Produção de Amônia

Fernado Bispo, Jeff Caponero

Sumário

Introdução

Com base nos dados disponibilizados no *dataset* "stackloss" (do R base), que apresenta dados de 21 dias de operação de um indústria que realiza oxidação de amônia (NH_3) em ácido nítrico (HNO_3). O ácido nítrico produzido é absorvido na torre de absorção contracorrente. As informações disponíveis na base de dados referem-se a:

-**Air fow**: que representa a taxa de operação da indústria (corrente de ar refrigerado);
-**Water Temp**: é a temperatura de resfriamento da água que circula nos canos da torre de absorção;
-**Acid.Conc.**: é a concentração do ácido [em porcentagem, após tratamento]; e
-**stack.loss** (variável dependente) é o percentual (após tratamento) de amônia introduzida no processo industrial que escapa da absorção (representando uma medida(inversa) de eficiência total da indústria).

Com base nestes dados, objetiva-se:

1. Ajustar um modelo linear múltiplo completo para estes dados. Avaliando as estimativas dos parâmetros, os resíduos e a influência das observações no ajuste do modelo, incluindo leverage, distância de Cook, DFBETAs, DFFITs e COVRATIOs.

2. Avaliar a partir de regressão parcial e dos resíduos parciais as variáveis no modelo, bem como o pressuposto de normalidade do resíduos.

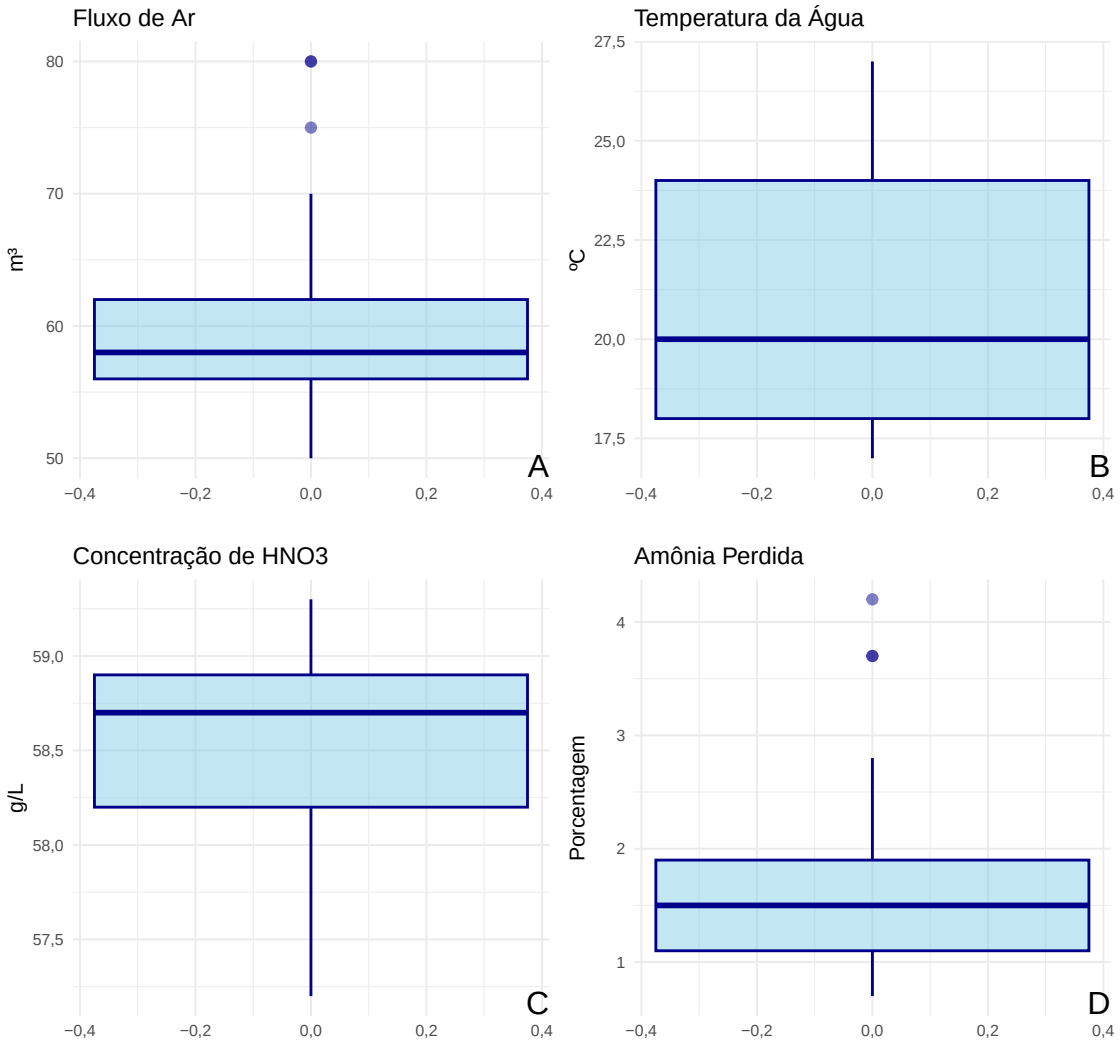
Resultados

Análise descritiva dos dados

Tabela 1: Medidas Resumo dos dados

	Mín	Q1	Med	Média	Q3	Máx	Desv.padrão	CV	Assimetria	Curtose
Amônia Perdida	0,7	1,1	1,5	1,75	1,9	4,2	1,02	0,58	1,16	0,13
Concentração de HNO3	57,2	58,2	58,7	58,63	58,9	59,3	0,54	0,01	-0,87	0,19
Fluxo de Ar	50,0	56,0	58,0	60,43	62,0	80,0	9,17	0,15	0,81	-0,26
Temperatura da Água	17,0	18,0	20,0	21,10	24,0	27,0	3,16	0,15	0,47	-1,23

Figura 1: BoxPlot das variáveis em análise.



