ME613 - Análise de Regressão

Parte 7

Samara F. Kiihl - IMECC - UNICAMP

Soma extra de quadrados

Motivação

- Verificar a redução na soma de quadrados do erro quando uma ou mais variáveis preditoras são adicionadas no modelo de regressão, dado que outras variáveis preditoras já estão incluídas no modelo.
- Equivalentemente, podemos utilizar a soma extra de quadrados para medir o aumento na soma de quadrados da regressão ao adicionarmos uma ou mais preditoras no modelo.
- Em resumo, a soma extra de quadrados pode nos auxiliar na decisão de inclusão ou retirada de variáveis no modelo.

Exemplo

Relação entre gordura corporal e 3 medidas corporais.

Subject i	Triceps Skinfold Thickness X_{i1}	Thigh Circumference X_{i2}	Midarm Circumference X_{i3}	Body Fat Y_i
1	19.5	43.1	29.1	11.9
2	24.7	49.8	28.2	22.8
3	30.7	51.9	37.0	18.7
...	...	...	...	...
18	30.2	58.6	24.6	25.4
19	22.7	48.2	27.1	14.8
20	25.2	51.0	27.5	21.1

Exemplo: Regressão de Y em X_1

(a) Regression of Y on X_1

$\hat{Y} = -1.496 + .8572X_1$

Source of Variation	SS	df	MS
Regression	352.27	1	352.27
Error	143.12	18	7.95
Total	495.39	19	

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_1	$b_1 = .8572$	$s\{b_1\} = .1288$	6.66

$$SQReg(X_1) = 352.27$$

$$SQE(X_1) = 143.12$$

Exemplo: Regressão de Y em X_1

```
dat = read.table('./dados/fat.txt')
colnames(dat) <- c("X1","X2","X3","Y")
X1 = dat[,1]
X2 = dat[,2]
X3 = dat[,3]
Y = dat[,4]

modelo1 <- lm(Y ~X1)
summary(modelo1)$coefficients
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -1.4961046  3.3192346 -0.4507378 6.575609e-01
## X1           0.8571865  0.1287808  6.6561675 3.024349e-06
```


Exemplo: Regressão de Y em X_1

```
anova(modelo1)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 352.27  352.27  44.305 3.024e-06 ***
## Residuals 18 143.12    7.95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Exemplo: Regressão de Y em X_2

(b) Regression of Y on X_2

$\hat{Y} = -23.634 + .8565X_2$

Source of Variation	SS	df	MS
Regression	381.97	1	381.97
Error	113.42	18	6.30
Total	495.39	19	

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_2	$b_2 = .8565$	$s\{b_2\} = .1100$	7.79

$$SQReg(X_2) = 381.97$$

$$SQE(X_2) = 113.42$$

Exemplo: Regressão de Y em X_2

```
modelo2 <- lm(Y ~X2)
summary(modelo2)$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -23.6344891  5.6574137 -4.177614 5.656662e-04
## X2            0.8565466  0.1100156  7.785681 3.599996e-07
```

```
anova(modelo2)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## X2         1 381.97  381.97  60.617 3.6e-07 ***
## Residuals 18 113.42    6.30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Exemplo: Regressão de Y em X_1 e X_2

(c) Regression of Y on X_1 and X_2

$\hat{Y} = -19.174 + .2224X_1 + .6594X_2$

Source of Variation	SS	df	MS
Regression	385.44	2	192.72
Error	109.95	17	6.47
Total	495.39	19	

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_1	$b_1 = .2224$	$s\{b_1\} = .3034$	.73
X_2	$b_2 = .6594$	$s\{b_2\} = .2912$	2.26

$$SQReg(X_1, X_2) = 385.44$$

$$SQE(X_1, X_2) = 109.95$$

Exemplo: Regressão de Y em X_1 e X_2

```
modelo12 <- lm(Y ~ X1 + X2)
summary(modelo12)$coefficients
```

```
##                Estimate Std. Error    t value   Pr(>|t|)
## (Intercept) -19.1742456  8.3606407 -2.2933943 0.03484327
## X1            0.2223526  0.3034389  0.7327755 0.47367898
## X2            0.6594218  0.2911873  2.2645969 0.03689872
```

```
anova(modelo12)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 352.27  352.27 54.4661 1.075e-06 ***
## X2         1  33.17   33.17  5.1284    0.0369 *
## Residuals 17 109.95    6.47
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Exemplo: Regressão de Y em X_1 e X_2

