Modelos Lineares I

Daniel dos Santos
Lyncoln Sousa Oliveira

16 de setembro de 2019

a) Especificando o modelo teórico que avalia o efeito do tempo de internação médio (em dias) no percentual de pacientes infectados em 100 estabelecimentos de saúde.

O modelo teórico é dado por:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad ; \quad i = 1, 2 \ldots, 100$$

Onde:
Y_i: Percentual de pacientes infectados do i-ésimo estabelecimento de saúde;
X_i: Tempo médio de internação em dias do i-ésimo estabelecimento de saúde;
β_0: Percentual de pacientes infectados quando não há tempo médio de internação;
β_1: Variação de percentual de pacientes infectados para cada unidade tempo médio de internação;
ε_i: Erro aleatório do i-ésimo estabelecimento de saúde;

Onde as hipóteses básicas são:
$E[\varepsilon_i] = 0; \quad i = 1, 2 \ldots, 100;$
$Var(\varepsilon_i) = \sigma^2; \quad i = 1, 2 \ldots, 100;$
$Cor(\varepsilon_i, \varepsilon_j) = 0; \quad i = 1, 2 \ldots, 100; \quad i \neq j;$
$\varepsilon_i \sim N(0, \sigma^2); \quad i = 1, 2 \ldots, 100;$

b) Ajustando o modelo pelo métodos dos mínimos quadrados.

O modelo ajustado é dado por:

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

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

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

$\hat{Y}_i$: É o valor estimado do percentual de pacientes infectados para i 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;

c) Testes de hipóteses para avaliar significância entre as variáveis do estudo. Para $\alpha = 5\%$ e $n = 100$.

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: Calculo da Estatística de teste sob H_0.

$$T = \frac{\hat{\beta_1}}{\sqrt{\hat{Var}(\hat{\beta_1})}} \sim T_{n-2}$$

onde,

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

desta forma,

$$t_{obs} = 5.993;$$

Veja a tabela 1

Passo 3:
Região Crítica.

$$RC = \{t \in R : t > t_{n-2,\frac{\alpha}{2}} = 1.984467 \quad ou \quad t < -t_{n-2,\frac{\alpha}{2}} = -1.984467\}$$

Passo 4: Tomada de decisão.

Como obteve-se um $t_{obs} = 5.993$, então $t_{obs} \in RC$ (verifique a figura 1), desta forma, rejeitamos a hipótese nula. Conclui-se. com base no teste T que existe uma relação estatisticamente significante entre tempo médio de internação (X) e percentual de pacientes infectados (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: Calculo 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 2), desta forma, rejeitamos a hipótese nula. Conclui-se. com base no teste F que existe uma relação estatisticamente significante entre tempo médio de internação (X) e percentual de pacientes infectados (Y).

d) Gráfico de dispersão incluindo o modelo ajustado.

Veja a figura 3.

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 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 4 e 5 para observar se há alguma violação nas hipóteses básicas do modelos, que são: Homecedasticidade, linearidade, normalidade e outliers. Será suposto a independência dos erros aleatórios do modelo.

- Homocedasticidade e Linearidade:
  Pela figura 4 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 linearidade não foi violada.
- Normalidade:
  Pela figura 5, pode-se notar que os quantis dos resíduos studentizados se aproximam dos quantis teóricos de uma distribuição normal com média 0 e variância σ^2, o que é um bom indicativo de normalidade dos erros.
- Outilers:
  Também pela figura 4 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 outilers.

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%.

Apêndice 1 - Tabelas

Tabela 1: Tabela resumo

Estimadores	Estimativa	Erro padrão	Valor da estatística de teste T	P-valor
$\hat{\beta_0}$	-0.14848	0.76072	-0.195	0.846
$\hat{\beta_1}$	0.47580	0.07939	5.993	3.42e-08

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	3.421e-08
Resíduos	120.871	98	1.233		
Total	165.176	99			

Apêndice 2 - Figuras

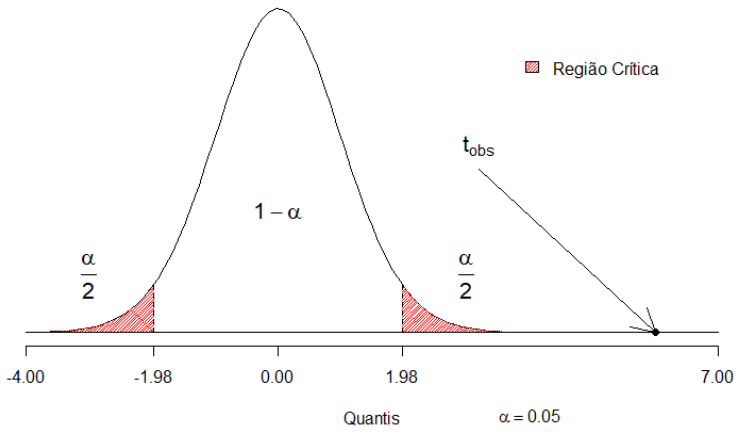

Figura 1: Densidadade e região crítica de T_{98}.

