ME613 - Análise de Regressão

Parte 10

Benilton S Carvalho e Rafael P Maia- 2S2020

Região de Confiança

Recordar é viver…

Modelo de regressão linear multipla

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \quad i = 1, \ldots, n$$

Notação matricial

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

$$\mathbf{Y}_{n\times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \mathbf{X}_{n\times p} = \begin{pmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \ldots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & X_{n1} & X_{n2} & \ldots & X_{n,p-1} \end{pmatrix} \boldsymbol{\beta}_{p\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} \boldsymbol{\varepsilon}_{n\times 1} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Estimador de Mínimos Quadrados

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

Recordar é viver…

$$\varepsilon_i \overset{iid}{\sim} \mathbf{N}(0, \sigma^2),\ i = 1, \ldots, n \equiv \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

E daí temos que

$$\hat{\boldsymbol{\beta}}_{p\times 1} \sim \mathcal{N}_p(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$

Um intervalo de $100(1-\alpha)\%$ de confiança para β_k é dado por:

$$IC(\beta_k, 1-\alpha) = \left[\hat{\beta}_k - t_{n-p,\alpha/2}\sqrt{\widehat{Var(\hat{\beta}_k)}}; \hat{\beta}_k + t_{n-p,\alpha/2}\sqrt{\widehat{Var(\hat{\beta}_k)}}\right]$$

Exemplo

Estudo sobre diversidade das espécies em Galápagos.

Conjunto de dados: 30 ilhas, 7 variáveis.

`Species`: the number of plant species found on the island

`Endemics`: the number of endemic species

`Area`: the area of the island (km^2)

`Elevation`: the highest elevation of the island (m)

`Nearest`: the distance from the nearest island (km)

`Scruz`: the distance from Santa Cruz island (km)

`Adjacent`: the area of the adjacent island (square km)

Exemplo

```r
library(faraway)
lmod <- lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent, gala)
sumary(lmod)
```

```
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  7.068221  19.154198  0.3690 0.7153508
## Area        -0.023938   0.022422 -1.0676 0.2963180
## Elevation    0.319465   0.053663  5.9532 3.823e-06
## Nearest      0.009144   1.054136  0.0087 0.9931506
## Scruz       -0.240524   0.215402 -1.1166 0.2752082
## Adjacent    -0.074805   0.017700 -4.2262 0.0002971
##
## n = 30, p = 6, Residual SE = 60.97519, R-Squared = 0.77
```

Exemplo

IC 95% para $\beta_{Adjacent}$:

```
confint(lmod)[6,]
```

```
##       2.5 %      97.5 %
## -0.11133622 -0.03827344
```

```
confint(lmod)
```

```
##                   2.5 %      97.5 %
## (Intercept) -32.4641006 46.60054205
## Area         -0.0702158  0.02233912
## Elevation     0.2087102  0.43021935
## Nearest      -2.1664857  2.18477363
## Scruz        -0.6850926  0.20404416
## Adjacent     -0.1113362 -0.03827344
```

Região de Confiança para β

Propriedades

$$\hat{\boldsymbol{\beta}}_{p\times 1} \sim \mathcal{N}_p(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$

$$\frac{1}{\sigma}(\mathbf{X}^T\mathbf{X})^{1/2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$$

$$\frac{1}{\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \sim \chi^2(p)$$

$$(n-p)\frac{s^2}{\sigma^2} \sim \chi^2(n-p)$$

Portanto:

$$\frac{1}{ps^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T\mathbf{X}^T\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \sim F(p, n-p)$$

Região de Confiança para β

Região de Confiança de $100 \times (1-\alpha)\%$:

$$\frac{1}{ps^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \mathbf{X}^T \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq F(p, n-p; 1-\alpha)$$

