

Produção de Amônia

Fernado Bispo, Jeff Caponero

Sumário

Apresentação	2
Atividade 1	3
Introdução	3
Resultados	3
Análise descritiva dos dados	3
Modelo de Regressão Linear Multipla	4
Significância do Modelo	5
Análise de Resíduos	6
Gráficos de Diagnóstico	7
Eliminação de observações anômalas	13
Conclusões	14
Atividade 2	15
Introdução	15
Resultados	16
Análise descritiva dos dados	16
Modelo de Regressão Linear Multipla	17

Apresentação

O relatório desta semana está dividido em duas atividades.

Atividade 1

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 flow**: 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.

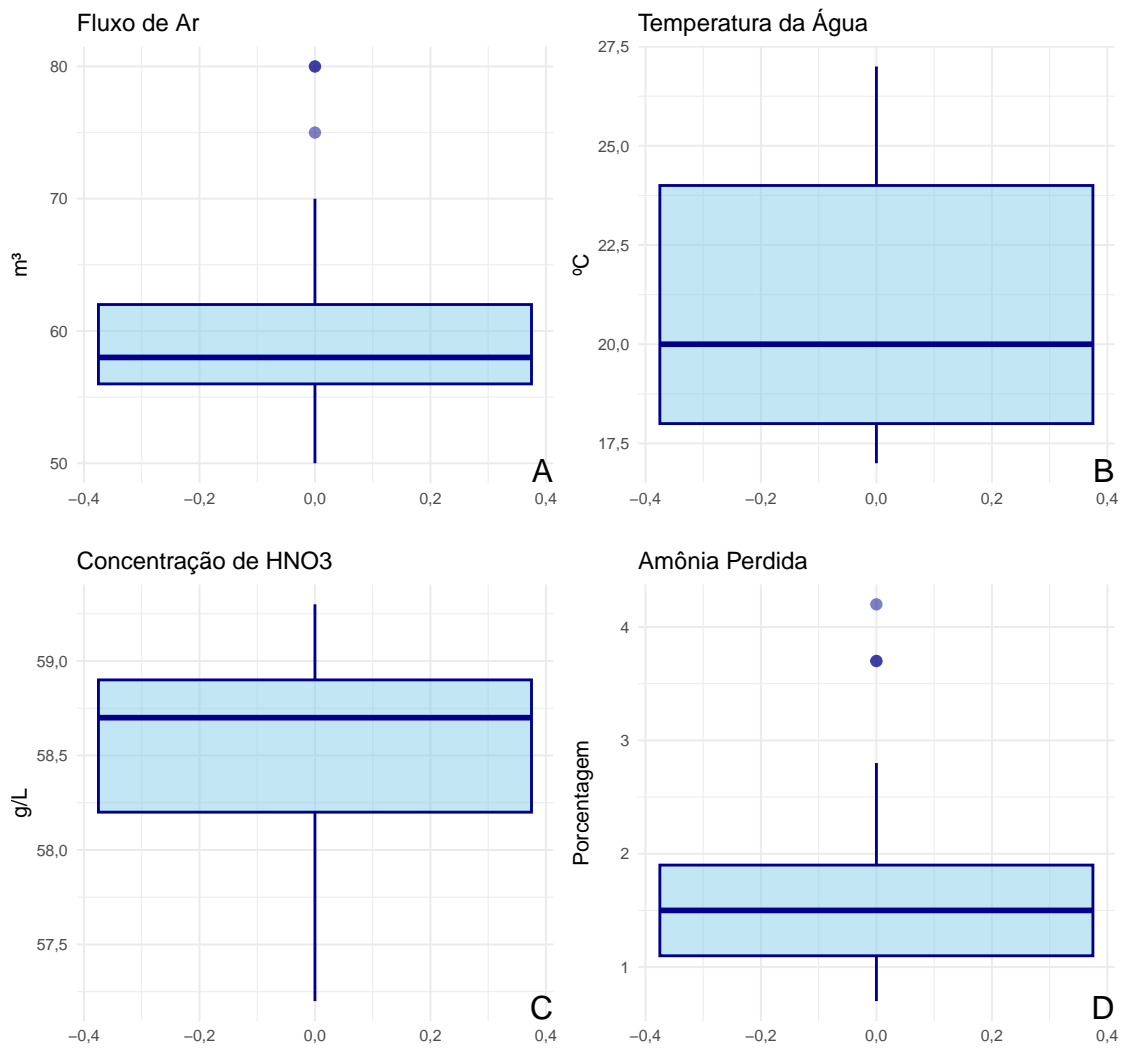
Resultados

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 HNO_3	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 Múltipla

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 variáveis 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 Celsius da temperatura da água provoca um 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%)

