TRABALHO FINAL DE REGRESSÃO LINEAR

Maria Luiza B. Quirino (190113456), Poliana Matos (190115670) e Rafael de Acypreste (200060023)

Professora Maria Theresa

Table of contents

Introdução

O Estudo sobre a Eficácia do Controle de Infecções Hospitalares (SENIC, Study on the Efficacy of Nosocomial Infection Control, em inglês) buscou avaliar se programas de controle e vigilância contra infecções reduziram as taxas de infecção hospitalar nos Estados Unidos. Também se desejou avaliar a relação entre algumas características dos hispitais e pacientes nas mudanças de taxa de infecção.

O estudo foi realizado entre 1975-76. Para e ste trabalho, será utilizada uma amostra aleatória de 113 hospitais, dos 338 hospitais que participaram do estudo.

Os dados coletados ajudarão a responder as seguintes perguntas:

1. O número de enfermeiros está relacionado às instalações e serviços do hospital e com a região? Em caso afirmativo, como?
2. A duração da internação está associada a quais fatores? Características do paciente, seu tratamento e hospital têm qual implicação?

Para responder a essas perguntas, será utilizado o arcabouço estatístico de regressões lineares, explicado na Seção 2.

1Objetivos

O objetivo geral do trabalho é avaliar como questões de estrutura dos hospitais se relacionaram com as infeções hospitalares em hospitais dos Estados Unidos no período de 1975-1976.

Os objetivos específicos são:

- Avaliar a relação entre o número de enfermeiros com respeito às instalações e região do hospital;
- Estudar se a duração da internação está associada a características do paciente, seu tratamento e as características do próprio hospital;
- Descrever o uso de modelos de regressão linear para a análise dos dados coletados na pesquisa.

2Metodologia

As principais fórmulas adotadas têm sua fundamentação especialmente determinada em Kutner et al. (2004).

Para o cumprimento dos objetivos de pesquisa, será usado o arcabouço teórico estatístico relacionado aos modelos de regressão linear. Em síntese, os modelos de regressão linear são modelos que buscam quantificar e qualificar as relações entre uma variável depentente — a ser explicada — e uma ou mais variáveis independentes, que auxiliam na explicação da variável dependente.

Como se trata de uma relação de dependência no sentido estatístico, não há necessariamente uma relação de causalidade entre as variáveis. Ainda assim, a relação de dependência pode ser usada para a previsão de valores da variável dependente, a partir de valores conhecidos das variáveis independentes.

A estrutura geral de um modelo de regressão linear é dada pela equação:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_n X_{ni} + \varepsilon_i \tag{2.1}$$

em que y é a variável dependente, $x_1, x_2, \dots, x_n$ são as variáveis independentes, $\beta_0, \beta_1, \dots, \beta_n$ são os parâmetros do modelo e ε_i é o erro aleatório.

2.1 Seleção de variáveis

O processo de seleção de variáveis envolve processos que ajudam a identificar as variáveis relevantes para o modelo. Antes, é preciso conhecer os tipos de variáveis que podem estar presentes no modelo para além dos formatos tradicional das variáveis como são coletadas.

Alguns critérios auxiliam na seleção das variáveis do model de regressão linear a ser utilizado, como a análise do R^2, R^2 ajustado, Critério de Pressão de Mallows (Cp) e Critério de Informação Bayesiano (BIC). Nessa seleção, busca-se uma boa relaçao entre capacidade explicativa/preditiva e parcimoniosidade do modelo.

O coeficiente de determinação (R^2) é uma medida de ajuste do modelo, que indica a proporção da variância da variável dependente que é explicada pelas variáveis independentes. Ele é calculado por:

$$R^2 = \frac{SQ_{reg}}{SQ_{tot}}$$

em que SQ_{reg} é a soma dos quadrados da regressão e SQ_{tot} é a soma dos quadrados totais.

O R^2 ajustado é uma medida de ajuste do modelo que parte do coeficiente de determinação, mas penaliza a inclusão de variáveis que não contribuem para a explicação da variável dependente. Sua fórmula é dada por

$$R^2_{ajustado} = 1 - \frac{(n-1)}{n-p}\frac{SQ_{erros}}{SQ_{tot}}$$

em que SQ_{erros} é a soma dos quadrados dos erros, n é o número de observações e p é o número de variáveis independentes.

O Critério de Pressão de Mallows (Cp) é uma medida de ajuste do modelo que penaliza a inclusão de variáveis que não contribuem para a explicação da variável dependente. É calculado por

$$Cp = \frac{SQ_{erros}}{MSE(X_1, \dots, X_{p-1})} - (n - 2p)$$

em que SQ_{erros} é a soma dos quadrados dos erros, MSE é o erro médio quadrático, n é o número de observações e p é o número de parâmetros.

Nesse caso, quando não há viés na regressão do modelo de base para comparação, o valor esperado de C_p é aproximadamente p (Kutner et al. 2004, 358).

O Critério de Informação Bayesiano (BIC) é uma medida de ajuste do modelo que penaliza a inclusão de variáveis que não contribuem para a explicação da variável dependente. É calculado por

$$BIC = n \ln(SQ_{erros,p}) - n \ln(n) + p \ln(n)$$

2.1 Modelos de mais de uma ordem

Os modelos de mais de uma ordem são aqueles em que a variável dependente é explicada por uma ou mais variáveis independentes que podem estar em forma de alguma potência inteira maior do que 1. São os chamados "modelos polinomiais" (Kutner et al. 2004, 294). Há duas razões principais para isso:

1. A relação entre a variável explicada e as variáveis explicativas é curvilínea; ou
2. Quando a relação entre as variáveis não é curvilínea, mas pode ser aproximada por uma curva.

Esta última razão tem aplicabilidade comum, e faz parte das hipóteses do presente estudo.

Um exemplo de modelo de mais de uma ordem é o modelo quadrático, dado pela equação:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_{1,1} X_{1i}^2 + \varepsilon_i \tag{2.2}$$

em que Y_i é a variável dependente, X_{1i} é a variável independente, β_0 é o intercepto, β_1 é o coeficiente da variável independente e $\beta_{1,1}$ é o coeficiente da variável independente elevada ao quadrado.

Entretanto, é preciso estar atento às complicações que fórmulas quadráticas ou superiores podem acrescer à interpretação dos resultados. A depender do sinal do coeficiente da variável independente elevada ao quadrado, a curva pode ter concavidade para cima ou para baixo. Em geral, a interpretação mais relevante está em torno de eventual ponto de inflexão (mínimo ou máximo), se este fizer parte do intervalo de observação da variável independente.

2.1 Variáveis categóricas

Variáveis categóricas também podem ser usadas em modelos de regressão linear, desde que sejam transformadas em variáveis binárias. A transformação é feita por meio da criação de novas colunas, que assumem o valor 1 quando a categoria está presente e 0 quando a categoria está ausente.

Para n categorias distintas, são necessárias $n-1$ colunas, pois a última categoria é a referência para as demais e estará representada pelo valor do intercepto do modelo quando as demais categorias assumirem valor 0. Nesse caso, há uma varição da reta de regressão para cada categoria, indicando uma alteração homogênea sobre o nível da variável resposta sob efeito de todas as demais variáveis.

Um exemplo de variável categórica é a filiação ou não a uma escola de medicina. Considerando X_1" como a variável categórica, X_2 outra variável quantitativa do modelo, a interpretação do modelo se dá da seguinte forma:

$$\begin{aligned} E[Y] &= \beta_0 + \beta_1(1) + \beta_2 X_2 = (\beta_0 + \beta_1) + \beta_2 X_2 \quad \text{, se } X_1 = 1 \\ E[Y] &= \beta_0 + \beta_1(0) + \beta_2 X_2 = \beta_0 + \beta_2 X_2 \quad \text{, se } X_1 = 0 \end{aligned} \tag{2.3}$$

Com essa construção, a interpretação do modelo se dá diretamente avaliando a presentaça ou não da variável categórica de interesse, mantendo as demais variáveis constantes.

2.1 Variáveis com interação

Quando um modelo de regressão linear possui variáveis sem interação entre elas, diz-se tratar de um "modelo aditivo" (Kutner et al. 2004). Entretanto, quando isso ocorre, as variáveis devem aparecer sob a forma de produto no modelo, como no exemplo a seguir:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_{1,2} X_{1i} X_{2i} + \varepsilon_i \tag{2.4}$$

Nesse caso, o efeito de X_1 sobre Y depende do valor de X_2, e vice-versa. A interpretação do modelo envolve fazer a análise de efeito de cada variável não aditiva a partir de um dado nível da outra variável com que ela se relaciona. Nesse caso, o efeito da variável X_1 sobre Y dado X_2 constante é dada por:

$$\beta_1 + \beta_{1,2} X_{2i}$$

.

Esse prodecidmento deve ser realizado para todas as formas de interação.

2.1 Procedimentos de seleção de variáveis (forward, backward e stepwise)

Há também procedimentos de seleção de variáveis que podem ser usados para a seleção de variáveis. São eles:

1. *Forward*: o procedimento parte de um modelo com apenas o intercepto e vai adicionando variáveis, uma a uma, até que não seja possível adicionar mais nenhuma variável com significância estatística. A adição de variáveis é feita com base no no menor p-valor.

2. *Backward*: fazendo o processo inverso do anterior, o procedimento parte de um modelo com todas as variáveis e vai retirando variáveis do modelo, uma a uma de acordo com seu p-valor, até que todas as variáveis sejam significativas do ponto de vista estatístico.

3. *Stepwise*: o procedimento parte de um modelo com apenas o intercepto e vai adicionando ou retirando variáveis, uma a uma, até que não seja possível adicionar ou seja necessário retirar alguma variável com significância estatística. O procedimento termina quando se encontra o "melhor" modelo (Kutner et al. 2004, 364–66).

2.2 Pressupostos de um modelo linear

Um modelo de regressão linear apresenta alguns pressupostos, que devem ser verificados para que o modelo seja considerado adequado. São eles:

1. Linearidade: a relação entre as variáveis deve ser linear. Caso contrário, é necessário transformar as variáveis para que a relação se torne linear. Este pressuposto pode ser verificado por meio de gráficos de dispersão entre as variáveis e os resíduos;
2. Normalidade: os erros devem ser normalmente distribuídos, o que se pode verificar ao analisar os resíduos do modelo. Pode ser verificado por meio do teste de Shapiro-Wilk e pela visualização do gráfico de distribuição normal dos resíduos;
3. Homocedasticidade: os erros devem ter variância constante, o que se pode verificar ao analisar os resíduos do modelo em relação às variáveis independentes. Costuma ser verificado por meio do teste de Breusch-Pagan; e
4. Independência: os erros devem ser independentes, o que também se pode verificar ao analisar os resíduos do modelo em relação às variáveis independentes. Pode ser verificado por meio do teste de Durbin-Watson;
5. Ausência de multicolinearidade entre as variáveis: as variáveis independentes não devem ser correlacionadas entre si, o que se pode verificar ao analisar a matriz de correlação entre as variáveis independentes. O caso da multicolinearidade perfeita pode fazer com que o modelo tenha múltiplas soluções, o que torna a estimação sem validade. O caso da multicolinearidade imperfeita pode fazer com que o modelo tenha solução, mas com variâncias muito grandes — com elevada chance de não rejeição da hipótese nula de parâmetro zero, estimativas com sinais em desacordo com toda a literatura existente, o que torna a estimação também problemática.

