Perda de Amônia e Evaporação de Água do Solo

Fernado Bispo, Jeff Caponero

Sumário

Apresentação

O relatório desta semana está dividido em duas atividades. Na primeira foi analisado um banco de dados sobre uma industria que realiza a oxidação de amônia, para o qual por meio de técnicas de regressão linear múltipla se elaborou um modelo para determinar a perda de amônia no processo. Na segunda atividade, se buscou determinar a quantidade de água perdida do solo, evaporação do solo, com base em mum banco de dados sobre propriedades do solo e do ar associadas. Nesta segunda atividade também foram utilizadas técnicas de regressão linear múltipla.

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 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

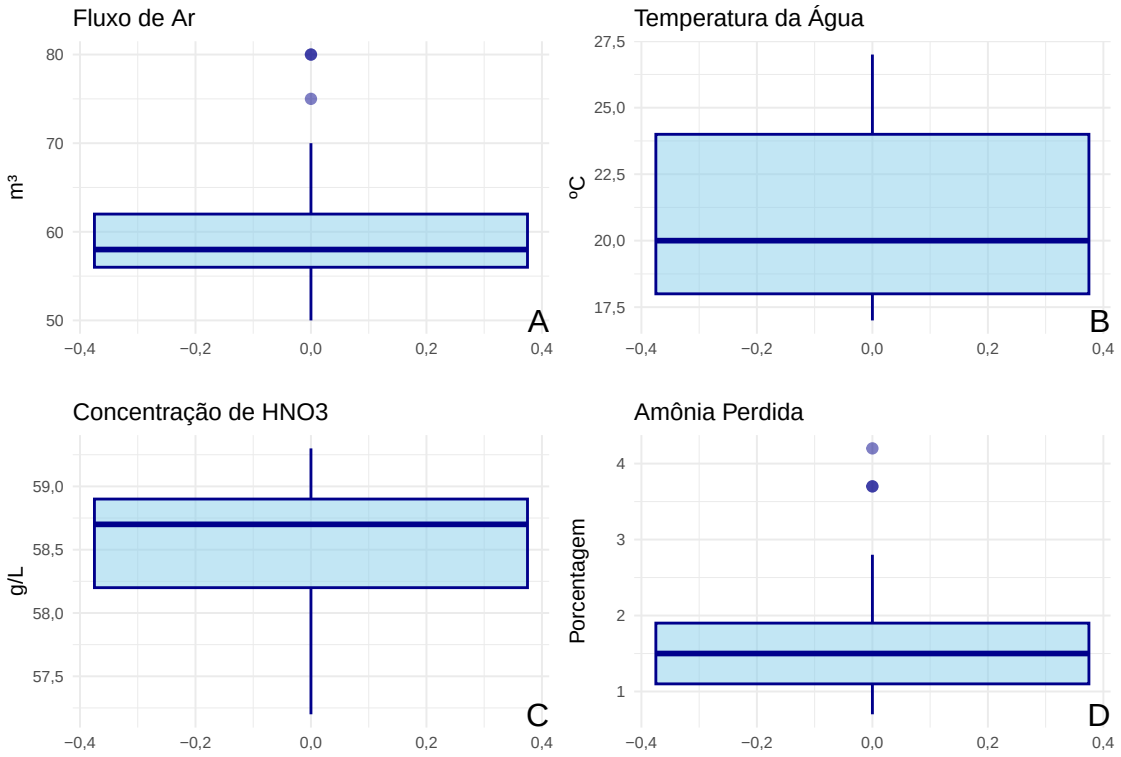
Análise descritiva dos dados

É possivel realizar uma descrição prévia dos dados por meio de medidas de resumo e de gráficos do tipo box-plot como vê-se a seguir:

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.



Nota-se uma assimetria nos dados apresentados e algumas observações que podem ser descritas como *outliers*. Entretanto é possivel propor um modelo de regressão como se segue.

Modelo de Regressão Linear Múltipla

O modelo de regressão múltipla 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 $R_a^2 = 0{,}898$.

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	p-valor
Fluxo de Ar	1	17,501	17,501	166,3707	0,0000
Temperatura da Água	1	1,303	1,303	12,3886	0,0026
Concentração de HNO3	1	0,100	0,100	0,9473	0,3440
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

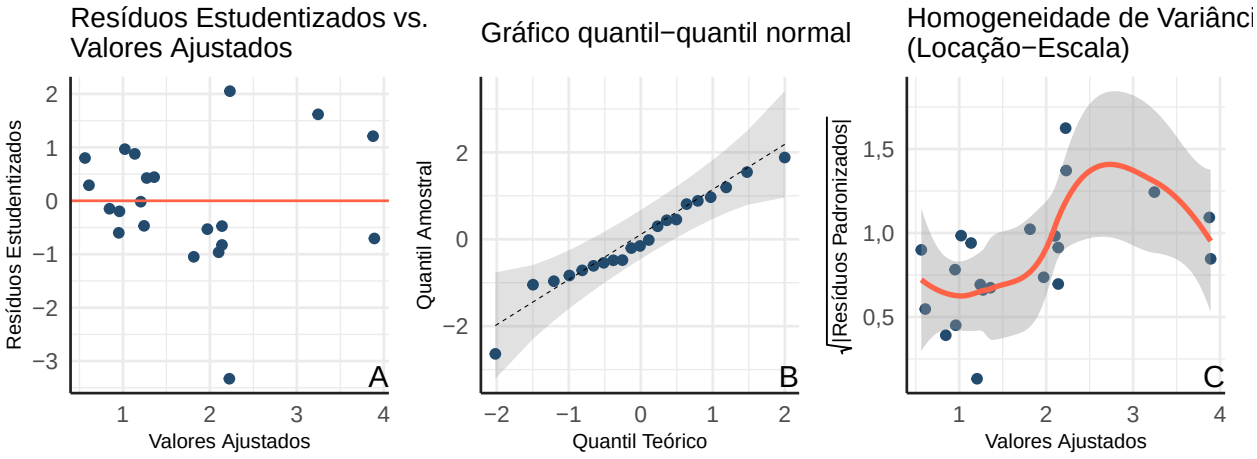
Legenda:
[1] LI: Limite Inferior (2,5%)
[2] LS: Limite Superior (97,5%)
[*] Nível de Significância de 5%.

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 $\hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3 = 0$. Porém, a introdução da concentração de HNO_3 não pode ser vista significativa ao modelo.

Através dos Intervalos de Confiança calculados (Tabela 3) 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

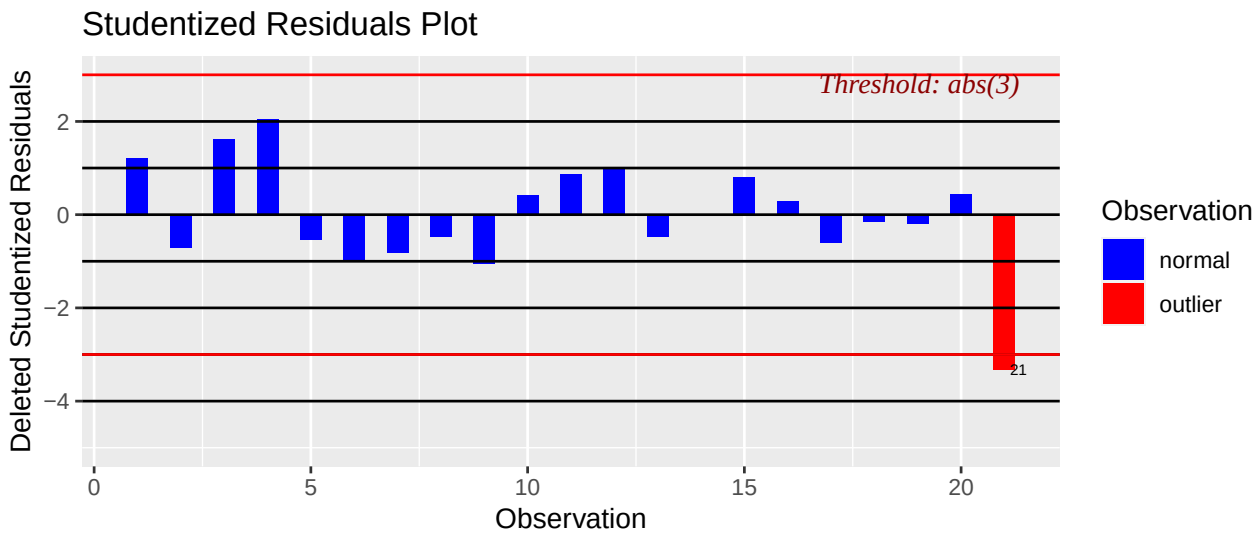


A Figura 2A apresenta um comportamento simétrico dos resíduos, podendo ser constatado uma pequena variabilidade inicial e um aumento desta à medida que os valores ajustados aumentam, caracterizando uma baixa heterocedasticidade. A Figura 2C, que trata da Homogeneidade de Variâncias (Locação-Escala) ressalta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

