UNIVERSIDADE FEDERAL DA BAHIA
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA

CAMILLE MENEZES PEREIRA DOS SANTOS
MICHEL MILER ROCHA DOS SANTOS

LABORATÓRIO 7: DIAGNÓSTICO DO MODELO DE REGRESSÃO LINEAR MÚLTIPLO E COLINEARIDADE

Salvador

2023

CAMILLE MENEZES PEREIRA DOS SANTOS
MICHEL MILER ROCHA DOS SANTOS

LABORATÓRIO 7: DIAGNÓSTICO DO MODELO DE REGRESSÃO LINEAR MÚLTIPLO E COLINEARIDADE

Atividades de laboratório apresentadas ao Instituto de Matemática e Estatística da Universidade Federal da Bahia como parte das exigências da disciplina Análise de Regressão ministrada pela professora Dra. Edleide de Brito.

Salvador

2023

SUMÁRIO

1 Introdução

A avaliação da eficiência em processos industriais desempenha um papel crucial no desenvolvimento e melhoria contínua das operações. Compreender os fatores que impactam a eficiência total de uma indústria é fundamental para identificar oportunidades de otimização e tomar decisões estratégicas.

Na primeira atividade, será realizada a avaliação da eficiência total de uma indústria que realiza oxidação de amônia em ácido nítrico. Inicialmente, será feita uma análise descritiva das variáveis de operação, incluindo a corrente de ar refrigerado, a temperatura de resfriamento, a concentração de ácido e a própria eficiência total. Serão observadas as distribuições, identificando possíveis discrepâncias, e analisadas as correlações entre as variáveis.

Em seguida, será aplicado um modelo de regressão linear múltipla para investigar como as variáveis de operação influenciam a eficiência industrial. Serão analisados os coeficientes estimados e suas interpretações, assim como o coeficiente de determinação e a significância global do modelo. Além disso, será realizada uma análise dos resíduos para verificar se as suposições do modelo são atendidas e identificar possíveis pontos influentes ou *outliers*.

Por meio dessas análises, buscará-se compreender quais variáveis de operação têm maior impacto na eficiência total da indústria e como essas informações podem contribuir para a melhoria dos processos industriais.

A segunda atividade tem como objetivo avaliar a multicolinearidade no modelo de regressão linear múltiplo. Com esse intuito, será, primeiramente, realizado uma análise descritiva dos dados, onde será possível detectar se as variáveis explicativas estão altamente correlacionadas entre si e se há pontos atípicos.

O próximo passo será ajustar um modelo de regressão linear múltiplo onde a evaporação do solo será a variável resposta e as demais variáveis climatológicas serão as variáveis explicativas. Através do teste F será observado se o modelo é significante para explicar a evaporação do solo. Será verificado os pressupostos dos resíduos (homoscedasticidade, normalidade e independência) e se há pontos atípicos ou influentes. E será avaliado a presença de multicolinearidade, usando o fator da inflação da variância (VIF).

Por fim, serão identificadas quais observações foram consideradas como atípicas ou influentes e quais variáveis explicativas apresentaram uma alta correlação com as demais variáveis explicativas (um alto valor de VIF). Essas observações e variáveis serão eliminadas e o modelo de regressão linear múltiplo será ajustado novamente. Todas as análises feitas para o primeiro modelo serão feitas para o segundo, com exceção da análise de multicolinearidade, pois com a diminuição do número de variáveis, a tendência é que os VIF's diminuam.

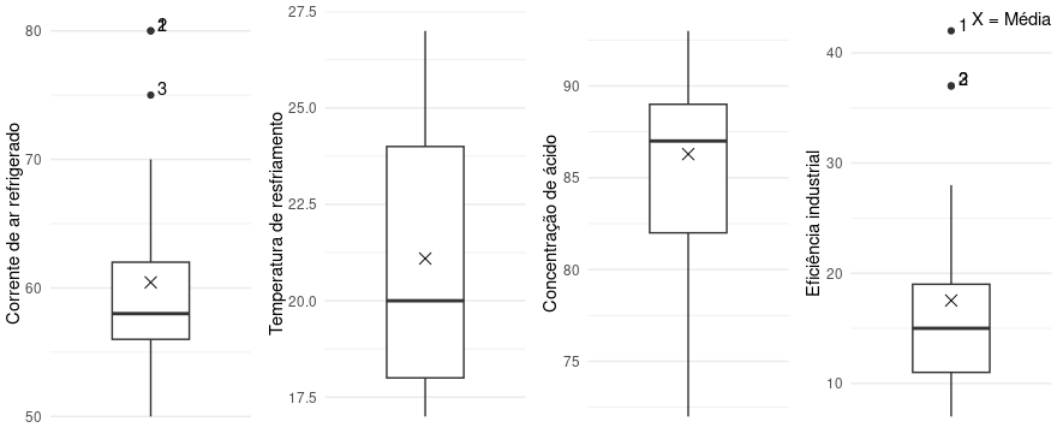
2 Avaliação da eficiência total da Indústria

2.1 Análise descritiva

Observando a Figura 2.1, a distribuição dos valores da variável corrente de ar refrigerado parece seguir uma distribuição aproximadamente simétrica, com a presença de dois pontos discrepantes.

A distribuição dos valores da variável temperatura de resfriamento sugere que a maioria das observações está entre o primeiro e o terceiro quartil, com uma mediana de 20. Parece haver uma ligeira assimetria positiva, uma vez que a média é um pouco maior do que a mediana.

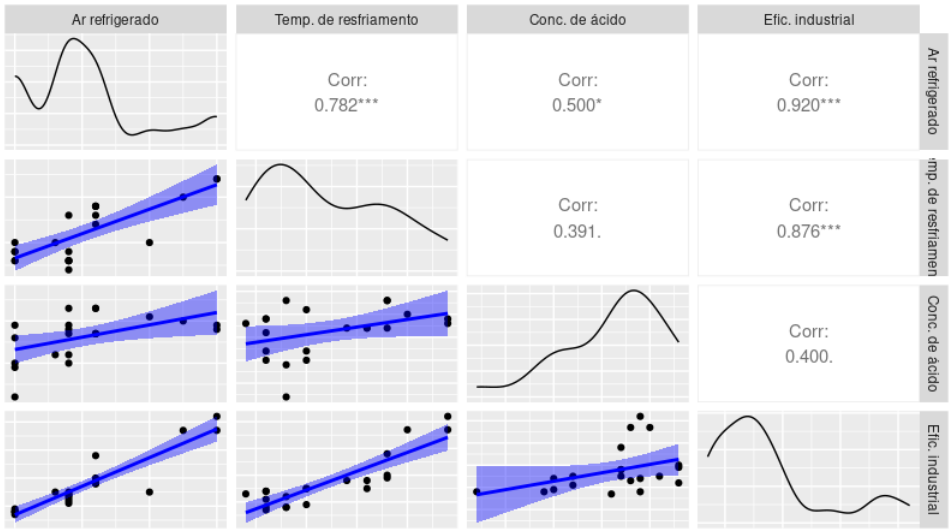
Figura 2.1: *Boxplots* das variáveis de operação de uma indústria que realiza oxidação de amônia em ácido nítrico



A distribuição dos valores da variável concentração de ácido indica que há uma maior variabilidade nas observações abaixo da mediana, embora média e a mediana estejam próximas, parece haver uma tendência clara de assimetria. A distribuição dos valores da variável eficiência total da indústria apresenta uma simetria, também com a presença de dois pontos discrepantes.

A correlações entre as variáveis na Figura 2.2, indica que as variáveis corrente de ar refrigerado e temperatura de resfriamento estão fortemente correlacionadas positivamente com a variável resposta (eficiência total). As demais variáveis também estão correlacionadas entre si, sobretudo a temperatura de resfriamento e a corrente de ar refrigerado,

Figura 2.2: Gráfico de dispersão, densidade e correlação das variáveis de operação de uma indústria que realiza oxidação de amônia em ácido nítrico



Considerando ainda os gráficos de dispersão, é possível notar que o ar refrigerando apresenta uma relação linear positiva relativamente forte com a variável resposta, apesar de haver poucas observações sendo duas delas não usuais. Assim como temperatura de resfriamento também apresenta uma relação linear positiva relativamente forte com a variável resposta. Mas, a concentração de ácido apresenta uma relação não linear com a variável resposta.

2.2 Modelo de regressão linear

A Tabela 2.1, apresenta uma sumarização da regressão linear em que a corrente de ar refrigerado, a temperatura de resfriamento e a concentração de ácido são utilizadas para predizer a eficiência industrial.

Tabela 2.1: Sumarização da regressão linear em que as variáveis de operação da indústria são utilizadas para predizer a eficiência industrial

Coeficientes	Coeficiente	EP	Est. T	Valor-p
Intercepto	-39,92	11,90	-3,36	**0,004**
Corrente de ar refrigerado	0,72	0,14	5,31	**<0,001**
Temperatura de resfriamento	1,30	0,37	3,52	**0,003**
Concentração de ácido	-0,15	0,16	-0,97	0,344

O coeficiente do intercepto aponta que, quando todas as variáveis independentes são iguais a zero, o valor estimado de da eficiência industrial é -39,92. E as interpretações para os coeficientes angulares estimados do modelo de regressão linear múltiplo são:

- Considerando todas as outras variáveis constantes, o aumento de uma unidade na corrente de ar refrigerado está associado a um aumento de 0,72 unidades no valor estimado da eficiência industrial;
- Considerando todas as outras variáveis constantes, o aumento de uma unidade na temperatura de resfriamento está associado a um aumento de 1,30 unidades no valor estimado da eficiência industrial;
- Considerando todas as outras variáveis constantes, o aumento de uma unidade na concentração do ácido não possui um efeito significativo no valor estimado da eficiência industrial, uma vez que o valor-p associado a esse coeficiente é 0,344 (acima do nível de significância de 0,05).

