Trabalho Final

Aline de Almeida Ramos e Juliana Magalhães Rosa

05/05/2022

Introdução

Este trabalho aborda aspectos relacionados à duração de internações em hospitais dos Estados Unidos.
Os dados analisados são compostos de uma amostra aleatória de 113 hospitais dos 338 que participaram da pesquisa entre 1975 e 1976.

Foram estudadas as seguintes variáveis:
- Número de Identificação (ID): variável qualitativa nominal;
- Duração da Internação (X1): variável quantitativa contínua;
- Idade (X2): variável quantitativa contínua;
- Risco de Infecção (X3): variável quantitativa contínua;
- Proporção de Culturas de Rotina (X4): variável quantitativa contínua;
- Proporção de Raio-X de Tórax de Rotina (X5): variável quantitativa contínua;
- Número de leitos (X6): variável quantitativa discreta;
- Filiação a Escola de Medicina (X7): variável qualitativa nominal;
- Região (X8): variável qualitativa nominal;
- Média diária de pacientes (X9): variável quantitativa discreta;
- Número de enfermeiro(s) (X10): variável quantitativa discreta;
- Facilidades e serviços disponíveis (X11): variável quantitativa contínua.

Objetivos

O intuito deste trabalho é investigar duas hipóteses de pesquisa:
- Hipótese 1: O número de enfermeira(o)s está relacionado às instalações e serviços disponíveis através de um modelo de segunda ordem. Suspeita-se também que varie segundo a região.
- Hipótese 2: A duração da internação está associada a características do paciente, do seu tratamento e do hospital.
Em relação à segunda hipótese, deseja-se verificar quais seriam essas características e também quantificar seu grau de associação com o tempo de internação.

Metodologia

O método utilizado para o desenvolvimento dessas análises é o de Regressão Linear Múltipla com as seguintes etapas de trabalho:
- Preparação dos dados;
- Seleção de variáveis (quando necessário);
- Análise exploratória;
- Seleção do modelo inicial;

- Diagnóstico do modelo;
- Medidas corretivas (se necessário);
- Validação do modelo;
- Modelo final.

A execução dessas etapas foi feita por meio do Software R.

Resultados

Hipótese 1

O número de enfermeira(o)s está relacionado às instalações e serviços disponíveis através de um modelo de segunda ordem. Suspeita-se também que varie segundo a região.

Preparação dos dados

Os dados utilizados estavam em formato xlsx e continham linhas vazias. Sendo assim, após a leitura desses dados no R, foi necessária a exclusão dessas linhas extras para obter apenas as 113 entradas que compõem a amostra.

```
## tibble [113 x 12] (S3: tbl_df/tbl/data.frame)
##  $ ID : num [1:113] 1 2 3 4 5 6 7 8 9 10 ...
##  $ X1 : num [1:113] 7.13 8.82 8.34 8.95 11.2 ...
##  $ X2 : num [1:113] 55.7 58.2 56.9 53.7 56.5 50.9 57.8 45.7 48.2 56.3 ...
##  $ X3 : num [1:113] 4.1 1.6 2.7 5.6 5.7 5.1 4.6 5.4 4.3 6.3 ...
##  $ X4 : num [1:113] 9 3.8 8.1 18.9 34.5 21.9 16.7 60.5 24.4 29.6 ...
##  $ X5 : num [1:113] 39.6 51.7 74 122.8 88.9 ...
##  $ X6 : num [1:113] 279 80 107 147 180 150 186 640 182 85 ...
##  $ X7 : num [1:113] 2 2 2 2 2 2 2 1 2 2 ...
##  $ X8 : num [1:113] 4 2 3 4 1 2 3 2 3 1 ...
##  $ X9 : num [1:113] 207 51 82 53 134 147 151 399 130 59 ...
##  $ X10: num [1:113] 241 52 54 148 151 106 129 360 118 66 ...
##  $ X11: num [1:113] 60 40 20 40 40 40 40 60 40 40 ...
```

Também foi necessária a modificação dos tipos de variáveis, uma vez que todas as colunas foram lidas como numéricas (decimais).
Vale comentar que as informações relativas à Região e à Filiação à Escola de Medicina foram codificadas, já nos dados originais. As regiões NE, NC, S e W são representadas por 1, 2, 3 e 4, respectivamente. Já a filiação tem 1 para as respostas positivas (sim) e 2 para as respostas negativas (não).

Para as variáveis qualitativas, existe a necessidade de criar variáveis *dummy* antes da construção do modelo. Esse tipo de variável tem como possíveis valores apenas o 0 e o 1.
Como a filiação à escola de medicina já é uma variável binária, bastou converter sua codifição para 0 (sim) e 1 (não).
Já para a região, foram criadas 3 variáveis *dummy* para NE, NC e S.

Para a construção dos modelos de regressão, foi extraída uma amostra de 57 hospitais. Sendo os 56 hospitais remanescentes deixados em uma amostra de validação que será utilizada nas etapas finais do trabalho.

Análise Exploratória

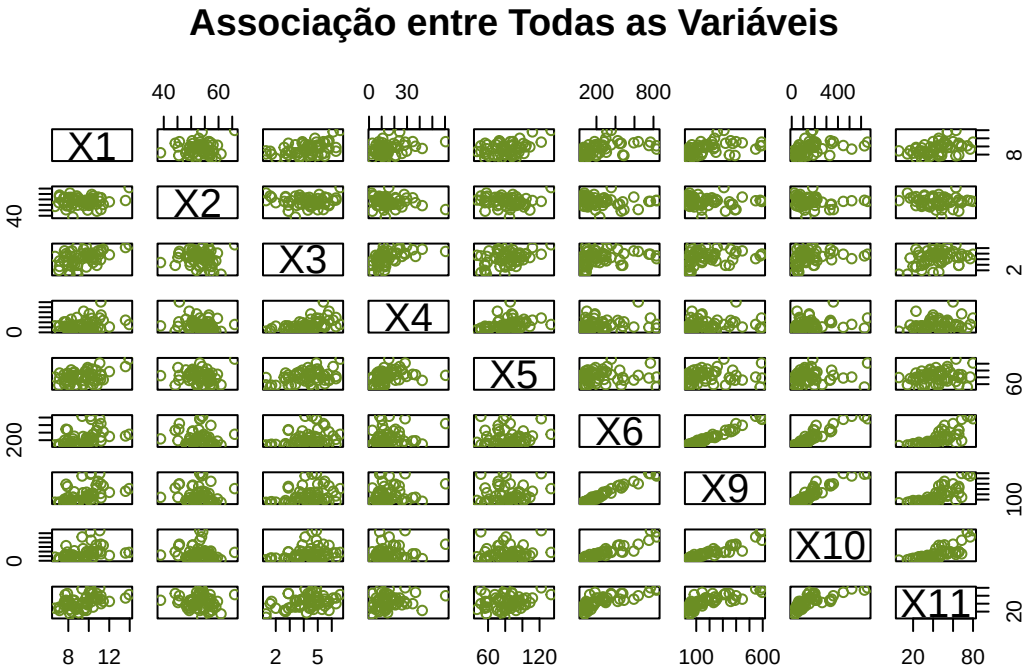


Table 1: Coeficientes de Correlação entre Todas as Variáveis

	X1	X2	X3	X4	X5	X6	X9	X10	X11
X1	1.0000	0.1326	0.4632	0.2981	0.4053	0.4409	0.4859	0.3869	0.4472
X2	0.1326	1.0000	-0.0442	-0.1962	-0.0472	-0.1342	-0.1205	-0.1406	-0.0939
X3	0.4632	-0.0442	1.0000	0.5341	0.3951	0.2853	0.3182	0.3011	0.3620
X4	0.2981	-0.1962	0.5341	1.0000	0.3458	0.2079	0.1933	0.2299	0.2320
X5	0.4053	-0.0472	0.3951	0.3458	1.0000	0.1011	0.1572	0.1230	0.2127
X6	0.4409	-0.1342	0.2853	0.2079	0.1011	1.0000	0.9871	0.9382	0.7860
X9	0.4859	-0.1205	0.3182	0.1933	0.1572	0.9871	1.0000	0.9381	0.7809
X10	0.3869	-0.1406	0.3011	0.2299	0.1230	0.9382	0.9381	1.0000	0.7626
X11	0.4472	-0.0939	0.3620	0.2320	0.2127	0.7860	0.7809	0.7626	1.0000

Observando os coeficientes de correlação entre as variáveis quantitativas, percebe-se que as únicas correlações fortes encontradas foram entre as variáveis relativas ao número de leitos (X6), à média diária de pacientes (X9), ao número de enfermeiro(s) (X10) e aos serviços disponíveis (X11). Todas essas características, pareadas, estão fortemente associadas.
A mesma conclusão pode ser retirada a partir dos gráficos de dispersão apresentados.

Table 2: Frequências Relativas para a Filiação à Escola de Medicina

Filiação	Frequência
sim	0.1578947
não	0.8421053

De acordo com a tabela acima, 15,79% dos hospitais estão filiados a escolas de medicina, enquanto 84,21% não estão.

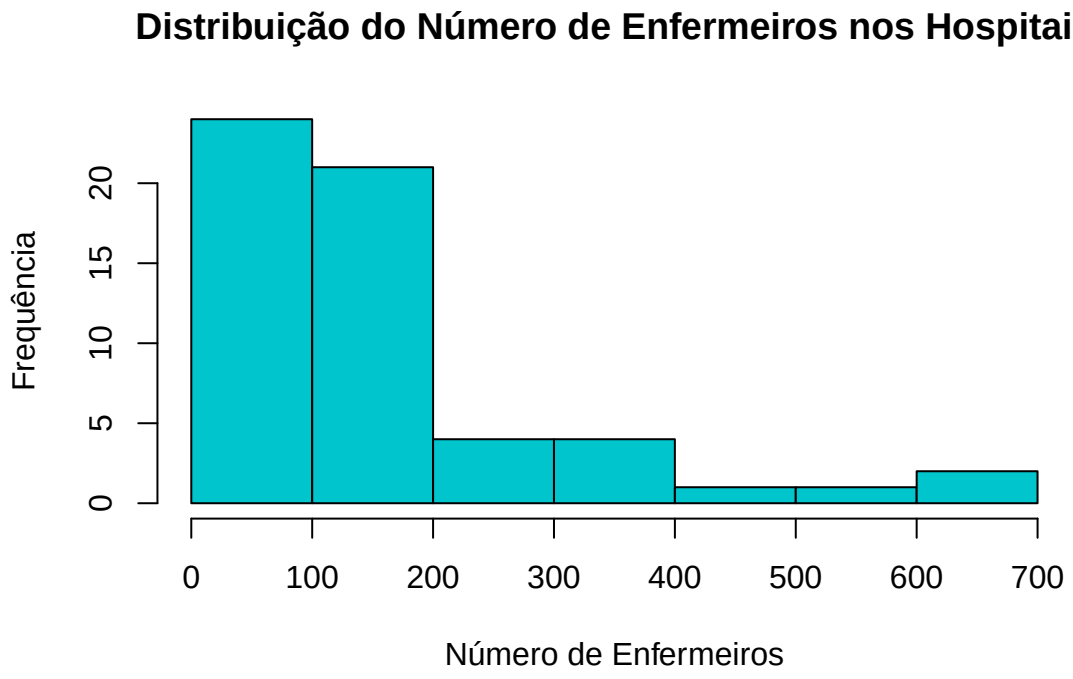
Table 3: Frequências Relativas para as Regiões

Região	Frequência
NE	0.2456140
NC	0.3333333
S	0.2456140
W	0.1754386

Em relação a nossa amostra, a região com maior porcentagem de hospitais é a NC (33,33%), seguida das regiões NE e S, ambas com 24,56% e, por fim, a região W, que contém 17,54% dos hospitais da amostra.

A seguir, é apresentada uma análise da variável resposta (X10 - Número de Enfermeiros).

```
##
##   The decimal point is 2 digit(s) to the right of the |
##
##   0 | 12444455555666677778899
##   1 | 011112222455557789
##   2 | 00000126
##   3 | 3456
##   4 | 5
##   5 | 2
##   6 | 36
```



Visualizando a distribuição do número de enfermeiros, percebe-se que não há normalidade desses dados, e sim uma assimetria à direita. O mais comum é que hajam até 200 enfermeiros por hospital.

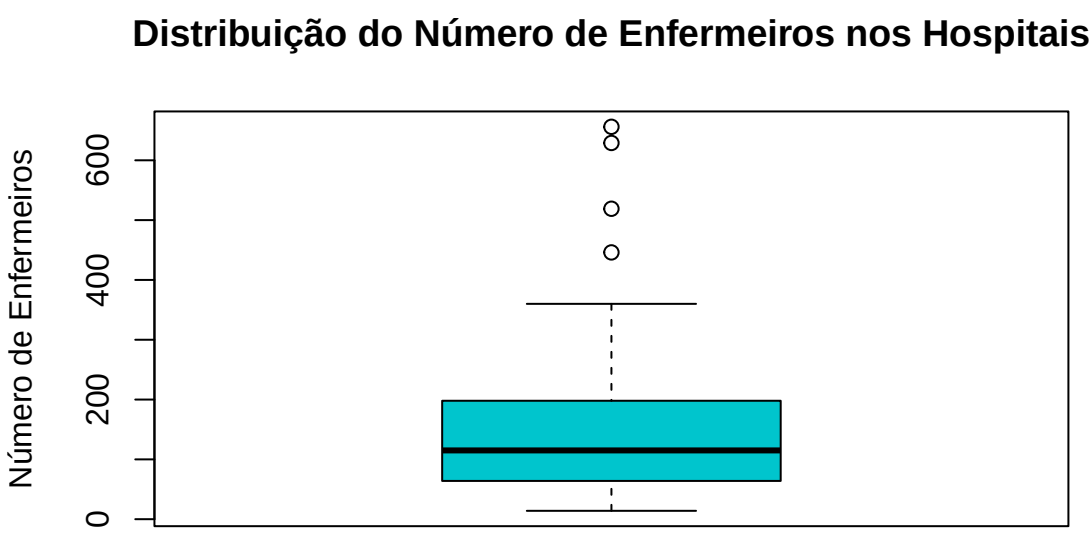
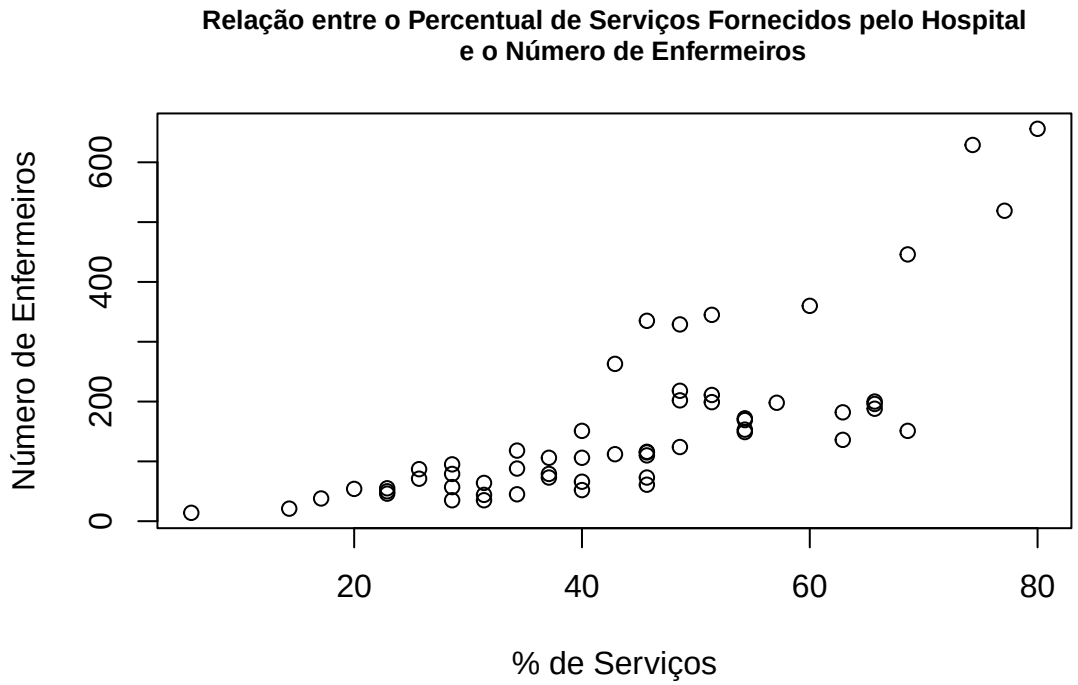