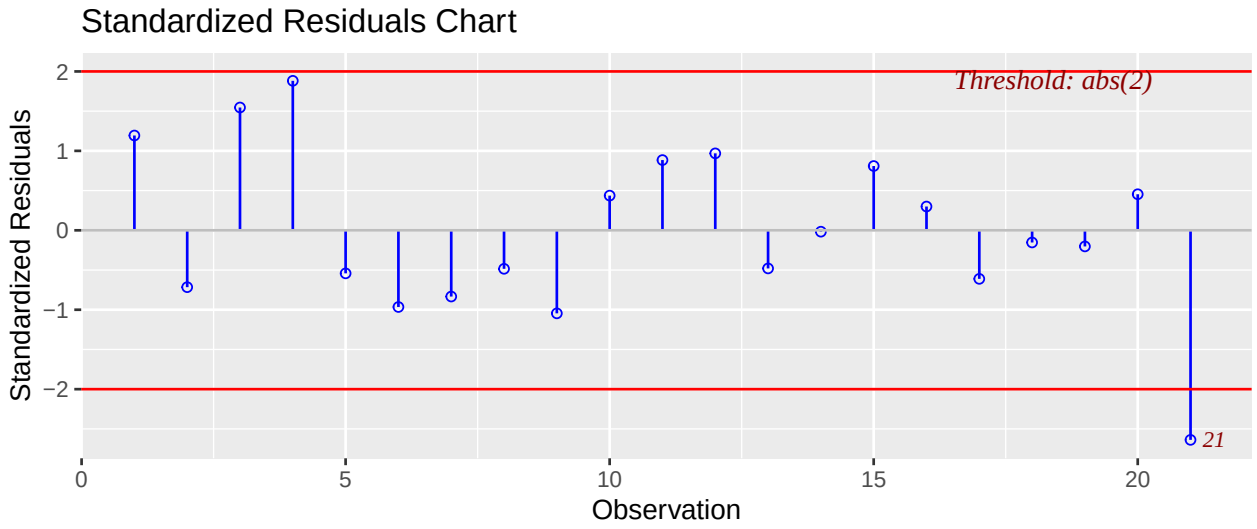
A análise dos gráficos de diagnóstico permite avaliar as observações realizadas e conhecer a influência de cada uma delas para o madelo de regressão proposto. Assim, com base no modelo, é possível fazer as seguintes análises:

Figura 3: Valores Ajustados e Resíduos Studentizados



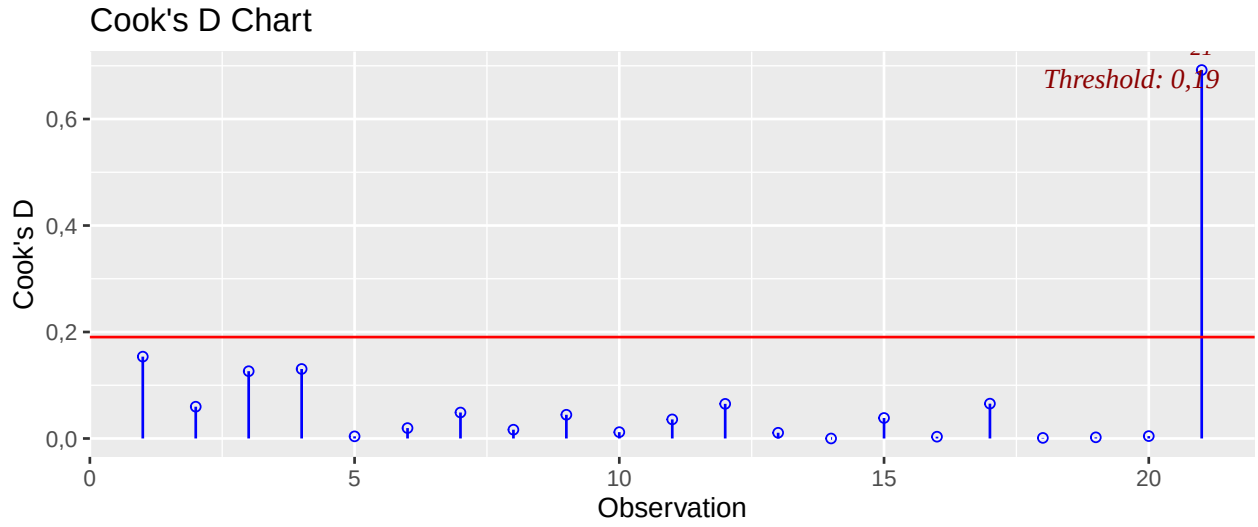
A Figura 3 demonstra que os resíduos estão todos dentro dos limites esperados, com exceção da observação 21 que por pouco ultrapassou o limite inferior. Não parece ser o caso de nenhuma intervenção por conta deste valor.

Figura 4: Valores Ajustados e Resíduos Padronizados.



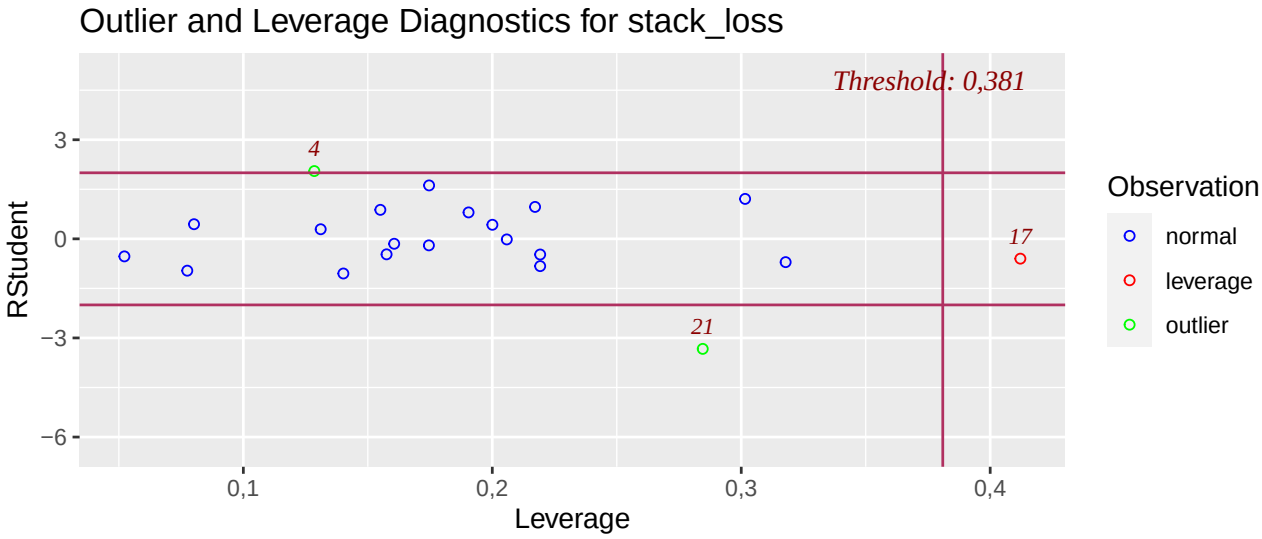
Na análise da Figura 4, onde os resíduos foram padronizados, verifica-se que apenas a observação 21 está fora do limite de aceitação.

Figura 5: Distância de Cook.