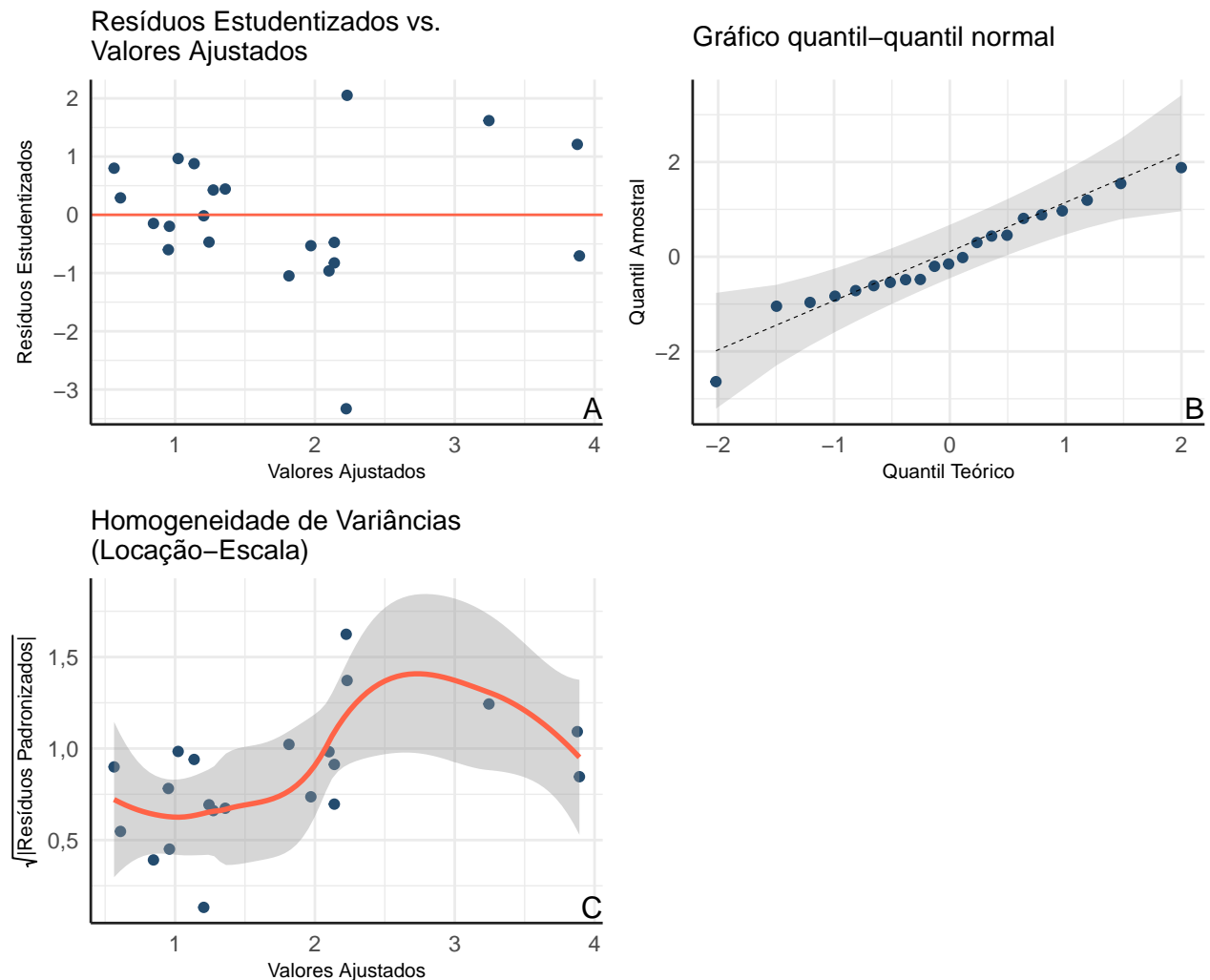
* Nível de Significância de 5%.

Com base na Tabela 2, avaliando o p-valor é possível afirmar que o modelo é significativo 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



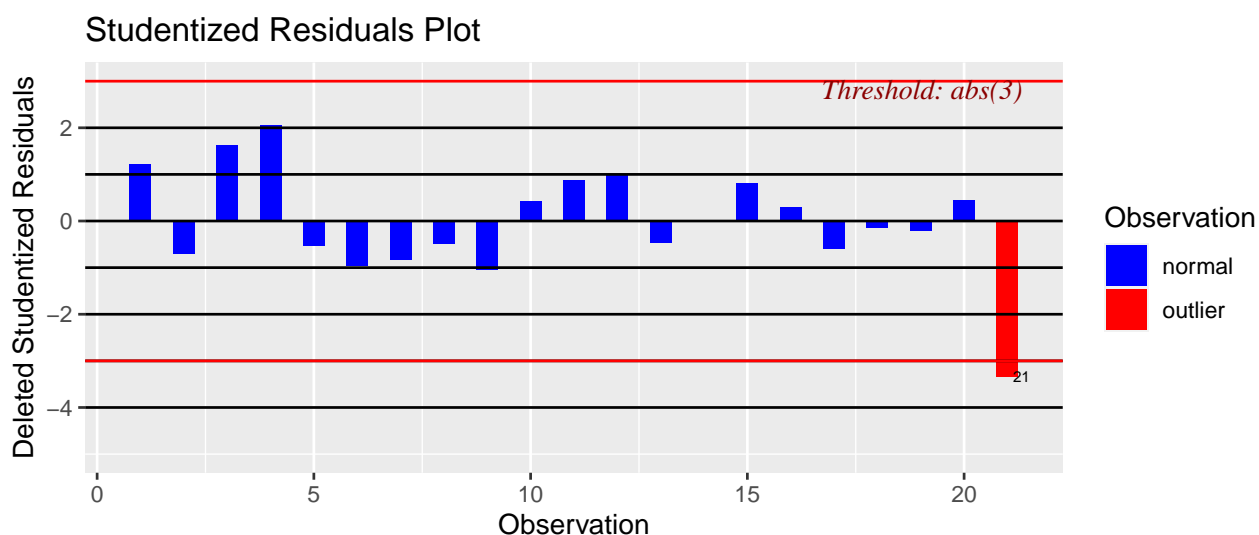
A Figura 2A apresenta um comportamento simétrico dos resíduos, podendo ser constatado uma pequena variabilidade inicial e um aumento desta à medida que os valores ajustados aumentam, caracterizando uma baixa heterocedasticidade. A Figura 2C, que trata da Homogeneidade de Variâncias (Locação-Escala) ressalta 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