Modelo de Regressão Linear Multipla

O modelo obtido pode ser representado por:

Y_i =3,614 + 0,072 X_{1i}+ 0,13 X_{2i} -0,152 X_{3i}

Onde:
Y_i - Amônia Perdida;
X_{1i} - Fluxo de Ar;
X_{2i} - Temperatura da Água;
X_{3i} - Concentração de HNO3;

Interpretando-se o modelo pode-se dizer que para cada variável, fixadas as demais condições (*Ceteris Paribus*), temos que a porcentagem de amônia perdida é de 3,614% caso todas as demais varíáveis tenham valor zero. Há um aumento de 0,072% na perda de amônia para cada metro cúbico de ar introduzido. O aumento de cada grau Ceusius da temperatura

da água provoca uma aumento de 0,13% de aumento na perda de amônia. O aumento em 1g/L na concentração do ácido nítrico reduz em 0,152% a perda de amônia. Neste modelo o coeficiente de determinação calculado foi de $R^2 = 0{,}914$, o que denota que 91,4% da variância dos dados é explicada pelo modelo. Pode-se calcular o coeficiente de determinação ajustado igual a $R_a^2 = 0{,}808$.

Significância do Modelo

Após o ajuste do modelo existe a necessidade de se avaliar a significância do mesmo, o teste de hipótese para tal situação será realizado, contendo as seguintes hipóteses:

$$H_0 : \hat{\beta}_1 = 0$$
$$H_1 : \hat{\beta}_1 \neq 0.$$

As Tabelas 2 e 3 trazem os principais resultados da tabela ANOVA e do Intervalo de Confiança para os parâmetros, possibilitando assim inferir sobre o modelo ajustado.

Tabela 2: Análise de Variância (ANOVA)

	GL[1]	Soma de Quadrados	Quadrado Médio	Estatística F-Snedecor
Fluxo de Ar	1	17,501	17,501	166,3707
Temperatura da Água	1	1,303	1,303	12,3886
Concentração de HNO3	1	0,100	0,100	0,9473
Resíduos	17	1,788	0,105	

Legenda:
[1] GL: Graus de Liberdade

Tabela 3: Intervalos de Confiança para os parâmetros estimados no MRLS.

	LI[1]	LS[2]
$\hat{\beta}_0$	-15,168	22,396
$\hat{\beta}_1$	0,043	0,100
$\hat{\beta}_2$	0,052	0,207
$\hat{\beta}_3$	-0,482	0,178

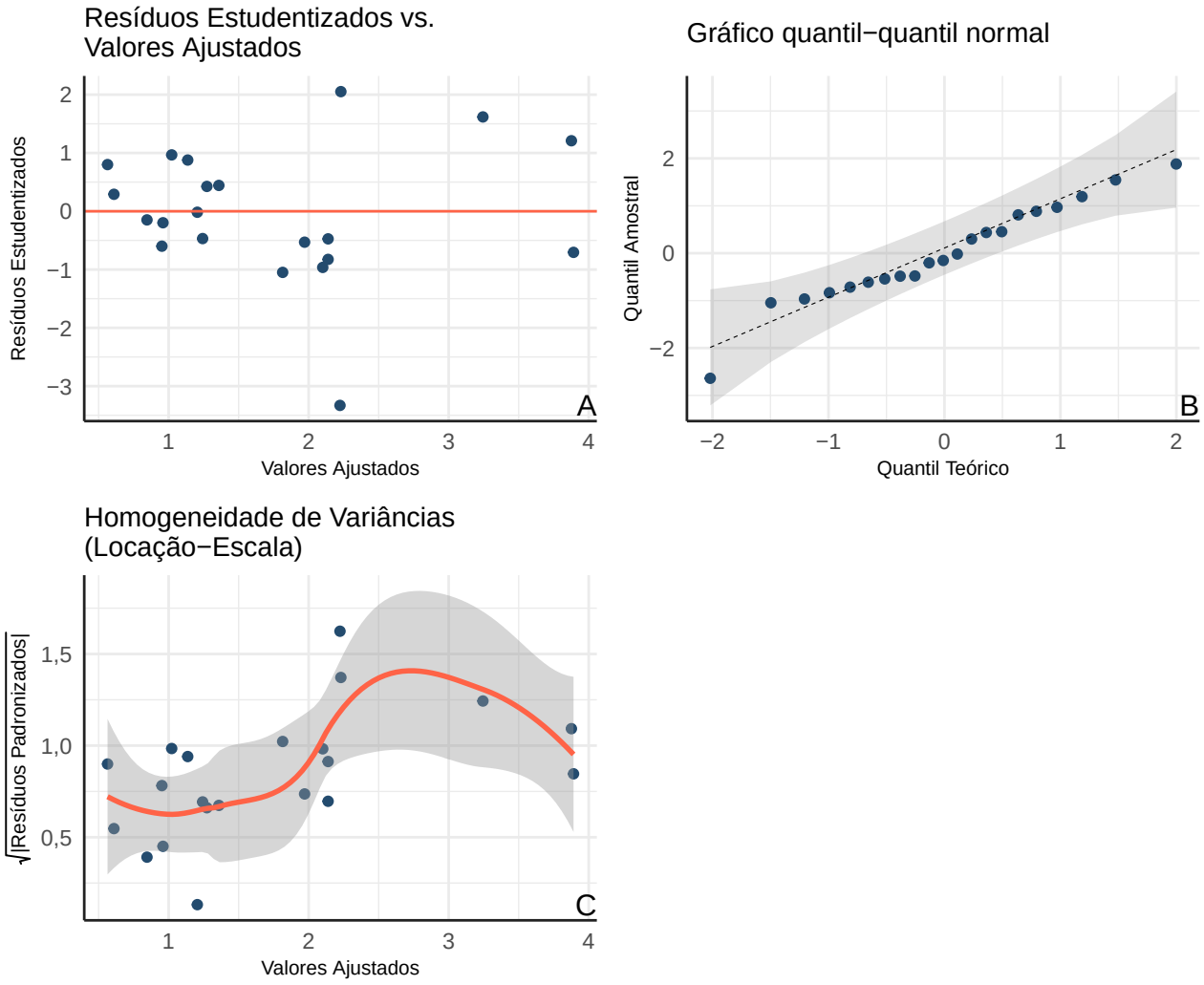
Legenda:
[1] LI: Limite Inferior (2,5%)
[2] LS: Limite Superior (97,5%)
[*] Nível de Significância de 5%.

