Lista 1 de Análise de Regressão

Alex dos Santos Lima, RA:230560

13 de Agosto de 2021

Resolução da lista (1) de Análise de Regressão.

Nesta resolução iremos usar o seguinte livro com os exercícios. Abaixo teremos a lista 1 com os exercícios da quinta edição do livro-texto: "Kutner, Nachtsheim, Neter and Li (2005). Applied Linear Statistical Models. Fifth Edition, McGraw-Hill" (KNNL2005).

1.6; 1.7; 2.4; 2.52; 2.64

1.6) Consider the normal error regression model (1.24). Suppose that the parameter values are $\beta_0 = 200, \beta_1 = 5.0$, and $\sigma = 4$.

a. Plot this normal error regression model in the fashion of Figura 1.6. Show the distributions of Y for X= 10, 20, and 40.
b. Explain the meaning of the parametrs β_0 and β_1. Assume that the scope of the model includes X=0.

1.7) In a simulation exercise, regression model (1.1) applies with $\beta_0 = 100, \beta_1 = 20$, and $\sigma^2 = 25$.

An observation on Y will be made for $X = 5$.

a. Can you state the exact probability tha Y will fall between 195 and 205? Explain.
b. If the normal error regression model (1.24) is applicable, can you now state the exact probability that Y will fall between 195 and 205? If so, state it.

2.4) Refer to **Grade point average** Problem 1.19.

a. Obtain a 99 percent confidence interval for β_1. Interpret your confidence interval. Does it include zero? Why might the director of admissions be interested in whether the confidence interval includes rero?
b. Test, using the test statistic t^*, whether or not a linear association exists between student's ACT score (X) and GPA at the end of the freshman year (Y). Use a level of significance of .Ol. State the alternatives, decision rule, and conclusion.
c. What is the P-value of your test in part (b)? How does it support the conclusion reached in part (b)?

2.52) Derive the expression in (2.22b) for the variance of b_0, making use of (2.31). Also explain how variande (2.22b) is a special case of variance (2.29b).

2.64) Refer to the **SENIC** data set in Appendix C.1 and Project 1.45. Using R^2 as the criterion, which predictor variable accounts for the largest reduction in the variability of the average length of stay?

Fazendo a tradução e resolvendo os exercícios logo abaixo.

1.6) Considerar o modelo de regressão de erro normal (1.24). Suponha que os valores dos parâmetros são $\beta_0 = 200, \beta_1 = 5,0$, e $\sigma = 4$.

a. Plotar este modelo de regressão de erro normal à modelo da Figura 1.6. Mostrar as distribuições de Y por X= 10, 20, e 40.

Resposta:

Seja o modelo de erro normal de regressão linear simple .

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Onde: Y_i é a observação da resposta do i

X_i é uma constante conhecida, o nivel da variável preditora no i

β_0 e β_1 são parametros

ϵ_i são independente em $N(0, \sigma^2)$

i=1,. . . ,n

Então temos:

$$Y_i = 200 + 5X_i + \epsilon_i$$

Sendo $X_i, i = 1, 2, 3$ para $X_1 = 10, X_2 = 20, X_3 = 40$

onde ϵ_i são independente em $N(0, \sigma^2) = N(0, 16)$ com $\sigma = 4$.

Da distribuição de probabilidade qual á média é a esperança :

