



Allan Victor Almeida Faria (190127180), Ananda Almeida de Sá (150117345),  
Bruno Marão Raposo (180098888)

## **Trabalho de Regressão Linear**

Trabalho de Regressão Linear de Análise  
de dados hospitalares.

Universidade de Brasília (UnB)  
Instituto de Ciências Exatas (IE)  
Departamento de Estatística (DE)

Brasília, DF

21/02/2021

# Sumário

<b>INTRODUÇÃO</b>	<b>3</b>
Objetivos	3
Metodologia	3
<b>ANÁLISE</b>	<b>5</b>
<b>RESULTADOS</b>	<b>11</b>
Dividir dados	11
Modelagem para Número de enfermeira(o)s	11
Pressupostos para um modelo inicial	12
Box-cox transformation	14
Modelo de Etapas	14
modelo hospital assumptions	15
Comparação	16
Duração da internação	17
Comparação entre modelos	20
<b>CONCLUSÃO</b>	<b>23</b>



## Introdução

### Objetivos

A fim de estudar sobre a duração da internação nos hospitais dos Estados Unidos no período de 1975-1976, foi retirada uma amostra aleatória de 113 hospitais selecionados entre 338 pesquisados, para isso foram propostas as seguintes hipóteses:

A primeira é verificar se o número de enfermeira(o)s está relacionado às instalações, ou seja, os números de leitos do hospital, e se há diferenças entre os serviços disponíveis pelos hospitais. Além de verificar se a mesma variável resposta mencionada anteriormente varia segundo a região.

Já a segunda é verificar se a duração da internação está associada a características do paciente, seu tratamento e do hospital.

### Metodologia

A metodologia neste estudo usa técnicas de modelagem de regressão linear, na tentativa de construir modelos para a predição do número de enfermeira(o)s e a duração da internação. Dentre a proposta do estudo, para avaliar entre diferentes tipos de modelos empregados, utiliza-se métodos de seleção de variáveis como “Feedforward”, “Backward” e “Stepwise” juntamente com métricas como RSS (Residual Sum of Squares),  $cp$  de mallows ( $c(p)$ ), Akaike information criterion (AIC) e testes estatísticos para avaliação destes.

Para a avaliar os pressupostos dos modelos lineares, como a homocedasticidade, normalidade e multicolinearidade, o uso da visualização dos gráficos de resíduos vs os valores preditos para avaliar a homocedasticidade, o teste de normalidade de Shapiro-Wilk e para multicolinearidade o uso do Variance inflation factor (VIF) ou GVIF para o caso generalizado. Sobre pontos influentes na estimação dos parâmetros, a utilização dos DFBetas e DFFits e correlação de pearson também foram utilizadas.

A qualidade do modelo é avaliada sobre os dados de validação com a métrica da raiz do erro quadrático medio (RSME) dado por



$$RMSE = \frac{\sqrt{\sum(\hat{Y} - Y)^2}}{N}$$

que avalia o desvio padrão médio das predições do modelo regressão  $\hat{Y}$ , sobre a variável resposta  $Y$ .

O programa utilizado para analisar os dados, modelar e testar com base nos dados será o R Studio, versão 4.2.0.



## Análise

Sobre a Tabela 1, temos a descrição das variáveis disponíveis pelo banco de dados disponibilizado pelo hospital, no qual apenas 2 variáveis, “Filiação a Escola de Medicina” (X7) e “Região” (X8) são variáveis qualitativas.

Tabela 1 – Descrição das variáveis do banco de dados.

Nome	Código	Descrição	Classificação
<b>Número de Identificação</b>	ID	1-113	Qualitativa ordinal
<b>Duração da Internação</b>	X1	Duração média da internação de todos os pacientes no hospital (em dias)	Quantitativa contínua
<b>Idade</b>	X2	Idade média dos pacientes	Quantitativa contínua
<b>Risco de Infecção</b>	X3	Probabilidade média estimada de adquirir infecção no hospital (em %)	Quantitativa contínua
<b>Proporção de Culturas de Rotina</b>	X4	Razão do número de culturas realizadas com relação ao número de pacientes sem sinais ou sintomas de infecção adquirida no hospital, vezes 100.	Quantitativa contínua



Tabela 1 – Descrição das variáveis do banco de dados.  
(continued)

Nome	Código	Descrição	Classificação
<b>Proporção de Raio-X de Tórax de Rotina</b>	X5	Razão do número de Raio-X de Tórax realizados com relação ao número de pacientes sem sinais ou sintomas de pneumonia, vezes 100.	Quantitativa contínua
<b>Número de leitos</b>	X6	Número médio de leitos no hospital durante o período de estudo	Quantitativa contínua
<b>Filiação a Escola de Medicina</b>	X7	1 – sim 2 – não	Qualitativa nominal
<b>Região</b>	X8	Região Geográfica, onde: 1 – NE 2- NC 3 – S e 4 – W	Qualitativa nominal
<b>Média diária de pacientes</b>	X9	Número médio de pacientes no hospital por dia durante o período do estudo	Quantitativa contínua
<b>Número de enfermeiro(s)</b>	X10	Número médio de enfermeiros(as) de tempo-integral ou equivalente registrados e licenciados durante o período de estudo ( número de tempos integrais+metade do número de tempo parcial)	Quantitativa contínua



Tabela 1 – Descrição das variáveis do banco de dados.  
(continued)

Nome	Código	Descrição	Classificação
<b>Facilidades e serviços disponíveis</b>	X11	% de 35 potenciais facilidades e serviços que são fornecidos pelo hospital	Quantitativa contínua

A Figura 1, temos que a frequência da variável “sim” (1) da Filiação a Escola de Medicina (X7) é desbalanceada ao comparar com a variável “não” (2). O mesmo observa-se para a “Região” (X8) com menor intensidade de desbalanceada entre suas variáveis. Estes fatores na amostragem podem ser prejudiciais na estimação dos parâmetros, assim opta-se a não utilizar X7 como variável explicativa para a modelagem dos problemas.

