

MAE0328 - Análise de Regressão

Fernando Henrique Ferraz Pereira da Rosa
Matheus Moreira Costa
Vagner Aparecido Pedro Junior

15 de maio de 2004

Lista 2¹

2. Um pesquisador deseja verificar se um instrumento para medir concentração de ácido láctico no sangue está bem calibrado. Para isto ele tomou 20 amostras de concentrações conhecidas e determinou a respectiva concentração através do instrumento. Como uma análise de regressão poderia auxiliar o pesquisador?

- (a) Modele o problema acima, especificando as variáveis independentes e dependentes e as hipóteses de interesse.

A variável dependente é a leitura do instrumento em questão, e a variável independente a concentração (conhecida) de ácido láctico. Um modelo linear do tipo $y = \beta_0 + \beta_1 x + \epsilon$ pode ser utilizado para estudar a calibração do instrumento. β_0 representaria um ajuste de escala e β_1 uma medida de quanto o instrumento está se desviando do valor real por unidade de concentração medida. Possíveis hipóteses nulas de interesse seriam $\beta_0 = 0$ e $\beta_1 = 1$.

- (b) Considerando os dados abaixo, ajuste o modelo $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ e teste a hipótese $H_0 : \beta_1 = 1$ contra a alternativa $H_1 : \beta_1 \neq 1$. Tire conclusões com base no resultado desse teste.

x	y				
1	1.1	0.7	1.8	0.4	
3	3.0	1.4	4.9	4.4	4.5
5	7.3	8.2	6.2		
10	12.0	13.1	12.6	13.2	
15	18.7	19.7	17.4	17.1	

Ajustando o modelo através dos estimadores de mínimos quadrados obtemos:

¹Powered by L^AT_EX 2_ε and R 1.8.1

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{646.01}{526.2} = 1.2277 \quad \text{e} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 8.38 - 1.2277(6.7) = 0.1595$$

Sob a suposição de normalidade dos erros, utilizamos a seguinte estatística baseada em uma distribuição t de Student:

$$t_0 = \frac{\hat{\beta}_1 - 1}{\sqrt{QMRes/S_{xx}}}$$

Onde

$$QMRes = \frac{SQT - \hat{\beta}_1 S_{xy}}{n - 2} = \frac{814.05 - 1.2277(646.01)}{18} = 1.1633$$

e

$$S_{xx} = 526.2$$

assim:

$$t_0 = \frac{1.2277 - 1}{\sqrt{1.1633/526.2}} = 4.8426$$

De onde segue, que o valor descritivo desse teste é de

$$\alpha^* = 2P(T_0 \geq t_0) = 0.0001$$

E a partir daí decidimos pela rejeição da hipótese nula, concluindo que o instrumento não está corretamente calibrado.

14. Sejam x uma variável independente e y uma dependente. Decidiu-se adotar o modelo de regressão $z = \beta_0 + \beta_1 w + \epsilon_i$, onde

$$w = \frac{(x - \bar{x})}{\sqrt{S_{xx}}} \quad \text{e} \quad z = \frac{(y - \bar{y})}{\sqrt{S_{yy}}}.$$

- (a) Prove que o estimador de mínimos quadrados de β_1 coincide com r , o coeficiente de correlação amostral entre x e y .

Para o modelo de regressão de w em z proposto, o estimador de mínimos quadrados de β_1 é dado por:

$$\hat{\beta}_1 = \frac{S_{wz}}{S_{ww}} = \frac{\sum w_i z_i - n^{-1} \sum w_i \sum z_i}{\sum w_i^2 - n^{-1} (\sum w_i)^2} \quad (1)$$

Começamos analisando a primeira parcela do denominador da igualdade à direita na equação acima:

$$\sum w_i^2 = \sum \left(\frac{x_i - \bar{x}}{\sqrt{S_{xx}}} \right)^2 = \frac{1}{S_{xx}} \sum (x_i - \bar{x})^2 = \frac{1}{S_{xx}} S_{xx} = 1$$

Calculando então a outra parcela:

$$-n^{-1} \left(\sum w_i \right)^2 = -n^{-1} \left(\sum \frac{x - \bar{x}}{\sqrt{S_{xx}}} \right)^2 = -(\sqrt{S_{xx}n})^{-1} \left(\sum x_i - \bar{x} \right)^2 = 0$$

De forma análoga temos no numerador que a primeira parcela pode ser desenvolvida da seguinte forma:

$$\begin{aligned} \sum w_i z_i &= \sum \frac{(x - \bar{x})(y - \bar{y})}{\sqrt{S_{xx}} \sqrt{S_{yy}}} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{S_{xx}S_{yy}}} \\ &= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{S_{xx}S_{yy}}} = \frac{S_{xy} + n\bar{x}\bar{y} - n\bar{x}\bar{y}}{\sqrt{S_{xx}S_{yy}}} \\ &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \end{aligned}$$