```
SQReg <- sum(anova(modelo12)[1:2,2])
SQReg
```

```
## [1] 385.4387
```

Exemplo: Soma extra de quadrados

Quando ambos X_1 e X_2 estão no modelo, temos que $SQE(X_1, X_2) = 109.95$, que é menor do que com apenas X_1 no modelo, $SQE(X_1) = 143.12$.

Esta diferença é denominada **soma extra de quadrados**:

$$SQReg(X_2 \mid X_1) = SQE(X_1) - SQE(X_1, X_2) = 143.12 - 109.95 = 33.17$$

Equivalentemente:

$$SQReg(X_2 \mid X_1) = SQReg(X_1, X_2) - SQReg(X_1) = 385.44 - 352.27 = 33.17$$

Exemplo: Soma extra de quadrados

```
modelo12 <- lm(Y ~ X1 + X2)
anova(modelo12)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 352.27  352.27 54.4661 1.075e-06 ***
## X2         1  33.17   33.17  5.1284    0.0369 *
## Residuals 17 109.95    6.47
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Na tabela, a linha X_2 contém $SQReg(X_2 \mid X_1)$.

Exemplo: Regressão de Y em X_1, X_2 e X_3

(d) Regression of Y on X_1, X_2, and X_3
$\hat{Y} = 117.08 + 4.334X_1 - 2.857X_2 - 2.186X_3$

Source of Variation	SS	df	MS
Regression	396.98	3	132.33
Error	98.41	16	6.15
Total	495.39	19	

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_1	$b_1 = 4.334$	$s\{b_1\} = 3.016$	1.44
X_2	$b_2 = -2.857$	$s\{b_2\} = 2.582$	−1.11
X_3	$b_3 = -2.186$	$s\{b_3\} = 1.596$	−1.37

$$SQReg(X_1, X_2, X_3) = 396.98$$

$$SQE(X_1, X_2, X_3) = 98.41$$

Exemplo: Regressão de Y em X_1, X_2 e X_3

```
modelo123 <- lm(Y ~ X1 + X2 + X3)
summary(modelo123)$coefficients
```

```
##               Estimate Std. Error   t value  Pr(>|t|)
## (Intercept) 117.084695  99.782403  1.173400 0.2578078
## X1            4.334092   3.015511  1.437266 0.1699111
## X2           -2.856848   2.582015 -1.106441 0.2848944
## X3           -2.186060   1.595499 -1.370142 0.1895628
```

```
anova(modelo123)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 352.27  352.27 57.2768 1.131e-06 ***
## X2         1  33.17   33.17  5.3931   0.03373 *
## X3         1  11.55   11.55  1.8773   0.18956
## Residuals 16  98.40    6.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exemplo: Soma extra de quadrados

Quando X_1, X_2 e X_3 estão no modelo, temos que $SQE(X_1, X_2, X_3) = 98.41$, que é menor do que com apenas X_1 e X_2 no modelo, $SQE(X_1, X_2) = 109.95$.

Esta diferença é denominada **soma extra de quadrados**:

$$\begin{aligned} SQReg(X_3 \mid X_1, X_2) &= SQE(X_1, X_2) - SQE(X_1, X_2, X_3) \\ &= 109.95 - 98.41 = 11.54 \end{aligned}$$

Equivalentemente:

$$\begin{aligned} SQReg(X_3 \mid X_1, X_2) &= SQReg(X_1, X_2, X_3) - SQReg(X_1, X_2) \\ &= 396.98 - 385.44 = 11.54 \end{aligned}$$

Exemplo: Soma extra de quadrados

```
modelo123 <- lm(Y ~ X1 + X2 + X3)
anova(modelo123)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 352.27  352.27 57.2768 1.131e-06 ***
## X2         1  33.17   33.17  5.3931   0.03373 *
## X3         1  11.55   11.55  1.8773   0.18956
## Residuals 16  98.40    6.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Na tabela, a linha X_2 contém $SQReg(X_2 \mid X_1)$.

Na tabela, a linha X_3 contém $SQReg(X_3 \mid X_1, X_2)$.

