TRABALHO FINAL DE REGRESSÃO LINEAR

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

Professora Maria Theresa

Table of contents

UnB

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.

1 Objetivos

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, $x_1, x_2, \ldots, x_n$ são as variáveis independentes, $\beta_0, \beta_1, \ldots, \beta_n$ 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:

$$\begin{aligned} E[Y] = \beta_0 + \beta_1(1) + \beta_2X_2 = (\beta_0 + \beta_1) + \beta_2X_2 \quad &, \text{se } X_1 = 1 \\ E[Y] = \beta_0 + \beta_1(0) + \beta_2X_2 = \beta_0 + \beta_2X_2 \quad &, \text{se } X_1 = 0 \end{aligned} \tag{2.3}$$

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:

$$Y_i = \beta_0 + \beta_1X_{1i} + \beta_2X_{2i} + \beta_{1,2}X_{1i}X_{2i} + \varepsilon_i \tag{2.4}$$

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 $\mathbf{Y}$ é o vetor de variáveis dependentes, $\mathbf{X}$ 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:

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

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:

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

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:

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

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.

3 Resultados

3.1 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 3.1.

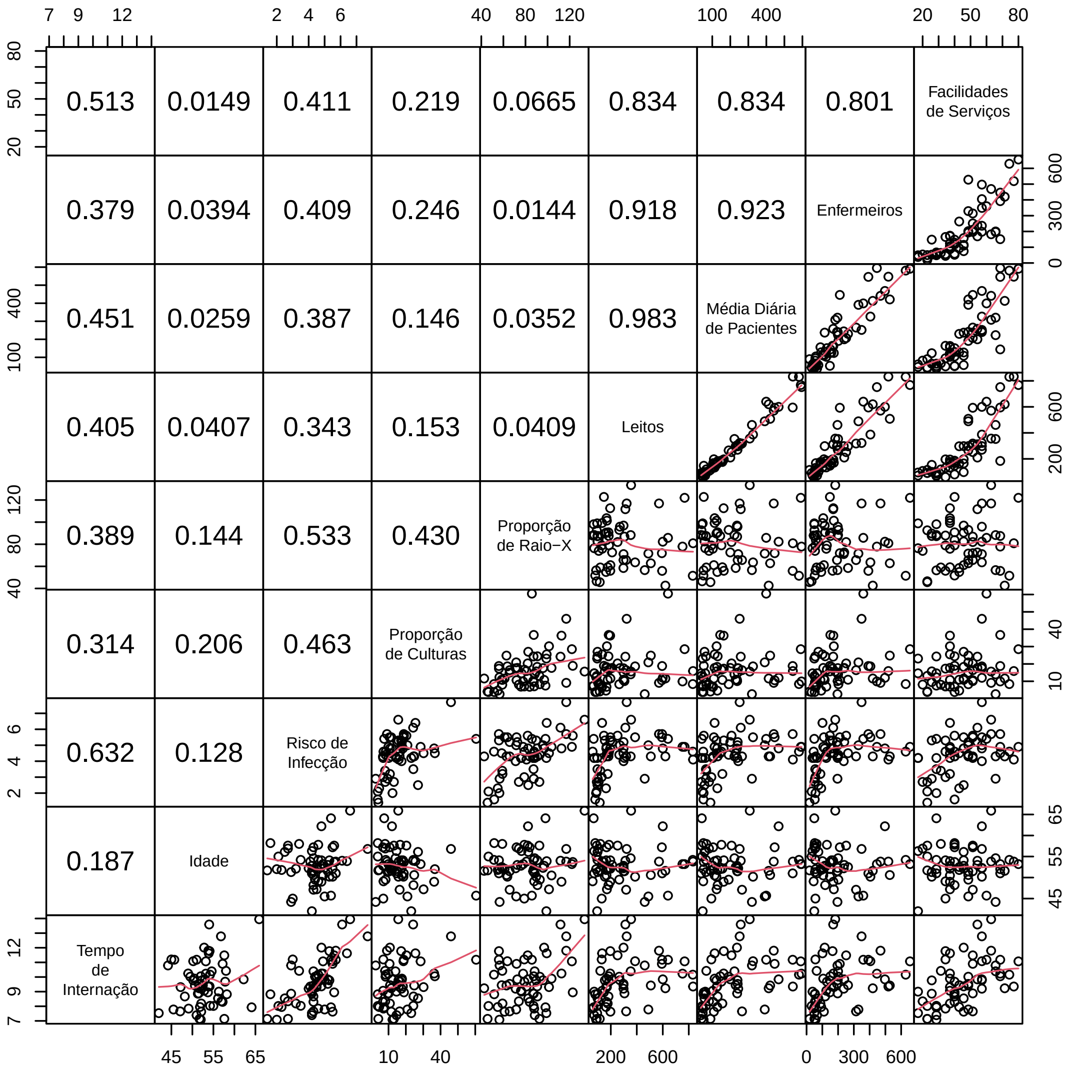


Figure 3.1: Correlações entre as variáveis quantitativas.

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 investigações 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.

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

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.

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.

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 3.2. 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.

A análise da Figura 3.3 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

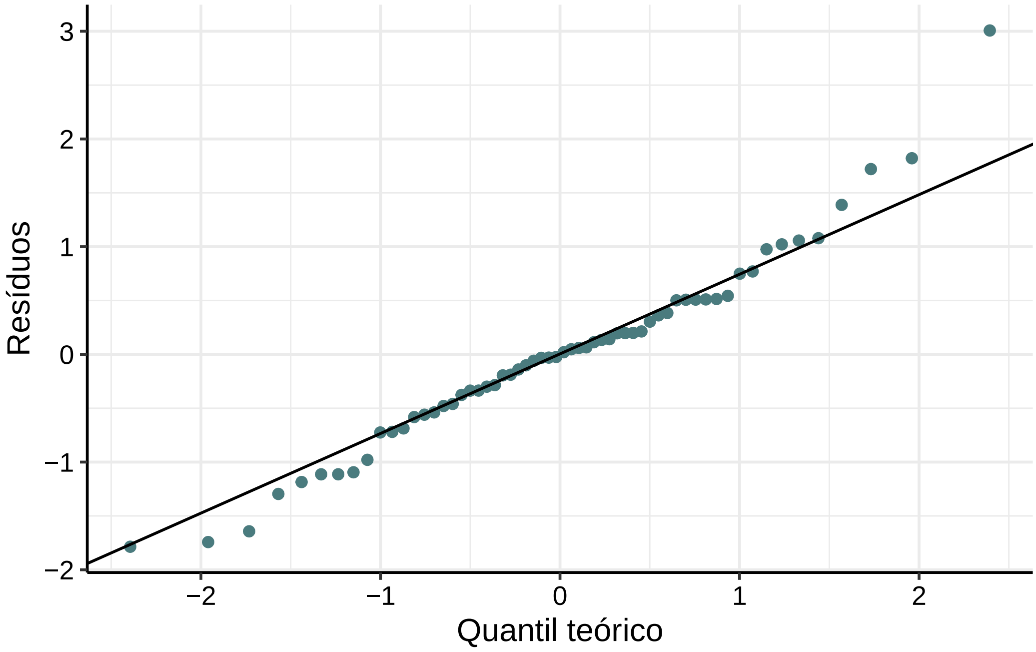


Figure 3.2: Gráfico de distribuição normal dos resíduos

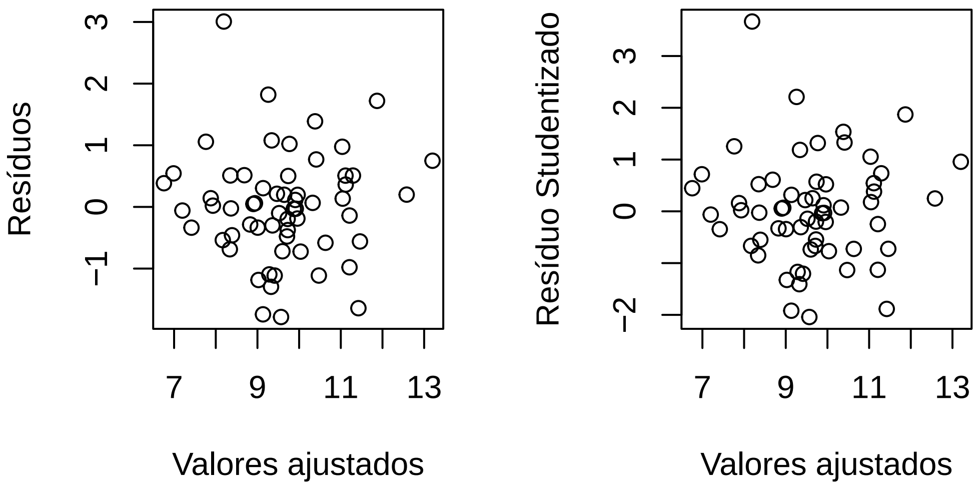


Figure 3.3: Gráfico de Resíduos e Resíduos Studentizados vs Valores Ajustados.

