Análise de Regressão

Davi Wentrick Feijó

2023-04-17

Exemplo 1 A taxa de metabolismo é importante em estudos sobre aumento de peso, dieta e exercício. Em um estudo com 19 indivíduos selecionados aleatoriamente entre os submetidos a um estudo de dieta, foram coletados dados sobre a massa do corpo sem gordura e a taxa metabólica em repouso. A massa do corpo sem gordura é o peso da pessoa, eliminada toda a gordura, e é dada em quilogramas. A taxa de metabolismo é medida em calorias queimadas a cada 24 horas e os pesquisadores acham que a massa do corpo sem gordura tem grande influência sobre ela.

massa	taxa
62.0	1792
62.9	1666
36.1	995
54.6	1425
48.5	1396
42.0	1418
47.4	1362
50.6	1502
42.0	1256
48.7	1614
40.3	1189
33.1	913
51.9	1460
42.4	1124
34.5	1052
51.1	1347
41.2	1204
51.9	1867
46.9	1439

Calcule o estimadores abaixo manualmente (com arredondamentos):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Os valores encontrados em sala foram os seguintes:

$$\hat{\beta}_0 = 117,44 \qquad\qquad \hat{\beta}_1 = 26,79$$

Com os coeficientes obtidos a nossa equacao ficou assim:

$$\hat{Y} = 117,44 + 26,79x$$

Aplicando a equacao encontrada nos dados podemos obter o valor estimado (sabendo que x = massa)

massa	taxa	estimado
62.0	1792	1778.420
62.9	1666	1802.531
36.1	995	1084.559
54.6	1425	1580.174
48.5	1396	1416.755
42.0	1418	1242.620
47.4	1362	1387.286
50.6	1502	1473.014
42.0	1256	1242.620
48.7	1614	1422.113
40.3	1189	1197.077
33.1	913	1004.189
51.9	1460	1507.841
42.4	1124	1253.336
34.5	1052	1041.695
51.1	1347	1486.409
41.2	1204	1221.188
51.9	1867	1507.841
46.9	1439	1373.891

Para verificar se esse valor estimado está condizente podemos calcular os residuos e soma-los, devemos encontrar que a soma dos reisudos é 0. Dado que a forma de se calcular os residuos segue essa funcao:

$$e_i = y_i - \hat{y_i}$$

onde e_i é o resíduo para a i-ésima observação, y_i é o valor observado da variável dependente para a i-ésima observação e $\hat{Y_i}$ é o valor previsto pela reta de regressão para a i-ésima observação.

massa	taxa	estimado	residuo
62.0	1792	1778.420	13.580
62.9	1666	1802.531	-136.531
36.1	995	1084.559	-89.559
54.6	1425	1580.174	-155.174
48.5	1396	1416.755	-20.755
42.0	1418	1242.620	175.380
47.4	1362	1387.286	-25.286
50.6	1502	1473.014	28.986
42.0	1256	1242.620	13.380
48.7	1614	1422.113	191.887
40.3	1189	1197.077	-8.077
33.1	913	1004.189	-91.189
51.9	1460	1507.841	-47.841
42.4	1124	1253.336	-129.336
34.5	1052	1041.695	10.305
51.1	1347	1486.409	-139.409
41.2	1204	1221.188	-17.188
51.9	1867	1507.841	359.159
46.9	1439	1373.891	65.109

```
erro
```

```
## [1] -2.559
```

$$\sum_{i=1}^{n} y_i - \hat{y_i} = -2,559$$

Calculando os parametros β_1 e β_0 computacionalmente: Com o resultado anterior podemos perceber que devido aos arredondamentos feitos em sala, a soma dos residuos nao deu zero! Agora vamos calcular β_1 e β_0 por meio do R. Fazendo os calculos necessarios temos que:

$$\bar{Y} = 1.369,526 \qquad \bar{X} = 46,74211$$

$$\sum_{n=1}^{n} X_i = 888,1 \qquad \sum_{n=1}^{n} Y_i = 26.021$$

$$\sum_{n=1}^{n} X_i^2 = 42747,03 \qquad \sum_{n=1}^{n} Y_i^2 = 36.829.995 \qquad \sum_{n=1}^{n} X_i Y_i = 1.249.481$$

Com esses resultados podemos aplicar na formual dos estimadores de β_1 e β_0

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

e assim encontramos:

$$\hat{\beta}_1 = \frac{1.249.481 - 19 \cdot 46,74211 \cdot 1.369,526}{42747,03 - 19 \cdot 46,74211^2} = 26,87857 \qquad \hat{\beta}_0 = 1.369,526 - 26.87857 \cdot 46,74211 = 113,1654$$

Com os novos coeficientes obtidos de β_1 e β_0 a nossa equacao ficou assim:

$$\hat{Y} = 113,1654 + 26,87857x$$

Agora podemos recalcular o valor estimado e depois fazer o calculo dos residuos.

