Allan Victor Almeida Faria (190127180), Ananda Almeida de Sá (150117345),
Bruno Kevyn Andrade de Souza

Trabalho de Regressão Linear

Brasília, DF

21/02/2021

Allan Victor Almeida Faria (190127180), Ananda Almeida de Sá (150117345),
Bruno Kevyn Andrade de Souza

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

Resumo

resumo aqui

Palavras-chaves: 1. Análise de dados.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
SAEB	Sistema de Avaliação da Educação Básica

Lista de símbolos

σ	Letra grega minúscula sigma
μ	Letra grega minúscula mu

Sumário

1 Introdução

Tipo de problema, tipo de dados, proposta para contornar o problema

1.1 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.

1.2 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 "Feedfoward", "Backfoward" 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 lineres, como a homocedasticidade, normalidade e multicolinearidade, os testes de homocedasticidade de bartlett, 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.

E para avaliar a qualidade do modelo, é utilizado os dados de validação sobre a raiz do erro quadrático medio (RSME) dado por

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

para avaliar nas predições do modelo regressão $\hat{Y}$.

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

2 Análise

Sobre a Tabela 1, temos a descrição das variáveis disponíveis pelo banco de dados disponibiilizado 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
Número de Identificação	ID	1-113	Qualitativa ordinal
Duração da Internação	X1	Duração média da internação de todos os pacientes no hospital (em dias)	Quantitativa contínua
Idade	X2	Idade média dos pacientes	Quantitativa contínua
Risco de Infecção	X3	Probabilidade média estimada de adquirir infecção no hospital (em %)	Quantitativa contínua
Proporção de Culturas de Rotina	X4	Razão do número de culturas realizadas com relação ao número de pacientes sem sinais ou sintomas de infeção adquirida no hospital, vezes 100.	Quantitativa contínua

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

Nome	Código	Descrição	Classificação
Proporção de Raio-X de Tórax de Rotina	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
Número de leitos	X6	Número médio de leitos no hospital durante o período de estudo	Quantitativa contínua
Filiação a Escola de Medicina	X7	1 – sim 2 – não	Qualitativa nominal
Região	X8	Região Geográfica, onde: 1 – NE 2- NC 3 – S e 4 – W	Qualitativa nominal
Média diária de pacientes	X9	Número médio de pacientes no hospital por dia durante o período do estudo	Quantitativa contínua
Número de enfermeiro(s)	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
Facilidades e serviços disponíveis	X11	% de 35 potenciais facilidades e serviços que são fornecidos pelo hospital	Quantitativa contínua

Para a Figura 1, temos que a frequência da váriavel "sim" (1) da Filiação a Escola de Medicina (X7) é desbalanceada ao comparar com a váriavel "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.