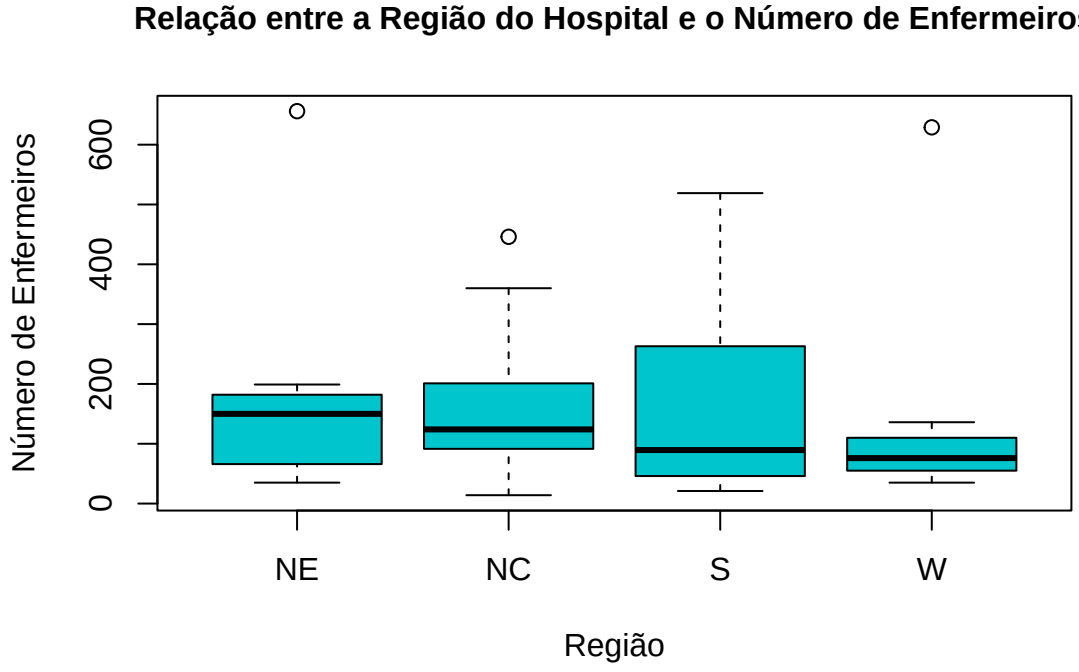
Table 4: Medidas Descritivas para o Número de Enfermeiros

media	q1	mediana	q3	q4	amplitude	iiq	variancia	dp	cv	Ap	Aq	k
158.7018	64	115	198	656	642	134	19930	141.1736	0.8896	0.9287	0.2388	0.2276

Pelo boxplot, nota-se que existem alguns valores extremos (*outliers*) e que a mediana da distribuição é próxima de 100 enfermeiros. A assimetria à direita observada no histograma e no ramo e folhas é confirmada tanto pelo boxplot como pelos coeficientes de assimetria de Pearson (Ap) e quartil de assimetria (Aq), que são ambos positivos.
Além disso, a mediana tem um valor de 115 e a média de 158.7. O coeficiente de variação de 0.89 mostra que existe bastante variabilidade nos dados.

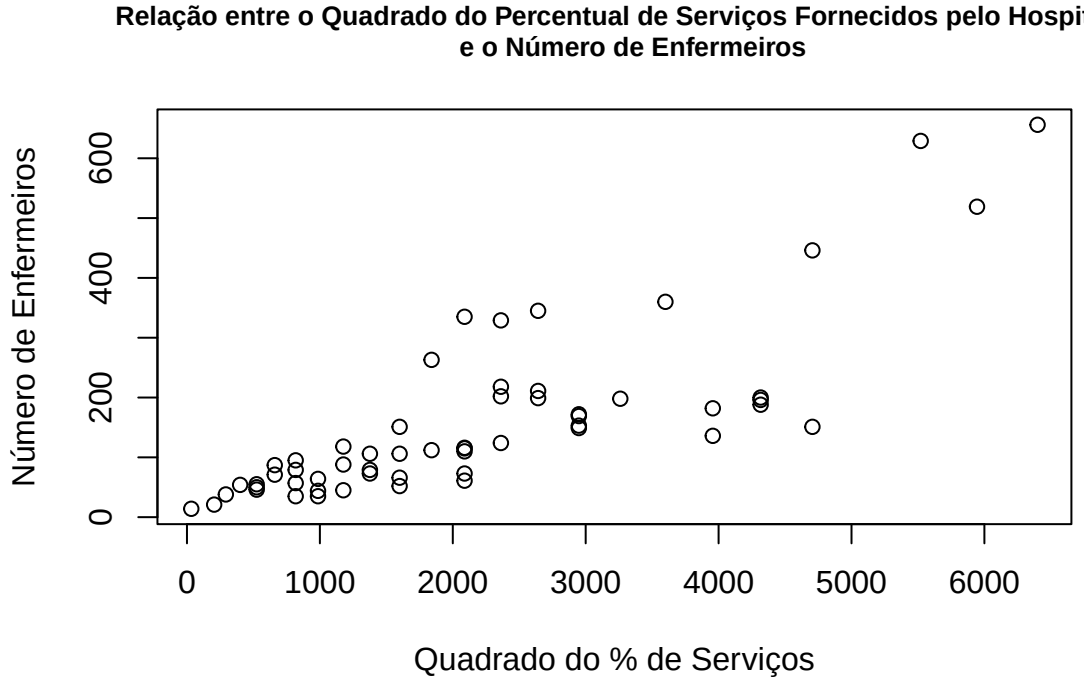


Com uma primeira visualização da relação entre o número de enfermeiros e o percentual de serviços fornecidos, percebe-se que existe uma correlação, mas que há um indício de uma curva nos pontos. Isso já é uma evidência a favor da hipótese 1, a qual sugere o modelo de segunda ordem para relacionar essas duas variáveis.



Analisando agora a distribuição do número de enfermeiros nos hospitais por região, nota-se que a variabilidade

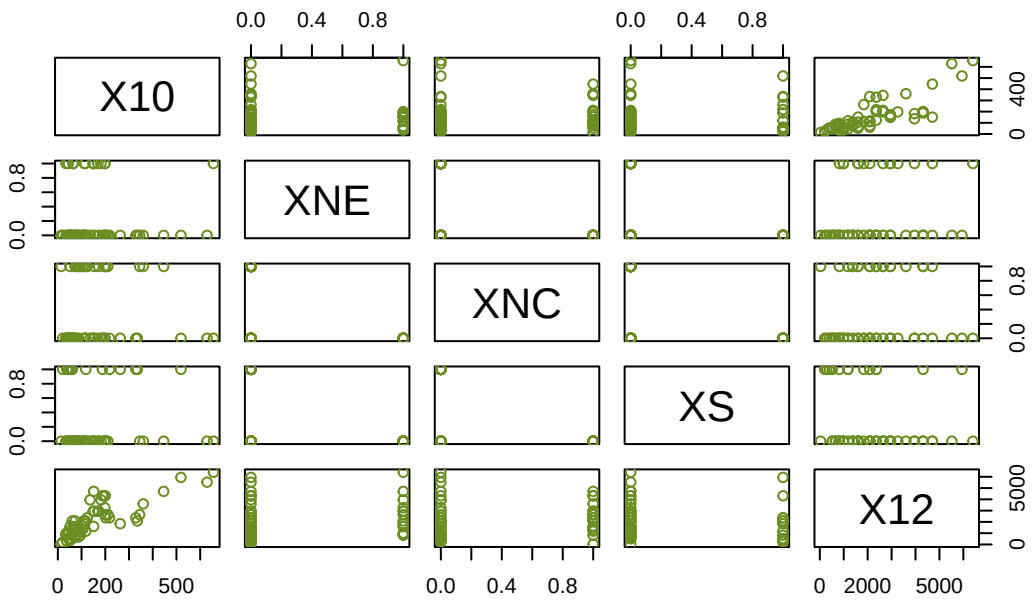
é maior na região S e menor na W.
Além disso, a região S é a única que não apresenta *outliers*.
Também é possível visualizar que a região NE apresenta assimetria à esquerda, o que contraria o padrão percebido nas demais regiões. Isso é perceptível pela distância entre a mediana e o primeiro quartil.



Utilizando a variável X12, que representa $X11^2$, vemos maior linearidade na correlação com a variável resposta (X10). Dessa forma, será utilizada no modelo a variável X12, como alternativa a um modelo de segunda ordem para X11.

Table 5: Coeficientes de Correlação entre as Variáveis do Modelo

	X10	XNE	XNC	XS	X12
X10	1.0000	0.0201	0.0326	0.0204	0.8091
XNE	0.0201	1.0000	-0.4035	-0.3256	0.1683
XNC	0.0326	-0.4035	1.0000	-0.4035	0.0868
XS	0.0204	-0.3256	-0.4035	1.0000	-0.1352
X12	0.8091	0.1683	0.0868	-0.1352	1.0000



Tendo selecionado as variáveis *dummy* XNE, XNC e XS (para regiões) e a variável X12 como explicativas e a X10 como resposta, já é possível propor o modelo inicial.