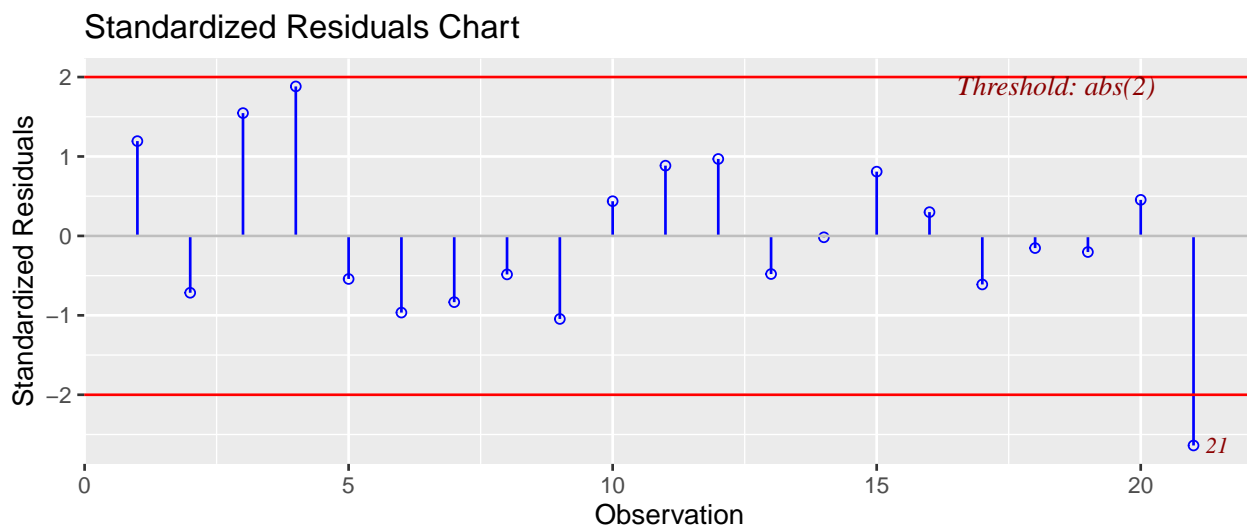
A análise dos gráficos de diagnóstico permite avaliar as observações realizadas e conhecer a influência de cada uma delas para o modelo de regressão proposto. Assim, com base no modelo, é possível fazer as seguintes análises:

Figura 3: Valores Ajustados e Resíduos Studentizados



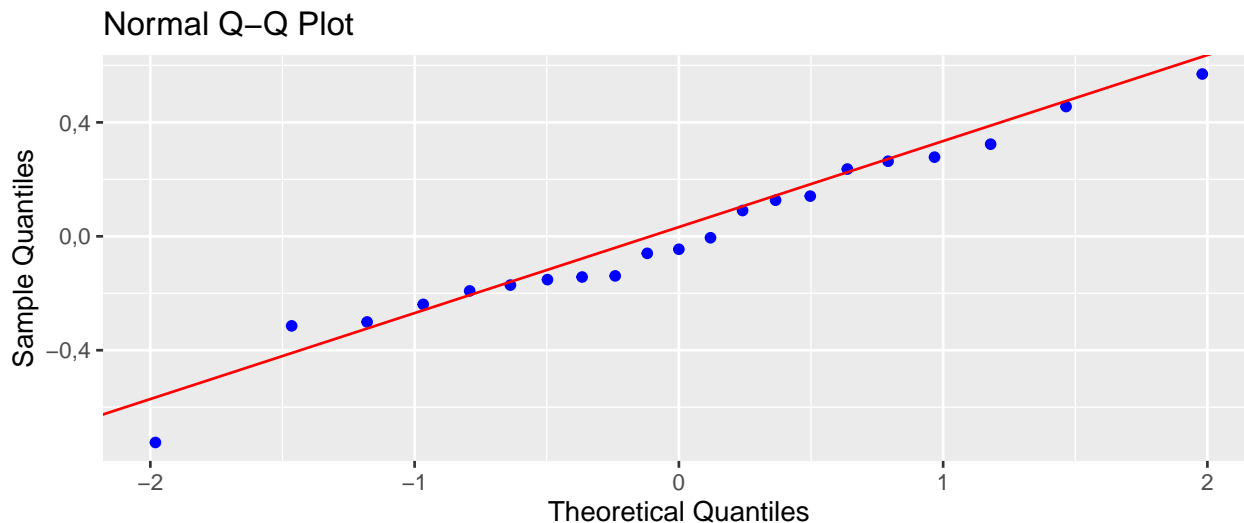
A Figura 3 demonstra que os resíduos estão todos dentro dos limites esperados, com exceção da observação 21 que por pouco ultrapassou o limite inferior. Não parece ser o caso de nenhuma intervenção por conta deste valor.

Figura 4: Valores Ajustados e Resíduos Padronizados.



