

Modelos Lineares I

*Daniel dos Santos
Lyncoln Sousa Oliveira*

16 de setembro de 2019

a) Especificando o modelo teórico.

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \beta_4 X_{i,4} + \varepsilon_i \quad ; \quad i = 1, 2, \dots, 100$$

Onde:

Y_i : Tempo de vida em escala logarítmica do i -ésimo paciente;

$X_{i,1}$: Idade do i -ésimo paciente;

$$X_{i,2} = \begin{cases} 1, & \text{se o } i\text{-ésimo paciente é do sexo feminino;} \\ 0, & \text{c.c.} \end{cases}$$

$$X_{i,3} = \begin{cases} 1, & \text{se o } i\text{-ésimo paciente tem histórico de uso de álcool severo;} \\ 0, & \text{c.c.} \end{cases}$$

$$X_{i,4} = \begin{cases} 1, & \text{se o } i\text{-ésimo paciente tem histórico de uso de álcool moderado;} \\ 0, & \text{c.c.} \end{cases}$$

β_0 : É intercepto do modelo; PERGUNTAR DPS

β_1 : Variação do tempo de vida em escala logarítmica para cada unidade de idade; β_2 : Efeito do tempo de vida em escala logarítmica quando o indivíduo é do sexo feminino;

β_3 : Efeito do tempo de vida em escala logarítmica quando o indivíduo possui histórico de álcool severo;

β_4 : Efeito do tempo de vida em escala logarítmica quando o indivíduo possui histórico de álcool moderado;

ε_i : Erro aleatório do i -ésimo paciente;

Onde as hipóteses básicas são:

$$E[\varepsilon_i] = 0; \quad i = 1, 2, \dots, 54;$$

$$Var(\varepsilon_i) = \sigma^2; \quad i = 1, 2, \dots, 54;$$

$$Cor(\varepsilon_i, \varepsilon_j) = 0; \quad i = 1, 2, \dots, 54; \quad i \neq j; \quad \varepsilon_i \sim N(0, \sigma^2); \quad i = 1, 2, \dots, 54;$$

Testes de hipóteses para avaliar significância entre as variáveis do estudo. Para $\alpha = 5\%$ e $n = 54$.

Teste de Hipótese para β_1

Passo 1: Definição das Hipóteses.

$$\begin{cases} H_0 : \beta_1 = 0; \\ H_1 : \beta_1 \neq 0; \end{cases}$$

Passo 2: Cálculo da Estatística de teste sob H_0 .

$$T = \frac{\hat{\beta}_1}{\sqrt{\hat{Var}(\hat{\beta}_1)}} \sim T_{49}$$

onde,

$$\hat{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{54} (X_{i,1} - \bar{X}_1)^2};$$

desta forma,

$$t_{obs} = -0,886;$$

Veja a [tabela 1](#)

Passo 3: Região Crítica.

$$RC = \{t \in R : t > t_{49,0,025} = 2.009575 \text{ ou } t < -t_{49,0,025} = -2.009575\}$$

Passo 4: Tomada de decisão.

Como obteve-se um $t_{obs} = -0,886$, então $t_{obs} \notin RC$ (verifique a [figura ??](#)), desta forma, não rejeitamos a hipótese nula a um nível de significância $\alpha = 5\%$. Conclui-se. com base no teste T que não existe uma relação estatisticamente significativa entre a idade (X_1) e o tempo de vida do paciente em escala logarítmica (Y).

Teste de Hipótese para β_2

Passo 1: Definição das Hipóteses.

$$\begin{cases} H_0 : \beta_2 = 0; \\ H_1 : \beta_2 \neq 0; \end{cases}$$

Passo 2: Calculo da Estatística de teste sob H_0 .

$$T = \frac{\hat{\beta}_2}{\sqrt{\hat{Var}(\hat{\beta}_2)}} \sim T_{49}$$

onde,

$$\hat{Var}(\hat{\beta}_2) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{54} (X_{i,2} - \bar{X}_2)^2};$$

desta forma,

$$t_{obs} = -2,012;$$

Veja a [tabela 1](#)

Passo 3: Região Crítica.

$$RC = \{t \in R : t > t_{49,0,025} = 2.009575 \text{ ou } t < -t_{49,0,025} = -2.009575\}$$

Passo 4: Tomada de decisão.