Com base na Tabela 2, avaliando o p-valor é possível afirmar que o modelo é significante rejeitando assim H_0 que tem como pressuposto $\hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3 = 0$.

Através dos Intervalos de Confiança calculados (Tabela 4) é possível afirmar com 95% de confiança que o verdadeiro valor de β_0 está entre (-15,1677; 22,3960); que o verdadeiro valor de β_1 está entre (0,0431; 0,1000); que o verdadeiro valor de β_2 está entre (0,0519; 0,2072); e que o verdadeiro valor de β_3 está entre (-0,4819; 0,1776).

Análise de Resíduos

Figura 2: Análise de resíduos do modelo ajustado



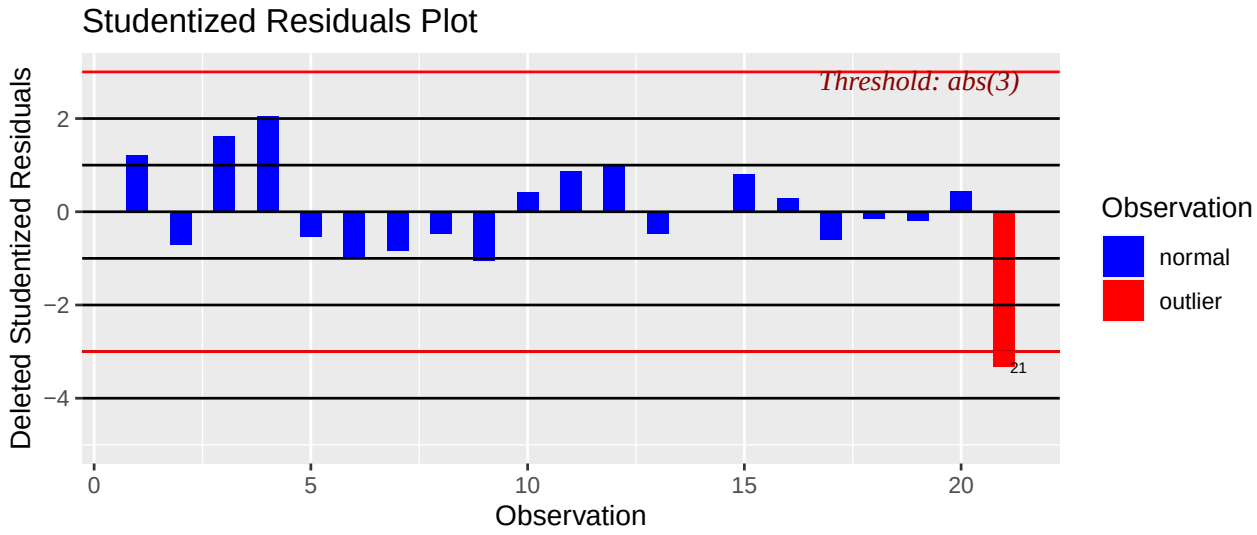
A Figura 2A apresenta um comportamento simétrico dos resíduos, podendo ser constatado uma pequena variabilidade inicial e um aumento desta à medida que os valores ajustados aumentam, caracterizando uma baixa heterocedasticidade. A Figura 2C, que trata da Homogeneidade de Variâncias (Locação-Escala) ressalta que há um problema na variabilidade dos dados, ampliando a interpretação feita na análise da Figura 2A, de que há uma mudança na variabilidade dos dados, caracterizando uma certa heterocedasticidade dos dados. A Figura 2B traz o gráfico para avaliação da normalidade dos dados, mostra

que apesar dos dados não estarem precisamente sobre a reta de referência, os mesmos estão contidos na região pertencente ao Intervalo de Confiança - IC, podendo assumir que há normalidade.