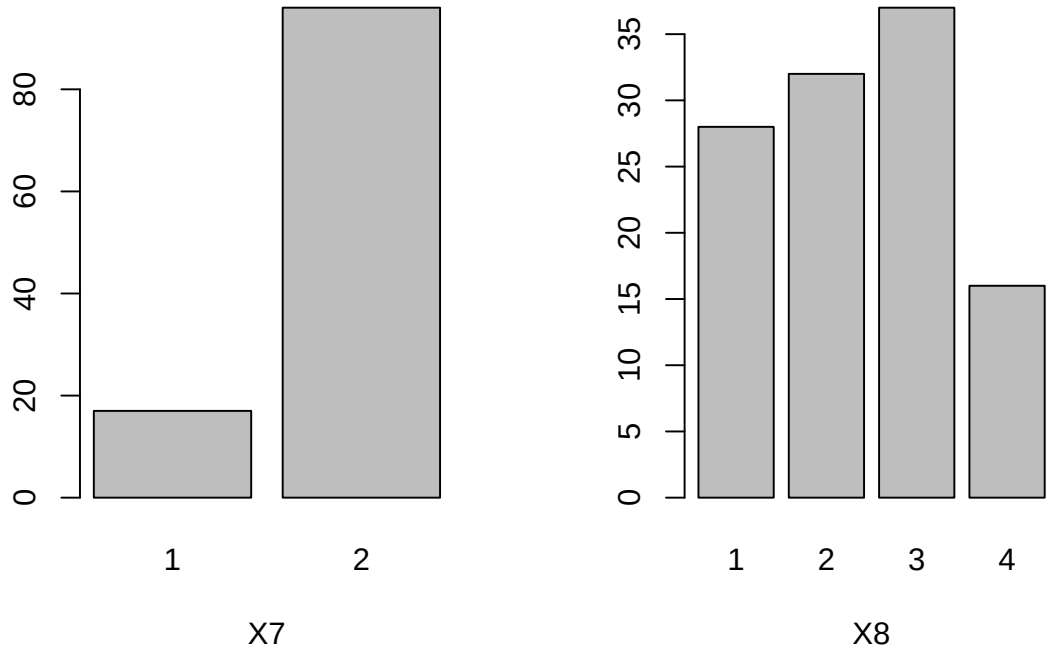

Figura 1 – Frequência de valores para as variáveis X7 e X8.

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 e feita de tal forma

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

logo, com esta padronização é possivel comparar entre as diferentes magnitudes das variáveis, com a característica de sua disperção 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 segundo ficou o restante das observações que compõe o banco. Para as duas hipóteses a modelagem é feita com modelos regressivos lineares múltiplos do tipo:

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

Onde tem-se,

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

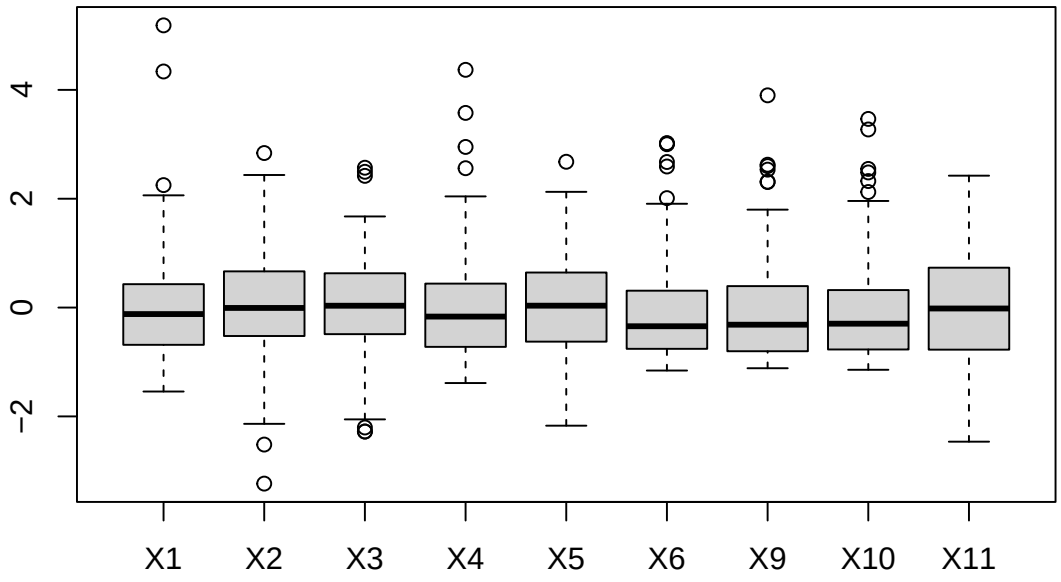

Figura 2 – Gráfico de box-plot das variaváveis quantitativas padronizadas.