0,38. A confirmação da normalidade, homocedasticidade e independência dos resíduos fortalece a robustez do modelo, oferecendo suporte à validade das inferências derivadas.

3.1 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.

3.1.1.1 tabela de variáveis?//

3.1.1.2 tabela dos modelos e estatísticas?//

```
Start:  AIC=18.81
t_internacao ~ idade + r_infeccao + prop_culturas + prop_raiox +
    leitos + escola_medicina + regiao + m_dia_pacientes + enfermeiros +
    facilidades_servicos + enfermeiros * regiao + I(enfermeiros^2) +
    I(facilidades_servicos^2)

                            Df Sum of Sq    RSS    AIC
- regiao:enfermeiros         3    2.9477 47.998 16.609
- m_dia_pacientes            1    0.0501 45.101 16.873
- facilidades_servicos       1    0.1024 45.153 16.942
- I(facilidades_servicos^2)  1    0.1743 45.225 17.038
- leitos                     1    0.2428 45.293 17.129
- escola_medicina            1    0.4343 45.485 17.382
- prop_culturas              1    0.6262 45.677 17.634
- idade                      1    0.8971 45.948 17.989
- I(enfermeiros^2)           1    1.1080 46.158 18.264
<none>                                  45.050 18.806
- prop_raiox                 1    2.1376 47.188 19.588
- r_infeccao                 1    7.2797 52.330 25.794

Step:  AIC=16.61
t_internacao ~ idade + r_infeccao + prop_culturas + prop_raiox +
    leitos + escola_medicina + regiao + m_dia_pacientes + enfermeiros +
    facilidades_servicos + I(enfermeiros^2) + I(facilidades_servicos^2)

                            Df Sum of Sq    RSS    AIC
- I(facilidades_servicos^2)  1    0.0006 47.999 14.610
- enfermeiros                1    0.0024 48.001 14.612
- facilidades_servicos       1    0.0173 48.015 14.631
- m_dia_pacientes            1    0.0440 48.042 14.664
- leitos                     1    0.1791 48.177 14.832
- I(enfermeiros^2)           1    0.2206 48.219 14.884
- idade                      1    0.5467 48.545 15.289
- prop_culturas              1    0.7951 48.793 15.595
- escola_medicina            1    1.0388 49.037 15.894
- prop_raiox                 1    1.2257 49.224 16.122
<none>                                  47.998 16.609
```

```
+ regiao:enfermeiros         3    2.9477 45.050 18.806
- r_infeccao                 1    8.4900 56.488 24.381
- regiao                     3   17.3565 65.355 29.129

Step:  AIC=14.61
t_internacao ~ idade + r_infeccao + prop_culturas + prop_raiox +
    leitos + escola_medicina + regiao + m_dia_pacientes + enfermeiros +
    facilidades_servicos + I(enfermeiros^2)

                            Df Sum of Sq    RSS    AIC
- enfermeiros                1    0.0042 48.003 12.615
- m_dia_pacientes            1    0.0436 48.042 12.664
- leitos                     1    0.1877 48.186 12.844
- facilidades_servicos       1    0.2563 48.255 12.929
- I(enfermeiros^2)           1    0.2693 48.268 12.945
- idade                      1    0.5475 48.546 13.290
- prop_culturas              1    0.7956 48.794 13.596
- escola_medicina            1    1.0977 49.096 13.967
- prop_raiox                 1    1.2339 49.233 14.133
<none>                                   47.999 14.610
+ I(facilidades_servicos^2)  1    0.0006 47.998 16.609
+ regiao:enfermeiros         3    2.7740 45.225 17.038
- r_infeccao                 1    8.5046 56.503 22.397
- regiao                     3   17.4328 65.432 27.200

Step:  AIC=12.62
t_internacao ~ idade + r_infeccao + prop_culturas + prop_raiox +
    leitos + escola_medicina + regiao + m_dia_pacientes + facilidades_servicos +
    I(enfermeiros^2)

                            Df Sum of Sq    RSS    AIC
- m_dia_pacientes            1    0.0397 48.043 10.665
- leitos                     1    0.2186 48.222 10.888
- facilidades_servicos       1    0.2966 48.300 10.985
- idade                      1    0.5901 48.593 11.348
- prop_culturas              1    0.9558 48.959 11.798
- escola_medicina            1    1.2677 49.271 12.179
- prop_raiox                 1    1.3873 49.390 12.325
<none>                                   48.003 12.615
- I(enfermeiros^2)           1    1.9218 49.925 12.970
+ enfermeiros                1    0.0042 47.999 14.610
+ I(facilidades_servicos^2)  1    0.0024 48.001 14.612
- r_infeccao                 1   10.0704 58.073 22.042
- regiao                     3   18.0275 66.031 25.746

Step:  AIC=10.66
t_internacao ~ idade + r_infeccao + prop_culturas + prop_raiox +
    leitos + escola_medicina + regiao + facilidades_servicos +
    I(enfermeiros^2)

                            Df Sum of Sq    RSS     AIC
- facilidades_servicos       1    0.2920 48.335  9.0282
- idade                      1    0.5691 48.612  9.3712
```

```
- prop_culturas               1    1.1919 49.235 10.1351
- prop_raiox                  1    1.3482 49.391 10.3252
- escola_medicina             1    1.4804 49.523 10.4856
<none>                                    48.043 10.6647
- I(enfermeiros^2)            1    1.8833 49.926 10.9718
- leitos                      1    2.0067 50.049 11.1199
+ m_dia_pacientes             1    0.0397 48.003 12.6151
+ enfermeiros                 1    0.0003 48.042 12.6643
+ I(facilidades_servicos^2)   1    0.0000 48.043 12.6647
- r_infeccao                  1   13.3172 61.360 23.3447
- regiao                      3   21.9945 70.037 27.2809

Step:  AIC=9.03
t_internacao ~ idade + r_infeccao + prop_culturas + prop_raiox +
    leitos + escola_medicina + regiao + I(enfermeiros^2)

                            Df Sum of Sq    RSS     AIC
- idade                       1    0.6086 48.943  7.7790
- prop_culturas               1    1.2845 49.619  8.6020
- prop_raiox                  1    1.3533 49.688  8.6851
<none>                                    48.335  9.0282
- I(enfermeiros^2)            1    2.1174 50.452  9.6008
- escola_medicina             1    2.1559 50.491  9.6464
+ facilidades_servicos        1    0.2920 48.043 10.6647
+ I(facilidades_servicos^2)   1    0.2639 48.071 10.6998
+ enfermeiros                 1    0.0683 48.266 10.9434
+ m_dia_pacientes             1    0.0351 48.300 10.9847
- leitos                      1    4.4524 52.787 12.3153