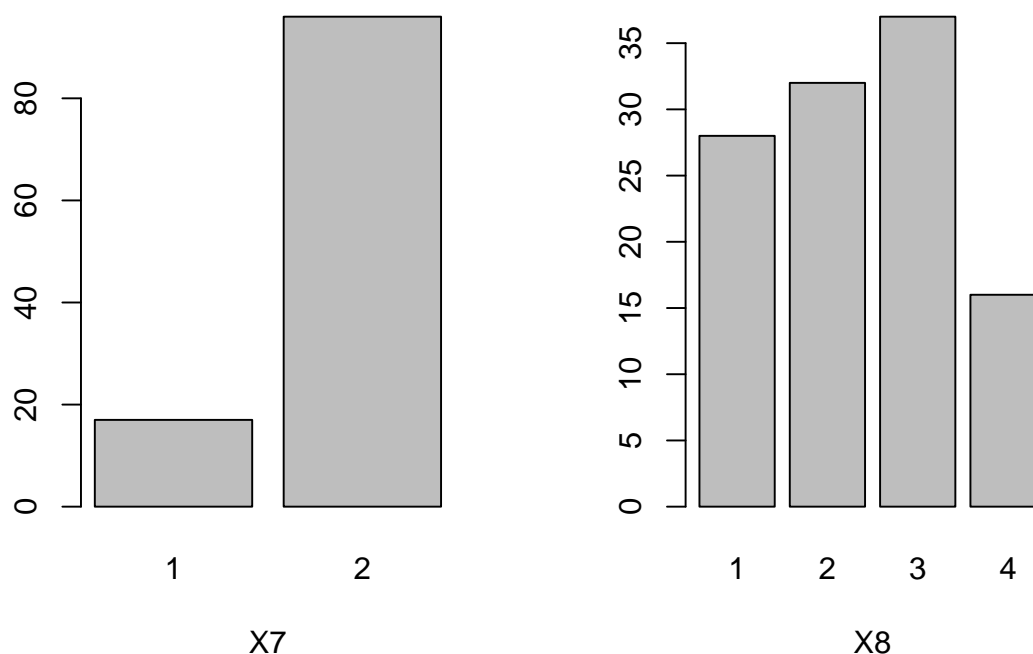


Figura 1 – Frequência dos níveis das variáveis X7 e X8.



Realizando uma breve análise descritiva das variáveis quantitativas, observa-se que na tabela 2 a média destas variáveis são diferentes em magnitude, para avaliar a dispersão destas variáveis, a Figura 2 resume estas variáveis padronizadas, ao passo que esta padronização é feita de tal forma

$$X_{adj} = \frac{X - \bar{X}}{\sigma_X}$$

logo, com esta padronização é possível comparar entre as diferentes magnitudes das variáveis, com a característica de sua dispersão em relação ao desvio padrão da variável, assim na parte de resultados utilizamos deste resultado para a modelagem do problema.

Tabela 2 – Medidas descritivas para boxplots

Variaveis	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Duração da Internação	6.700	8.340	9.420	9.648	10.470	19.560
Idade	38.80	50.90	53.20	53.23	56.20	65.90
Risco de Infecção	1.300	3.700	4.400	4.355	5.200	7.800
Proporção de Culturas de Rotina	1.60	8.40	14.10	15.79	20.30	60.50
Proporção de Raio-X de Tórax de Rotina	39.60	69.50	82.30	81.63	94.10	133.50
Número de leitos	29.0	106.0	186.0	252.2	312.0	835.0
Média diária de pacientes	20.0	68.0	143.0	191.4	252.0	791.0
Número de enfermeiro(s)	14.0	66.0	132.0	173.2	218.0	656.0
Facilidades e serviços disponíveis	5.70	31.40	42.90	43.16	54.30	80.00

As etapas para o estudo da internação dos hospitais foram separadas em duas maneiras, a primeira é a construção e a segunda é a validação do modelo. Para a primeira etapa, foi selecionada uma amostra aleatória simples com 57 observações, para a segunda ficou o restante das observações que compõe o banco. Para as duas hipóteses a modelagem é feita com modelos regressivos lineares múltiplos do tipo:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + e_i, \forall i = 1, \dots, n$$

Onde tem-se,

- $Y_{ij}$  - variável resposta;
- $X_{i1}, X_{i2}, \dots, X_{ik}$  - k variáveis explicativas ou independentes;
- $\beta_0, \beta_1, \beta_2, \dots, \beta_k$  - parâmetros do modelo;
- $e_i$  - são independentes e  $N(0, \sigma^2)$ .



