Modelos de Regressão e Previsão

Análise de Regressão Linear Multipla

Prof. Carlos Trucíos
carlos.trucios@facc.ufrj.br
ctruciosm.github.io

Faculdade de Administração e Ciências Contábeis
Universidade Federal do Rio de Janeiro

Aula 5

Introdução

Introdução

- Pensar que uma única variável x pode explicar y é bastante ingenuo

Introdução

- Pensar que uma única variável x pode explicar y é bastante ingenuo
- Dizer que todos os outros fatores que afetam y (e incorporados em u) são não correlacionados com x é bastante irrealista

Introdução

- Pensar que uma única variável x pode explicar y é bastante ingenuo
- Dizer que todos os outros fatores que afetam y (e incorporados em u) são não correlacionados com x é bastante irrealista
- Um modelo que inclua mais do que uma variavél explicativa parece ser bastante mais razoavel.

Introdução

- Pensar que uma única variável x pode explicar y é bastante ingenuo
- Dizer que todos os outros fatores que afetam y (e incorporados em u) são não correlacionados com x é bastante irrealista
- Um modelo que inclua mais do que uma variavél explicativa parece ser bastante mais razoavel.
- Se incluirmos no nosso modelo mais variáveis que sejam úteis para explicar y, mais da variabilidade de y poderá ser explicada

Introdução

- Pensar que uma única variável x pode explicar y é bastante ingenuo
- Dizer que todos os outros fatores que afetam y (e incorporados em u) são não correlacionados com x é bastante irrealista
- Um modelo que inclua mais do que uma variavél explicativa parece ser bastante mais razoavel.
- Se incluirmos no nosso modelo mais variáveis que sejam úteis para explicar y, mais da variabilidade de y poderá ser explicada

Introdução

- Pensar que uma única variável x pode explicar y é bastante ingenuo
- Dizer que todos os outros fatores que afetam y (e incorporados em u) são não correlacionados com x é bastante irrealista
- Um modelo que inclua mais do que uma variavél explicativa parece ser bastante mais razoavel.
- Se incluirmos no nosso modelo mais variáveis que sejam úteis para explicar y, mais da variabilidade de y poderá ser explicada

MRLM

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

RLM

RLM: Estimação

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{1,1} + \beta_2 x_{1,2} + \cdots + \beta_k x_{1,k} + u_1 \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + \cdots + \beta_k x_{n,k} + u_n
\end{aligned} \tag{1}
$$

RLM: Estimação

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{1,1} + \beta_2 x_{1,2} + \cdots + \beta_k x_{1,k} + u_1 \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + \cdots + \beta_k x_{n,k} + u_n
\end{aligned} \tag{1}
$$

$$
\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,k} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \vdots \\ y_k \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}
$$

RLM: Estimação

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{1,1} + \beta_2 x_{1,2} + \cdots + \beta_k x_{1,k} + u_1 \\
\vdots \\
y_n &= \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + \cdots + \beta_k x_{n,k} + u_n
\end{aligned} \tag{1}
$$

$$
\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,k} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \vdots \\ y_k \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}
$$

Em forma matricial

$$Y = X\beta + u$$

RLM: Estimação

Ideia: Minimizar a soma de quadrados do erro.

$$\hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k]' = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n} u_i^2$$

$$= \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n} (\underbrace{y_i - b_0 - b_1 x_{i,1} - \cdots - b_k x_{i,k}}_{u_i})^2 \quad (2)$$

RLM: Estimação

Ideia: Minimizar a soma de quadrados do erro.

$$\hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k]' = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n} u_i^2$$
$$= \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n} (\underbrace{y_i - b_0 - b_1 x_{i,1} - \cdots - b_k x_{i,k}}_{u_i})^2 \quad (2)$$

Ou equivalentemente,

$$\hat{\beta} = \underset{b}{\operatorname{argmin}}\, u'u = \underset{b}{\operatorname{argmin}} (Y - Xb)'(Y - Xb)$$