- r_infeccao                  1   14.1692 62.504 22.4530
- regiao                      3   24.3458 72.680 27.5036

Step:  AIC=7.78
t_internacao ~ r_infeccao + prop_culturas + prop_raiox + leitos +
    escola_medicina + regiao + I(enfermeiros^2)

                            Df Sum of Sq    RSS     AIC
- prop_raiox                  1    1.6345 50.578  7.7500
<none>                                    48.943  7.7790
- I(enfermeiros^2)            1    1.9624 50.906  8.1377
- escola_medicina             1    1.9779 50.921  8.1560
- prop_culturas               1    2.4045 51.348  8.6566
+ idade                       1    0.6086 48.335  9.0282
+ facilidades_servicos        1    0.3315 48.612  9.3712
+ I(facilidades_servicos^2)   1    0.3303 48.613  9.3727
+ enfermeiros                 1    0.0152 48.928  9.7604
+ m_dia_pacientes             1    0.0148 48.928  9.7609
- leitos                      1    4.3344 53.278 10.8704
- r_infeccao                  1   15.5036 64.447 22.2897
- regiao                      3   27.1215 76.065 28.2344

Step:  AIC=7.75
t_internacao ~ r_infeccao + prop_culturas + leitos + escola_medicina +
    regiao + I(enfermeiros^2)
```

```
                             Df Sum of Sq    RSS     AIC
- I(enfermeiros^2)            1    1.6771 52.255  7.7072
<none>                                    50.578  7.7500
+ prop_raiox                  1    1.6345 48.943  7.7790
- prop_culturas               1    1.7667 52.344  7.8100
- escola_medicina             1    2.1252 52.703  8.2195
+ idade                       1    0.8898 49.688  8.6851
+ I(facilidades_servicos^2)   1    0.4644 50.113  9.1965
+ facilidades_servicos        1    0.3468 50.231  9.3371
+ enfermeiros                 1    0.0827 50.495  9.6518
- leitos                      1    3.4475 54.025  9.7064
+ m_dia_pacientes             1    0.0190 50.559  9.7275
- regiao                      3   27.9339 78.512 28.1341
- r_infeccao                  1   26.4002 76.978 30.9505

Step:  AIC=7.71
t_internacao ~ r_infeccao + prop_culturas + leitos + escola_medicina +
    regiao

                             Df Sum of Sq    RSS    AIC
<none>                                    52.255  7.707
+ I(enfermeiros^2)            1    1.6771 50.578  7.750
+ enfermeiros                 1    1.5884 50.666  7.855
- escola_medicina             1    1.9778 54.233  7.936
- prop_culturas               1    1.9907 54.245  7.950
- leitos                      1    2.0282 54.283  7.992
+ prop_raiox                  1    1.3492 50.906  8.138
+ idade                       1    0.6870 51.568  8.913
+ facilidades_servicos        1    0.5599 51.695  9.061
+ I(facilidades_servicos^2)   1    0.5267 51.728  9.099
+ m_dia_pacientes             1    0.1070 52.148  9.584
- regiao                      3   29.0340 81.289 28.220
- r_infeccao                  1   28.9259 81.181 32.140


Call:
lm(formula = t_internacao ~ r_infeccao + prop_culturas + leitos +
    escola_medicina + regiao, data = base_new)

Residuals:
     Min       1Q   Median       3Q      Max
-1.95948 -0.47967 -0.03284  0.50181  2.93722

Coefficients:
                    Estimate Std. Error t value           Pr(>|t|)
(Intercept)         7.967780   0.751259  10.606 0.0000000000000131 ***
r_infeccao          0.687316   0.128108   5.365 0.0000018964125281 ***
prop_culturas      -0.021608   0.015352  -1.407            0.16524
leitos              0.001121   0.000789   1.421            0.16138
escola_medicinaNão -0.601445   0.428710  -1.403            0.16658
regiaoNC           -0.724846   0.367959  -1.970            0.05419 .
regiaoS            -1.159967   0.382956  -3.029            0.00382 **
```

```
regiaoW           -2.333658   0.445522  -5.238 0.0000029789393262 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.002 on 52 degrees of freedom
Multiple R-squared:  0.6487,    Adjusted R-squared:  0.6014
F-statistic: 13.72 on 7 and 52 DF,  p-value: 0.0000000006689


Start:  AIC=18.81
t_internacao ~ idade + r_infeccao + prop_culturas + prop_raiox +
    leitos + escola_medicina + regiao + m_dia_pacientes + enfermeiros +
    facilidades_servicos + enfermeiros * regiao + I(enfermeiros^2) +
    I(facilidades_servicos^2)



Call:
lm(formula = t_internacao ~ idade + r_infeccao + prop_culturas +
    prop_raiox + leitos + escola_medicina + regiao + m_dia_pacientes +
    enfermeiros + facilidades_servicos + enfermeiros * regiao +
    I(enfermeiros^2) + I(facilidades_servicos^2), data = base_new)

Residuals:
     Min       1Q   Median       3Q      Max
-1.78601 -0.49401 -0.00211  0.50312  3.00704

Coefficients:
                              Estimate   Std. Error t value Pr(>|t|)
(Intercept)                5.286782505  2.471613531   2.139  0.03829 *
idade                      0.033110928  0.036205218   0.915  0.36566
r_infeccao                 0.501030791  0.192323177   2.605  0.01265 *
prop_culturas             -0.015217111  0.019916689  -0.764  0.44912
prop_raiox                 0.013007636  0.009214298   1.412  0.16541
leitos                     0.002180888  0.004584341   0.476  0.63674
escola_medicinaNão        -0.357753850  0.562219118  -0.636  0.52802
regiaoNC                  -0.449397144  0.694364256  -0.647  0.52102
regiaoS                   -0.741832136  0.635035808  -1.168  0.24933
regiaoW                   -2.651450379  0.754361773  -3.515  0.00107 **
m_dia_pacientes            0.001321221  0.006113252   0.216  0.82994
enfermeiros                0.003343730  0.007328944   0.456  0.65057
facilidades_servicos      -0.021252007  0.068791305  -0.309  0.75890
I(enfermeiros^2)          -0.000009917  0.000009757  -1.016  0.31529
I(facilidades_servicos^2)  0.000292527  0.000725653   0.403  0.68890
regiaoNC:enfermeiros      -0.001280466  0.002913130  -0.440  0.66252
regiaoS:enfermeiros       -0.001562367  0.002694930  -0.580  0.56518
regiaoW:enfermeiros        0.003007525  0.002803330   1.073  0.28947
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.036 on 42 degrees of freedom
Multiple R-squared:  0.6971,    Adjusted R-squared:  0.5746
F-statistic: 5.687 on 17 and 42 DF,  p-value: 0.000002271


Start:  AIC=18.81
```

