CC0288 - Inferência Estatística I

Exemplo de Regressão 14/04/2023.

Prof. Maurício

Exemplo do livro Estatística Básica de Daniel Furtado Ferreira (página 593)

1. Os dados referem-se a uma amostra de tamanho $n = 11$, na qual se aplicou CO_2 em diferentes concentrações em folhas de trigo (X) à temperatura de 35^oC, a quantidade de CO_2 absorvida (Y) em $cm^3/dm^2/hora$ foi avaliada.

   Os resultados estão apresentados a seguir.

Variável	1	2	3	4	5	6	7	8	9	10	11
X	75	100	100	120	130	130	160	190	200	240	250
Y	0,00	0,65	0,50	1,00	0,95	1,30	1,80	2,80	2,50	4,30	4,50

   a. Faça um diagrama de dispersão explicando quem é a variável explicativa e a variável resposta. Parece haver uma relação linear entre elas?

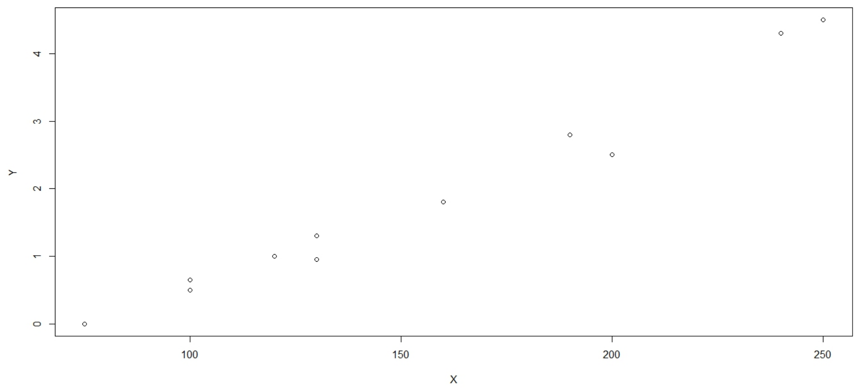

Figura 1:

   b. Calcule os seguintes somatórios:

$$SX = \sum_{i=1}^{n} X_i \ ; SY = \sum_{i=1}^{n} Y_i \ ; \ SXY = \sum_{i=1}^{n} X_i \, Y_i,$$

$$SX2 = \sum_{i=1}^{n} X_i^2 \ ; SY2 = \sum_{i=1}^{n} Y_i^2.$$

   c. O modelo de regressão linear é dado por:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \ i = 1, 2, \ldots, n,$$

   com os valores de bX fixados inicialmente e

$$E(\epsilon_i) = 0 \quad e \quad Var(\epsilon_i) = \sigma^2,\ i = 1, 2, \ldots, n.$$

Note que

$$E(Y_i) = \beta_0 + \beta_1 X_i, i = 1, 2, \ldots, n,$$

que é a nossa reta de regressão populacional.
Vamos estimar a reta de regressão:

$$\hat{Y}_i = b_0 + b_1 X_i$$

O coeficiente angular β_1 é estimado por:

$$b_1 = \frac{nSXY - SXSY}{nSX2 - SX^2}$$

O coeficiente linear β_0 é estimado por:

$$b_0 = \bar{Y} - b_1 \bar{X}.$$

Mostre que o modelo ajustado de regressão é:

$$\bar{Y}_i = -2,0759 + 0,0254 X_i.$$

d. Como estimar o parâmetro σ^2?
Seja

$$e_i = Y_i - \hat{Y}_i, , i = 1, 2, \ldots, n$$

a diferença entre o valor real observado Y_i e o valor predito pelo modelo de regressão $\hat{Y}_i$.
Vamos usar o estimador

$$S^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SQRes}{n-2},$$

que é um estimador não viciado para σ^2.
A soma de quadrados residual mede então a parte da variabilidade da variável resposta Y que não é explicada pelo modelo.
Considere

$$SQT = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

que é chamada de soma de quadrados total que nada mais é a soma dos quadrados dos erros cometida quando se ignora a importância da variável explicativa X no modelo.
Quando se incorpora a variável temos a nossa previsão $\hat{Y}_i$ e surge uma nova soma de quadrados

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

que mede a importância da variável explicativa no modelo ajustado.

Um fato relevante que no modelo vale:

$$SQT = SQReg + SQRes.$$

Note que:

$$SQT = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2.$$

$$SQReg = b_1\left[nSXY - nSXSY\right]$$

e

$$SQRes = SQT - SQReg,$$

que é a maneira mais fácil de calcular a soma de quadrados residual.

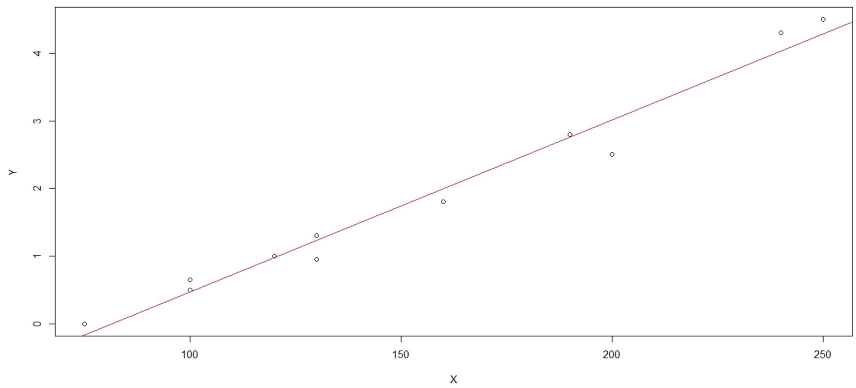

Figura 2:

Calcule as 3 somas de quadrados. Estime σ^2

e. Calcule o coeficiente de determinação do modelo.

$$R^2 = \frac{SQReg}{SQT} = 1 - \frac{SQRes}{SQT},$$

que mede a proporção da variabilidade da variável resposta Y que é explicada pela variável explicativa X. Quanto mais perto de um melhor o ajuste.!!!!

f. O incremento de uma unidade de gás carbônico prova um aumento médio de quanto na absorção pelas folhas à temperatura de 35^oC?

g. O valor da estimativa do coeficiente linear negativo tem algum sentido prático? Diante disso que modelo de regressão o estatístico deve usar?

h. Ajuste o modelo do item anterior.

```
>
>
> ###Exemplo do Daniel:
>
```

```
>
> i=1:11
>
> X=c(75,100,100,120,130,130,160,190,200,240,250)
> Y=c(0,65,50,100,95,130,180,280,250,430,450)/100
> n=length(X);n
[1] 11
>
>
>
> plot(X,Y)
> SX=sum(X);SX;SX2=sum(X^2);SX2;SXY=sum(X*Y);SXY;SY=sum(Y);SY;SY2=sum(Y^2);SY2
[1] 1695
[1] 295625
[1] 4004.5
[1] 20.3
[1] 60.335
>
> Xb=mean(X);Xb;SX/n
[1] 154.0909
[1] 154.0909
>
> Yb=mean(Y);Yb;SY/n
[1] 1.845455
[1] 1.845455
>
> num=n*SXY-SX*SY;num
[1] 9641
>
> den=n*SX2-SX^2;den
[1] 378850
>
> b_1=num/den;b_1
[1] 0.02544807
> round(b_1,4)
[1] 0.0254
>
> b_0=Yb-b_1*Xb;b_0
[1] -2.075861
>
> round(b_0,4)
[1] -2.0759
>
>
> #####Reta ajustada Yprev_i= -2,0759 +0,0254 X_i
>
> plot(X,Y)
> abline(c(b_0,b_1),col="red")
>
> ###Os valores previstos são dados por:
>
> Y_prev=b_0+b_1*X;Y_prev
```

```
[1] -0.1672562  0.4689455  0.4689455  0.9779068  1.2323875  1.2323875
[7]  1.9958295  2.7592715  3.0137521  4.0316748  4.2861555
>
>
>
> ###Os erros de previsão
>
> e=Y-Y_prev;e
[1]  0.16725617  0.18105451  0.03105451  0.02209318 -0.28238749  0.06761251
[7] -0.19582948  0.04072852 -0.51375214  0.26832519  0.21384453
>
>
> ### Soma de quadrados
>
> SQT=sum( (Y-mean(Y))^2);SQT
[1] 22.87227
> SQReg=b_1*(SXY -SX*SY/n);SQReg
[1] 22.30407
>
> SQRes=sum(e^2);SQRes;SQT-SQReg
[1] 0.5681992
[1] 0.5681992
>
>
>
>
> ###Estimativa de sigma^2
>
>
> S2=SQRes/(n-2);S2
[1] 0.06313324
>
> ####Coeficiente de Determinação
>
>
> R2=SQReg/SQT;R2
[1] 0.9751577
>
> mod1=lm(Y~X);mod1

Call:
lm(formula = Y ~ X)

Coefficients:
(Intercept)            X
-2.07586      0.02545

>
> summary(mod1)

Call:
lm(formula = Y ~ X)
```

```
Residuals:
Min       1Q   Median       3Q      Max
-0.51375 -0.08687  0.04073  0.17416  0.26833

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.075861   0.221956  -9.353 6.23e-06 ***
X            0.025448   0.001354  18.796 1.57e-08 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 0.2513 on 9 degrees of freedom
Multiple R-squared:  0.9752,    Adjusted R-squared:  0.9724
F-statistic: 353.3 on 1 and 9 DF,  p-value: 1.569e-08

> anova(mod1)
Analysis of Variance Table

Response: Y
Df  Sum Sq Mean Sq F value    Pr(>F)
X          1 22.3041 22.3041  353.29 1.569e-08 ***
Residuals  9  0.5682  0.0631
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
>
>
> mod2=lm(Y~X -1);mod2

Call:
lm(formula = Y ~ X - 1)

Coefficients:
X
0.01355

>
> summary(mod2)

Call:
lm(formula = Y ~ X - 1)

Residuals:
Min       1Q   Median       3Q      Max
-1.01594 -0.75778 -0.46096  0.00855  1.11353

Coefficients:
Estimate Std. Error t value Pr(>|t|)
X 0.013546   0.001435   9.437  2.7e-06 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

```
Residual standard error: 0.7804 on 10 degrees of freedom
Multiple R-squared:  0.8991,    Adjusted R-squared:  0.889
F-statistic: 89.06 on 1 and 10 DF,  p-value: 2.696e-06

> anova(mod1)
Analysis of Variance Table

Response: Y
Df  Sum Sq Mean Sq F value    Pr(>F)
X          1 22.3041 22.3041  353.29 1.569e-08 ***
Residuals  9  0.5682  0.0631
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
>
>
```