Massa	Taxa	Novo estimado	Novo residuo
62.0	1792	1779.637	12.363305
62.9	1666	1803.827	-137.827407
36.1	995	1083.482	-88.481753
54.6	1425	1580.735	-155.735283
48.5	1396	1416.776	-20.776011
42.0	1418	1242.065	175.934688
47.4	1362	1387.210	-25.209585
50.6	1502	1473.221	28.778994
42.0	1256	1242.065	13.934689
48.7	1614	1422.152	191.848275
40.3	1189	1196.372	-7.371744
33.1	913	1002.846	-89.846046
51.9	1460	1508.163	-48.163146
42.4	1124	1252.817	-128.816739
34.5	1052	1040.476	11.523957
51.1	1347	1486.660	-139.660291
41.2	1204	1220.562	-16.562456
51.9	1867	1508.163	358.836854
46.9	1439	1373.770	65.229700

realizando o calculo da soma de reisiduos novamente podemos verificar que dessa vez devido a maior precisao atingimos o 0

```
erro
```

```
## [1] -1.250555e-12
```

```
round(erro, 4)
```

```
## [1] 0
```

Podemos rodar o modelo completo no R para ter nocao dos resultado!

```
lm(dados$taxa ~ dados$massa) %>%
    summary()
```

```
##
## Call:
## lm(formula = dados$taxa ~ dados$massa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -155.74  -89.16  -16.56   21.36  358.84
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  113.165    179.587   0.630    0.537
## dados$massa   26.879      3.786   7.099 1.78e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 133.1 on 17 degrees of freedom
## Multiple R-squared:  0.7478, Adjusted R-squared:  0.7329
## F-statistic:  50.4 on 1 and 17 DF,  p-value: 1.784e-06
```

Calcule o estimador da variancia dos erros : Podemos estimar o σ^2 da variancia dos erros por meio da seguinte formula:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

Realizando os calculos no R temos:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{301.051,1}{19-2} = \frac{301.051,1}{17} = 17.708,89 \qquad \hat{\sigma} = \sqrt{\hat{\sigma}^2} = 133,0747$$

Podemos calcular o erro padrao por meio da seguinte formula:

$$SE = \frac{s}{\sqrt{n}}$$

onde “s” representa o desvio padrão da amostra e “n” é o tamanho da amostra. O símbolo “SE” é usado para indicar o erro padrão.

Realizando os calculos obtemos o seguinte valor:

$$SE = \frac{s}{\sqrt{n}} = \frac{133,0747}{\sqrt{19}} = 30,52944$$

para ter uma nocao da variabilidade pode calcular o coeficiente de variação (CV) que é uma medida relativa de variabilidade que é frequentemente usada para comparar a variabilidade relativa entre diferentes conjuntos de dados. É definido como o desvio padrão dividido pela média, expresso como uma porcentagem.

$$CV = \frac{s}{\overline{x}} \times 100\%$$

onde “s” representa o desvio padrão da amostra e $\overline{x}$ representa a média da amostra. O símbolo “%” indica que o resultado é expresso como uma porcentagem.

Utilizando o R para o calculo obtemos o seguinte resultado:

$$CV = \frac{s}{\overline{x}} \times 100 = \frac{133,0747}{1.369,526} \times 100 = 9,716845\%$$

Calculando o estimador da variancia de β_1 e β_0: As formulas para o calculo sao as seguintes:

$$s^2(\beta_0) = V(\beta_0) = \hat{\sigma}^2[\frac{\sum_{i=1}^{n} X_i^2}{n(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2)}]$$

$$s^2(\beta_1) = V(\beta_1) = \frac{\hat{\sigma}^2}{(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2)}$$

Como já temos esses valores calculados é só substituir na formula!

$$s^2(\beta_1) = \frac{17.708,89}{(42.747,03 - 19 \cdot 46,74211^2)} = 14,33493 \qquad s(\beta_1) = \sqrt{s^2(\beta_1)} = \sqrt{14,33493} = 3,78615$$

$$s^2(\beta_0) = 17.708,89[\frac{42.747,03}{19 \cdot (42.747,03 - 19 \cdot 46,74211^2)}] = 32.251,35 \quad s(\beta_0) = \sqrt{s^2(\beta_0)} = \sqrt{32.251,35} = 179,5866$$