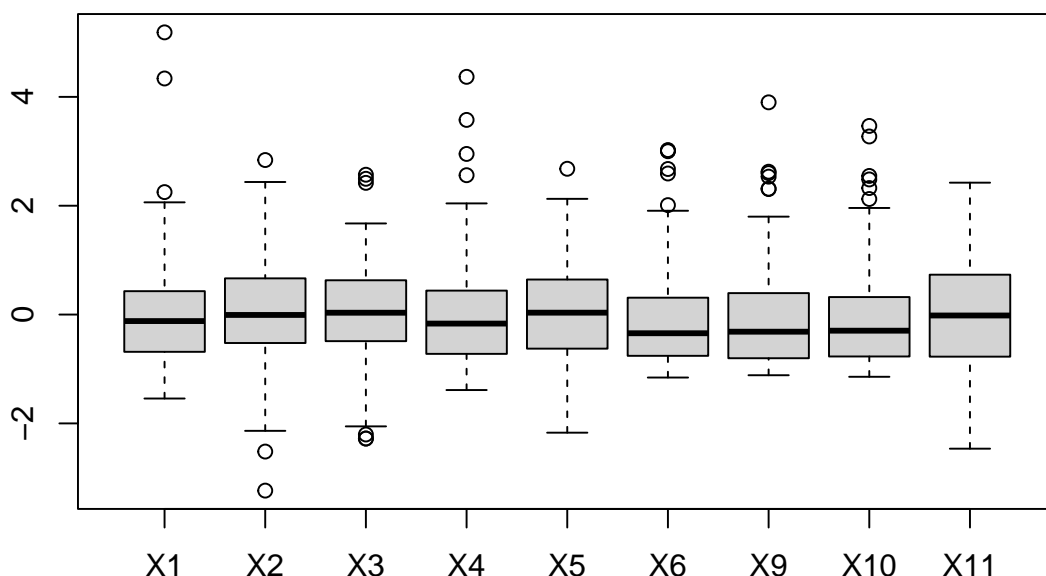


Figura 2 – Gráfico de box-plot das variáveis quantitativas padronizadas.

Para a primeira hipótese, define-se como modelo I aquele que relaciona a variável resposta, Número de enfermeiro(s) (X10), com as variáveis explicativas, instalações (X6), serviços disponíveis pelos hospitais (X11) e a região (X8).

Já o modelo II é definido como aquele que relaciona a variável resposta, Duração da internação (X1), com as variáveis explicativas, a características do paciente (X2), seu tratamento (X4 e X5) e do hospital (X3).

Para verificar a natureza e a força da relação entre as variáveis à identificar lacunas e pontos discrepantes no conjunto de dados, utiliza-se a matriz de correlação.

A figura 3, tem-se que as variáveis que estão nas três extremidades externas dos dois eixos apresentam uma correlação forte, então, X10 com X11, X6 com X11 e X10 e X9 com X11, X10 e X6. A maior correlação é apresentada entre as variáveis X6 e X9, que é o número de leitos e a média diária de pacientes, respectivamente, para as variáveis que possuem o “X” marcando-a, pelo teste de correlação de pearson, estas variáveis, com 95%

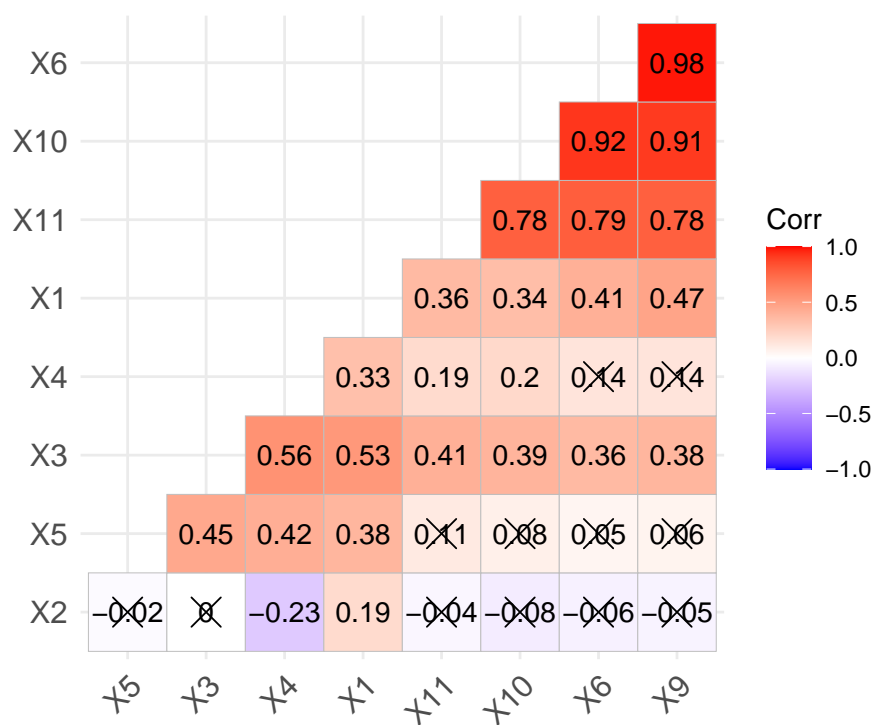


Figura 3 – Gráfico de calor da correlação entre as variáveis dos dados.

de confiança, não possuem correlação.



## Resultados

### Dividir dados

Para efetuar estimar os parâmetros do modelo, separa-se o banco em teste e treino no qual 57 observações foram para o treino, dentre elas na Tabela 3, observa-se que os dados amostrados são proporcionais, assim, ao passo da modelagem, a estimação dos parâmetros será mais representativa.

Tabela 3 – Proporção de observações dos dados no banco de treino e validação sobre a variável X8.

X8	Train Size	Valid Size
1	14	14
2	17	15
3	18	19
4	8	8

### Modelagem para Número de enfermeira(o)s

Para o pressuposto da interação com a região, verifica-se pelo teste ANOVA a dispersão destes dados sobre o número de enfermeiros, no qual a Figura 4, observa-se que os valores centrais estão bem próximos e sobre a Tabela 4, percebe-se que há evidência de não rejeitar a hipótese de igualdade das médias de cada grupo da variável X8,  $\mu_1 = \dots = \mu_4$ , no qual indica que estes valores não influenciam na resposta do número de enfermeiros.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X8	3	14239.94	4746.65	0.22	0.8798
Residuals	53	1126692.10	21258.34		

Tabela 4 – ANOVA para o X8 vs X10.