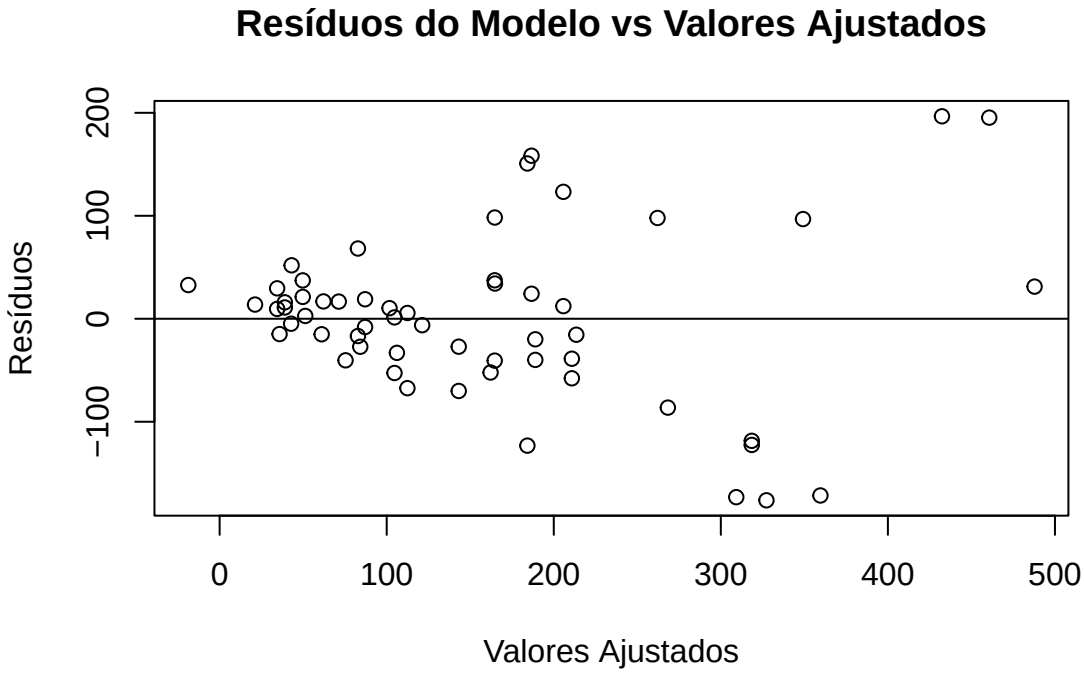
Seleção do Modelo Inicial

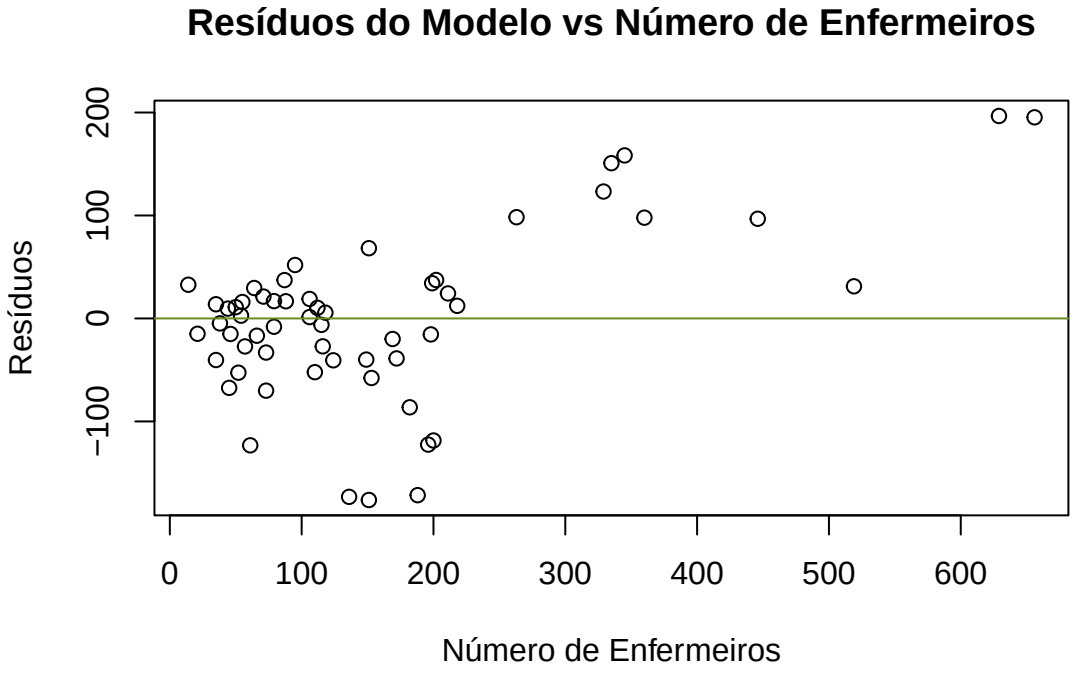
```
## Analysis of Variance Table
##
## Response: amostra$X10
##              Df Sum Sq Mean Sq  F value    Pr(>F)
## amostra$X12   1 730698  730698 106.5662 3.427e-14 ***
## amostra$XNE   1  15472   15472   2.2565    0.1391
## amostra$XNC   1  10529   10529   1.5356    0.2208
## amostra$XS    1   2830    2830   0.4127    0.5234
## Residuals    52 356551    6857
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Call:
## lm(formula = amostra$X10 ~ amostra$X12 + amostra$XNE + amostra$XNC +
##     amostra$XS)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -176.304  -39.940    2.725   31.217  196.624
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

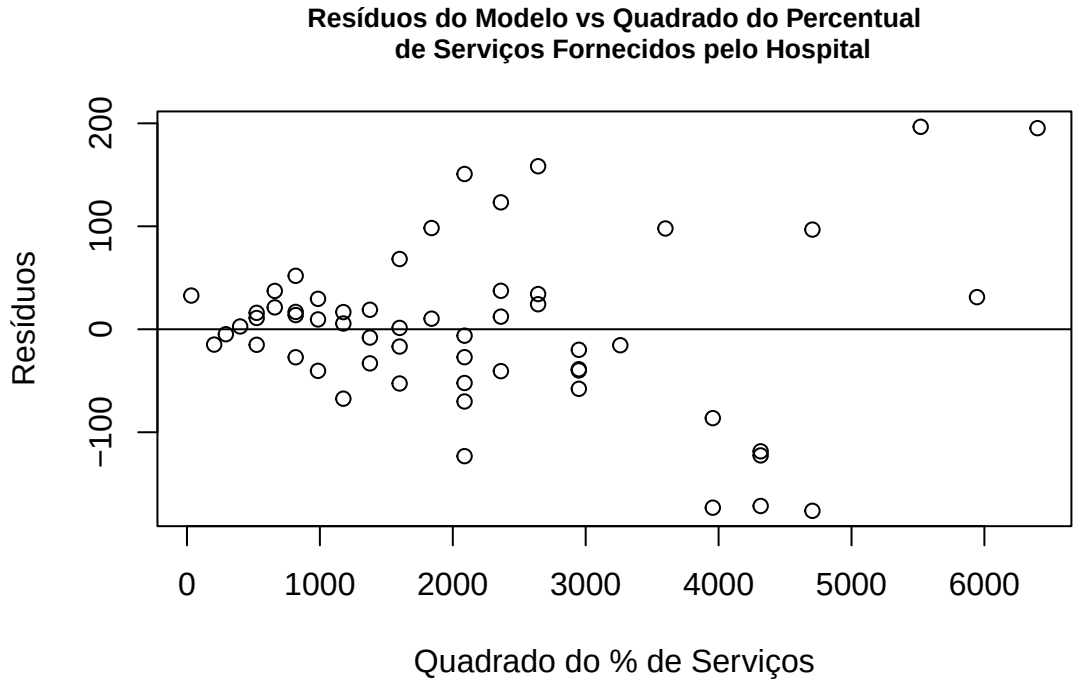
```
## (Intercept)  -2.249666  29.179032  -0.077    0.939
## amostra$X12   0.078730   0.007522  10.467  2.1e-14 ***
## amostra$XNE -40.943982  34.969645  -1.171    0.247
## amostra$XNC -19.059162  32.725683  -0.582    0.563
## amostra$XS   22.032557  34.295946   0.642    0.523
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82.81 on 52 degrees of freedom
## Multiple R-squared:  0.6805, Adjusted R-squared:  0.656
## F-statistic: 27.69 on 4 and 52 DF,  p-value: 2.437e-12
```

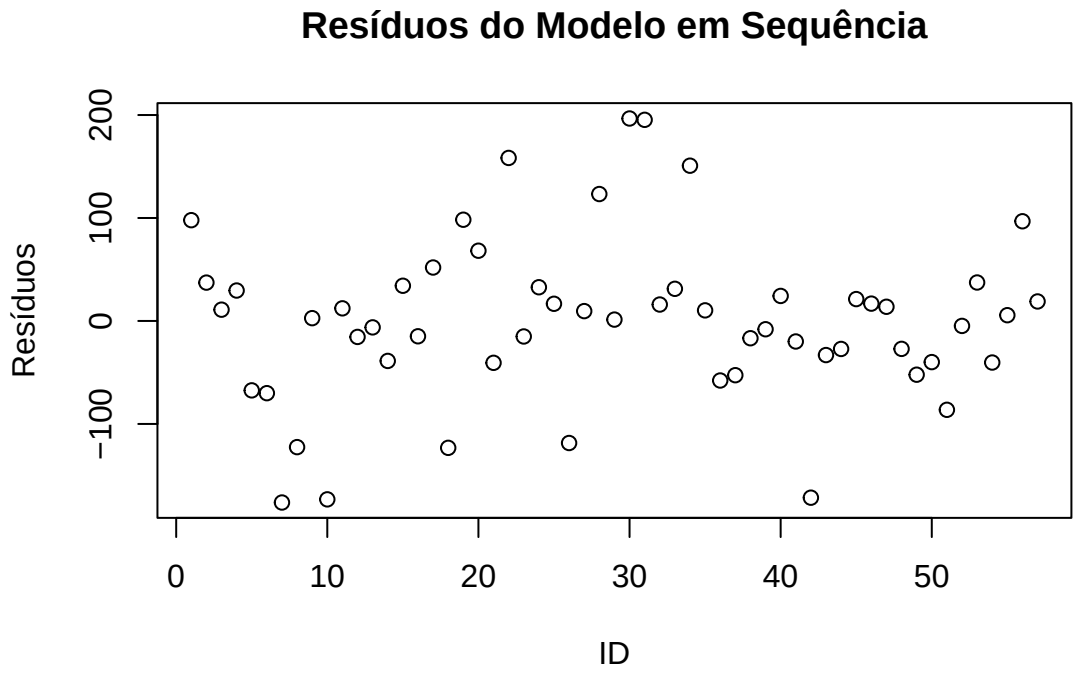
Pela estatística F e seu p-valor menor do que 0,05, conclui-se que, de fato, há regressão.
De acordo com os resultados do modelo, um aumento unitário na variável X12 (quadrado do percentual de serviços fornecidos) resulta em um aumento médio de 0.08 no número de enfermeiros.
Em relação ao número de enfermeiros nos hospitais da região W, há uma diminuição média de aproximadamente 40 enfermeiros na região NE e 20 enfermeiros na região NC. Em contrapartida, ao passar da região W para a região S, há um aumento de 22 enfermeiros em média.

Diagnóstico do Modelo

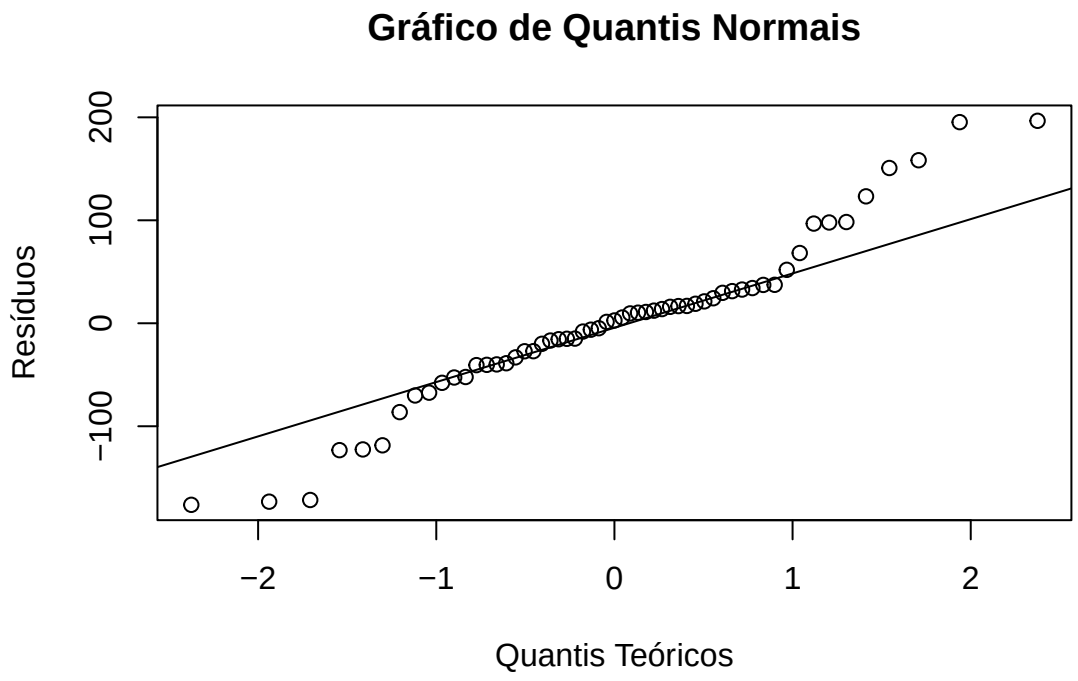








Os gráficos residuais apresentam evidências de heterocedasticidade, por causa da mudança na variabilidade dos resíduos com o aumento dos valores ajustados, ou das variáveis explicativas, etc.



Pelo qqplot, observa-se maiores indícios de não normalidade dos resíduos e, consequentemente, da variável resposta.

```
##
##  studentized Breusch-Pagan test
##
## data:  modcomp
## BP = 29.467, df = 4, p-value = 6.282e-06
```

Realizado o teste de Breusch-Pagan para verificação da homocedasticidade, rejeita-se a hipótese nula, com uma estatística qui-quadrado de 29,47, e portanto, conclui-se que a variância não é constante.

Medidas Corretivas

Devido à heterocedasticidade e à falta de normalidade da distribuição do número de enfermeiros — pressupostos do modelo — foi necessária a adoção de uma medida corretiva. Para isso, a estratégia escolhida foi a da aplicação do logaritmo na variável resposta, técnica que possui o potencial de solucionar ambos os problemas.

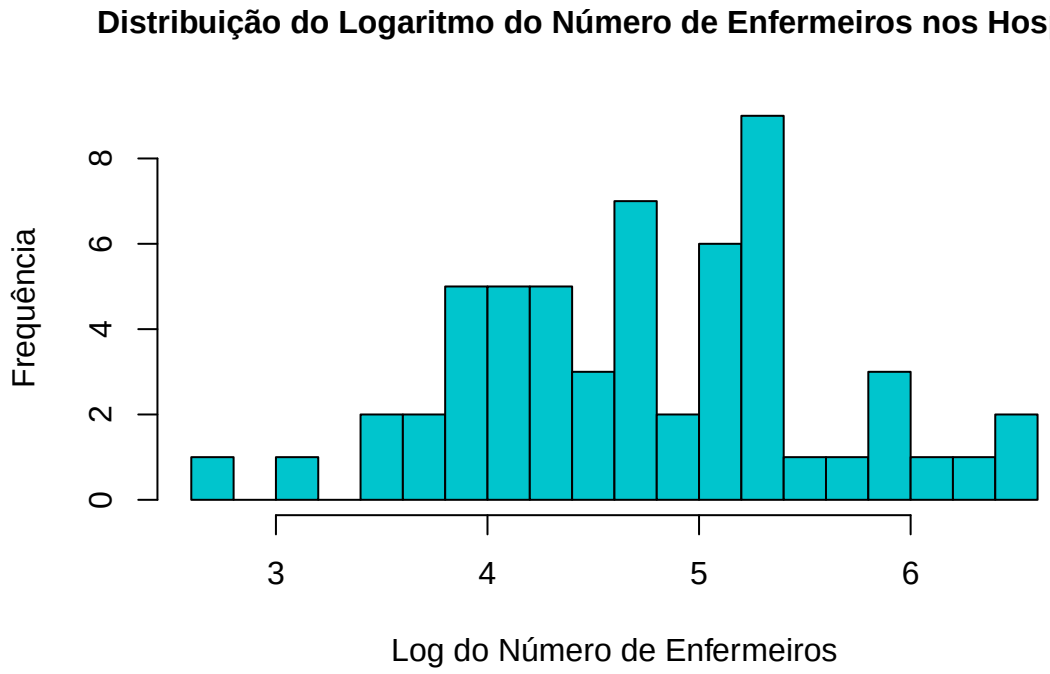


Table 6: Coeficientes de Correlação entre as Variáveis do Modelo Corrigido

	lnX10	XNE	XNC	XS	X12
lnX10	1.0000	0.0609	0.1072	-0.0542	0.8230
XNE	0.0609	1.0000	-0.4035	-0.3256	0.1683
XNC	0.1072	-0.4035	1.0000	-0.4035	0.0868
XS	-0.0542	-0.3256	-0.4035	1.0000	-0.1352
X12	0.8230	0.1683	0.0868	-0.1352	1.0000

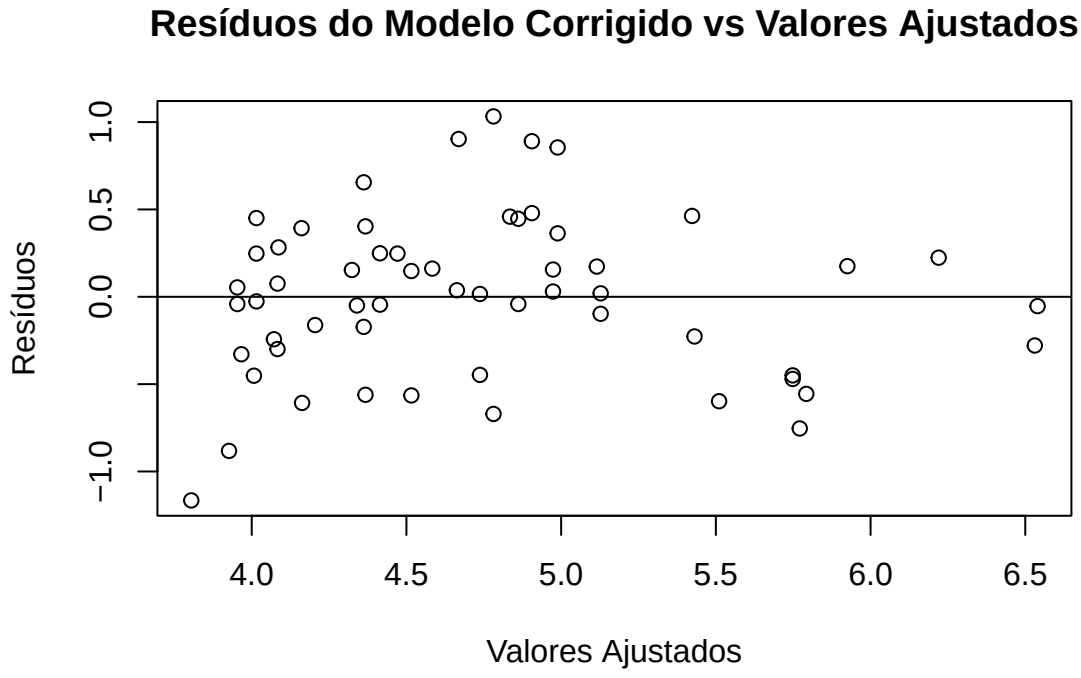
O histograma acima indica que houve, de fato, uma redução na assimetria à direita que a variável resposta apresentava anteriormente.

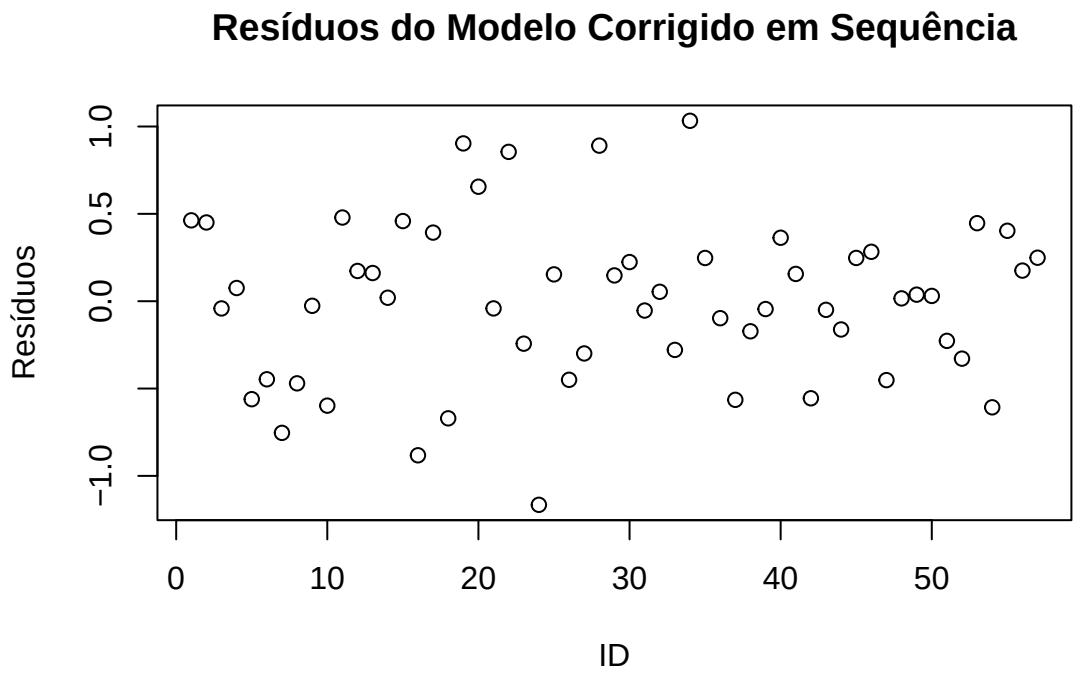
```
##
## Call:
## lm(formula = amostra$lnX10 ~ amostra$XNC + amostra$XNE + amostra$XS +
##     amostra$X12)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16570 -0.29918  0.01978  0.24896  1.03273
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.716e+00  1.688e-01  22.006  < 2e-16 ***
## amostra$XNC  7.449e-02  1.894e-01   0.393    0.696
## amostra$XNE -7.948e-02  2.023e-01  -0.393    0.696
```