Por fim calculemos a segunda parcela do numerador da equação 1:

$$-n^{-1} \sum w_i \sum z_i = \frac{\sum (x - \bar{x}) \sum (y - \bar{y})}{\sqrt{S_{xx}S_{yy}}} = 0$$

Assim:

$$\hat{\beta}_1 = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = r_{x,y}.$$

(b) Prove que neste caso $SQT = 1$ e $SQReg = r^2$.

$$\begin{aligned} SQT &= \sum (z_i - \hat{z}_i)^2 = \sum \left(\frac{y_i - \bar{y}}{\sqrt{S_{yy}}} - \underbrace{\frac{1}{n} \sum \frac{y_i - \bar{y}}{\sqrt{S_{yy}}}}_0 \right)^2 \\ &= \sum \left(\frac{y_i - \bar{y}}{\sqrt{S_{yy}}} \right)^2 = (S_{yy})^{-1} \sum (y_i - \bar{y})^2 \\ &= (S_{yy})^{-1} S_{yy} = 1 \\ SQReg &= \hat{\beta}_1 S_{wz} = r \cdot r = r^2. \end{aligned}$$

(c) Calcule o estimador de mínimos quadrados de β_0 .

$$\hat{\beta}_0 = \bar{z} - \hat{\beta}_1 \bar{w} = 0 - r \cdot 0 = 0$$

16. Suponha que a regressão de y em x é da forma $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ com $\epsilon_i \sim N(0, \sigma^2)$, ϵ_i e ϵ_j independentes, $i \neq j$, $i = 1, \dots, k$.

Sejam $y_{i1}, y_{i2}, \dots, y_{in_i}$ v.a. independentes, associadas ao mesmo valor x_i , $i = 1, \dots, k$ e $\bar{y}_i = 1/n_i \sum_{j=1}^{n_i} y_{ij}$.

- (a) Como é a regressão de \bar{y} em x ?

Ela é da forma $\bar{y}_i = \beta_0 + \beta_1 x_i + \epsilon_i$, com $\epsilon \sim N(0, \sigma_i^2)$, de forma que a variância não é constante. Em particular, temos, pelas propriedades estatísticas da média que:

$$\text{Var}(\bar{y}_i) = \text{Var}(\epsilon_i) = \sigma^2/n_i. \quad (2)$$

- (b) Determine os estimadores de mínimos quadrados ponderados para o modelo obtido em (a), baseado na amostra $(x_1, \bar{y}_1), \dots, (x_k, \bar{y}_k)$.

Notemos que dada a variância para cada \bar{y} (equação 2) a escolha natural da função peso para cada observação será dada por:

$$w_i = n_i$$

de forma que o peso escolhido é naturalmente inversamente proporcional a variância para cada observação. A função quadrática que precisamos minimizar então fica dada por:

$$S(\beta_0, \beta_1) = \sum_{i=1}^n n_i (\bar{y}_i - \beta_0 - \beta_1 x_i)^2$$

Resultando nas equações:

$$\begin{aligned} \hat{\beta}_0 \sum_{i=1}^n n_i + \hat{\beta}_1 \sum_{i=1}^n n_i x_i &= \sum_{i=1}^n n_i y_i \\ \hat{\beta}_0 \sum_{i=1}^n n_i x_i + \hat{\beta}_1 \sum_{i=1}^n n_i x_i^2 &= \sum_{i=1}^n n_i x_i y_i \end{aligned}$$

Isolando $\hat{\beta}_0$ e $\hat{\beta}_1$ obtemos:

$$\hat{\beta}_1 = \frac{\sum n_i x_i y_i - \frac{\sum n_i x_i \sum n_i y_i}{\sum n_i}}{\sum n_i x_i^2 - \frac{(\sum n_i x_i)^2}{\sum n_i}} \quad (3)$$

$$\hat{\beta}_0 = \frac{\sum n_i y_i - \hat{\beta}_1 \sum n_i x_i}{\sum n_i} \quad (4)$$

Que são portanto os estimadores de mínimos quadrados ponderados pedidos.

- (c) A que se reduzem estes estimadores se $n_1 = n_2 = \dots = n_k = N$?
 É imediata a verificação de que nesse caso os estimadores se reduzem aos estimadores tradicionais de mínimos quadrados do modelo linear simples. Isso é esperado pois se $n_1 = n_2 = \dots = n_k = N$ então pela equação 2 todo \bar{y}_i tem a mesma variância e portanto estamos num modelo de regressão linear simples, com variâncias constantes.

21. (*Exercício 2.4 de Montgomery, Peck e Vining*) A tabela B3 apresenta dados da performance da milhagem de gasolina para 32 automóveis diferentes.