Gráficos de Diagnóstico

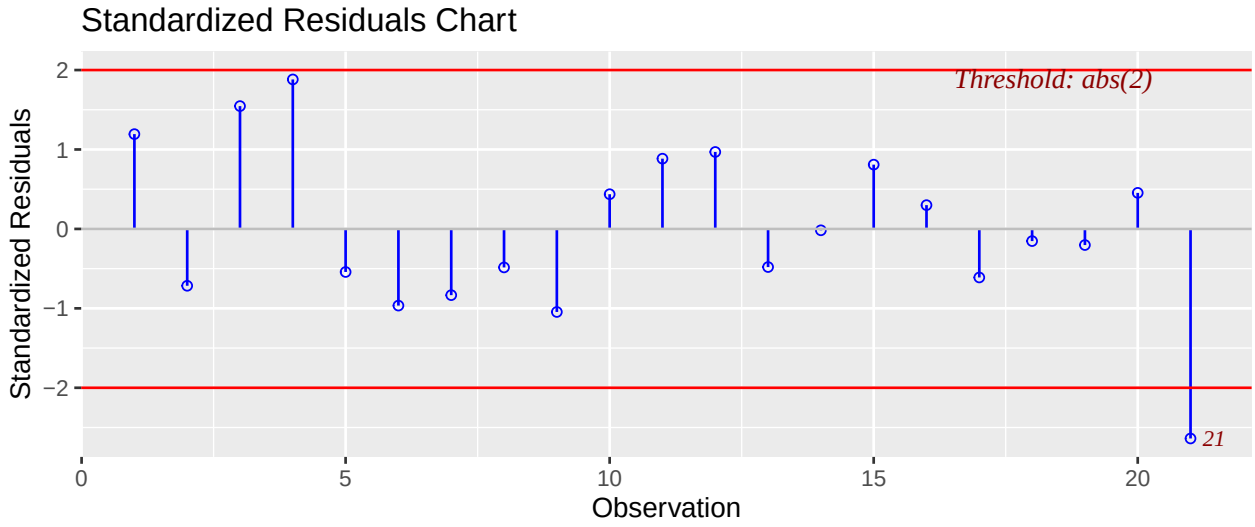
A análise dos gráficos de diagnóstico permite avaliar as observações realizadas e conhecer a influência de cada uma delas para o madelo de regressão proposto. Assim, com base no modelo, é possível fazer as seguintes análises:

Figura 2: Valores Ajustados e Resíduos Studentizados



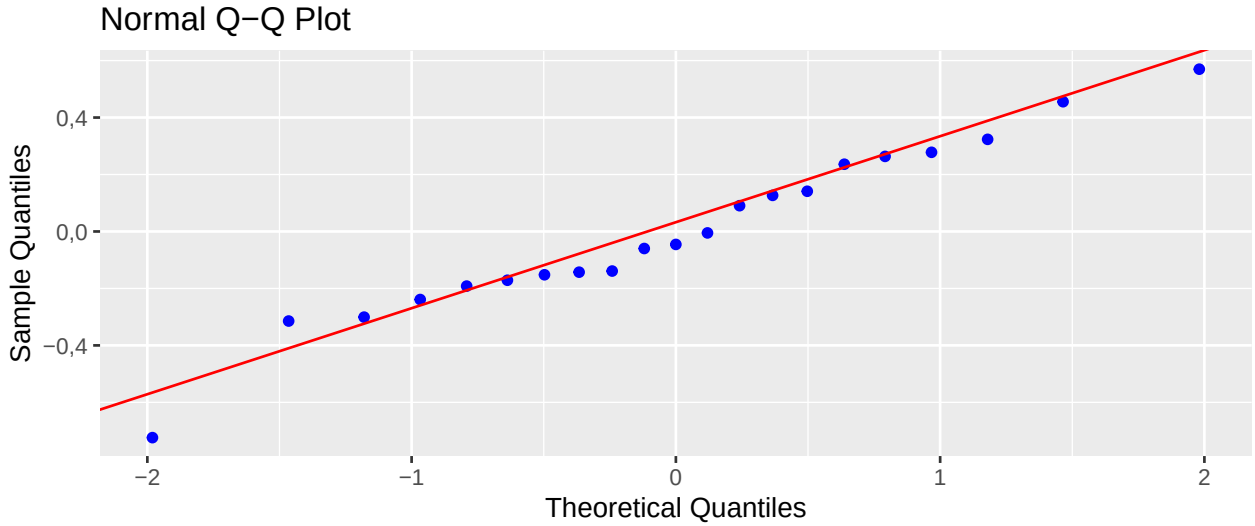
A Figura 2 demonstra que os resíduos estão todos dentro dos limites esperados, com exceção da observação 7 que por pouco ultrapassou o limite inferior. Não parece ser o caso de nenhuma intervenção por conta deste valor.

