Análise de Dados em Regressão Linear Múltipla

Arthur Silva, Gustavo Braga e Romulo Freitas

Universidade Federal do Ceará

Professor/Orientador:
Prof. Dr. Ronald Targino Nojosa

Modelos de Regressão Linear I
28 de dezembro de 2023

Sumário

Introdução

Métricas

Métodos

Aplicação

Análise de Resíduos

Conclusões

Introdução

Um problema importante em muitas aplicações da análise de regressão envolve selecionar o conjunto de variáveis independentes ou regressores a serem usadas no modelo [2].

É importante destacar que nem todos os candidatos a regressores são necessários para modelar adequadamente a variável resposta $\boldsymbol{Y}$, portanto, devemos selecionar um subconjunto apropriado de variáveis explicativas a partir de um conjunto que inclua provavelmente todas as variáveis importantes.

Esta seleção é feita por meio de métricas e comparações entre os possíveis modelos, considerando várias possibilidades de formação com k variáveis disponíveis.

Métricas

Coeficiente de Determinação

O **Coeficiente de Determinação (R^2)** é uma medida que varia no intervalo [0, 1] e indica a proporção da variação na variável dependente que é explicada pela(s) variável(is) regressora(s) [3]. A fórmula do $\boldsymbol{R^2}$ é dada por:

$$R^2 = \frac{SQ_{REG}}{SQ_{TOT}} = 1 - \frac{SQRES}{SQ_{TOT}}.$$

No entanto, deve-se usá-lo com cautela, pois um $\boldsymbol{R^2}$ alto não garante um modelo robusto, já que esta medida é inflacionada com o acréscimo de novas variáveis no modelo.

Métricas

Coeficiente de Determinação Ajustado

O **Coeficiente de Determinação Ajustado (R_a^2)** apresenta-se como uma métrica alternativa ao R^2. A inclusão indiscriminada de variáveis pode aumentar o R^2, prejudicando assim, o princípio da parcimônia. Nesse contexto, o R_a^2 atua como um ajuste ao penalizar a incorporação de regressores pouco explicativos, promovendo, assim, uma abordagem mais criteriosa na seleção de variáveis [2]. A fórmula do $\boldsymbol{R_a^2}$ é dada por:

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)(1-R^2).$$

Métricas

Quadrado Médio Residual

O **Quadrado Médio Residual** QM_{Res} é uma medida que avalia a dispersão dos resíduos. Essencialmente, ele quantifica a média dos quadrados dos erros residuais, os quais correspondem às discrepâncias entre os valores observados e os previstos pelo modelo. Assim, a diminuição do valor do QM_{Res} indica uma melhor capacidade do modelo em se adaptar aos dados. A fórmula do QM_{Res} é dada por:

$$\hat{\sigma}^2 = QM_{RES} = \frac{SQ_{RES}}{n-p},$$

onde p é o número de parâmetros no modelo.

Métricas

Estatística C_p de *Mallows*

A **Estatística C_p de Mallows** é calculado como a soma dos quadrados dos resíduos do modelo, ajustado para o número de preditores no modelo. O objetivo é encontrar um equilíbrio entre um modelo que se ajusta bem aos dados e um modelo que seja simples [3]. A fórmula do C_p é dada por:

$$C_p = \frac{SQ_{RES}(p)}{\hat{\sigma}^2} - n + 2p.$$

Métricas

AIC

O **Critério de Informação de Akaike** (AIC) é comumente utilizado para comparar modelos estatísticos e selecionar o que oferece o melhor equilíbrio entre o ajuste aos dados e simplicidade.

Hirotugu Akaike propôs o AIC, baseado na maximização da entropia esperada do modelo. Essencialmente, o AIC é uma medida de probabilidade logarítmica penalizada. Seja $\hat{L}$ o ponto de máximo da função de verossimilhança para um modelo específico. O AIC é

$$AIC = -2\ln(\hat{L}) + 2p,$$

onde p é o número de parâmetros do modelo [3].

Métricas

BIC

O **Critério de Informação Bayesiano** é uma medida estatística utilizada para selecionar modelos estatísticos. Similar ao AIC, o BIC busca encontrar um equilíbrio entre o ajuste do modelo aos dados e a sua complexidade. O BIC penaliza a complexidade do modelo de maneira mais rigorosa do que o AIC [3].
A fórmula geral para o BIC é:

$$BIC = -2\ln(\hat{L}) + p\ln(n),$$

onde $\ln(n)$ representa o logaritmo natural do tamanho da amostra. O modelo que resulta no menor valor de BIC é considerado preferível.

Resumo

Principais métricas para seleção de variáveis

Métrica	**Fórmula**	**Interpretação**
R_a^2	$1 - \left(\frac{n-1}{n-p}\right)\frac{SQRes}{SQTot}$	Valor ajustado que mostra o quanto a variação de Y é explicada pelo modelo. Quanto maior, melhor.
$QMRes$	$\frac{SQRes}{n-p}$	Estimativa da variância populacional. Quanto menor, melhor.
AIC	$-2\ln(\hat{L}) + 2p$	Uso da parcimônia e maximização de entropia. Quanto menor, melhor.
BIC	$-2\ln(\hat{L}) + p\ln(n)$	Mesmo princípio do AIC, com maior penalidade. Quanto menor, melhor.
C_p	$\frac{SQRes}{QMres} - n + 2(p+1)$	Alternativa complementar ao AIC e BIC. Quanto menor, melhor

Todas as Possíveis Regressões (*Best Subset Selection*)

A partir do modelo nulo, este método cria todos os possíveis modelos com as k variáveis disponíveis, e faz uma comparação entre seus respectivos R^2 ou SQ_{RES}, selecionando aqueles com os valores mais baixos destas medidas.

Em seguida, usando métricas pré-estabelecidas, define o melhor modelo dentre os selecionados no passo anterior.

Pontos a considerar:

1. Com k variáveis, haverá 2^k modelos, o que pode ser custoso computacionalmente para muitas variáveis.
2. Por considerar todos os modelos e ser menos criterioso na seleção dos "melhores", pode levar ao caso de *overfitting*.

Seleção Progressiva (*Forward Stepwise*)

Adiciona variáveis gradativamente ao modelo nulo, e avalia com uma métrica, a influência da variável adicionada. Se ela for siginificante para a regressão, a mesma é mantida, e novo teste é realizado com outra variável, considerando o novo modelo. O procedimento é repetido k vezes.

Eliminação Regressiva (*Backward Stepwise*)

Sendo o oposto do método anterior, o *Backward* inicia com um modelo completo (todas as variáveis), e vai gradativamente realizando deleções.

Na primeira iteração, é verificado com uma métrica, se a remoção da variável "melhora" ou "piora" o modelo. Caso melhore, a mesma é retirada, do contrário, ela é mantida, e a iteração vai para o próximo passo, considerando o novo modelo.

O processo é repetido k vezes.

Forward-Backward Stepwise

O último método é uma combinação dos dois métodos anteriores, onde a partir do modelo nulo ou do modelo completo, são feitas várias combinações de adição/retirada de variáveis, avaliando-se a influência das variáveis na regressão a partir de métrica(s) estabelecida(s).

Desempenho dos Times de 1976 da Liga Nacional de Futebol Americano

A base de dados escolhida é referente às estatísticas do desempenho dos times de 1976 da Liga Nacional de Futebol Americano, LNFA *(National Football League, Fonte: The Sporting News)*. Ela foi retirada do livro ***Estatística Aplicada e Probabilidade para Engenheiros*** [2, pág. 237].

Time	y	x1	x2	x3	x4	x5	x6	x7	x8	x9
Washington	10	2113	1985	38.9	64.7	4	868	59.7	2205	1917
Minnesota	11	2003	2855	38.8	61.3	3	615	55	2096	1575
New England	11	2957	1737	40.1	60	14	914	65.6	1847	2175
Oakland	13	2285	2905	41.6	45.3	-4	957	61.4	1903	2476
Pittsburgh	10	2971	1666	39.2	53.8	15	836	66.1	1457	1866
Baltimore	11	2309	2927	39.7	74.1	8	786	61	1848	2339
Los Angeles	10	2528	2341	38.1	65.4	12	754	66.1	1564	2092
Dallas	11	2147	2737	37	78.3	-1	797	58.9	2476	2254
Atlanta	4	1689	1414	42.1	47.6	-3	714	57	2577	2001
Buffalo	2	2566	1838	42.3	54.2	-1	797	58.9	2476	2254

Tabela com as 10 primeiras observações.