A Tabela 2.2 apresenta a Tabela ANOVA do modelo, o valor-p do teste F obtido foi menor que o nível 0,05. Portanto, rejeita-se a hipótese nula de igualdade de todos os coeficientes angulares do modelo a zero, a um nível de significância de 5%. Ou seja, rejeita-se a hipótese de que o modelo não tem nenhuma validade para explicar a variável resposta eficiência energética.

Tabela 2.2: Tabela ANOVA da regressão linear em que as variáveis de operação da indústria são utilizadas para predizer a eficiência industrial

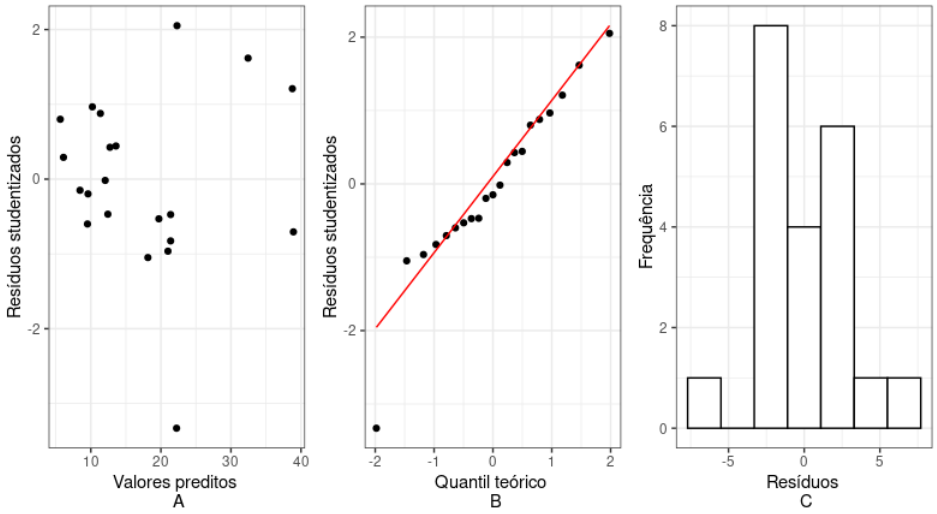
Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrado médio	Estatística F	Valor-p
Regressão	3	1890,41	630,14	59,9	<0,001
Resíduos	17	178,83	10,52		
Total	20	2069,24			

O coeficiente de determinação foi 0,91 e o coeficiente de determinação ajustado foi 0,90. Isso significa que, 91% da variabilidade da variável eficiência industrial pode ser explicada pelo modelo de regressão linear, ou seja, o seu ajuste é satisfatório.

2.3 Análise dos resíduos

O gráfico de resíduos contra valores preditos na Figura 2.3 (A), mostra que parece haver uma maior variabilidade nos resíduos entre os maiores valores preditos. Além disso, há um possível *outlier*, com valor absoluto maior que três. Realizando o teste de *Goldfeld-Quandt*, foi obtido um valor-p de aproximadamente 0,93, não rejeitando a hipótese nula, ao nível de significância de 5%, de que a variância dos resíduos é constante. Logo, os resíduos não violam a suposição de homocedasticidade.

Figura 2.3: Gráfico dos resíduos versus valores ajustados, resíduos estudentizados versus quantil teórico e histograma dos resíduos da regressão linear em que as variáveis de operação da indústria são utilizadas para predizer a eficiência industrial



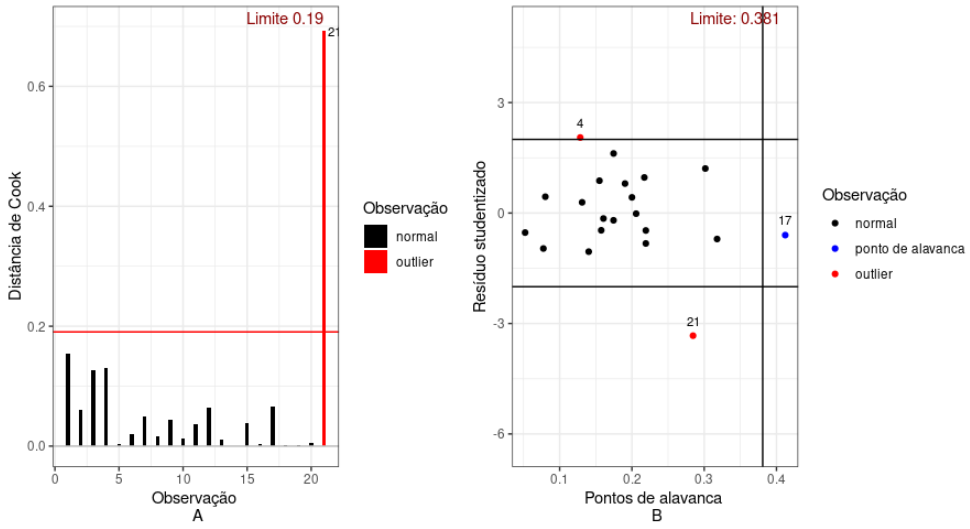
O *qq-plot* dos resíduos estudentizados, Figura 2.3 (B), mostra que a normalidade dos resíduos está sendo prejudicada pela presença do ponto atípico observado no gráfico anterior. O histograma dos resíduos, Figura 2.3 (C), reforça essa ideia, variável aparenta ter uma distribuição com a cauda esquerda mais pesada. Mas, com exceção dessa observação, a distribuição dos resíduos parece ser normalmente distribuída. O teste de *Shapiro-Wilk* indicou um valor-p de aproximadamente 0,82, não rejeitando a hipótese nula, a um nível de significância de 5%, de que os resíduos vem de uma distribuição normal. Então, podemos afirmar que a suposição de normalidade dos erros não foi violada pelo modelo.

Já o teste de *Durbin-Watson* resultou num valor-p de aproximadamente 0,04, apontando que há autocorrelação serial nos resíduos a um nível de 5%. Ou seja, há violação do pressuposto de independência.

A Figura 2.4 (A) indica que há um ponto influente no modelo (observação 21), uma vez que ele apresenta uma distância de Cook maior que o limite considerado de 0,19. Em outras palavras, se essa observação fosse removida da base de dados, isso causaria uma

mudança expressiva nas estimativas de mínimos quadrados obtidas para os coeficientes.

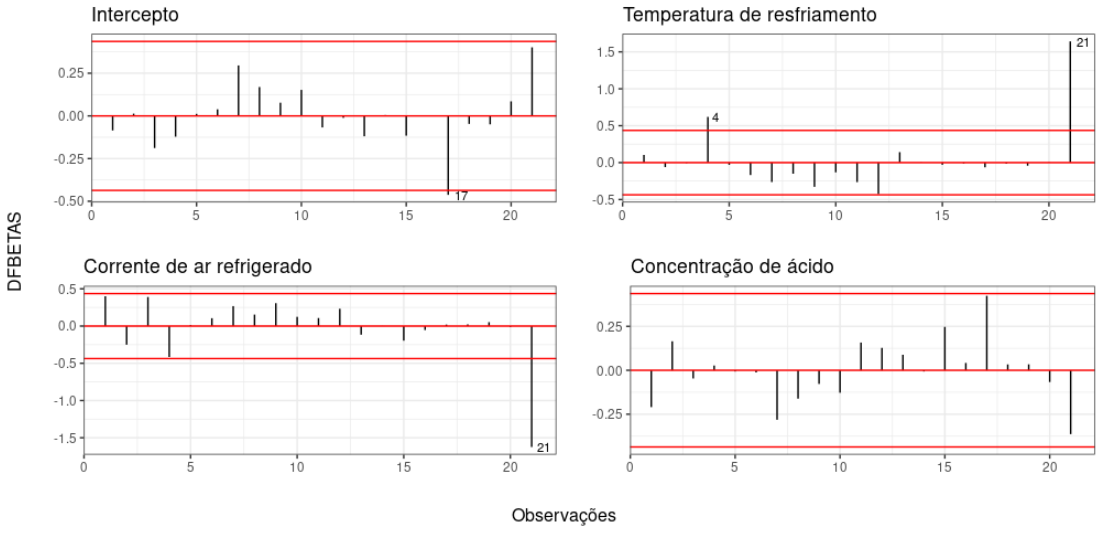
Figura 2.4: Gráfico de Distância de Cook, gráfico dos pontos de Alavanca e Resíduo estudentizado da regressão linear em que as variáveis de operação da indústria são utilizadas para predizer a eficiência industrial



Na Figura 2.4 (B), é possível notar que há uma observação (17) que é extrema no espaço das variáveis explicativas — ponto de alavanca — e duas (4 e 21) que são extremas na variável resposta — *outlier*.