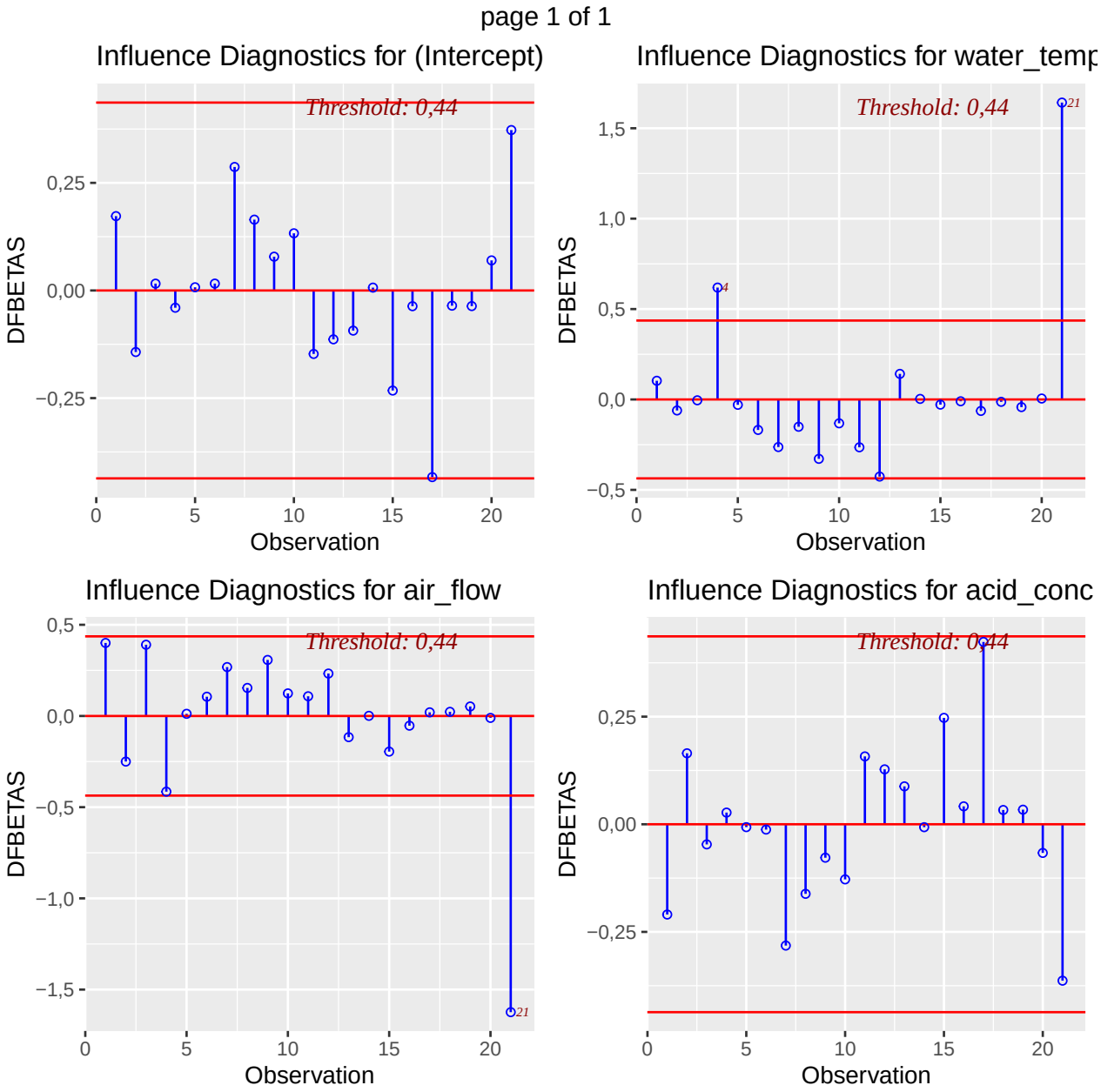
A analise da distância de Cook apresentada na Figura 5 demonstra que novamente apenas a observação 21 destoa do conjunto de observações e tem uma distância expressiva.

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



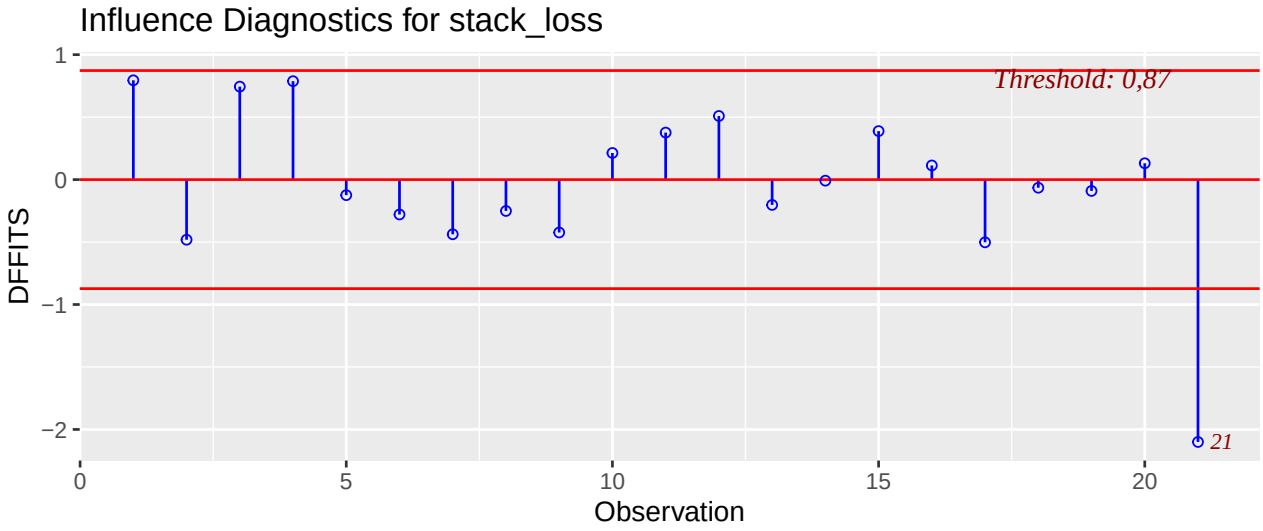
Pela Figura 6, observamos 2 observações que podem ser consideradas como *Outliers* e uma como ponto de alavanca. Interessante notar que apesar da observação 21 ter aparecido nos gráficos anteriores como uma observação anômala, aqui ela é classificada como um *outlier* o que denota uma influência menos prejudicial que a da observação 17 que aparece pela primeira vez.

Figura 7: DFBetas para as variáveis do modelo.



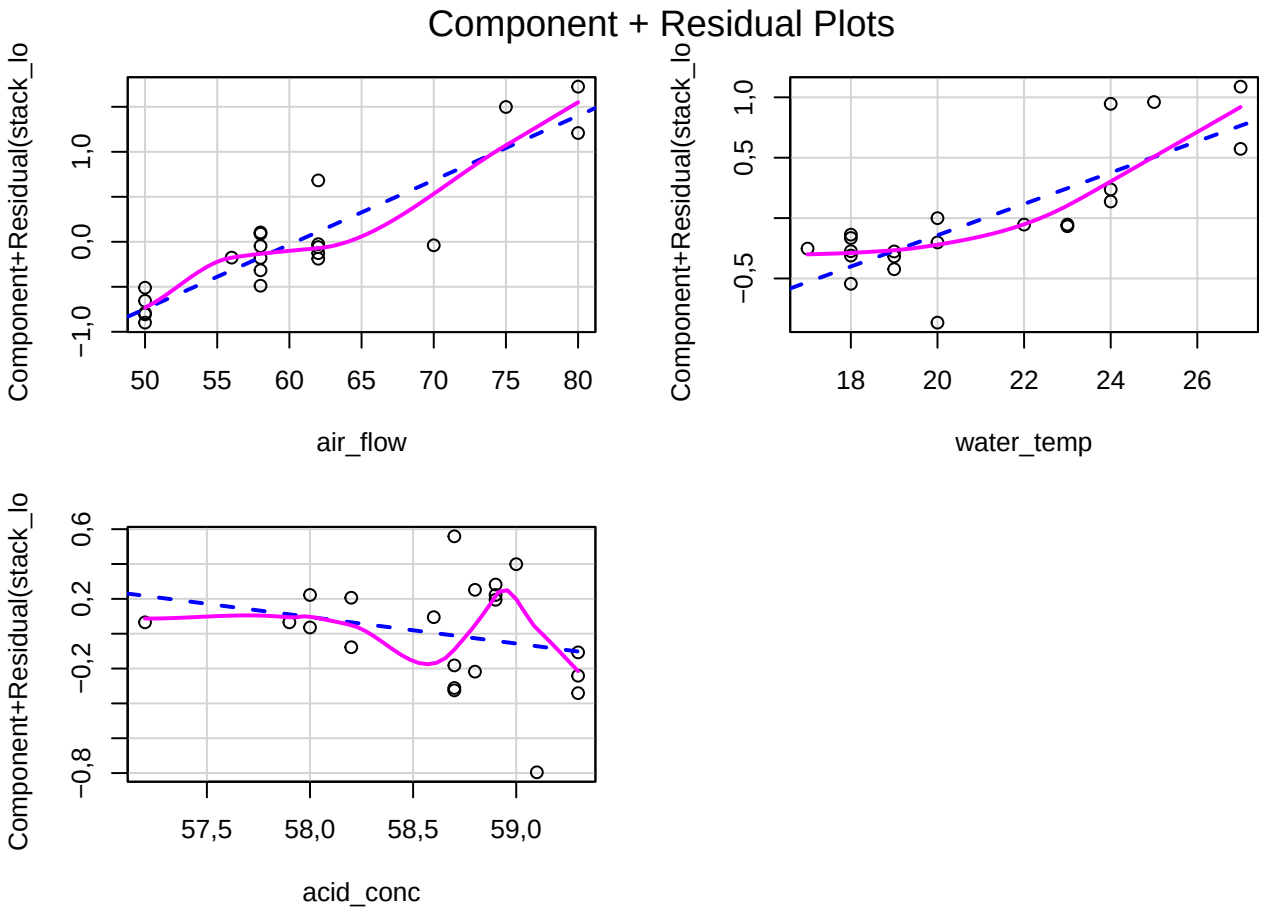
A Figura 7 apresenta os DFBetas para cada uma das variáveis utilizadas no modelo. Nota-se que mais uma vez a observação 21 tem comportamento anômalo.