Variáveis

onde:

- $\boldsymbol{Y}$: Jogos ganhos (por 14 jogos na temporada);
- $\boldsymbol{X_1}$: Jardas conquistadas na corrida (temporada);
- $\boldsymbol{X_2}$: Jardas conquistadas na passagem da bola (temporada);
- $\boldsymbol{X_3}$: Jardas conquistadas antes de dar um chute na bola (jardas/chute);
- $\boldsymbol{X_4}$: Percentagem de gols feitos com bola chutada (gol feitos/gols tentados - temporada);
- $\boldsymbol{X_5}$: Diferença de perda de bolas (bolas tomadas - bolas perdidas);

Variáveis

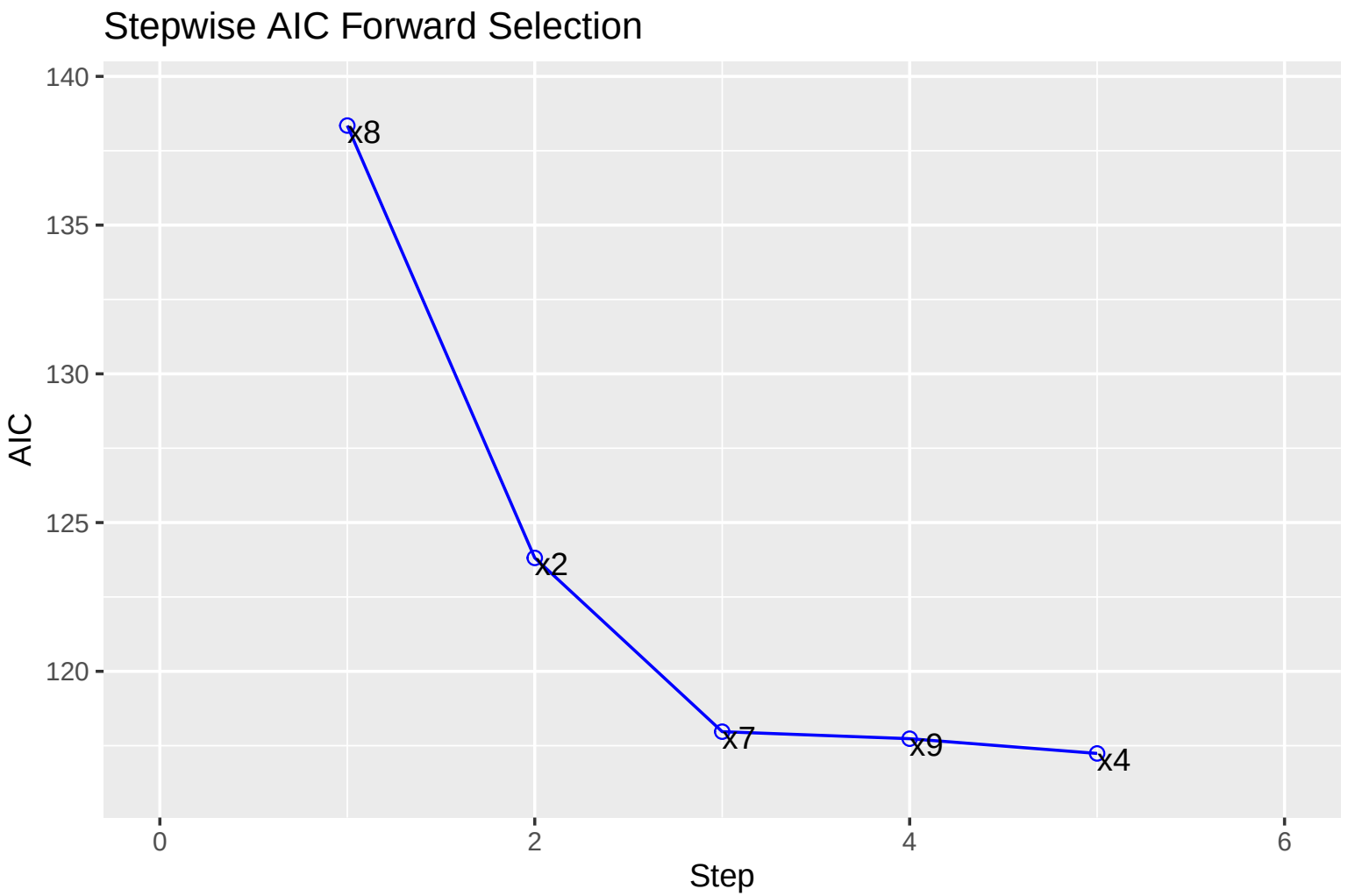
- $\boldsymbol{X_6}$: Jardas conquistadas antes de perder a bola (temporada);
- $\boldsymbol{X_7}$: Percentagem de corrida (jogadas na corrida/jogadas total);
- $\boldsymbol{X_8}$: Jardas conquistadas pelo oponenente na corrida (temporada);
- $\boldsymbol{X_9}$: Jardas conquistadas pelo oponente na passagem da bola (temporada).

Seleção do Melhor Subconjunto de Variáveis

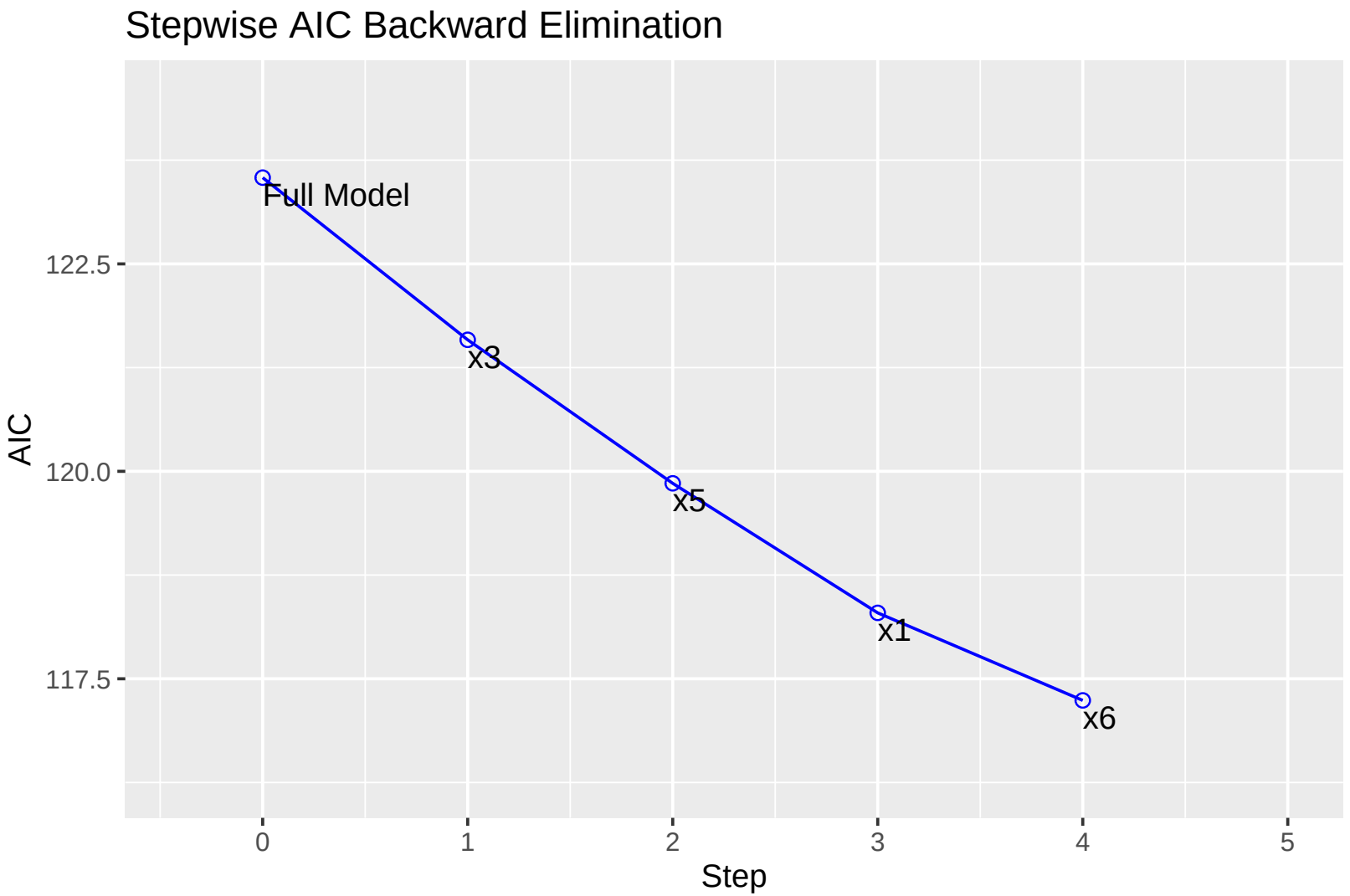
A função `ols_step_best_subset()` do pacote **olsrr** [1] aponta os melhores modelos, com as seguintes variáveis regressoras:

- **Variáveis Regressoras do Modelo 1:** X_2, X_4, X_7, X_8, X_9.
- **Variáveis Regressoras do Modelo 2:** X_2, X_7, X_8, X_9.
- **Variáveis Regressoras do Modelo 3:** X_2, X_7, X_8.