Vamos calcular os Coeficientes de variacao dos valores estimados da variancia do erro.

$$CV\beta_1 = \frac{s(\beta_1)}{\beta_1} \times 100 = \frac{3,78615}{26,87857} \times 100 = 14,08613\%$$
$$CV\beta_0 = \frac{s(\beta_0)}{\beta_0} \times 100 = \frac{179,5866}{113,1654} \times 100 = 158,6939\%$$

Vale notar que a variancia de β_0 é maior pois incorpara a variancia de β_1 (no caso ele representa somente o intercepto entao nao faz muita diferenca nesse exemplo em especifico)

Vamos calcular um intervalo de confianca para β_0 e β_1: Sabemos que a estatistica :

$$\frac{\hat{\beta}_1 - \beta_1}{S(\hat{\beta}_1)} \sim t(n-2)$$

$$IC = P(-t_{1-\alpha/2} < \frac{\hat{\beta}_1 - \beta_1}{S(\beta_1)} < t_{1-\alpha/2})$$

Isolando o β_1 obtemos:

$$IC = P(\hat{\beta}_1 - t_{1-\alpha/2} \cdot S(\hat{\beta}_1) < \beta_1 < \hat{\beta}_1 + t_{1-\alpha/2} \cdot S(\hat{\beta}_1)) = \hat{\beta}_1 \pm t_{1-\alpha/2} \cdot S(\hat{\beta}_1)$$

Sabendo que $S(\hat{\beta}_1)$ é igual a $\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}}$

Chegamos na formula final para o intervalo de confianca:

$$IC = \hat{\beta}_1 \pm t_{1-\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}})$$

Seguindo a mesma deducao e sabendo que $S(\hat{\beta}_0) = \frac{\hat{\sigma} \cdot \sum_{i=1}^{n} X_i^2}{\sqrt{n(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2)}}$ obtemos:

$$IC = \hat{\beta}_0 \pm t_{1-\alpha/2} \cdot \frac{\hat{\sigma} \cdot \sum_{i=1}^{n} X_i^2}{\sqrt{n(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2)}})$$

Fazendo os calculos no R obtemos os seguintes intervaloes para β_0 e β_1:

```
# estimando intervalo de confianca para beta 1 com 95% de confianca
alfa = 0.05

ic_beta_1 = c(beta_1 - qt(alfa/2, n - 2, lower.tail = FALSE) * s_beta_1,
    beta_1 + qt(alfa/2, n - 2, lower.tail = FALSE) * s_beta_1)

# estimando intervalo de confianca para beta 0 com 95% de confianca
alfa = 0.05

ic_beta_0 = c(beta_0 - qt(alfa/2, n - 2, lower.tail = FALSE) * s_beta_0,
    beta_0 + qt((alfa/2), n - 2, lower.tail = FALSE) * s_beta_0)
```

```
ic_beta_1
```

```
## [1] 18.89049 34.86665
```

```
ic_beta_0
```

```
## [1] -265.7292  492.0600
```

Podemos verificar isso por meio da funcao lm() que roda a regressao por completa!

```r
fit <- lm(dados$taxa ~ dados$massa)

# Cálculo do intervalo de confiança para beta1 com nível de
# significância de 0.05
alpha <- 0.05
se_b1 <- summary(fit)$coefficients[2, 2]
beta1 <- summary(fit)$coefficients[2, 1]
t_crit <- qt(1 - alpha/2, df = fit$df.residual)
lower_b1 <- beta1 - t_crit * se_b1
upper_b1 <- beta1 + t_crit * se_b1
cat("Intervalo de confiança para beta1: [", lower_b1, ", ", upper_b1,
    "]\n")
```

```
## Intervalo de confiança para beta1: [ 18.89049 ,  34.86665 ]
```

```r
# Cálculo do intervalo de confiança para beta0 com nível de
# significância de 0.05
alpha <- 0.05
se_b0 <- summary(fit)$coefficients[1, 2]
beta0 <- summary(fit)$coefficients[1, 1]
t_crit <- qt(1 - alpha/2, df = fit$df.residual)
lower_b0 <- beta0 - t_crit * se_b0
upper_b0 <- beta0 + t_crit * se_b0
cat("Intervalo de confiança para beta0: [", lower_b0, ", ", upper_b0,
    "]\n")
```

```
## Intervalo de confiança para beta0: [ -265.7292 ,  492.06 ]
```