Figura 3: Valores Ajustados e Resíduos Padronizados.



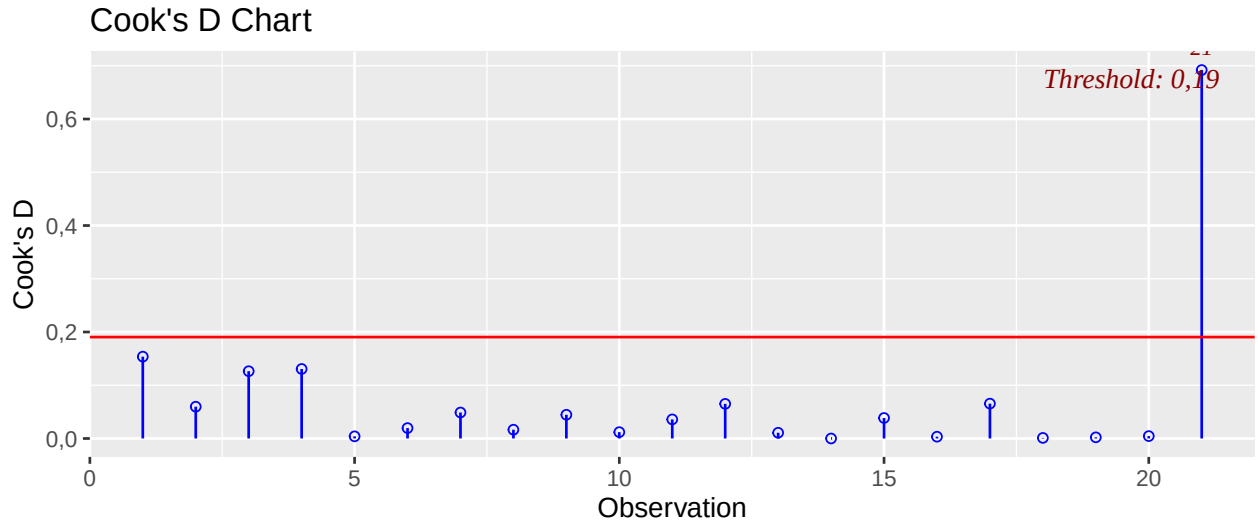
Já a análise da Figura 3, onde os resíduos foram padronizados, o número de observações que ultrapassaram os limites chegou a 4,8% do total o que é condizente com uma confiança de 95%.

Figura 4: Análise dos quantis teóricos e amostrais



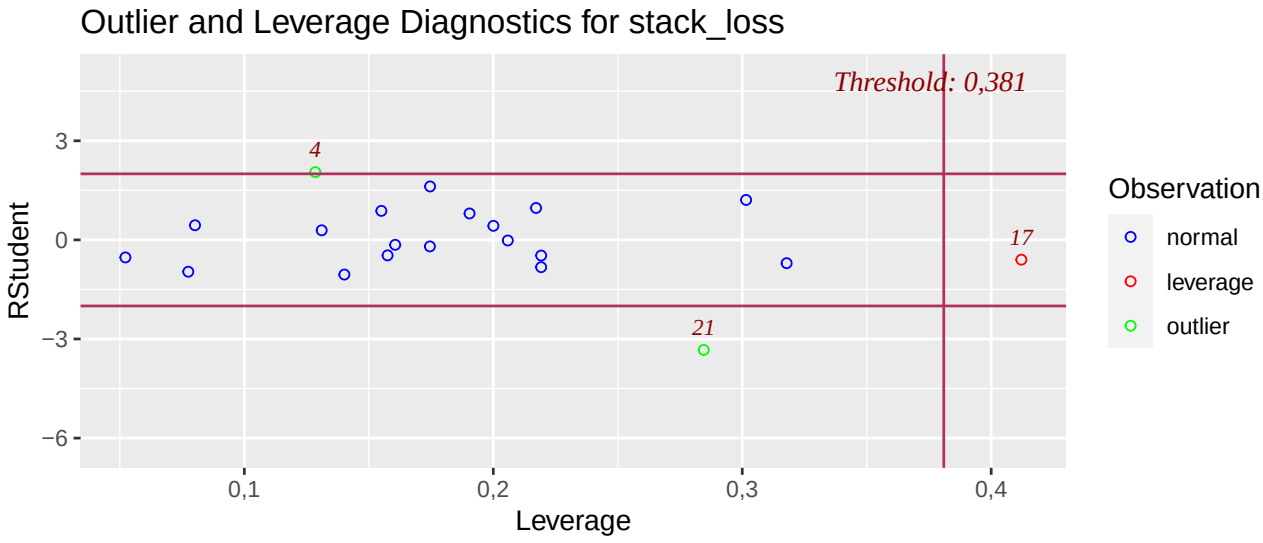
A Figura 4 apresentou um bom ajuste dos resíduos à distribuição normal, sendo um pouco pior nas caudas da distribuição por uma qauntidade pequena de pontos.

Figura 5: Distância de Cook.