Exemplo: Soma extra de quadrados

Podemos avaliar, também, a adição de mais de uma variável ao mesmo tempo. Por exemplo, podemos avaliar o efeito de incluir X_2 e X_3 a um modelo com apenas X_1:

$$\begin{aligned} SQReg(X_2, X_3 \mid X_1) &= SQE(X_1) - SQE(X_1, X_2, X_3) \\ &= 143.12 - 98.41 = 44.71 \end{aligned}$$

Equivalentemente:

$$\begin{aligned} SQReg(X_2, X_3 \mid X_1) &= SQReg(X_1, X_2, X_3) - SQReg(X_1) \\ &= 396.98 - 352.27 = 44.71 \end{aligned}$$

Exemplo: Soma extra de quadrados

```
modelo1 <- lm(Y ~X1)
modelo123 <- lm(Y ~X1 + X2 + X3)

anova(modelo1,modelo123)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1
## Model 2: Y ~ X1 + X2 + X3
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1     18 143.120
## 2     16  98.405  2    44.715 3.6352 0.04995 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$SQReg(X_2, X_3 \mid X_1) = 44.71$$

Soma extra de quadrados

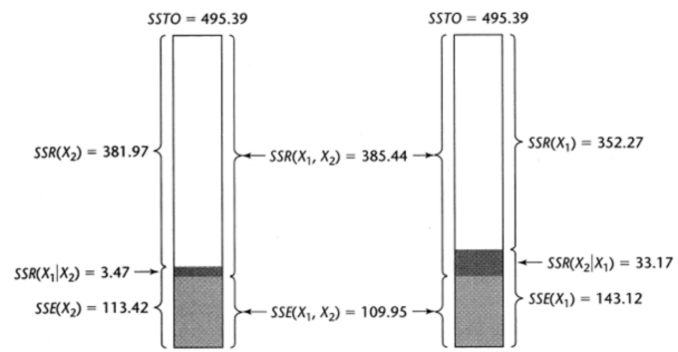
Em geral, se temos X_1 e X_2 no modelo, podemos escrever:

$$SQReg(X_1, X_2) = SQReg(X_1) + SQReg(X_2 \mid X_1)$$

ou, dado que a ordem de entrada das variáveis é arbitrária no modelo, temos:

$$SQReg(X_1, X_2) = SQReg(X_2) + SQReg(X_1 \mid X_2)$$

Exemplo



Soma extra de quadrados

Se temos X_1, X_2 e X_3 no modelo, podemos escrever, por exemplo:

$$SQReg(X_1, X_2, X_3) = SQReg(X_1) + SQReg(X_2 \mid X_1) + SQReg(X_3 \mid X_1, X_2)$$

$$SQReg(X_1, X_2, X_3) = SQReg(X_2) + SQReg(X_3 \mid X_2) + SQReg(X_1 \mid X_2, X_3)$$

$$SQReg(X_1, X_2, X_3) = SQReg(X_1) + SQReg(X_2, X_3 \mid X_1)$$

Teste para β_k usando soma extra de quadrados

- $H_0 : \beta_k = 0.$
- $H_1 : \beta_k \neq 0.$

Vimos que podemos usar a seguinte estatística do teste:

$$t^* = \frac{\hat{\beta}_k}{\sqrt{\widehat{Var(\hat{\beta}_k)}}} \overset{\text{sob } H_0}{\sim} t_{n-p}$$

Teste para β_k usando soma extra de quadrados

Equivalentemente, podemos utilizar soma extra de quadrados para o mesmo teste de hipóteses.

Estatística do teste:

$$F^* = \frac{SQReg(X_k \mid X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_{p-1})}{1} \div \frac{SQE(X_1, \ldots, X_{p-1})}{n-p}$$
$$\overset{\text{sob } H_0}{\sim} F_{1,n-p}$$

Exemplo: Regressão de Y em X_1, X_2 e X_3

Queremos testar se X_3 pode ser excluída do modelo.

```
modelo12 <- lm(Y ~X1 + X2)
modelo123 <- lm(Y ~X1 + X2 + X3)
anova(modelo12,modelo123)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X2
## Model 2: Y ~ X1 + X2 + X3
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     17 109.951
## 2     16  98.405  1    11.546 1.8773 0.1896
```

$F^* = 1.88$. Não encontramos evidências para rejeitar $H_0: \beta_3 = 0$.

Teste para vários β_k 's usando soma extra de quadrados

- $H_0: \beta_q = \beta_{q+1} = \ldots = \beta_{p-1} = 0$.
- H_1: pelo menos um $\beta_q, \ldots, \beta_{p-1}$ não é zero.

(por conveniência, a notação assume que os últimos $p - q$ coeficientes do modelo serão testados)

Estatística do teste:

$$F^* = \frac{SQReg(X_q, \ldots, X_{p-1} \mid X_1, \ldots, X_{q-1})}{p-q} \div \frac{SQE(X_1, \ldots, X_{p-1})}{n-p}$$
$$\overset{\text{sob } H_0}{\sim} F_{p-q,n-p}$$

Exemplo: Regressão de Y em X_1, X_2 e X_3

Queremos testar se X_2 e X_3 podem ser excluídas do modelo.