$$E[Y_i] = 200 + 5X_i$$

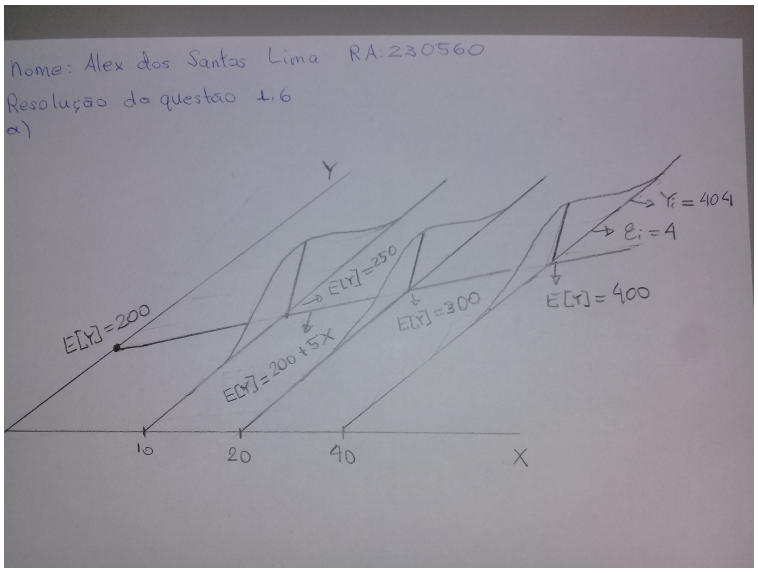
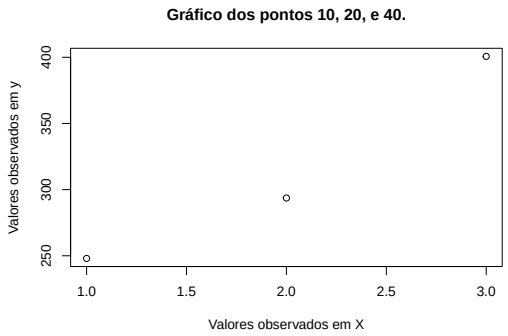


Figura 1: Distribuição da esperança de cada valor e sua média



b. Explicar o significado dos parametros β_0 e β_1. Assumir que o âmbito do modelo inclui X=0.

Resposta:

O β_0 é o intercepto do modelo de regressão linear simples, ou seja, quando X assume 0 (zero). O valor de Y é exatamente igual a 0 e temos que $\beta_0 = 200$.

Desta maneira, β_1 é coeficiente angular do modelo de regressão linear simples, ou seja, a cada incremento unitário em X aumentamos β_1 unidade(s) em X, logo o $\beta_1 = 5, 0$.

1.7) Num exercício de simulação, o modelo de regressão (1.1) aplica-se com $\beta_0 = 100, \beta_1 = 20$, e $\sigma^2 = 25$. Uma observação sobre Y será feita por $X = 5$.

a. Pode indicar a probabilidade exacta de Y cair entre 195 e 205? Explique.

Resposta:

Não, podemos fazer porque não é definido uma definição para o modelo.

b. Se o modelo de regressão de erro normal (1,24) for aplicável, pode agora indicar a probabilidade exacta de Y cair entre 195 e 205? Em caso afirmativo, declare-o.

Resposta:

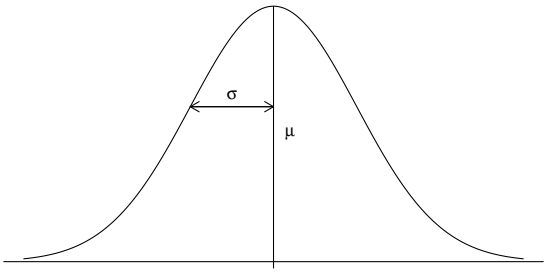
Sendo assim a fórmula logo abaixo é calculada com probabilidade de valores da média μ, como Y não é media para cair entre 195 e 205. Sendo assim a função de densidade conjunta é dada pela probabilidade que cálculo o valor centrado na média é essa.

$$f_i = \frac{1}{\sqrt{2\pi\sigma}} exp\left[-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2\right]$$

Logo vamos usar outra fórmula que faz os cálculos entre dois valores. Vamos agora assumir um erro na distribuição normal em ϵ. Sendo a probabilidade da distribuição de Y para cada valor de $X = x$ e para normal com média e variância dada. Para $X = 5$ testem aos valores da esperança $E[X] = \beta_0 + \beta_1 X = 100 + 20 * 5 = 200$ e $\sigma^2 = 25$ respectivamente. A exata probabilidade seguir.

$$P(195 \leq Y \leq 205) = P\left(\frac{195 - 200}{5} \leq \frac{X - \mu}{\sigma} \leq \frac{205 - 200}{5}\right)$$
$$= P(-1 \leq Z \leq 1) = 2 * P(Z \leq 1) - 1 = 0.6826$$

Pela tabela da normal.