```
## amostra$XS   1.183e-01  1.985e-01   0.596    0.554
## amostra$X12  4.537e-04  4.352e-05  10.424 2.43e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4791 on 52 degrees of freedom
## Multiple R-squared:  0.6857, Adjusted R-squared:  0.6615
## F-statistic: 28.36 on 4 and 52 DF,  p-value: 1.606e-12
```

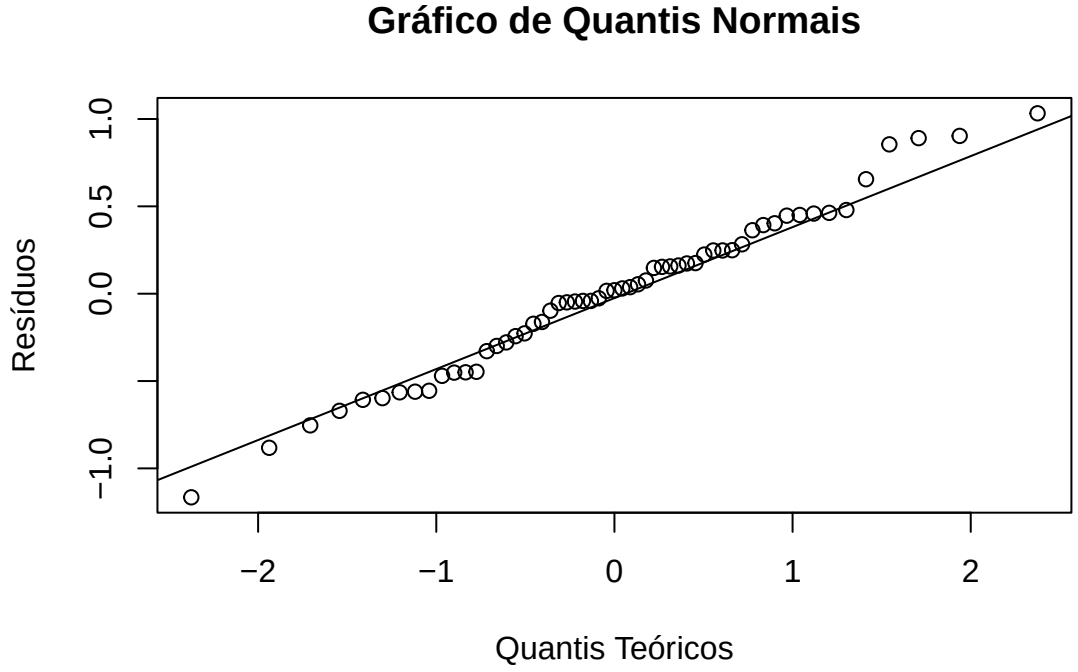
```
## Analysis of Variance Table
##
## Response: amostra$lnX10
##             Df  Sum Sq Mean Sq  F value    Pr(>F)
## amostra$XNC  1  0.4366  0.4366   1.9017    0.1738
## amostra$XNE  1  0.4926  0.4926   2.1456    0.1490
## amostra$XS   1  0.1711  0.1711   0.7452    0.3920
## amostra$X12  1 24.9452 24.9452 108.6547 2.431e-14 ***
## Residuals   52 11.9383  0.2296
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De acordo com os resultados do novo modelo, um aumento unitário na variável X12 (quadrado do percentual de serviços fornecidos) resulta em um aumento médio de 0.00045 no log de X10.
Em relação ao log do número de enfermeiros nos hospitais da região W, há uma diminuição média de aproximadamente 0.079 na região NE. Por outro lado, ao passar da região W para a região NC, há um aumento médio de 0.074 e para a região S, o aumento médio é de 0.12.





Como esperado, os gráficos residuais não mais indicam heterocedasticidade, uma vez que os pontos agora se apresentam espalhados de forma aleatória pelo diagrama.



Observa-se que, com a aplicação do log, há maior adesão à distribuição normal.

Validação do Modelo

Será utilizada a amostra de validação composta pelos 56 hospitais remanescentes.

Table 7: Coeficientes de Correlação entre as Variáveis na Amostra de Validação

	lnX10	XNE	XNC	XS	X12
lnX10	1.0000	0.1738	0.0742	-0.2108	0.8497
XNE	0.1738	1.0000	-0.3175	-0.4820	0.0642
XNC	0.0742	-0.3175	1.0000	-0.4590	0.1180
XS	-0.2108	-0.4820	-0.4590	1.0000	-0.1647
X12	0.8497	0.0642	0.1180	-0.1647	1.0000

Os coeficientes de correlação estão próximos ao que eram na amostra anterior. Permanece a forte correlação entre o quadrado do percentual de serviços fornecidos e o log do número de enfermeiros.

```
##
## Call:
## lm(formula = amostra_validacao$lnX10 ~ amostra_validacao$XNC +
##     amostra_validacao$XNE + amostra_validacao$XS + amostra_validacao$X12)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04357 -0.26192  0.01391  0.30313  0.94470
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.704e+00  2.107e-01  17.580  < 2e-16 ***
## amostra_validacao$XNC  7.176e-03  2.211e-01   0.032    0.974
## amostra_validacao$XNE  2.173e-01  2.183e-01   0.995    0.324
## amostra_validacao$XS  -2.763e-02  2.055e-01  -0.134    0.894
## amostra_validacao$X12  5.940e-04  5.163e-05  11.504 8.78e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4473 on 51 degrees of freedom
## Multiple R-squared:  0.7366, Adjusted R-squared:  0.716
## F-statistic: 35.66 on 4 and 51 DF,  p-value: 3.322e-14
```

Tendo construído o mesmo modelo, mas a partir da amostra de validação, é possível comparar esses resultados aos já obtidos.
Observando os valores das estimativas dos coeficientes de regressão, nota-se que permanecem próximos do zero, com valores similares aos encontrados anteriormente.

MSPR
0.3214

Como segunda etapa de validação, ajustou-se o modelo construído pela amostra de treinamento aos dados da amostra de validação. A partir das predições obtidas, foi calculado o erro quadrático médio que resultou em 0.32, valor próximo ao MSE do modelo (0.23).
Com isso, conclui-se que há um bom ajustamento do modelo formulado.

Modelo Final

Tendo validado o modelo construído, utiliza-se a amostra completa para a formulação do modelo final.

```
##
## Call:
## lm(formula = dados_hosp$lnX10 ~ dados_hosp$XNC + dados_hosp$XNE +
##     dados_hosp$XS + dados_hosp$X12)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16333 -0.27865  0.00723  0.34861  1.27013
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.739e+00  1.394e-01  26.824   <2e-16 ***
## dados_hosp$XNC 4.693e-02  1.524e-01   0.308    0.759
## dados_hosp$XNE 8.623e-02  1.562e-01   0.552    0.582
## dados_hosp$XS  7.358e-02  1.480e-01   0.497    0.620
## dados_hosp$X12 4.968e-04  3.511e-05  14.151   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4946 on 108 degrees of freedom
## Multiple R-squared:  0.6603, Adjusted R-squared:  0.6478
## F-statistic: 52.49 on 4 and 108 DF,  p-value: < 2.2e-16
```