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, estás variáveis, com 95%

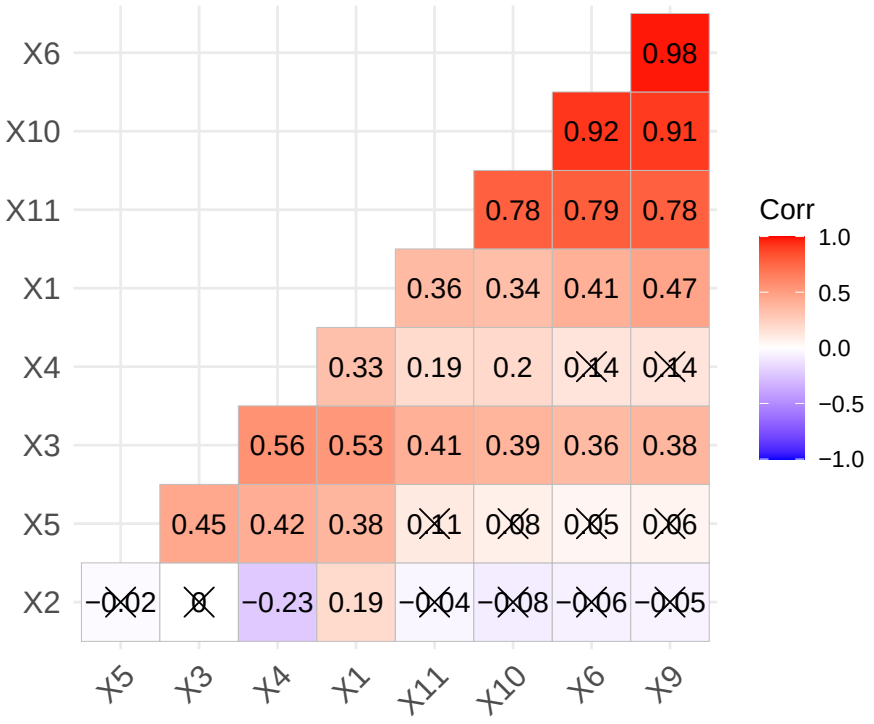

Figura 3 – Gráfico de calor da correlação entre as variaváveis dos dados.

de confiança, não possuem correlação.

3 Resultados

3.1 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

3.2 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, $\mu_1 = ... = \mu_4$, no qual indica que estes valores não influênciam 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.

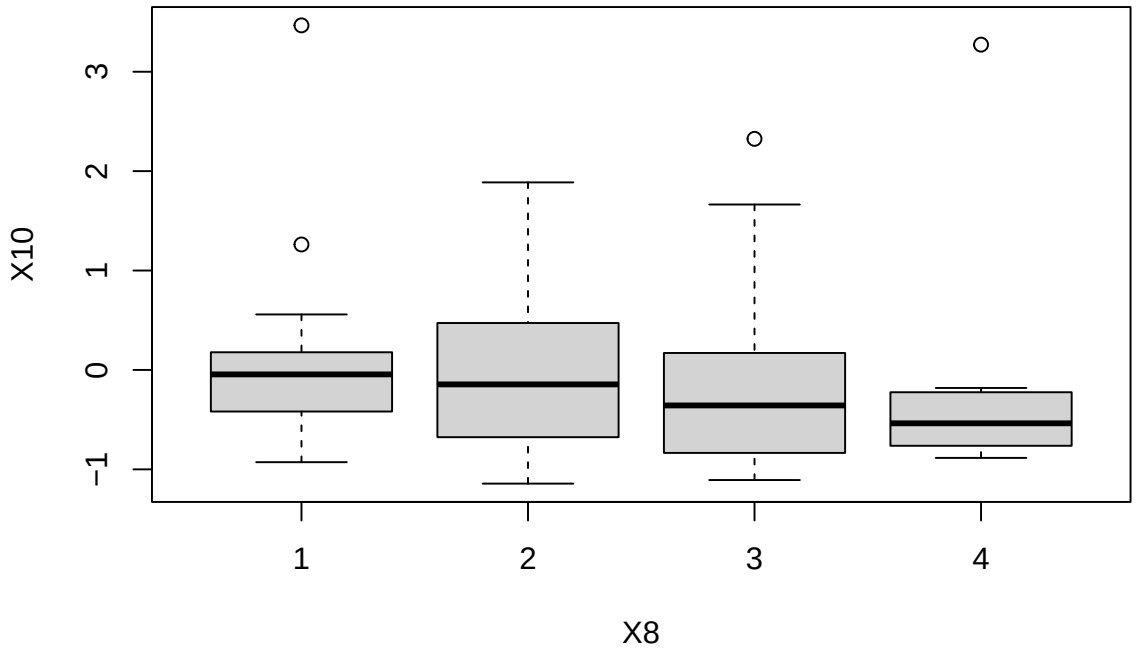

Figura 4 – Box-plot das variaveil resposta X8 com base na X10.