- (a) Ajuste um modelo de regressão linear simples relacionando a milhagem de gasolina y (milhas por galão) com o deslocamento do pistão x_1 .

Vamos ajustar o modelo:

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

Através dos estimadores de mínimos quadrados obtemos:

$$\hat{\beta}_1 = -0.04736 \quad \text{e} \quad \hat{\beta}_0 = 33.7226$$

- (b) Construa a tabela de análise de variância e teste a significância da regressão.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	955.72	1	955.72	101.74
Residual	281.82	30	9.39	
Total	1237.54	31		

$$\alpha^* = P(F_0 \geq 101.74) = 3.743 \times 10^{-11}$$

O que garante a rejeição da hipótese nula de não significância da regressão em qualquer $\alpha \geq 3.743 \times 10^{-11}$.

- (c) Qual porcentagem da variabilidade total da milhagem da gasolina é explicada pela relação linear com o deslocamento do pistão?

Esse valor é dado por $R = 0.772274$, de forma que 70% da variabilidade total da milhagem da gasolina está sendo explicada pela relação linear com o deslocamento do pistão.

- (d) Encontre um intervalo de confiança de 95% para a média de milhagem da gasolina se o deslocamento do pistão é 275 in^3 .

Um intervalo de confiança com $\gamma = 1 - \alpha$ para $E(y|x_0)$ nesse modelo é dado por:

$$\left[\hat{y}_0 - t_{\alpha/2, n-2} \sqrt{MSE \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}, \hat{y}_0 + t_{\alpha/2, n-2} \sqrt{MSE \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} \right]$$

Substituindo os valores conhecidos temos que o intervalo de confiança pedido será dado por:

$$IC(y_0, 0.95) = [273.8893, 276.1107].$$

- (e) Suponhamos que desejamos prever o consumo de gasolina de um carro com um motor de 275 in^3 . Dê uma estimativa pontual da milhagem. Ache um intervalo de predição com 95% de confiança. Uma estimativa pontual é obtida colocando 275 no modelo com os parâmetros estimados. Obtendo:

$$y = 33.72268 - 0.04736(275) = 20.74604$$

Obtendo um intervalo de predição, através de fórmula parecida com a do intervalo para a média, notando que na variância teremos 1 dentro do parênteses multiplicando o MSE, temos:

$$IP(y_0, 0.95) = [268.6427, 281.3573]$$

- (f) Compare os dois intervalos obtidos nas partes *d* e *e*. Explique a diferença entre eles. Qual é maior, e porque?
O intervalo de predição é maior, pois estamos prevendo a ocorrência em específico de *uma* observação da variável. No intervalo de confiança para a média, como o próprio nome diz, estamos prevendo onde a média estará centrada, e portanto, o intervalo será menor, já que a média reduz a variabilidade.

22. (*Exercício 2.5 de Montgomery, Peck e Vining*) Considere novamente o problema do consumo de gasolina. Repita o exercício anterior (partes a, b e c) usando o peso do veículo x_{10} como a variável regressora. Baseado na comparação dos dois modelos, você pode concluir que x_1 é um melhor regressor que x_{10} ?

Vamos ajustar o modelo:

$$y = \beta_0 + \beta_1 x_{10} + \epsilon$$

Através dos estimadores de mínimos quadrados obtemos:

$$\hat{\beta}_1 = -0.005752 \quad \text{e} \quad \hat{\beta}_0 = 40.852431$$

Construindo a tabela de análise de variância:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	921.53	1	921.53	87.482
Residual	316.02	30	10.53	
Total	1237.54	31		

Obtemos o nível descritivo:

$$\alpha^* = P(F_0 \geq 87.482) = 2.121 \times 10^{-10}$$

O que garante a rejeição da hipótese nula de não significância da regressão em qualquer $\alpha \geq 2.121 \times 10^{-10}$.

Com esse regressor a proporção da variabilidade explicada pela relação linear entre y e x_{10} é dada por $R = 0.7446$, o que significa que 74% da variabilidade total da milhagem da gasolina está sendo explicada pelo regressor x_{10} isolado. Como tanto os valores de R como α^* foram menores no caso do regressor x_{10} , concluímos que x_1 é um melhor regressor que x_{10} nesse modelo linear.

Sobre

A versão eletrônica desse arquivo pode ser obtida em <http://www.feferraz.net>

Copyright (c) 1999-2005 Fernando Henrique Ferraz Pereira da Rosa.
 É dada permissão para copiar, distribuir e/ou modificar este documento sob os termos da Licença de Documentação Livre GNU (GFDL), versão 1.2, publicada pela Free Software Foundation;
 Uma cópia da licença em está inclusa na seção intitulada "Sobre / Licença de Uso".