```

Exemplo #2: restaurantes no Alasca

Estimando as regressões

```
ex2_1 <- lm(nota_inspecao ~ num_estab,
            data = ex2_dados)
ex2_2 <- lm(nota_inspecao ~ num_estab + ano,
            data = ex2_dados)
```

Mostrando os resultados *(output no próximo slide)*

```
ex2 <- list(ex2_1, ex2_2)
msummary(ex2,
         stars = TRUE,
         output = "markdown",
         gof_omit = c('BIC|AIC|Log'))
```

Exemplo #2: restaurantes no Alasca

	Model 1	Model 2
(Intercept)	94.866***	225.333***
	(0.046)	(12.411)
num_estab	-0.019***	-0.019***
	(0.0004)	(0.0004)
ano		-0.065***
		(0.006)
Num.Obs.	27178	27178
R2	0.065	0.068
R2 Adj.	0.065	0.068
F	1876.705	997.386
RMSE	6.05	6.04

Note: ^^ + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Exemplo #2: restaurantes no Alasca

```
ex2_3a <- lm(nota_inspecao ~ ano, data = ex2_dados)
ex2_3b <- lm(num_estab ~ ano, data = ex2_dados)
ex2_dados$resid_nota <- resid(ex2_3a)
ex2_dados$resid_estab <- resid(ex2_3b)
ex2_3 <- lm(resid_nota ~ resid_estab, data = ex2_dados)
ex2 <- append(ex2, list(ex2_3))
msummary(ex2, stars = TRUE, output = "markdown",
         gof_omit = c('BIC|AIC|Log|F'))
```

Output no próximo slide.

Exemplo #2: restaurantes no Alasca

	Model 1	Model 2	Model 3
(Intercept)	94.866***	225.333***	-5e-15
	(0.046)	(12.411)	(0.037)
num_estab	-0.019***	-0.019***	
	(0.0004)	(0.0004)	
ano		-0.065***	
		(0.006)	
resid_estab			-0.019***
			(0.0004)
Num.Obs.	27178	27178	27178
R2	0.065	0.068	0.067
R2 Adj.	0.065	0.068	0.067
RMSE	6.05	6.04	6.04

Note: + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Pressupostos para inferência

1. Linearidade: modelo populacional é $E(y \mid x_1, \ldots, x_k) = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k$
2. Exogeneidade estrita: $E(\epsilon \mid x_1, \ldots, x_k) = 0$
3. Sem multicolinearidade: independência linear entre os $x_1, \ldots, x_k$
4. Erros esféricos: sem autocorrelação e com homoscedasticidade: $E(\epsilon_i^2 \mid x_1, \ldots, x_k) = \sigma_2$
5. Erros com distribuição normal condicional aos regressores: $\epsilon \mid x_1, \ldots, x_k \sim N(0, \sigma^2 I_n)$

- Ausência de viés: 1 + 2;
- Modelo estimável: 1 + 2 + 3
- Inferência e eficiência: 1 + 2 + 3 + 4 + 5
- Pressuposto #5 não é necessário em amostras grandes

Distribuição amostral e variância

As fórmulas são mais complicadas, mas assim como antes:

- Calculamos a variância de cada coeficiente, que depende de σ, das variâncias e covariâncias dos x e do tamanho da amostra n.
- A distribuição amostral dos coeficientes é (aproximadamente) normal
- Podemos construir ICs para cada coeficiente usando a distribuição t
- Podemos fazer testes de hipótese para cada coeficiente

O que muda?

- Podemos fazer teste de hipótese global da regressão e/ou testar vários coeficientes de uma vez só

(Fórmulas omitidas, mas fáceis de achar no Google ou nos livros).

Testes de hipóteses (i)

Teste global da regressão

Tipicamente, os *softwares* estatísticos reportam a **estatística F** e o **p-valor** para o teste global do tipo:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0 \qquad H_1 : \text{pelo menos um } \beta_k \neq 0$$

A interpretação do p-valor é como nos testes parciais.

Testes de hipóteses (i)

Teste global da regressão

Tipicamente, os *softwares* estatísticos reportam a **estatística F** e o **p-valor** para o teste global do tipo:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0 \qquad\qquad H_1 : \text{pelo menos um } \beta_k \neq 0$$

A interpretação do p-valor é como nos testes parciais.

Outros testes coletivos

É possível testar qualquer conjunto de restrições lineares com o teste F. Por exemplo, podemos ter $H_0 : \beta_1 = 1, \beta_2 = 10, \beta_3 = \ldots = \beta_k = 0$ etc.

Para mais detalhes, ver *Wooldrige, Introductory Econometrics - A Modern Approach, 2a edição*, p. 139-150.

Testes de hipóteses (ii)

Para coeficientes individuais, os testes de hipóteses são de novo muito parecidos com os de regressões bivariadas.

Se $H_0 : \beta_k = c$, então a **estatística t** é:

$$t = \frac{b - c}{se(b)} \quad \text{com} \quad df = n - p = n - k + 1$$

Pode-se usar o teste t também para comparações entre dois coeficientes, por exemplo: $H_0 : \beta_1 = \beta_2$.

Intervalos de confiança

De novo, assim como em modelos bivariados, podemos construir intervalos de confiança para cada coeficiente parcial, com a mesma interpretação de antes. Os intervalos vão ser construídos por:

$$b \pm t \cdot se(b) \qquad \text{em que } t \text{ possui } df = n - p = n - k + 1$$

Uma diferença é que, em modelo multivariado, os IC e as estatísticas t também estão "controlando" pelas demais variáveis do modelo, isto é, na prática, são calculados somente com a variância "única" de cada variável.

Exemplo #3: escolas na Califórnia

Estatísticas descritivas

