ME613 - Análise de Regressão

Parte 8

Benilton S Carvalho e Rafael P Maia - 2S2020

Multicolinearidade

Video anterior

- Sempre teremos algum grau multicolinearidade entre os dados
- Alteração nas estimativas dos parâmetros
- Menor soma extra de quadrados de regressão
- Aumento dos erros padrões das estimativas do parâmetros

Introdução

Vamos considerar o modelo de regressão linear múltipla:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \;\; i = 1, \ldots, n, \;\; \epsilon_i \; iid \; \sim N(0, \sigma^2)$$

em que X_1 e X_2 são variáveis tranformadas via transformação de correlação (ou seja estão centradas no 0 e limitadas entre -1 e 1).

A matrix de desenho do modelo é data por

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \\ \vdots & \vdots \\ X_{n1} & X_{n2} \end{pmatrix} \;\; e \;\; \mathbf{X}^T\mathbf{X} = \begin{pmatrix} \sum X_{i1}^2 & \sum X_{i1}X_{i2} \\ \sum X_{i1}X_{i2} & \sum X_{i2}^2 \end{pmatrix} = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}$$

Daí temos que

$$Var(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} \frac{\sigma^2}{1-r_{12}^2} & -\frac{\sigma^2 r_{12}}{1-r_{12}^2} \\ -\frac{\sigma^2 r_{12}}{1-r_{12}^2} & \frac{\sigma^2}{1-r_{12}^2} \end{pmatrix}$$

Portanto quanto forte a correlação entre X_1 e X_2 maior as $Varbe\hat{t}a_1$ e $Varbe\hat{t}a_1$.