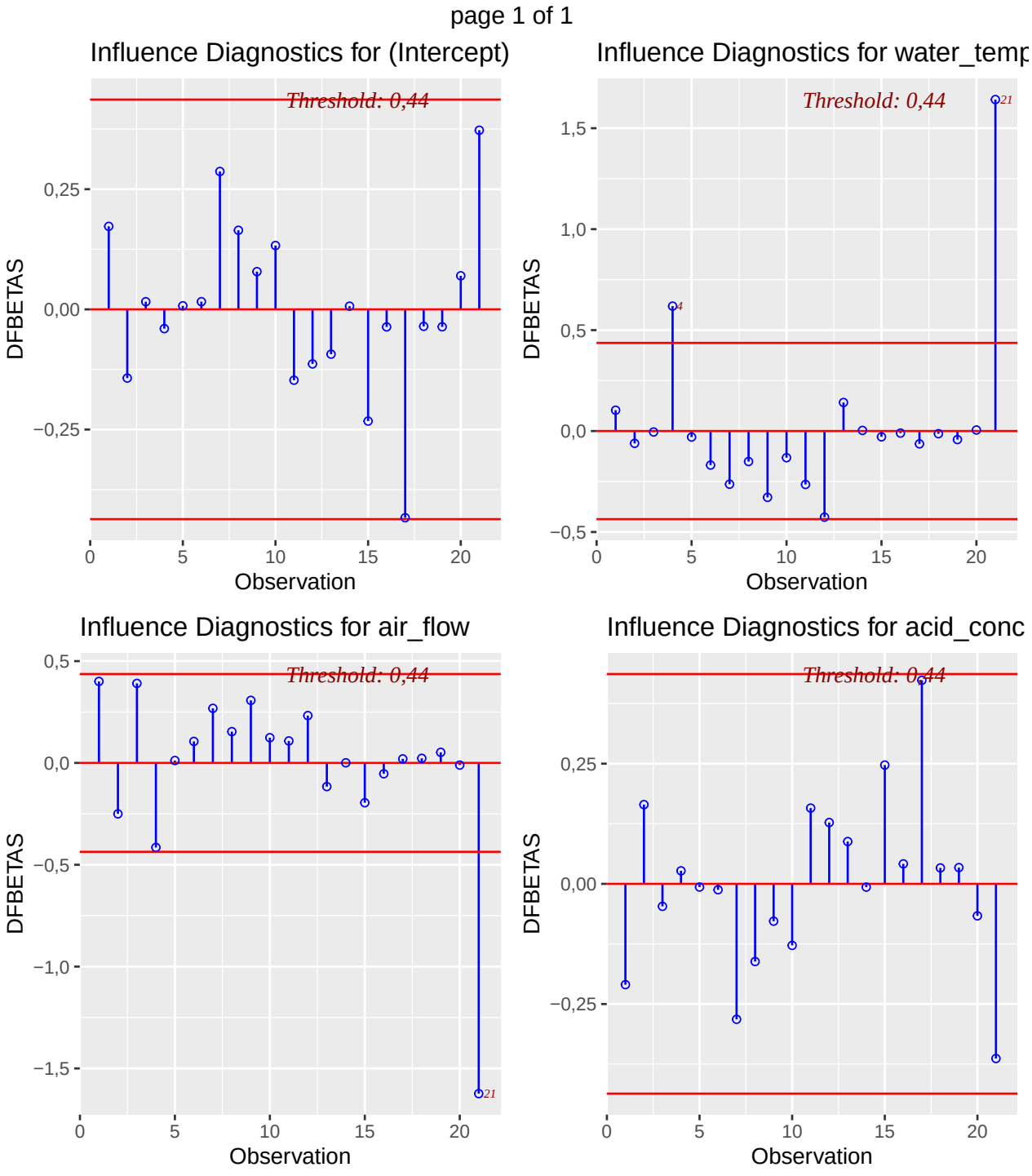
A analise da distância de Cook apresentada na Figura 5 demonstra que 25 (6,6%) observações tem uma distância expressiva, mas apenas sete deles estão acima da distância de 0,025. O tretamento destes pontos pode manter os resíduos dentro do esperado com uma confiança de 95%.

Figura 6: Análise dos pontos de Alavanca e Resíduo Studentizado.



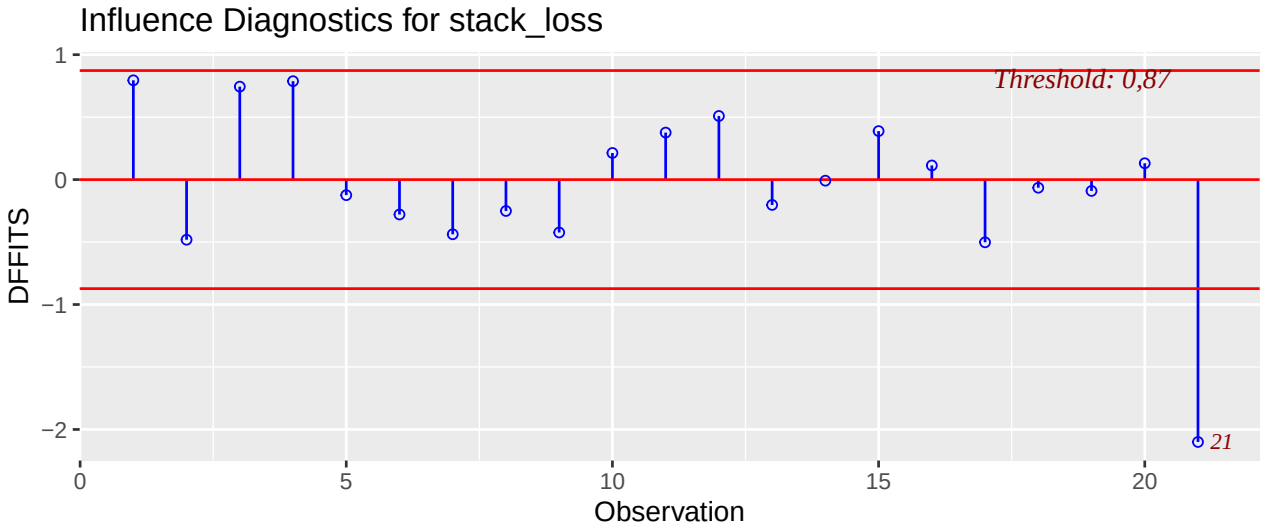
Pela Figura 6, observamos 15 observações que podem ser consideradas como *Outliers* e 21 como observações de alavanca, além de 3 com as duas características, o que representa um total de 10,3% das observações. Uma quantidade tão expressiva de dados não pode ser descartada sem o amparo de um especialista na área.