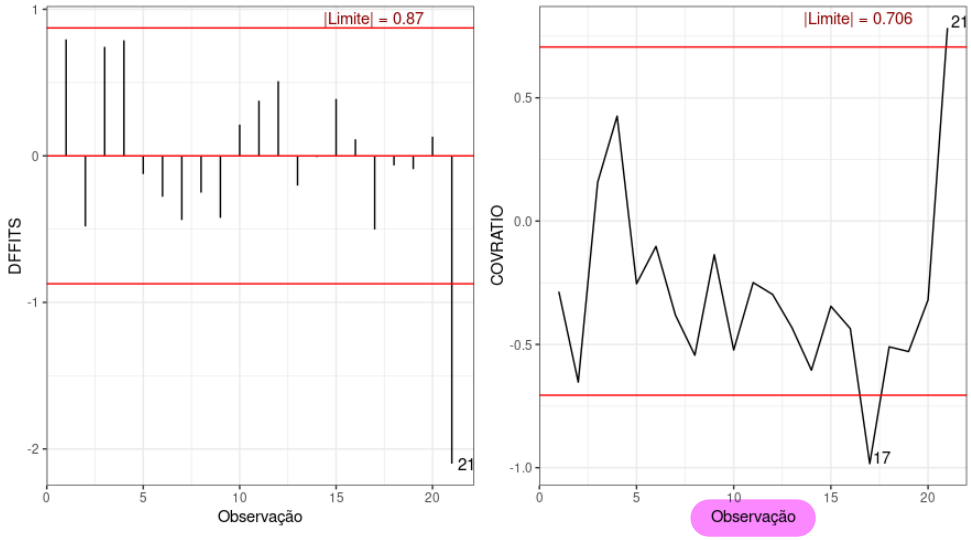
Os gráficos dos DFBETAS, Figura 2.5, também indicam a presença de pontos influentes no modelo de regressão para cada parâmetro do modelo, sendo considerados influentes aqueles com $|DFBETA_{i,j}| > 2/\sqrt{21}$. A observação 21 novamente se apresentou como observação influente para duas variáveis explicativas.

Figura 2.5: Gráfico de DFBETAS da regressão linear em que as variáveis de operação da indústria são utilizadas para predizer a eficiência industrial



A Figura 2.6 aponta que a observação 21 mais uma vez é indicada como uma observação influente no gráfico dos DFFITS e no gráfico do COVARATIO.

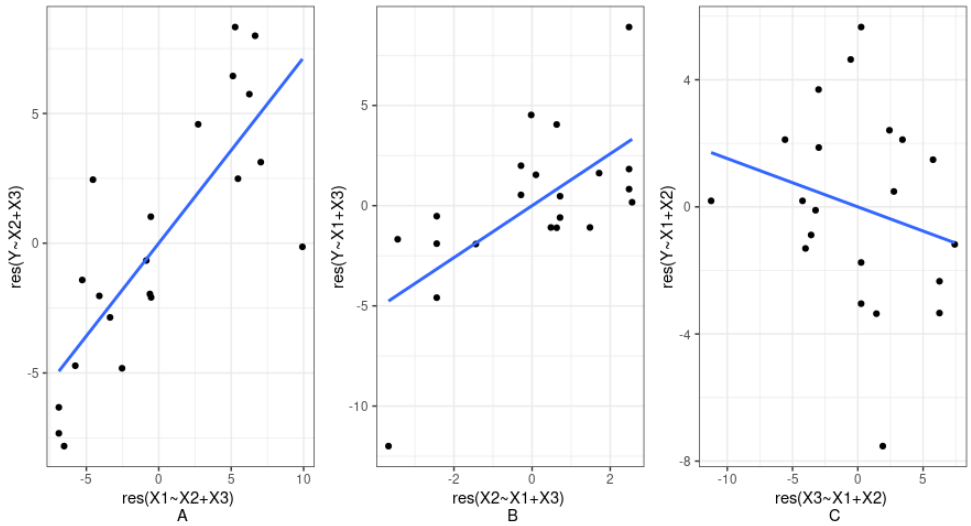
Figura 2.6: Gráfico de DFFITS e gráfico do COVARATIO da regressão linear em que as variáveis de operação da indústria são utilizadas para predizer a eficiência industrial



Como essa observação foi considerado uma observação influente por diferentes métricas, é aconselhável ajustar novamente o modelo de regressão linear múltiplo sem ela e verificar o impacto da retirada dessa observação.

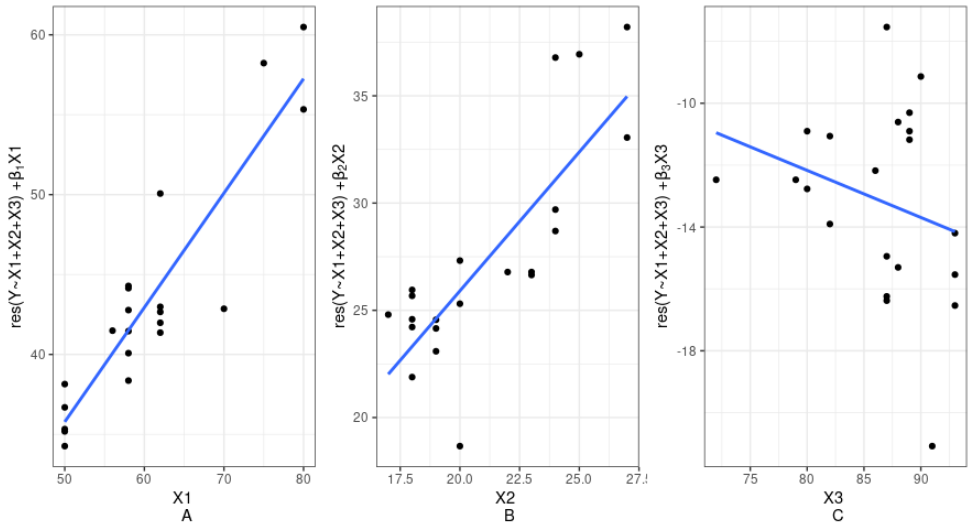
2.4 Regressão parcial e resíduos parciais

Figura 2.7: Gráfico da regressão parcial da regressão linear em que as variáveis de operação da indústria são utilizadas para predizer a eficiência industrial



Os gráficos da regressão parcial na Figura 2.7 (A) e (B), mostram que parece que ser necessário incluir as variáveis X1: corrente de ar refrigerado e X2: temperatura de resfriamento no modelo de regressão, já que os pontos estão em torno de uma reta. Mas, os pontos na Figura 2.7 (C), estão bem dispersos e não há um padrão claro. Logo, a variável X3: concentração de ácido não é necessária no modelo de regressão linear múltiplo.

Figura 2.8: Gráfico dos resíduos parciais da regressão linear em que as variáveis de operação da indústria são utilizadas para predizer a eficiência industrial



Os gráficos dos resíduos parciais, Figura 2.8 (A) e (B), assim como nos gráficos anteriores mostram a necessidade de incluir as variáveis corrente de ar refrigerado e temperatura de resfriamento no modelo de regressão. Uma vez que os pontos também estão em torno de uma reta, sobretudo na variável corrente de ar refrigerado. Mas, a Figura 2.8 (C) aponta que não há necessidade de incluir a variável concentração de ácido modelo.

3 Análise dos dados diários sobre evaporação do solo

3.1 Descrição dos dados

O conjunto de dados utilizado para essa atividade são sobre as condições climáticas e temporais de determinada região. Os dados contém 46 observações com 14 variáveis. os dados possuem três variáveis, que representam o índice da observação, o dia e o mês em que a observação foi coletada, que não são relevantes para explicar a evaporação do solo no contexto de regressão linear, portanto essas variáveis não serão utilizadas na análise.

As 11 variáveis presentes nesse estudo são:

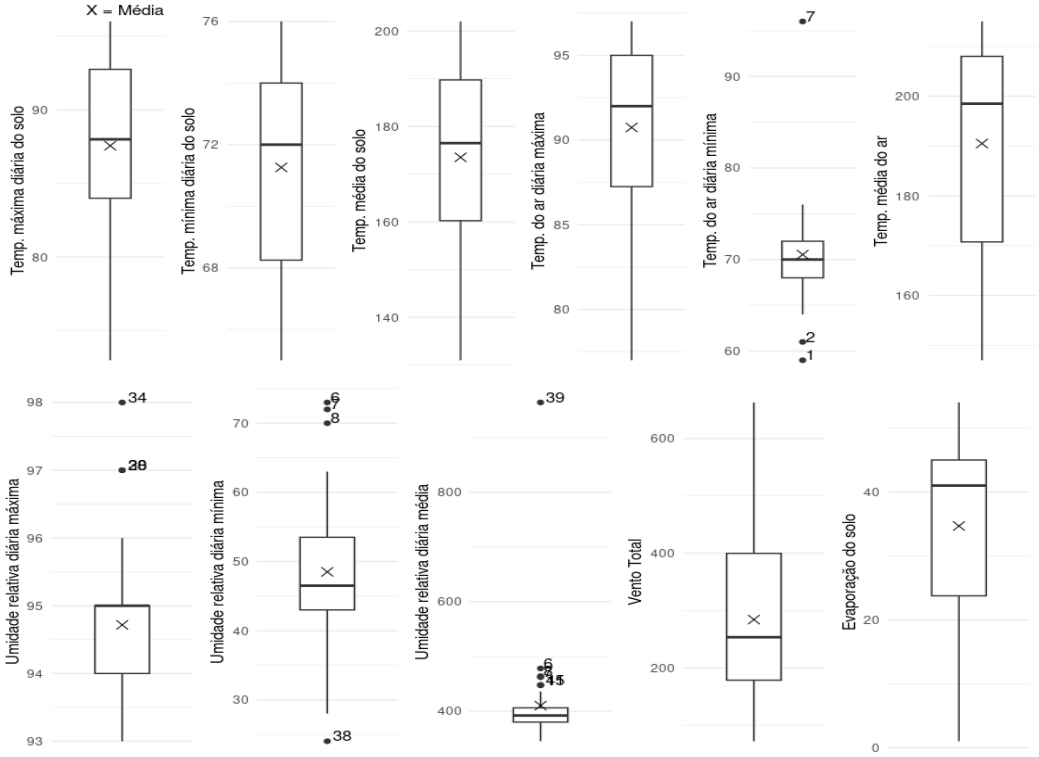
- Evaporação do solo ($EVAP$)
- Temperatura do ar diária máxima ($MAXAT$)
- Temperatura do ar diária mínima ($MINAT$)
- Temperatura média do ar ($AVAT$)
- Temperatura máxima diária do solo ($MAXST$)
- Temperatura mínima diária do solo ($MINST$)
- Temperatura média do solo ($AVST$)
- Umidade relativa diária máxima ($MAXH$)
- Umidade relativa diária mínima ($MINH$)
- Umidade relativa média (AVH)
- Vento total ($WIND$)

3.2 Análise descritiva

Nos *boxplots* da Figura 3.1, a temperatura máxima e mínima diária do solo além da temperatura média do solo apresentam uma maior variabilidade abaixo da mediana do que acima, o que poderia indicar que a distribuição dessas variáveis apresentam assimetria à esquerda.

A temperatura máxima diária do ar e a temperatura média do ar apresentam uma variabilidade bem maior quando comparada com a temperatura mínima diária do ar. Mas, a temperatura mínima diária do solo apresenta *outliers* (as observações 1,2 e 7).

Figura 3.1: *Boxplots* das variáveis climatológicas



Nas três variáveis que medem a umidade relativa apresentam *outliers*, com destaque para a variável umidade relativa média que apresenta a observação 39 como um valor atípico bem discrepante dos demais *outliers*. A umidade relativa diária máxima apresenta pequena variabilidade entre a mediana e o 3º quartil, a linha da mediana está logo acima da linha do 3º quartil.

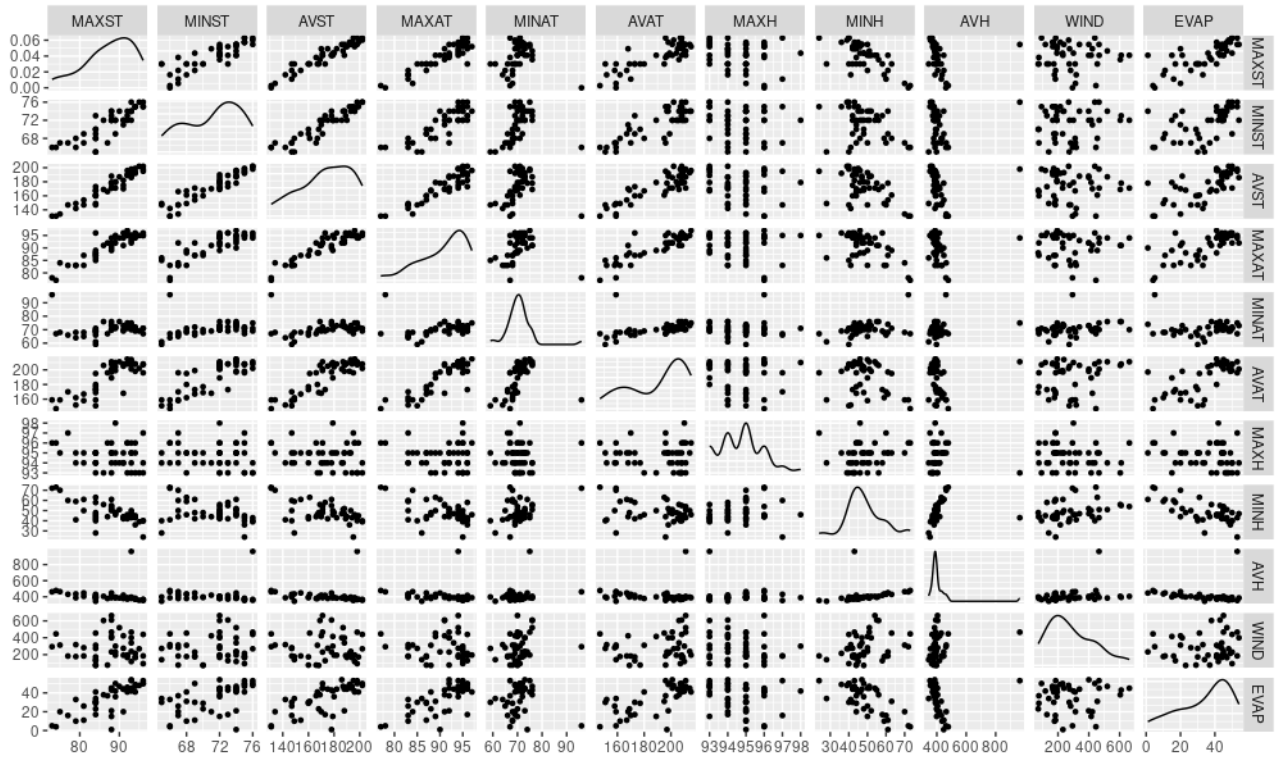
O vento total apresenta uma maior variabilidade acima da mediana enquanto que a evaporação do solo apresenta uma maior variabilidade abaixo da mediana. Portanto, o primeiro apresenta uma distribuição mais assimétrica à direita, enquanto o segundo apresenta uma distribuição mais assimétrica à esquerda.

Observando os gráficos de dispersão da Figura 3.2 e a matriz de correlação das variáveis

na Tabela 3.1, é possível notar que a temperatura máxima e mínima diária do solo (*MAXST* e *MINST*) e temperatura média do solo (*AVST*) estão bem correlacionadas positivamente entre si, apresentando relações bastante lineares.

A temperatura máxima diária do ar (*MAXAT*) é correlacionada positivamente com a temperatura média do ar (*AVAT*), apresentando um coeficiente de correlação linear de *Pearson* bem alto. Isso é observável na gráfico de dispersão entre elas. Entretanto, a temperatura mínima diária do ar (*MINAT*) não apresentou coeficientes de correlação alto com essas duas variáveis, apesar de aparentar ter uma relação bastante linear nos gráficos de dispersão. Estes valores mais baixos de coeficientes de correlação se devem, muito provavelmente, a observação 7 que apareceu como *outlier* no *boxplot* da variável temperatura mínima diária do ar na Figura 3.1 e que aparece isoladamente nos gráficos de dispersão.

Figura 3.2: Gráficos de dispersão e densidade das variáveis climatológicas



As variáveis de umidade relativa diária máxima e mínima (*MAXH* e *MINH*), junto a umidade relativa média, apresentam baixo coeficiente de correlação linear entre si. Isto pode ser verificado no gráfico de dispersão entre as variáveis umidade relativa diária máxima e mínima. Entretanto, para a variável umidade relativa média (*AVH*), o coeficiente de correlação parece está sendo afetado pela observação 39, que também se apresentou como *outlier* no *boxplot* desta variável.

O vento total (*WIND*) parece pouco correlacionado com as outras variáveis. A

evaporação do solo apresenta (*EVAP*) um alto coeficiente de correlação linear positivo (acima de 0,6) com as variáveis temperatura máxima diária do solo, temperatura média do solo, temperatura máxima diária do ar e temperatura média do solo. A umidade relativa diária mínima apresenta uma alta correlação linear negativa com a variável evaporação do solo. Estas variáveis podem ser boas preditoras para a variável resposta evaporação do solo.