Figura 8: DfFit para as variáveis do modelo.



A Figura 8 acompanha os gráficos anteriores apresentando mais uma vez a observação 21 como discrepante.

Figura 9: COVRatio para as variáveis do modelo.

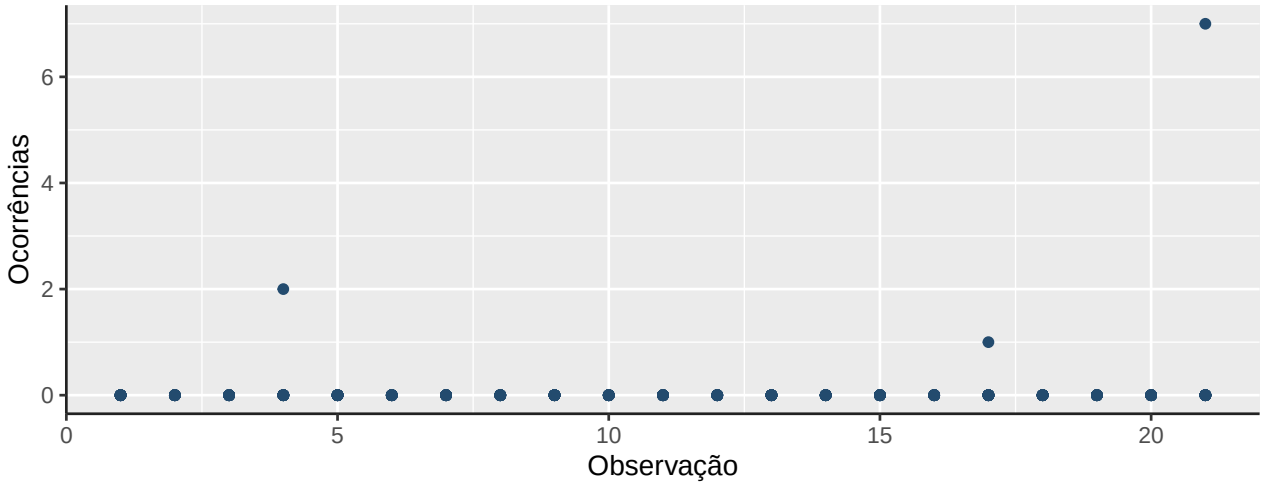


Da Figura 9 verifica-se que as variáveis estão diretamente correlacionadas com os resíduos da Amônia Perdida, o que por sua vez indica que a inclusão de observações destas variáveis apresentam bom impacto ao modelo.

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 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 e com base na análise da variância feita na Tabela 2, a eliminação da variável "Concentração de HNO_3", chega-se a um novo modelo dado por:

Y_i^* = -5,042+ 0,094 X_{1i}^*+ 0,052 X_{2i}^*

Onde:
Y_i^* - Amônia Perdida;

X^*_{1i} - Fluxo de Ar;
X^*_{2i} - Temperatura da Água;

Neste novo modelo o coeficiente de determinação calculado foi de $R^2 = 0{,}966$, o que denota que 96,6% da variância dos dados é explicada pelo modelo. O valor deste novo coeficiente permite concluir que a eliminação das observações com maior impacto e da variável com pouca relevância ao modelo foi benéfica. Pode-se calcular o coeficiente de determinação ajustado igual a $R^2_a = 0{,}961$

Conclusões

Verificou-se que embora os dados apresentem certa assimetria é possível propor um modelo de regressão linear múltipla para conhecer a quantidade de amônia perdida no processo.

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.

O modelo final obtido, embora muito mais simples que o inicial, foi capaz de ter um resultado expressivamente melhor confirmando a eficácia do uso das técnicas adotadas.

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. Realisando 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 obseervações anômalas e pela definição de variáveis preditoras a serem incluídas no modelo.

Resultados

Tratamento dos dados

A fim de reduzir o número de variáveis do modelo, com uma perda de informação aceitável as variáveis de máximo e mínimo foram substituídas por uma variável de amplitude, isto é, a amplitude térmica representa a diferença entre o valor máximo e mínimo observados, criando assim as seguintes variáveis:

- **AMPAT** - Amplitude térmica do ar diária;
- **AMPST** - Amplitude térmica do solo diária;
- **AMPH** - Amplitude da umidade relativa diária.

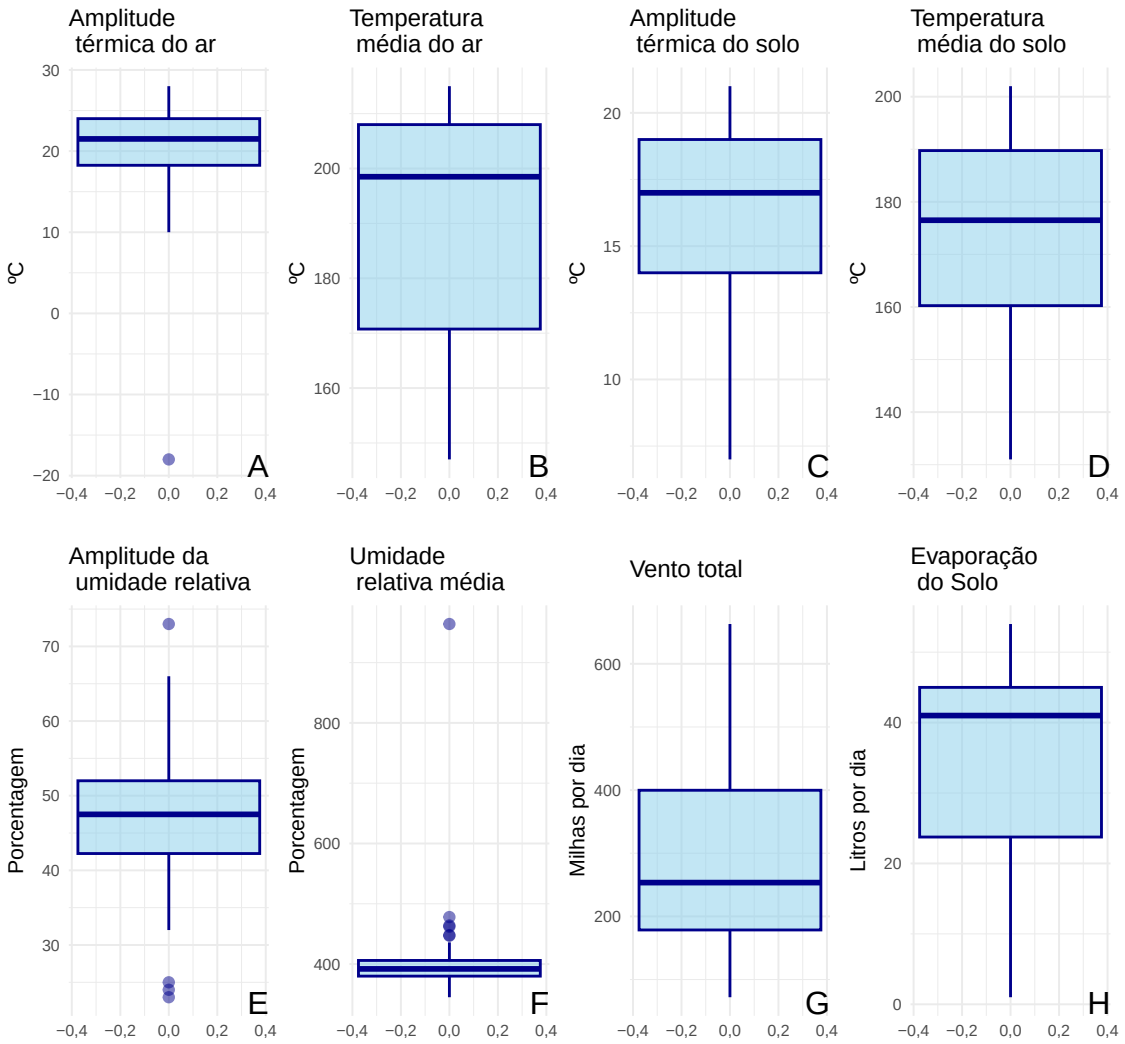
Análise descritiva dos dados

Similarmente ao que foi feito na primeira atividade, é possivel realizar uma descrição prévia dos dados por meio de medidas de resumo e de gráficos do tipo box-plot como vê-se a seguir:

Tabela 5: Medidas Resumo dos dados

	Mín	Q1	Med	Média	Q3	Máx	Desv.padrão	CV	Assimetria	Curtose
Amplitude da umidade relativa	23	42	47,5	46,22	52	73	9,88	0,21	-0,24	0,61
Amplitude témica do solo	7	14	17,0	16,30	19	21	3,69	0,23	-0,82	-0,20
Amplitude térmica do ar	-18	18	21,5	20,20	24	28	6,93	0,34	-3,59	17,28
Evaporação do Solo	1	23	41,0	34,67	45	54	14,64	0,42	-0,71	-0,68
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
Umidade relativa média	345	380	392,0	410,00	406	964	88,40	0,22	5,34	30,57
Vento total	72	177	253,5	284,20	400	663	149,45	0,53	0,70	-0,34

Figura 11: BoxPlot das variáveis em análise.