Figura 7: DFBetas para as variáveis do modelo.



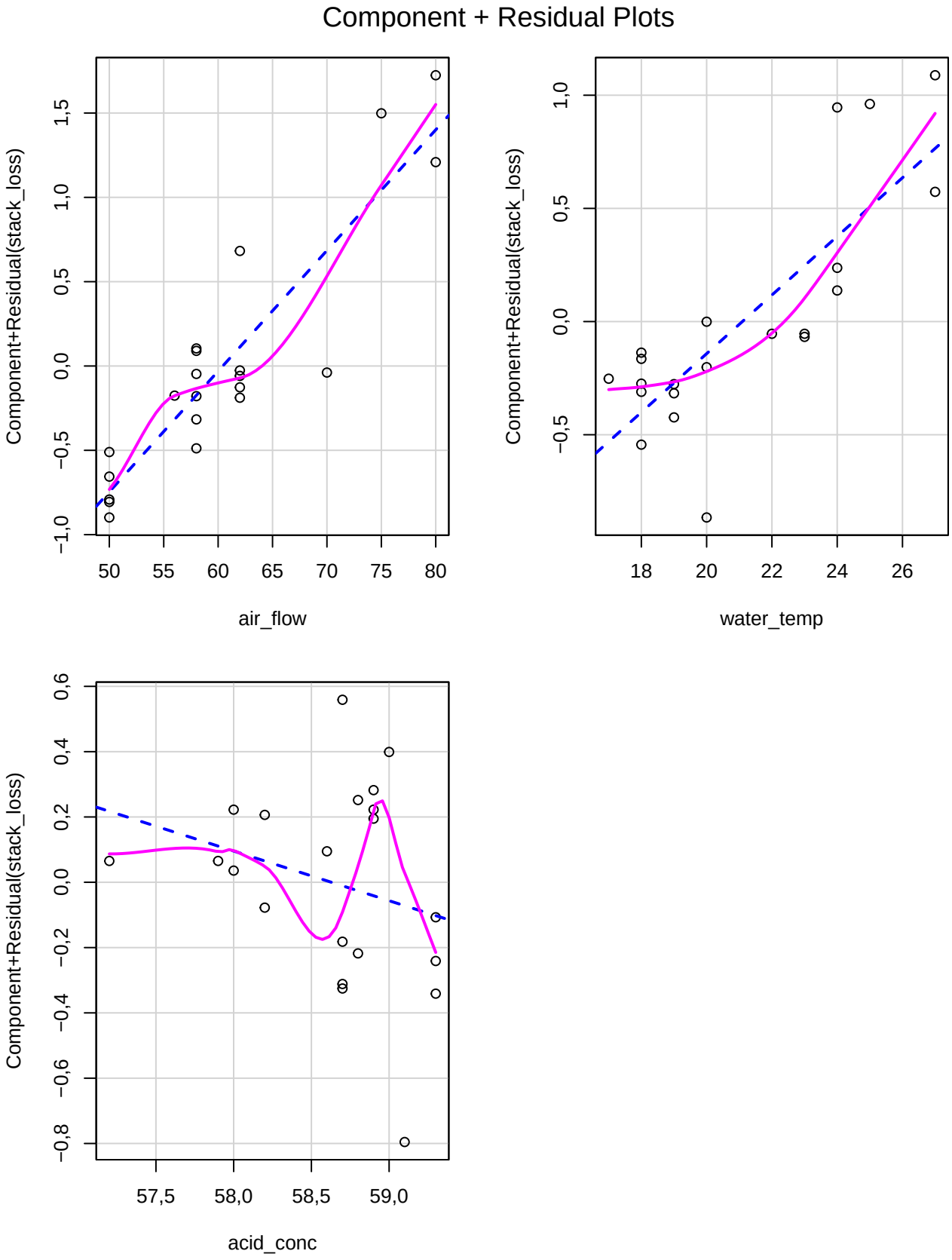
A Figura 7 apresenta os DFBetas para cada uma das variáveis utilizadas no modelo, com uma média de 6,1% de observações discrepantes com cerca de 4 observações críticas, isto é, valores mais extemos. O tratamento destas observações podem trazer o modelo para uma situação mais compatível com a confiança estabelecida.