~~Tabela 3.1:~~ Matriz de correlação das variáveis climatológicas

	MAXST	**MINST**	**AVST**	**MAXAT**	**MINAT**	**AVAT**	**MAXH**	**MINH**	**AVH**	**WIND**	**EVAP**
MAXST	1,00	0,85	0,95	0,91	0,05	0,82	-0,19	-0,67	-0,12	-0,03	0,77
MINST	0,85	1,00	0,93	0,84	0,29	0,82	-0,17	-0,34	0,06	0,09	0,54
AVST	0,95	0,93	1,00	0,91	0,18	0,87	-0,16	-0,53	-0,04	-0,03	0,69
MAXAT	0,91	0,84	0,91	1,00	0,10	0,87	-0,10	-0,53	-0,12	-0,04	0,72
MINAT	0,05	0,29	0,18	0,10	1,00	0,35	0,01	0,39	0,19	0,25	-0,01
AVAT	0,82	0,82	0,87	0,87	0,35	1,00	-0,04	-0,30	-0,03	0,17	0,71
MAXH	-0,19	-0,17	-0,16	-0,10	0,01	-0,04	1,00	0,17	-0,12	-0,17	-0,19
MINH	-0,67	-0,34	-0,53	-0,53	0,39	-0,30	0,17	1,00	0,22	0,31	-0,67
AVH	-0,12	0,06	-0,04	-0,12	0,19	-0,03	-0,12	0,22	1,00	0,24	-0,09
WIND	-0,03	0,09	-0,03	-0,04	0,25	0,17	-0,17	0,31	0,24	1,00	0,10
EVAP	0,77	0,54	0,69	0,72	-0,01	0,71	-0,19	-0,67	-0,09	0,10	1,00

3.3 Modelo de regressão linear

A Tabela 3.2 apresenta uma sumarização da ANOVA, onde é possível notar resultado do teste-F, que avalia a bondade global do ajuste do modelo completo. O teste rejeitou a hipótese nula ao nível de significância de 5%, ou seja, o teste indicou que o modelo de regressão linear múltiplo tem um ajuste significativo na evaporação do solo.

O coeficiente de determinação do modelo (R^2) foi 0,7984. Ou seja, 79,84% da variabilidade da evaporação do solo pode ser explicada pelo modelo de regressão linear múltiplo.

Tabela 3.2: Tabela ANOVA da regressão linear em que variáveis climatológicas são utilizadas para predizer a evaporação do solo

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrado médio	Estatística F	Valor-p
Regressão	10	7698,08	769,81	13,86	<0,001
Resíduos	35	19,03	55,54		
Total	45	9642,11			

Através da Tabela 3.3, é possível notar que apenas as variáveis temperatura média do ar (*AVAT*) e umidade relativa diária mínima (*MINH*) foram significativas no modelo, ao nível de significância de 5%. Portanto, grande parte das variáveis que estão no modelo não estão sendo relevantes para predizer a evaporação do solo.

Tabela 3.3: Sumário da Regressão: variáveis climatológicas para predizer a evaporação do solo

Coeficientes	Coeficiente	EP	Est. T	Valor-p
Intercepto	117,578	122,858	0,957	0,345
MAXAT	0,336	0,679	0,495	0,623
MINAT	0,009	0,291	0,034	0,973
AVAT	0,495	0,166	2,975	**0,005**
MAXST	0,240	0,903	0,265	0,792
MINST	-1,318	1,208	-1,091	0,283
AVST	-0,116	0,335	-0,346	0,732
MAXH	-0,877	1,011	-0,868	0,392
MINH	-0,823	0,231	-3,559	**0,001**
AVH	0,008	0,013	0,579	0,566
WIND	0,016	0,009	1,606	0,117

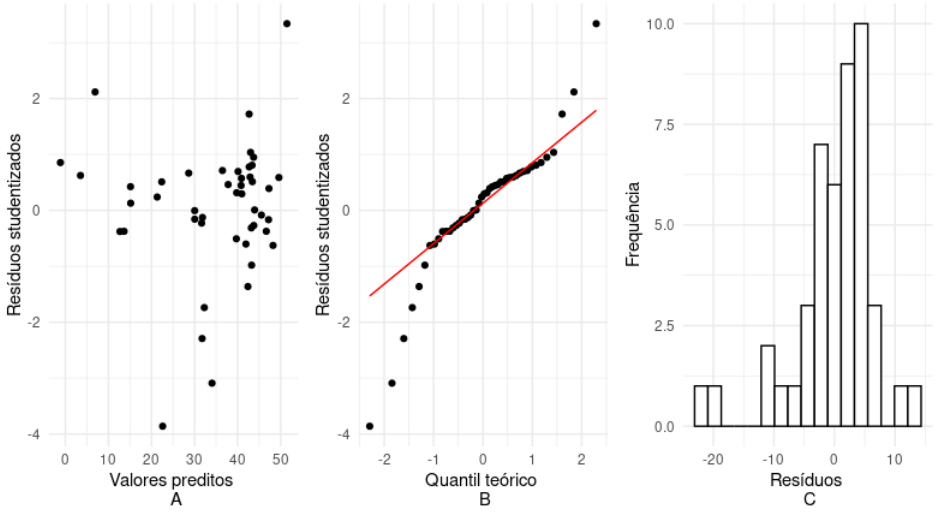
O coeficiente do intercepto aponta que, quando todas as variáveis independentes são iguais a zero, o valor estimado da evaporação do solo é 117,58. A interpretação dos coeficientes angulares do modelo, que são significativos ao nível de 5% são:

- Considerando todas as outras variáveis constantes, o aumento de uma unidade na temperatura média do ar está associado a um aumento de 0,495 unidades no valor estimado da evaporação do solo;
- Considerando todas as outras variáveis constantes, o aumento de uma unidade na umidade relativa diária mínima está associado a diminuição de 0,823 unidades no valor estimado da evaporação do solo.

3.4 Análise dos resíduos

O gráfico dos resíduos contra valores preditos na Figura 3.3, mostra que parece haver uma maior variabilidade nos resíduos nos valores intermediários dos valores preditos. Isso se deve a algumas observações que possivelmente podem ser consideradas como não usuais. O teste de *Goldfeld-Quandt* apresentou um valor-p de 0.25, ou seja, não rejeitamos a hipótese de variância contante dos resíduos a um nível de 5%. Logo, os resíduos desse modelo de regressão linear múltiplo não violam a suposição de homocedasticidade.

Figura 3.3: Gráfico dos resíduos versus valores ajustados, resíduos estudentizados versus quantil teórico e histograma dos resíduos da regressão linear em que as variáveis climatológicas são utilizadas para predizer a evaporação do solo



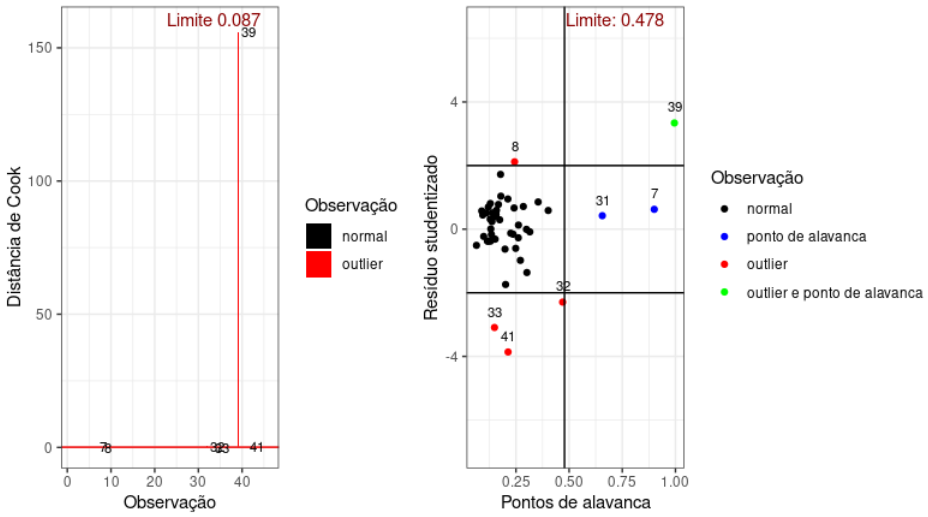
O *qq-plot* dos resíduos estudentizados, Figura 3.3 (B), indica que a normalidade dos resíduos pode estar sendo violada, sobretudo por causa da presença das observações atípicas nas duas caudas da distribuição. Essas observações tornam a caudas da distribuição dos resíduos mais pesada, como pode ser visto na Figura 3.3 (C). O teste de *Shapiro-Wilk* apontou que os resíduos não são normalmente distribuídos, com um valor-p de aproximadamente zero. Logo, é possível afirmar que os resíduos modelo de regressão em questão violam a suposição de normalidade.

Já o teste de *Durbin-Watson* resultou num valor-p de aproximadamente zero, apontando que há autocorrelação serial nos resíduos a um nível de 5%. Ou seja, há violação do pressuposto de independência.

Pelo gráfico de distância de Cook na Figura 3.4, podemos notar que há observações que estão impactando as estimativas dos coeficientes do modelo. Em especial a observação 39, que apresenta uma distância de Cook bem discrepante das demais. No gráfico de resíduos estudentizados versus pontos de alavanca também na Figura 3.4, há quatro observações que são indicadas como *outliers*, duas observações que estão sendo indicadas como ponto

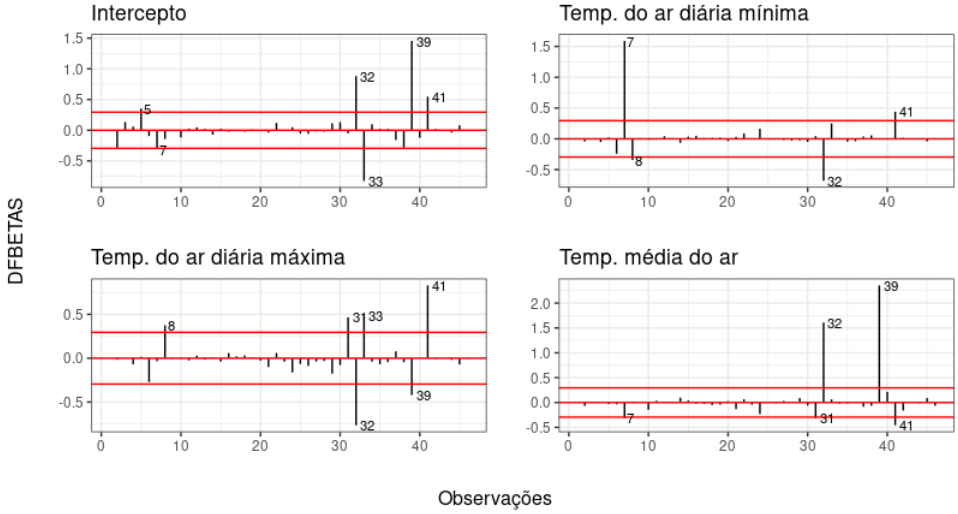
de alavanca e a observação 39 que está sendo indicada como sendo *outlier* e ponto de alavanca.