```
modelo1 <- lm(Y ~ X1)
modelo123 <- lm(Y ~ X1 + X2 + X3)
anova(modelo1,modelo123)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1
## Model 2: Y ~ X1 + X2 + X3
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1     18 143.120
## 2     16  98.405  2    44.715 3.6352 0.04995 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$F^* = 3.64$.

Coeficiente de Determinação Parcial

Motivação

Para avaliar o modelo: observar quanto da SQT está contida em $SQReg$ e quanto está na SQE.

Podemos utilizar para avaliar o modelo:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{SQReg}{SQT}$$

conhecido como **coeficiente de determinação**, que é a proporção da variabilidade total explicada pelo modelo de regressão ajustado.

O **coeficiente de determinação parcial** irá avaliar a contribuição marginal de alguma(s) preditora(s), dado que as demais já estão no modelo.

Caso de duas variáveis preditoras

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

- Coeficiente de determinação parcial entre Y e X_1, dado que X_2 já está no modelo:

$$R^2_{Y1|2} = \frac{SQE(X_2) - SQE(X_1, X_2)}{SQE(X_2)} = \frac{SQReg(X_1 \mid X_2)}{SQE(X_2)}$$

- Coeficiente de determinação parcial entre Y e X_2, dado que X_1 já está no modelo:

$$R^2_{Y2|1} = \frac{SQE(X_1) - SQE(X_1, X_2)}{SQE(X_1)} = \frac{SQReg(X_2 \mid X_1)}{SQE(X_1)}$$

Exemplos

$$R^2_{Y1|23} = \frac{SQReg(X_1 \mid X_2, X_3)}{SQE(X_2, X_3)}$$

$$R^2_{Y2|13} = \frac{SQReg(X_2 \mid X_1, X_3)}{SQE(X_1, X_3)}$$

$$R^2_{Y3|12} = \frac{SQReg(X_3 \mid X_1, X_2)}{SQE(X_1, X_2)}$$

$$R^2_{Y4|123} = \frac{SQReg(X_4 \mid X_1, X_2, X_3)}{SQE(X_1, X_2, X_3)}$$

Exemplo: Gordura corporal

```
SQE1 <- deviance(modelo1) #SQE modelo só com X1
SQE2 <- deviance(modelo2) #SQE modelo só com X2
SQE12 <- deviance(modelo12) #SQE modelo com X1 e X2
SQE123 <- deviance(modelo123) #SQE modelo com X1 X2 e X3

SQReg2.1 <- SQE1-SQE12 # SQReg(X2|X1)
SQReg3.12 <- SQE12-SQE123 # SQReg(X3|X1,X2)


RY2.1  <- SQReg2.1/SQE1 # Coef. det. parcial de Y com X2 dado X1 no modelo
RY2.1


## [1] 0.2317564


RY3.12  <- SQReg3.12/SQE12 # Coef. det. parcial de Y com X3 dado X1 e X2 no modelo
RY3.12


## [1] 0.1050097
```


Exemplo: Gordura corporal

Quando X_2 é adicionada ao modelo contendo apenas X_1, a $SQE(X_1)$ é reduzida em 23%. A inclusão de X_2 no modelo explica 23% da variação em Y que não pode ser explicada apenas por X_1.

Quando X_3 é adicionada ao modelo contendo X_1 e X_2, a $SQE(X_1, X_2)$ é reduzida em 10%. Isto é, 10% da variação em Y que não pode ser explicada pelo modelo com X_1 e X_2 é explicada pela inclusão de X_3 no modelo.

Propriedades

O coeficiente de determinação parcial assume valores entre 0 e 1.

Outra maneira de obter $R^2_{Y1|2}$:

- Obtenha os resíduos da regressão de Y em X_2: $e_i(Y \mid X_2)$.
- Obtenha os resíduos da regressão de X_1 em X_2: $e_i(X_1 \mid X_2)$.
- Calcule R^2 entre $e_i(Y \mid X_2)$ e $e_i(X_1 \mid X_2)$.

O diagrama de dispersão de $e_i(Y \mid X_2)$ versus $e_i(X_1 \mid X_2)$ fornece uma representação gráfica da relação entre Y e X_1, ajustada por X_2. É também chamado de *added variable plot* ou **gráfico de regressão parcial**.