2.3 Estimação dos parâmetros

Os parâmetros do modelo são estimados por meio do método dos mínimos quadrados ordinários (MQO). O método consiste em minimizar a soma dos quadrados dos resíduos, ou seja, a soma dos quadrados das diferenças entre os valores observados e os valores estimados pelo modelo.

A estrutura geral do modelo de regressão linear em forma matricial é dada pela equação:

$$\mathbf{Y}_{[n\times 1]} = \mathbf{X'}_{[n\times p]}\beta_{[p\times 1]} + \varepsilon_{[n\times 1]} \tag{2.5}$$

em que **Y** é o vetor de variáveis dependentes, **X** é a matriz de variáveis independentes, β é o vetor de parâmetros e ε é o vetor de erros aleatórios.

Os parâmetros são estimados por meio da equação:

$$b = (\mathbf{X'X})^{-1}\mathbf{X'Y} \tag{2.6}$$

em que b é o vetor de parâmetros estimados.

Para fazer inferências e os testes de hipóteses, é necessário estimar a matriz de variância e covariância dos parâmetros. Isso pode ser feito a partir da estimação dos parâmetros e por meio da equação:

$$\mathbf{Var}(b) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{2.7}$$

em que $\mathbf{V}(b)$ é a matriz de variância e covariância dos parâmetros e σ^2 é a variância dos erros aleatórios.

Por fim, como os erros aleatórios não são observados, é preciso estimar a variância dos erros aleatórios. Isso pode ser feito por meio da equação:

$$\mathbf{Var}(b) = \frac{SQ_{erros}}{n-p}\sigma^2(\mathbf{X}'\mathbf{X})^{-1} = MSE(\mathbf{X}'\mathbf{X})^{-1} \tag{2.8}$$

em que $\hat{\sigma}^2$ é a estimativa da variância dos erros aleatórios, SQ_{erros} é a soma dos quadrados dos erros, $n-p$ é o número de graus de liberdade do modelo, e MSE é o erro médio quadrático.

2.3 Testes de ausência de regressão e de significância dos parâmetros

A primeira análise de um modelo consiste em testar a hipótese nula de ausência de regressão. Isso é feito por meio de um teste F, cuja estatística é dada por:

$$F = \frac{SQ_{reg}}{p-1} \div \frac{SQ_{erros}}{n-p} \tag{2.9}$$

em que SQ_{reg} é a soma dos quadrados da regressão, $p-1$ é o número de graus de liberdade da regressão, SQ_{erros} é a soma dos quadrados dos erros e $n-p$ é o número de graus de liberdade dos erros.

Se os erros tiverem distribuição normal, a estatística F segue uma distribuição F com $p-1$ e $n-p$ graus de liberdade.

Já os a validade estatística dos parâmetros pode ser testada por meio de um teste $t-student$, cuja estatística é dada por:

$$t^* = \frac{b_j - \beta_j}{\sqrt{\mathbf{s^2}(b_j)}} \tag{2.10}$$

em que b_j é o parâmetro estimado, β_j é o parâmetro teórico, $\mathbf{s^2}(b_j)$ é a variância do parâmetro estimado e $t*$ é a estatística do teste que tem distribuição $t-student_{(n-p)}$. Normalmente, testa-se a hipótese nula de que o parâmetro é igual a zero, ou seja, $H_0 : \beta_j = 0$.

2.4 Validação do modelo

Por fim, após todos os procedimentos acima indicados, deve-se testar a validade do modelo e sua capacidade de generalização. Para isso, é preciso testar o modelo em uma amostra diferente daquela usada para a estimação dos parâmetros.

No caso em análise, será calculado um modelo com uma amostra de tamanho 60.

Em primeiro lugar, será calculado um novo modelo com os demais dados do problema, que fazem parte do conjunto de validação. Os parâmetros deste modelo são comparados aos parâmetros do modelo do conjunto de treinamento. Caso haja estabilidade dos parâmetros, pode-se dizer que o modelo é consistente com toda a população.

Em seguida, será calculado o erro médio quadrático (MSE) do modelo originalmente treinado no conjunto de validação. O MSE é calculado por meio da equação:

$$MSE = \frac{SQ_{erros}}{n-p} \tag{2.11}$$

em que SQ_{erros} é a soma dos quadrados dos erros, $n-p$ é o número de graus de liberdade do modelo.

Espera-se, com isso, que o modelo tenha um MSE próximo ao MSE do modelo de treinamento. Teste semelhante pode ser feito com o coeficiente de determinação (R^2) e o R^2 ajustado.

Por fim, caso o modelo escolhido se comporte bem nas duas análises indicadas acima, estima-se o mesmo modelo, desta vez utilizando o conjunto de dados completo. Por se tratar de um tamanho de amostra maior, espera-se que a precisão do modelo seja mais elevada. Com isso, tem-se um modelo final, que pode ser usado para a explicar a relação entre as variáveis e para a previsão de valores da variável dependente a partir de valores conhecidos das variáveis independentes.

3Resultados

3.1 Análise Exploratória

Nesta seção, foi feita uma análise exploratória dos dados, é uma etapa essencial para compreender e interpretar. Foram feitas análises gráficas e de medidas resumo das 11 variáveis do banco de dados, com o objetivo de visualizar individualmente, compreender melhor os dados, verificar as possíveis hipóteses para o problema e verificar relação entre elas.

Variável	Descrição
Duração da Internação	Duração média da internação dos pacientes no hospital em dias
Idade	Idade média dos pacientes
Risco de Infecção	Probabilidade média estimada de adquirir infecção no hospital (em %)
Proporção de Culturas de Rotina	Razão do número de culturas realizadas com relação ao número de pacientes sem sinais ou sintomas de infecção adquirida no hospital, vezes 100
Proporção de Raio-X de Tórax de Rotina	Razão do número de Raio-X de tórax realizados com relação ao número de pacientes sem sinais ou sintomas de pneumonia, vezes 100
Número de Leitos	Número médio de leitos no hospital durante o período de estudo
Filiação a Escola de Medicina	1-sim 2-não
Região	Região Geográfica, onde 1- NE 2- NC 3-S E 4- W
Média diária de Pacientes	Número médio de pacientes no hospital por dia durante o período do estudo
Número de enfermeiro(s)	Número de enfermeiros de tempo-integral ou equivalente registrados e licenciados durante o período de estudo
Facilidades e Serviços Disponíveis	% de 35 potenciais facilidades e serviços que são fornecidos pelo hospital

Variáveis quantitativas	Mínimo	1° quartil	Medina	3° quartil	Máximo	Média	Desvio Padrão
Duração da internação	7.08	8.33	9.43	10.51	13.95	9.6	1.59
Idade	42.00	50.90	52.65	54.70	65.90	52.88	4.42
Risco de Infecção	1.40	4.08	4.50	5.23	7.70	4.41	1.25
Prop. de Cultura de Rotina	2.60	9.40	14.80	19.25	60.50	16.38	10.45
Prop. de Raio-X de Tórax	42.60	63.40	82.15	92.70	133.50	81.01	21.05
Leitos	60	127.2	195.5	353.8	833	283.4	211.13
Média diária de Pacientes	38	88	159.5	275.8	595	212.8	161.02
Enfermeiros	19	66.25	161.50	254	656	200.37	159.65
Facilidades e Serviços disponíveis	17.10	34.30	45.70	57.10	80	45.57	15.84

No banco de dados, existem 11 variáveis, das quais 9 são quantitativas, dentre elas, duração da internação, idade, risco de infecção, proporção de culturas de rotina, proporção de raio-x de tórax de rotina, facilidades e serviços disponíveis são contínuas e leitos, média de pacientes e enfermeiros são discretas. Completando a quantidade de variáveis, filiação a escola de medicina e região, que são qualitativas nominais.

Ao analisar o nosso gráfico de boxplot (Figura 3.1), percebe-se a existência de valores que se diferenciam drasticamente da normalidade (*outliers*), mas eles não são semelhantes para todas as variáveis. Em alguns casos, como idade de pacientes e risco de infecção, eles aparecem tanto abaixo como acima do esperado. Em outros casos, como proporção de