Como obteve-se um $t_{obs} = -2,012$, então $t_{obs} \in RC$ (verifique a [figura ??](#)), desta forma, rejeitamos a hipótese nula a um nível de significância $\alpha = 5\%$. Conclui-se. com base no teste T que existe uma relação estatisticamente significativa entre o sexo do paciente (X_2) e o tempo de vida em escala logarítmica (Y).

Teste de Hipótese para β_3

Passo 1: Definição das Hipóteses.

$$\begin{cases} H_0 : \beta_3 = 0; \\ H_1 : \beta_3 \neq 0; \end{cases}$$

Passo 2: Calculo da Estatística de teste sob H_0 .

$$T = \frac{\hat{\beta}_3}{\sqrt{\hat{Var}(\hat{\beta}_3)}} \sim T_{49}$$

onde,

$$\hat{Var}(\hat{\beta}_3) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{54} (X_{i,3} - \bar{X}_3)^2};$$

desta forma,

$$t_{obs} = -2,1891;$$

Veja a [tabela 1](#)

Passo 3: Região Crítica.

$$RC = \{t \in R : t > t_{49,0,025} = 2.009575 \text{ ou } t < -t_{49,0,025} = -2.009575\}$$

Passo 4: Tomada de decisão.

Como obteve-se um $t_{obs} = -2,1891$, então $t_{obs} \in RC$ (verifique a [figura ??](#)), desta forma, rejeitamos a hipótese nula a um nível de significância $\alpha = 5\%$. Conclui-se. com base no teste T que existe uma relação estatisticamente significativa entre o consumo moderado de álcool (X_3) e o tempo de vida em escala logarítmica (Y).

Teste de Hipótese para β_4

Passo 1: Definição das Hipóteses.

$$\begin{cases} H_0 : \beta_4 = 0; \\ H_1 : \beta_4 \neq 0; \end{cases}$$

Passo 2: Calculo da Estatística de teste sob H_0 .

$$T = \frac{\hat{\beta}_4}{\sqrt{\hat{Var}(\hat{\beta}_4)}} \sim T_{49}$$

onde,

$$\hat{Var}(\hat{\beta}_4) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{54} (X_{i,4} - \bar{X}_4)^2};$$

desta forma,

$$t_{obs} = -2,012;$$

Veja a [tabela 1](#)

Passo 3: Região Crítica.

$$RC = \{t \in R : t > t_{49,0,025} = 2.009575 \text{ ou } t < -t_{49,0,025} = -2.009575\}$$

Passo 4: Tomada de decisão.

Como obteve-se um $t_{obs} = -2,012$, então $t_{obs} \in RC$ (verifique a [figura ??](#)), desta forma, rejeitamos a hipótese nula a um nível de significância $\alpha = 5\%$. Conclui-se. com base no teste T que existe uma relação estatisticamente significativa entre o consumo severo da álcool (X_4) e o tempo de vida em escala logarítmica (Y).

Teste F da tabela ANOVA

Passo 1: Definição das Hipóteses.

$$\begin{cases} H_0 : \beta_1 = 0; \\ H_1 : \beta_1 \neq 0; \end{cases}$$

Passo 2: Cálculo da Estatística de teste sob H_0 .

$$F = \frac{QMReg}{QMRes} \sim F_{1,n-2}$$

onde,

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

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

desta forma,

$$f_{obs} = 35.922;$$

Veja a tabela 2

Passo 3: Região Crítica.

$$RC = \{f \in R : f > f_{1,n-2;\alpha} = 3.938111\};$$

Passo 4: Tomada de decisão.

Como obteve-se um $f_{obs} = 35.922$, então $f_{obs} \in RC$ (verifique a figura ??), desta forma, rejeitamos a hipótese nula a um nível de significância $\alpha = 5\%$. Conclui-se. com base no teste F que existe uma relação estatisticamente significativa entre tempo médio de internação (X) e percentual de pacientes infectados (Y).

d) Gráfico de dispersão incluindo o modelo ajustado.

Veja a figura ??.

Observa-se uma relação estatística linear positiva entre as variáveis tempo médio de internação (em dias) e percentual de pacientes infectados, isto é, quanto maior o tempo médio de internação maior tende a ser o percentual de pacientes infectados.

e) Coeficiente de correlação linear de Pearson e coeficiente de determinação do modelo (R^2).

- Coeficiente de correlação linear de Pearson;
Coeficiente de correlação linear é de 0.5179087, isto é, existe uma correlação linear positiva moderada, ou seja, quanto maior o tempo médio de internação maior tende a ser o percentual de pacientes infectados.
- Coeficiente de determinação do modelo (R^2);
O coeficiente de determinação do modelo (R^2) é de 0.2682294, isto é, o modelo ajustado explica aproximadamente 26,82% da variação do percentual de pacientes infectados.

f) Verificação de das hipóteses básicas do modelo.