Regressão Múltipla Padronizada

Motivação

Erros de precisão numérica quando

- $\mathbf{X}^T\mathbf{X}$ tem determinante próximo de 0.
- elementos de $\mathbf{X}^T\mathbf{X}$ diferem substancialmente em ordem de magnitude.

Para cada um dos problemas, há soluções propostas.

Veremos inicialmente o problema de ordem de magnitude.

Transformação de correlação

Ao utilizarmos a transformação de correlação, obtemos que todos os elementos de $\mathbf{X}^T\mathbf{X}$ variam entre 1 e -1.

Isto acarreta menos problemas de arredondamento para inverter $\mathbf{X}^T\mathbf{X}$.

Falta de comparabilidade entre coeficientes

Em geral, não podemos compara os coeficientes de regressão entre si, dado que não estão nas mesmas unidades.

Exemplo:

$$\hat{Y} = 200 + 20000X_1 + 0.2X_2$$

Pode-se pensar que apenas X_1 é relevante no modelo.

Mas suponha que:

Y: dólares

X_1: milhares de dólares

X_2: centavos de dólares

Falta de comparabilidade entre coeficientes

O efeito na resposta média do aumento de 1000 dólares em X_1 (1 unidade de aumento, X_1 está em milhares) quando X_2 é constante, é de 20000 dólares.

O efeito na resposta média do aumento de 1000 dólares em X_2 (100000 unidades de aumento, X_2 está em centavos) quando X_1 é constante, é de 20000 dólares.

Transformação de correlação evita este tipo de comparação equivocada.

Transformação de correlação

Padronização usual:

$$\frac{Y_i - \bar{Y}}{s_Y}$$

$$\frac{X_{ik} - \bar{X}_k}{s_k}, \quad k = 1, 2, \ldots, p-1$$

em que:

$$s_Y = \sqrt{\frac{\sum_i (Y_i - \bar{Y})^2}{n-1}}$$

$$s_k = \sqrt{\frac{\sum_i (X_{ik} - \bar{X}_k)^2}{n-1}}, \quad k = 1, 2, \ldots, p-1$$

Transformação de correlação

A transformação de correlação é uma função das variáveis padronizadas:

$$Y_i^* = \frac{1}{\sqrt{n-1}} \left(\frac{Y_i - \bar{Y}}{s_Y} \right)$$

$$X_{ik}^* = \frac{1}{\sqrt{n-1}} \left(\frac{X_{ik} - \bar{X}_k}{s_k} \right), \quad k = 1, 2, \ldots, p-1$$

Modelo de Regressão Padronizado

$$Y_i^* = \beta_1^* X_{i1}^* + \ldots + \beta_{p-1}^* X_{i,p-1}^* + \varepsilon_i^*$$

Relação com modelo de regressão múltipla usual:

$$\beta_k = \left(\frac{s_Y}{s_k} \right) \beta_k^*, \quad k = 1, 2, \ldots, p-1$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}_1 - \ldots - \beta_{p-1} \bar{X}_{p-1}$$

Modelo de Regressão Padronizado

$$\mathbf{X}^*_{n \times p-1} = \begin{pmatrix} X_{11}^* & X_{12}^* & \ldots & X_{1,p-1}^* \\ X_{21}^* & X_{22}^* & \ldots & X_{2,p-1}^* \\ \vdots & \vdots & \vdots & \\ X_{n1}^* & X_{n2}^* & \ldots & X_{n,p-1}^* \end{pmatrix}$$

Seja a matriz de correlação de $\mathbf{X}$:

$$r_{XX\, p-1 \times p-1} = \begin{pmatrix} 1 & r_{12} & \ldots & r_{1,p-1} \\ r_{21} & 1 & \ldots & r_{2,p-1} \\ \vdots & \vdots & \vdots & \\ r_{p-1,1} & r_{p-1,2} & \ldots & 1 \end{pmatrix}$$

Modelo de Regressão Padronizado

em que r_{jk} é o coeficiente de correlação entre X_j e X_k.

$$\begin{aligned}
\sum X_{ij}^* X_{ik}^* &= \sum \left[\frac{1}{\sqrt{n-1}} \left(\frac{X_{ij} - \bar{X}_j}{s_j} \right) \right] \frac{1}{\sqrt{n-1}} \left(\frac{X_{ik} - \bar{X}_k}{s_k} \right) \\
&= \frac{1}{n-1} \frac{\sum (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)}{s_j s_k} \\
&= \frac{\sum (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)}{\sqrt{\sum (X_{ij} - \bar{X}_j)^2 \sum (X_{ik} - \bar{X}_k)^2}} \\
&= r_{jk}
\end{aligned}$$