Na maioria das variáveis percebe-se uma assimetria nos dados. O número de *outliers* é razoável e precisará ser investigado mais adiante. Entretanto, já pode se verificar uma inconsistência nos dados que já pode ser eliminada: há uma observação negativa na amplitude térmica do ar, o que é inconcebível. Embora as condições não sejam ideiais pode-se propor um modelo de regressão para se melhor conhecer a evaporação de água pelo solo.

Modelo de Regressão Linear Múltipla

O modelo de regressão lineal múltipla inicialmente obtido pode ser representado por:

Y_i =-72,854 + 0,198 X_{1i}+ 0,463 X_{2i}+ 0,874 X_{3i} -0,236 X_{4i}+ 0,711 X_{5i}+ 0,011 X_{6i}+ 0,017 X_{7i}

Onde:
Y_i - Evaporação do Solo;

X_{1i} - Amplitude térmica do ar;
X_{2i} - Temperatura do ar média;
X_{3i} - Amplitude térmica do solo;;
X_{4i} - Temperatura do solo média;
X_{5i} - Amplitude da umidade relativa;
X_{6i} - Umidade relativa média;
X_{7i} - Vento Total.

Interpretando-se o modelo pode-se dizer que para cada variável, fixadas as demais condições (*Ceteris Paribus*), temos que:

- um aumento de 198mL na evaporação para cada grau Ceusius na amplitude térmica do ar;
- um aumento de 463mL na evaporação para cada grau Ceusius na temperatura média do ar;
- um aumento de 874mL na evaporação para cada grau Ceusius na amplitude térmica do solo;
- uma redução de 236mL na evaporação para cada grau Ceusius na temperatura média do solo;
- um aumento de 711mL na evaporação para ponto porcentual de amplitude da umidade relativa;
- um aumento de 11mL na evaporação para ponto porcentual de umidade relativa média;
- um aumento de 17mL na evaporação para cada milha de vento total.

Neste modelo o coeficiente de determinação calculado foi de $R^2 = 0{,}758$, o que denota que 75,8% da variância dos dados é explicada pelo modelo. Pode-se calcular o coeficiente de determinação ajustado igual a $R_a^2 = 0{,}704$.

Testes de Diagnósticos do Modelo

Para avaliar se o modelo atende aos pressupostos, além da análise gráfica podem ser realizados testes de diagnósticos, que são testes de hipóteses para avaliação dos pressupostos que são:

- Normalidade;

  H_0 : Os resíduos possuem normalidade.

  H_1 : Os resíduos **não** possuem normalidade.

- Homoscedasticidade (Homogeneidade de Variância);

  H_0 : Os resíduos possuem variância constante.

  H_1 : Os resíduos **não** possuem variância constante.

- Linearidade;

- Independência.

  H_0: Existe correlação serial entre os resíduos.

  H_1: **Não** existe correlação serial entre os resíduos.

Para tanto serão utilizados os seguintes testes:

- Shapiro-Wilk, para avaliar a Normalidade;
- Breush-Pagan, para avaliar a Homoscedasticidade;
- Durbin-Watson, para avaliar a Independência.

Tabela 6: Testes de Diagnósticos dos Resíduos

	Estatística de teste	p-valor
Shapiro-Wilk	0,9299	0,0093
Breush-Pagan	15,0727	0,0351
Durbin-Watson	1,3778	0,0041

A Tabela 6 traz os testes de diagnósticos realizados para avaliar o modelo de regressão ajustado. Verifica-se que a hipótese de nula da homocedasticidade deve ser rejeitada com um nível de significância de 5%, uma vez que o teste de Breush-Pagam obteve um p-valor menor que 0.05. A normalidade da distribuição foi também rejeitada como indica o p-valor do teste de Shapiro-Wilk. Nota-se ainda que há dependência entre as características confirmado pelo p-valor do teste de Durbin-Watson. Este codependência era esperada uma vez que os valores médios de tempeartura do ar e do solo são dependente bem como a unidade relativa. O que pode ser conferdo na matriz de correlação a seguir.

Correlação entre as variáveis do modelo

A correlação entre as variáveis do modelo pode ser medida pelo coeficiente de correlação entre elas.

Tabela 7: Matriz de correlação das variáveis do modelo

	Evaporação do Solo
Amplitude térmica do ar	0,538
Temperatura média do ar	0,692
Amplitude térmica do solo	0,756
Temperatura média do solo	0,654
Amplitude da umidade relativa	0,619
Umidade relativa média	-0,062
Vento total	0,105
Evaporação do Solo	1,000

Nota-se que as variaveis mais fortemente correlacionadas à evaporação de água no solo no modelo proposto são as temperaturas média do ar e a amplitude térmica do solo, e as menos correlacionadas são a umidade realtiva média e a quantidade total de ventos.

Significância das variáveis 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 8: Análise de Variância (ANOVA)

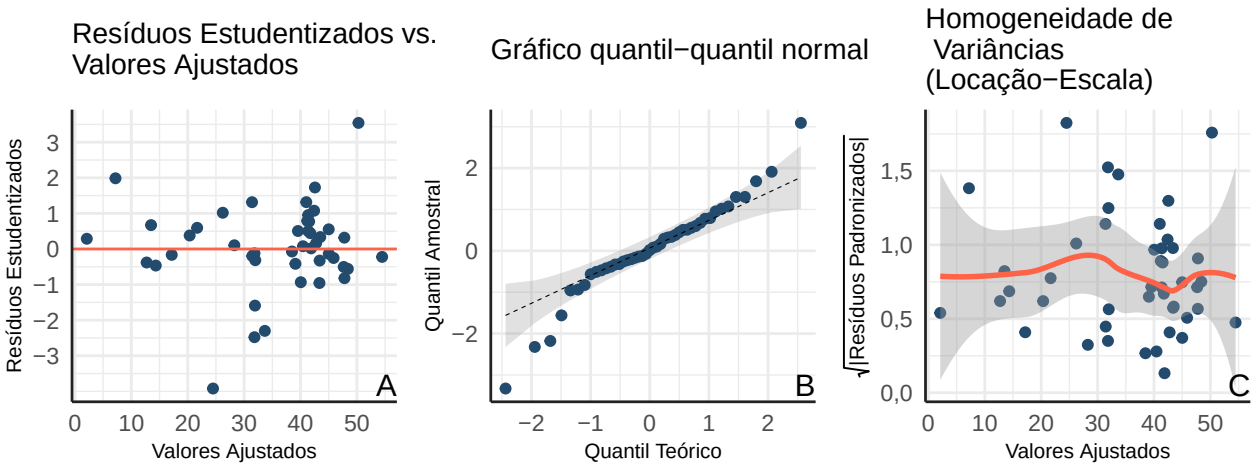
	GL[1]	Soma de Quadrados	Quadrado Médio	Estatística F-Snedecor	p-valor
Amplitude térmica do ar	1	2.532,010	2.532,010	44,2194	0,0000
Temperatura média do ar	1	2.572,696	2.572,696	44,9300	0,0000
Amplitude térmica do solo	1	705,547	705,547	12,3218	0,0012
Temperatura média do solo	1	101,905	101,905	1,7797	0,1903
Amplitude da umidade relativa	1	445,690	445,690	7,7836	0,0083
Umidade relativa média	1	68,808	68,808	1,2017	0,2801
Vento total	1	196,719	196,719	3,4355	0,0718
Resíduos	37	2.118,625	57,260		

Legenda:
[1] GL: Graus de Liberdade

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 $\hat{\beta}_j = 0$. Porém, a introdução das variáveis: Temperatura média do solo, Umidade Relativa média e Vento total não podem ser vistas como significativas ao modelo.