```
ex3_dados %>% descr(stats = c('mean', 'sd', 'med', 'min','max'),
                    transpose = TRUE)
```

```
## Descriptive Statistics
## ex3_dados
## N: 420
##
##                           Mean   Std.Dev   Median      Min      Max
## -------------------- -------- --------- -------- -------- ------
--
##          gasto_aluno     5.31      0.63     5.21     3.93     7.71
##             nota_ler   654.97     20.11   655.75   604.50   704.00
##             pc_aluno     0.14      0.06     0.13     0.00     0.42
##        pct_ajuda_alm    44.71     27.12    41.75     0.00   100.00
##       pct_out_idioma    15.77     18.29     8.78     0.00    85.54
##           renda_dist    15.32      7.23    13.73     5.34    55.33
```

Exemplo #3: escolas na Califórnia

Modelos

```
ex3_1 <- lm(nota_ler ~ pc_aluno, data = ex3_dados)
ex3_2 <- lm(nota_ler ~ pc_aluno + pct_out_idioma +
            pct_ajuda_alm + gasto_aluno, data = ex3_dados)
ex3_3 <- lm(nota_ler ~ pc_aluno + pct_out_idioma +
            pct_ajuda_alm + gasto_aluno + renda_dist,
            data = ex3_dados)
```

Exemplo #3: escolas na Califórnia

Modelos

```
ex3_1 <- lm(nota_ler ~ pc_aluno, data = ex3_dados)
ex3_2 <- lm(nota_ler ~ pc_aluno + pct_out_idioma +
            pct_ajuda_alm + gasto_aluno, data = ex3_dados)
ex3_3 <- lm(nota_ler ~ pc_aluno + pct_out_idioma +
            pct_ajuda_alm + gasto_aluno + renda_dist,
            data = ex3_dados)
```

Mostrando os resultados *(output no próximo slide)*

```
ex3 <- list(ex3_1, ex3_2, ex3_3)
msummary(ex3, output = "markdown", stars = TRUE,
         gof_omit = c('BIC|AIC|Log|Num. Obs|F'))
```

	Model 1	Model 2	Model 3
(Intercept)	643.140***	656.101***	652.607***
	(2.189)	(3.624)	(3.515)
pc_aluno	87.036***	12.757+	12.179+
	(14.530)	(6.896)	(6.603)
pct_out_idioma		-0.205***	-0.260***
		(0.030)	(0.030)
pct_ajuda_alm		-0.548***	-0.430***
		(0.020)	(0.027)
gasto_aluno		4.684***	2.995***
		(0.685)	(0.709)
renda_dist			0.531***
			(0.085)
Num.Obs.	420	420	420
R2	0.079	0.822	0.838
R2 Adj.	0.077	0.821	0.836
RMSE	19.27	8.46	8.09

Note: ^^ + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Exemplo #3: escolas na Califórnia