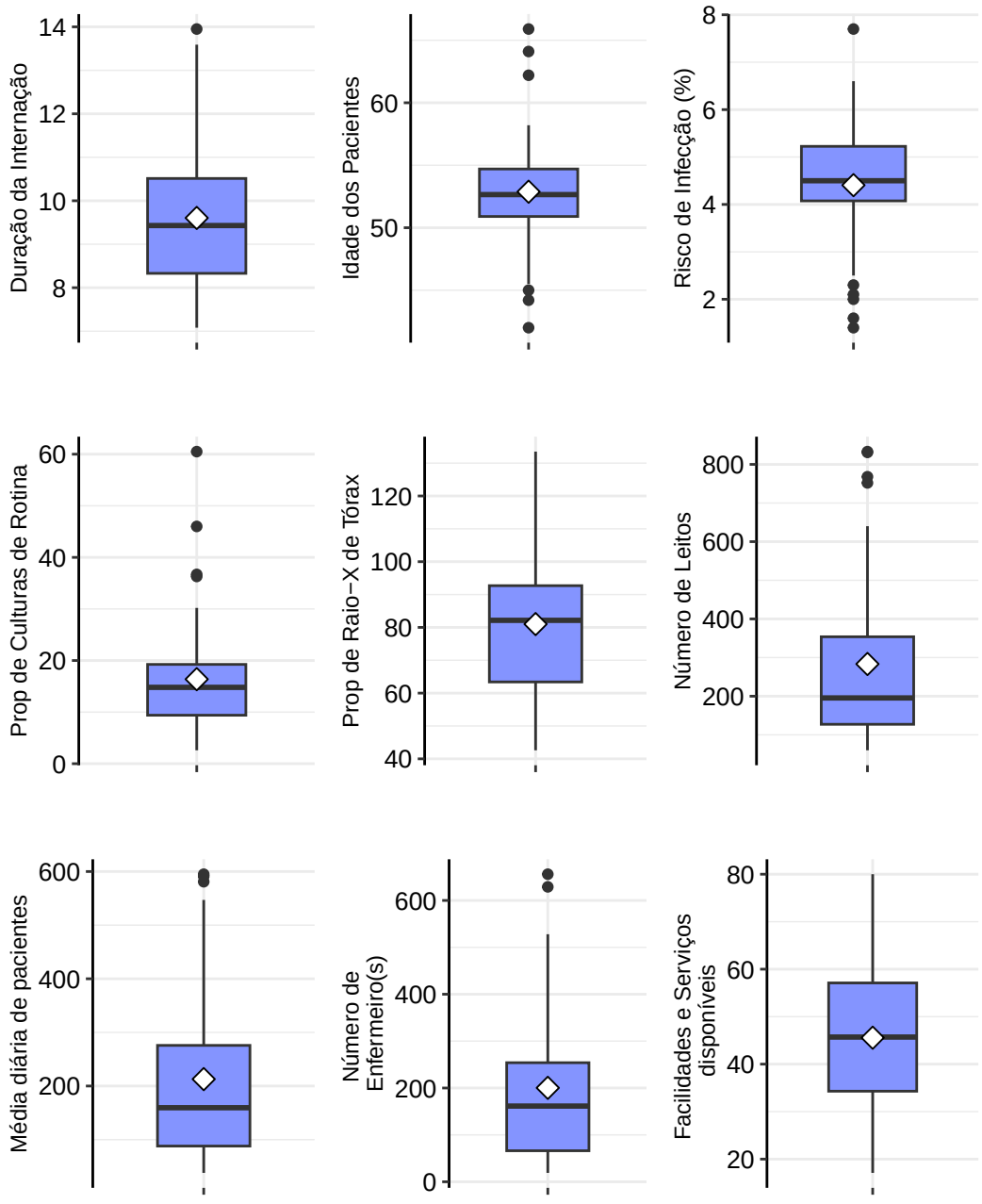

Figure 3.1: Boxplots e das variáveis quantitativas.

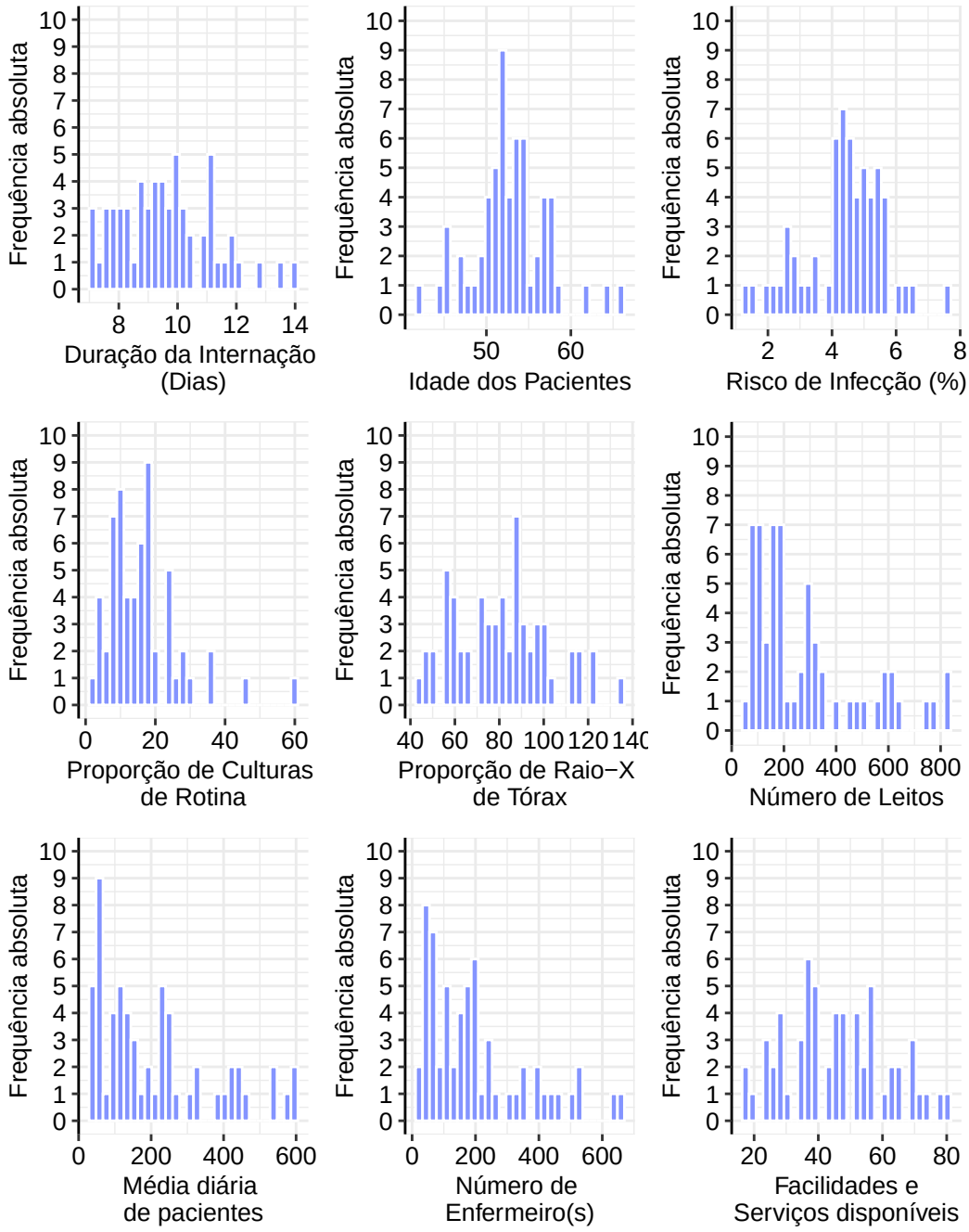

Figure 3.2: Histogramas das variáveis quantitativas.

raio-x de tórax e facilidades e serviços disponíveis, não há valores discrepantes. E em sua maioria existem *outliers* que estão acima do limite superior.

Observando então a Figura 3.2, percebe-se que a influência dos *outliers* é significativa, dado que, nas variáveis que possuem apenas valores acima do limite superior, é visível uma assimetria positiva a direita. Já as que possuem acima do limite superior e abaixo do inferior, e as que não se observa *outliers*, são um pouco mais simétricas. Essa condição que afeta diretamente também nos valores de média e desvio padrão.

Observando as variáveis qualitativas, temos como característica, que todas são nominais. No histograma (Figura 3.3), temos que a maioria não possui filiação com escola de medicina, observação que é confirmada através da Tabela 3, já que, 81,67% diz não ter.

Variáveis qualitativas		Frequência	Porcentagem
Filiação a Escola de Medicina	Sim	11	18.33%
	Não	49	81,67%
Região	NE	17	28,33%
	NC	17	28,33%
	S	18	30%
	W	8	13,34%

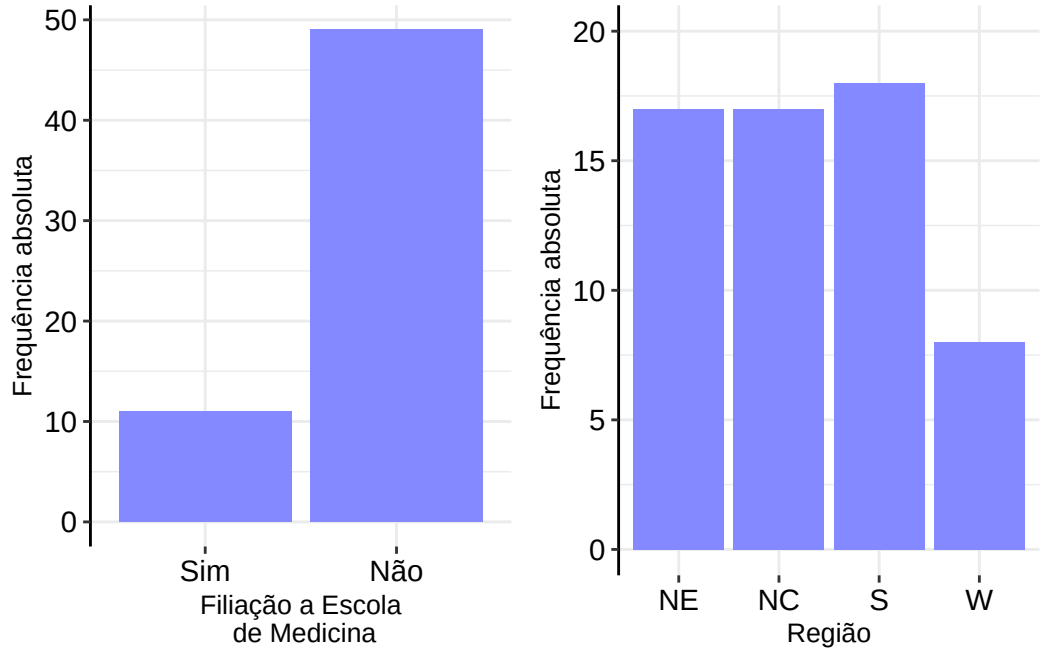

Figure 3.3: Histogramas das variáveis discretas.

E, para a variável região, existe uma maior frequência nas regiões NE, NC e S do que na região W. Observação que é ressaltada na Tabela 3, já que somente 13,34% se encontra na região W, dividindo então, com valores bem parecidos para o restante das regiões.

3.2 Modelo Completo e Seleção de Variáveis

Nesta etapa da pesquisa, realizou-se a análise preliminar das variáveis em estudo e a construção do modelo de regressão linear múltipla para investigar os fatores associados à duração da internação hospitalar. A abordagem foi enriquecida pela exploração de relações

de segunda ordem, com foco específico na influência do número de enfermeiros(as) nas instalações e serviços disponíveis. Além disso, foram incorporadas variáveis regionais para examinar possíveis variações geográficas na duração da internação. Essa abordagem permite uma compreensão mais abrangente dos fatores que contribuem para a complexidade do tempo de internação hospitalar, considerando não apenas características individuais do paciente, mas também aspectos relacionados ao tratamento e ao contexto hospitalar.

Na avaliação das correlações entre variáveis quantitativas, foram observadas associações positivas e significativas entre a duração da internação e variáveis como risco de infecção, número de leitos, média diária de pacientes, quantidade de enfermeiros(as) e a disponibilidade de facilidades e serviços hospitalares. Essas correlações sugerem a possível influência dessas variáveis na variabilidade da duração da internação, como se pode ver na Figura Figura 3.4.