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 disperção não explicada sobre essas retas, com a densidade destes pontos focalizadas na origem.

3.2.1 Pressupostos para um modelo inicial

Agora presumindo um modelo inicial para explicar a variável de número de enfermeiros $X10$ é dada por

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

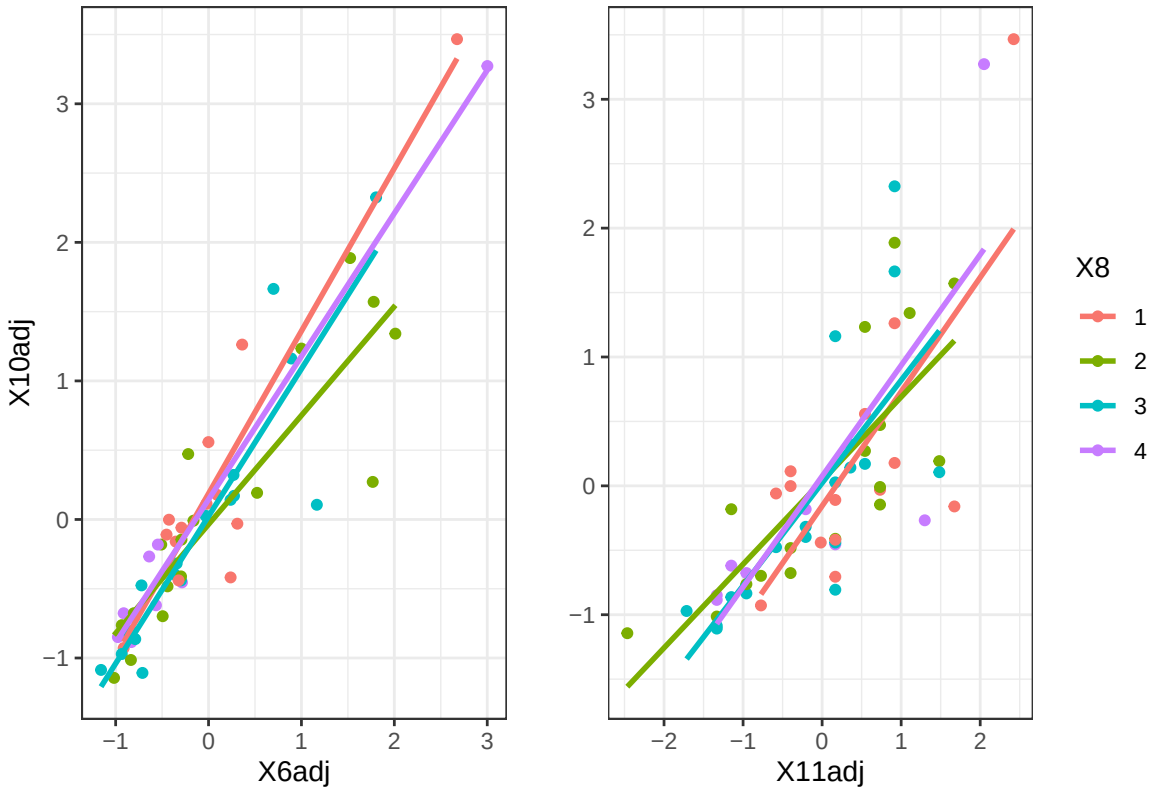

Figura 5 – Gráficos de disperção das variáveis explicativas normalizadas $X6_{adj}$, $X8_{adj}$ e $X11_{adj}$ sobre a variável $X10_{adj}$, número de enfermeira(o)s.