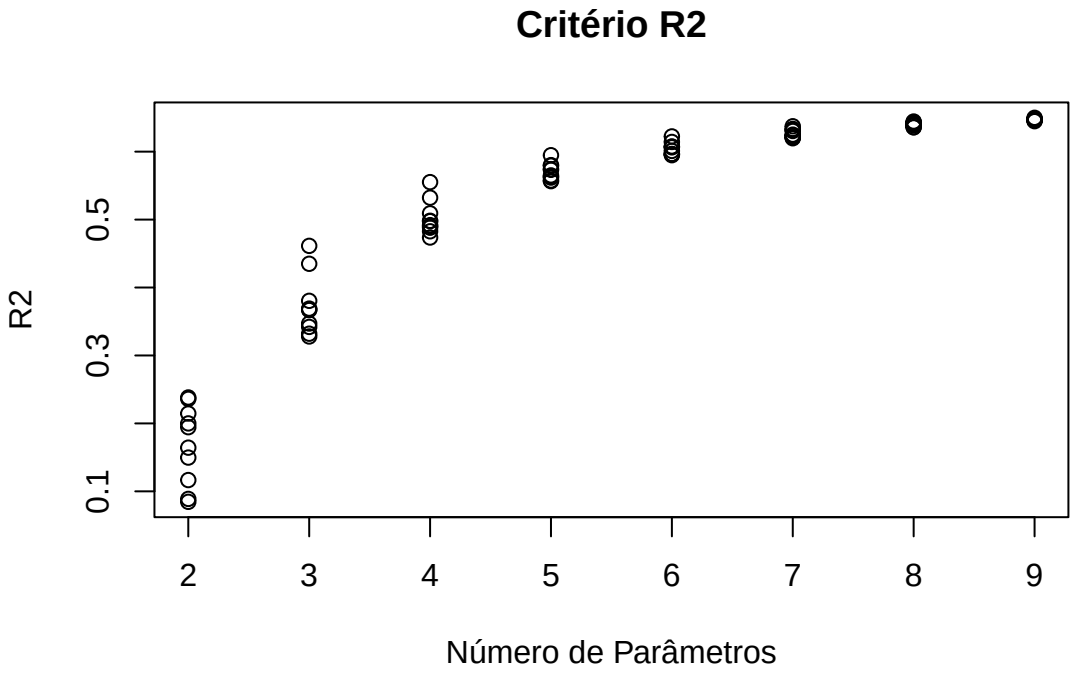
```
## Analysis of Variance Table
##
## Response: dados_hosp$lnX10
##                 Df Sum Sq Mean Sq  F value  Pr(>F)
## dados_hosp$XNC   1  0.462   0.462   1.8880 0.17227
## dados_hosp$XNE   1  1.884   1.884   7.6997 0.00651 **
## dados_hosp$XS    1  0.030   0.030   0.1231 0.72640
## dados_hosp$X12   1 48.998  48.998 200.2579 < 2e-16 ***
## Residuals      108 26.425   0.245
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

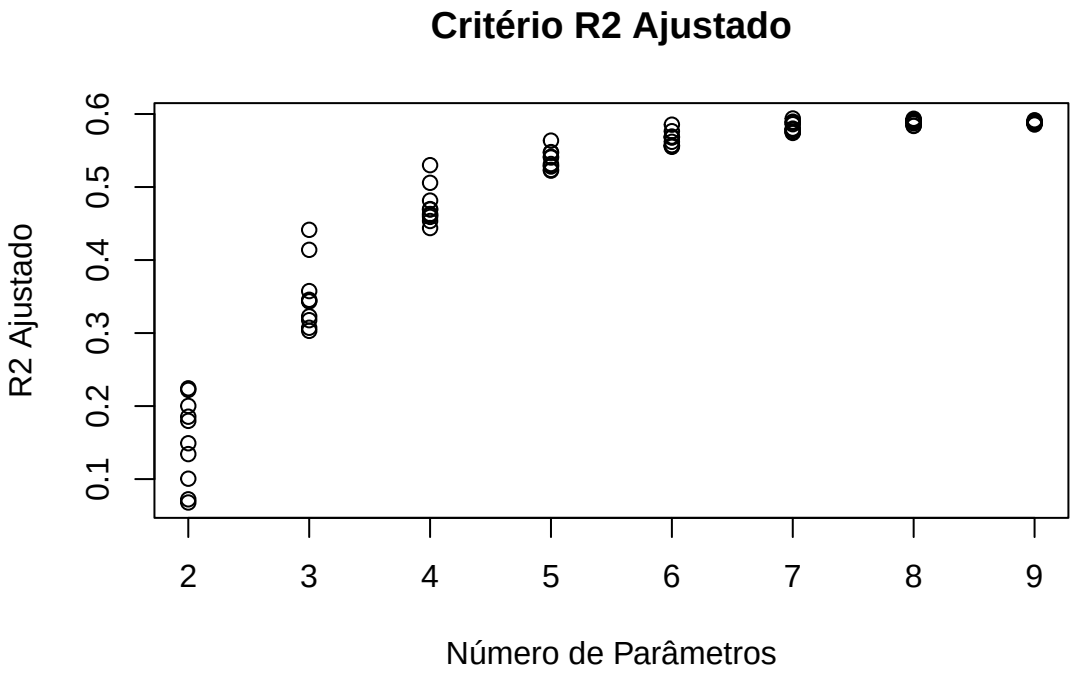
Hipótese 2

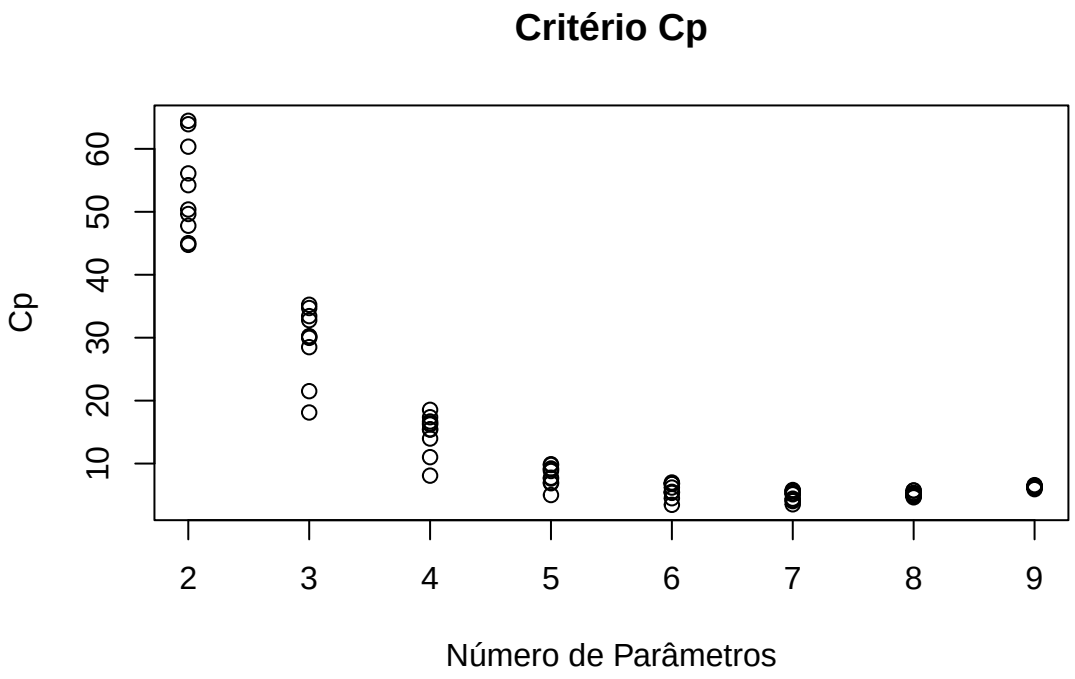
A duração da internação está associada a características do paciente, do seu tratamento e do hospital.

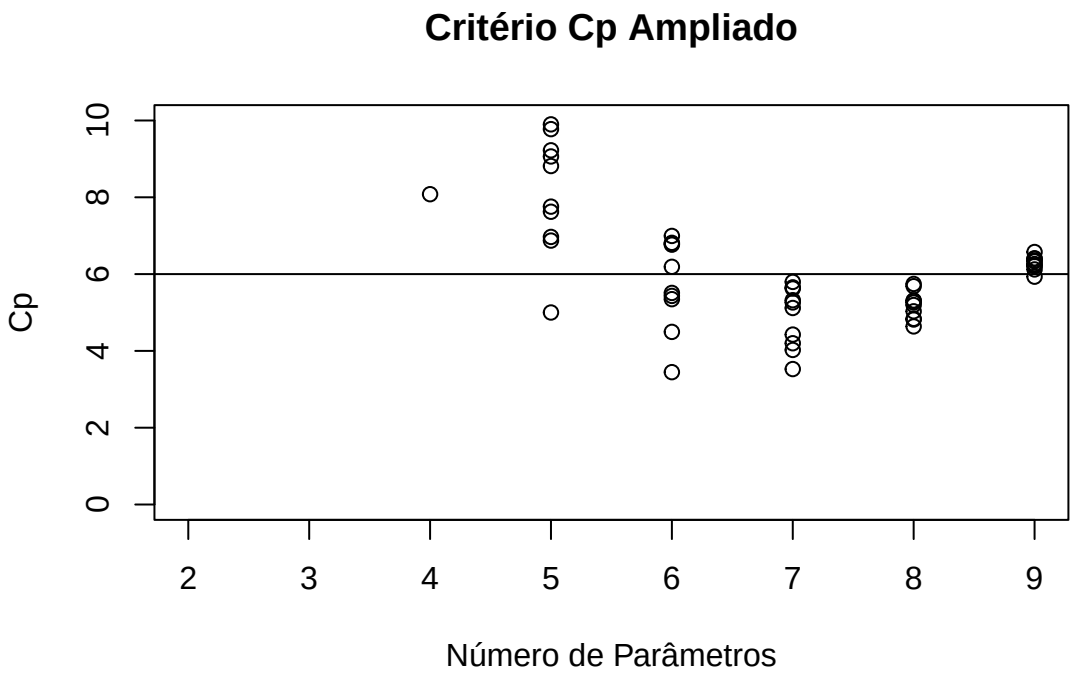
Seleção de Variáveis

A variável resposta a ser investigada é a duração da internação (X1).
Antes de montar o modelo de regressão para o estudo, é necessária a etapa de seleção de variáveis, para que possam ser escolhidas as características que influenciam mais na resposta.


Critério R2
R2
0.1
0.3
0.5
2
3
4
5
6
7
8
9
Número de Parâmetros


Critério R2 Ajustado
R2 Ajustado
0.1
0.2
0.3
0.4
0.5
0.6
2
3
4
5
6
7
8
9
Número de Parâmetros





Modelo com 6 parâmetros

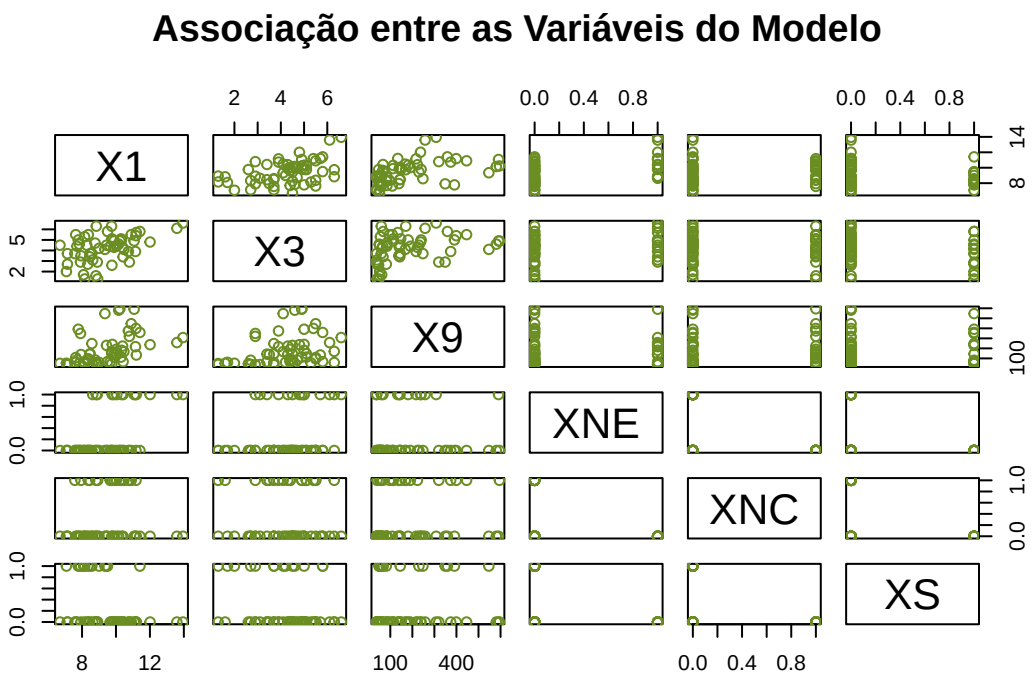
A partir dos gráficos para o R^2 e o $R^2 ajustado$, percebe-se que modelos com 6 parâmetros seriam uma escolha plausível, já que os valores para essas duas medidas são altos, mas a quantidade de variáveis nos

modelos não seria tão grande.
Analisando o gráfico do Cp, chega-se à mesma conclusão. Além disso, observando mais de perto os modelos de 6 parâmetros, localiza-se o que possui o Cp mais próximo do número de parâmetros. Esse modelo inclui as variáveis explicativas X3, X9, XNC, XNE e XS.

Análise Exploratória

Table 9: Coeficientes de Correlação entre as Variáveis do Modelo

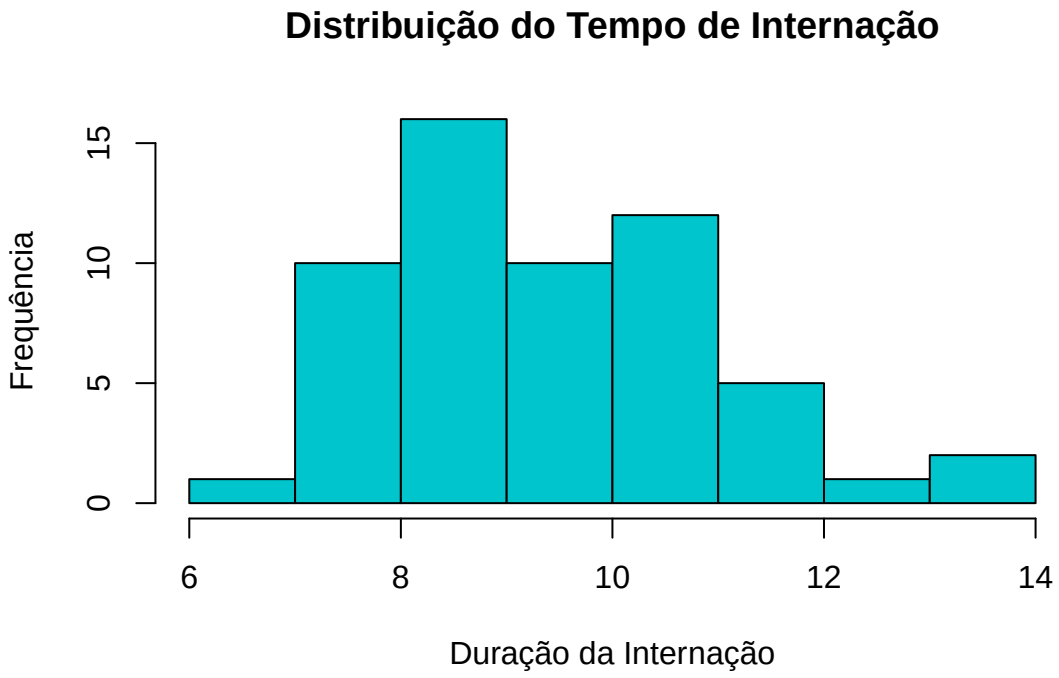
	X1	X3	X9	XNE	XNC	XS
X1	1.0000	0.4632	0.4859	0.4881	0.1321	-0.2908
X3	0.4632	1.0000	0.3182	0.2346	-0.0059	-0.2341
X9	0.4859	0.3182	1.0000	0.0281	0.0961	0.0384
XNE	0.4881	0.2346	0.0281	1.0000	-0.4035	-0.3256
XNC	0.1321	-0.0059	0.0961	-0.4035	1.0000	-0.4035
XS	-0.2908	-0.2341	0.0384	-0.3256	-0.4035	1.0000



Em geral, as variáveis explicativas selecionadas parecem ter uma correlação moderada com a duração de internação.
Olhando para a distribuição da variável resposta:

```
##
##   The decimal point is at the |
##
##    6 | 7
##    7 | 1116678899
##    8 | 122334556688899
```