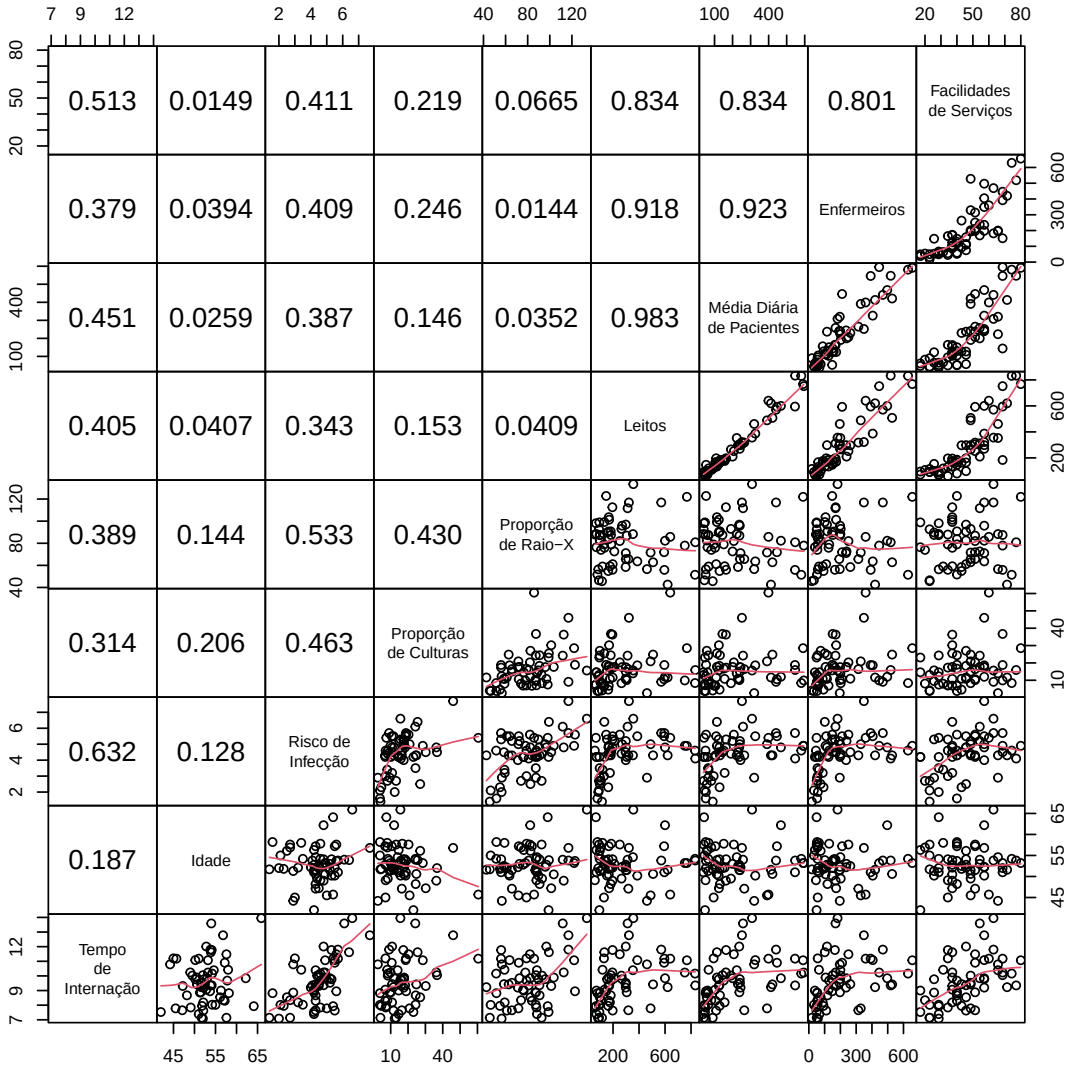

Figure 3.4: Correlações entre as variáveis quantitativas.

Ao explorar a relação entre a idade média dos pacientes e o tempo internação, foi observada uma associação modesta de 0,187, o que sugere, em geral, que pacientes mais idosos podem demandar internações mais prolongadas. O risco de infecção mostrou uma correlação substancial (0,632), o que indica que hospitais com maiores índices de risco infeccioso podem enfrentar internações mais extensas. A análise das proporções de culturas de rotina (0,314) e Raio-X de Tórax de rotina (0,389) revelou associações interessantes. Hospitais que realizam mais culturas de rotina e exames de Raio-X de Tórax de rotina parecem enfrentar internações mais longas, o que sugere uma possível relação entre a extensão das investi-

gações diagnósticas e a duração do tratamento.

Destaca-se que a disponibilidade de leitos, o número de enfermeiros(as) e o percentual de facilidades e serviços apresentaram correlações positivas e consistentes (de 0,405, 0,379 e 0,513 respectivamente) com a duração da internação. Esses resultados ressaltam a importância crítica desses fatores na gestão eficaz do tempo de internação, evidenciando a necessidade de estruturas hospitalares adequadas e recursos humanos suficientes. Além disso, a média diária de pacientes apresentou uma correlação positiva de 0,451 com o tempo de internação, o que indica um aumento no tempo das internações em hospitais que apresentam maior demanda diária.

Ao considerar a influência de variáveis relacionadas ao paciente, tratamento e hospital na duração da internação, o modelo de regressão linear múltipla fornece insights valiosos, que podem ser vistos na Tabela 3.1. Esse modelo, embora seja mais simples por não incorporar interações ou termos de segunda ordem, é notável em sua capacidade de explicar 67,55% da variação na duração da internação. Isso sugere que, mesmo sem levar em conta complexidades adicionais nas relações entre as variáveis, as características básicas do paciente, seu tratamento e o ambiente hospitalar desempenham um papel crucial na determinação da duração da internação.

$$\begin{aligned} Y_i =& \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \beta_7 X_{7i} + \beta_8 X_{8i} \\ &+ \beta_9 X_{9i} + \beta_{10} X_{10i} + \beta_{11} X_{11i} + \beta_{11,11} X_{11i}^2 + \beta_{12} X_{12i} + \beta_{12,12} X_{12i}^2 \\ &+ \beta_{7,12} X_{7i} X_{12i} + \beta_{8,12} X_{8i} X_{12i} + \beta_{9,12} X_{9i} X_{12i} + \varepsilon_i \end{aligned} \tag{3.1}$$

O modelo inicial ressalta a relevância de certas variáveis na explicação da duração da internação. Por exemplo, o coeficiente associado à idade não é estatisticamente significativo (p-valor: 0,5294), sugerindo que a idade média dos pacientes não está fortemente associada à duração da internação.

Em contrapartida, o coeficiente para o Risco de Infecção é estatisticamente significativo (p-valor: 0,001043), indicando que um aumento no risco de infecção está associado a um aumento na duração da internação. Além disso, os coeficientes negativos para as Regiões (regiaoNC, regiaoS, regiaoW) indicam que essas regiões têm internações mais curtas em comparação com a região NE. As demais variáveis do modelo não apresentam significância estatística uniforme, ressaltando a importância de variáveis específicas, como o Risco de Infecção, na explicação da variação na duração da internação.

Considerando a complexidade das relações entre variáveis, foi incorporarado ao modelo interacões e termos de segunda ordem. Essa abordagem visa explorar nuances que podem ser negligenciadas em um modelo linear.

A consideração específica da interação entre enfermeiros e a região W, por exemplo, apresenta um coeficiente negativo significativo (-2.651), o que indica que a região W demonstra uma associação substancial entre o aumento do número de enfermeiros e a redução mais acentuada na duração da internação em comparação com outras regiões geográficas.

O termo quadrático para o número de enfermeiros (X_{11i}^2) apresenta um coeficiente negativo, sugerindo a possibilidade de saturação no impacto positivo do aumento no número de enfermeiros, indicando que, após um certo ponto, um aumento adicional pode não ter o mesmo efeito positivo na redução da duração da internação. Por outro lado, o termo quadrático para facilidades e servicos disponíveis (X_{12i}^2) não demonstra significância estatística (p-valor: 0,6889), sugerindo que a relação quadrática entre essa variável e a duração da internação não é estatisticamente robusta.

Table 3.1: Modelo de regressão linear múltipla com interação e termos de segunda ordem.

	Dependent variable:	
	t_internacao	
	Modelo simples	Modelo segunda ordem e interações
idade	0.022	0.033
	p = 0.530	p = 0.366
r_infeccao	0.579***	0.501**
	p = 0.002	p = 0.013
prop_culturas	−0.014	−0.015
	p = 0.447	p = 0.450
prop_raiox	0.008	0.013
	p = 0.329	p = 0.166
leitos	0.001	0.002
	p = 0.782	p = 0.637
escola_medicinaNão	−0.441	−0.358
	p = 0.362	p = 0.529
regiaoNC	−0.615	−0.449
	p = 0.141	p = 0.522
regiaoS	−0.991**	−0.742
	p = 0.021	p = 0.250
regiaoW	−2.064***	−2.651***
	p = 0.0003	p = 0.002
m_dia_pacientes	0.002	0.001
	p = 0.715	p = 0.830
enfermeiros	−0.003	0.003
	p = 0.211	p = 0.651
facilidades_servicos	0.015	−0.021
	p = 0.393	p = 0.759
I(enfermeiros^2)		−0.00001
		p = 0.316
I(facilidades_servicos^2)		0.0003
		p = 0.689
regiaoNC:enfermeiros		−0.001
		p = 0.663
regiaoS:enfermeiros		−0.002
		p = 0.566
regiaoW:enfermeiros		0.003
		p = 0.290
Constant	5.764***	5.287**
	p = 0.008	p = 0.039
Observations	60	60
R^2	0.676	0.697
Adjusted R^2	0.593	0.575
Residual Std. Error	1.013 (df = 47)	1.036 (df = 42)
F Statistic	8.154*** (df = 12; 47)	5.687*** (df = 17; 42)
Note:		*p<0.1; **p<0.05; ***p<0.01

Na avaliação global do modelo, o coeficiente de determinação (R^2) de 0,6971 destaca sua eficácia ao explicar aproximadamente 69,71% da variabilidade na duração da internação, evidenciando sua robusta capacidade preditiva. No âmbito específico do modelo, destaca-se que o risco de infecção apresenta uma associação positiva e estatisticamente significativa (coeficiente: 0,501, p-valor: 0,0126) com a duração da internação. Essa constatação sugere que um aumento no risco de infecção está relacionado a um período mais longo de internação.