Figura 3.4: Gráfico de distância de Cook, gráfico dos pontos de Alavanca e Resíduo estudentizado da regressão linear em que as variáveis climatológicas são utilizadas para predizer a evaporação do solo



No gráfico dos DFBETAS na Figura 3.5, as observações 32 e 41 aparecem nos quatro gráficos (intercepto, temperatura do ar diária mínima, temperatura diária máxima, e temperatura média do ar). Ou seja, essas observações podem ser consideradas como ponto influentes para essas variáveis, o que significa que elas impactam na estimativa dos coeficientes das mesmas.

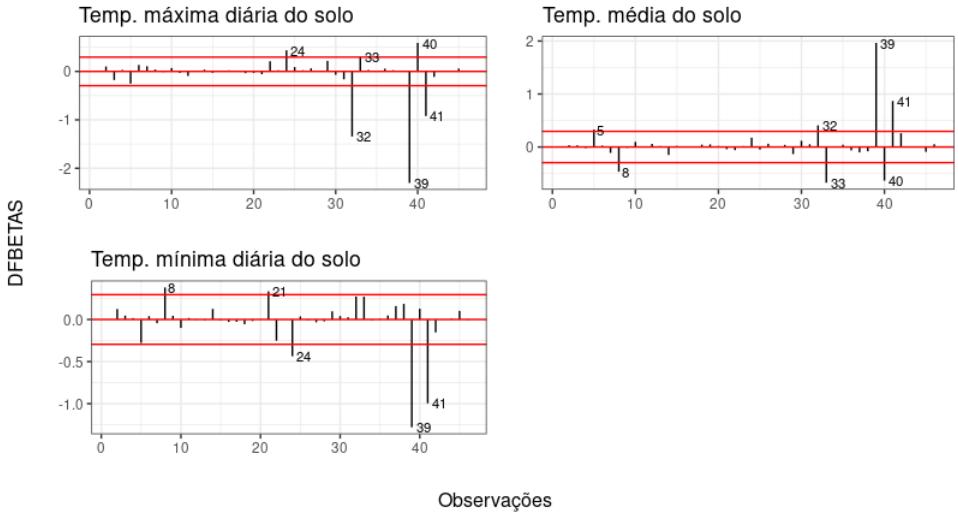
Figura 3.5: Gráfico dos DFBETAS da regressão linear em que as variáveis climatológicas são utilizadas para predizer a evaporação do solo



Nos gráficos DFBETAS para a temperatura máxima diária do solo, temperatura média do solo e temperatura mínima diária do solo na Figura 3.6, as observações 39 e 41 foram

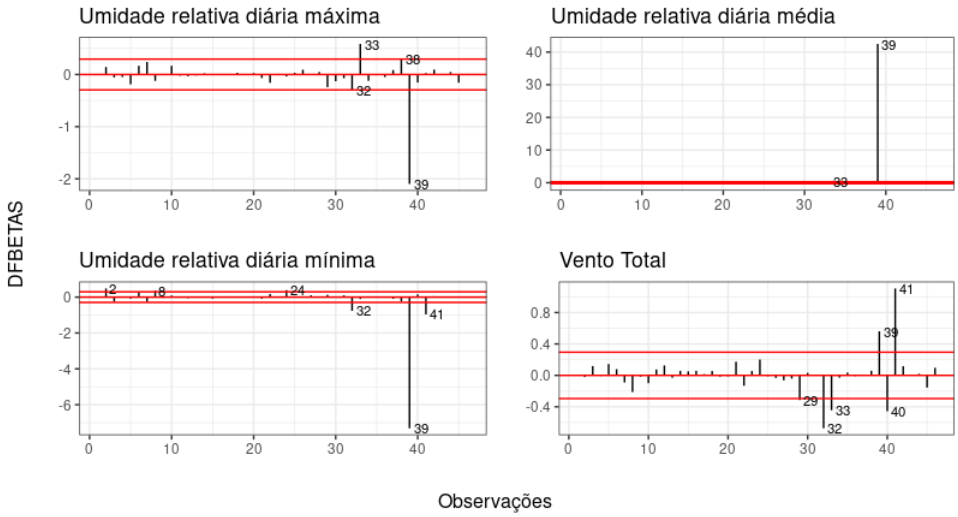
consideradas como pontos influentes.

Figura 3.6: Gráfico dos DFBETAS da regressão linear em que os dados climatológicos são utilizadas para predizer a evaporação do solo



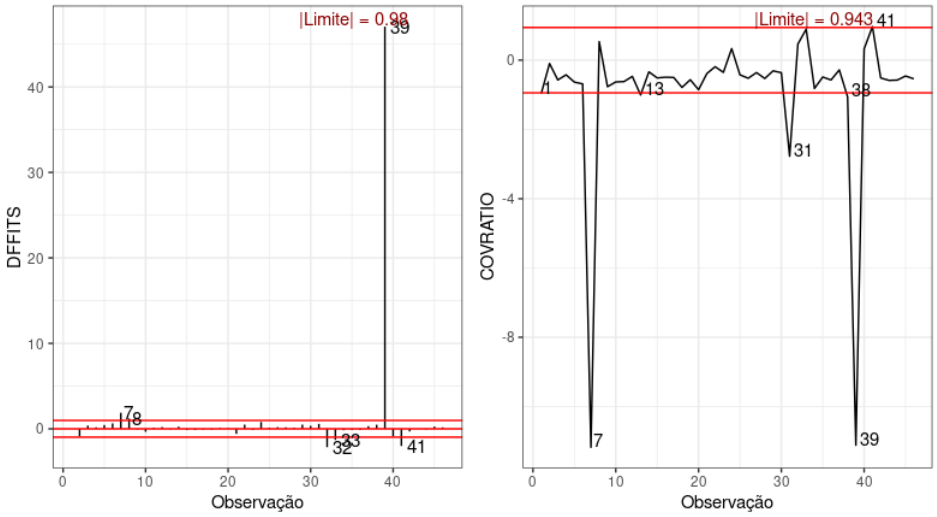
Para o gráfico dos DFBETAS da umidade relativa diária máxima, umidade relativa média e umidade relativa diária mínima da Figura 3.7, a observação 39 destoa como sendo a que mais influencia as estimativas dos coeficientes dessas variáveis. O gráfico de DFBETAS do vento total, também na Figura 3.7, mostra as observações 39 e 41 como sendo as mais influentes para esta variável

Figura 3.7: Gráfico dos DFBETAS da regressão linear em que os dados climatológicos são utilizadas para predizer a evaporação do solo



Nos gráficos de DFFITS e do COVARATIO, a observação 39 se apresenta novamente como um forte ponto atípico, afetando os valores preditos e a covariância das estimativas dos coeficientes. Menção também para a observação 7 que se apresentou fortemente como ponto atípico no gráfico COVARATIO.

Figura 3.8: Gráfico de DFFITS e gráfico do COVARATIO da regressão linear em que as variáveis climatológicas são utilizadas para predizer a evaporação do solo



3.5 Colinearidade

Para avaliar se há colinearidade no modelo, é possível avaliar o R_j^2 dos modelos onde cada variável explicativa está em função das demais variáveis explicativas. Com esse R_j^2, é possível calcular o fator de inflação de variância (VIF). Em um cenário ideal, todos os VIF's deveriam ser iguais ou bem próximos a 1, entretanto, muita das vezes, as variáveis explicativas estão bem correlacionadas entre si.

Na Tabela 3.4, é possível observar que há VIF maiores 10, o que indica que há multicolinearidade no modelo.

Tabela 3.4: Coeficiente de determinação e valores VIF

Variável	R_j^2	VIF
Temp. do ar diária máxima	0,81	5,30
Temp. do ar diária mínima	0,31	1,45
Temp. média do ar	0,82	5,43
Temp. máxima diária do solo	0,92	12,64
Temp. mínima diária do solo	0,86	6,93
Temp. média do solo	0,95	18,89
Umidade relativa diária máxima	0,09	1,10
Umidade relativa diária mínima	0,63	2,68
Umidade relativa média	0,09	1,10
Vento Total	0,26	1,35

3.6 Novo modelo de regressão linear

Com base nas análises descritivas e de pontos de influência no modelo e na análise de multicolinearidade, foram realizados alguns ajustes na base de dados a fim de melhorar o modelo de regressão linear múltiplo. Optou-se por excluir a observação 39 que se mostrou como um ponto influente em quase todas as métricas utilizadas. E foram eliminadas as variáveis explicativas com um R_j^2 maior que 0,85 (ou um VIF maior que 6,66), ou seja, as variáveis temperatura máxima e mínima diária do solo e temperatura média do solo foram retiradas do modelo.