RLM: Estimação

Após um pouco de Cálculo Matricial,

$$\hat{\beta} = [X'X]^{-1}X'Y$$

RLM: Estimação

Após um pouco de Cálculo Matricial,

$$\hat{\beta} = [X'X]^{-1}X'Y$$

Note que,

$$\begin{aligned}\hat{\beta} &= [X'X]^{-1}X'Y = [X'X]^{-1}X'(\underbrace{X\beta + u}_{Y}) \\ &= \underbrace{[X'X]^{-1}X'X}_{I}\,\beta \qquad (3) \\ &= \beta + [X'X]^{-1}X'u\end{aligned}$$

RLM: Estimação

```
WAGE1 = read.table("./DadosMRP/wage1.txt")[,1:4]
colnames(WAGE1) = c("salario", "educacao",
                    "experiencia","anos_empresa")
coef(lm(log(salario) ~ educacao +
          experiencia + anos_empresa, data = WAGE1))
```

```
##  (Intercept)    educacao  experiencia anos_empresa
##  0.284359545 0.092028987  0.004121109  0.022067218
```

RLM: Estimação

```
WAGE1 = read.table("./DadosMRP/wage1.txt")[,1:4]
colnames(WAGE1) = c("salario", "educacao",
                    "experiencia","anos_empresa")
coef(lm(log(salario) ~ educacao +
          experiencia + anos_empresa, data = WAGE1))
```

```
##   (Intercept)     educacao   experiencia anos_empresa
##   0.284359545  0.092028987  0.004121109  0.022067218
```

- Mantendo os fatores *experiencia* e *anos_empresa* fixos, quando os anos de educação formal aumentam em 1, espera-se que o salário aumente em 9.2%(100×0.092028987)

RLM: Qualidade de ajuste - R2

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^n}_{SQT} = \underbrace{\sum_{i=1}^{2}(\hat{y}_i - \bar{y})^2}_{SQE} + \underbrace{\sum_{i=1}^{n}\hat{u}_i^2}_{SQR}$$

R2

$$R^2 = 1 - SQR/SQT = SQE/SQT$$

RLM: Qualidade de ajuste - R2

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^n}_{SQT} = \underbrace{\sum_{i=1}^{2}(\hat{y}_i - \bar{y})^2}_{SQE} + \underbrace{\sum_{i=1}^{n}\hat{u}_i^2}_{SQR}$$

R2

$$R^2 = 1 - SQR/SQT = SQE/SQT$$

- R^2 já foi introduzido na RLS

RLM: Qualidade de ajuste - R2

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^n}_{SQT} = \underbrace{\sum_{i=1}^{2}(\hat{y}_i - \bar{y})^2}_{SQE} + \underbrace{\sum_{i=1}^{n}\hat{u}_i^2}_{SQR}$$

R2

$$R^2 = 1 - SQR/SQT = SQE/SQT$$

- R^2 já foi introduzido na RLS
- R^2 : proporção da variabilidade de y explicada pelo modelo

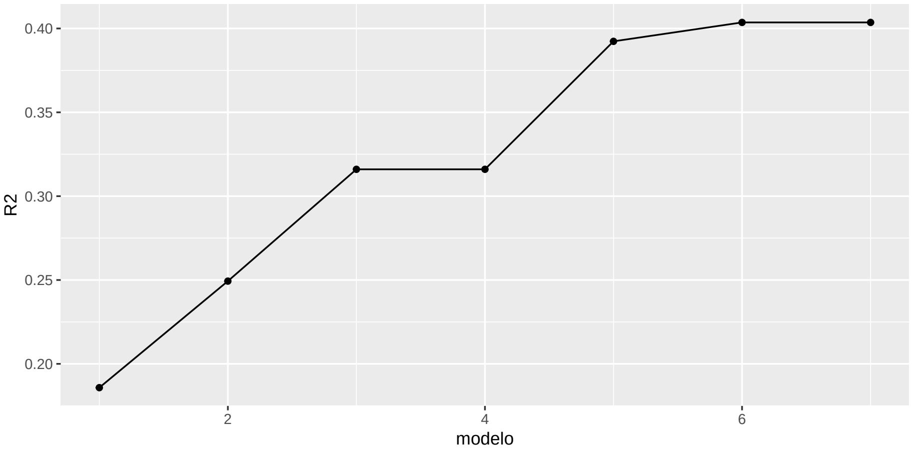
RLM: Qualidade de ajuste - R2

```
modelo1 = lm(log(wage)~educ,data = WAGE1)
modelo2 = lm(log(wage)~educ+exper,data = WAGE1)
modelo3 = lm(log(wage)~educ+exper+tenure,data = WAGE1)
modelo4 = lm(log(wage)~educ+exper+tenure+nonwhite,data = WAGE1)
modelo5 = lm(log(wage)~educ+exper+tenure+nonwhite
            +female,data = WAGE1)
modelo6 = lm(log(wage)~educ+exper+tenure+nonwhite
            +female + married,data = WAGE1)
modelo7 = lm(log(wage)~educ+exper+tenure+nonwhite+female
            +married+numdep,data = WAGE1)
```