Em contraste, variáveis como idade e filiação a escola de medicina não apresentaram significância estatística, indicando que a idade média dos pacientes e a vinculação à escola de medicina não estão fortemente associadas à duração da internação. Além disso, outras variáveis, como proporção de culturas, leitos e facilidades e servicos diponíveis, também não apresentaram significância estatística uniforme, sugerindo que sua inclusão no modelo pode não contribuir significativamente para explicar a variabilidade observada na duração da internação.

Entretanto, surge a necessidade de uma análise mais profunda para assegurar a robustez do modelo. Questões sobre a presença de multicolinearidade entre as variáveis independentes são pertinentes, considerando o potencial impacto na precisão das estimativas. Como dito, a multicolinearidade, oriunda da alta correlação entre variáveis independentes, pode dificultar a identificação de seus efeitos individuais, comprometendo a interpretação dos resultados.

Dessa forma, a etapa seguinte compreenderá a condução de diagnósticos específicos para avaliar a multicolinearidade e, se necessário, efetuar ajustes no modelo. Paralelamente, serão realizadas análises dos pressupostos da regressão linear múltipla, tais como a normalidade dos resíduos, homocedasticidade e independência, visando garantir a validade das inferências.

Continuando a análise do modelo combinado, foram realizados testes importantes para verificar a adequação dos resíduos ao pressuposto da normalidade e homocedasticidade.

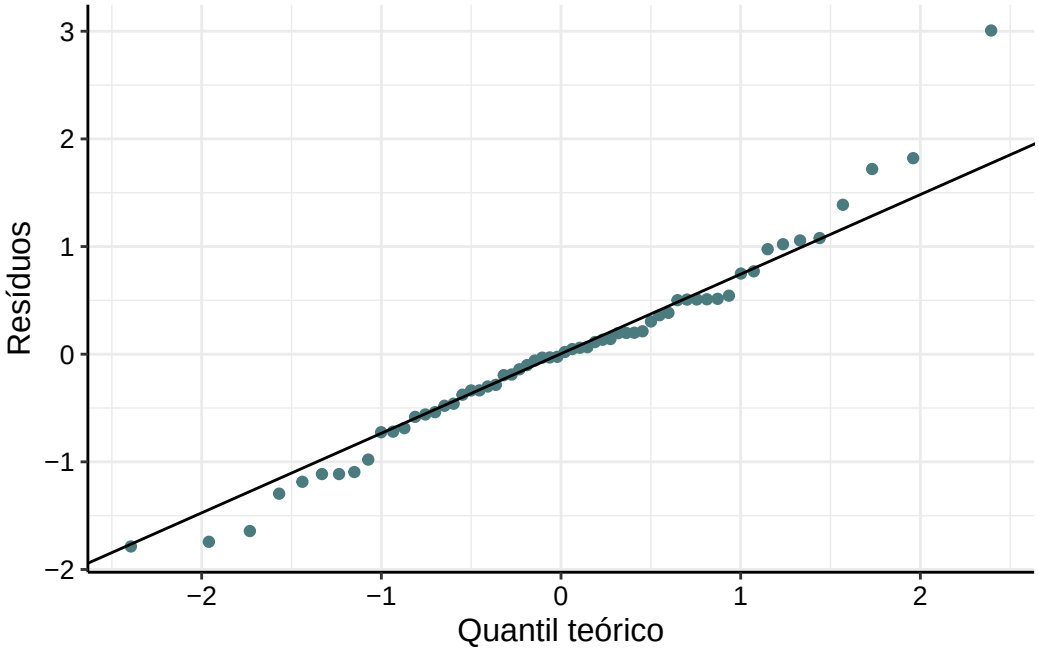

Figure 3.5: Gráfico de distribuição normal dos resíduos

O teste de normalidade de Shapiro-Wilk foi aplicado aos resíduos do modelo combinado, resultando em uma estatística W de 0,9675 e um p-valor de 0,1094. O valor de W próximo a 1 e o p-valor superior a 0.05 sugerem que não há evidências suficientes para rejeitar a

hipótese nula de normalidade dos resíduos. Conclusão semelhante pode ser identificada na Figura 3.5. Assim, os resíduos do modelo parecem seguir uma distribuição normal.

Além disso, foi realizado o teste de Breusch-Pagan para avaliar a homocedasticidade dos resíduos. O teste resultou em uma estatística BP de 17,1716, com 17 graus de liberdade e um p-valor de 0,4428. O p-valor maior que 0.05 indica que não há evidências significativas para rejeitar a hipótese nula de homocedasticidade. Portanto, os resíduos parecem exibir homocedasticidade, indicando que a variância dos erros é constante em diferentes níveis de preditores.

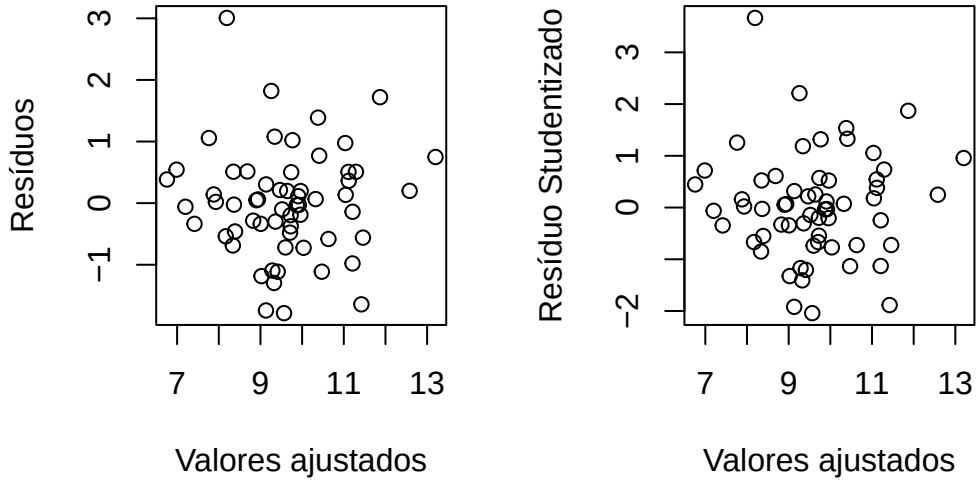

Figure 3.6: Gráfico de Resíduos e Resíduos Studentizados vs Valores Ajustados.

A análise da Figura 3.6 reforça que a independência dos resíduos foi atendida, consolidando a confiabilidade dos resultados obtidos no modelo. Além disso, o teste de Durbin-Watson para independência também não rejeitou a hipótese nula de independência, com p-valor de 0,324. A confirmação da normalidade, homocedasticidade e independência dos resíduos fortalece a robustez do modelo, oferecendo suporte à validade das inferências derivadas.

3.2 Seleção de Variáveis

Na presente seção, a análise das variáveis do modelo será aprofundada, visando à seleção criteriosa do modelo reduzido que melhor se adequa à explicação do tempo de internação. Métodos de seleção de variáveis serão utilizados com o intuito de assegurar uma abordagem mais precisa e refinada na identificação dos fatores mais relevantes para o tempo de internação.

3.2.1.1 tabela de variáveis?//

3.2.1.2 tabela dos modelos e estatísticas?//

Como se pode ver na Figura 3.7, ao analisar o crescimento nos níveis dos gráficos em relação ao R^2, R^2 ajustado, Critério de Pressão de Mallows (Cp) e Critério de Informação Bayesiano (BIC) até 6 parâmetros, observa-se um aumento em ambos os indicadores até atingir esse ponto. Contudo, ao aplicar métodos de seleção de variáveis, como o *Backward* e o *Stepwise*, é escolhido o modelo que mantém 7 parâmetros, enquanto no método *Forward*, nenhuma variável é removida. Diante dessa divergência, optou-se por trabalhar com dois modelos distintos, buscando determinar qual deles melhor atende aos objetivos específicos do trabalho.

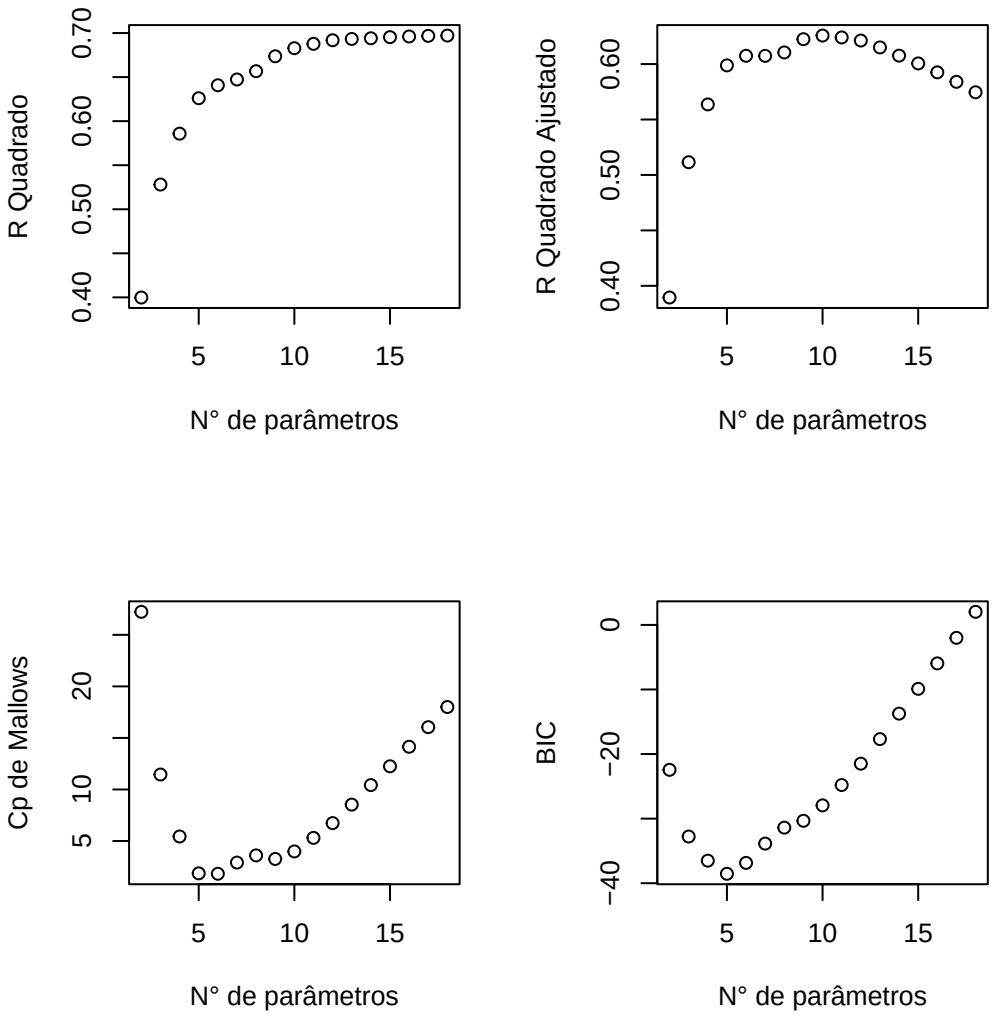

Figure 3.7: Gráficos de R^2, R^2 ajustado, Cp de Mallows e BIC.