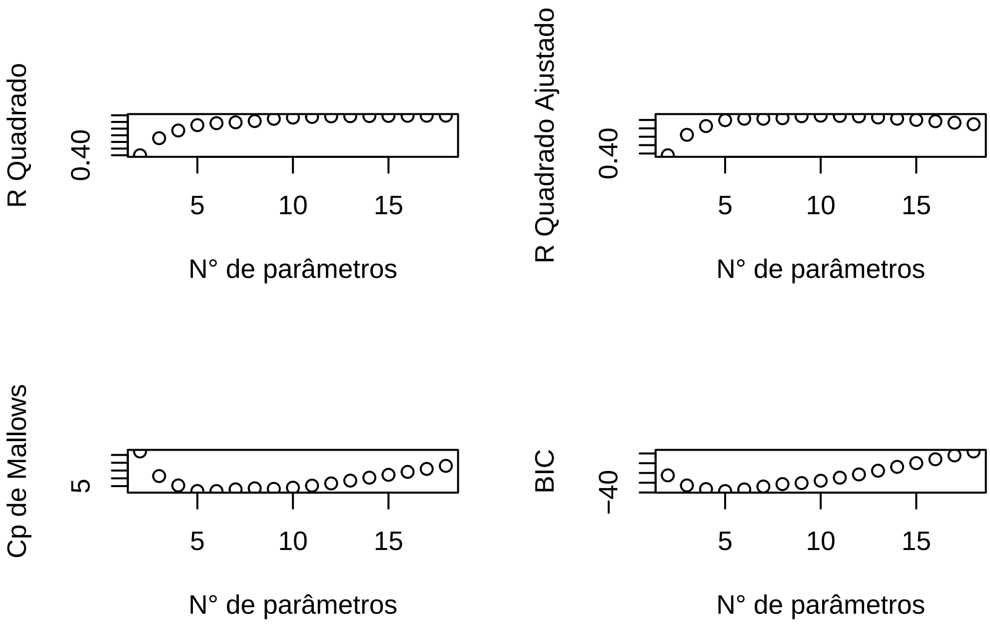
```
t_internacao ~ idade + r_infeccao + prop_culturas + prop_raiox +
    leitos + escola_medicina + regiao + m_dia_pacientes + enfermeiros +
    facilidades_servicos + enfermeiros * regiao + I(enfermeiros^2) +
    I(facilidades_servicos^2)

                            Df Sum of Sq    RSS    AIC
- regiao:enfermeiros         3    2.9477 47.998 16.609
- m_dia_pacientes            1    0.0501 45.101 16.873
- facilidades_servicos       1    0.1024 45.153 16.942
- I(facilidades_servicos^2)  1    0.1743 45.225 17.038
- leitos                     1    0.2428 45.293 17.129
- escola_medicina            1    0.4343 45.485 17.382
- prop_culturas              1    0.6262 45.677 17.634
- idade                      1    0.8971 45.948 17.989
- I(enfermeiros^2)           1    1.1080 46.158 18.264
<none>                                   45.050 18.806
- prop_raiox                 1    2.1376 47.188 19.588
- r_infeccao                 1    7.2797 52.330 25.794

Step:  AIC=16.61
t_internacao ~ idade + r_infeccao + prop_culturas + prop_raiox +
    leitos + escola_medicina + regiao + m_dia_pacientes + enfermeiros +
    facilidades_servicos + I(enfermeiros^2) + I(facilidades_servicos^2)

                            Df Sum of Sq    RSS    AIC
- I(facilidades_servicos^2)  1    0.0006 47.999 14.610
- enfermeiros                1    0.0024 48.001 14.612
- facilidades_servicos       1    0.0173 48.015 14.631
- m_dia_pacientes            1    0.0440 48.042 14.664
- leitos                     1    0.1791 48.177 14.832
- I(enfermeiros^2)           1    0.2206 48.219 14.884
- idade                      1    0.5467 48.545 15.289
- prop_culturas              1    0.7951 48.793 15.595
- escola_medicina            1    1.0388 49.037 15.894
- prop_raiox                 1    1.2257 49.224 16.122
<none>                                   47.998 16.609
- r_infeccao                 1    8.4900 56.488 24.381
- regiao                     3   17.3565 65.355 29.129

Step:  AIC=14.61
t_internacao ~ idade + r_infeccao + prop_culturas + prop_raiox +
    leitos + escola_medicina + regiao + m_dia_pacientes + enfermeiros +
    facilidades_servicos + I(enfermeiros^2)

                       Df Sum of Sq    RSS    AIC
- enfermeiros           1    0.0042 48.003 12.615
- m_dia_pacientes       1    0.0436 48.042 12.664
- leitos                1    0.1877 48.186 12.844
- facilidades_servicos  1    0.2563 48.255 12.929
- I(enfermeiros^2)      1    0.2693 48.268 12.945
- idade                 1    0.5475 48.546 13.290
- prop_culturas         1    0.7956 48.794 13.596
- escola_medicina       1    1.0977 49.096 13.967
```

```
- prop_raiox           1    1.2339 49.233 14.133
<none>                            47.999 14.610
- r_infeccao           1    8.5046 56.503 22.397
- regiao               3   17.4328 65.432 27.200

Step:  AIC=12.62
t_internacao ~ idade + r_infeccao + prop_culturas + prop_raiox +
    leitos + escola_medicina + regiao + m_dia_pacientes + facilidades_servicos +
    I(enfermeiros^2)

                       Df Sum of Sq    RSS    AIC
- m_dia_pacientes       1    0.0397 48.043 10.665
- leitos                1    0.2186 48.222 10.888
- facilidades_servicos  1    0.2966 48.300 10.985
- idade                 1    0.5901 48.593 11.348
- prop_culturas         1    0.9558 48.959 11.798
- escola_medicina       1    1.2677 49.271 12.179
- prop_raiox            1    1.3873 49.390 12.325
<none>                            48.003 12.615
- I(enfermeiros^2)      1    1.9218 49.925 12.970
- r_infeccao            1   10.0704 58.073 22.042
- regiao                3   18.0275 66.031 25.746

Step:  AIC=10.66
t_internacao ~ idade + r_infeccao + prop_culturas + prop_raiox +
    leitos + escola_medicina + regiao + facilidades_servicos +
    I(enfermeiros^2)

                       Df Sum of Sq    RSS     AIC
- facilidades_servicos  1    0.2920 48.335  9.0282
- idade                 1    0.5691 48.612  9.3712
- prop_culturas         1    1.1919 49.235 10.1351
- prop_raiox            1    1.3482 49.391 10.3252
- escola_medicina       1    1.4804 49.523 10.4856
<none>                            48.043 10.6647
- I(enfermeiros^2)      1    1.8833 49.926 10.9718
- leitos                1    2.0067 50.049 11.1199
- r_infeccao            1   13.3172 61.360 23.3447
- regiao                3   21.9945 70.037 27.2809

Step:  AIC=9.03
t_internacao ~ idade + r_infeccao + prop_culturas + prop_raiox +
    leitos + escola_medicina + regiao + I(enfermeiros^2)

                   Df Sum of Sq    RSS     AIC
- idade             1    0.6086 48.943  7.7790
- prop_culturas     1    1.2845 49.619  8.6020
- prop_raiox        1    1.3533 49.688  8.6851
<none>                        48.335  9.0282
- I(enfermeiros^2)  1    2.1174 50.452  9.6008
- escola_medicina   1    2.1559 50.491  9.6464
- leitos            1    4.4524 52.787 12.3153
- r_infeccao        1   14.1692 62.504 22.4530
```

```
- regiao             3   24.3458 72.680 27.5036

Step:  AIC=7.78
t_internacao ~ r_infeccao + prop_culturas + prop_raiox + leitos +
    escola_medicina + regiao + I(enfermeiros^2)

                   Df Sum of Sq    RSS     AIC
- prop_raiox        1    1.6345 50.578  7.7500
<none>                          48.943  7.7790
- I(enfermeiros^2)  1    1.9624 50.906  8.1377
- escola_medicina   1    1.9779 50.921  8.1560
- prop_culturas     1    2.4045 51.348  8.6566
- leitos            1    4.3344 53.278 10.8704
- r_infeccao        1   15.5036 64.447 22.2897
- regiao            3   27.1215 76.065 28.2344

Step:  AIC=7.75
t_internacao ~ r_infeccao + prop_culturas + leitos + escola_medicina +
    regiao + I(enfermeiros^2)

                   Df Sum of Sq    RSS     AIC
- I(enfermeiros^2)  1    1.6771 52.255  7.7072
<none>                          50.578  7.7500
- prop_culturas     1    1.7667 52.344  7.8100
- escola_medicina   1    2.1252 52.703  8.2195
- leitos            1    3.4475 54.025  9.7064
- regiao            3   27.9339 78.512 28.1341
- r_infeccao        1   26.4002 76.978 30.9505

Step:  AIC=7.71
t_internacao ~ r_infeccao + prop_culturas + leitos + escola_medicina +
    regiao

                  Df Sum of Sq    RSS    AIC
<none>                         52.255  7.707
- escola_medicina  1    1.9778 54.233  7.936
- prop_culturas    1    1.9907 54.245  7.950
- leitos           1    2.0282 54.283  7.992
- regiao           3   29.0340 81.289 28.220
- r_infeccao       1   28.9259 81.181 32.140



Call:
lm(formula = t_internacao ~ r_infeccao + prop_culturas + leitos +
    escola_medicina + regiao, data = base_new)

Residuals:
     Min       1Q   Median       3Q      Max
-1.95948 -0.47967 -0.03284  0.50181  2.93722

Coefficients:
                    Estimate Std. Error t value           Pr(>|t|)
(Intercept)         7.967780   0.751259  10.606 0.0000000000000131 ***
```

```
r_infeccao          0.687316   0.128108   5.365 0.0000018964125281 ***
prop_culturas      -0.021608   0.015352  -1.407            0.16524    
leitos              0.001121   0.000789   1.421            0.16138    
escola_medicinaNão -0.601445   0.428710  -1.403            0.16658    
regiaoNC           -0.724846   0.367959  -1.970            0.05419 .  
regiaoS            -1.159967   0.382956  -3.029            0.00382 ** 
regiaoW            -2.333658   0.445522  -5.238 0.0000029789393262 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.002 on 52 degrees of freedom
Multiple R-squared:  0.6487,    Adjusted R-squared:  0.6014 
F-statistic: 13.72 on 7 and 52 DF,  p-value: 0.0000000006689
```