```
##     9 | 04457788899
##    10 | 012233445789
##    11 | 12224
##    12 | 0
##    13 | 6
##    14 | 0
```



Pelos gráficos acima, é possível notar que a distribuição da variável resposta possui comportamento semelhante ao de uma distribuição normal.

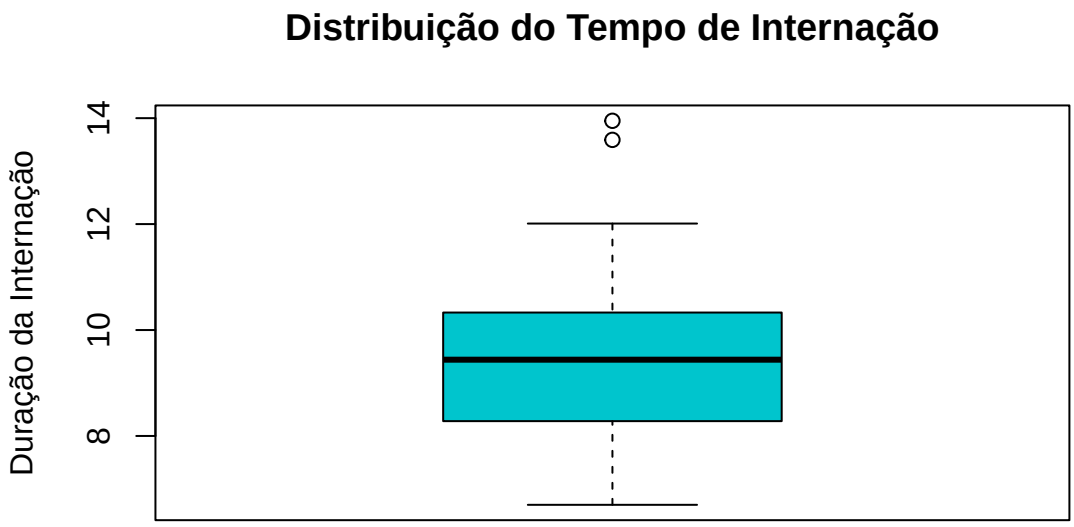
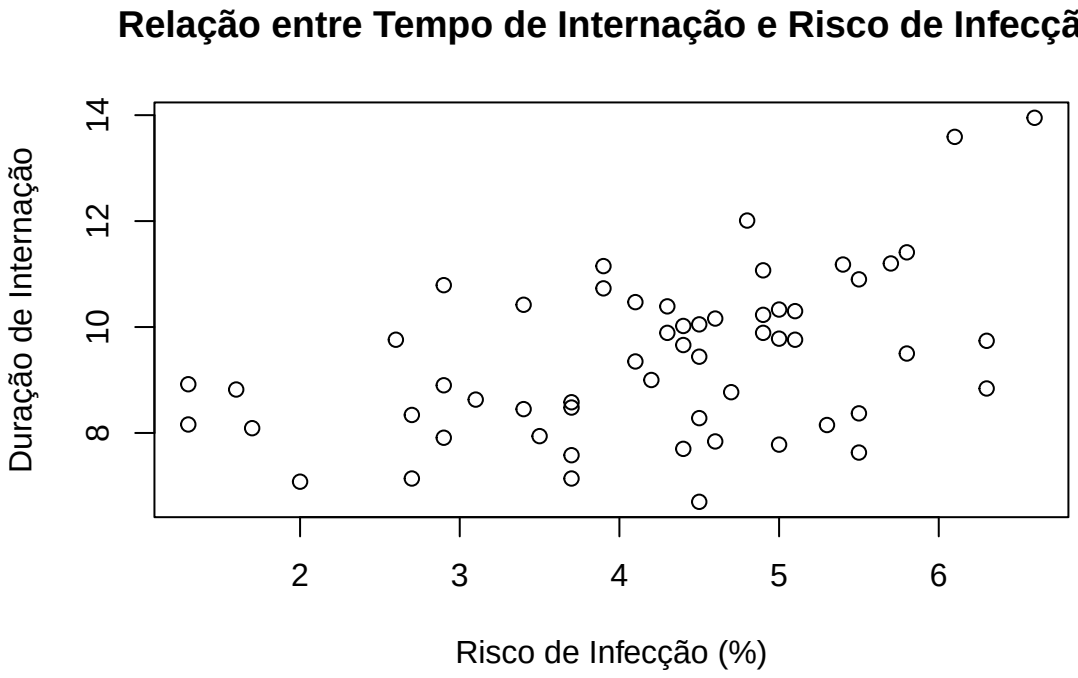


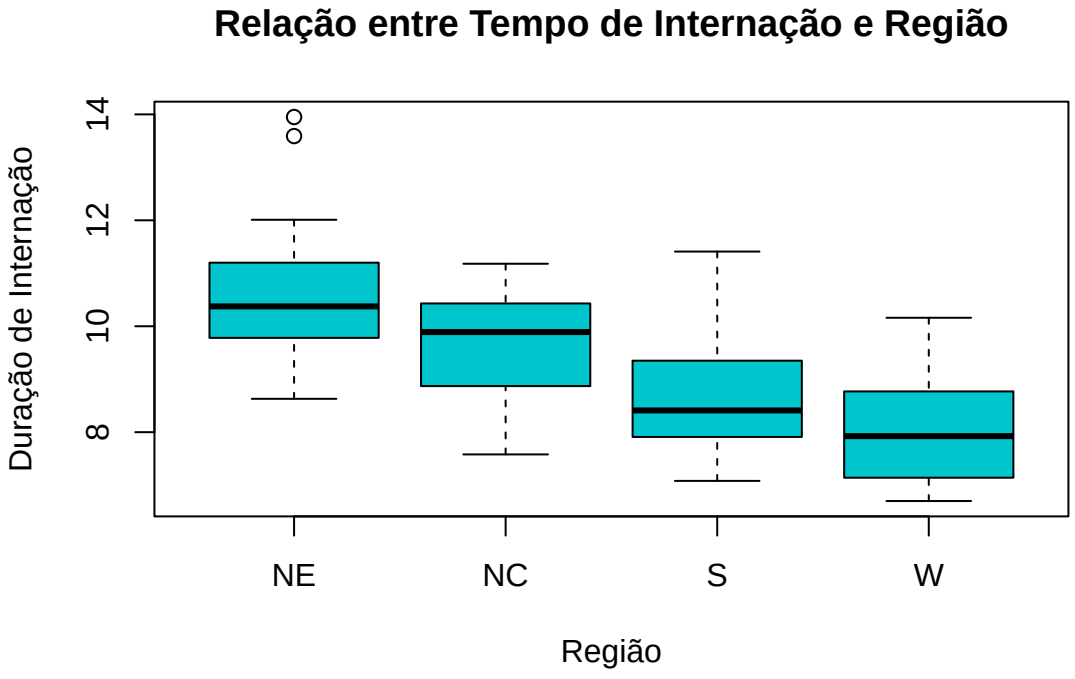
Table 10: Medidas Descritivas para o Tempo de Internação

media	q1	mediana	q3	q4	amplitude	iiq	variancia	dp	cv	Ap	Aq	k
9.4095	8.28	9.44	10.33	13.95	7.25	2.05	2.3197	1.523	0.1619	-0.0601	-0.1317	0.0035

Pelo boxplot, nota-se que existem dois *outliers* e que a mediana da distribuição é próxima de 9 dias. Os coeficientes de assimetria de Pearson (Ap) e quartil de assimetria (Aq) são próximos de zero, ratificando a evidência de simetria da distribuição.
Além disso, a mediana tem um valor de 9.44 e a média de 9.41, valores próximos um do outro devido à simetria já mencionada. O coeficiente de variação de 0.16 mostra que existe relativamente pouca variabilidade nos tempos de internação.

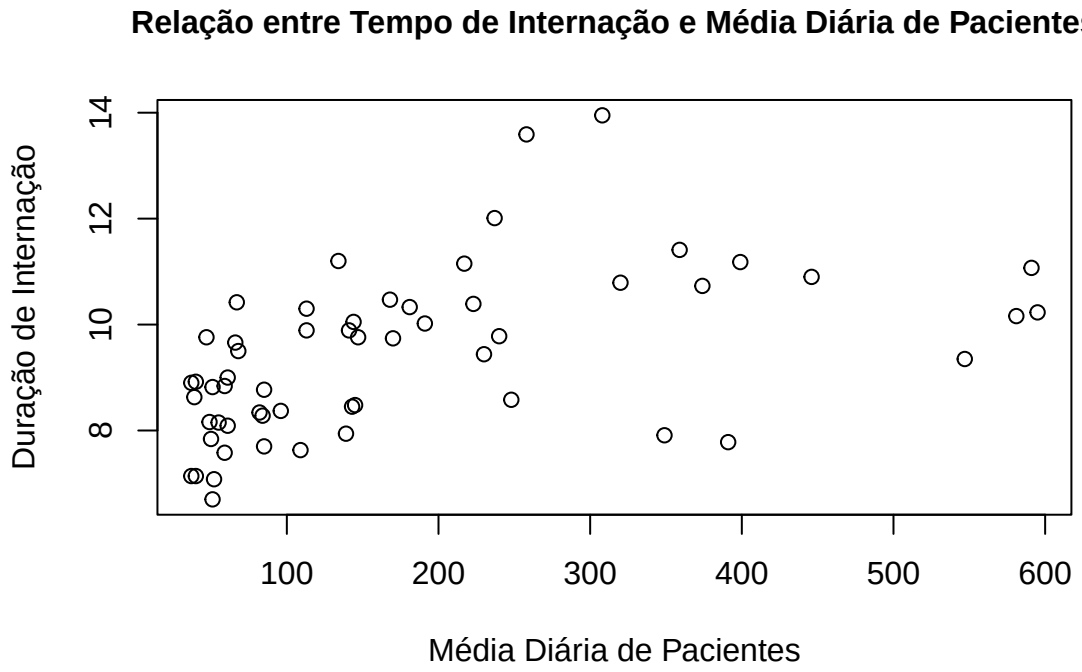


Analisando graficamente a relação entre as variáveis X1 e X3, nota-se que para pacientes com baixo risco de infecção, a duração de internação é menor. Já para pacientes com risco em torno de 6%, esse tempo aumenta.



É evidente que o tempo de internação dos pacientes varia de acordo com a região em que o hospital se

encontra. Na região NE, as internações são mais longas (em torno de 11 dias), enquanto na região W são mais curtas (em torno de 8 dias).



O diagrama acima indica que existe uma relação positiva entre o número médio de pacientes por dia no hospital e o tempo de internação, principalmente para hospitais com média de 100 a 300 pacientes diários.

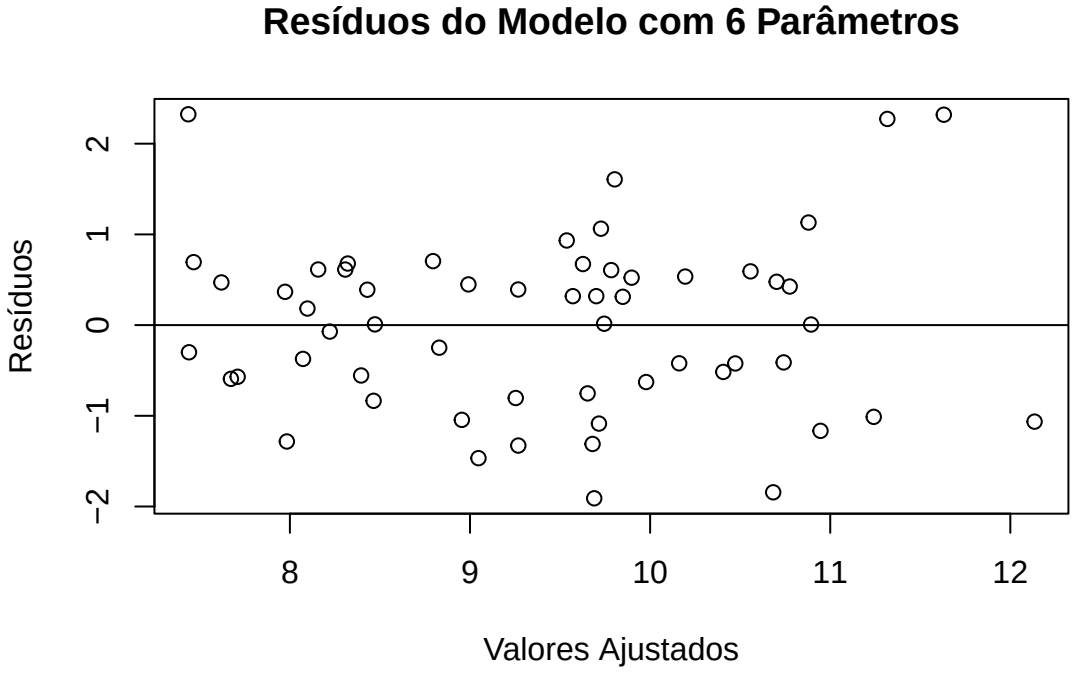
Seleção do Modelo Inicial

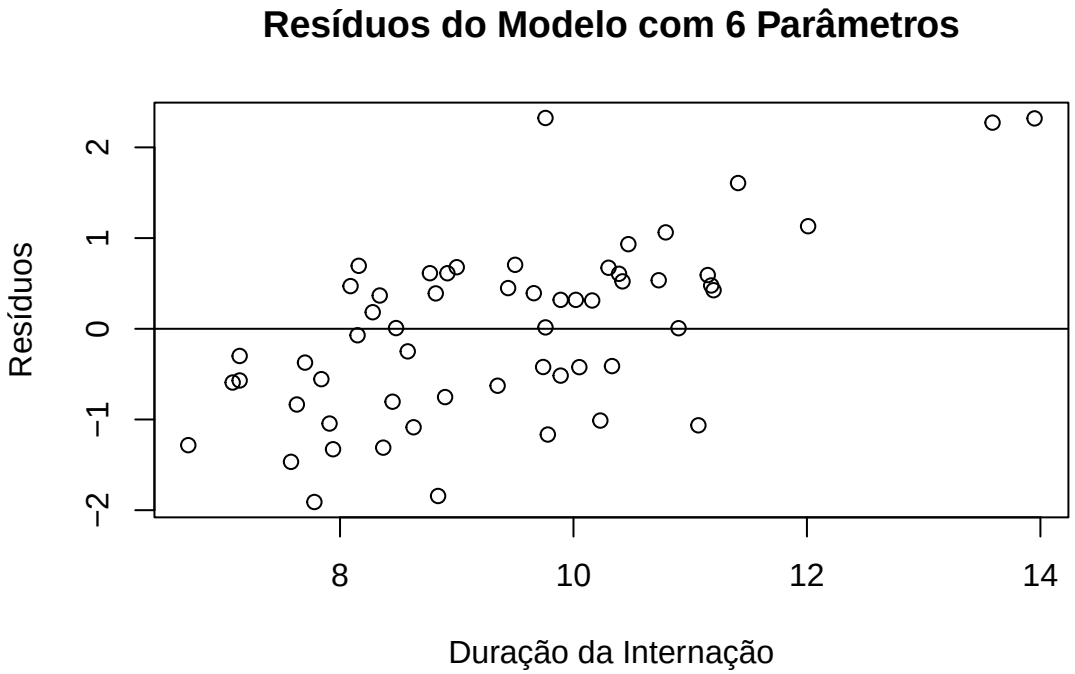
```
## Analysis of Variance Table
##
## Response: amostra$X1
##             Df Sum Sq Mean Sq F value    Pr(>F)
## amostra$X3   1 27.868 27.8682 27.4256 3.122e-06 ***
## amostra$X9   1 16.562 16.5619 16.2989  0.000182 ***
## amostra$XNE  1 21.716 21.7156 21.3708 2.614e-05 ***
## amostra$XNC  1 11.111 11.1112 10.9347  0.001734 **
## amostra$XS   1  0.820  0.8205  0.8074  0.373102
## Residuals   51 51.823  1.0161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Call:
## lm(formula = amostra$X1 ~ amostra$X3 + amostra$X9 + amostra$XNE +
##     amostra$XNC + amostra$XS)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.90963 -0.62776  0.01511  0.59253  2.32412
```

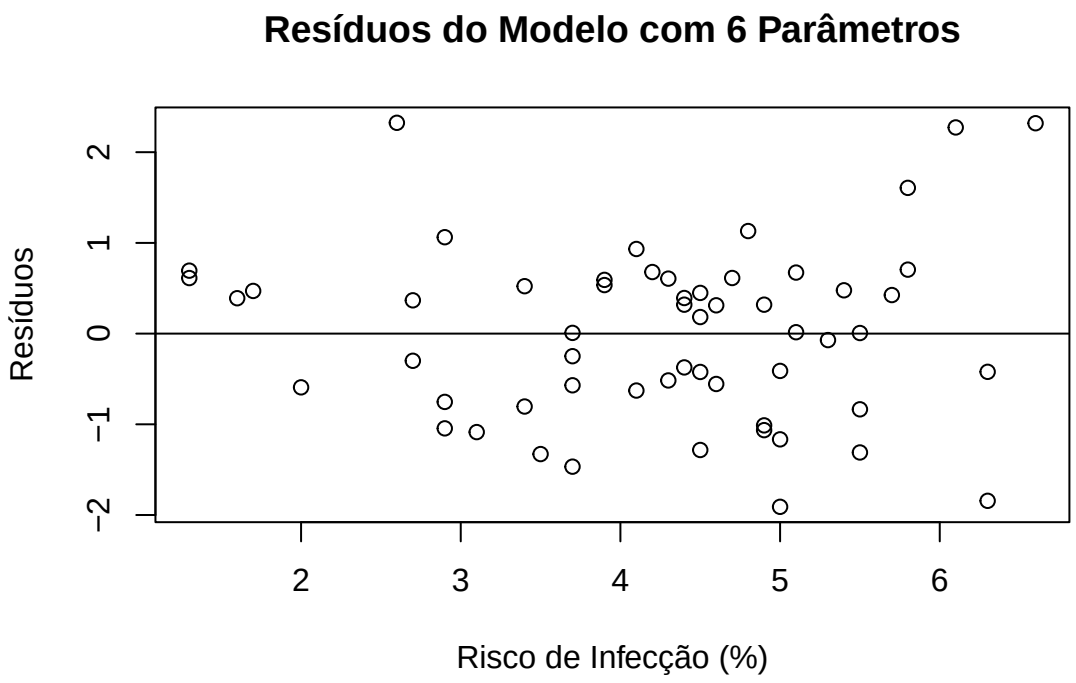
```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.5432419  0.5709883  11.460 1.01e-15 ***
## amostra$X3  0.2806658  0.1176745   2.385 0.020829 *
## amostra$X9  0.0034662  0.0009459   3.664 0.000591 ***
## amostra$XNE 2.1674704  0.4237455   5.115 4.79e-06 ***
## amostra$XNC 1.2607289  0.4022672   3.134 0.002857 **
## amostra$XS  0.3877899  0.4315612   0.899 0.373102
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.008 on 51 degrees of freedom
## Multiple R-squared:  0.6011, Adjusted R-squared:  0.5619
## F-statistic: 15.37 on 5 and 51 DF,  p-value: 3.351e-09
```

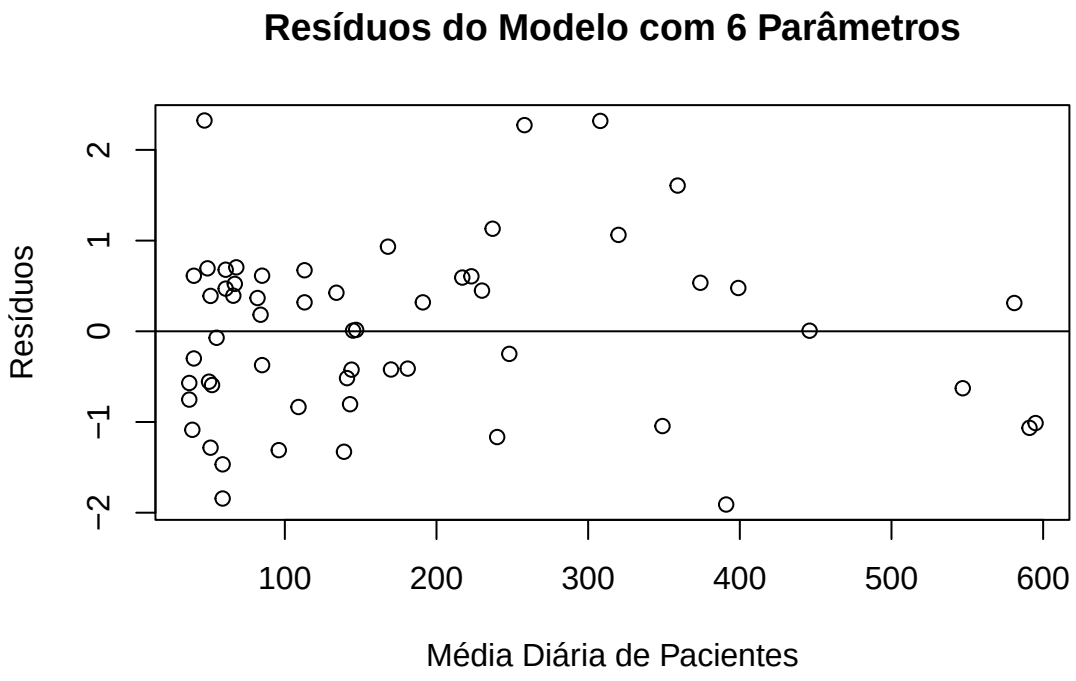
Pela estatística F e seu p-valor menor do que 0,05, conclui-se que, de fato, há regressão. De acordo com os resultados do modelo, um aumento de 1% no risco de infecção resulta em um aumento médio de 0.28 no tempo de duração da internação. Em relação ao tempo de internação nos hospitais da região W, há um aumento médio de aproximadamente 2.17 dias na região NE e 1.26 dias na região NC. Ao passar da região W para a região S, há um aumento de 0.39 na resposta média.

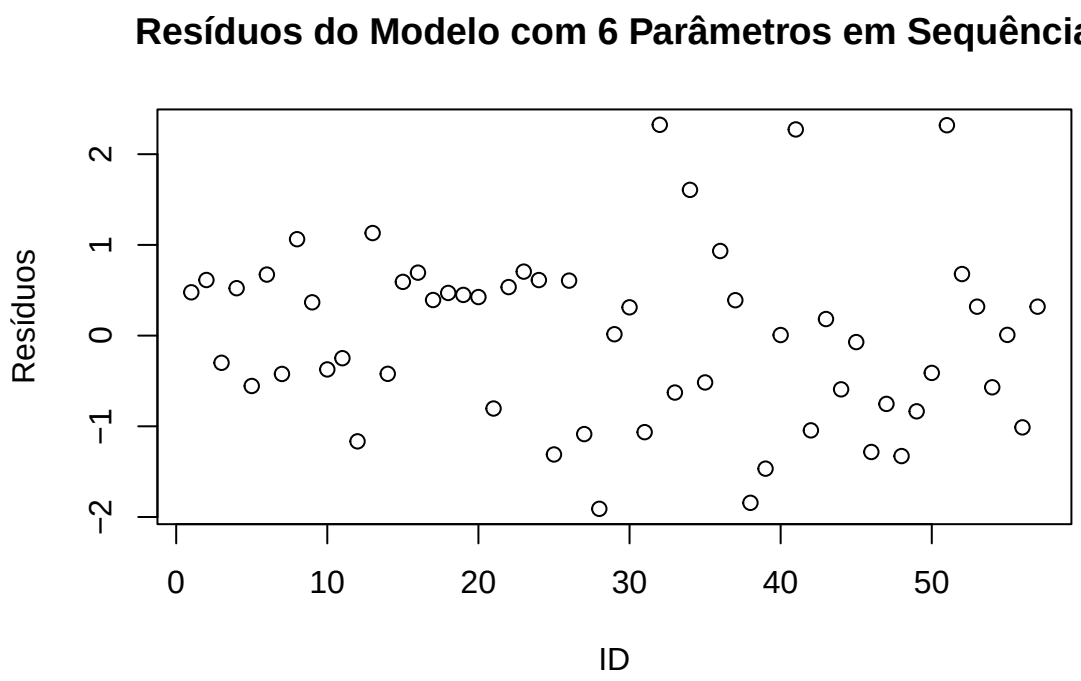
Diagnóstico do Modelo



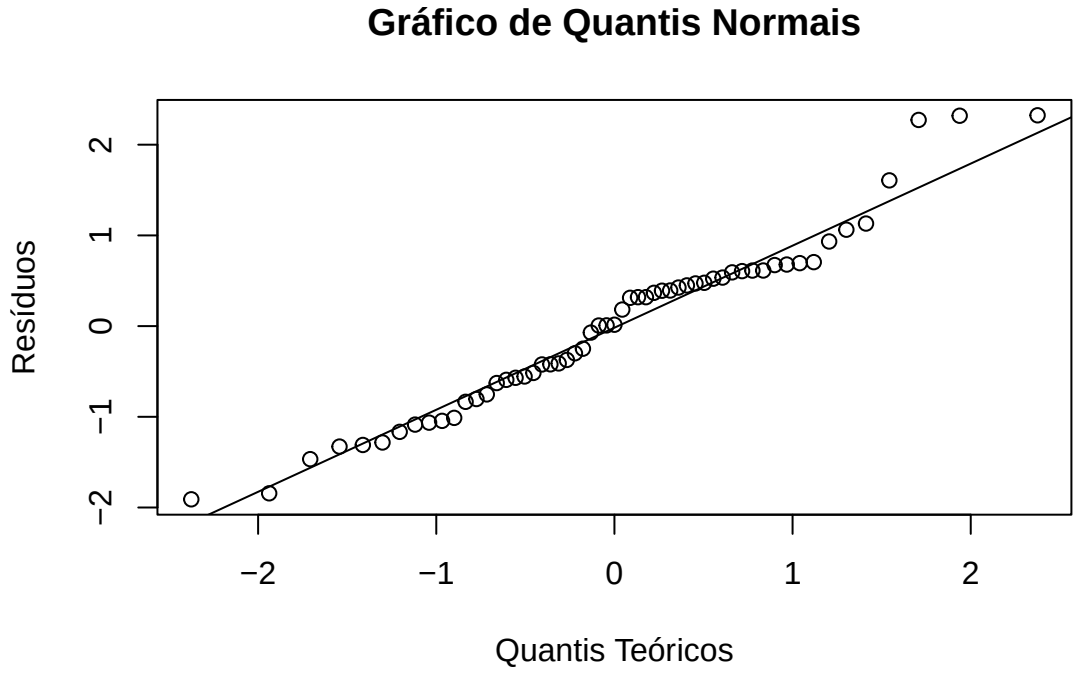



Resíduos do Modelo com 6 Parâmetros
Resíduos
−2
−1
0
1
2
2
3
4
5
6
Risco de Infecção (%)


Resíduos do Modelo com 6 Parâmetros
Resíduos
−2
−1
0
1
2
100
200
300
400
500
600
Média Diária de Pacientes



De modo geral, os gráficos residuais não apontam descumprimento dos pressupostos do modelo. Apenas no diagrama dos resíduos vs a variável resposta, destacam-se os *outliers* já mencionados da distribuição de X1.



O gráfico dos quantis normais também indica ajuste à distribuição normal.