A sumarização da ANOVA do novo modelo na Tabela 3.5 indica que este modelo de regressão linear múltiplo tem um ajuste global a variável dependente ao nível de significância de 5%.

O coeficiente de determinação (R^2) indica que 80,05% da variabilidade da evaporação do solo é explicada por esse novo modelo de regressão linear múltiplo. Esse valor é levemente maior que o coeficiente determinação obtido no modelo anterior. Isso é devido a eliminação da observação 39. Em um cenário que apenas as variáveis explicativas fossem eliminadas, o coeficiente de determinação do completo sempre será maior que o reduzido.

Tabela 3.5: Tabela ANOVA da nova regressão linear

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrado médio	Estatística F	Valor-p
Regressão	7	7453,00	1064,71	21,34	<0,001
Resíduos	37	1845,80	49,89		
Total	44	9298,80			

Na Tabela 3.6, é possível observar que apenas as variáveis umidade relativa média (*AVH*) e vento total (*WIND*) foram significativas ao nível de significância de 5%. As variáveis temperatura média do ar (AVAT) e umidade relativa diária mínima (MINH) que foram significativas no primeiro modelo, deixaram de ser significativas nesse segundo modelo.

O coeficiente do intercepto aponta que, quando todas as variáveis independentes são iguais a zero, o valor estimado da evaporação do solo é 69,13. A interpretação dos coeficientes angulares do modelo, que são significativos ao nível de 5% são:

- Considerando todas as outras variáveis constantes, o aumento de uma unidade na umidade relativa média está associado a uma diminuição de 0,36 unidades no valor estimado da evaporação do solo;
- Considerando todas as outras variáveis constantes, o aumento de uma unidade

Tabela 3.6: Sumário da Regressão: variáveis climatológicas para predizer a evaporação do solo

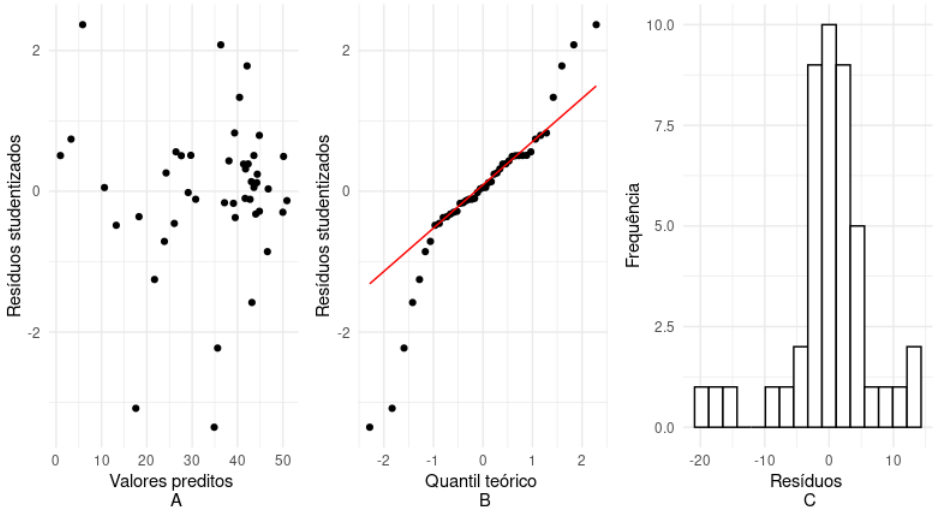
Coeficientes	Coeficiente	EP	Est. T	Valor-p
Intercepto	69,13	100,69	0,69	0,50
MAXAT	0,11	0,58	0,20	0,85
MINAT	-0,11	0,27	-0,40	0,69
AVAT	0,20	0,16	1,24	0,22
MAXH	0,62	1,03	0,60	0,55
MINH	0,06	0,36	0,17	0,87
AVH	-0,36	0,13	-2,72	**0,01**
WIND	0,02	0,01	2,14	**0,04**

no vento total está associado ao aumento de 0,02 unidades no valor estimado da evaporação do solo.

3.7 Resíduos do novo modelo de regressão

Observando a Figura 3.9 (A), é possível notar que ainda há pontos afetando a suposição de variância constante dos resíduos. Mas, o teste de *Goldfeld-Quand*, assim como no modelo anterior, indicou que os resíduos são homocedásticos (valor-p = 0,21) a um nível de 5%. Ou seja, é possível afirmar que essa suposição não está sendo violada pelo modelo de regressão linear múltiplo.

Figura 3.9: Gráfico dos resíduos versus valores ajustados, resíduos estudentizados versus quantil teórico e histograma dos resíduos da regressão linear em que as variáveis climatológicas são utilizadas para predizer a evaporação do solo

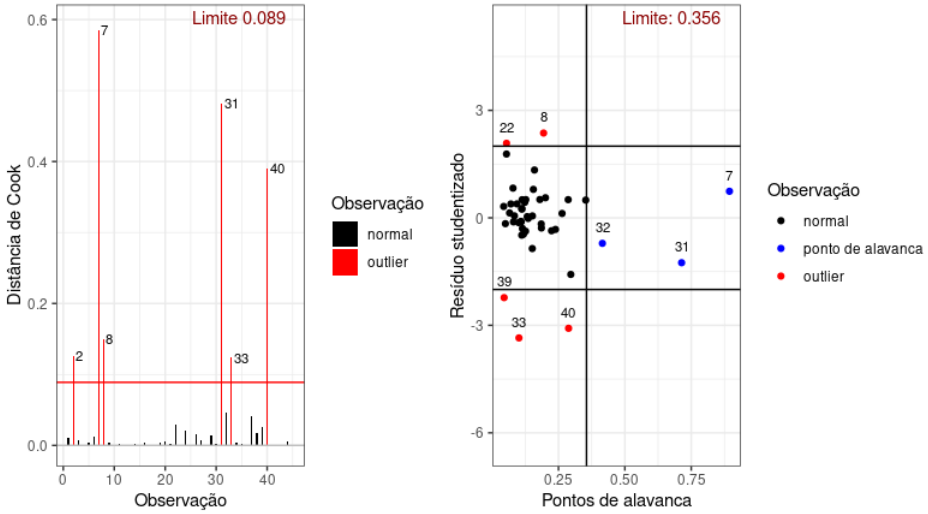


Observando ainda a Figura 3.9 (B), é perceptível ainda há observações que podem ser consideradas não usuais, mas são menos observações que o registrado no modelo anterior. As caudas da distribuição dos resíduos continuam pesadas, como também pode ser visto na Figura 3.9 (C). O teste de *Shapiro-Wilk* apresentou um valor-p de aproximadamente zero, rejeitando a hipótese de que os resíduos são normalmente distribuídos. Então, assim como no modelo anterior, os resíduos desse modelo de regressão violam a suposição de normalidade.

Já o teste de *Durbin-Watson* resultou num valor-p de aproximadamente 0,07, apontando que não há autocorrelação serial nos resíduos a um nível de 5%. Ou seja, não há violação do pressuposto de independência.

No gráfico da distância de Cook na Figura 3.10, podemos notar que há observações que são pontos influentes nas estimativas de todos os coeficientes de modo geral. Também na Figura 3.10, o gráfico dos resíduos estudentizados contra os pontos de alavanca indica que há cinco pontos que são considerados apenas como *outliers* e três pontos que são considerados apenas como pontos de alavanca.

Figura 3.10: Gráfico de distância de Cook, gráfico dos pontos de Alavanca e Resíduo estudentizado da regressão linear em que as variáveis climatológicas são utilizadas para predizer a evaporação do solo



Nos gráficos dos DFBETAS do intercepto e das variáveis temperatura do ar diária mínima e máxima e temperatura média do ar na Figura 3.11, podemos observar que nenhuma observação está presente como ponto influente simultaneamente nos quatro gráficos. Entretanto, as observações 7, 31 e 33 aparecem como pontos influentes em três dos quatro gráficos.