A probabilidade exata de Y cair entre 195 e 205 é quando o valor esperado seja X=5 em $E[Y] = 200$. Sim, pode cair com uma probabilidade de 0.68

2.4) Consultar **Média de pontos** Problema 1.19.

O banco de dados em **Média de pontos** do Problema 1.19.

Table 1: Conjunto de dados com as 6 primeiras linhas

V1	V2
3.897	21
3.885	14
3.778	28
2.540	22
3.028	21
3.865	31

a. Obter um intervalo de confiança de 99 por cento por β_1. Interprete o seu intervalo de confiança. Inclui zero? Porque estaria o director de admissões interessado em saber se o intervalo de confiança intervalo inclui zero?

Resposta:

Um intervalo dde $100(1-\alpha)\%$ de confiança para β_1 é dado pela seguinte formula:

$$100(1-\alpha)\% = \left[\hat{\beta}_1 - t_{n-2,\alpha/2}\sqrt{\frac{S^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}}; \hat{\beta}_1 + t_{n-2,\alpha/2}\sqrt{\frac{S^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}}\right]$$

Em que S^2 é descrito pela seguinte fórmula logo abaixo

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i-\hat{Y})^2}{n-2}$$

Vamos cálcular o valor de $t_{n-2,\alpha/2} = t_{120-2,\frac{0.01}{2}} = t_{118,5*10^{-3}}$

Como encontrar $t_{n-2,\alpha/2}$

$$P(-t_{n-2,\alpha/2} < T < t_{n-2,\alpha/2}) = 1-\alpha$$

$P(-t_{n-2,\alpha/2} < T < t_{n-2,\alpha/2}) = 1 - 0.99 = 0.01$

Com esses valores aqui visto na tabela de t

Gral de liberdade = 118

Probabilidade para um teste bicaudal =0.995

$P(-2.618137 < T < 2.618137) = 0.01$

Desvio padrão:

$\sigma = 0.01277$

Estatísitca do teste t:

$t_{(1-\frac{\alpha}{2},n-2)} = 2.6181$

O nível de significância.:

$\alpha = 99\%$

Calcálo do modelo ajustado em Regressão Linear Simples:

$Y_i = 2.114 + 0.0388X_i$

Table 2: I.C. para β_1 em 99%

	0.5 %	99.5 %
(β_0)	1.273902675	2.95419590
β_1	0.005385614	0.07226864

Table 3: Estatística da Regressão Linear Simples.

	Estimativa	S. Q. dos Erros	t value	Pr(>|t|)
(β_0)	2.1140	0.3209	6.59	0.0000
β_1	0.0388	0.0128	3.04	0.0029

99% I.C. para $\beta_1 : b_1 - t_{(1-\frac{\alpha}{2},n-2)}s(b_1) \leq \beta_1 \leq b_1 + t_{(1-\frac{\alpha}{2},n-2)}s(b_1) = 0.00538 \leq \beta_1 \leq 0.07226$

$t(.995; 118) = 2.61814, .03883ś2.61814(.01277))$

No intervalo do I.C. caso seja incluído zero, o β_1 pode ser zero e $\beta_1 = 0$

$b_1 == 0.00538 \leq \beta_1 \leq 0.07226$

Interpretação do Intervalo de Confiança paraβ_1:

Se repetirmos o experimento com n=120 e os mesmos valores fixos de X dessa amostra e um número muito grande de vezes, o intervalo assim construído cobrirá o verdadeiro valor de β_1 em 99% das amostras. Assim, caso inclui o 0, X no caso os pontos médios podam não ter uma importância estatisticamente para estimar Y sendo a resposta de interesse. Estamos dizemos com basamento com no intervalo de confiança com 99% por cento.

b. Teste, utilizando a estatística do teste t^*, quer exista ou não uma associação linear entre a pontuação ACT do estudante (X) e GPA no final do ano de caloiro (Y). Utilizar um nível de significância de 0.01. Indique as alternativas, a regra da decisão e a conclusão.