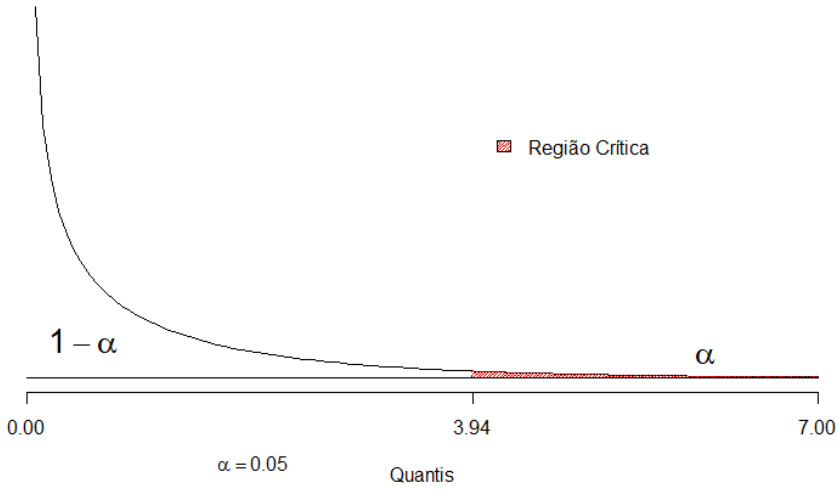

Figura 2: Densidadade e região crítica de $F_{1,98}$.

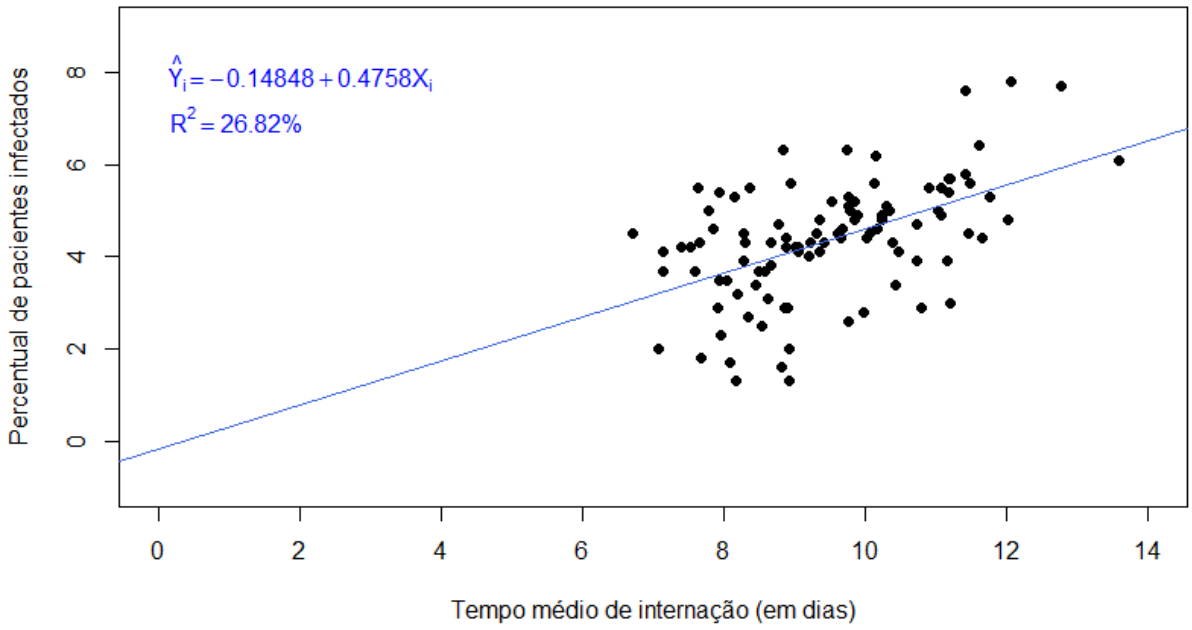

Figura 3: Densidadade e região crítica de $F_{1,98}$.