Deseja-se estudar se o número de enfermeira(o)s está relacionado às instalações, ou seja, os números de leitos do hospital, e se há diferenças entre os serviços disponíveis pelos hospitais. Neste caso, a variável resposta é o número de enfermeira(o)s e as duas outras variáveis são explicativas.

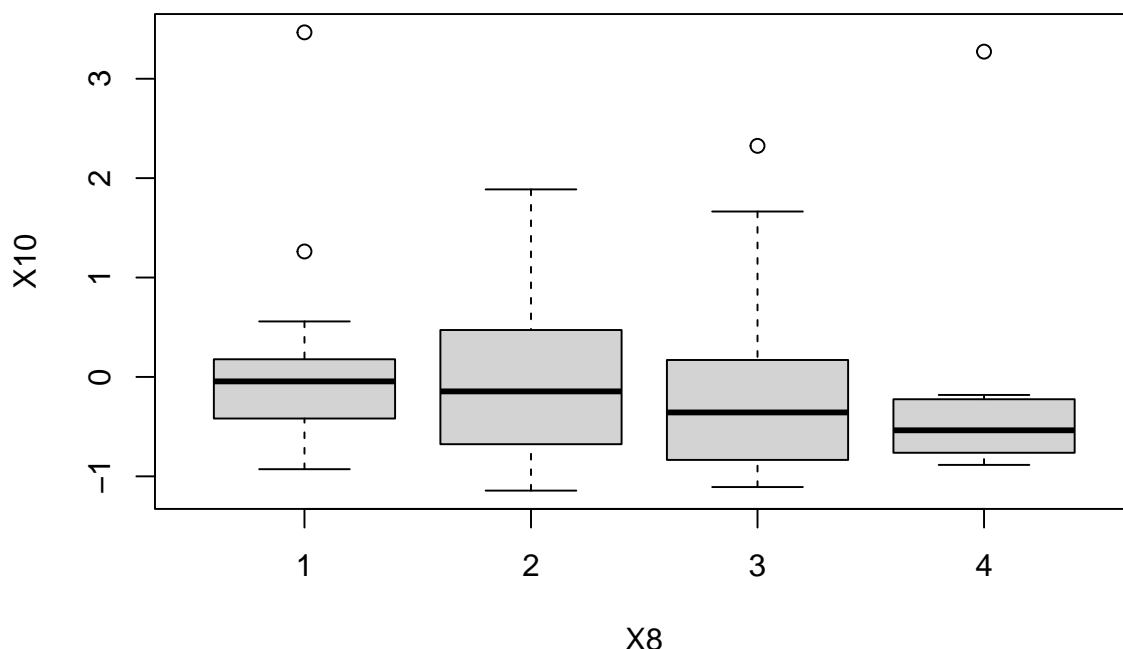


Figura 4 – Box-plot das variáveis resposta X8 com base na X10.

Para isso, faz-se necessário a aplicação da regressão linear múltipla. No qual, na Figura 4, o gráfico da dispersão sobre as variáveis explicativas normalizadas  $X6_{adj}$  e  $X11_{adj}$ , sobre a influência do número de enfermeiros  $X10_{adj}$ , correlacionada com a  $X8_{adj}$ , verifica-se que as aproximações podem ser retas de primeira ordem mas há dispersão não explicada sobre essas retas, com a densidade destes pontos focalizadas na origem.

## Pressupostos para um modelo inicial

Para primeira etapa é definido um modelo inicial para explicar a variável de número de enfermeiros  $X10$  que é dada por

$$\begin{aligned}\hat{y}_{X10} = & \beta_0 + \beta_{X1}X1 + \beta_{X6}X6 + \beta_{X8}X8 + \beta_{X11}X11 \\ & + \beta_{X1,X8}(X1X8) + \beta_{X6,X8}(X6X8) + \beta_{X8}(X8) + \beta_{X11,X8}(X11X8)\end{aligned}$$