```
##
##  studentized Breusch-Pagan test
##
## data:  modh2
## BP = 6.4253, df = 5, p-value = 0.267
```

O teste de Breusch-Pagan resultou em um p-valor elevado (0.267). Sendo assim, não é possível descartar a hipótese de homocedasticidade.

Validação do Modelo

Table 11: Coeficientes de correlação entre as Variáveis do Modelo na Amostra de Validação

	X1	X3	X9	XNE	XNC	XS
X1	1.0000	0.5770	0.4746	0.4148	-0.0548	-0.1379
X3	0.5770	1.0000	0.4344	0.2032	0.0653	-0.2536
X9	0.4746	0.4344	1.0000	0.1147	0.0613	-0.0707
XNE	0.4148	0.2032	0.1147	1.0000	-0.3175	-0.4820
XNC	-0.0548	0.0653	0.0613	-0.3175	1.0000	-0.4590
XS	-0.1379	-0.2536	-0.0707	-0.4820	-0.4590	1.0000

Os coeficientes de correlação para o modelo ajustado aos dados de validação continuam apresentando associação moderada para a maioria das variáveis explicativas.

```
##
## Call:
## lm(formula = amostra_validacao$X1 ~ amostra_validacao$X3 + amostra_validacao$X9 +
##     amostra_validacao$XNE + amostra_validacao$XNC + amostra_validacao$XS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2169 -0.9052 -0.3116  0.5685  6.8738
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.369231   0.988561   4.420 5.32e-05 ***
## amostra_validacao$X3  0.700241   0.176196   3.974 0.000227 ***
## amostra_validacao$X9  0.003319   0.001579   2.102 0.040566 *
## amostra_validacao$XNE 2.859243   0.790621   3.616 0.000694 ***
## amostra_validacao$XNC 1.325527   0.797821   1.661 0.102885
## amostra_validacao$XS  1.693009   0.746848   2.267 0.027756 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.6 on 50 degrees of freedom
## Multiple R-squared:  0.5309, Adjusted R-squared:  0.484
## F-statistic: 11.32 on 5 and 50 DF,  p-value: 2.486e-07
```

Em geral, as estimativas dos coeficientes de regressão estão próximas aos valores encontrados no modelo de treinamento. A exceção é a variável indicadora da região S, cujo coeficiente era originalmente igual a 0.39 e agora foi estimado como 1.69.

MSPR
2.9517

Ao comparar o erro quadrático médio (2.95) ao quadrado médio do erro do modelo de treinamento (1.02), percebe-se que houve um aumento no seu valor. Com isso, conclui-se que 2.95 provavelmente será uma melhor estimativa do erro de predição do modelo quando este for aplicado a outros dados futuramente.

Modelo Final

Agora, será utilizada a amostra completa dos 113 hospitais para a formulação do modelo final.

```
##
## Call:
## lm(formula = dados_hosp$X1 ~ dados_hosp$X3 + dados_hosp$X9 +
##     dados_hosp$XNC + dados_hosp$XNE + dados_hosp$XS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6578 -0.7590 -0.1698  0.6072  7.4087
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.4132648  0.5523903   9.800  < 2e-16 ***
## dados_hosp$X3  0.5167578  0.1070699   4.826 4.64e-06 ***
## dados_hosp$X9  0.0034365  0.0009158   3.753 0.000285 ***
## dados_hosp$XNC 1.2816467  0.4201311   3.051 0.002880 **
## dados_hosp$XNE 2.4409517  0.4295494   5.683 1.16e-07 ***
## dados_hosp$XS  1.0970302  0.4136199   2.652 0.009212 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.35 on 107 degrees of freedom
## Multiple R-squared:  0.5232, Adjusted R-squared:  0.5009
## F-statistic: 23.48 on 5 and 107 DF,  p-value: 7.32e-16
```

Modelo com 5 parâmetros

Um segundo modelo, agora com 5 parâmetros, foi selecionado para fins de análise. Essa seleção foi feita ao se avaliar conjuntamente o R^2, o $R^2 ajustado$ e o Cp. Inclui as variáveis explicativas X3, X9, XNC, XNE.

Seleção do Modelo Inicial

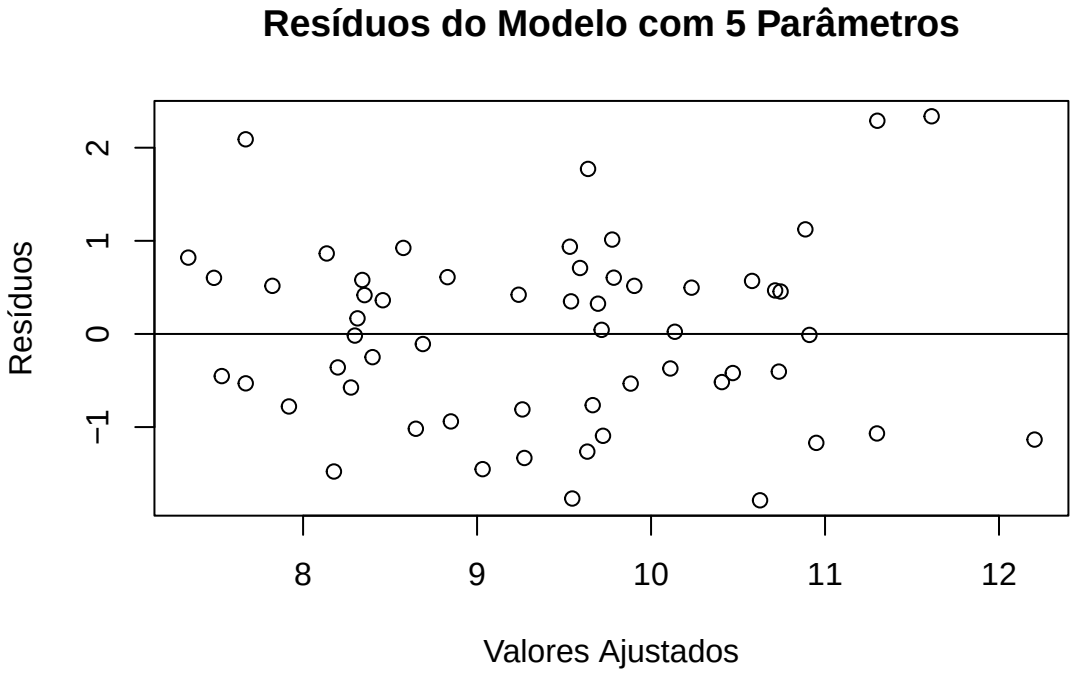
```
## Analysis of Variance Table
##
## Response: amostra$X1
##             Df Sum Sq Mean Sq F value    Pr(>F)
## amostra$X3   1 27.868 27.8682  27.527 2.889e-06 ***
## amostra$X9   1 16.562 16.5619  16.360 0.0001742 ***
## amostra$XNE  1 21.716 21.7156  21.450 2.464e-05 ***
## amostra$XNC  1 11.111 11.1112  10.975 0.0016852 **
```

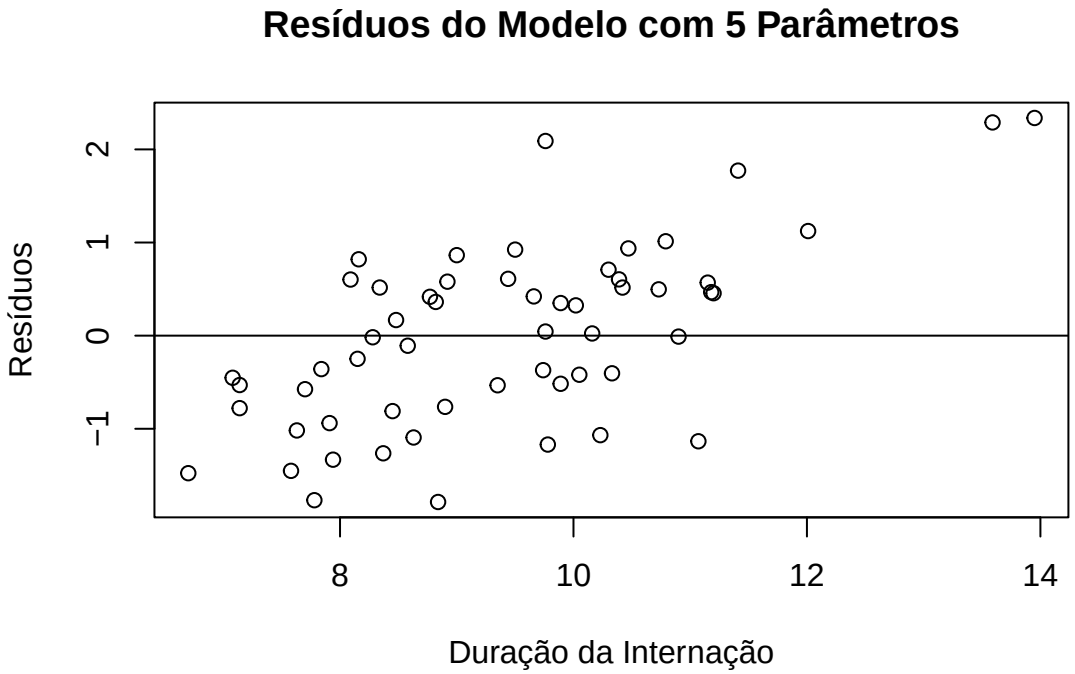
```
## Residuals   52 52.644  1.0124
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