RLM: Qualidade de ajuste - R2

```
x = 1:7
y = c(summary(modelo1)$r.squared, summary(modelo2)$r.squared,
  summary(modelo3)$r.squared, summary(modelo4)$r.squared,
  summary(modelo5)$r.squared, summary(modelo6)$r.squared,
  summary(modelo7)$r.squared)
dados = data.frame(R2 = y, modelo = x)
library(ggplot2)
ggplot(dados, aes(y = R2, x= modelo)) +
  geom_line() + geom_point()
```

RLM: Qualidade de ajuste - R2



RLM: Qualidade de ajuste -R2 ajustado

- R^2 tem a desvantagem que nunca diminui quando incluimos uma nova variavel no modelo (mesmo que esta não seja importante)

RLM: Qualidade de ajuste -R2 ajustado

- R^2 tem a desvantagem que nunca diminui quando incluimos uma nova variavel no modelo (mesmo que esta não seja importante)
- Uma alternativa é usar uma nova medida de qualidade de ajuste.

RLM: Qualidade de ajuste -R2 ajustado

- R^2 tem a desvantagem que nunca diminui quando incluimos uma nova variavel no modelo (mesmo que esta não seja importante)
- Uma alternativa é usar uma nova medida de qualidade de ajuste.

RLM: Qualidade de ajuste -R2 ajustado

- R^2 tem a desvantagem que nunca diminui quando incluimos uma nova variavel no modelo (mesmo que esta não seja importante)
- Uma alternativa é usar uma nova medida de qualidade de ajuste.

R2-Ajustado

$$R_A^2 = 1 - \frac{n-1}{n-(k+1)}(1-R^2)$$

RLM: Qualidade de ajuste -R2 ajustado

```
yhat = fitted.values(modelo1)
ytrue = log(WAGE1$wage)
n = length(ytrue)
k = 1
R2 = cor(yhat,ytrue)^2
R2a = 1- (n-1)/(n-(k+1))*(1-R2)
R2a
```

```
## [1] 0.1842527
```

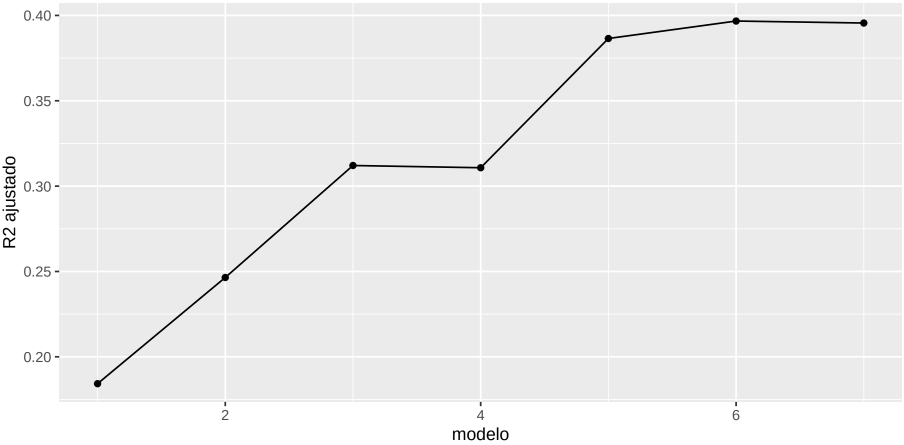
```
summary(modelo1)$adj.r.squared
```

```
## [1] 0.1842527
```