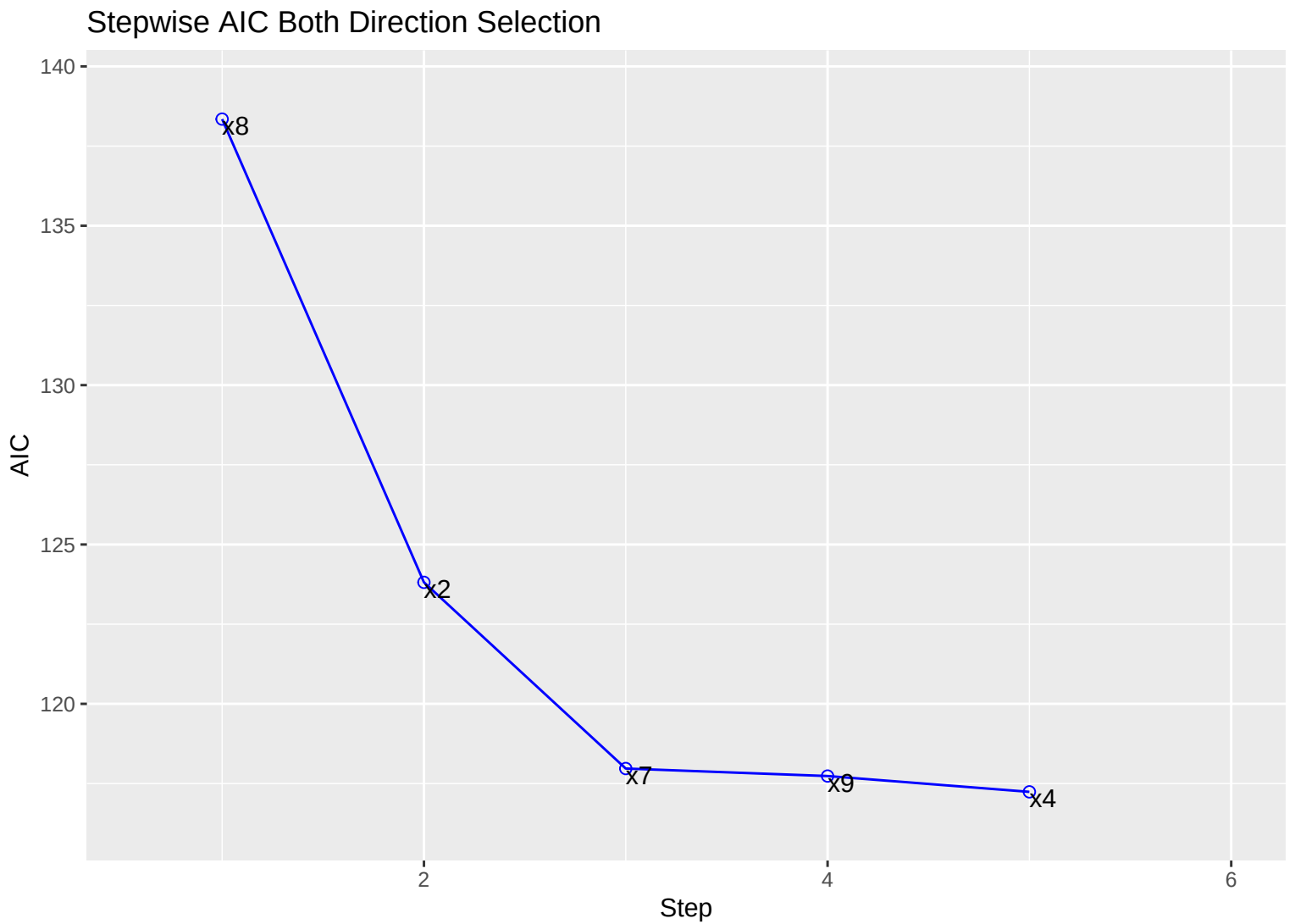
Seleção *Forward Stepwise*



Seleção *Backward Stepwise*



Seleção *Forward-Backward Stepwise*



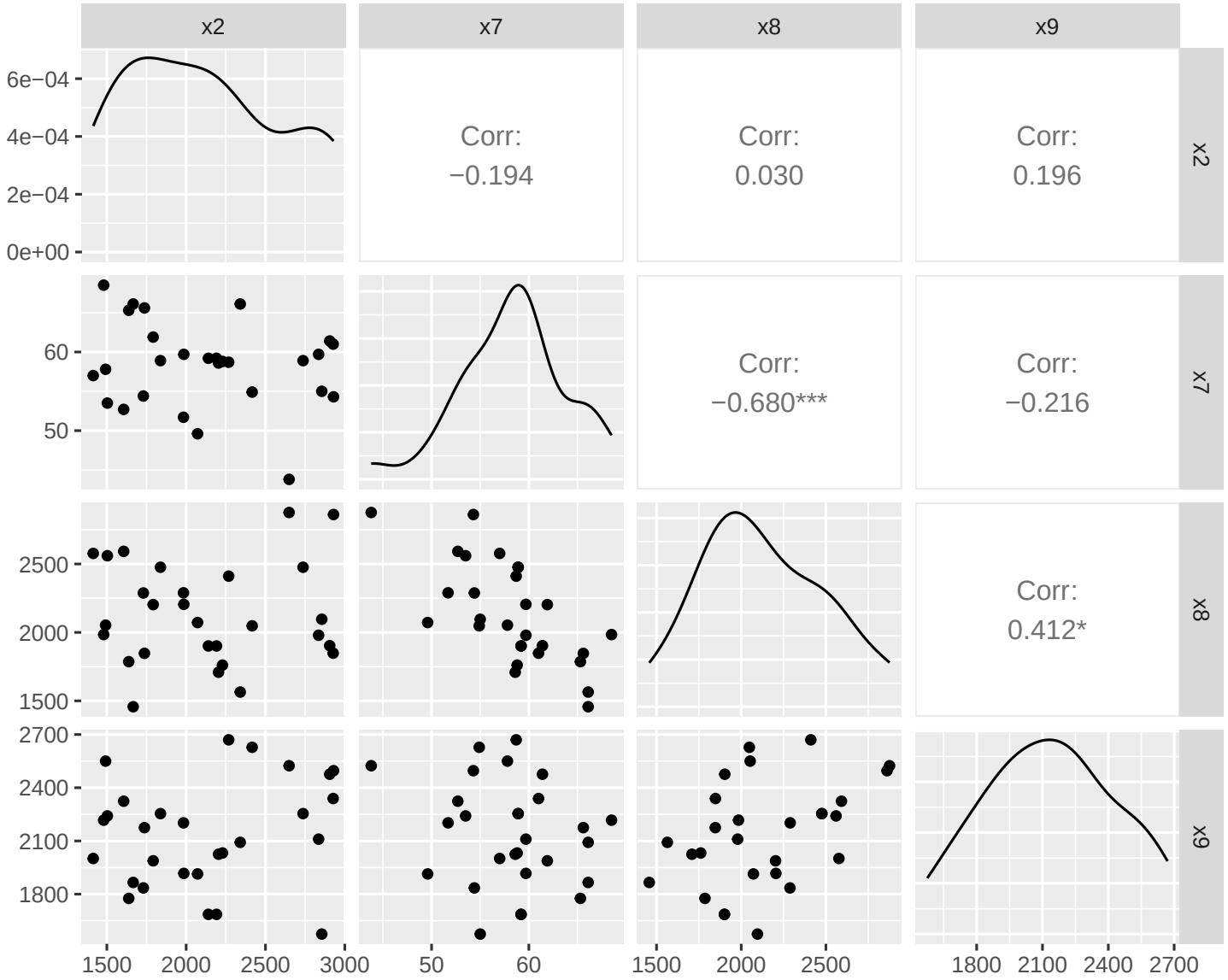
Modelos Selecionados

Após a etapa de seleção das variáveis para a construção do melhor modelo de regressão linear possível, foram selecionados **três** modelos baseados nas métricas que foram estabelecidas, sendo eles:

1. **Modelo 1:** $Y = \beta_0 + \beta_1 X_2 + \beta_2 X_4 + \beta_3 X_7 + \beta_4 X_8 + \beta_5 X_9 + \varepsilon$;
2. **Modelo 2:** $Y = \beta_0 + \beta_1 X_2 + \beta_2 X_7 + \beta_3 X_8 + \beta_4 X_9 + \varepsilon$;
3. **Modelo 3:** $Y = \beta_0 + \beta_1 X_2 + \beta_2 X_7 + \beta_3 X_8 + \varepsilon$.