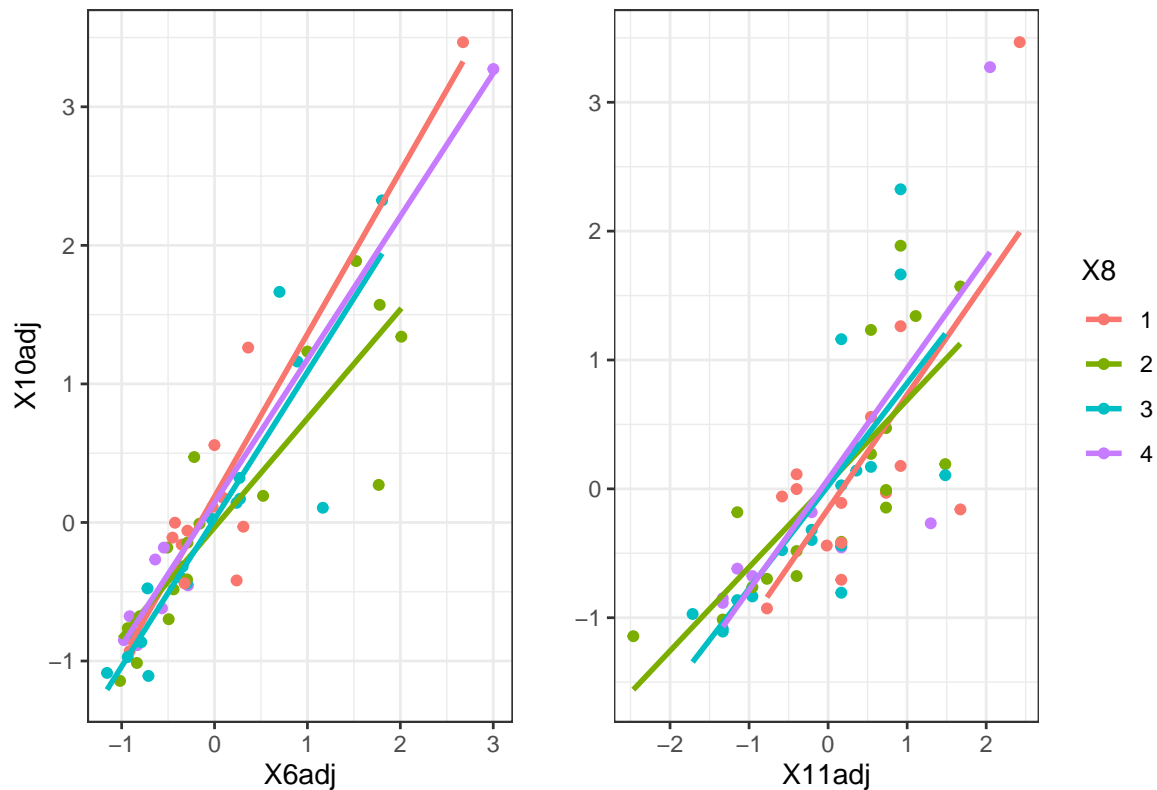


Figura 5 – Gráficos de dispersão das variáveis explicativas normalizadas  $X6_{adj}$ ,  $X8_{adj}$  e  $X11_{adj}$  sobre a variável  $X10_{adj}$ , número de enfermeira(o)s.

no qual presume que o modelo é explicado pela “duração da internação” ( $X1$ ), “Número de leitos” ( $X6$ ), “Facilidades e serviços disponíveis” ( $X11$ ) com a “Região” ( $X8$ ).

Aplicando o teste de ausência de Regressão, obtem-se que as variáveis explicativas com diferença significativa para o modelo são:  $X1$ ,  $X8$ , e  $X6$ . As interações com a região  $X8$ , foram descartadas por estar perto do limite do p-value 0.05.

Agora construindo um novo modelo de regressão

$$\hat{y}_{X10} = \beta_0 + \beta_{X1}X1 + \beta_{X6}X6$$

diferença significativa so em  $X6$ ,

$$X10 = \beta_0 + \beta_{X6adj}X6adj \quad (1)$$



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1adj	1	219271.01	219271.01	79.35	0.0000
X8	3	22596.11	7532.04	2.73	0.0564
X6adj	1	727219.51	727219.51	263.18	0.0000
X11adj	1	6462.84	6462.84	2.34	0.1339
X1adj:X8	3	32650.88	10883.63	3.94	0.0147
X8:X6adj	3	15046.20	5015.40	1.82	0.1595
X8:X11adj	3	4395.21	1465.07	0.53	0.6641
Residuals	41	113290.29	2763.18		

Tabela 5 – Modelo - ANOVA

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	180.6618	7.6879	23.50	0.0000
X1adj	9.6295	10.6282	0.91	0.3689
X6adj	132.1270	8.8946	14.85	0.0000

Tabela 6 – Modelo de regressão -  $X10 = X1_{adj} + X6_{adj}$

Testando a normalidade do modelo atual encontramos um (p-value: 0.0026539), no qual rejeitamos a normalidade. Dessa forma sera aplicado uma transformação nos dados para ajustar sua normalidade.

$$X10 = \beta_0 + \beta_{X6adj}X6adj \quad (1)$$

## Box-cox transformation

A transformação box-cox é dada por  $Y_{cox} = \frac{(Y^\lambda - 1)}{\lambda}$  no qual para o Número de enfermeira(o)s ( $X10$ ) esta transformação é dada por  $X10_{cox} = \frac{(X10^\lambda - 1)}{\lambda}$  onde lambda é igual  $\lambda = 0.5858586$  e assim o shapiro teste para os residuos do modelo com box-cox (p-value:0.3074418).

Logo o modelo é dado por

$$X10_{cox} = \beta_0 + \beta_{X6adj}X6adj \quad (\text{Modelo 1.1})$$

## Modelo de Etapas

(PENSANDO EM RETIRAR ESSA SECAO)



## modelo hospital assumptions

Levando em consideração a Hipotese do Hospital, o modelo será definido como:

$$X10_{\text{cox}} = \beta_0 + \beta_{X6\text{adj}}X6\text{adj} + X6\text{adj}^2 + \beta_{X11\text{adj}}X11\text{adj} + \beta_{X11\text{adj}}X11\text{adj}^2 + \beta_{X8}X8 \quad (\text{Modelo 1.3})$$

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	34.6154	1.7953	19.28	0.0000
X6adj	14.0058	1.6170	8.66	0.0000
I(X6adj^2)	-0.5905	0.8758	-0.67	0.5033
X3adj	2.8509	0.9815	2.90	0.0055
I(X3adj^2)	-0.8250	0.5256	-1.57	0.1229
X82	-3.0844	2.1414	-1.44	0.1561
X83	-1.5135	2.1923	-0.69	0.4932
X84	-1.2894	2.7867	-0.46	0.6456

Tabela 7 – Regressão do modelo 1.3

Ajustando o modelo de acordo com a Regressão:

$$X10_{\text{cox}} = \beta_0 + X6\text{adj} + \beta_{X3\text{adj}}X3\text{adj} \quad (\text{Modelo 1.4})$$

temos que a normalidade do modelo é atendida (p-value 0.0576297), não existe multicolinearidade (VIF= (1.15,1.15)) e  $R^2 = 0.8534112$ , ou seja, o modelo atende as pressuposições básicas.