Modelos	CP	RSQ	ADJR2	BIC
1	27.232843	0.3998261	0.3894783	-22.443459
2	11.451693	0.5280421	0.5114822	-32.768894
3	5.441333	0.5858030	0.5636139	-36.507443
4	1.865170	0.6260115	0.5988124	-38.540099
5	1.824131	0.6407290	0.6074632	-36.854636
6	2.910643	0.6473160	0.6073895	-33.870556
7	3.607358	0.6567137	0.6105021	-31.396673
8	3.261344	0.6736303	0.6224351	-30.334368
9	3.990342	0.6827952	0.6256984	-27.949016
10	5.305425	0.6877340	0.6240062	-24.796203
11	6.738989	0.6918185	0.6211935	-21.491839
12	8.534518	0.6932928	0.6149846	-17.685234
13	10.427219	0.6940666	0.6076071	-13.742438
14	12.248461	0.6953555	0.6005773	-9.901425
15	14.143281	0.6961140	0.5925165	-5.956640
16	16.046710	0.6968103	0.5839956	-1.999943
17	18.000000	0.6971471	0.5745638	2.027710



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.

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:

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

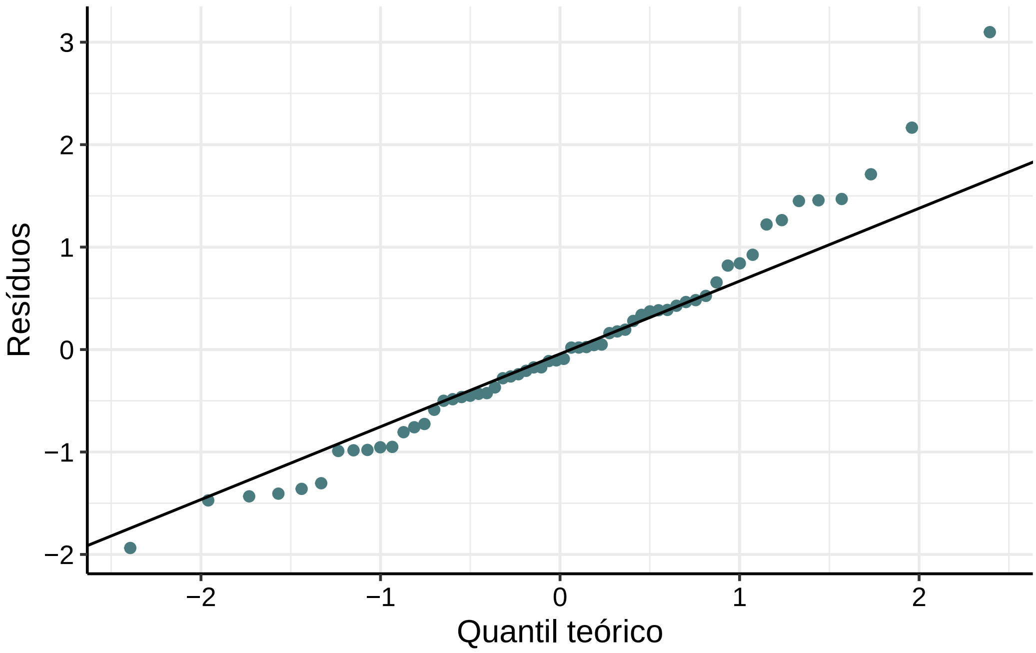
Além disso, o Modelo com 6 Variáveis expande a abordagem ao incorporar as variáveis de idade, risco de infecção, regiaoW, termo quadrático de facilidades e servicos disponíveis, interação entre regiaoNC e número de enfermeiros e interação entre a regiaoS e número de enfermeiros. A equação do modelo é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

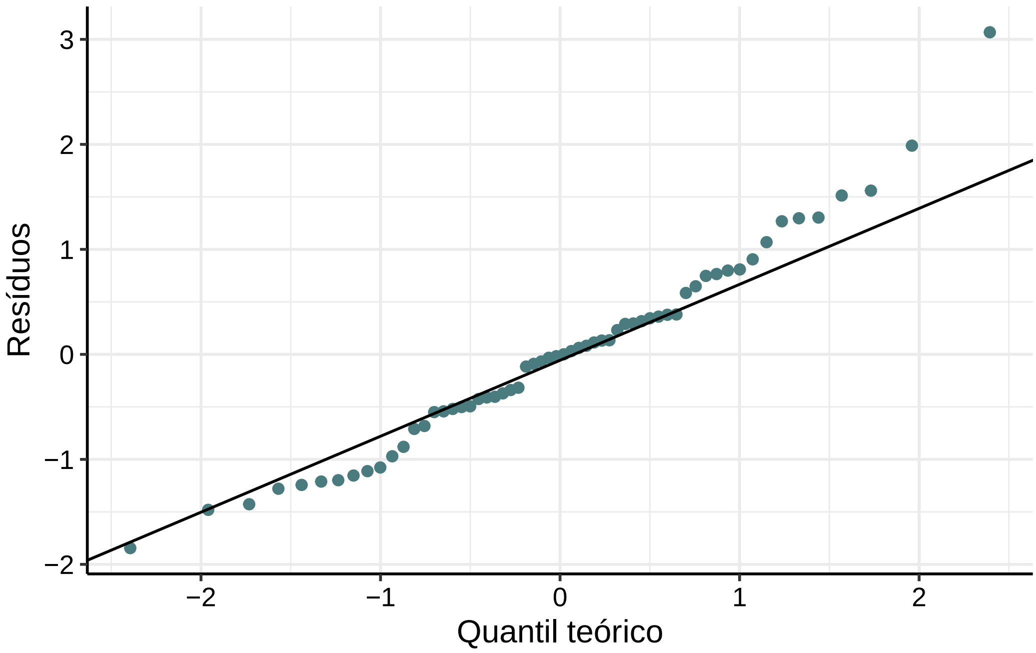
$$Y_i = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \beta_3 X_{10} + \beta_4 X_{13}^2 + \beta_5 X_8 X_{12} + \beta_6 X_9 X_{12} + \varepsilon_i$$

```
Call:
lm(formula = t_internacao ~ idade + r_infeccao + regiaoW + facilidades_servicos +
    regiaoS * enfermeiros, data = base_mods)

Residuals:
     Min       1Q   Median       3Q      Max
-1.93684 -0.52194 -0.09834  0.43673  3.09817

Coefficients:
                      Estimate Std. Error t value   Pr(>|t|)
(Intercept)          3.6708164  1.7112121   2.145     0.0366 *
idade                0.0416580  0.0304472   1.368     0.1771
r_infeccao           0.6340303  0.1223387   5.183 0.00000362 ***
regiaoW             -1.7254126  0.4056108  -4.254 0.00008777 ***
facilidades_servicos 0.0310965  0.0145468   2.138     0.0373 *
regiaoS             -0.2846061  0.4729830  -0.602     0.5500
enfermeiros         -0.0003107  0.0015084  -0.206     0.8376
regiaoS:enfermeiros -0.0019694  0.0018300  -1.076     0.2868
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.01 on 52 degrees of freedom
Multiple R-squared:  0.6433,    Adjusted R-squared:  0.5953
F-statistic:  13.4 on 7 and 52 DF,  p-value: 0.000000009751
```

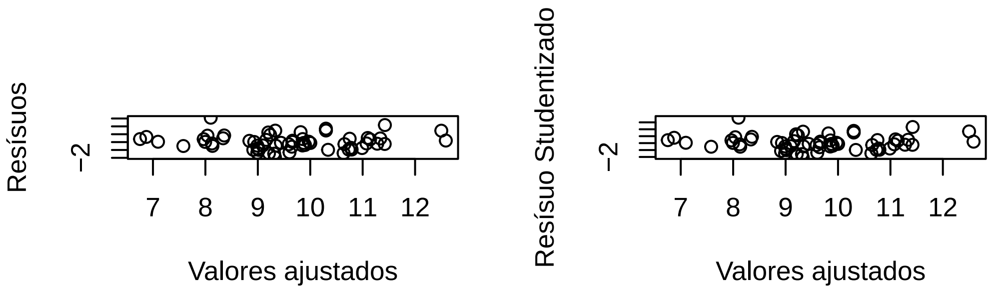
```
Call:
lm(formula = t_internacao ~ idade + r_infeccao + regiaoW + I(facilidades_servicos^2) +
    regiaoNC * enfermeiros + regiaoS * enfermeiros, data = base_mods)

Residuals:
```