Das 09 variáveis presentes na base de dados, 5 (X_2, X_4, X_7, X_8 e X_9) foram selecionadas.

Matriz de Correlação



Análise de Diagnóstico dos Modelos

Modelos estatísticos são utilizados como aproximações de processos complexos e são construídos sobre um conjunto de suposições [4].

Os resíduos são utilizados para avaliar a qualidade do ajuste do modelo de regressão e para validar as suposições de normalidade, homocedasticidade, independência dos erros e checar a existência de *outliers*.

Nesse sentido, será realizada uma análise de resíduos dos modelos selecionados, a fim de encontrar o mais "robusto" se utilizando do critério da parcimômia.

Modelo 1

Estimação

A partir da seleção de variáveis por meio das métricas estabelecidas, foram estimados os parâmetros do modelo de regressão e a seguinte função estimada:

$$\hat{Y} = -10,584 + 0,0043X_2 + 0,048X_4 + 0,284X_7 - 0,003X_8 - 0,002X_9$$

SUMÁRIO - MODELO 1

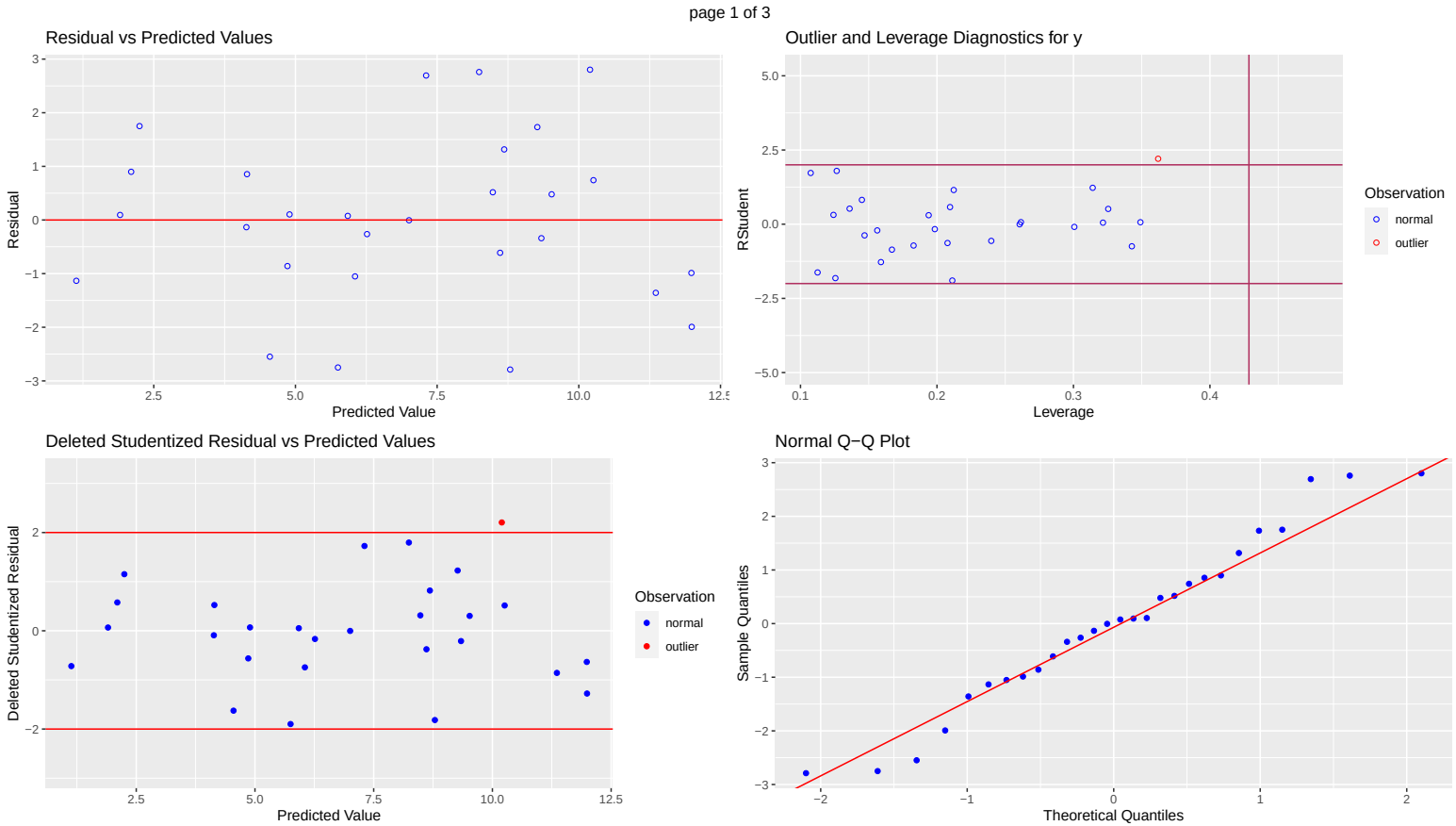
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-10.5840	7.9392	-1.33	0.1961
x2	0.0040	0.0007	5.47	0.0000
x4	0.0480	0.0335	1.43	0.1662
x7	0.2843	0.0873	3.26	0.0036
x8	-0.0028	0.0014	-2.01	0.0564
x9	-0.0020	0.0013	-1.61	0.1227

ANOVA - MODELO 1

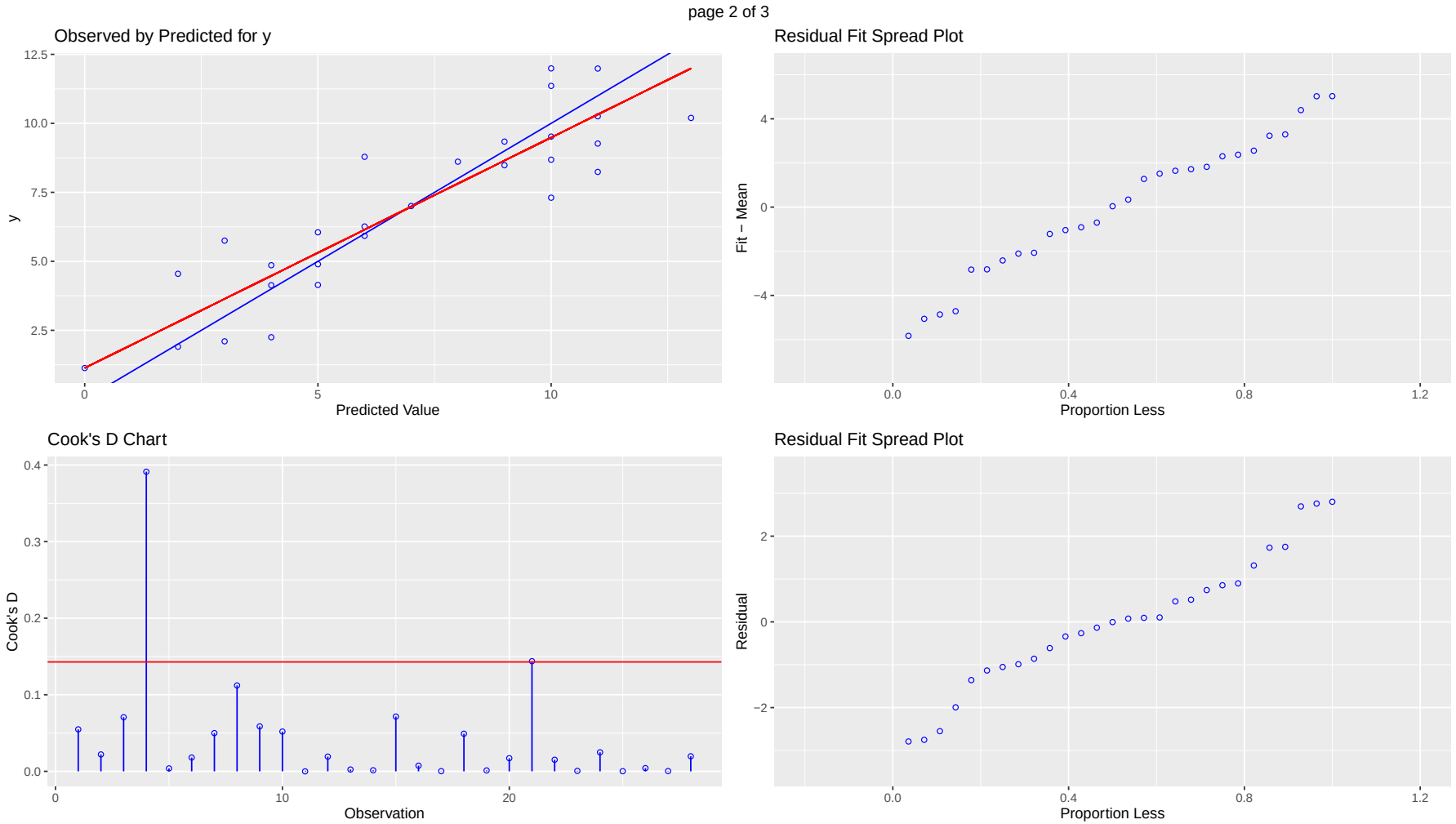
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	5	261.50	52.30	17.58	0.0000
Residuals	22	65.46	2.98		