Na análise dos modelos de regressão escolhidos, destaca-se o Modelo com 5 Variáveis, que incorpora as características de idade, risco de infecção, regiaoW, facilidades e servicos disponíveis e interaçao da regiaoS com o número de enfermeiros. Expresso pela equação:

$$Y_i = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \beta_3 X_{10} + \beta_4 X_{13} + \beta_5 X_9 X_{12} + \varepsilon_i$$

Além disso, o Modelo com 6 Variáveis expande a abordagem ao incorporar as variáveis de idade, risco de infecção, regiaoW, termo quadrático de facilidades e servicos disponíveis, interação entre regiaoNC e número de enfermeiros e interação entre a regiaoS e número de enfermeiros. A equação do modelo é expressa por:

$$Y_i = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \beta_3 X_{10} + \beta_4 X_{13}^2 + \beta_5 X_8 X_{12} + \beta_6 X_9 X_{12} + \varepsilon_i$$

Ambos os modelos estão representados na Tabela 3.2. O Modelo 1 apresenta um coeficiente de determinação de 0,6433, enquanto o Modelo 2 apresenta um coeficiente de determinação de 0,6535. Esses resultados indicam que o Modelo 1 explica 64,33% da variabilidade na duração da internação, enquanto o Modelo 2 explica 65,35% da variabilidade na duração da internação. Assim, o Modelo 1 parece ser mais parcimonioso, pois explica uma proporção maior da variabilidade na duração da internação com menos variáveis.

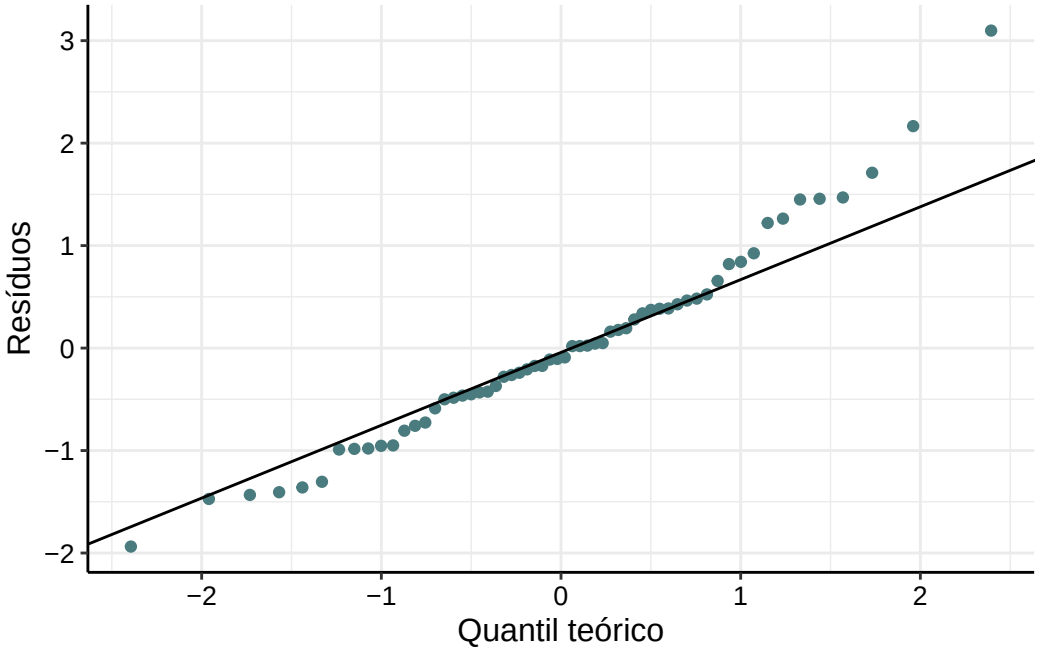

Figure 3.8: Gráfico de Distribuição de Probabilidade Normal do Modelo 1.

Os testes de normalidade também podem ser realizados. O teste de Shapiro-Wilk para o modelo 1 não rejeitou a hipótese nula de não normalidade, com p-valor de 0,1468. O teste de Shapiro-Wilk para o modelo 2 também não rejeitou a hipótese nula de não normalidade, com p-valor de 0,2411. Além disso, os gráficos de resíduos e resíduos studentizados para ambos os modelos não apresentam padrões claros, sugerindo que os resíduos podem ser considerados normalmente distribuídos, como se pode ver nas Figuras 3.8 e 3.9.

Com respeito à homogeneidade da variância, o teste de Breusch-Pagan para o modelo 1 não rejeitou a hipótese nula de homogeneidade, com p-valor de 0,405. O teste de Breusch-Pagan para o modelo 2 também não rejeitou a hipótese nula de homogeneidade, com p-valor de 0,395. Conclusões semelhantes podem ser visualizadas na Figura 3.10.

Table 3.2: Estatísticas dos Modelos 1 e 2.

	Dependent variable:	
	t_internacao	
	Modelo 1	Modelo 2
idade	0.042	0.032
	p = 0.178	p = 0.318
r_infeccao	0.634***	0.645***
	p = 0.00001	p = 0.00001
regiaoW	−1.725***	−1.973***
	p = 0.0001	p = 0.00005
facilidades_servicos	0.031**	
	p = 0.038	
I(facilidades_servicos^2)		0.0003**
		p = 0.035
regiaoNC		−0.318
		p = 0.611
regiaoS	−0.285	−0.513
	p = 0.550	p = 0.351
regiaoNC:enfermeiros		−0.0005
		p = 0.829
enfermeiros:regiaoS		−0.002
		p = 0.342
enfermeiros	−0.0003	−0.0005
	p = 0.838	p = 0.782
regiaoS:enfermeiros	−0.002	
	p = 0.287	
Constant	3.671**	5.006***
	p = 0.037	p = 0.008
Observations	60	60
R^2	0.643	0.654
Adjusted R^2	0.595	0.591
Residual Std. Error	1.010 (df = 52)	1.015 (df = 50)
F Statistic	13.398*** (df = 7; 52)	10.480*** (df = 9; 50)
Note:		*p<0.1; **p<0.05; ***p<0.01

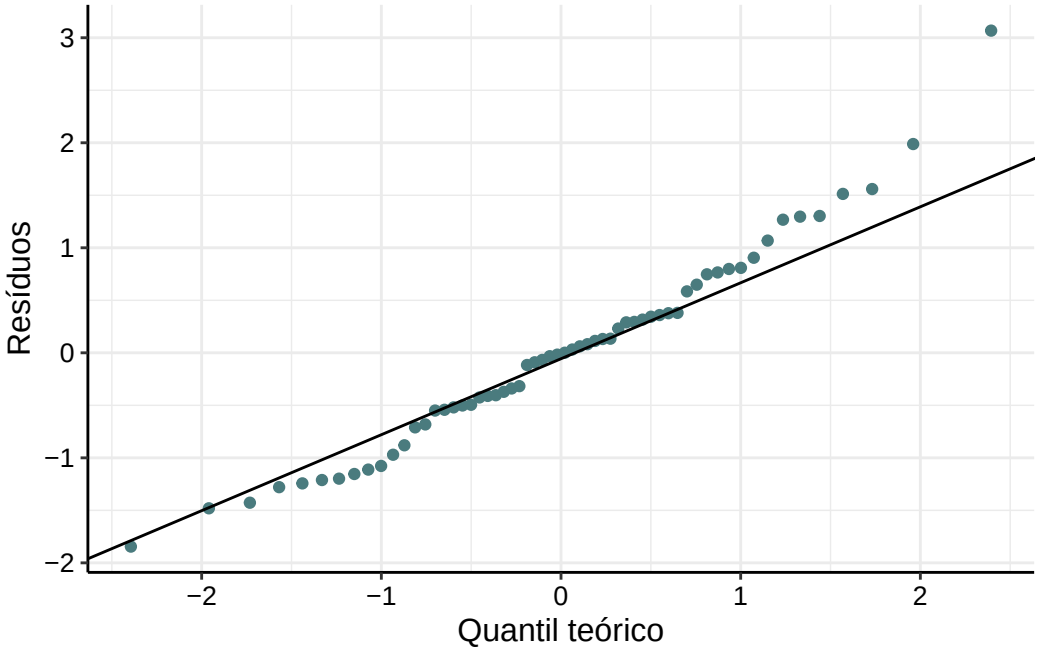

Figure 3.9: Gráfico de Distribuição de Probabilidade Normal do Modelo 2.

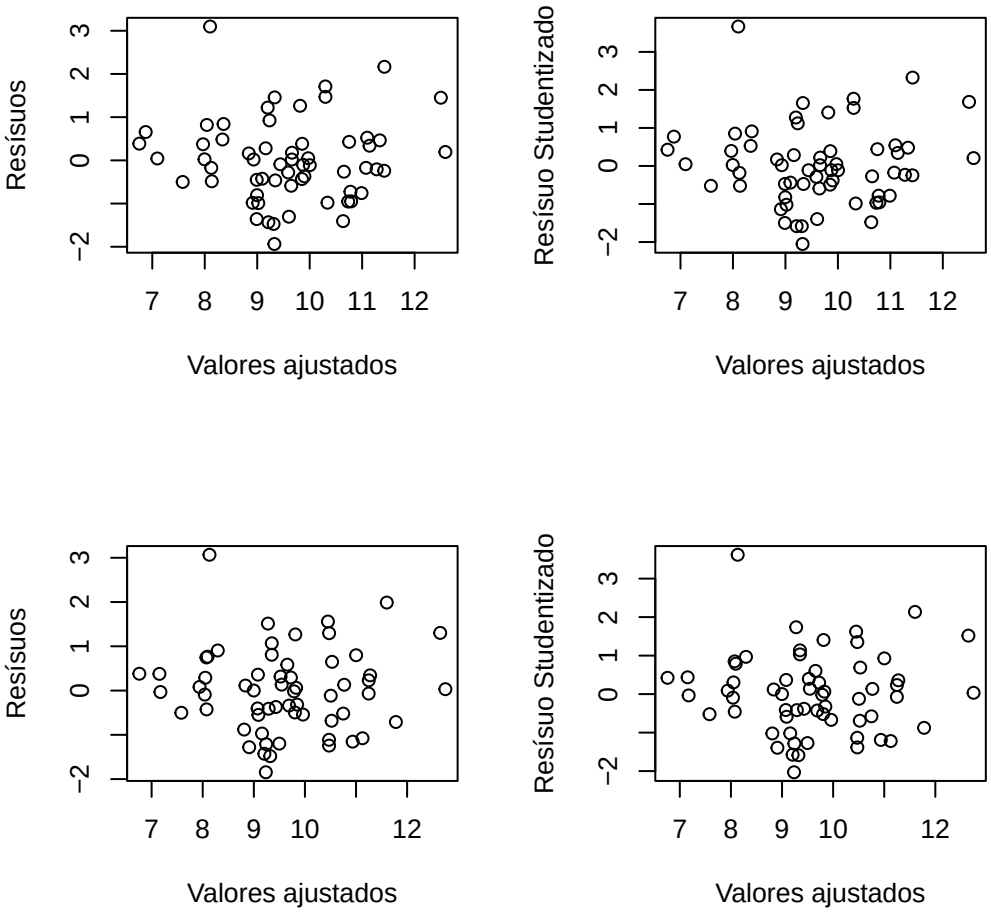

Figure 3.10: Gráficos de Resíduos e Resíduos Studentizados vs Valores Ajustados dos Modelos 1 e 2.