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

	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

agora os resultados obtidos pela ANOVA, na Tabela 5, temos que pelos testes de igualdade de parâmetros iguais a zero, deu diferença significativo as variáveis explicativas

$X1$, $X8$, $X6$ e $X11$. 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

$$\hat{y}_{X11} = \beta_0 + \beta_{X1}X1 + \beta_{X6}X6 + \beta_{X8}X8 + \beta_{X11}X11$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	139.7070	43.2170	3.23	0.0022
X1adj	8.8045	12.4757	0.71	0.4836
X6adj	120.0347	13.5505	8.86	0.0000
X82	-22.4495	22.8693	-0.98	0.3310
X83	-9.3718	23.6867	-0.40	0.6940
X84	9.2307	29.9392	0.31	0.7591
X11	1.1173	0.7993	1.40	0.1683

Tabela 6 – Modelo de regressão - $X10 = X1_{adj} + X6_{adj} + X8 + X11_{adj}$

temos que os resultados da regressão na tabela 6 com valor do F-statistics, para o teste linear geral, percebe-se que o teste de regressão é significativo, indicando que há regressão nesses dados, os parâmetros $X6$ e $X11$ tem diferenças significativas, podendo descartar os outros. Mas ao teste de normalidade, (p-value: 0.0026539), no qual rejeitamos a normalidade, um novo é essencial para adequar aos préssupostos do regressão.

$$X10 = \beta_0 + \beta_{X6adj}X6adj \tag{1}$$

3.2.2 Box-cox transformation

A transformação box-cox é dada por

$$Y_{\text{cox}} = \frac{(Y^{\lambda} - 1)}{\lambda}$$

no qual para o Número de enfermeira(o)s ($X10$) esta transformação é dada por

$$X10_{\text{cox}} = \frac{(X10^{\lambda} - 1)}{\lambda}$$

onde lambda é igual $\lambda = 0.5858586$ e assim o shapiro teste para o box-cox (p-value:0.3074418)

agora avalindo este modelo temos que o erro medio das previsões é baixo e o R2 no banco de teste é alto, assim sendo um bom modelo para começar e avaliar com as suposições do hospital

logo o modelo é dado por

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

Agora avalindo o modelo no banco de teste, temos que a raiz do erro quadratico médio e dado por

3.2.3 Modelo de Etapas

(PENSANDO EM RETIRAR ESSSA SECAO)

3.2.4 modelo hospital assumptions

Agora como a primeira modelagem sonre o modelo baseado nas ideias do hospital temos que,

$$X10_{\text{cox}} = \beta_0 + \beta_{X6adj} X6adj + X6adj^2 + \beta_{X11adj} X11adj + \beta_{X11adj} X11adj^2 + \beta_{X8} X8 \qquad \text{(Modelo 1.3)}$$

mas ao analisar a tabela 7, temos que o modelo mais parcimonioso é dado pelo modelo

	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

$$X10_{\text{cox}} = \beta_0 + X6adj + \beta_{X3adj} X3adj \qquad \text{(Modelo 1.4)}$$

a normalidade do modelo é atendida (p-value 0.0576297) e a multicolinearidade dada pela tabela 8, observa-se que é um modelo que atende as pressuposições básicas e possui o $R^2 = 0.8534112$, indicando que seja um bom modelo.

	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

Agora avaliando o 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.

3.2.5 Comparação

Agora comparando os modelos 1.4 e 1.1, o teste linear geral sob o p-value 3.264323×10^{-4} temos que evidência para acrediatar que existe diferença siginificativa entre os dois modelos, e assim, na tabela 9, observa-se que o modelo 4 possui vallores 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 – Metricas avaliadas nos dados de validação sob os modelos 1.1 e 1.4.

assim na figura 6 temos que o modelo é adequado e segue os pressupostos da regressão.