Análise de Resíduos - Modelo 1

Já com a função `ols_plot_diagnostics()`, temos gráficos mais completos e precisos:



Análise de Resíduos - Modelo 1



Normalidade dos Resíduos - Modelo 1

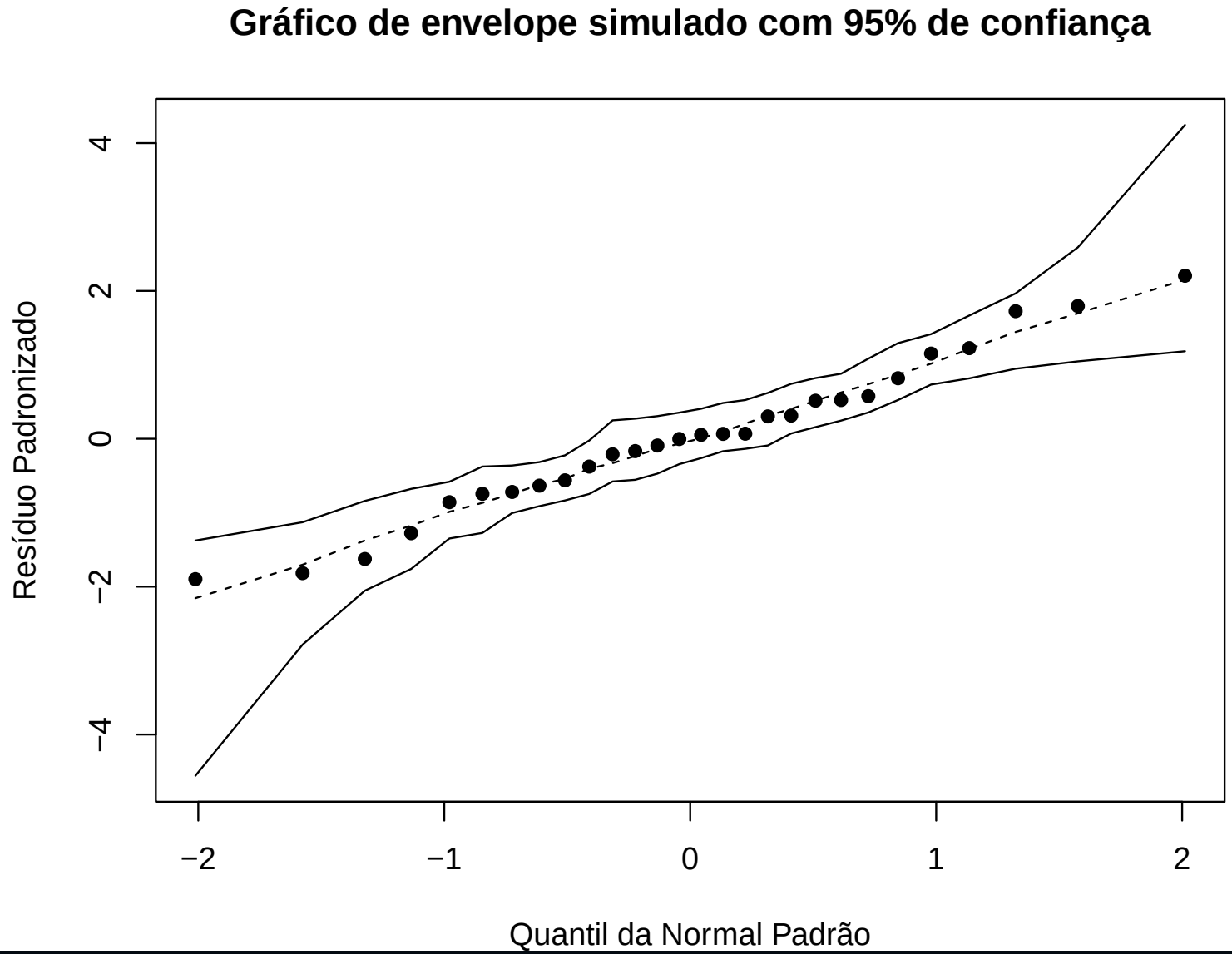
Através da função `ols_test_normality()` do pacote **olsrr** [1], podemos testar a normalidade dos resíduos por meio de testes não paramétricos como o teste de Shapiro-Wilk e o teste de Kolmogorov-Smirnov:

Teste	**Estatística**	**p-valor**
Shapiro-Wilk	0,9688	0,5500
Kolmogorov-Smirnov	0,0805	0,9867

Resultados dos Testes

Nenhum dos testes realizados rejeitou a hipótese de normalidade dos resíduos.

Envelopes Simulados - Modelo 1



Pressupostos - Modelo 1

Homocedasticidade

Função `bptest()` no R:

```
data: m1 BP = 6.9601, df = 5, p-value = 0.2236
```

Ausência de Multicolinearidade

Função `vif()` no R:

X_2	X_4	X_7	X_8	X_9
1.187897	1.143936	2.059484	2.335104	1.312865

Modelo 2

Estimação

A partir da seleção de variáveis por meio das métricas estabelecidas, foram estimados os parâmetros do modelo de regressão e a seguinte função estimada:

$$\hat{Y} = -7,043 + 0,004X_2 + 0,266X_7 - 0,003X_8 - 0,002X_9$$

SUMÁRIO - MODELO 2

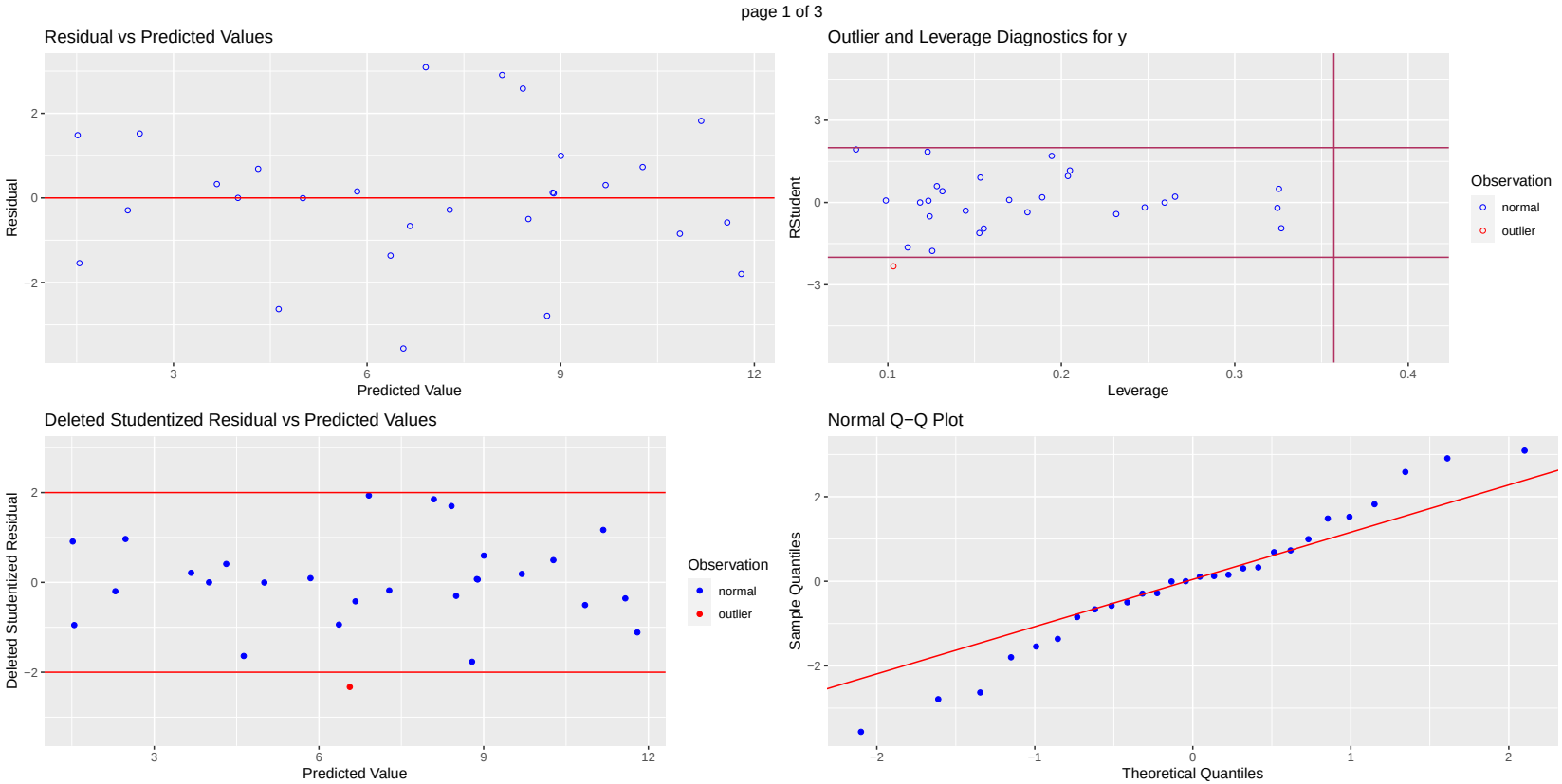
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-7.0427	7.7147	-0.91	0.3708
x2	0.0042	0.0007	5.84	0.0000
x7	0.2663	0.0883	3.01	0.0062
x8	-0.0031	0.0014	-2.24	0.0347
x9	-0.0018	0.0013	-1.38	0.1800