Modelo de Regressão Padronizado

Portanto, temos que:

$$\mathbf{X}^{*T} \mathbf{X}^* = r_{XX} \, .$$

De maneira similar:

$$\mathbf{X}^{*T} \mathbf{Y}^*_{p-1 \times 1} = r_{YX}$$

em que r_{YX} é o vetor de correlações entre $\mathbf{Y}$ e cada coluna de $\mathbf{X}$.

Modelo de Regressão Padronizado

Equações normais:

$$\mathbf{X}^{*T} \mathbf{X}^* \hat{\boldsymbol{\beta}}^* = \mathbf{X}^{*T} \mathbf{Y}$$

Estimador de mínimos quadrados:

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}^{*T} \mathbf{X}^*)^{-1} \mathbf{X}^{*T} \mathbf{Y}$$

Equivalentemente:

$$\hat{\boldsymbol{\beta}}^* = r_{XX}^{-1} r_{YX} \, .$$

Exemplo

```
ds = read.csv("http://www.math.smith.edu/r/data/help.csv")
female = subset(ds, female==1)

lm1 = lm(pcs ~ mcs + homeless, data=female)

summary(lm1)$coefficients
```

```
##               Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 39.6261938 2.49829796 15.861276 1.595217e-29
## mcs          0.2194469 0.07643551  2.871007 4.958451e-03
## homeless    -2.5690667 1.95078674 -1.316939 1.907536e-01
```

Exemplo

```
library(QuantPsyc)
lm.beta
```

```
## function (MOD)
## {
##     b <- summary(MOD)$coef[-1, 1]
##     sx <- sapply(MOD$model[-1], sd)
##     sy <- sapply(MOD$model[1], sd)
##     beta <- b * sx/sy
##     return(beta)
## }
## <environment: namespace:QuantPsyc>
```


Exemplo

```
lm.beta(lm1)
```

```
##        mcs   homeless
##  0.2691888 -0.1234776
```

Uma mudança de 1 desvio-padrão em `mcs` tem mais do que o dobro de impacto de uma mudança de 1 desvio-padrão em `homeless`.

Exemplo: Dwaine Studios

Y: vendas

X_1: população

X_2: renda per capita

```
dados <- read.table("./dados/CH07TA05.txt")
colnames(dados) <- c("Y","X1","X2")
dados
```

```
##        Y   X1   X2
## 1  174.4 68.5 16.7
## 2  164.4 45.2 16.8
## 3  244.2 91.3 18.2
## 4  154.6 47.8 16.3
## 5  181.6 46.9 17.3
## 6  207.5 66.1 18.2
## 7  152.8 49.5 15.9
## 8  163.2 52.0 17.2
## 9  145.4 48.9 16.6
```


Exemplo: Dwaine Studios

Modelo usual, sem padronização:

```
modelo <- lm(Y ~ X1+X2,data=dados)
summary(modelo)$coefficients
```

```
##               Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) -68.85707 60.0169532 -1.147294 2.662817e-01
## X1            1.45456  0.2117817  6.868201 2.001691e-06
## X2            9.36550  4.0639581  2.304527 3.332136e-02
```

Exemplo: Dwaine Studios

Modelo padronizado:

```
dadosPadrao <- as.data.frame(scale(dados)/sqrt(dim(dados)[1]-1))
modeloPadrao <- lm(Y ~ X1+X2-1,data=dadosPadrao)
summary(modeloPadrao)$coefficients
```

```
##     Estimate Std. Error  t value     Pr(>|t|)
## X1 0.7483670   0.106055 7.056406 1.025522e-06
## X2 0.2511039   0.106055 2.367676 2.866468e-02
```


Exemplo: Dwaine Studios

Ou, diretamente, pelo comando:

```
lm.beta(modelo)
```

```
##        X1        X2
## 0.7483670 0.2511039
```

Note que o comando apenas libera as estimativas (sem erro-padrão, testes, etc…)

Leitura

- Applied Linear Statistical Models: Seções 7.1-7.5.
- Draper & Smith - Applied Regression Analysis: Capítulo 6.
- Weisberg - Applied Linear Regression: Seções 6.1-6.3
- Faraway - Linear Models with R: Seções 3.1 e 3.2.