Vamos formular os teste de hipoteses sobre β_0 e β_1

$$H_0 : \beta_1 = 0 \quad \sim \quad H_0 : \text{Ausencia de regressao}$$

$$H_1 : \beta_1 \neq 0 \quad \sim \quad H_1 : \text{Existe de regressao}$$

A estatistica do teste é dado por:

$$T = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} : \text{distribuição de Student com (n-2) g.l}$$

```
# Ho: beta_1 = 0 HO: beta_1 > 0 (ou != 0)
alfa = 0.05
teste_stat = beta_1/s_beta_1

rc = qt(alfa, n - 2)  #regiao critica

pvalor = pt(teste_stat, n - 2, lower.tail = FALSE)  #pvalor

result = pvalor < alfa  #pvalor menor que o alfa de 5% logo rejeitamos HO
cat("O P-Valor observado foi:", pvalor, "O valor observado pertence a regiao crítica de",
    alfa, "?:", result)
```

```
## O P-Valor observado foi: 8.918394e-07 O valor observado pertence a regiao crítica de 0.05 ?: TRUE
```

Logo existe uma regressao ou seja β_1 contribui para a relacao da massa sem gordura e o metabolismo

O teste de hipotese sobre β_0 segue os mesmos parametros do teste anterior:

$$H_0 : \beta_0 = \beta$$

$$H_1 : \beta_0 \neq \beta$$

A estatistica do teste é dado por:

$$T = \frac{\hat{\beta}_0}{s(\hat{\beta}_0)} : \text{distribuição de Student com (n-2) g.l}$$

Esse teste busca testar um valor de β_0 contra algum outro valor de interesse

Intervalo de confianca sobre σ^2 Temos que:

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} : \text{distribuição de Qui-Quadrado com (n-2) g.l}$$

Fixada uma probabilidade de de $1-\alpha$ podemos determinar $\chi_{\alpha/2}$ e $\chi_{1-\alpha/2}$ para chegar na seguinte formula:

$$P(\chi_{\alpha/2} < \frac{(n-2)\hat{\sigma}^2}{\sigma^2} < \chi_{1-\alpha/2}) = 1-\alpha$$

$$P(\frac{(n-2)\hat{\sigma}^2}{\chi_{\alpha/2}} < \sigma^2 < \frac{(n-2)\hat{\sigma}^2}{\chi_{1-\alpha/2}}) = 1-\alpha$$

Obtemos o seguinte intervalo:

$$\sigma^2 \in (\frac{(n-2)\hat{\sigma}^2}{\chi_{\alpha/2}}, \frac{(n-2)\hat{\sigma}^2}{\chi_{1-\alpha/2}})$$

Relembrando que:

$$(n-2)\hat{\sigma}^2 = \sum_{i=1}^{n} e_i^2$$

Podemos chegar nesse novo intervalo:

$$\sigma^2 \in (\frac{\sum_{i=1}^{n} e_i^2}{\chi_{\alpha/2}}, \frac{\sum_{i=1}^{n} e_i^2}{\chi_{1-\alpha/2}})$$

```
gl = n - 2
alfa = 0.05
# intervalo usando a formula do primeiro intervalo encontrado
ic_sigma_quadrado = c(((n - 2) * sigma_quadrado/qchisq(alfa/2, gl, lower.tail = FALSE)),
    ((n - 2) * sigma_quadrado/qchisq(1 - alfa/2, gl, lower.tail = FALSE)))
cat("Intervalo de confiança para sigma quadrado: [", ic_sigma_quadrado[1],
    ", ", ic_sigma_quadrado[2], "]\n")
```

```
## Intervalo de confiança para sigma quadrado: [ 9971.548 ,  39799.53 ]
```

```
# intervalo usando a soma dos residuos ao quadrado
ic_sigma_quadrado = c(erro_quadrado/qchisq(alfa/2, gl, lower.tail = FALSE),
    erro_quadrado/qchisq(1 - (alfa/2), gl, lower.tail = FALSE))

cat("Intervalo de confiança para sigma quadrado: [", ic_sigma_quadrado[1],
    ", ", ic_sigma_quadrado[2], "]\n")
```

```
## Intervalo de confiança para sigma quadrado: [ 9971.548 ,  39799.53 ]
```