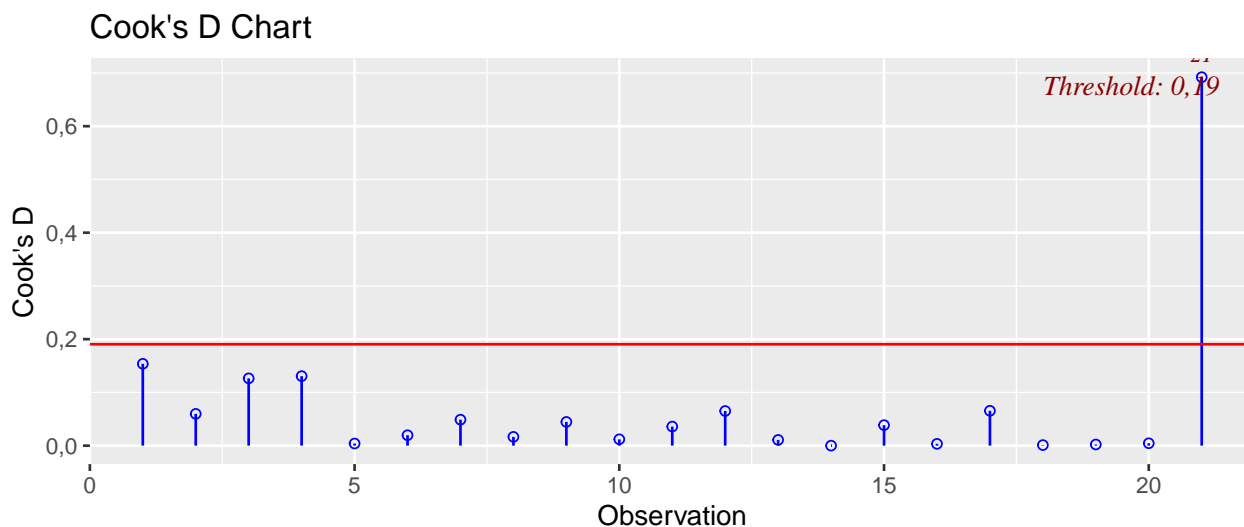
Na análise da Figura 4, onde os resíduos foram padronizados, verifica-se que apenas a observação 21 está fora do limite de aceitação.

Figura 5: Análise dos quantis teóricos e amostrais



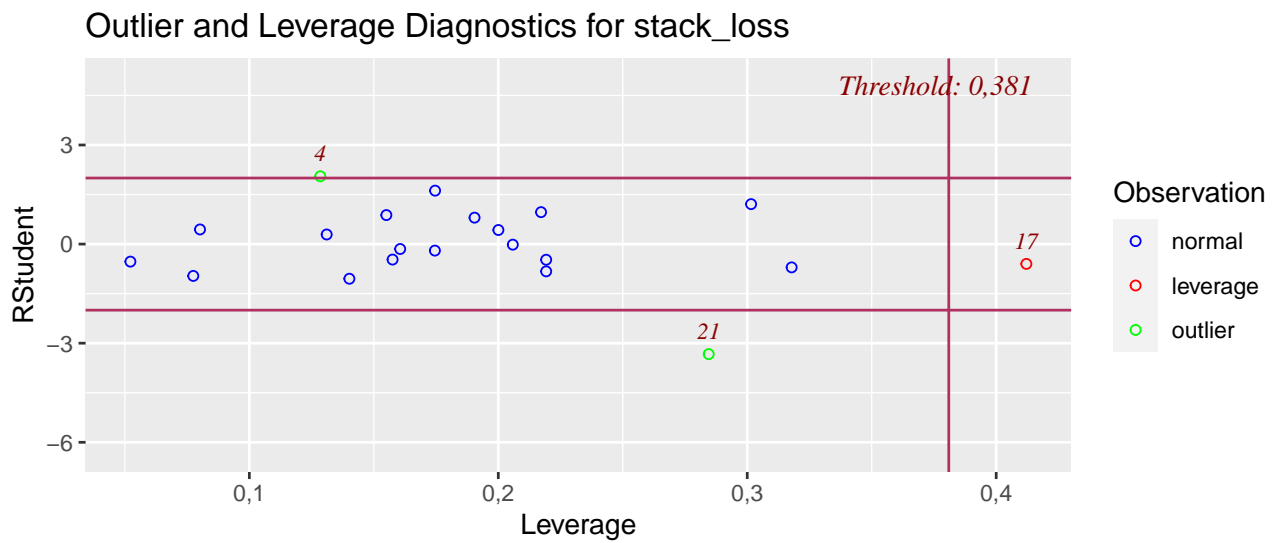
A Figura 5 apresentou um bom ajuste dos resíduos à distribuição normal, sendo apenas um pouco pior nas caudas da distribuição.

Figura 6: Distância de Cook.