Serão utilizadas as figuras ?? e ?? para observar se há alguma violação nas hipóteses básicas do modelos, que são: Homocedasticidade, lineariedade, normalidade e outliers. Será suposto a independência dos erros aleatórios do modelo.

- Homocedasticidade e Lineariedade:
Pela figura ?? é possível notar uma nuvem de pontos aleatórios em torno de 0(zero), o que indica que os erros aleatórios possuem variâncias constante. Também é possível notar que não há presença de um padrão sistemático dos pontos, o que indica que a hipótese de lineariedade não foi violada.
- Normalidade:
Pela figura ??, pode-se notar que os quantis dos resíduos studentizados se aproximam dos quantis teóricos de uma distribuição normal, o que é um bom indicativo de normalidade dos erros.
- Outliers:
Também pela figura ?? é possível notar que apenas 6 observações dos resíduos studentizados são maiores que $|2|$, o que é apenas 6% de toda a amostra. Logo é uma quantidade aceitável de outliers.

g) O modelo é adequando para os dados apresentados?

Por não apresentar nenhuma violação nas hipóteses básicas, o modelo é adequado para representar os dados observados, porém seu coeficiente de determinação (R^2) é de apenas 26.82%.

d) Ajustando o modelo pelo métodos dos mínimos quadrados.

O modelo ajustado é dado por:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad ; \quad i = 1, 2, \dots, 100$$

Modelo ajustado com suas devidas estimativas: Veja a tabela 1 no apêndice.

$$\hat{Y}_i = -0.14848 + 0.47580 X_i \quad ; \quad i = 1, 2, \dots, 100$$

\hat{Y}_i : É o valor estimado do percentual de pacientes infectados para o i-ésimo estabelecimento de saúde;

X_i : É o valor observado de tempo médio de internação em dias para o i-ésimo estabelecimento de saúde;

$\hat{\beta}_0$: É o intercepto estimado do modelo, que possui valor de -0.14848 . Não possui interpretação prática para o problema;

$\hat{\beta}_1$: É o coeficiente angular estimado do modelo, que possui valor de 0.47580 . Isto é, o valor estimado do percentual de pacientes infectados é acrescentado em 0.47580 para cada unidade do tempo médio de internação;

Apêndice 1 - Tabelas

Tabela 1: Tabela resumo

Estimadores	Estimativa	Erro padrão	Valor da estatística de teste T	P-valor
$\hat{\beta}_0$	6,956899	0,314901	22,092	< 0,001
$\hat{\beta}_1$	-0,004994	0,005638	-0,886	0,38011
$\hat{\beta}_2$	0,248113	0,123322	2,012	0,04974
$\hat{\beta}_3$	-0,533359	0,184475	-2,891	0,00571
$\hat{\beta}_4$	-0,438230	0,167344	-2,619	0,01171

Tabela 2: Tabela ANOVA

Fontes de variação	Soma dos quadrados	gl	Quadrado médio	Valor da estatística de teste F	P-valor
Regressão	44.305	1	44.305	35.922	< 0.001
Resíduos	120.871	98	1.233		
Total	165.176	99			

Apêndice 3 - Códigos

Resolução das questões

```
#Lendo a base de dados
BD = foreign::read.spss("internacoes.sav")
tabela = dplyr::as_tibble(BD); tabela

#Questão a)
modelo1 = lm(tabela$Percentual_infectados~tabela$Tempo_internacao)
summary(modelo1)

#Questão b)
modelo1
anova(modelo1)

#Questão c)
qt(0.025, 98, lower.tail = F)
qf(0.05, 1, 98, lower.tail = F)

# Questão d)
plot(tabela$Percentual_infectados~tabela$Tempo_internacao, pch = 19,
      ylab = "Percentual de pacientes infectados",
      xlab = "Tempo médio de internação (em dias)",
      ylim = c(-1,9),
      xlim = c(0,14))
abline(modelo1, col = "Royalblue")
text(2,8,expression(hat(Y[i]) == -0.14848 + 0.47580*X[i]),col = "blue")
text(1.1,7, expression(R^2 == "26.82%"), col = "blue")

# Questão e)
R = cor(tabela$Percentual_infectados,tabela$Tempo_internacao); R
R2 = R^2; R2

# Questão f)
yichapeu = fitted(modelo1)
ei = rstandard(modelo1)
plot(ei~tabela$Tempo_internacao, pch = 20,
      ylim = c(-3,3), ylab = "Resíduos studentizados",
      xlab = "Tempo médio de internação (em dias)")
abline(h = c(-2,0,2), col = c("red","black","red"),lty = c(2,1,2))
plot(ei~yichapeu, pch = 20,
      ylim = c(-3,3), ylab = "Resíduos studentizados",
      xlab = "Percentual de pacientes infectados estimado")
abline(h = c(-2,0,2), col = c("red","black","red"),lty = c(2,1,2))
qqnorm(ei, pch = 20, main="",
        ylab = "Quantis teóricos normal padrão",
        xlab = "Quantis dos resíduos studentizados")
abline(0,1, col = "blue")
```