Com base apenas na avaliação da multicolinearidade, o modelo 1 pode ser considerado um pouco mais robusto em relação a esse aspecto, apresentando uma média (2,2028) de VIF (Fator de Inflação da Variância) inferior em comparação com o modelo 2 (3,0219).

Além disso, os resultados dos testes lineares gerais entre os modelos reduzidos e completos indicam que a inclusão das variáveis adicionais não resulta em uma melhoria estatisticamente significativa na explicação do tempo de internação para ambos os modelos selecionados.

Em ambas as comparações, o p-valor associado ao teste F é maior que o nível de significância de 5%, levando à não rejeição da hipótese nula (H_0) de que o modelo reduzido é suficiente. Assim, considerando a robustez em relação à multicolinearidade, evidenciada pela média de VIF, e a adequação estatística dos modelos, optou-se por escolher o modelo 1 como a abordagem mais parcimoniosa para explicar a variabilidade no tempo de internação. Diante disso, inicia-se a análise do modelo a partir da Figura 3.11, que apresenta seu correlograma.

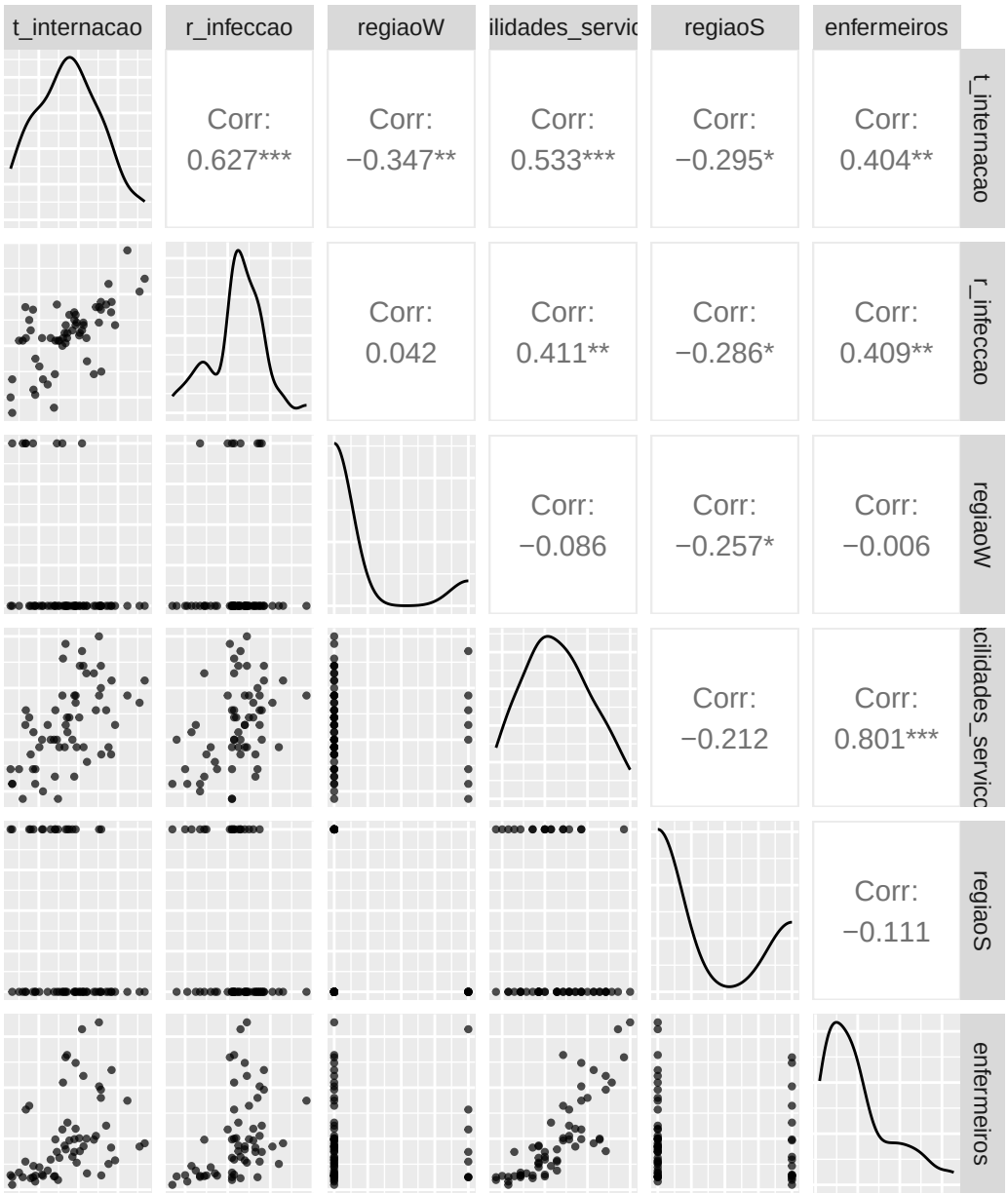

Figure 3.11: Correlograma do Modelo 1.

3.3 Valores Influentes

Na presente seção, foram adotadas medidas destinadas a ressaltar a influência de observações individuais nos parâmetros do modelo. As métricas utilizadas compreendem DFBetas, DFFits, *Cook's Distance*, *Leverage* e *Influence Total*.

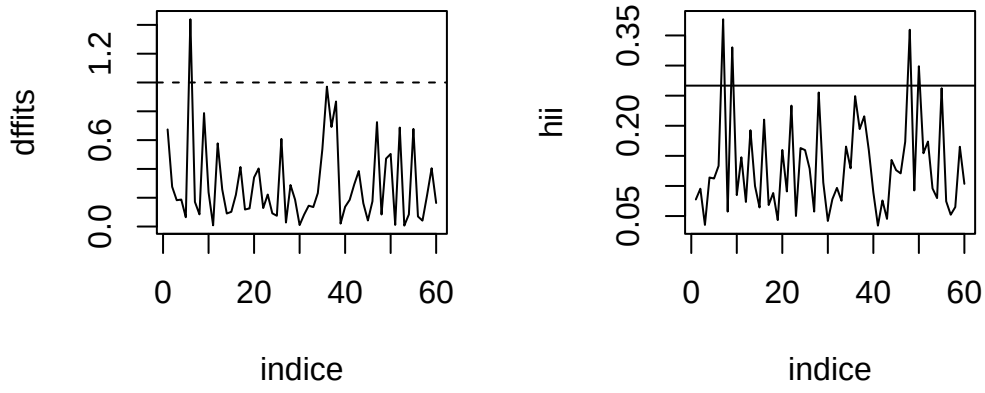

Figure 3.12: Gráficos de DFFits e Hii.

Os resultados indicam que algumas observações exercem influência substancial sobre o modelo. As observações de número 6, 35, 36, 38, 47 e 50 demonstraram um impacto significativo em diversas variáveis, evidenciando influência em diferentes aspectos do modelo. Por exemplo, a observação 6 revelou um efeito significativo nas variáveis idade e facilidades e servicos disponíveis.

No que diz respeito à DFFits (Figura 3.12), destaca-se a observção 6, indicando uma influência considerável sobre a estimativa ajustada da resposta. Analisando Cook's Distance (Figura 3.14), observou-se que as observações 6, 38 e outras se destacaram, sugerindo uma influência global no modelo. Essas observações podem ter um impacto desproporcional nas estimativas do modelo.

No contexto da avaliação de Leverage (*Hat*), as observações 6 e 38 foram identificadas com valores elevados, indicando que estão "distantes" das demais em termos das variáveis independentes. Isto pode ser observado na Figura 3.15.

Finalmente, ao avaliar a Influence Total, observações 6, 7, 28 e 38 foram identificadas como globalmente influentes, destacando a necessidade de uma análise mais aprofundada desses pontos específicos.

3.4 Validação

O MSPR (Mean Squared Prediction Error ou Erro Quadrático Médio de Previsão) é uma métrica essencial para avaliar a precisão das previsões de um modelo. Neste estudo, o MSPR calculado para o modelo 1 selecionado no conjunto de validação foi de 6,5303. Quanto menor o valor do MSPR, melhor o desempenho do modelo em realizar previsões precisas. Essa métrica representa a média dos quadrados dos erros de previsão para as observações na amostra de validação.

As métricas de desempenho do modelo selecionado foram analisadas nos conjuntos de treinamento e validação. No treinamento, o modelo apresentou um R^2 de 0,6433 e um MSE de 0,8843 1.0011, indicando um ajuste relativamente bom aos dados. Na validação, o R^2 foi de aproximadamente 0,4871 e o MSE foi de cerca de 3,2651. Essas métricas fornecem uma avaliação do ajuste do modelo aos dados, sendo desejáveis valores mais altos de R^2 e mais baixos de MSE.