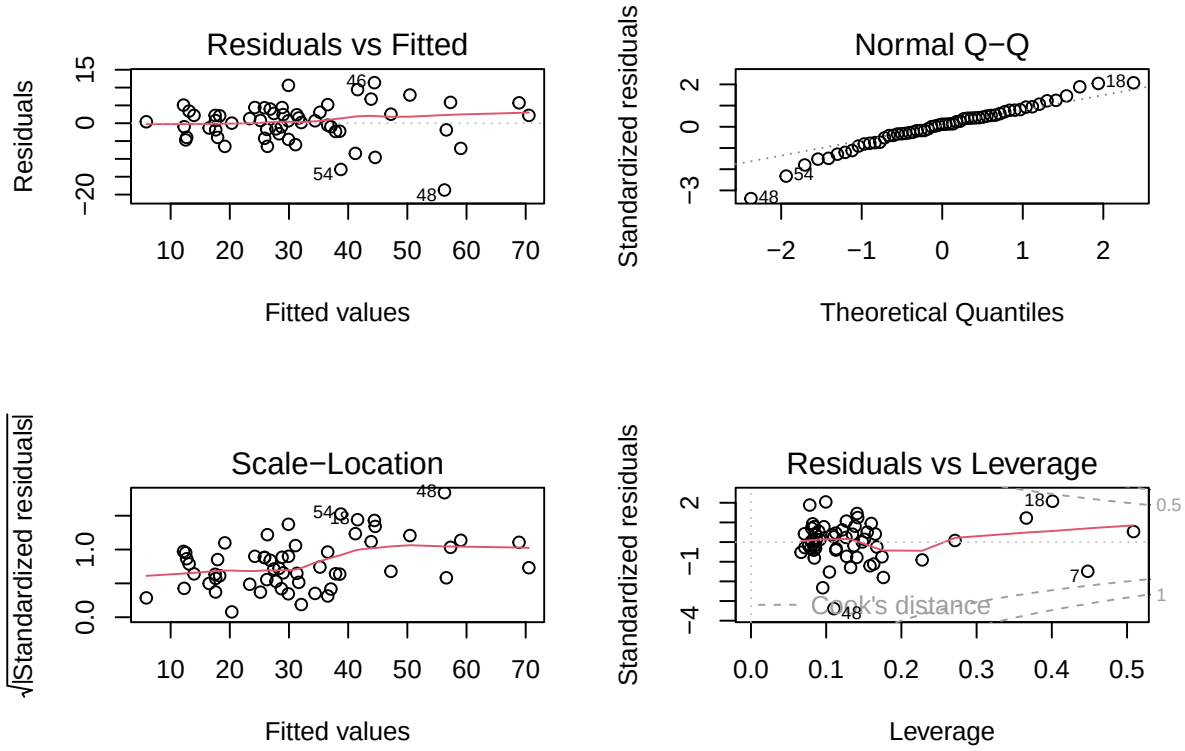

Figura 6 – Análise de valores residuais do modelo 1.4.

3.3 Duração da internação

A duração da internação está associada a características do paciente, seu tratamento e do hospital

características do paciente:X2, seu tratamento:X4,X5 hospital:X3,X6,X7,X9,X10, X11

Deseja-se estudar se a Duração da internação está associada a características do paciente, seu tratamento e do hospital, ou seja, a duração da internação, 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. Para isso, faz-se necessário a aplicação da regressão linear múltipla realizada no script a seguir:

3.3.1 MODELO POR HIPOTESE

O modelo por hipotese de ANOVA:

	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 10 – ANOVA - model pre

O modelo inicial dado por:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0351	0.0743	-0.47	0.6388
X3adj	0.4330	0.0810	5.34	0.0000
X6adj	-0.9105	0.4259	-2.14	0.0372
X9adj	1.1407	0.4359	2.62	0.0115

Tabela 11 – Regressao - model 2.1

O VIF do modelo inicial:

	VIF
X3adj	1.19
X6adj	30.10
X9adj	30.74

Tabela 12 – VIF Modelo inicial

Correlação entre os dados:

	corr	value
1	X3,X9	0.39
2	X3,X6	0.36
3	X9,X6	0.98
4	X1,X9	0.48
5	X1,X6	0.43

Dois possiveis modelos para avaliar qual prediz melhor:

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

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0433	0.0766	-0.57	0.5739
X3adj	0.4512	0.0832	5.42	0.0000
X9adj	0.2268	0.0880	2.58	0.0127

Tabela 13 – mod 2.1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0459	0.0781	-0.59	0.5588
X3adj	0.4707	0.0839	5.61	0.0000
X6adj	0.1825	0.0876	2.08	0.0420

Tabela 14 – mod 2.4

Teste linear geral do modelos 2.1 e 2.4 NA

Metricas aplicadas nos bancos de validação com os modelos 2.1 e 2.4

	df	AIC	RSS	RMSE	R2	Normal Test
Modelo 2.1	4.00	104.23	18.05	33.68	0.47	0.91
Modelo 2.4	4.00	106.44	18.77	9.99	0.26	0.91

Tabela 15 – Metricas avaliadas nos dados de validação sob os modelos 2.1 e 2.4.