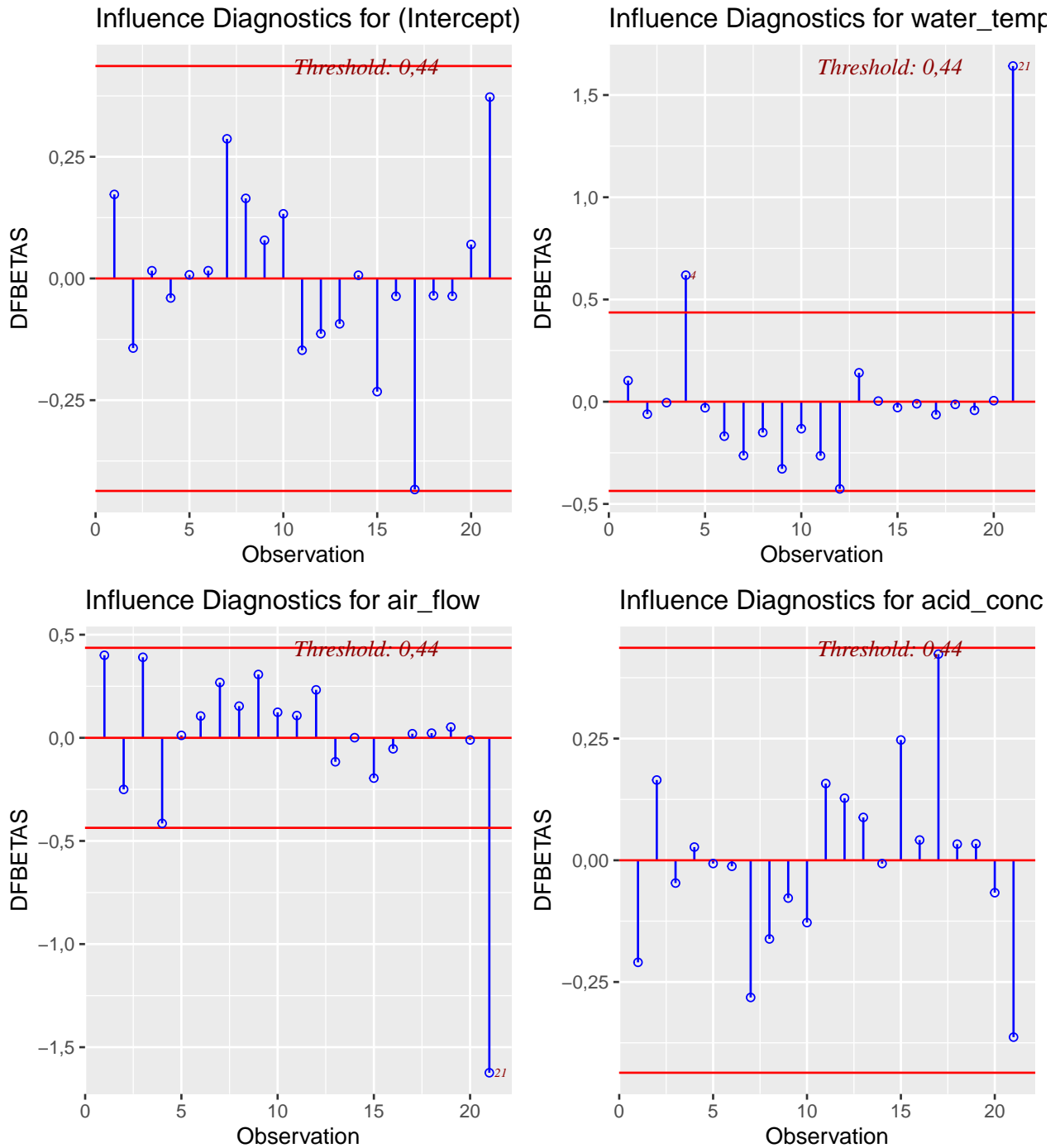
A análise da distância de Cook apresentada na Figura 6 demonstra que novamente apenas a observação 21 destoa do conjunto de observações e tem uma distância expressiva.

Figura 7: Análise dos pontos de Alavanca e Resíduo Studentizado.



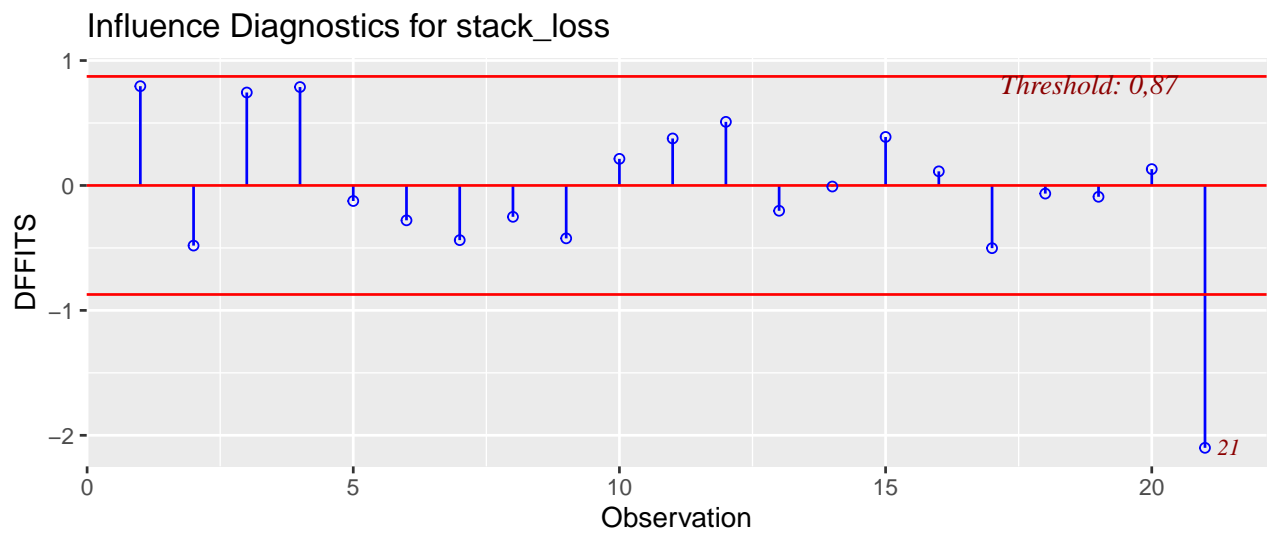
Pela Figura 7, observamos 2 observações que podem ser consideradas como *Outliers* e uma como ponto de alavanca. Interessante notar que apesar da observação 21 ter aparecido nos gráficos anteriores como uma observação anômala, aqui ela é classificada como um *outlier* o que denota uma influência menos prejudicial que a da observação 17 que aparece pela primeira vez.