ANOVA - MODELO 2

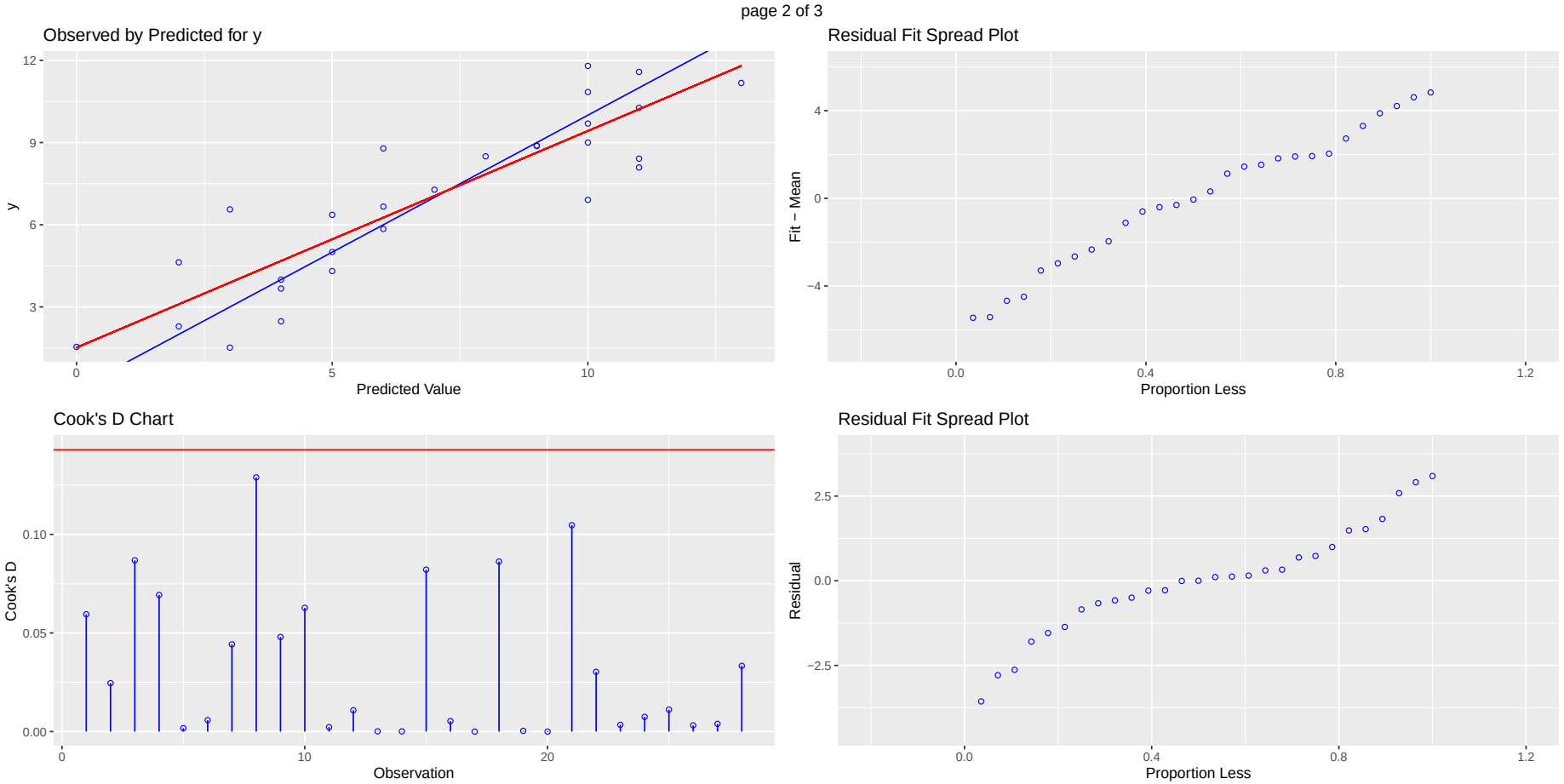
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	4	255.40	63.85	20.52	0.0000
Residuals	23	71.56	3.11		

Análise de Resíduos - Modelo 2

Já com a função `ols_plot_diagnostics()`, temos gráficos mais completos e precisos:



Análise de Resíduos do Modelo 2



Testes de Normalidade dos Resíduos - Modelo 2

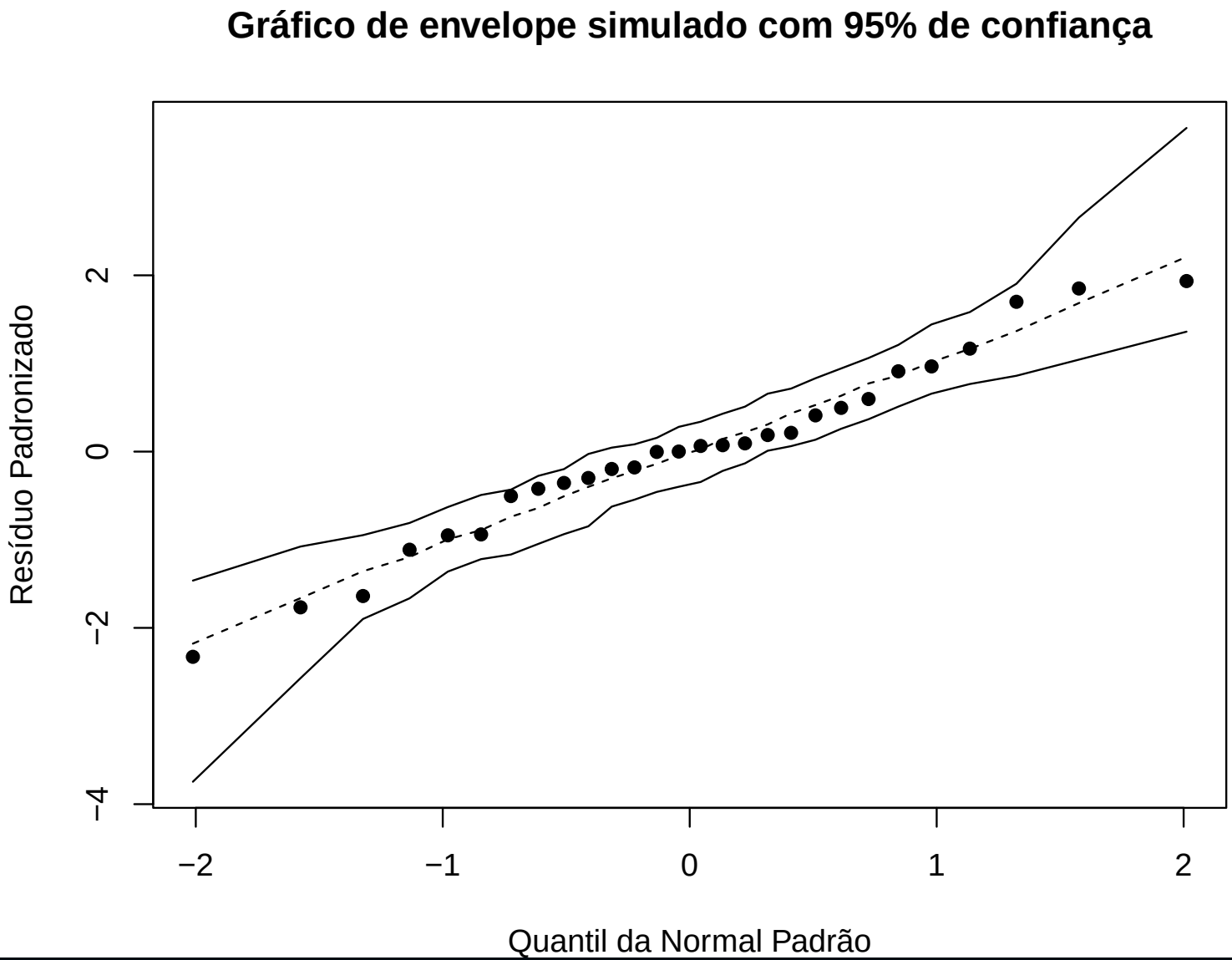
A seguir o resultado dos testes de Shapiro-Wilk e Kolmogorov-Smirnov para testar a normalidade dos resíduos do modelo 2:

Teste	**Estatística**	**p-valor**
Shapiro-Wilk	0.9776	0.7905
Kolmogorov-Smirnov	0.0987	0.9232

Resultados dos Testes

Nenhum dos testes realizados rejeitou a hipótese de normalidade dos resíduos.