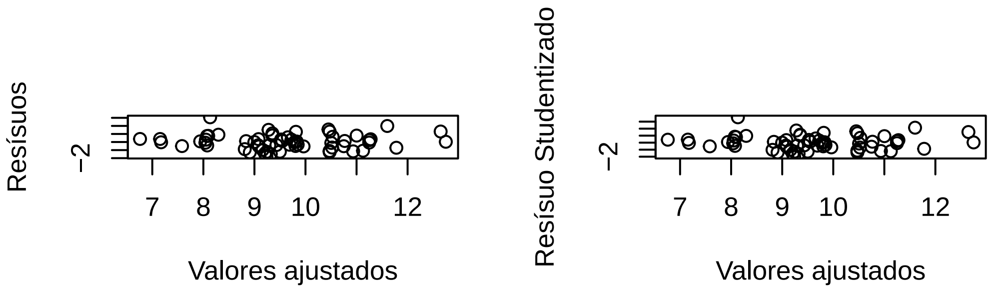
```
     Min       1Q   Median       3Q      Max 
-1.84507 -0.54434 -0.00822  0.43141  3.06766 

Coefficients:
                            Estimate Std. Error t value   Pr(>|t|)    
(Intercept)                5.0057901  1.7928804   2.792     0.0074 ** 
idade                      0.0316902  0.0313640   1.010     0.3172    
r_infeccao                 0.6452175  0.1269580   5.082 0.00000559 ***
regiaoW                   -1.9731008  0.4437199  -4.447 0.00004864 ***
I(facilidades_servicos^2)  0.0003485  0.0001601   2.176     0.0343 *  
regiaoNC                  -0.3179298  0.6194729  -0.513     0.6101    
enfermeiros               -0.0004626  0.0016573  -0.279     0.7813    
regiaoS                   -0.5129229  0.5447945  -0.941     0.3510    
regiaoNC:enfermeiros      -0.0004915  0.0022558  -0.218     0.8284    
enfermeiros:regiaoS       -0.0018981  0.0019777  -0.960     0.3418    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.015 on 50 degrees of freedom
Multiple R-squared:  0.6535,	Adjusted R-squared:  0.5912 
F-statistic: 10.48 on 9 and 50 DF,  p-value: 0.00000000681


	Shapiro-Wilk normality test

data:  modelo_select1$residuals
W = 0.97005, p-value = 0.1468


	Shapiro-Wilk normality test

data:  modelo_select2$residuals
W = 0.97451, p-value = 0.2411
```






```
	studentized Breusch-Pagan test

data:  modelo_select1
BP = 7.2329, df = 7, p-value = 0.405


	studentized Breusch-Pagan test

data:  modelo_select2
BP = 9.4704, df = 9, p-value = 0.395
```

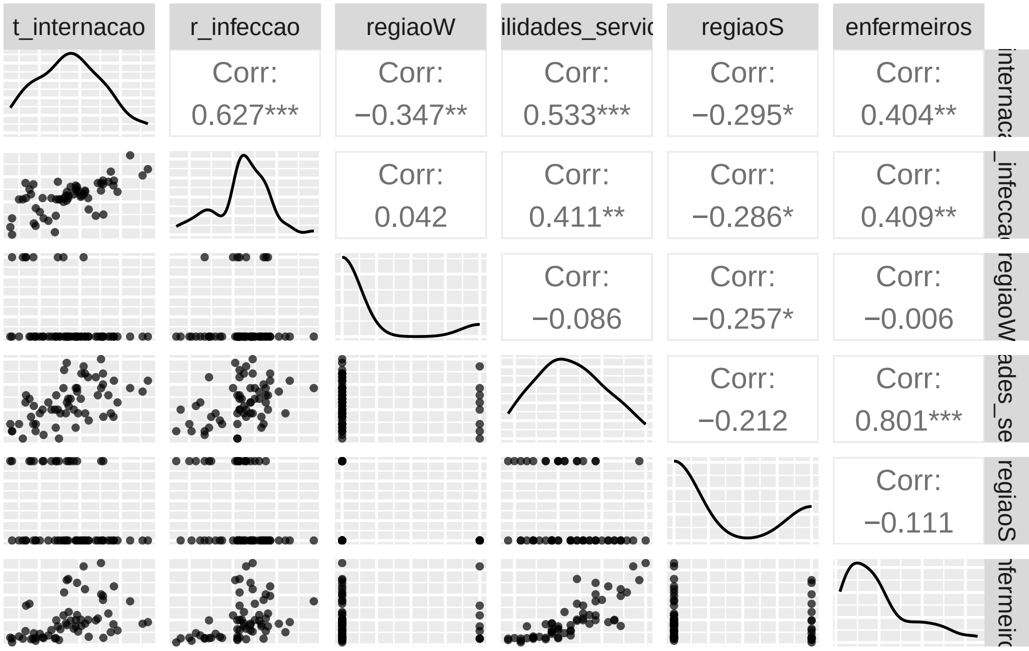


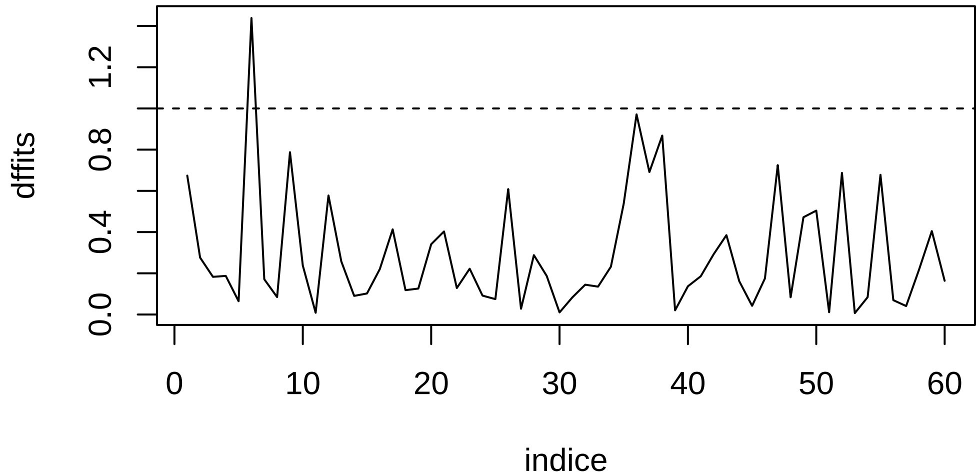


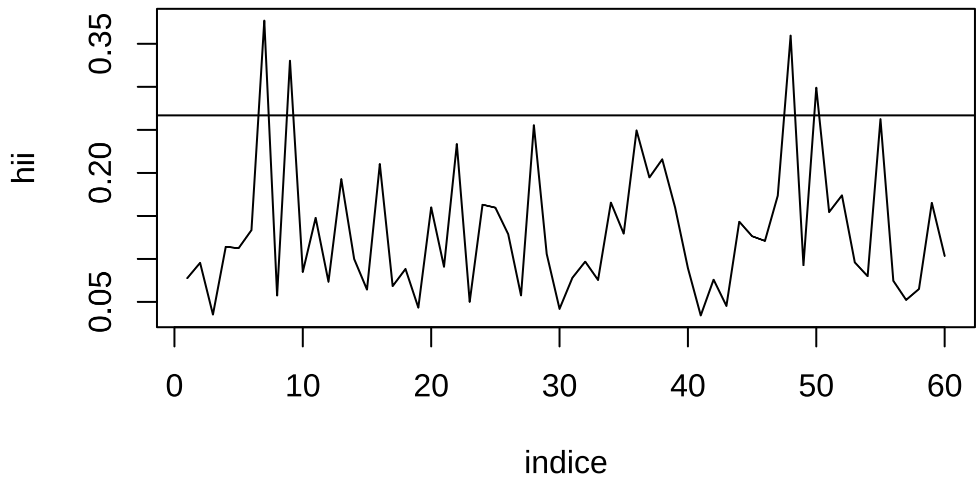

```
Analysis of Variance Table

Model 1: t_internacao ~ idade + r_infeccao + regiaoW + facilidades_servicos +
```