Resposta:

1.O nível de significância:

$\alpha = 0.01$

2. Hipóteses para ser testada

H_0:$\beta_1 = 0$

H_1:$\beta_1 \neq 0$

3. Estatística do Teste

Resultado através do modelo ajustado em (a) onde 0.03883 é a soma dos quadrados da regressão para o valor do β_1 testando quando o $\beta_1 = 0$ na hipótese nula e alternativa temos que $\beta_1 \neq 0$, e no dominador é somatória dos quadrados médio dos erros sendo 0.01277. $T_0 = \dfrac{b_1 - \beta_{10}}{s(b_1)} = \dfrac{b_1}{s(b_1)} = \dfrac{(0.03883 - 0)}{0.01277} = 3.04072$

4. Interpretação da questão b)

Decisão: Rejeita H_0 se $|T_0| > t_{(1-\frac{\alpha}{2},n-2)}, 3.04$
Se $|t^*_{(0.995,118)}| \leq 2.618137$ concluimos H_0, de outra forma H_a, Concluimos H_a

O que faz rejeitar H_0

A um nível de significância de 1% porcento, considerando o teste t, rejeitamos a hipótese nula, ou seja, X é estatisticamente significante para estimar Y.

c. Qual é o P-valor do seu teste em parte (b)? Como é que apoia a conclusão alcançada em parte (b)?

Resposta:

Table 4: Anova para valor-p

	Gral de L.	Soma S. Q.	Média dos S. Q	Valor de F	Pr(>F)
β_1	1	3.59	3.59	9.24	0.0029
Residuals	118	45.82	0.39		

Cálcular o p-valor ele é calculado na saida do modelo ajustado em (a);

Como p-valor=0.00291 < 0.01, o que concluimos que devemos rejeitar H_0

2.52) Derivar a expressão em (2.22b) para a variação de b_0, fazendo uso de (2.31). Explicar também como variande (2.22b) é um caso especial de variância (2.29b).

Resposta:

A expressão em (2.22b) é essa logo abaixo.

$$\sigma^2\{b_0\} = \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2}\right]$$

Resolvando temos que.

$$\sigma^2\{b_0\} = \sigma^2\left\{\overline{Y} - b_1\overline{X}\right\} = \sigma^2\left\{\overline{Y}\right\} + \overline{X}^2\sigma^2\{b_1\} - 2\overline{X}\sigma\left\{\overline{Y}, b_1\right\}$$
$$= \frac{\sigma^2}{n} + \overline{X}\frac{\sigma^2}{\sum(X_i - \overline{X})^2} - 0$$

$$= \sigma^2\left[\frac{1}{n} + \frac{\overline{X^2}}{\sum(X_i - \overline{X}^2)}\right]$$

Na expressão da variânça em (2.22b) é

$$\sigma^2\{b_0\} = \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2}\right]$$

e a variânça em (2.29b) é

$$\sigma^2\left\{\hat{Y_H}\right\} = \sigma^2\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum(X_i - \overline{X})^2}\right]$$

Em (2.22b) a função de $\sigma^2\{b_0\}$ monótonas decrescentes em S_{XX}, por exemplo, num desenho experimental em que S_{XX} forem os maiores possíveis valores de X ao definir um tamanho da amostra possível com os valores de X_{min} e X_{max}, por linearidade, parece ser razoável fixar o valor de X como X_{mim} em metade e X_{max} na outra metade. Agora em (2.29b) a variância da questão é a predição de uma nova observação, tanto podemos querer predizer o valor para um nível de X_h presente no experimento como para um novo nível de predição. No caso de $E[Y_h]$ temos uma quantidade determinística que só é desconhecida por não sabermos os valores de β_0 e β_1. Assim, $\sigma\{Y_h\} \to 0$ quadno $n \to \infty$, pois $\hat{\beta_0}$ e $\hat{\beta_1}$ são consistentes. Sendo assim divide-se o problema de predição nos casos de parâmetros: conhecidos; e desconhecidos.

2.64) Consultar o conjunto de dados **SENIC** no Apêndice C.1 e Projecto 1.45. Usando R^2 como critério, que prevê a maior redução na variabilidade da duração média da estadia?

Resposta:

Sendo assim, os maiores R^2 que prevê a maior redução na variablidade são esses logo abaixo.