RLM: Qualidade de ajuste - R2 ajustado

```
x = 1:7
yadj = c(summary(modelo1)$adj.r.squared,
         summary(modelo2)$adj.r.squared,
         summary(modelo3)$adj.r.squared,
         summary(modelo4)$adj.r.squared,
         summary(modelo5)$adj.r.squared,
         summary(modelo6)$adj.r.squared,
         summary(modelo7)$adj.r.squared)
dados = data.frame(R2 = yadj, modelo = x)
library(ggplot2)
ggplot(dados, aes(y = R2, x= modelo)) + geom_line() +
  geom_point() + ylab("R2 ajustado")
```

RLM: Qualidade de ajuste - R2 ajustado



RLM: Qualidade de ajuste - R2 ajustado

```r
#R2
round(y,4)
```

```
## [1] 0.1858 0.2493 0.3160 0.3160 0.3923 0.4036 0.4036
```

```r
#R2-Ajustado
round(yadj,4)
```

```
## [1] 0.1843 0.2465 0.3121 0.3108 0.3865 0.3967 0.3956
```

RLM: Propriedades

HRLM1: Linear nos parâmetros

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u \tag{4}$$

RLM: Propriedades

HRLM1: Linear nos parâmetros

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u \tag{4}$$

HRLM2: Amostragem aleatória

$(y_1, x_{1,1}, \ldots, x_{1,k}), \cdots, (y_n, x_{n,1}, \ldots, x_{n,k})$ constituem uma a.a. de tamanho n do modelo populacional (4)

RLM: Propriedades

HRLM1: Linear nos parâmetros

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u \tag{4}$$

HRLM2: Amostragem aleatória

$(y_1, x_{1,1}, \ldots, x_{1,k}), \cdots, (y_n, x_{n,1}, \ldots, x_{n,k})$ constituem uma a.a. de tamanho n do modelo populacional (4)

HRLM3: Colinearidade não perfeita

Não há relações linerares exatas entre as variáveis independêntes e nenhuma das variáveis independentes é constante.

RLM: Propriedades

HRLM1: Linear nos parâmetros

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u \tag{4}$$

HRLM2: Amostragem aleatória

$(y_1, x_{1,1}, \ldots, x_{1,k}), \cdots, (y_n, x_{n,1}, \ldots, x_{n,k})$ constituem uma a.a. de tamanho n do modelo populacional (4)

HRLM3: Colinearidade não perfeita

Não há relações linerares exatas entre as variáveis independêntes e nenhuma das variáveis independentes é constante.

HRLM4: Média condicional zero

$$E(u|X) = 0$$

RLM: Propriedades

Teorema: Inexistência do viés MQO

Sob HRLM1–HRLM4,

$$E(\hat{\beta}) = \beta$$

Prova

$$\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + u) = \underbrace{(X'X)^{-1}X'X}_{I}\beta + (X'X)^{-1}X'u$$

$$E(\hat{\beta}|X) = E(\beta + (X'X)^{-1}X'u|X) = \underbrace{E(\beta|X)}_{\beta} + \underbrace{E((X'X)^{-1}X'u|X)}_{(X'X)^{-1}X'\underbrace{E(u|X)}_{0}} = \beta$$