Análise de Resíduos

Figura 12: Análise de resíduos do modelo ajustado

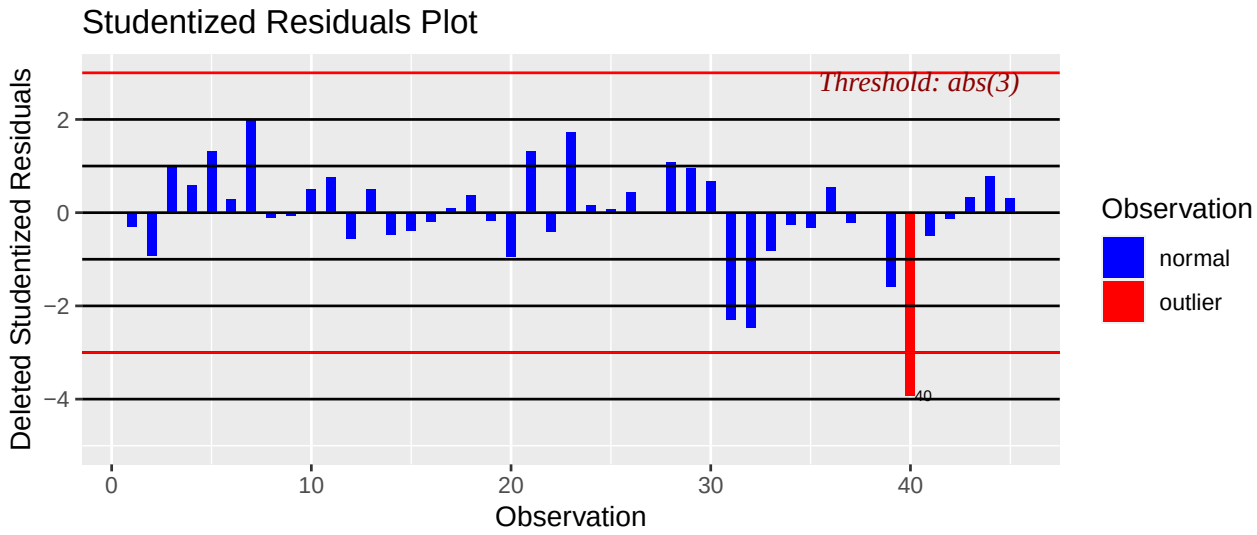


A Figura 12A apresenta um comportamento assimétrico dos resíduos, podendo ser constatado uma pequena variabilidade inicial e um aumento desta à medida que os valores ajustados aumentam, caracterizando uma maior heterocedasticidade. A Figura 12C, que trata da Homogeneidade de Variâncias (Locação-Escala) ressalta que há um problema na variabilidade dos dados, ampliando a interpretação feita na análise da Figura 12A, de que há uma mudança na variabilidade dos dados, caracterizando uma heterocedasticidade dos dados. A Figura 12B traz o gráfico para avaliação da normalidade dos dados, mostra que os dados não estão precisamente sobre a reta de referência, especialmente nas caudas da distribuição onde fogem inclusive da região pertencente ao Intervalo de Confiança - IC, podendo assumir que não há normalidade.

Gráficos de Diagnóstico

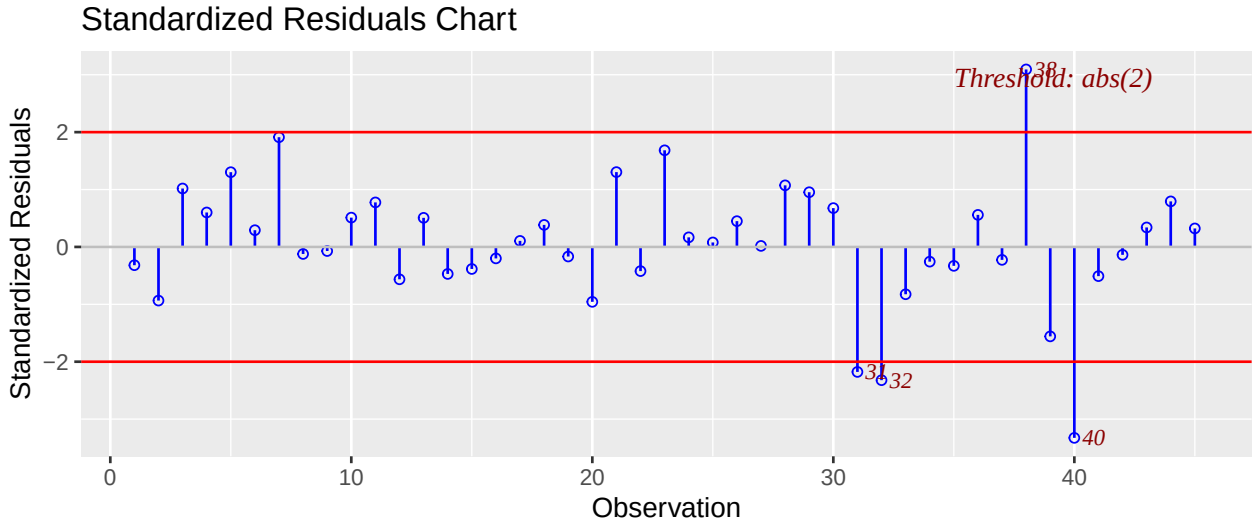
A análise dos gráficos de diagnóstico permite avaliar as observações realizadas e conhecer a influência de cada uma delas para o madelo de regressão proposto. Assim, com base no modelo, é possível fazer as seguintes análises:

Figura 13: Valores Ajustados e Resíduos Studentizados



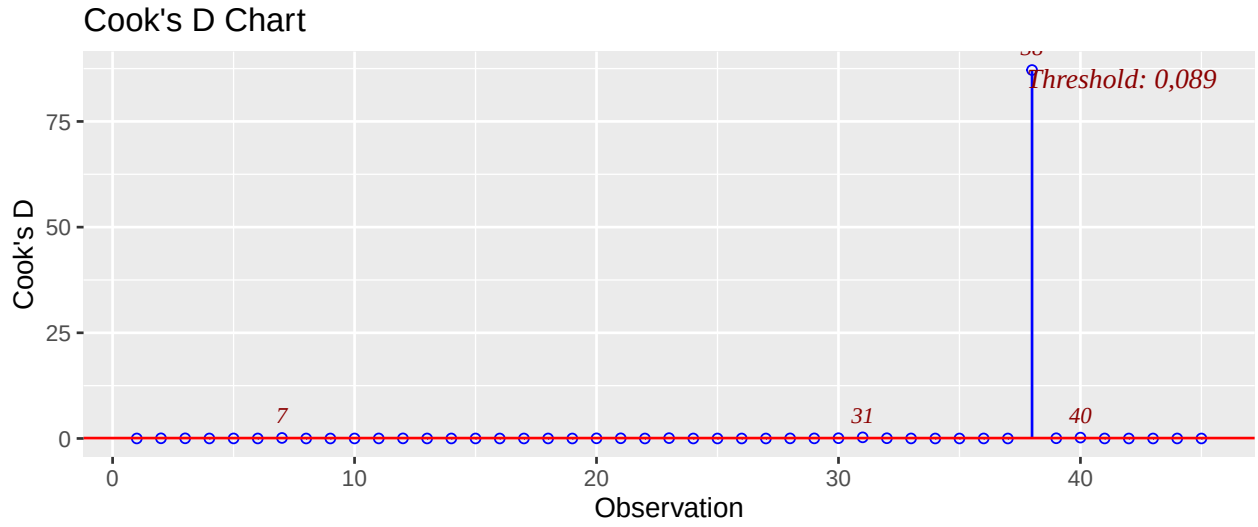
A Figura 13 demonstra que os resíduos estão todos dentro dos limites esperados, com exceção da observação 40 que ultrapassa o limite inferior. Não parece ser o caso de nenhuma intervenção por conta deste valor.

Figura 14: Valores Ajustados e Resíduos Padronizados.