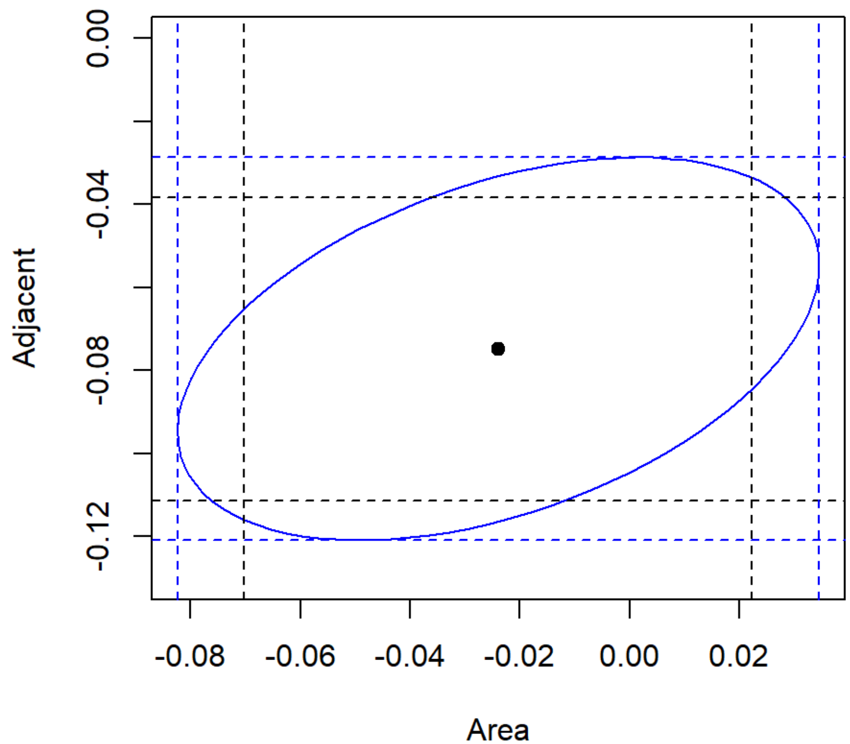
Exemplo

RC 95% para $\beta_{Adjacent}$ e β_{Area}

```
library(ellipse)
aa <- ellipse(lmod,which=c(2,6),level=0.95)
plot(aa,type="l",ylim=c(-0.13,0),col="blue")
points(coef(lmod)[2], coef(lmod)[6], pch=19)
abline(v=confint(lmod)[2,],lty=2)
abline(h=confint(lmod)[6,],lty=2)
abline(h=c(max(aa[,2]),min(aa[,2])),lty=2,col="blue")
abline(v=c(max(aa[,1]),min(aa[,1])),lty=2,col="blue")
```

Exemplo

RC 95% para $\beta_{Adjacent}$ e β_{Area}



Teste de Hipótese Linear

Teste de Hipótese Linear

Teste de hipótese linear:

$H_0\colon \mathbf{R}_{r\times p}\boldsymbol{\beta}_{p\times 1} = \mathbf{q}_{r\times 1}$

$H_1\colon \mathbf{R}_{r\times p}\boldsymbol{\beta}_{p\times 1} \neq \mathbf{q}_{r\times 1}$

Para testar, começamos pensando no vetor de discrepância com relação à H_0:

$$\mathbf{R}_{r\times p}\hat{\boldsymbol{\beta}}_{p\times 1} - \mathbf{q}_{r\times 1} = \mathbf{m}_{r\times 1}$$

queremos medir quão longe $\mathbf{m}$ está de $\mathbf{0}$.

Teste de Wald para Hipótese Linear

Precisamos então conhecer a distribuição de $\mathbf{m}$, sob H_0:

$$E(\mathbf{m}) = E(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}) = \mathbf{R}E(\hat{\boldsymbol{\beta}}) - \mathbf{q} = \mathbf{R}\boldsymbol{\beta} - \mathbf{q} \overset{H_0}{=} \mathbf{0}$$

$$Var(\mathbf{m}) = Var(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}) = \mathbf{R}Var(\hat{\boldsymbol{\beta}})\mathbf{R}^T = \sigma^2\mathbf{R}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{R}^T$$

Estatística do teste:

$$\begin{aligned} W &= \mathbf{m}^T[Var(\mathbf{m})]^{-1}\mathbf{m} \\ &= (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})^T[\sigma^2\mathbf{R}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{R}^T]^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}) \\ &\overset{H_0}{\sim} \chi^2(r) \end{aligned}$$

Teste de Wald para Hipótese Linear

Problema: não conhecemos σ^2. Temos que utilizar um estimador para σ^2: s^2.