Figura 8: DFBetas para as variáveis do modelo.



A Figura 8 apresenta os DFBetas para cada uma das variáveis utilizadas no modelo. Nota-se que mais uma vez a observação 21 tem comportamento anômalo.

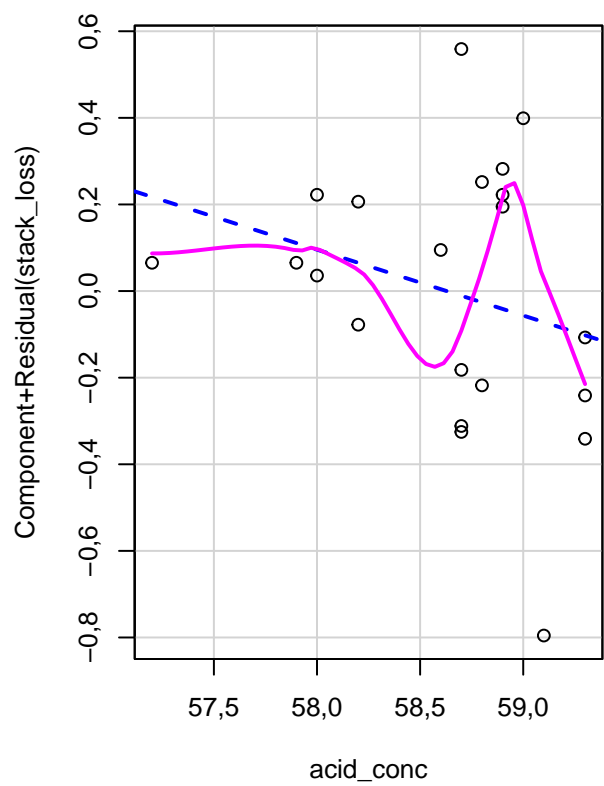
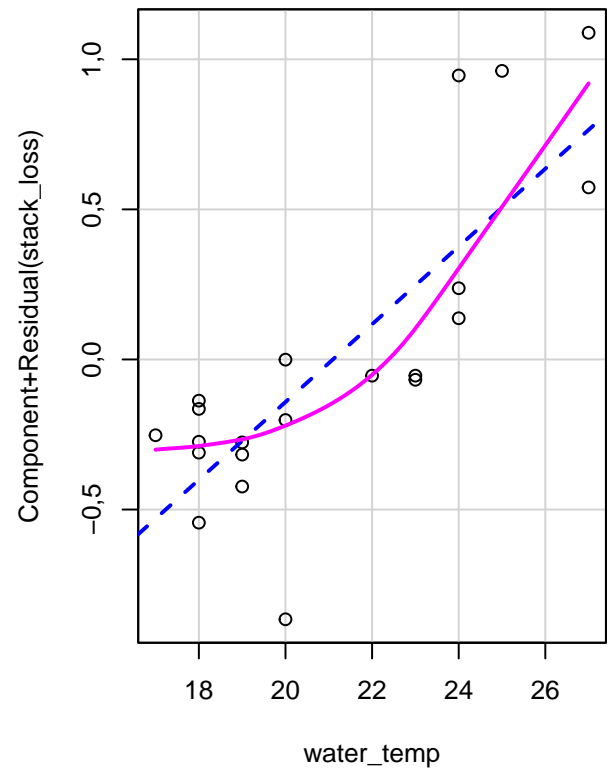
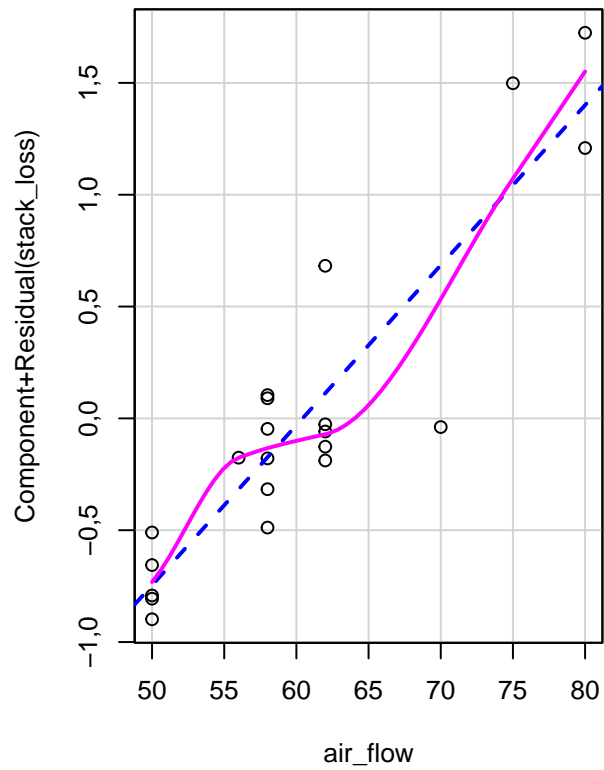
Figura 9: DfFit para as variáveis do modelo.