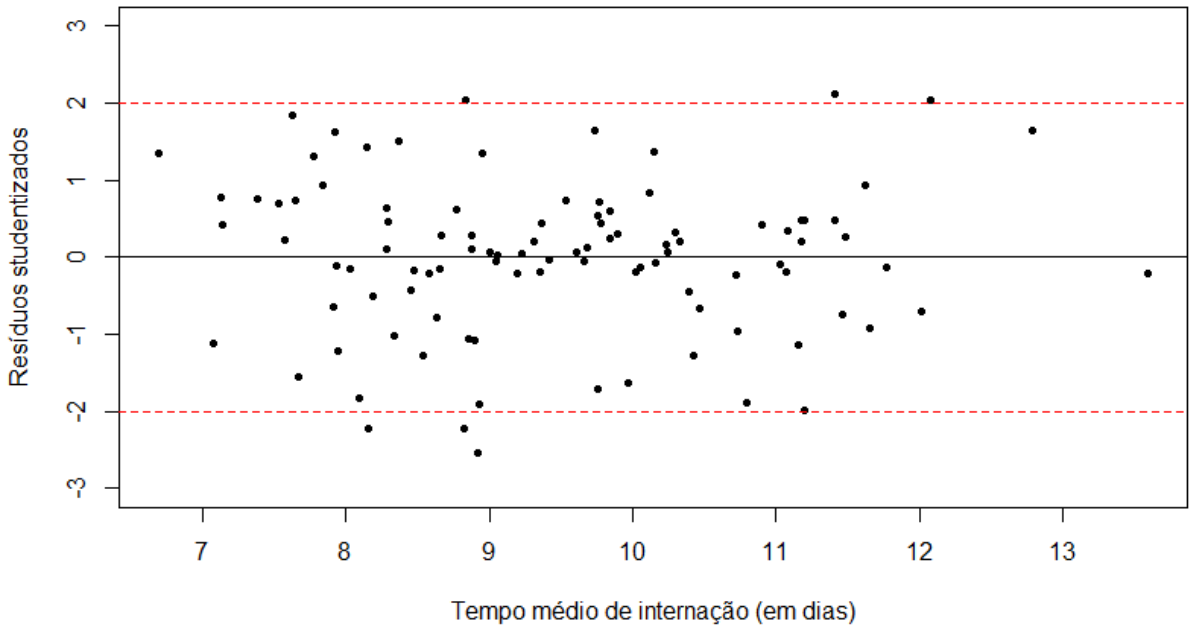

Figura 4: Gráfico dos resíduos stduentizados vs tempo médio de internação em dias.

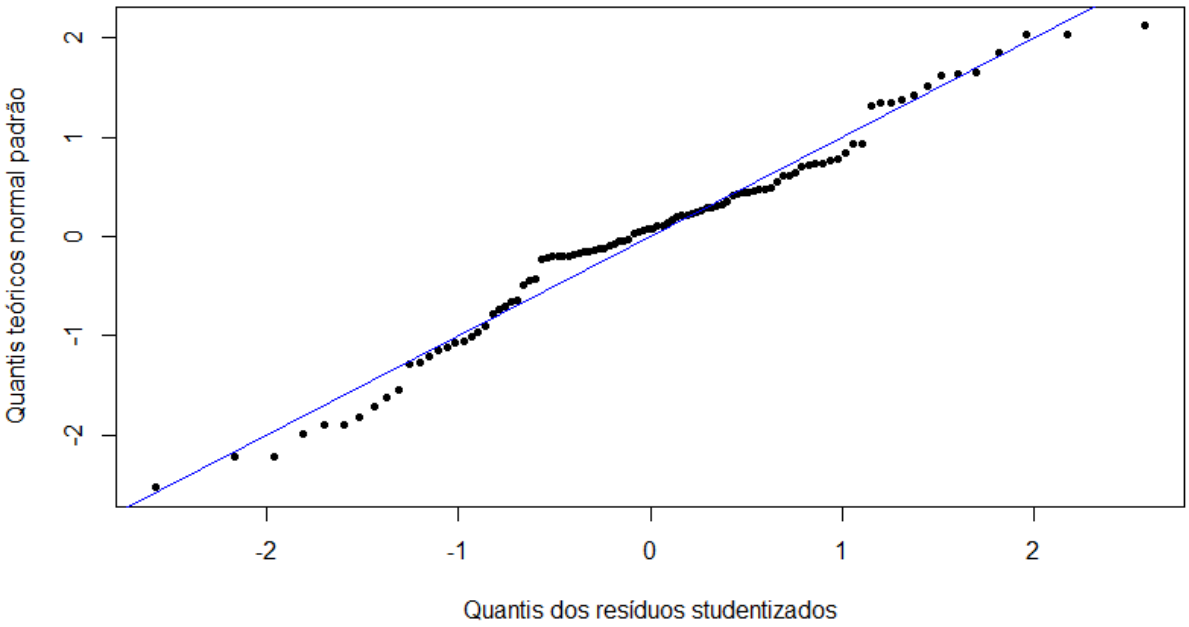

Figura 5: Qqplot entre os quantis teóricos de uma $N(0,1)$ e os resíduos studentizados.

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))
```

```r
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)
```