Logo, $E[E(\hat{\beta}|X)] = E[\hat{\beta}] = \beta$

RLM: Propriedades

HRLM5: Variância constante

$$V(u|X) = E[uu'|X] = \sigma^2 I$$

RLM: Propriedades

HRLM5: Variância constante

$$V(u|X) = E[uu'|X] = \sigma^2 I$$

Variância dos EMQO

Sob HRLM1–HRLM5,

$$V(\hat{\beta}|X) = \sigma^2(X'X)^{-1}$$

RLM: Propriedades

Prova Sabemos que $\hat{\beta} = \beta + [X'X]^{-1}X'u$

RLM: Propriedades

Prova Sabemos que $\hat{\beta} = \beta + [X'X]^{-1}X'u$

$$
\begin{aligned}
V(\hat{\beta}|X) &= E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'|X\right] \\
&= E\left[(X'X)^{-1}X'u((X'X)^{-1}X'u)'|X\right] \\
&= E\left[(X'X)^{-1}X'uu'X(X'X)^{-1}|X\right] \\
&= (X'X)^{-1}X' \underbrace{E\left[uu'|X\right]}_{\sigma^2 I} X(X'X)^{-1} \\
&= \sigma^2 \underbrace{(X'X)^{-1}X'X}_{I}(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}
\end{aligned} \tag{5}
$$

RLM: Propriedades

- Na prática, σ^2 não é conhecido, precisamos estima-lo

RLM: Propriedades

- Na prática, σ^2 não é conhecido, precisamos estima-lo
- $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} \hat{u}_i^2}{n - (k+1)}$

RLM: Propriedades

- Na prática, σ^2 não é conhecido, precisamos estima-lo
- $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} \hat{u}_i^2}{n - (k+1)}$

RLM: Propriedades

- Na prática, σ^2 não é conhecido, precisamos estima-lo
- $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} \hat{u}_i^2}{n - (k+1)}$

Hipóteses de Gauss-Markov

HRLM1–HRLM5

- Sob as hipóteses de Gauss-Markov, $E(\hat{\sigma}^2) = \sigma^2$

RLM na prática

Incluir variáveis irrelevantes

Suponha que o modelo populacional seja

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

e que estimemos o modelo

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

Qual é o efeito de incluir a variavel irrelevante x_3?

- Em termos da inexistencia do viés não há ifeito nenhum $E(\hat{\beta}) = [\beta_0, \beta_1, \beta_2, 0]'$

Incluir variáveis irrelevantes

Suponha que o modelo populacional seja

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

e que estimemos o modelo

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

Qual é o efeito de incluir a variavel irrelevante x_3?

- Em termos da inexistencia do viés não há ifeito nenhum $E(\hat{\beta}) = [\beta_0, \beta_1, \beta_2, 0]'$
- Contudo, incluir variáveis irrelevantes pode ter efeitos indesejáveis nas variâncias dos estimadores MQO.

Incluir variáveis irrelevantes

```
library(MASS)
simular_dados_ir = function(n,rho,betas){
  u = rnorm(n)
  mu = c(0,0)
  Sigma = matrix(c(1,rho,rho,1),ncol=2)
  x = mvrnorm(2*n, mu, Sigma)
  y = betas[1] + betas[2]*x[,1] + u
  dados = data.frame(y,x1 = x[,1], x2 = x[,2])
  return(dados)
}
```

Incluir variáveis irrelevantes

```
dados = simular_dados_ir(1000,0,c(2,1.2))
summary(lm(y~x1, data = dados))$coefficients
```

```
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 1.995039 0.02304782 86.56087        0
## x1          1.185456 0.02283101 51.92305        0
```

```
summary(lm(y~x1+x2, data = dados))$coefficients
```

```
##                 Estimate Std. Error    t value  Pr(>|t|)
## (Intercept)  1.995117268 0.02305624 86.5326383 0.0000000
## x1           1.185503354 0.02283756 51.9102399 0.0000000
## x2          -0.004835569 0.02272230 -0.2128116 0.8314957
```

Incluir variáveis irrelevantes

```
dados = simular_dados_ir(1000,0.8,c(2,1.2))
summary(lm(y~x1, data = dados))$coefficients
```

```
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 2.058491 0.02173913 94.69059        0
## x1          1.222505 0.02158600 56.63416        0
```

```
summary(lm(y~x1+x2, data = dados))$coefficients
```

```
##               Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)  2.0583523 0.02174762 94.6472384  0.000000e+00
## x1           1.2327190 0.03657242 33.7062440 1.469904e-197
## x2          -0.0125775 0.03634975 -0.3460133  7.293691e-01
```

Omissão de variáveis

Suponha que o modelo populacional seja

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

e que estimemos o modelo

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Qual é o efeito de omitir a variavel x_3?