A Figura 9 acompanha os gráficos anteriores apresentando mais uma vez a observação 21 como discrepante.

Figura 10: COVRatio para as variáveis do modelo.

Component + Residual Plots

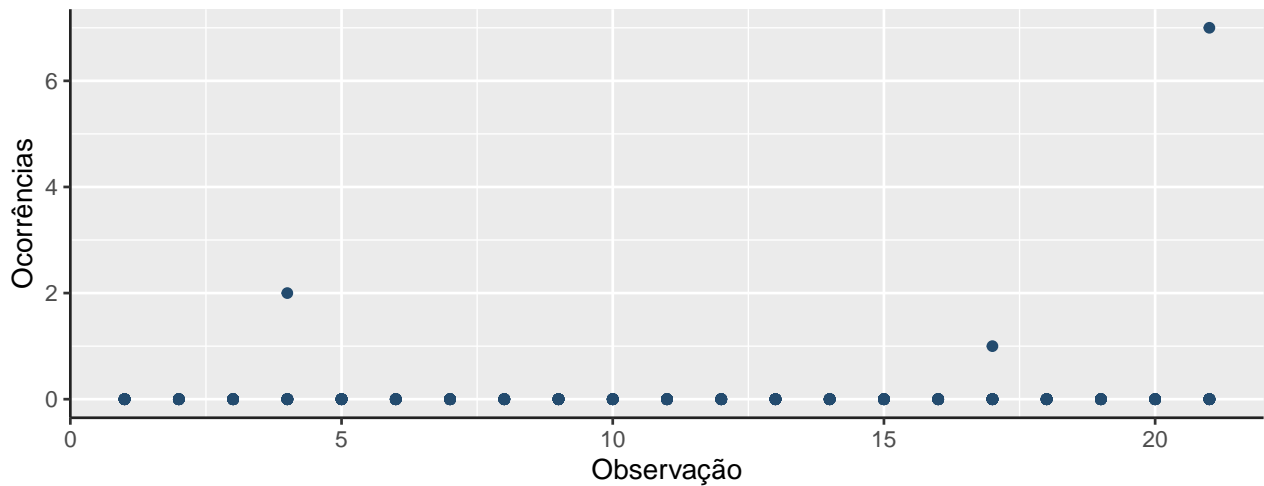


Da Figura 10 verifica-se que as variáveis estão diretamente correlacionadas com os resíduos da Amônia Perdida, o que por sua vez indica que a inclusão de observações destas variáveis apresentam bom impacto ao modelo.

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 11.

Figura 11: 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 analisados 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^* - Amônia Perdida;

X_{1i}^* - Fluxo de Ar;

X_{2i}^* - Temperatura da Água;
 X_{3i}^* - Concentração de HNO₃;

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

Atividade 2

Introdução

Para análise dos dados diários sobre evaporação do solo (EVAP), Freund (1979) identificou as seguintes variáveis preditoras:

- MAXAT** - temperatura do ar diária máxima;
- MINAT** - temperatura do ar diária mínima;
- AVAT** - medida de temperatura média do ar;
- MAXST** - temperatura máxima diária do solo;
- MINST** - temperatura mínima diária do solo;
- AVST** - medida de temperatura média do solo;
- MAXH** - umidade relativa diária máxima;
- MINH** - umidade relativa diária mínima;
- AVH** - umidade relativa diária média;
- WIND** - vento total, medido em milhas por dia.

Com base nestes dados, objetiva-se:

1. Ajustar um modelo completo sobre evaporação do solo (EVAP), definindo os fatores significativamente associados, o coeficiente de determinação, a avaliação da bondade do modelo completo.
2. Determinar a correlação entre todos os preditores e a resposta. Realizando uma análise dos resíduos e dos pontos de alavanca e influentes.
3. Investigar possíveis problemas de colinearidade, avaliando o R_j^2 para cada preditor e seu VIF.
4. Melhorar o modelo pela exclusão de observações anômalas e pela definição de variáveis preditoras a serem incluídas no modelo.