	VIF
X6adj	1.15
X3adj	1.15

Tabela 8 – VIF para o modelo 1.4

	df	AIC
Modelo 1.3	9.00	372.76
Modelo 1.4	4.00	369.36

Comparando os modelos com teste linear geral e o AIC, o teste linear geral dado por 0.3214056, o que indica que não há diferença entre o modelo linear e o modelo de segunda ordem. Temos que o modelo proposto pelo hospital não tem diferença significativa, e assim, o modelo escolhido foi o que possui primeira ordem (1.4).



## Comparação

Agora comparando os modelos 1.4 e 1.1, o teste linear geral sob o p-value  $3.264323 \times 10^{-4}$  temos que evidência para acreditar que existe diferença significativa entre os dois modelos, e assim, na tabela 9, observa-se que o modelo 4 possui valores mais baixos nas métricas AIC, RSS e RMSE onde representa o erro quadrático médio do modelo aplicado no banco de validação.

	df	AIC	RSS	RMSE	R2
Modelo 1.1	3.00	381.12	2407.13	6.56	0.83
Modelo 1.4	4.00	369.36	1891.07	6.61	0.83

Tabela 9 – Métricas avaliadas nos dados de validação sob os modelos 1.1 e 1.4.

Comparando os modelos 1.4 e 1.1, o teste linear geral sob o p-value  $3.264323 \times 10^{-4}$  temos que evidência para acreditar que existe diferença significativa entre os dois modelos, e assim, na tabela 9, observa-se que o modelo 4 possui valores mais baixos nas métricas AIC, RSS e RMSE onde representa o erro quadrático médio do modelo aplicado no banco de validação. Dessa forma, na figura 6 temos que o modelo é adequado e segue os pressupostos da regressão.

$$X10_{\text{cox}} = \beta_0 + \beta_{X6\text{adj}}X6\text{adj} + \beta_{X3\text{adj}}X3\text{adj} \quad (\text{Modelo 1.4})$$

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	31.6669	0.7839	40.40	0.0000
X6adj	13.2987	0.8796	15.12	0.0000
X3adj	3.2335	0.8423	3.84	0.0003

Tabela 10 – Regressão do modelo

$$X10_{\text{cox}} = 31.6669 + 13.2987X6\text{adj} + 3.2335X3\text{adj} \quad (\text{Modelo 1.4})$$



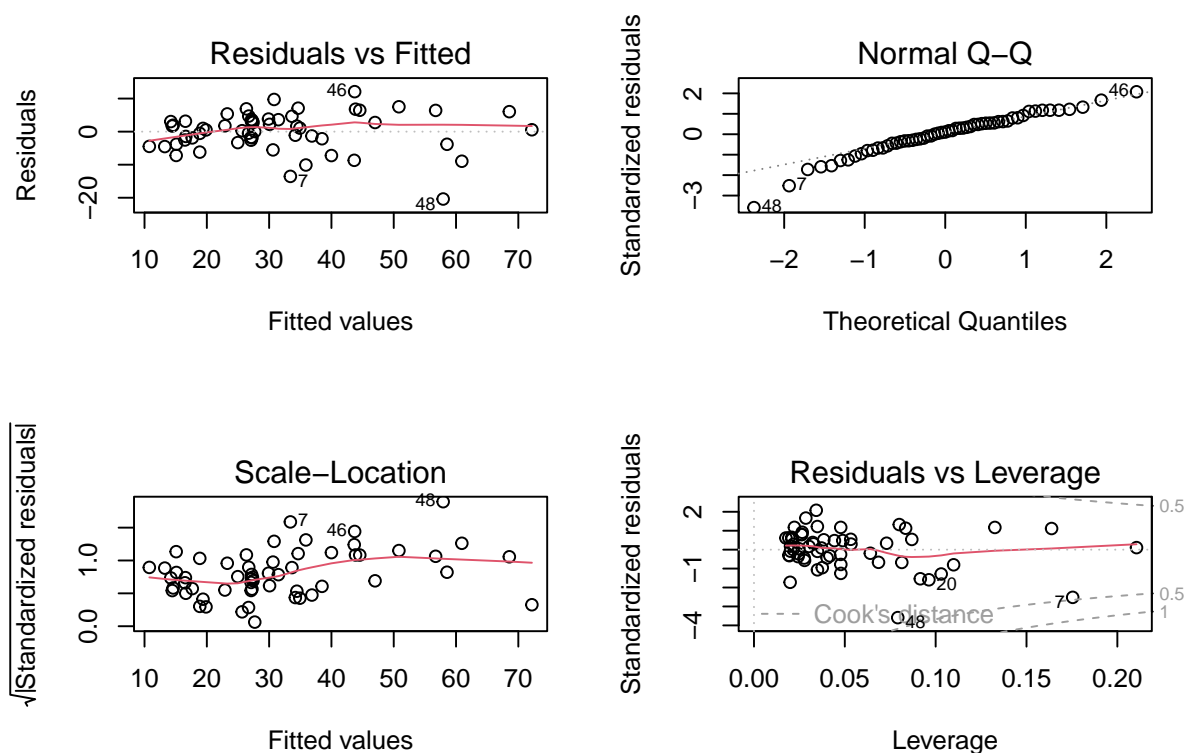


Figura 6 – Análise de valores residuais do modelo 1.4.

## Duração da internação

Como já mencionado anteriormente, o intuito do modelo II seria avaliar as possíveis relações que as variáveis explicativas apresentam em relação a duração de internação ( $X_1$ ). Para isso, realizando o teste ANOVA, a dispersão destes dados sobre a duração de internação, apresentados os resultados na Tabela 10 e considerando o nível de significância de 5%, temos que pelos testes de igualdade de parâmetros iguais a zero, deu diferença significativa as variáveis explicativas  $X_3$ ,  $X_6$  e  $X_9$ .