Figura 3.11: Gráfico dos DFBETAS da regressão linear em que as variáveis climatológicas são utilizadas para predizer a evaporação do solo

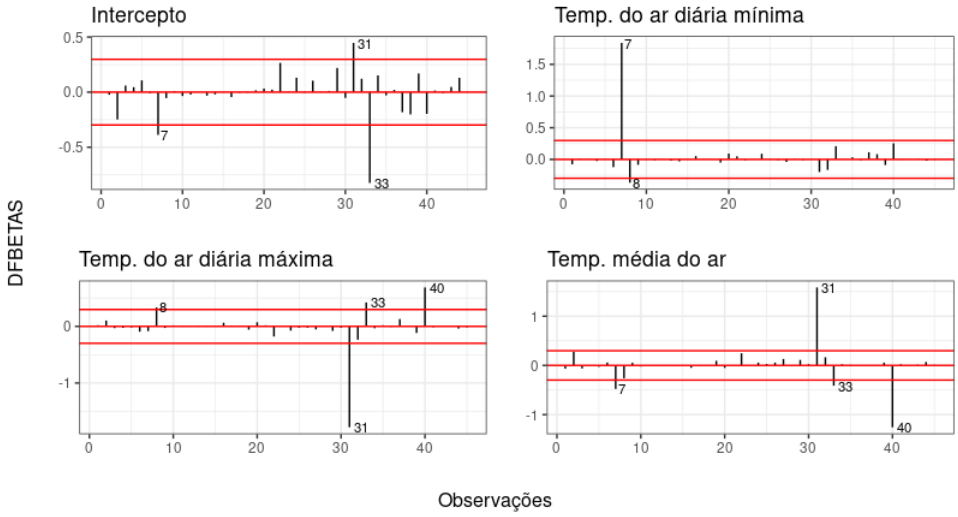
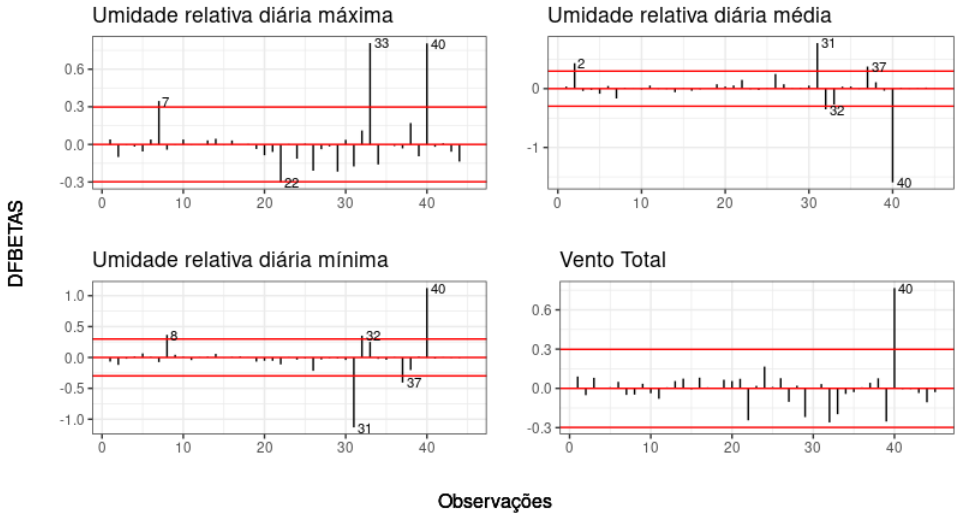


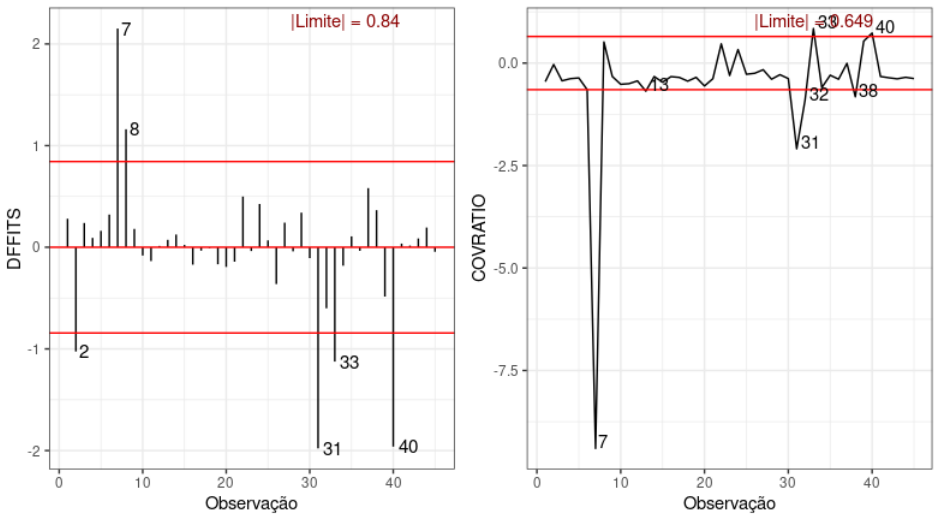
Figura 3.12: Gráfico dos DFBETAS da regressão linear em que as variáveis climatológicas são utilizadas para predizer a evaporação do solo



A Figura 3.12 apresenta os gráficos DFBETAS para as variáveis umidade relativa diária máxima e mínima e umidade relativa média. Podemos observar que a observação 40 está presente como ponto influente nos quatro gráficos. Portanto, esta observação está afetando a estimativa dos coeficientes das quatro variáveis. Importante ressaltar que essa observação 40 era a observação 41 no modelo anterior.

No gráfico de DFFITS da Figura 3.13, podemos observar que as observações 7, 31 e 40 destoam como pontos influentes, ou seja, são pontos que afetam bastante os valores preditos. No gráfico do COVARATIO também na Figura 3.13, é possível notar que a observação 7 também destoa como ponto influente, ou seja, como uma observação que afeta bastante a covariância das estimativas dos coeficientes.

Figura 3.13: Gráfico de DFFITS e gráfico do COVARATIO da regressão linear em que as variáveis climatológicas são utilizadas para predizer a evaporação do solo



4 Conclusão

4.1 Atividade 1

Em conclusão, a análise descritiva dos dados revelou que as variáveis corrente de ar refrigerado e temperatura de resfriamento apresentam uma forte correlação positiva com a variável resposta, eficiência total da indústria. Por outro lado, a concentração de ácido não demonstrou um efeito significativo na eficiência total.

O modelo de regressão linear múltipla confirmou esses resultados, mostrando que a corrente de ar refrigerado e a temperatura de resfriamento são preditores significativos da eficiência total da indústria, enquanto a concentração de ácido não apresentou um efeito significativo.

A análise dos resíduos indicou que os pressupostos de homocedasticidade e normalidade dos erros foram atendidos, embora tenha sido observada autocorrelação serial nos resíduos. Além disso, algumas observações foram identificadas como influentes, destacando a necessidade de uma análise mais aprofundada desses casos.

Com base na análise de regressão parcial e resíduos parciais, foi confirmada a importância das variáveis corrente de ar refrigerado e temperatura de resfriamento no modelo de regressão, enquanto a variável concentração de ácido não demonstrou ser relevante.

Em suma, este estudo mostrou que a corrente de ar refrigerado e a temperatura de resfriamento são os principais fatores que afetam a eficiência total da indústria, enquanto a concentração de ácido não exerce um papel significativo nesse contexto. Essas descobertas podem ser úteis para orientar estratégias de otimização e melhorias na eficiência operacional da indústria em questão.

4.2 Atividade 2

Em conclusão, a análise dos dados diários sobre a evaporação do solo revelou informações importantes sobre as variáveis climatológicas relacionadas a esse processo. Os *boxplots* permitiram identificar a presença de observações não usuais, ou *outliers*, em algumas variáveis, como a temperatura mínima diária do solo e a umidade relativa média. Essas observações discrepantes podem indicar situações climáticas atípicas ou erros na coleta de dados, além de que essas observações podem ser indicadas como ponto influentes no modelo, o que aconteceu.

Além disso, na matriz de correlação foi observado que as temperaturas do solo e do ar estão fortemente correlacionadas entre si, assim como a umidade relativa diária máxima e mínima. Portanto, as variáveis explicativas estão correlacionadas entrei si. Por outro lado, a evaporação do solo apresentou uma alta correlação positiva com as temperaturas máximas diárias do ar e do solo e com as temperaturas médias do ar e do solo, além de uma correlação negativa com a umidade relativa mínima.

O modelo de regressão linear múltiplo com a base completa foi significativo para explicar a evaporação do solo com nível de significância de 5%. Entretanto apenas duas variáveis climatológicas, a temperatura média do ar e a umidade relativa diária mínima, mostraram-se estatisticamente significativas para predizer a evaporação do solo. Na análise de resíduos, mostrou-se que as suposições de normalidade e independência foram violadas. As observações 39 e 41 foram indicadas como pontos influentes na maior parte das métricas utilizadas. E na análise de colinearidade, foi indicado, através do VIF, que havia multicolinearidade no modelo.

Com base nas análises feitas com o modelo ajustada na base de dados completa, decidiu em retirar a observação 39, que se apresentou como ponto influente, e as variáveis temperatura máxima e mínima diária do solo além da temperatura média do solo. Ajustou-se novamente o modelo de regressão linear múltiplo. O modelo foi significativo como foi anteriormente, porém apenas duas variáveis climatológicas, a umidade relativa media e o vento total, mostraram-se estatisticamente significativas para predizer a evaporação do solo. Dessa vez apenas a suposição de normalidade foi violada. As observações 7,31 e 40 se apresentaram como pontos influentes, sendo que a observação 40 era a observação 41 para o 1° modelo.

Comparado com o 1° modelo, o modelo novo obteve um coeficiente de determinação (R^2) levemente maior que o anterior, o que se deve a eliminação da observação 39. Entretanto, o novo modelo não foi capaz de aumentar a quantidade de variáveis significativas no modelo, apenas mudou quais foram as duas variáveis significativas. Portanto, se faz necessário verificar se eliminar outra observação considerada influente ou se mudar o

critério para seleção das variáveis pode melhor o modelo de regressão linear múltiplo. Se não, o modelo de regressão linear não é adequado para esses dados.