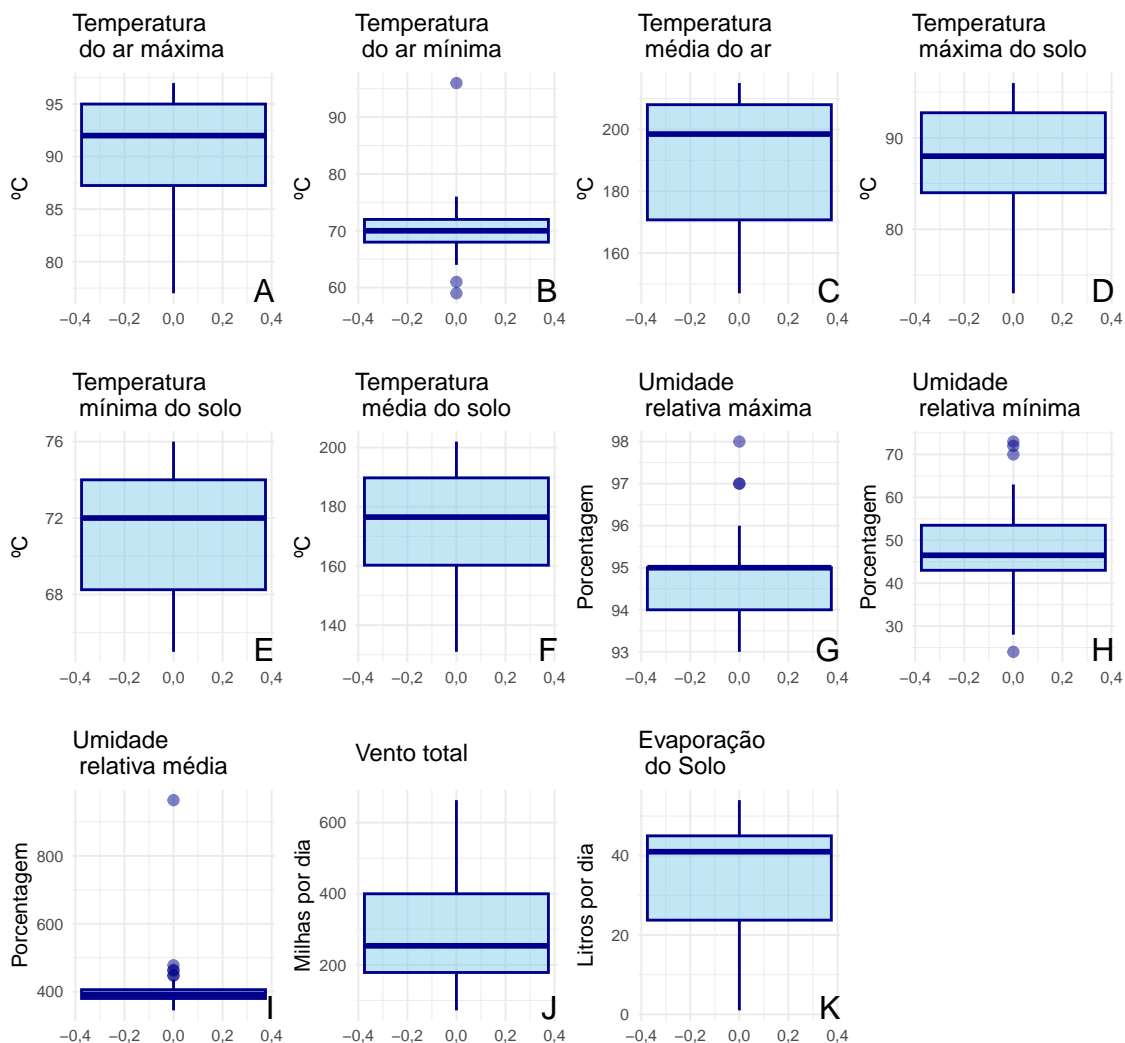
Resultados

Análise descritiva dos dados

Tabela 5: Medidas Resumo dos dados

	Mín	Q1	Med	Média	Q3	Máx	Desv.padrão	CV	Assimetria	Curtose
Evaporação do Solo	1	23	41,0	34,67	45	54	14,64	0,42	-0,71	-0,68
Temperatura do ar máxima	77	87	92,0	90,74	95	97	5,07	0,06	-0,92	-0,05
Temperatura do ar mínima	59	68	70,0	70,54	72	96	5,25	0,07	2,10	9,97
Temperatura máxima do solo	73	84	88,0	87,57	93	96	6,06	0,07	-0,66	-0,39
Temperatura média do ar	147	170	198,5	190,50	208	215	20,97	0,11	-0,66	-1,13
Temperatura média do solo	131	160	176,5	173,52	190	202	20,08	0,12	-0,49	-0,82
Temperatura mínima do solo	65	68	72,0	71,26	74	76	3,30	0,05	-0,39	-1,15
Umidade relativa máxima	93	94	95,0	94,72	95	98	1,20	0,01	0,39	-0,27
Umidade relativa média	345	380	392,0	410,00	406	964	88,40	0,22	5,34	30,57
Umidade relativa mínima	24	43	46,5	48,50	54	73	10,02	0,21	0,41	0,48
Vento total	72	177	253,5	284,20	400	663	149,45	0,53	0,70	-0,34

Figura 1: BoxPlot das variáveis em análise.



Modelo de Regressão Linear Multipla

O modelo obtido pode ser representado por:

$$Y_i = 117,578 + 0,24 X_{1i} - 1,318 X_{2i} - 0,116 X_{3i} + 0,336 X_{4i} + 0,01 X_{5i} + 0,495 X_{6i} - 0,877 X_{7i} - 0,823 X_{8i} + 0,008 X_{9i} + 0,016 X_{10i}$$

Onde:

- Y_i - Evaporação do Solo;
- X_{1i} - Temperatura do ar máxima;
- X_{2i} - Temperatura do ar mínima;
- X_{3i} - Temperatura do ar média;
- X_{4i} - Temperatura do solo máxima;
- X_{5i} - Temperatura do solo mínima;
- X_{6i} - Temperatura do solo média;
- X_{7i} - Umidade relativa máxima;

X_{8i} - Umidade relativa mínima;

X_{9i} - Umidade relativa média;

X_{10i} - Vento Total;

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 variáveis 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 Celsius da temperatura da água provoca um 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 $R^2 = 0,798$, o que denota que 79,8% da variância dos dados é explicada pelo modelo. Pode-se calcular o coeficiente de determinação ajustado igual a $R_a^2 = 0,733$.