Considerando as variável explicativa, tem-se o modelo inicial dado por:

$$Y_{X1adj} = \beta_0 + \beta_1 X_{3adj} + \beta_2 X_{6adj} + \beta_3 X_{9adj} + e_i, \forall i = 1, \dots, n$$

Utilizando o Fator de Inflação da Variância (VIF), obtém-se os valores para  $X_3$ ,  $X_6$  e  $X_9$ , através da Tabela 11, como 1,19, 30,10 e 30,74, respectivamente. Pode-se analisar que o máximo dele é maior que 10, indicando assim a presença de multicolinearidade.



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X2adj	1	0.08	0.08	0.29	0.5902
X3adj	1	16.33	16.33	58.71	0.0000
X4adj	1	1.10	1.10	3.97	0.0521
X5adj	1	0.12	0.12	0.43	0.5132
X6adj	1	1.49	1.49	5.35	0.0251
X9adj	1	2.97	2.97	10.69	0.0020
X10adj	1	0.61	0.61	2.19	0.1458
X11adj	1	0.40	0.40	1.43	0.2380
Residuals	48	13.35	0.28		

Tabela 11 – ANOVA - model pre

Conclui-se que a variável X6 e X9 estão correlacionadas, o que é verificado anteriormente através da força do resultado do coeficiente de correlação.

	VIF
X3adj	1.19
X6adj	30.10
X9adj	30.74

Tabela 12 – VIF Modelo inicial

Com isso, não é necessário o uso das duas variáveis, basta uma para representar no modelo. Para selecionar qual a variável é a melhor para representar, avalia-se primeiramente as correlações da variável resposta com as duas explicativas, através da Figura 3 obtida anteriormente, o valor para X1 e X6 é de 0,43, enquanto o de X1 e X9 é de 0,48. Com isso, temos que pela análise dos coeficientes de correlação, a variável X9 apresenta uma maior correlação com a variável resposta, então é a sugerida para a validação. Ainda na seleção de variáveis entre X6 e X9, tem-se os modelos para avaliar qual prediz melhor:



$$X1 = \beta_0 + \beta_{X3}X3adj + \beta_{X9}X9adj \quad (\text{Modelo 2.1})$$

$$X1 = \beta_0 + \beta_{X3}X3adj + \beta_{X6}X6adj \quad (\text{Modelo 2.4})$$

	df	AIC	RSS	RMSE	R2	Normal Test
Modelo 2.1	4.00	104.23	18.05	9.98	0.30	0.91
Modelo 2.4	4.00	106.44	18.77	9.99	0.26	0.91

Tabela 13 – Metricas avaliadas nos dados de validação sob os modelos 2.1 e 2.4.

Com a tabela acima, é possível verificar através do coeficiente de determinação, também conhecido como  $R^2$ , que, novamente, a variável explicativa que melhor se adequa ao modelo é a variável X9, visto que o valor do coeficiente foi maior que o de X6, tanto no modelo completo como no ajustado. Por fim, para finalizar a seleção das variáveis, tem-se o modelo ajustado:

$$Y_i = \beta_0 + \beta_1 X3adj + \beta_2 X9adj + e_i, \forall i = 1, \dots, n$$

Onde a segunda variável explicativa irá variar entre a X6 e a X9, comparando o modelo inicial com o ajustado, aplicando o Cp de Mallows, tem-se que os valores obtidos para X6 e X9 são aproximadamente 6,57 e 8,85, respectivamente. Com isso, para realizar a escolha entre os modelos, utiliza-se o que apresentou o menor valor do critério, novamente, a variável explicativa X9 é escolhida. Portanto, tem-se o modelo ajustado com a escolha da segunda variável explicativa é determinado pelo Modelo 2.1.

Analisando o modelo ajustado e utilizando o Fator de Inflação da Variância (VIF), obtém-se os valores para X3 e X9 como 1,17 e 1,17, respectivamente. Pode-se analisar que o máximo dele não é maior que 10, indicando assim a ausência de multicolinearidade. Verifica-se a normalidade do modelo através do teste de Shapiro-Wilk, tem-se que o p-valor obtido é de aproximadamente 0,9, o que considerando o nível de significância de 5%, o modelo apresenta distribuição normal. Aplicando o  $R^2$ , tem-se ainda que o valor obtido foi de 30%, o que é um valor fraco. Por fim, outra maneira analisar o modelo é através da raiz do erro quadrático médio (RMSE) que apresentou valor de 9.98, que é considerado um modelo bom.

Logo, tem-se que a duração de internação está relacionada tanto com o risco de infecção como com a média diária de pacientes. Mas ainda entrando na questão de



	RMSE	R2
Modelo 2.1	9.98	0.30

Tabela 14 – Metricas avaliadas nos dados de validação sob o modelo 2.1

avaliar se existem outras variáveis explicativas que estão correlacionadas com a duração, é recomendado utilizar outras maneiras de fazer essa modelagem, como os métodos Stepwise, Forward e o Backward. Aplicando os três métodos, conclui-se que todos os geraram o mesmo modelo:

$$X1adj = \beta_0 + \beta_{X3}X3adj + \beta_{X8}X8 + \beta_{X9}X9adj + \beta_{X11}X11adj \quad (\text{Modelo 2.2})$$

Verifica-se a normalidade do modelo através do teste de Shapiro-Wilk, tem-se que o p-valor obtido é de aproximadamente 0,7, o que considerando o nível de significância de 5%, o modelo apresenta distribuição normal. Com auxílio da Tabela 17, analisando  $R^2$ , tem-se ainda que o valor obtido foi de 46%, o que é um valor moderado. E finalmente, a raiz do erro quadrático médio (RMSE) apresentou valor de 1,64, que é considerado um modelo bom.