Sabemos que $(n-p)\frac{s^2}{\sigma^2} \sim \chi^2(n-p)$.

A estatística do teste é:

$$
\begin{aligned}
F &= \frac{W}{r}\frac{\sigma^2}{s^2} \\
&= \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})^T[\mathbf{R}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{R}^T]^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})}{r\sigma^2}\frac{\sigma^2}{s^2} \\
&= \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})^T[s^2\mathbf{R}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{R}^T]^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})}{r} \\
&\overset{H_0}{\sim} F_{r,n-p}
\end{aligned}
$$

Exemplo

- Para H_0: $\beta_j = 0$, definimos

$$\mathbf{R} = [0 \quad 0 \quad \ldots \quad 1 \quad 0 \quad \ldots \quad 0]$$ e $\mathbf{q} = 0$.

- Para H_0: $\beta_k = \beta_j$, definimos

$$\mathbf{R} = [0 \quad 0 \quad 1 \quad \ldots \quad -1 \quad 0 \quad \ldots \quad 0]$$ e $\mathbf{q} = 0$.

- Para H_0: $\beta_1 + \beta_2 + \beta_3 = 1$, definimos

$$\mathbf{R} = [0 \quad 1 \quad 1 \quad 1 \quad 0 \quad \ldots \quad 0]$$ e $\mathbf{q} = 1$.

Exemplo

- Para H_0: $\beta_0 = 0$, $\beta_1 = 0$ e $\beta_2 = 0$, definimos

$$\mathbf{R} = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \end{pmatrix} \qquad \mathbf{q} = \mathbf{0}$$

- Para H_0: $\beta_1 + \beta_2 = 1$, $\beta_3 + \beta_5 = 0$ e $\beta_4 + \beta_5 = 0$, definimos

$$\mathbf{R} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \qquad \mathbf{q} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

Exemplo

```r
library(systemfit)
data( "Kmenta" )
modelo <- lm(consump ~ price + income,data=Kmenta)
summary(modelo)$coef
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 99.8954229 7.51936214 13.285093 2.090605e-10
## price       -0.3162988 0.09067741 -3.488177 2.815290e-03
## income       0.3346356 0.04542183  7.367285 1.099860e-06
```

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

X_1: price

X_2: income

Exemplo

$H_0 \colon \beta_1 = 0$

```
R <- matrix( c(0,1,0),ncol=length(coef(modelo)),byrow=TRUE)
linearHypothesis(modelo,R)
```

```
## Linear hypothesis test
##
## Hypothesis:
## price = 0
##
## Model 1: restricted model
## Model 2: consump ~ price + income
##
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     18 108.660
## 2     17  63.332  1    45.328 12.167 0.002815 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exemplo

H_0: $\beta_1 = 2$

```
R <- matrix( c(0,1,0),ncol=length(coef(modelo)),byrow=TRUE)
q=2
linearHypothesis(modelo,R,q)
```

```
## Linear hypothesis test
##
## Hypothesis:
## price = 2
##
## Model 1: restricted model
## Model 2: consump ~ price + income
##
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     18 2494.21
## 2     17   63.33  1    2430.9 652.52 5.311e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exemplo

H_0: $\beta_1 = 2$ e $\beta_2 = 1$.

```
R <- matrix( c(0,1,0,
               0,0,1),ncol=length(coef(modelo)),byrow=TRUE)
q=matrix(c(2,1),ncol=1)
linearHypothesis(modelo,R,q)
```

```
## Linear hypothesis test
##
## Hypothesis:
## price = 2
## income = 1
##
## Model 1: restricted model
## Model 2: consump ~ price + income
##
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1     19 7146.8
## 2     17   63.3  2    7083.4 950.7 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Agradecimento

- Slides criados por Samara F Kiihl / IMECC / UNICAMP
- Editado por Rafael P Maia / IMECC / UNICAMP

Leituras

- The Matrix Cookbook
- Faraway - Linear Models with R: Seção 3.5.
- Draper & Smith - Applied Regression Analysis: Seção 9.1.