```
ex3_coeftest <- coeftest(ex3_3)
round(ex3_coeftest, digits = 4)
```

```
##
## t test of coefficients:
##
##                 Estimate Std. Error  t value Pr(>|t|)
## (Intercept)     652.6067     3.5146 185.6863   <2e-16 ***
## pc_aluno         12.1795     6.6025   1.8447   0.0658 .
## pct_out_idioma  -0.2596     0.0304  -8.5331   <2e-16 ***
## pct_ajuda_alm   -0.4303     0.0271 -15.8534   <2e-16 ***
## gasto_aluno       2.9947     0.7095   4.2209   <2e-16 ***
## renda_dist        0.5306     0.0852   6.2287   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

Exemplo #3: escolas na Califórnia

```
# Teste de significancia global
summary(ex3_3)
```

```
F-statistic: 427.2 on 5 and 414 DF,  p-value: < 2.2e-16
```

Exemplo #3: escolas na Califórnia

```
# Teste de significancia global
summary(ex3_3)
```

```
F-statistic: 427.2 on 5 and 414 DF,  p-value: < 2.2e-16
```

```
# Teste para coeficientes especificos
linearHypothesis(ex3_3, c('pct_out_idioma = pct_ajuda_alm'))
```

```
## Linear hypothesis test
##
## Hypothesis:
## pct_out_idioma - pct_ajuda_alm = 0
##
## Model 1: restricted model
## Model 2: nota_ler ~ pc_aluno + pct_out_idioma + pct_ajuda_alm + gasto
##     renda_dist
##
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1    415 28217
## 2    414 27503  1    713.58 10.741 0.001136 **
```

Comparando modelos

F-statistic: 35.88 on 1 and 418 DF, p-value: 4.531e-09Mais uma vez, queremos saber também quão o grau de aderência do modelo aos dados. Afinal, estamos impondo uma relação linear, reduzindo a complexidade do mundo para obter os padrões que julgamos mais importantes.

Além disso, nossas teorias nem sempre são precisas sobre quais variáveis devemos incluir no modelo. O que fazer para decidir qual nossa melhor opção?

- Quanto mais variáveis incluímos, maior a variância dos estimadores, sem necessariamente melhor o ajuste.
- "Tudo mais constante", modelos mais parcimoniosos são sempre preferíveis.

Estatísticas de ajuste (i)

Erro padrão da regressão σ_ϵ

Igual ao que vimos antes: quantifica, na escala de y, o tamanho médio dos erros. A fórmula é:

$$s_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-p}} = \sqrt{\frac{SSE}{n-p}}$$

Em que n é o tamanho da amostra e $p = k + 1$, isto é, o número de parâmetros.

Estatísticas de ajuste (ii)

r^2 ou coeficiente de determinação

É a proporção da variância de y explicada por $x_1, x_2, ...x_k$:

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Em regressões múltiplas, o r^2 equivale ao quadrado da correlação entre y e o valor predito $\hat{y} = \hat{a} + \hat{b}_1 x_1 + ...\hat{b}_k x_k$

Mais uma vez, o r^2 pode variar entre zero (quando não há correlação entre y e o conjunto dos regressores) e 1.

Estatísticas de ajuste (iii)

r^2 ajustado

O problema do r^2 é que ele nunca diminui quando acrescentamos novas variáveis, mesmo que elas sejam irrelevantes para y. Por isso, o chamado r^2 ajustado incorpora uma penalidade ao r^2 tradicional, podendo diminuir com a adição de novos regressores:

$$r^2 \text{ ajustado} = 1 - \frac{SSR/(n-p)}{SST/(n-1)} = 1 - (1-r^2)\frac{n-1}{n-p}$$

Mesmo uma variável relevante para y pode ter efeito nulo sobre o r^2 se ela for altamente correlacionada com alguma outra variável já incluída no modelo. Nesse caso, é possível até que o r^2 ajustado caia.

- Observe que $r^2 \text{ ajustado} \leq r^2$

Exemplo #1: alturas (cont.)

```
ex1_glance <- bind_rows(glance(ex1_1),
                        glance(ex1_2),
                        glance(ex1_3))
ex1_compara <- bind_cols(mod = c('ex1_1','ex1_2', 'ex1_3'),
                         ex1_glance)
ex1_compara %>% select(mod, r.squared, adj.r.squared, sigma)
```

```
## # A tibble: 3 x 4
##   mod   r.squared adj.r.squared sigma
##   <chr>     <dbl>         <dbl> <dbl>
## 1 ex1_1    0.0405        0.0395  8.91
## 2 ex1_2    0.105         0.103   8.61
## 3 ex1_3    0.112         0.109   8.58
```