Efeito nos coeficientes de regressão

X_1: tríceps

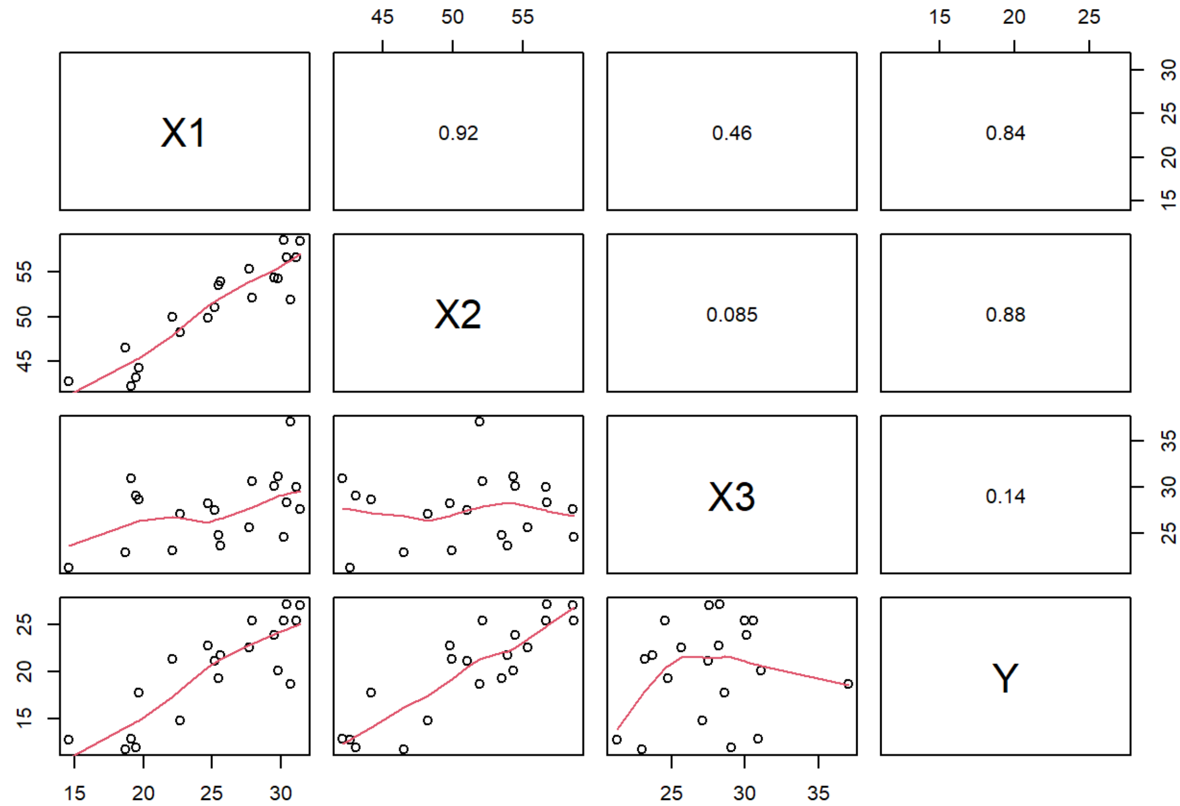
X_2: coxa

X_3: antebraço

Y: gordura corporal

```
##      X1   X2   X3    Y
## 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
## 4  29.8 54.3 31.1 20.1
## 5  19.1 42.2 30.9 12.9
## 6  25.6 53.9 23.7 21.7
## 7  31.4 58.5 27.6 27.1
## 8  27.9 52.1 30.6 25.4
## 9  22.1 49.9 23.2 21.3
## 10 25.5 53.5 24.8 19.3
## 11 31.1 56.6 30.0 25.4
## 12 30.4 56.7 28.3 27.2
## 13 18.7 46.5 23.0 11.7
## 14 19.7 44.2 28.6 17.8
```

Exemplo



Efeito no desvio-padrão da estimativa

Variável no modelo	$\hat{\beta}_1$	$\hat{\beta}_2$
X_1	0.129	
X_2		0.11
X_1, X_2	0.303	0.291
X_1, X_2, X_3	3.016	2.582

Efeito nos valores ajustados e preditos

Variável no modelo	QME
X_1	7.95
X_1, X_2	6.47
X_1, X_2, X_3	6.15

QME diminui conforme variáveis são adicionadas ao modelo (caso usual).

Efeito nos valores ajustados e preditos

A precisão do valor ajustado não é tão afetada quando inserimos ou não uma variável preditora muito correlacionada com outra já no modelo.

Por exemplo, se considerarmos apenas o modelo com X_1, o valor estimado de gorduta corporal para $X_1 = 25$ é:

$$\hat{Y} = 19.934 \qquad \sqrt{\widehat{Var(\hat{Y})}} = 0.632$$

Quando incluímos X_2, altamente correlacionada à X_1, temos:

$$\hat{Y} = 19.356 \qquad \sqrt{\widehat{Var(\hat{Y})}} = 0.624$$

quando $X_1 = 25$ e $X_2 = 50$, por exemplo.

Efeito nos testes simultâneos de β_k

Considere os dados sobre gordura corporal e o modelo com X_1 e X_2 no modelo.

Queremos testar H_0: $\beta_1 = \beta_2 = 0$.

Calculamos:

$$t_1 = \frac{\hat{\beta}_1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \qquad t_2 = \frac{\hat{\beta}_2}{\sqrt{\widehat{Var}(\hat{\beta}_2)}}$$

e não rejeitamos H_0 se ambos $|t_1|$ e $|t_2|$ forem menores do que $t_{n-3,\alpha/4} = 2.46$

para $\alpha = 0.05$.

Exemplo

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9469 -1.8807  0.1678  1.3367  4.0147
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.1742     8.3606  -2.293   0.0348 *
## X1            0.2224     0.3034   0.733   0.4737
## X2            0.6594     0.2912   2.265   0.0369 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.543 on 17 degrees of freedom
## Multiple R-squared:  0.7781, Adjusted R-squared:  0.7519
## F-statistic:  29.8 on 2 and 17 DF,  p-value: 2.774e-06
```

Não rejeitamos H_0.

Exemplo

```
## 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
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ 1
## Model 2: Y ~ X1 + X2
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     19 495.39
## 2     17 109.95  2    385.44 29.797 2.774e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exemplo

Se utilizarmos o teste F para $H_0 : \beta_1 = \beta_2 = 0$, temos:

$$F_{obs} = \frac{QMReg}{QME} = \frac{385.44/2}{109.95/17} = 29.8$$

Sob H_0 a estatística do teste tem distribuição $F(2, 17)$, de maneira que o valor crítico para $\alpha = 0.05$ é 3.59.

Encontramos evidências para rejeitar H_0.

Resultado contrário ao obtido com os testes t com correção de Bonferroni.

Como lidar com a multicolinearidade

- Coletar mais dados
- Eliminação de variáveis explicativas
- Transformação de variáveis explicativas
- Componentes principais

Agradecimento

- Slides criados por Samara F Kiihl / IMECC / UNICAMP

Leitura

- Applied Linear Statistical Models: Seção 7.6.
- Faraway - Linear Models with R: Seção 7.3.