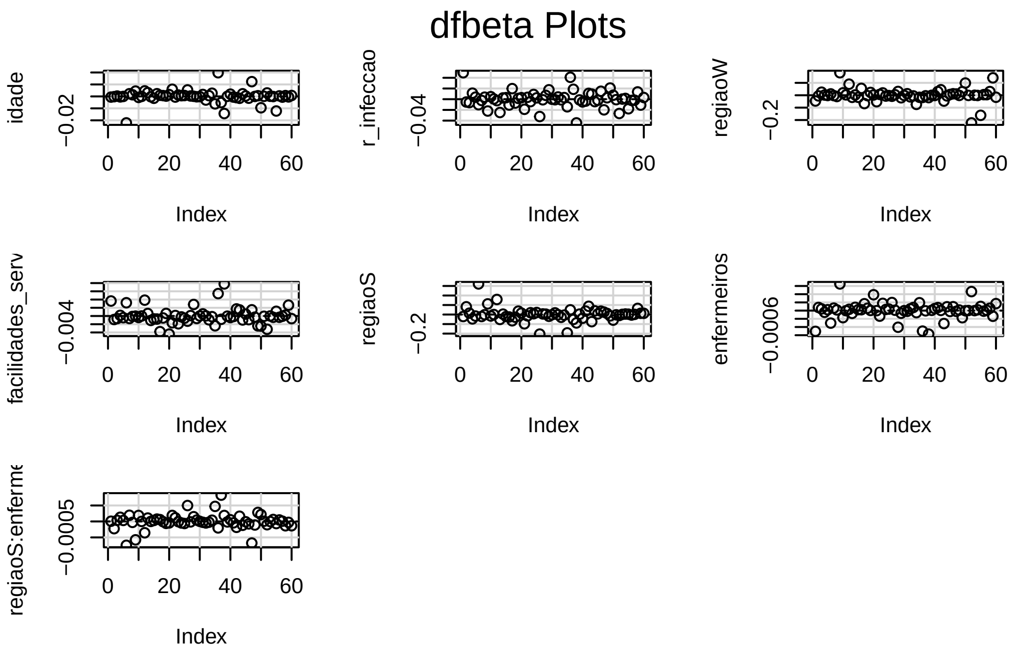
```
    regiaoS * enfermeiros
Model 2: t_internacao ~ idade + r_infeccao + prop_culturas + prop_raiox +
    leitos + escola_medicina + regiao + m_dia_pacientes + enfermeiros +
    facilidades_servicos + enfermeiros * regiao + I(enfermeiros^2) +
    I(facilidades_servicos^2)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     52 53.058
2     42 45.050 10    8.0075 0.7465 0.6773

Analysis of Variance Table

Model 1: t_internacao ~ idade + r_infeccao + regiaoW + I(facilidades_servicos^2) +
    regiaoNC * enfermeiros + regiaoS * enfermeiros
Model 2: t_internacao ~ idade + r_infeccao + prop_culturas + prop_raiox +
    leitos + escola_medicina + regiao + m_dia_pacientes + enfermeiros +
    facilidades_servicos + enfermeiros * regiao + I(enfermeiros^2) +
    I(facilidades_servicos^2)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     50 51.536
2     42 45.050  8    6.4856 0.7558 0.6427

               idade          r_infeccao             regiaoW
            1.046602            1.343286            1.117928
facilidades_servicos             regiaoS         enfermeiros
            3.071925            2.762576            3.353421
 regiaoS:enfermeiros
            2.723749

[1] 2.202784

                    idade                r_infeccao                  regiaoW
                 1.099401                  1.432078                 1.324399
I(facilidades_servicos^2)                  regiaoNC              enfermeiros
                 3.379157                  4.535964                 4.007632
                  regiaoS      regiaoNC:enfermeiros      enfermeiros:regiaoS
                 3.628231                  4.641548                 3.149104

[1] 3.021946
```

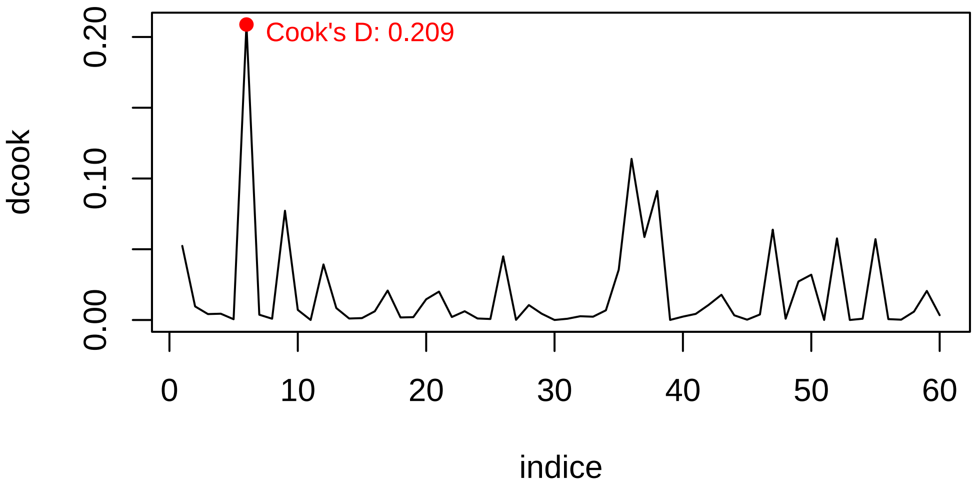
Com base apenas na avaliação da multicolinearidade, o modelo 1 pode ser considerado um pouco mais robusto em relação a esse aspecto, apresentando uma média (2.202784) de VIF (Fator de Inflação da Variância) inferior em comparação com o modelo 2 (3.021946). Além disso, os resultados dos testes lineares 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 (H0) 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, optamos por escolher o modelo 1 como a abordagem mais parcimoniosa para explicar a variabilidade no tempo de internação.


t_internacao
r_infeccao
regiaoW
ilidades_servic
regiaoS
enfermeiros
Corr:
0.627***
Corr:
−0.347**
Corr:
0.533***
Corr:
−0.295*
Corr:
0.404**
Corr:
0.042
Corr:
0.411**
Corr:
−0.286*
Corr:
0.409**
Corr:
−0.086
Corr:
−0.257*
Corr:
−0.006
Corr:
−0.212
Corr:
0.801***
Corr:
−0.111