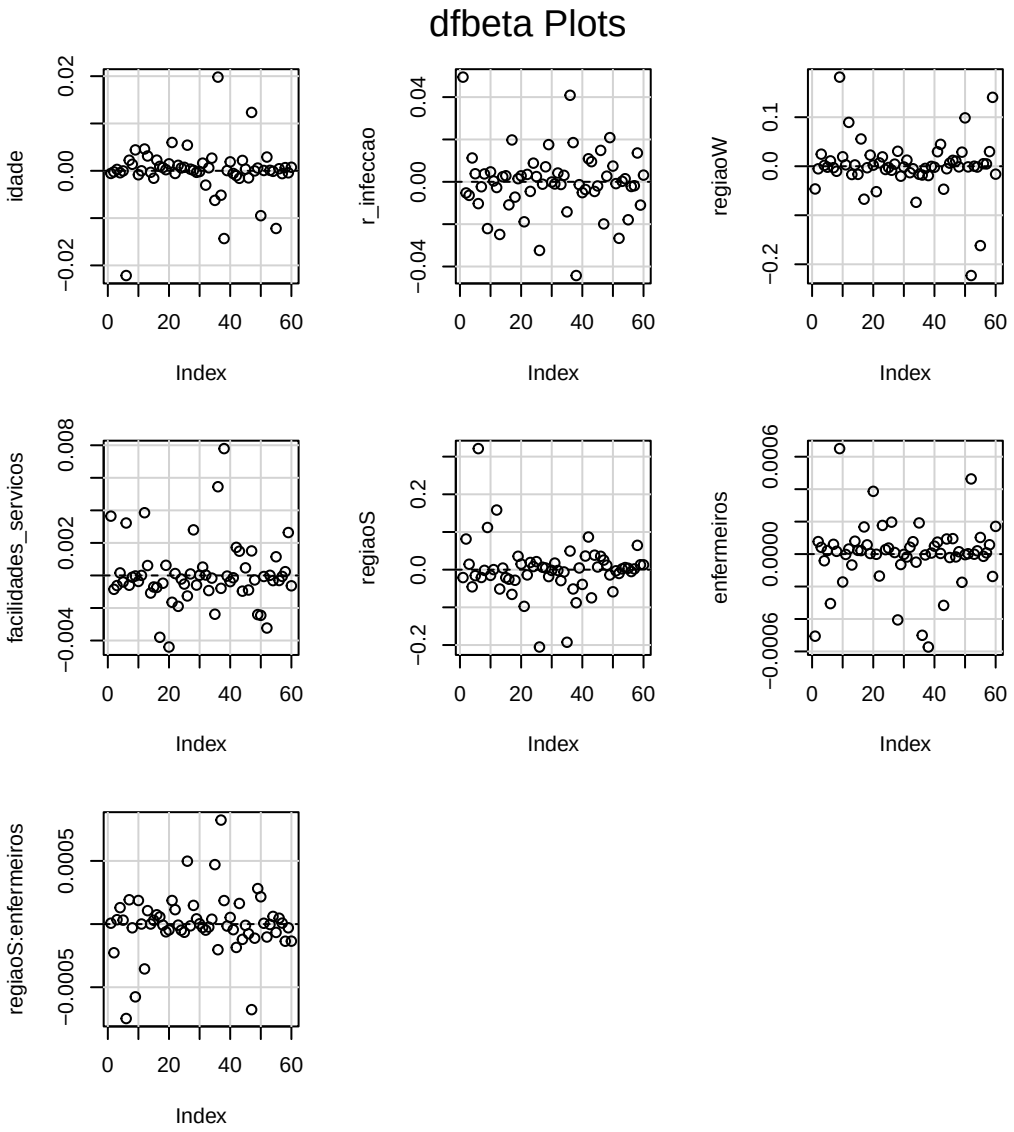

Figure 3.13: Gráficos de DFBetas e DFBetas Studentizados.

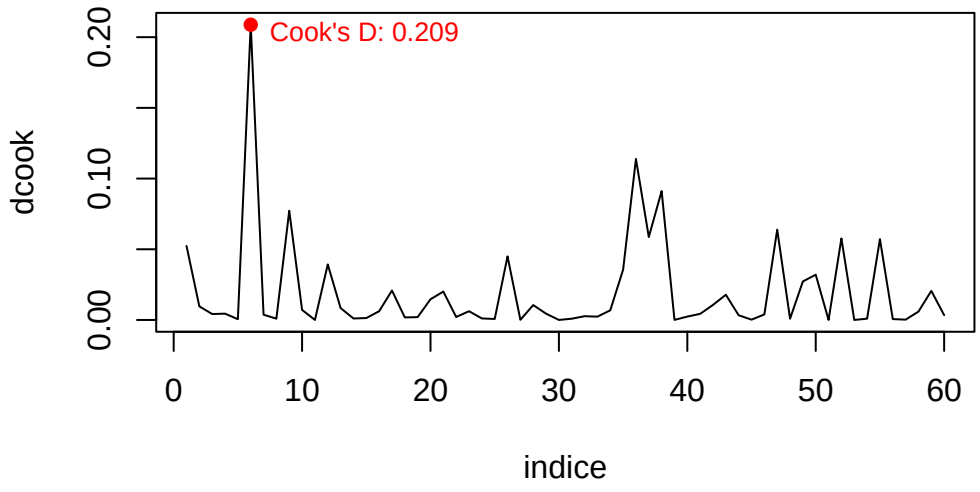

Figure 3.14: Gráfico de Cook's Distance.

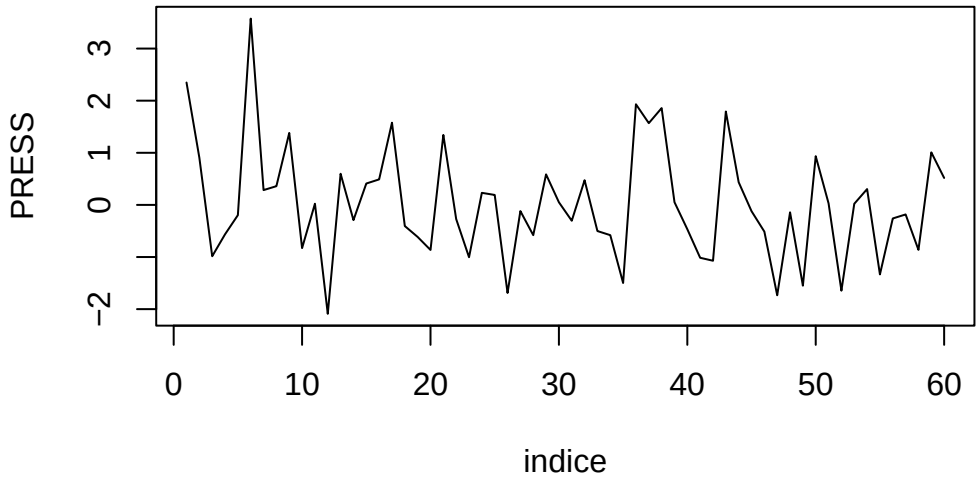

Figure 3.15: Gráfico de Leverage (Hat).

Table 3.3: Estatísticas do Modelo de Validação.

	Dependent variable:	
	t_internacao	
	Modelo de treinamento	Modelo de validação
idade	0.042	0.143**
	p = 0.178	p = 0.013
r_infeccao	0.634***	0.724***
	p = 0.00001	p = 0.001
regiaoW	−1.725***	−2.048***
	p = 0.0001	p = 0.007
facilidades_servicos	0.031**	−0.033
	p = 0.038	p = 0.221
regiaoS	−0.285	0.117
	p = 0.550	p = 0.908
enfermeiros	−0.0003	0.009**
	p = 0.838	p = 0.018
regiaoS:enfermeiros	−0.002	−0.007
	p = 0.287	p = 0.200
Constant	3.671**	−0.440
	p = 0.037	p = 0.888
Observations	60	53
R^2	0.643	0.487
Adjusted R^2	0.595	0.407
Residual Std. Error	1.010 (df = 52)	1.722 (df = 45)
F Statistic	13.398*** (df = 7; 52)	6.106*** (df = 7; 45)
Note:		*p<0.1; **p<0.05; ***p<0.01

Essas métricas são fundamentais para avaliar o desempenho do modelo, fornecendo insights valiosos para possíveis ajustes e melhorias.

4Conclusões

A presente pesquisa abrangeu uma análise ampla dos fatores associados à duração da internação hospitalar, incorporando variáveis individuais do paciente, características do tratamento e contexto hospitalar. A exploração de relações de segunda ordem e a inclusão de variáveis regionais enriqueceram a compreensão da complexidade desse fenômeno.

As correlações identificadas entre a duração da internação e variáveis como risco de infecção, número de leitos, média diária de pacientes, quantidade de enfermeiros(as) e a disponibilidade de facilidades e serviços hospitalares sugerem influências significativas nesse tempo. Destaca-se a importância crítica de fatores como a disponibilidade de leitos, o número de enfermeiros(as) e a percentagem de facilidades e serviços na gestão eficaz da duração da internação.

O modelo de regressão linear múltipla, mesmo sem incorporar interações ou termos de segunda ordem, demonstrou uma capacidade notável ao explicar aproximadamente 67.55% da variação na duração da internação. Resultados indicaram que o risco de infecção, região geográfica e características básicas do paciente e tratamento são elementos cruciais na determinação desse tempo. O modelo evidenciou a falta de associação estatisticamente significativa entre a idade média dos pacientes e a duração da internação.

Ao considerar a complexidade das relações entre variáveis, a inclusão de interações e termos de segunda ordem proporcionou insights adicionais. A interação entre enfermeiros e a região W, por exemplo, destacou uma associação substancial entre o aumento do número de enfermeiros e uma redução mais acentuada na duração da internação. Entretanto, a análise apontou a necessidade de uma investigação mais profunda sobre a presença de multicolinearidade entre variáveis independentes.

Na fase de seleção de variáveis, dois modelos foram considerados, com o Modelo com 5 Variáveis sendo escolhido devido à sua maior robustez em relação à multicolinearidade e adequação estatística. A influência de observações individuais nos parâmetros do modelo foi avaliada, destacando pontos específicos que exercem impacto considerável.

A validação do modelo foi realizada utilizando o MSPR, R^2 e MSE nos conjuntos de treinamento e validação. O modelo apresentou um desempenho satisfatório nos dados de treinamento, com um R^2 de 0.6433 e um MSE de 1.0011. No entanto, no conjunto de validação, o R^2 foi de aproximadamente `r format(round(summary(modelo_validacao)$r.squared, digits = 4), decimal.mark = ",")` e o MSE foi de cerca de 3.265, indicando espaço para melhorias e ajustes.

Em resumo, este estudo fornece uma visão abrangente dos fatores que influenciam a duração da internação hospitalar, destacando a importância de variáveis específicas e ressaltando a necessidade contínua de aprimoramento do modelo para melhor compreensão e previsão desse fenômeno complexo. I

Referências

Kutner, M., C. Nachtsheim, J. Neter, and W. Li. 2004. *Applied Linear Statistical Models*. McGraw-Hill Companies,Incorporated. https://books.google.com.br/books?id=0Qq-swEACAAJ.