Gerando o gráfico da densidade da F.

```
degree_1 = 1
degree_2 = 98
quantile = qf(0.05, df1 = degree_1, df2 = degree_2, lower.tail = F)
test_stat = 35.922

rc_values = seq(quantile, 7, length = 100)
denisty_rc_values = df(rc_values, df1 = degree_1, df2 = degree_2)
ic_values = seq(0, quantile, length = 100)
denisty_ic_values = df(ic_values, df1 = degree_1, df2 = degree_2)

plot(
  function(x)
    df(x,
      df1 = degree_1,
      df2 = degree_2),
  xlim = c(0, 7),
  ylab = '',
  xlab = 'Quantis',
  bty="n",
  yaxt='n',
  xaxt='n'
)
axis(side=1, at=round(c(0, quantile, 7), 2))

polygon(
  x = c(quantile, rc_values, 7),
  y = c(0, denisty_rc_values, 0),
  border = FALSE,
  col = 'red',
  density = 50
)

lines(
  x = c(quantile, quantile),
  y = c(0, denisty_rc_values[length(denisty_rc_values)]),
  lty = 2
)

lines(x=c(0, 7), y=c(0,0))

text(6, 0.09, expression(alpha), cex=1.7)
text(0.5, 0.15, expression(1 - alpha), cex=1.7)

par(xpd=TRUE)
text(2, -0.33, expression(alpha == 0.05))
legend(4, 1, legend = c('Região Crítica'), box.col = "white",
  fill = c('red'),
  density = 50)
par(xpd=FALSE)
```


Gerando o gráfico da densidade da T.

```
degree = 98
quantile = qt(0.975, df = degree)
b0_test_stat = -0.195
b1_test_stat = 5.993
rc_values = seq(-4, -quantile, length = 100)
denisty_rc_values = dt(rc_values, df = 46)
ic_values = seq(-quantile, quantile, length = 100)
denisty_ic_values = dt(ic_values, df = 46)

plot(
  function(x)
    dt(x, df = 46),
  xlim = c(-4, 7),
  ylab = '',
  xlab = 'Quantis',
  bty="n",
  yaxt='n',
  xaxt='n'
)
axis(side=1, at=round(c(-4, -quantile, 0 ,quantile, 7), 2))

polygon(
  x = c(-4, rc_values, -quantile),
  y = c(0, denisty_rc_values, 0),
  border = FALSE,
  col = 'red',
  density = 50
)
polygon(
  x = c(quantile, sort(-1 * rc_values), 7),
  y = c(0, sort(denisty_rc_values, decreasing = TRUE), 0),
  border = FALSE,
  col = 'red',
  density = 50
)

lines(
  x = c(-quantile, -quantile),
  y = c(0, denisty_rc_values[length(denisty_rc_values)]),
  lty = 2
)
lines(
  x = c(quantile, quantile),
  y = c(0, denisty_rc_values[length(denisty_rc_values)]),
  lty = 2
)

lines(x=c(-4, 7), y=c(0,0))
```

```

text(3.2, 0.23, expression(t[obs]), cex=1.3)
points(b1_test_stat, 0, pch=16)
arrows(3.2, 0.2,b1_test_stat,0)

text(-3, 0.07, expression(frac(alpha, 2)), cex=1.3)
text(3, 0.07, expression(frac(alpha,2)), cex=1.3)
text(0, 0.15, expression(1 - alpha), cex=1.3)

par(xpd=TRUE)
text(4, -0.1, expression(alpha == 0.05))
legend(3.7,0.35,legend = c('Região Crítica'), box.col = "white",
      fill = c('red'),
      density = 50)
par(xpd=FALSE)

```