dffits
0.0
0.4
0.8
1.2
0
10
20
30
40
50
60
indice

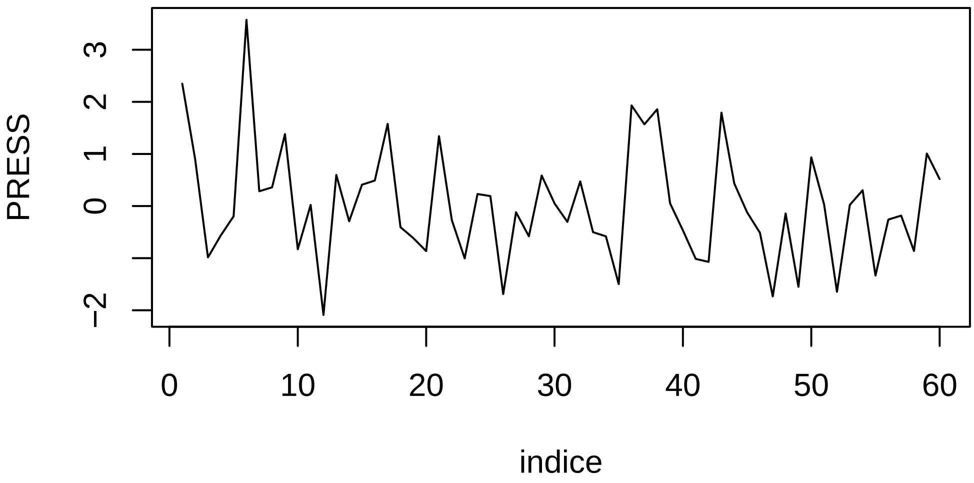

hii
0.05
0.20
0.35
0
10
20
30
40
50
60
indice





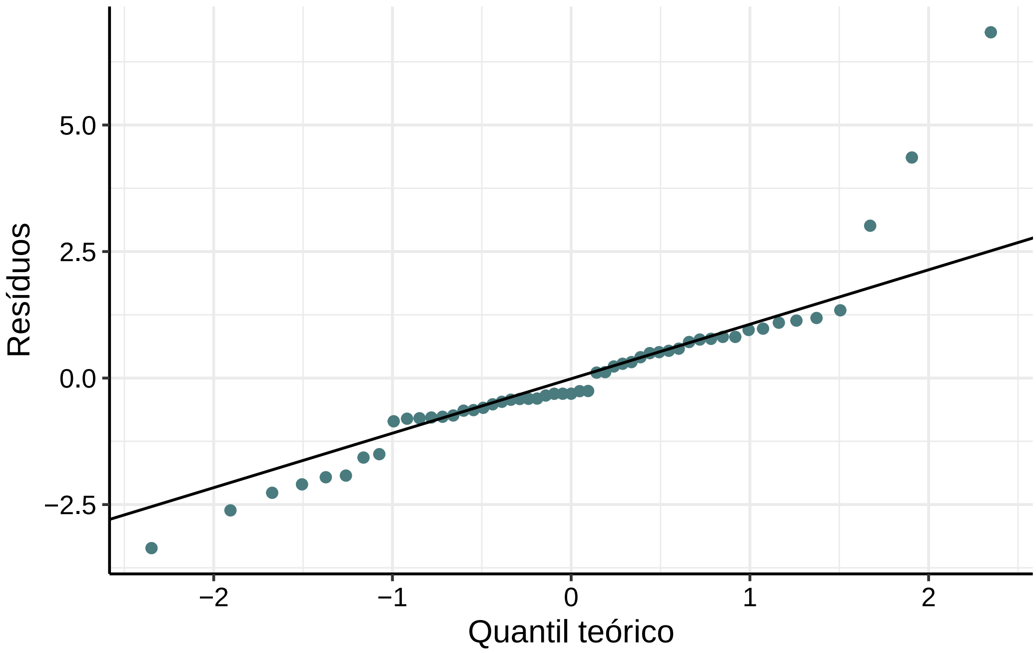

```
          1           2           3           4           5           6
 2.34807615  0.90612034 -0.98462031 -0.56422560 -0.19641850  3.57496127
          7           8           9          10          11          12
 0.28396955  0.35934974  1.38141275 -0.82866642  0.02243118 -2.09025439
         13          14          15          16          17          18
 0.59761975 -0.29169226  0.41009280  0.48952197  1.57750595 -0.40555957
         19          20          21          22          23          24
-0.61472651 -0.86465757  1.34274981 -0.27089858 -1.00401872  0.23154099
         25          26          27          28          29          30
 0.19085623 -1.68957138 -0.11885879 -0.58044079  0.58563618  0.05098920
         31          32          33          34          35          36
-0.30381958  0.47354480 -0.50153228 -0.58067362 -1.49843380  1.93126528
         37          38          39          40          41          42
 1.56837465  1.85771488  0.05193752 -0.46691543 -1.01450886 -1.07130236
         43          44          45          46          47          48
 1.79206817  0.43507041 -0.12086555 -0.51266792 -1.73357282 -0.14220276
         49          50          51          52          53          54
-1.54982927  0.93501630  0.02881104 -1.64587801  0.02178569  0.30346421
         55          56          57          58          59          60
-1.33339361 -0.26014179 -0.18318594 -0.86243331  1.00809668  0.51729884
```

UnB

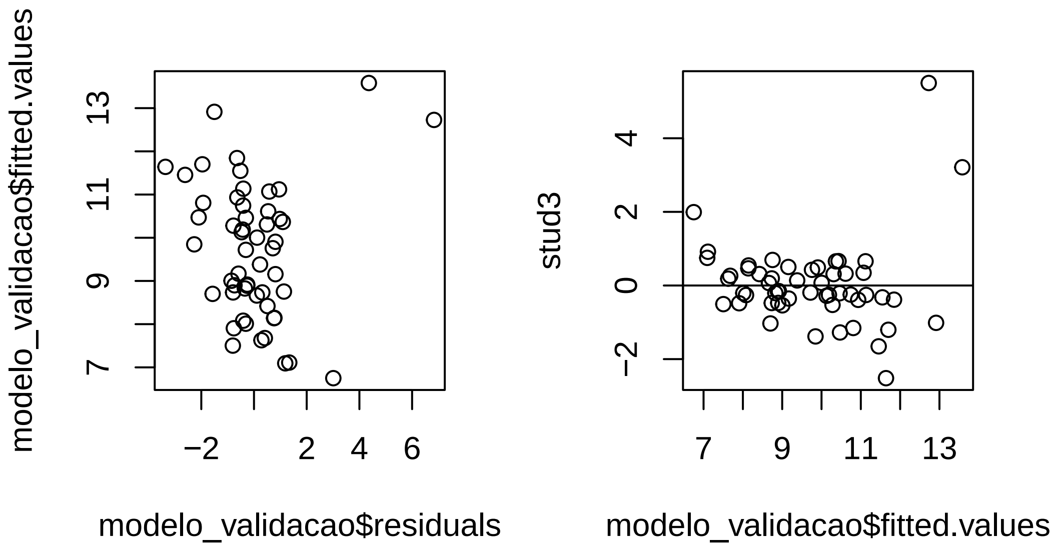



```
Call:
lm(formula = t_internacao ~ idade + r_infeccao + regiaoW + facilidades_servicos +
    regiaoS * enfermeiros, data = valid)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3605 -0.7397 -0.3095  0.7123  6.8329

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.439635   3.085988  -0.142 0.887351
idade                 0.142892   0.054958   2.600 0.012565 *
r_infeccao            0.724047   0.197974   3.657 0.000665 ***
regiaoW              -2.048333   0.718759  -2.850 0.006577 **
facilidades_servicos -0.033250   0.026747  -1.243 0.220260
regiaoS               0.116508   1.000129   0.116 0.907780
enfermeiros           0.009362   0.003807   2.459 0.017834 *
regiaoS:enfermeiros  -0.006750   0.005178  -1.304 0.199007
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.722 on 45 degrees of freedom
Multiple R-squared:  0.4871,    Adjusted R-squared:  0.4074
F-statistic: 6.106 on 7 and 45 DF,  p-value: 0.00004816


Call:
   aov(formula = modelo_validacao)

Terms:
                   idade r_infeccao   regiaoW facilidades_servicos   regiaoS
Sum of Squares    9.61694   64.18955  27.83662              0.20923   6.11207
Deg. of Freedom         1          1         1                    1         1
                enfermeiros regiaoS:enfermeiros Residuals
Sum of Squares     13.75793             5.03940 133.44933
Deg. of Freedom           1                   1        45

Residual standard error: 1.722075
Estimated effects may be unbalanced
```

```
Analysis of Variance Table

Response: t_internacao
                     Df  Sum Sq Mean Sq F value     Pr(>F)    
idade                 1   9.617   9.617  3.2429   0.078438 .  
r_infeccao            1  64.190  64.190 21.6451 0.00002898 ***
regiaoW               1  27.837  27.837  9.3867   0.003685 ** 
facilidades_servicos  1   0.209   0.209  0.0706   0.791747    
regiaoS               1   6.112   6.112  2.0610   0.158023    
enfermeiros           1  13.758  13.758  4.6393   0.036641 *  
regiaoS:enfermeiros   1   5.039   5.039  1.6993   0.199007    
Residuals            45 133.449   2.966                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Analysis of Variance Table

Response: t_internacao
                     Df Sum Sq Mean Sq F value         Pr(>F)    
idade                 1  5.204   5.204  5.1004       0.028143 *  
r_infeccao            1 55.976  55.976 54.8596 0.000000001112 ***
regiaoW               1 19.005  19.005 18.6264 0.000071456978 ***
facilidades_servicos  1  8.848   8.848  8.6720       0.004823 ** 
regiaoS               1  5.121   5.121  5.0187       0.029374 *  
enfermeiros           1  0.360   0.360  0.3524       0.555309    
regiaoS:enfermeiros   1  1.182   1.182  1.1581       0.286820    
Residuals            52 53.058   1.020                           
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


	Shapiro-Wilk normality test

data:  modelo_validacao$residuals
W = 0.85983, p-value = 0.00001781
```




```
    studentized Breusch-Pagan test

data:  modelo_validacao
BP = 12.382, df = 7, p-value = 0.08867
```




```
[1] 6.242633
```

PIPIPIPÓPÓPÓ.......

PIPIPIPÓPÓPÓ.......

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.