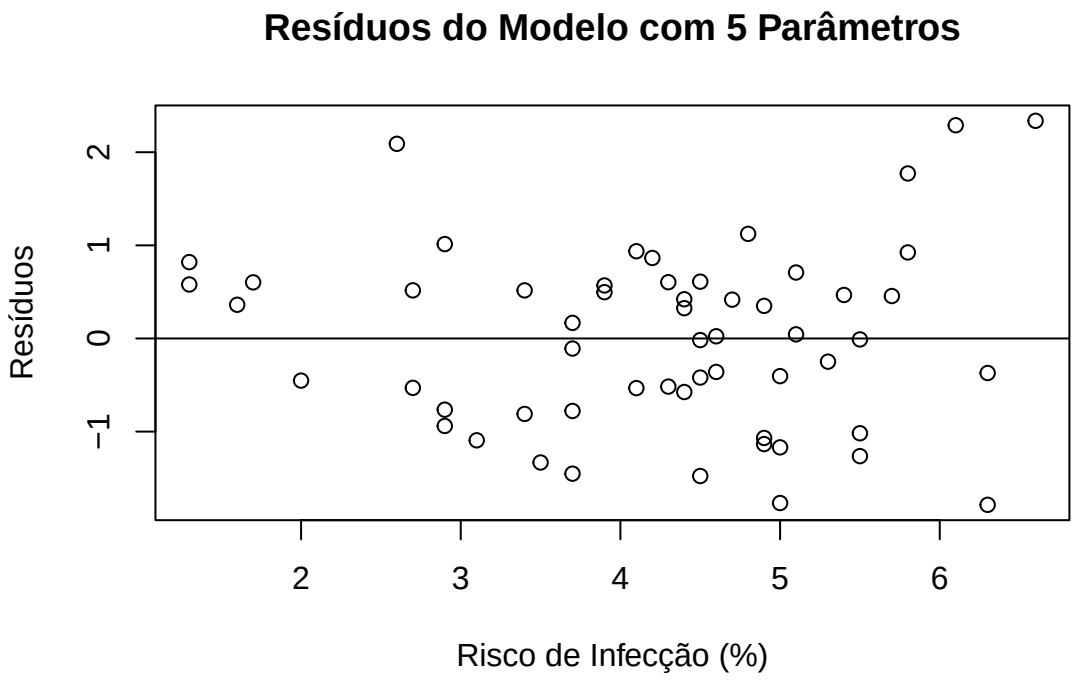
```
##
## Call:
## lm(formula = amostra$X1 ~ amostra$X3 + amostra$X9 + amostra$XNE +
##     amostra$XNC)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.78650 -0.76470  0.02337  0.57967  2.33731
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.8243489  0.4767767  14.314  < 2e-16 ***
## amostra$X3  0.2592770  0.1150283   2.254 0.028435 *
## amostra$X9  0.0036482  0.0009223   3.956 0.000232 ***
## amostra$XNE 1.9534653  0.3498355   5.584 8.67e-07 ***
## amostra$XNC 1.0329912  0.3118077   3.313 0.001685 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.006 on 52 degrees of freedom
## Multiple R-squared:  0.5947, Adjusted R-squared:  0.5636
## F-statistic: 19.08 on 4 and 52 DF,  p-value: 1.041e-09
```

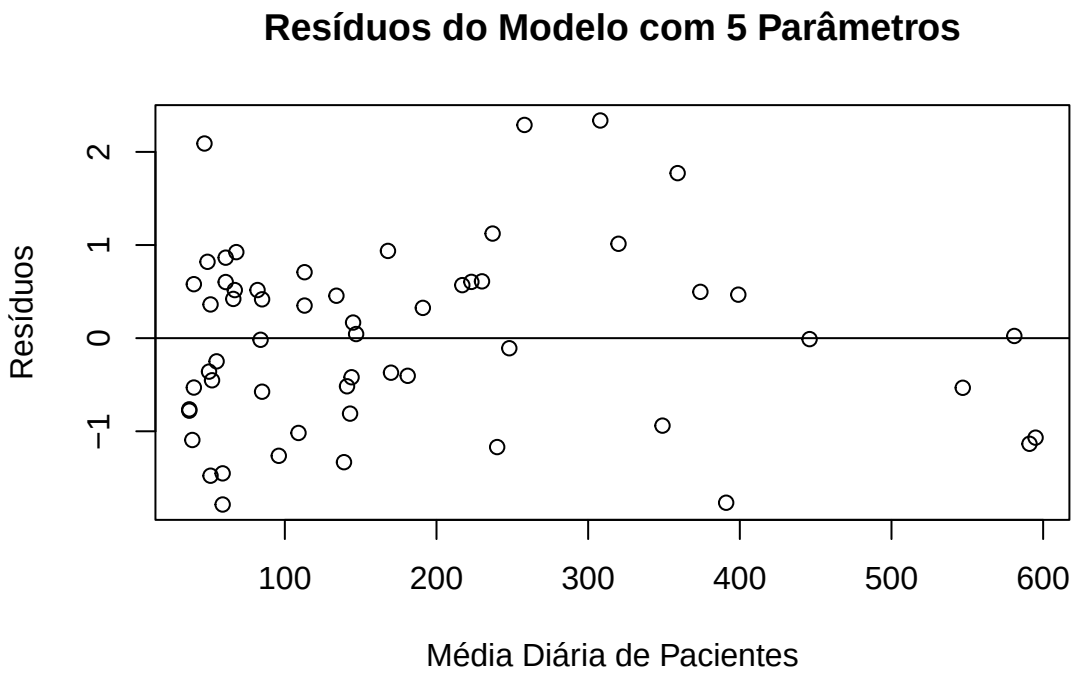
Pela estatística F e seu p-valor menor do que 0,05, conclui-se que, de fato, há regressão.
De acordo com os resultados do modelo, um aumento de 1% no risco de infecção resulta em um aumento médio de 0.26 no tempo de duração da internação. Já um aumento de 1 paciente na média diária do hospital causa um aumento de 0.004 na média da variável resposta.
Em relação ao tempo de internação nos hospitais da região W, há um aumento médio de aproximadamente 1.95 dias na região NE e 1.03 dias na região NC.

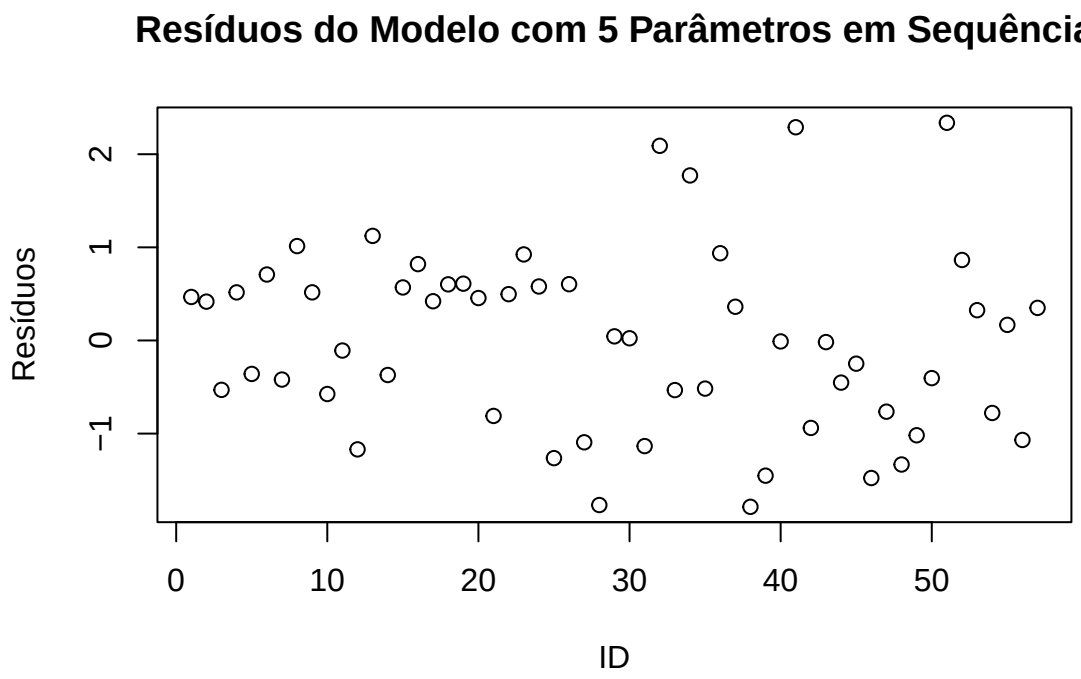
Diagnóstico do Modelo



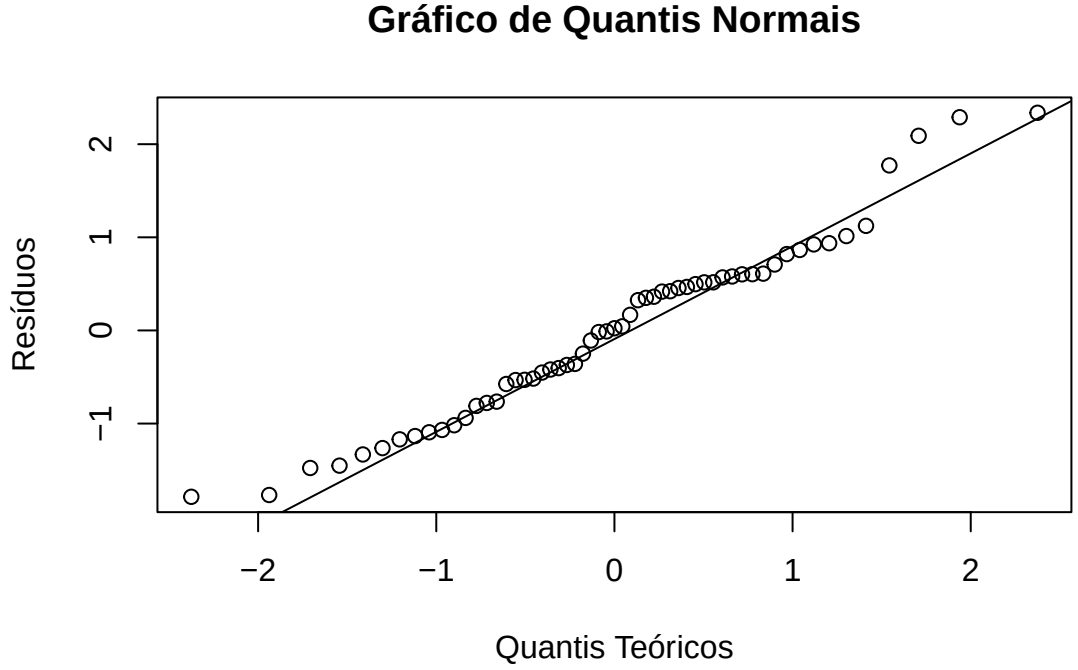



Resíduos do Modelo com 5 Parâmetros
Resíduos
−1
0
1
2
2
3
4
5
6
Risco de Infecção (%)


Resíduos do Modelo com 5 Parâmetros
Resíduos
−1
0
1
2
100
200
300
400
500
600
Média Diária de Pacientes



De maneira geral, os gráficos residuais parecem respeitar os pressupostos do modelo (normalidade, heterocedasticidade, etc).



##

```
##  Shapiro-Wilk normality test
##
## data:  modh2_2$residuals
## W = 0.97022, p-value = 0.1719
```

E novamente, o gráfico indica um bom ajustamento à distribuição normal, evidência que pode ser confirmada com o teste de Shapiro-Wilk.
Outra suposição que deve-se verificar é se há a homogeneidade de variâncias:

```
##
##  studentized Breusch-Pagan test
##
## data:  modh2_2
## BP = 7.7951, df = 4, p-value = 0.09938
```

Assim como para o pressuposto de normalidade, com o teste de Breusch-Pagan não existem evidências para rejeitar a hipótese nula. Isto é, existe homocedasticidade.

Validação do Modelo

```
##
## Call:
## lm(formula = amostra_validacao$X1 ~ amostra_validacao$X3 + amostra_validacao$X9 +
##     amostra_validacao$XNE + amostra_validacao$XNC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3643 -0.7671 -0.3166  0.5922  6.9546
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5.897412   0.751809   7.844 2.55e-10 ***
## amostra_validacao$X3    0.632705   0.180567   3.504 0.000964 ***
## amostra_validacao$X9    0.003857   0.001623   2.377 0.021264 *
## amostra_validacao$XNE   1.542481   0.557687   2.766 0.007886 **
## amostra_validacao$XNC -0.006055   0.561357  -0.011 0.991435
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.663 on 51 degrees of freedom
## Multiple R-squared:  0.4827, Adjusted R-squared:  0.4422
## F-statistic:  11.9 on 4 and 51 DF,  p-value: 6.67e-07
```

MSPR
2.5194

Ao comparar o erro quadrático médio (2.51) ao quadrado médio do erro do modelo de treinamento (1.01), percebe-se que houve um aumento no seu valor. Então, assim como para o modelo com 6 parâmetros, conclui-se, que 2.51 provavelmente será uma melhor estimativa do erro de predição do modelo quando este for aplicado a outros dados futuramente.

Modelo Final

Agora, será utilizada a amostra completa dos 113 hospitais para a formulação do modelo final de 5 parâmetros.

```
##
## Call:
## lm(formula = dados_hosp$X1 ~ dados_hosp$X3 + dados_hosp$X9 +
##     dados_hosp$XNC + dados_hosp$XNE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3797 -0.8449 -0.1156  0.6365  7.4605
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.3005748  0.4516731  13.949  < 2e-16 ***
## dados_hosp$X3  0.4680556  0.1083904   4.318 3.50e-05 ***
## dados_hosp$X9  0.0038844  0.0009249   4.200 5.51e-05 ***
## dados_hosp$XNC 0.5147368  0.3131960   1.643    0.103
## dados_hosp$XNE 1.6961481  0.3340023   5.078 1.60e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.388 on 108 degrees of freedom
## Multiple R-squared:  0.4918, Adjusted R-squared:  0.473
## F-statistic: 26.13 on 4 and 108 DF,  p-value: 3.668e-15
```

Conclusão

Em relação à primeira hipótese levantada, foi selecionado o seguinte modelo:

$$\widehat{log(X_{10i})} = 3.739 + 0.047X_{NCi} + 0.086X_{NEi} + 0.074X_{Si} + 0.0005X_{11i}^2$$

Com isso, conclui-se que o percentual de serviços fornecidos pelo hospital está relacionado com o número de enfermeira(o)s através de um modelo de segundo grau, conforme a hipótese. Além disso, o número de enfermeira(o)s varia conforme a região.
Como a aplicação do logaritmo foi usada como medida corretiva para garantir a normalidade e heterocedasticidade dos dados, os coeficientes de regressão do modelo tiveram seus valores reduzidos, ficando próximos de zero.

Em relação à segunda hipótese, foram selecionados 2 modelos:

$$\widehat{X_{1i}} = 5.413 + 0.517X_{3i} + 0.003X_{9i} + 1.282X_{NCi} + 2.441X_{NEi} + 1.097X_{Si}$$

$$\widehat{X_{1i}} = 6.301 + 0.468X_{3i} + 0.004X_{9i} + 0.515X_{NCi} + 1.696X_{NEi}$$

Foi descoberto que as variáveis de maior influência sobre a duração das internações são a região, o risco de infecção do paciente e a média diária de pacientes do hospital.