Envelopes Simulados - Modelo 2



Pressupostos - Modelo 2

Homocedasticidade

Função `ols_test_breusch_pagan()` no R:

```
Chi2 = 0.03953339, DF = 1, Prob > Chi2 = 0.8423957
```

Ausência de Multicolinearidade

Função `ols_vif_tol()` no R:

X_2	X_7	X_8	X_9
1.121421	2.016836	2.265781	1.285705

Modelo 3

Estimação

A partir da seleção de variáveis por meio das métricas estabelecidas, foram estimados os parâmetros do modelo de regressão e a seguinte função estimada:

$$\hat{Y} = -7,634 + 0,004X_2 + 0,248X_7 - 0,004X_8$$

SUMÁRIO - MODELO 3

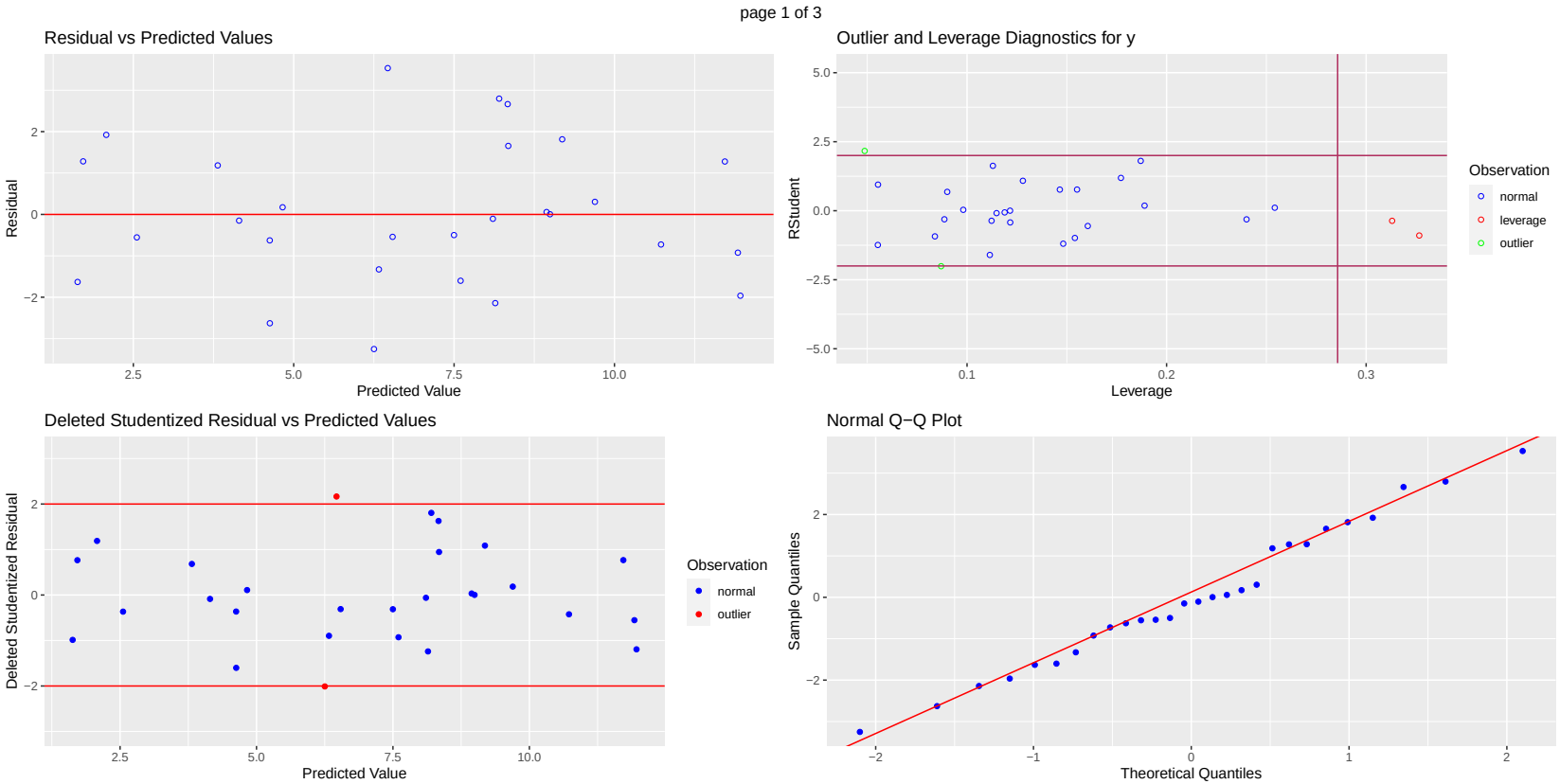
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-7.6345	7.8478	-0.973	0.3403
x2	0.0040	0.0007	5.572	0.0000
x7	0.2478	0.0890	2.785	0.0103
x8	-0.0039	0.0013	-3.005	0.0061

ANOVA - MODELO 3

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	3	249.454	83.151	25.747	0.0000
Residuals	24	77.511	3.230		

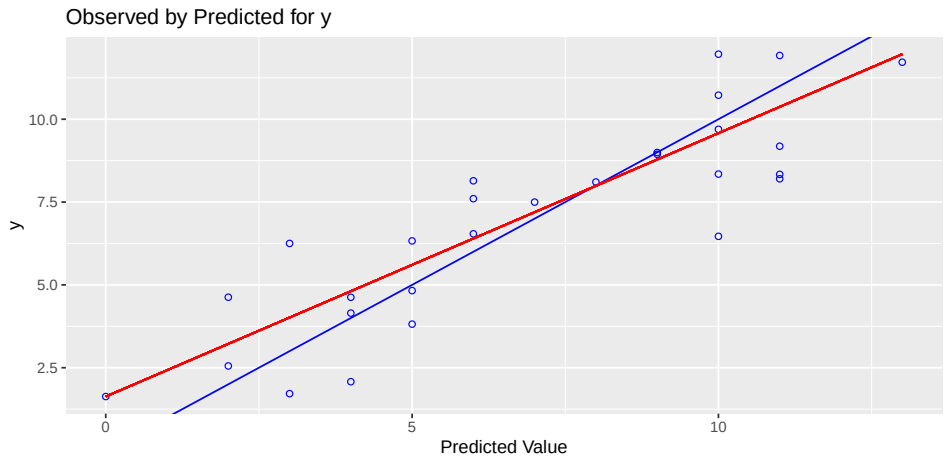
Análise de Resíduos - Modelo 3

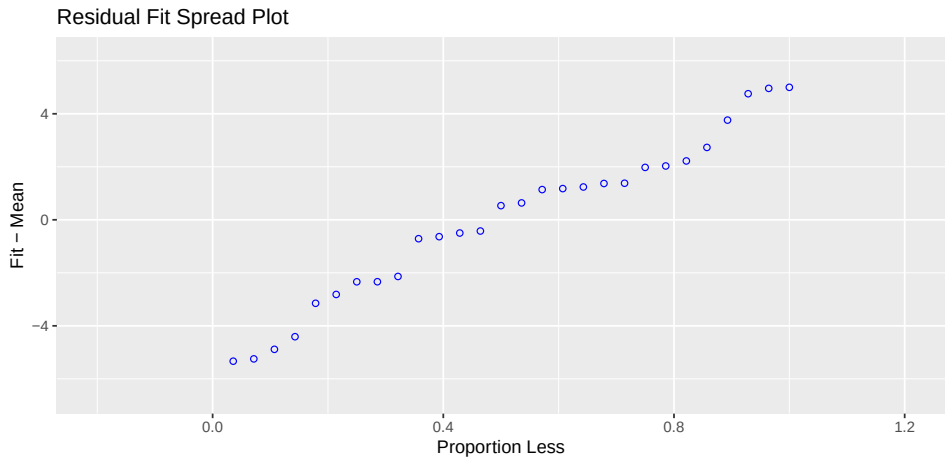
Gráficos contruídos por meio da função `ols_plot_diagnostics()` do pacote olsrr [1].