Exemplo #2: restaurantes (cont.)

```
ex2_glance <- bind_rows(glance(ex2_1),
                        glance(ex2_2))
ex2_compara <- bind_cols(mod = c('ex2_1','ex2_2'),
                         ex2_glance)
ex2_compara %>% select(mod, r.squared, adj.r.squared, sigma)
```

```
## # A tibble: 2 x 4
##   mod   r.squared adj.r.squared sigma
##   <chr>     <dbl>         <dbl> <dbl>
## 1 ex2_1    0.0646        0.0646  6.05
## 2 ex2_2    0.0684        0.0683  6.04
```

Exemplo #3: escolas (cont.)

```
ex3_glance <- bind_rows(glance(ex3_1),
                        glance(ex3_2),
                        glance(ex3_3))
ex3_compara <- bind_cols(mod = c('ex3_1','ex3_2', 'ex3_3'),
                         ex3_glance)
ex3_compara %>% select(mod, r.squared, adj.r.squared, sigma)
```

```
## # A tibble: 3 x 4
##   mod   r.squared adj.r.squared sigma
##   <chr>     <dbl>         <dbl> <dbl>
## 1 ex3_1    0.0790        0.0768 19.3
## 2 ex3_2    0.822         0.821   8.51
## 3 ex3_3    0.838         0.836   8.15
```

Seleção de modelos (i)

Modelos aninhados

(1) $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

(2) $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \epsilon$

O modelo reduzido (1) está aninhado no completo (2) $\rightarrow$ idênticos se os coeficientes das variáveis adicionais forem zero.

Como escolher? Estimamos as duas regressões e fazemos um **teste F** em que H_0 é que os coeficientes adicionais são zero.

$$F = \frac{(SSE_r - SSE_c)/df_1}{SSE_c/df_2} = \frac{(R_c^2 - R_r^2)/df_1}{(1 - R_c^2)/df_2}$$

Em que df_1 é o número de variáveis adicionais no modelo completo e df_2 é são os graus de liberdade residuais do modelo completo.

Seleção de modelos (ii)

Modelos não aninhados

(1) $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$

(2) $y = \alpha + \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_k z_k + \epsilon$

Não podemos fazer o mesmo **teste F** de antes, mas há alternativas um pouco mais complicadas (ex., estatística AIC).

Na prática, para o mesmo y (sem transformações), a opção mais intuitiva é comparar o r^2 ajustado entre modelos.

Se houver diferença grande entre eles, o modelo com r^2 ajustado mais elevado de ajusta melhor aos dados.

Exemplo #1: alturas

```
anova(ex1_1, ex1_3)
```

```
## Analysis of Variance Table
##
## Model 1: filhos ~ mae
## Model 2: filhos ~ mae + pai + num_irmaos
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    932 73989
## 2    930 68470  2    5518.7 37.479 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

```
linearHypothesis(ex1_3, c('pai = 0', 'num_irmaos = 0'))
```

```
## Linear hypothesis test
##
## Hypothesis:
```

Exemplo #2: restaurantes no Alasca

```
anova(ex2_1, ex2_2)

## Analysis of Variance Table
##
## Model 1: nota_inspecao ~ num_estab
## Model 2: nota_inspecao ~ num_estab + ano
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1  27176 995216
## 2  27175 991186  1    4030.6 110.5 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

linearHypothesis(ex2_2, c('ano = 0'))

## Linear hypothesis test
##
## Hypothesis:
```

Exemplo #3: escolas

```
anova(ex3_1, ex3_3)

## Analysis of Variance Table
##
## Model 1: nota_ler ~ pc_aluno
## Model 2: nota_ler ~ pc_aluno + pct_out_idioma + pct_ajuda_alm
##     renda_dist
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    418 156022
## 2    414  27503  4    128519 483.65 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

Recapitulação

Introdução à regressão multivariada

Preparação para a aula

Estimativas de ponto

Inferência

Próxima aula

Próxima aula

Atividade

Entrega da atividade #6 até às 8h30 do dia **12/12**

Leituras obrigatórias (idênticas às leituras das aulas 9 e 10)

Agresti 2018, cap.9 a 10

Agresti 2018, cap. 11 a 13

Leituras optativas (idênticas às leituras das aulas 9 e 10)

Agresti 2018, cap. 14

Bussab e Morettin 2010 cap. 16

Huntington-Klein, cap. 13