- Em geral, produz estimadores viesados

Omissão de variáveis

```r
# Simulando os dados
library(MASS)
simular_dados = function(n,rho,betas){
  u = rnorm(n)
  mu = c(0,0)
  Sigma = matrix(c(1,rho,rho,1),ncol=2)
  x = mvrnorm(2*n, mu, Sigma)
  y = betas[1] + betas[2]*x[,1] + betas[3]*x[,2] + u
  dados = data.frame(y,x1 = x[,1], x2 = x[,2])
  return(dados)
}
```

Omissão de variáveis

```
dados = simular_dados(1000,0,c(2,1.2,0.6))
coef(lm(y~x1+x2, data = dados))
```

```
## (Intercept)          x1          x2
##   1.9703174   1.2013451   0.5788781
```

```
coef(lm(y~x1, data = dados))
```

```
## (Intercept)          x1
##    1.946137    1.204578
```

Omissão de variáveis

```
dados = simular_dados(1000,0.7,c(2,1.2,0.6))
coef(lm(y~x1+x2, data = dados))
```

```
## (Intercept)          x1          x2
##   2.0099273   1.1782008   0.6327452
```

```
coef(lm(y~x1, data = dados))
```

```
## (Intercept)          x1
##    1.999019    1.630289
```

Omissão de variáveis

```
dados = simular_dados(1000,-0.7,c(2,1.2,0.6))
coef(lm(y~x1+x2, data = dados))
```

```
## (Intercept)          x1          x2
##   2.0090776   1.1486131   0.5597226
```

```
coef(lm(y~x1, data = dados))
```

```
## (Intercept)          x1
##   2.0213655   0.7702578
```

Omissão de variáveis

```r
dados = simular_dados(1000,-0.8,c(2,0.5,1.2))
coef(lm(y~x1+x2, data = dados))
```

```
## (Intercept)          x1          x2
##   1.9530441   0.4890481   1.1813204
```

```r
coef(lm(y~x1, data = dados))
```

```
## (Intercept)          x1
##    1.972377   -0.440604
```

Teorema de Gauss-Markov

Teorema de Gauss-Markov

Sob as hipóteses HRLM1–HRLM5, $\hat{\beta}$ é o melhor estimador linear não viesado (Best Linear Unbiased Estimator –BLUE) de β.

Teorema de Gauss-Markov

Teorema de Gauss-Markov

Sob as hipóteses HRLM1–HRLM5, $\hat{\beta}$ é o melhor estimador linear não viesado (Best Linear Unbiased Estimator –BLUE) de β.

- $\tilde{\beta}$ é **linear** se $\tilde{\beta} = A'Y$, onde $A_{n\times(k+1)}$ função de X

Teorema de Gauss-Markov

Teorema de Gauss-Markov

Sob as hipóteses HRLM1–HRLM5, $\hat{\beta}$ é o melhor estimador linear não viesado (Best Linear Unbiased Estimator –BLUE) de β.

- $\tilde{\beta}$ é **linear** se $\tilde{\beta} = A'Y$, onde $A_{n\times(k+1)}$ função de X
- **Melhor:** menor variância. $V(\hat{\beta}|X) \leq V(\tilde{\beta}|X)$, para qualquer estimador linear não viesado $\tilde{\beta}$

Multicolinearidade

- A HRLM3 nos diz que não existe colinearidade perfeita entra as variáveis independentes

Multicolinearidade

- A HRLM3 nos diz que não existe colinearidade perfeita entra as variáveis independentes
- Contudo, na prática podemos ter variaveis independentes fortemente correlacionadas (mas $\neq \pm 1$)