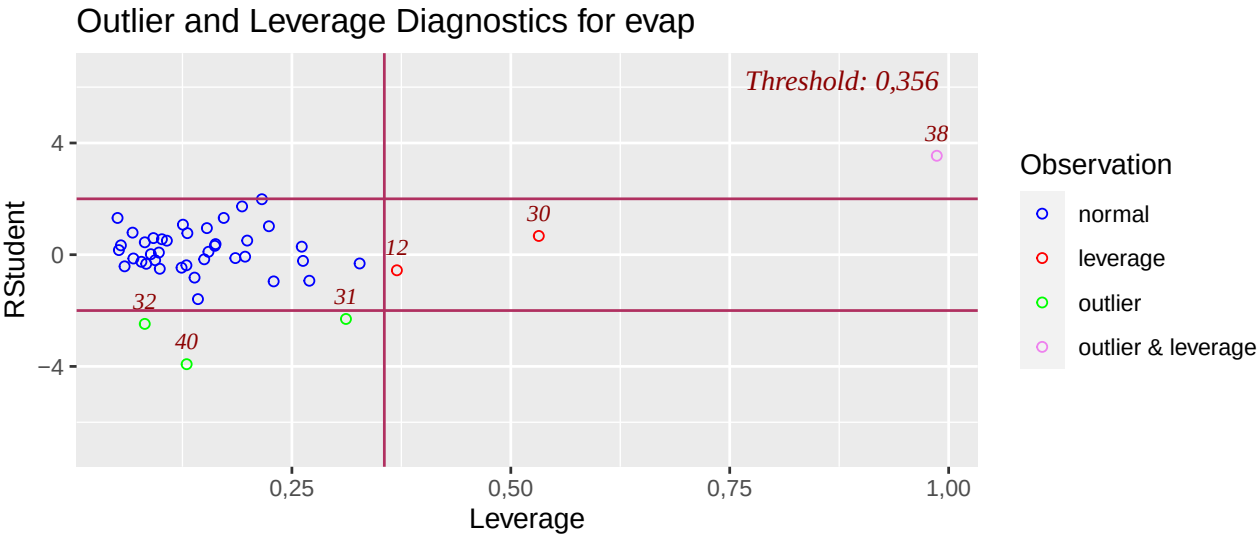
Na análise da Figura 14, onde os resíduos foram padronizados, verifica-se que a mesma observação anterior (40) está além do limite de aceitação, bem como outras três foram detectadas: 31, 32 e 38.

Figura 15: Distância de Cook.



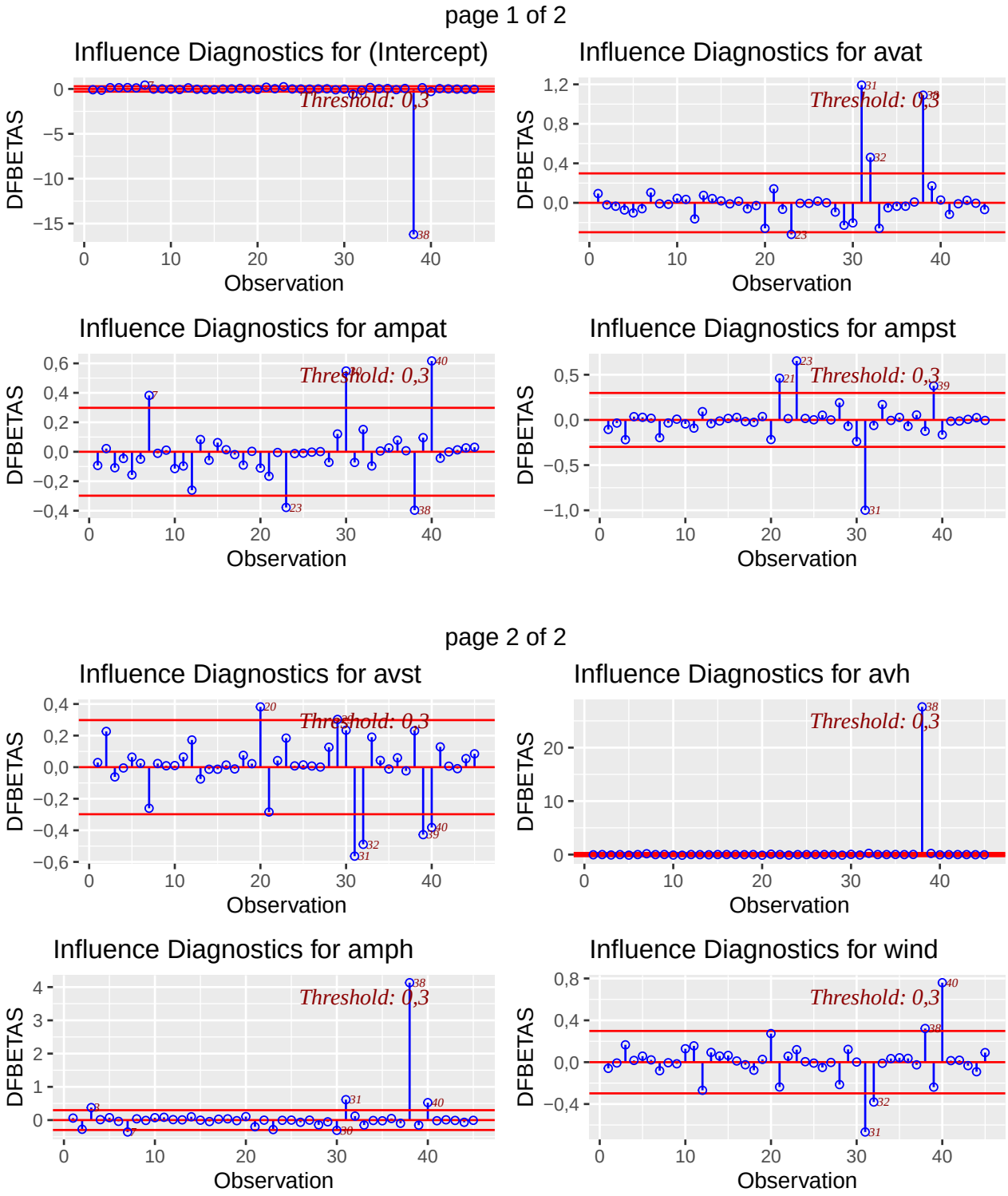
A analise da distância de Cook apresentada na Figura 15 demonstra que a observação 38 destoa do conjunto de observações e tem uma distância muito expressiva.

Figura 16: Análise dos pontos de Alavanca e Resíduo Studentizado.



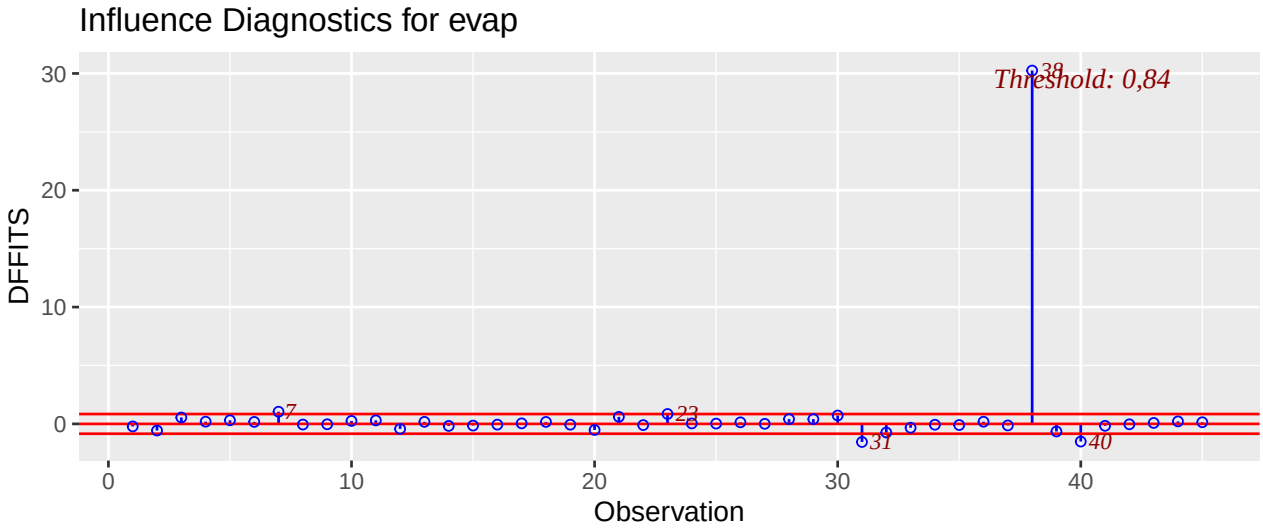
Pela Figura 16, observamos quatro observações que podem ser consideradas como *Outliers* e uma como ponto de alavanca e a observação 38 com as duas características sendo possívelmente uma das observações com maior influência negativa ao modelo.

Figura 17: DFBetas para as variáveis do modelo.