	RMSE	R2
Modelo 2.2	9.92	0.46

Tabela 15 – Metricas avaliadas nos dados de validação sob o modelo 2.2

Analisando o modelo ajustado e utilizando o Fator de Inflação da Variância (VIF), obtém-se os valores para X3, X8, X9 e X11 os 4 deram abaixo de 10. Pode-se analisar que o máximo dele não é maior que 10, indicando assim a ausência de multicolinearidade.

## Comparação entre modelos

Agora, comparando os modelos obtidos 2.1 e 2.2, tem-se os seguintes valores:

	df	AIC	RSS	RMSE	R2	Normal Test
Modelo Step-wise	8.00	88.29	11.86	9.92	0.46	0.70
Modelo Secionado H.	4.00	104.23	18.05	9.98	0.30	0.91

Tabela 16 – Metricas avaliadas nos dados de validação sob os modelos 2.1 e 2.4.

Por fim, analisando o modelo obtido através da ausência de regressão e o novo modelo, aplicando o teste de ANOVA entre eles e analisando soma de quadrados residual



(RSS), obtém-se os valores 18.05 e 11.86, respectivamente. Como, considerando o p-valor do teste, 0,0002, é menor que o nível de significância de 5%, é possível realizar uma escolha entre os dois métodos. Dessa forma, escolhe-se o modelo novo, já que, apresentou o RSS menor. Analisando ainda o Critério da Informação de Akaike (AIC), obtém-se para o modelo de ausência de regressão e para o novo modelo foi de 104.23 e 88.29, respectivamente, com isso, escolhe-se o novo modelo por apresentar o menor valor.

$$X1adj = \beta_0 + \beta_{X3}X3adj + \beta_{X8}X8 + \beta_{X9}X9adj + \beta_{X11}X11adj \quad (\text{Modelo 2.2})$$

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.4685	0.1363	3.44	0.0012
X3adj	0.4388	0.0743	5.91	0.0000
X82	-0.5549	0.1807	-3.07	0.0035
X83	-0.5910	0.1843	-3.21	0.0023
X84	-1.0330	0.2195	-4.71	0.0000
X9adj	0.3711	0.1092	3.40	0.0013
X11adj	-0.2210	0.1031	-2.14	0.0369

Tabela 17 – Regressão do modelo

$$X1adj = 0.4685 + 0.4388X3adj + \beta_{X8}X8 + 0.3711X9adj - 0.2210X11adj \quad (\text{Modelo 2.2})$$

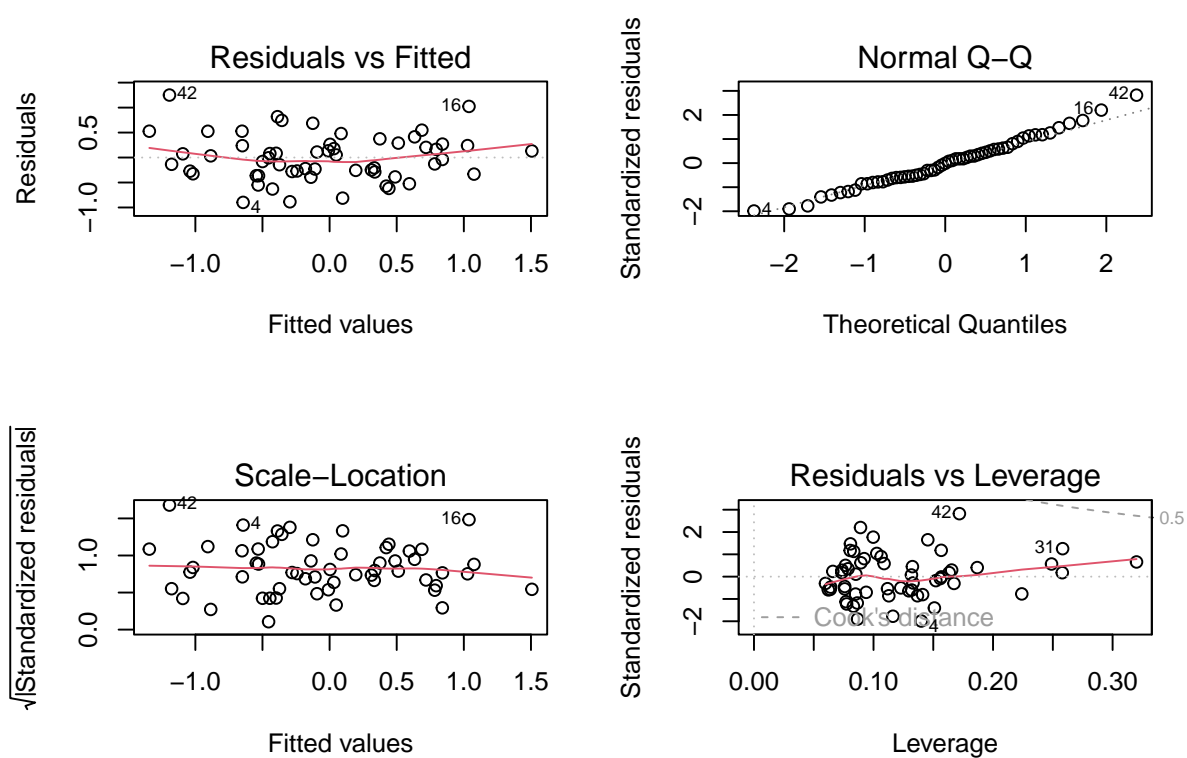


Figura 7 – Análise de valores residuais do modelo 1.4.



## Conclusão