Multicolinearidade

- A HRLM3 nos diz que não existe colinearidade perfeita entra as variáveis independentes
- Contudo, na prática podemos ter variaveis independentes fortemente correlacionadas (mas $\neq \pm 1$)
- Este fenômeno é conhecido na literatura como **multicolinearidade**

Multicolinearidade

- A HRLM3 nos diz que não existe colinearidade perfeita entra as variáveis independentes
- Contudo, na prática podemos ter variaveis independentes fortemente correlacionadas (mas $\neq \pm 1$)
- Este fenômeno é conhecido na literatura como **multicolinearidade**
- Multicolinearidade tem consequencias tanto na estimação dos parâmetros quanto na estimação das suas respectivas variâncias.

Multicolinearidade

```r
dados = simular_dados_ir(1000,0.99,c(2,1.2))
summary(lm(y~x1, data = dados))$coefficients
```

```
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 2.026173 0.02218281 91.33979        0
## x1          1.226712 0.02221425 55.22184        0
```

```r
summary(lm(y~x1+x2, data = dados))$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  2.0271272 0.02218174 91.387208 0.000000e+00
## x1           1.4769665 0.15648755  9.438236 1.016722e-20
## x2          -0.2520998 0.15604626 -1.615545 1.063507e-01
```

Multicolinearidade

```r
dados = simular_dados_ir(1000,0.995,c(2,1.2))
summary(lm(y~x1, data = dados))$coefficients
```

```
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 1.983651 0.02197578 90.26535        0
## x1          1.177893 0.02181073 54.00523        0
```

```r
summary(lm(y~x1+x2, data = dados))$coefficients
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 1.98337840  0.0219893 90.1974434 0.000000e+00
## x1          1.08540739  0.2151364  5.0452062 4.940465e-07
## x2          0.09313229  0.2155232  0.4321219 6.656995e-01
```

Multicolinearidade: VIF

VIF: Fator de Inflação de Variância (**V**ariance **I**nflation **F**actor)

VIF

$$VIF_j = \frac{1}{1 - R_j^2}$$

onde R_j^2 é o R^2 da regressão da j-ésima variavel (x_j) sobre as outras variáveis regressoras.

Multicolinearidade: VIF

VIF: Fator de Inflação de Variância (**V**ariance **I**nflation **F**actor)

VIF

$$VIF_j = \frac{1}{1 - R_j^2}$$

onde R_j^2 é o R^2 da regressão da j-ésima variavel (x_j) sobre as outras variáveis regressoras.

- Valores grandes de VIF_j podem indicar multicolinearidade

Multicolinearidade: VIF

VIF: Fator de Inflação de Variância (**V**ariance **I**nflation **F**actor)

VIF

$$VIF_j = \frac{1}{1 - R_j^2}$$

onde R_j^2 é o R^2 da regressão da j-ésima variavel (x_j) sobre as outras variáveis regressoras.

- Valores grandes de VIF_j podem indicar multicolinearidade
- **Quanto é grande?** Algumas vezes 10 é considerado grande, mas nao existe uma regra (10 representaria um $R_j^2 = 0.9$)

Multicolinearidade: VIF

```
library(car)
```

```
## Loading required package: carData
```

```
dados = simular_dados_ir(1000,0.995,c(2,1.2))
modelo = lm(y~x1+x2, data = dados)
vif(modelo)
```

```
##       x1       x2
## 102.6393 102.6393
```

Multicolinearidade: VIF

```r
WAGE1 = read.table("./DadosMRP/wage1.txt")[,1:4]
colnames(WAGE1) = c("salario", "educacao",
                    "experiencia","anos_empresa")
modelo = lm(log(salario) ~ educacao +
           experiencia + anos_empresa, data = WAGE1)
vif(modelo)
```

```
##     educacao  experiencia anos_empresa
##     1.112771     1.477618     1.349296
```

Leituras recomendadas

Leituras recomendadas

- Wooldridge, Jeffrey M. *Introdução à Econometria: Uma abordagem moderna*. (2016). Cengage Learning. – **Cap 3**