Taxa de infecção: $R^2 = 0,2846$

Instalações: $R^2 = 0,1264$

X-raio:$R^2 = 0,1463$

Usando o coeficiente de determinação o R^2 sendo a formula logo abaixo.

Calculando usando o R^2

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{Y_i} - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2/(n-1)}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2/(n-1)}$$

Banco de daddos em **SENIC** no Apêndice C.1 e Projecto 1.45

Table 5: Conjunto de dados com as 6 primeiras linhas

V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
7.13	55.7	4.1	9.0	39.6	279	2	4	207	241	60
8.82	58.2	1.6	3.8	51.7	80	2	2	51	52	40
8.34	56.9	2.7	8.1	74.0	107	2	3	82	54	20
8.95	53.7	5.6	18.9	122.8	147	2	4	53	148	40
11.20	56.5	5.7	34.5	88.9	180	2	1	134	151	40
9.76	50.9	5.1	21.9	97.0	150	2	2	147	106	40

Códigio usado para realizer a Lista 1

Código usado na questão 1.6

Gráficos de pontos

- b0 = 200;b1 = 5
- sigma = 4
- e = rnorm(3, sd = sigma)
- x = c(10,20,40);y = b1*x + b0 + e
- plot(y, main = “Gráfico dos pontos 10, 20, e 40.” ,
  xlab = “Valores observados em X”,
  ylab = “Valores observados em y”)

Código usado na questão 1.7

Gráfico da normal

- x <- seq(-3, 3, length = 501)
- plot(x, dnorm(x), axes = FALSE, type = ‘l’, xlab = ’‘,ylab =”) ; abline(h = 0)
- x <- 0 ; lines(c(0, 0), c(dnorm(x), -0.01))
- x <- -1 ; lines(c(-1, 0), c(dnorm(x), dnorm(x))) arrows(-1, dnorm(x), 0, dnorm(x), code= 3, length = 0.1) text(0.2, 0.2, expression(italic(mu))) text(-0.5, 0.26, expression(italic(sigma)))

Código uado na questão 2.4

Leitura do conjunto de dados

- NO banco de dados em **Média de pontos** do Problema 1.19
- dados1 <- read.table(“~/dados1.txt”, quote=“"”, comment.char="")
- knitr::kable(head(dados1), caption = “Conjunto de dados com as primeira linhas”)

Biblioteca usada para resover a questão 2.4

- library(readxl)
- library(readr)
- library(gdata)

Leitura de conjunto de dados em Média de pontos do Problema 1.19

- dados1 <- read.table(“~/dados1.txt”, quote=“"”, comment.char="")

Ajuste do modelo de Regressão Linear Simples

- fit <- lm(dados1$V1~dados1$V2)
- a<-summary(fit)
- apha <- 0.995
- tvalor<- qt(apha, df= 120-2)

- intvl<- confint(fit, level = apha,df =fit$df) - aqui alpha= 0.99 pegamos 1-0.99=0.01
- s2 <- sum((dados1$V1-mean(dados1$V1))^2)

b)

- intervalo <- confint(fit, level = 0.01 ,df =fit$df)
- alfa <-1-(0.01/2) obtendo o alpha
- ttab<-qt(alfa,df=120-2) obtendo o valor do t tabelado

Questão 2.64

Leiturado do conunto de dados para questão 2.64

- dados <- read.table("~/dados.txt", quote="""", comment.char="")
- dados <- dados[,-(1)] - retirando uma coluna de index
- knitr::kable(head(dados),caption = "Conjunto de dados com as primeiras linhas")

Cálculo do R^2

- dados <- read.table("~/dados.txt", quote="""", comment.char="")
- dados <- dados[,-(1)]
- anv<- anova(lm(dados))
- sy <- summary(lm(dados))
- y <-dados$V2
- x <- c(dados$V4)
- beta1 <- cor(y, x) * sd(y) / sd(x)
- beta0 <- mean(y) - beta1 * mean(x)
- y_chapeu <- beta0 +beta1*x
- y_barra <- mean(y)
- r_quadrado <- sum((y_chapeu-y_barra)$^{2)/\text{sum}((\text{y-y_barra})}$2)