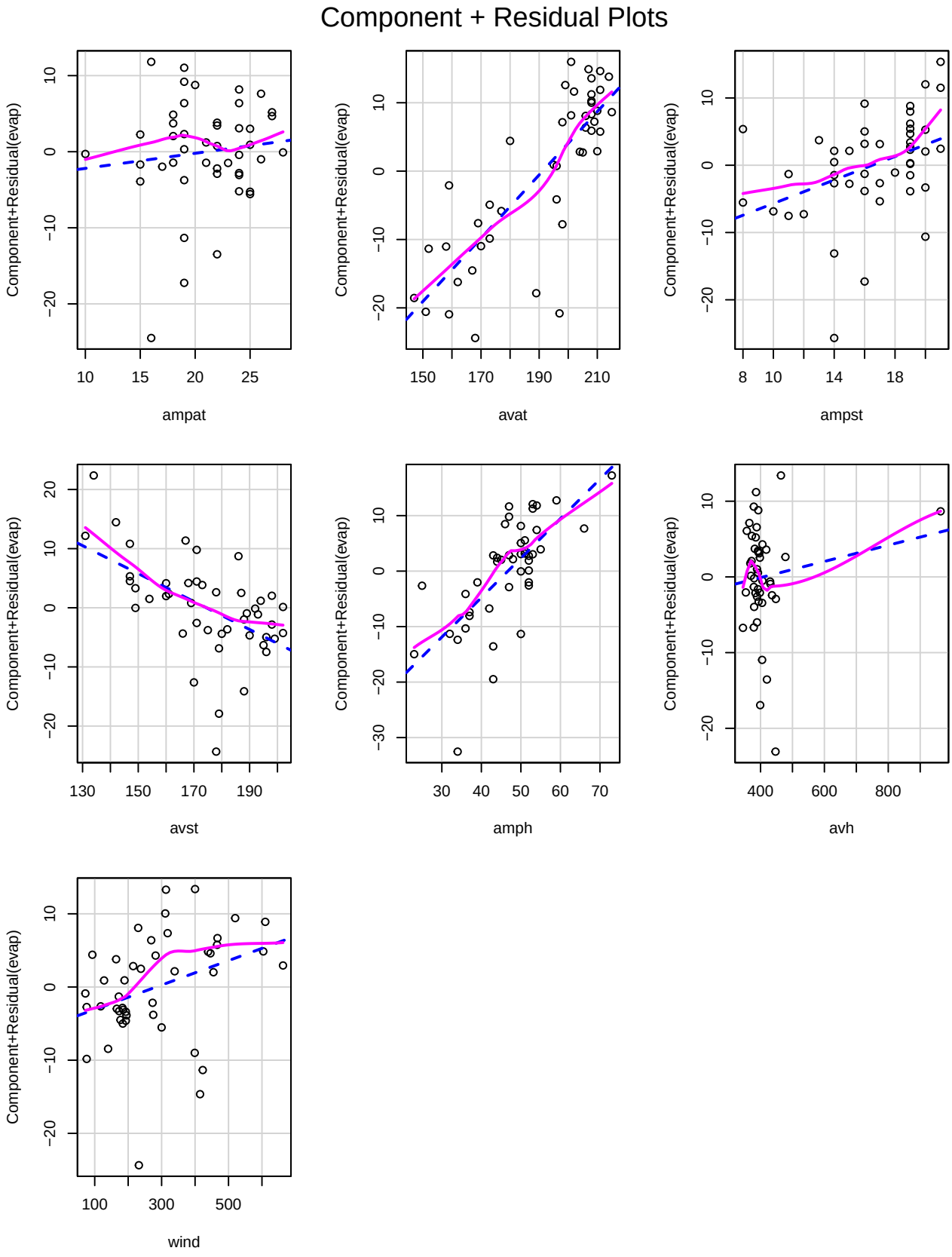
A Figura 17 apresenta os DFBetas para cada uma das variáveis utilizadas no modelo. Nota-se que as observações já identificadas como anômalas pelos gráficos anteriores se repetem com maior frequência na Figura 17.

Figura 18: DfFit para as variáveis do modelo.



A Figura 18 acompanha os gráficos anteriores apresentando mais uma vez a observação 38 como discrepante.

Figura 19: COVRatio para as variáveis do modelo.



Da Figura 19 verifica-se que as variáveis estão diretamente correlacionadas com os resíduos da evaporação da água do solo, o que por sua vez indica que a inclusão de observações destas variáveis apresentam bom impacto ao modelo.

Análise de Coliearidade dos Preditores

A colinearidade das variáveis preditoras pode ser avaliada por meio do cálculo dos R_j^2 e VIF_j destas variáveis, ver tabela 8.

Tabela 9: R_j^2 e VIF dos Preditores

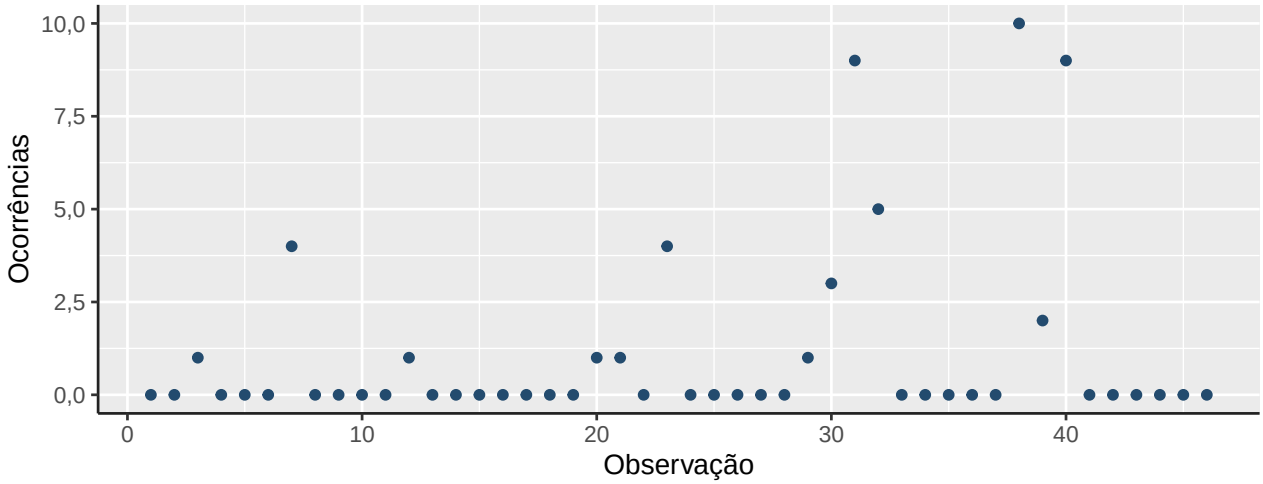
	R2_j	VIF
Amplitude térmica do ar	0,653	2,879
Temperatura média do ar máxima	0,818	5,505
Amplitude térmica do solo	0,780	4,551
Temperatura média do solo	0,830	5,875
Amplitude da umidade relativa	0,751	4,014
Umidade relativa média	0,161	1,192
Vento total	0,298	1,424

Uma regra empírica estabelece que valores de VIF_j acima de dez implica em colinearidade entre as variáveis preditoras. Analisando a Tabela 8 observa-se que nenhuma das variáveis chegou proximo a esse valor, indicando, contrariamente ao senso comum, que não há uma colinearidade expressiva entre as variáveis preditoras.

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 20: Número de ocorrências para cada observação



Considerando apenas as observações com 2 ou mais ocorrências temos a Tabela a seguir.

Tabela 10: Observações com maior número de ocorrências.

Observação	Ocorrências
7	4
23	4
30	3
31	9
32	5
38	10
39	2
40	9

Podemos intuir que essas são as observações com maior impacto negativo no modelo. Logo, eliminando-as do conjunto de dados analizados, bem como as variáveis descritas como pouco relevantes pela Tabela 8, chegamos a um novo modelo dado por:

Y_i^* =-59,921 -0,702 X_{1i}^*+ 0,324 X_{2i}^*+ 1,462 X_{3i}^*+ 0,534 X_{4i}^*

Onde:
Y_i^* - Evaporação do Solo;
X_{1i}^* - Amplitude térmica do ar;
X_{2i}^* - Temperatura do ar média;
X_{3i}^* - Amplitude térmica do solo;;
X_{4i}^* - Amplitude da umidade relativa.

Neste novo modelo o coeficiente de determinação calculado foi de $R^2 = 0{,}865$, o que denota que 86,5% da variância dos dados é explicada pelo modelo. O valor deste novo coeficiente permite concluir que a eliminação das observações com maior impacto e das variáveis pouco

relevantes ao modelo foi benéfica. Pode-se calcular o coeficiente de determinação ajustado igual a $R_a^2 = 0{,}848$

Conclusões

Verificou-se que embora os dados apresentem assimetria e bom número de *outliers* foi possível propor um modelo de regressão linear múltipla para conhecer a quantidade de evaporação de água do solo.

O uso de amplitudes no lugar de mínimo e máximos pareceu eficiente para simplificar o modelo sem perda significativa de informação ao modelo de regressão linear multipla realizado. É de se supor ainda que essa transformação eliminou alguma colinearidade entre as variáveis preditoras.

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.

O modelo final obtido, embora mais simples que o inicial, foi capaz de ter um resultado expressivamente melhor confirmando a eficácia do uso das técnicas adotadas.