Escolhendo novo modelo e avaliando regressao

	VIF
X3adj	1.17
X9adj	1.17

Tabela 16 – Novo Modelo utilizado.

3.3.2 Modelo por etapas

Todos os algorítmos de, “Stepwise”, “Forward” e “Backward” obtiveram do mesmo modelo

$$X1 = \beta_0 + \beta_{X3}X3 + \beta_{X8}X8 + \beta_{X9}X9 + \beta_{X11}X11 \qquad \text{(Modelo 2.2)}$$

Resultado do modelo selecionado pelo stepwise:

	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 – mod 2.4

Resultado do teste shapiro wilk (normalidade): (p-value: 0.703443)

vif do modelo step

	VIF
X3adj	1.15
X8	1.03
X9adj	1.60
X11adj	1.61

Tabela 18 – Modelo modstep

3.3.3 Comparação entre modelos

Pelo teste linear geral, temos o valor (p-value: 2.6360176×10^{-4})

Logo comparando as metricas no banco de validação temos que:

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

modelo verificação

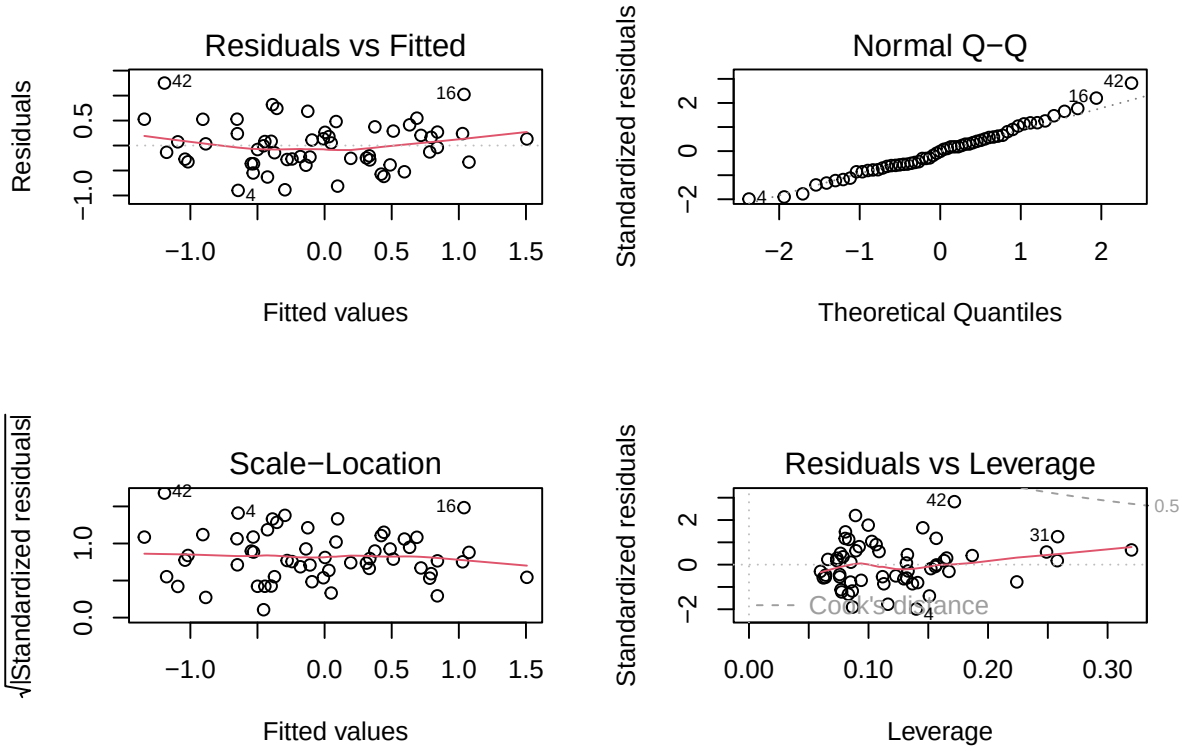

Figura 7 – Avaliação do resuduo do modelo step

4 Conclusão