Figura 8: DfFit para as variáveis do modelo.



A Figura 8 acompanha os gráficos anteriores apresentando 6,9% de observações discrepantes, mas apenas seis valores são extremos, desta forma pode-se da mesma forma tratá-los e manter a confiança do modelo.

Figura 9: COVRatio para as variáveis do modelo.

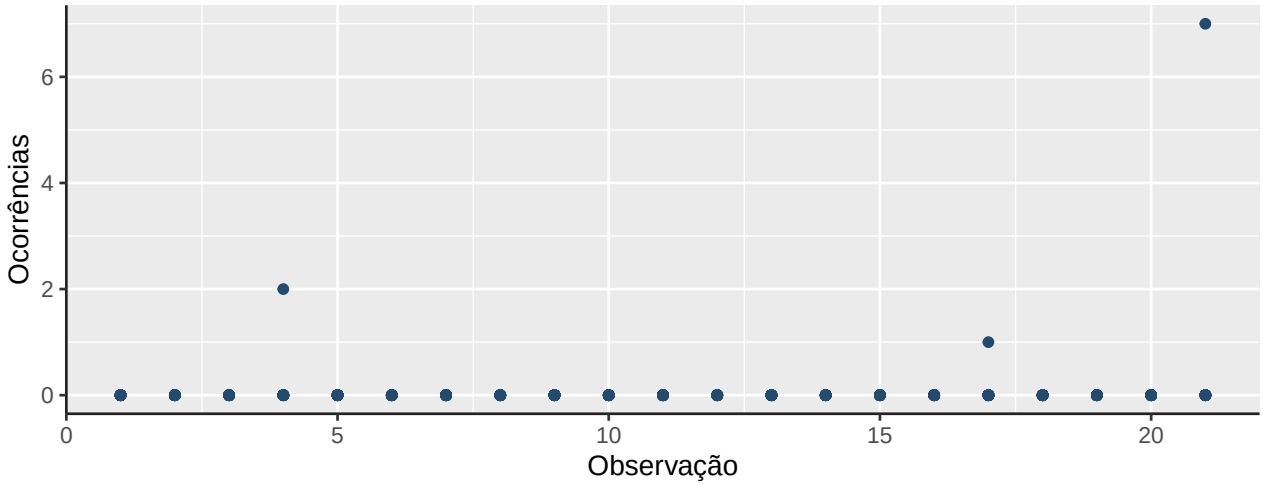

Component + Residual Plots
Component+Residual(stack_loss)
air_flow
water_temp
acid_conc

Da Figura 9 verifica-se que as variáveis “Altura”, “Cintura” e “Quadil” estão mais diretamente correlacionadas com os resíduos do “Peso”, o que por sua vez indica que a inclusão de observações destas variáveis apresentam maior impacto ao modelo.

Eliminação de observações anômalas

Avaliando as observações que apresentaram comportamento anômalo nos diagnósticos dos valores ajustados e resíduos studentizados, valores ajustados e resíduos padronizados, distância de CooK, pontos de alavanca e *outliers*, análise de DfFit e todas as análises de BFBetas, chegamos as frequências de observações anômalas apresentadas na Figura 10.

Figura 10: Número de ocorrências para cada observação



Considerando apenas as observações com 1 ou mais ocorrências temos a Tabela a seguir.

Tabela 4: Observações com maior número de ocorrências.

Observação	Ocorrências
4	2
17	1
21	7

Podemos intuir que essas são as observações com maior impacto negativo no modelo. Logo, eliminando-as do conjunto de dados analizados chegamos a um novo modelo dado por:

$Y_i^{¤} = 1{,}47 + 0{,}096\ X_{1i}^{¤} + 0{,}056\ X_{2i}^{¤} - 0{,}114\ X_{3i}^{¤}$

Onde:
$Y_i^{¤}$ - Peso;
$X_{1i}^{¤}$ - Colesterol total;

$X_{2i}^{¤}$ - Glicose estabilizada;
$X_{3i}^{¤}$ - Lipoproteína de alta densidade;

Neste novo modelo o coeficiente de determinação calculado foi de $R^2 = 0{,}968$, o que denota que 96,8% da variância dos dados é explicada pelo modelo. O valor deste novo coeficiente permite concluir que a eliminação das observações com maior impacto no modelo foi benéfica. Pode-se calcular o coeficiente de determinação ajustado igual a $R_a^2 = 0{,}909$

Conclusões

Verificou-se que a análise de variâncias foi um teste mais poderoso para identificar variáveis desnecessárias ao modelo que a analise individual das significancias das variáveis ao modelo.

Embora se não se tenha um conhecimento específico da área estudada, foi possível realizar uma avaliação dos dados recebidos e propor um tratamento que efetivamente melhorou o modelo de regressão linear multipla realizado.

As anomalias relatadas em cada um dos gráficos de diagnóstico elaborados foram tratadas de igual maneira contabilizando para cada observação o número de ocorrências observadas. Por este método se elencou as observações com maior potencial de prejuízo ao modelo e ao descartá-las do rol de dados avaliados obteve-se uma expressiva melhora no modelo.