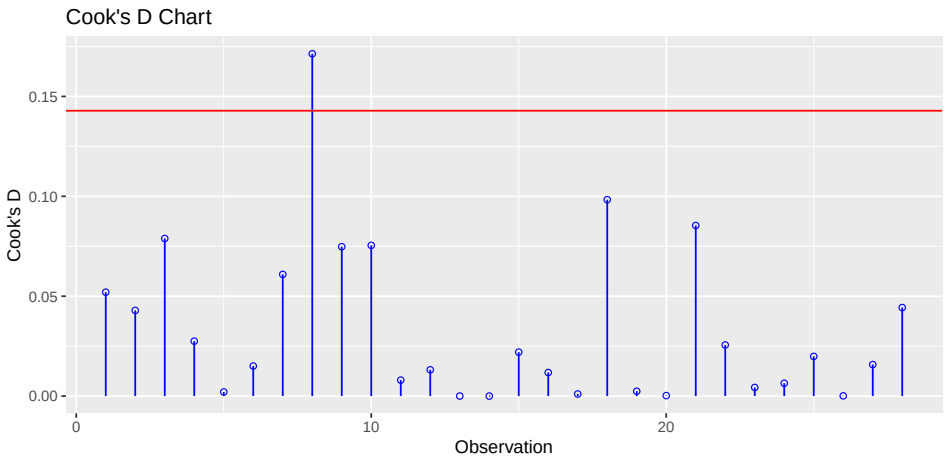


Análise de Resíduos - Modelo 3

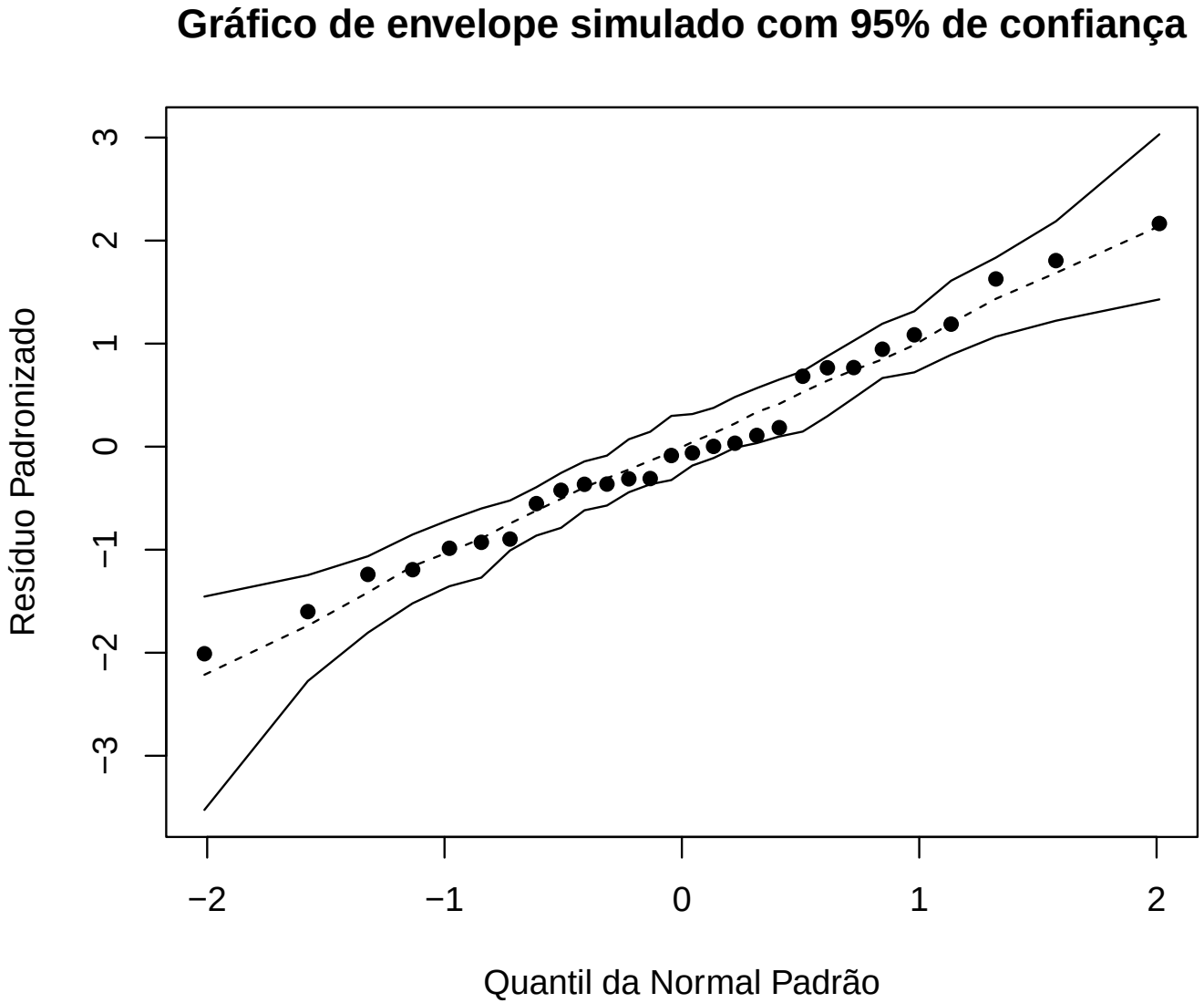
page 2 of 3







Envelopes Simulados - Modelo 3



Testes de Normalidade dos Resíduos - Modelo 3

A seguir o resultado dos testes de Shapiro-Wilk e Kolmogorov-Smirnov para testar a normalidade dos resíduos do modelo 3:

Teste	**Estatística**	**p-valor**
Shapiro-Wilk	0,9817	0,8892
Kolmogorov-Smirnov	0,1073	0,8704

Resultados dos Testes

Nenhum dos testes realizados rejeitou a hipótese de normalidade dos resíduos.

Pressupostos - Modelo 3

Homocedasticidade

Função `ols_test_breusch_pagan()` no R:

```
Chi2 = 0.0151, DF = 1, Prob > Chi2 = 0.9023
```

Ausência de Multicolinearidade

Função `ols_vif_tol()` no R:

X_2	X_7	X_8
1,0606	1,9704	1,8978

Conclusões

Ao longo desta análise, exploramos detalhadamente a seleção de variáveis e a construção de modelos, destacando métricas como R^2, C_p, AIC, BIC e o teste de significância F. A aplicação de métodos como, *Forward Stepwise*, *Backwar Stepwise* e *Forward-Backward Stepwise* resultou na identificação de três possíveis modelos.

Conclusões

Os valores das principais métricas encontradas para cada modelo são:

	Modelo 1	Modelo 2	Modelo 3
R_a^2	0,7543	0,7431	0,7333
AIC	117,2391	117,7348	117,9705
BIC	126,5645	125,728	124,6315
C_p	3,1263	2,9094	2,6474
QM_{Res}	2,975	3,111	3,230

Conclusões

Ao compararmos os valores de AIC dos três modelos, junto com as métricas QM_{Res} e R_a^2, o modelo 1 se apresenta como o mais favorável já que seu valor AIC é levemente inferior aos demais. Além disso, ele possui um maior R_a^2 e um menor QM_{Res}.

Conclusões

Quando comparamos o valor BIC, o modelo 3 se apresenta como o mais favorável, visto que ele apresenta uma valor inferior considerável em relação aos demais. É importante pontuar que ele também foi o mais significante, segundo testes realizados por meio da estatística F. Além disso, o modelo 3 é o mais parcimonioso.

Conclusões

"Essencialmente, todos os modelos estão errados, mas alguns são úteis."
George Box.

Referências

[1] Aravind Hebbali. *olsrr: Tools for Building OLS Regression Models*. Fev. de 2020. URL: https://cran.r-project.org/web/packages/olsrr/index.html (acedido em 25/11/2023).

[2] Douglas C. Montgomery. *Estatística Aplicada E Probabilidade Para Engenheiros*. pt-BR. 7ª ed. Rio de Janeiro, RJ: Ltc-Livros Tecnicos E Cientificos Editora Lda, jul. de 2022. ISBN: 9788521637332.

[3] Douglas C. Montgomery, Elizabeth A. Peck e G. Geoffrey Vining. *Introduction to linear regression analysis*. 5th ed. Wiley series in probability and statistics 821. Hoboken, NJ: Wiley, 2012. ISBN: 9780470542811.

[4] J.M. Singer, J.S. Nobre e F.M.M. Rocha. *Análise de Dados Longitudinais (versão parcial preliminar)*. Disponível para download em https://www.ime.usp.br/~jmsinger/MAE0610/Singer&Nobre&Rocha2018jun.pdf. 2018.