Roteiro

Camille Menezes

Junho de 2023

1 Introdução - Atv 1

- A primeira atividade consistiu na análise de regressão linear múltipla dos dados de uma indústria que realiza oxidação de amônia em ácido nítrico;
- As variáveis explicativas no estudo são: a corrente de ar refrigerado, a temperatura de resfriamento da água e a concentração de ácido;
- A variável resposta é o percentual de amônia que escapa da absorção, que é uma medida inversa da eficiência da industrial;
- Fizemos uma análise descritiva, ajustamos o modelo de regressão linear, analisamos dos resíduos para verificar se as suposições do modelo são atendidas e identificar possíveis pontos influentes ou *outliers*;
- Além de alguns gráficos da regressão parcial e dos resíduos parciais.

2 Análise descritiva

- Esses são os *boxplots* das variáveis presentes no estudo;
- A distribuição dos valores da variável corrente de ar refrigerado parece seguir uma distribuição aproximadamente simétrica, com a presença de dois pontos discrepantes.
- Parece haver uma ligeira assimetria a direita na temperatura de resfriamento, há uma maior variabilidade acima da mediana;
- O contrário da concentração de ácido que apresenta maior variabilidade abaixo da mediana;
- Já a distribuição da variável resposta é aparentemente simétrica, mas há dois pontos que fogem da tendência;
- Os coeficiente de correlação indicam que as variáveis corrente de ar refrigerado e temperatura de resfriamento estão fortemente correlacionadas positivamente com a variável resposta (eficiência total). Já a concentração de ácido apresenta uma correlação moderada com a variável resposta;
- Observando ainda os gráficos de dispersão, é possível notar que o ar refrigerado e a temperatura de resfriamento apresentam uma relação linear positiva relativamente forte com a variável resposta;
- Já a eficiência industrial apresenta uma relação não linear com a variável resposta.

3 Modelo de regressão linear I

- Essa tabela sumariza a regressão linear em que essas variáveis são utilizadas para predizer a eficiência industrial;
- Apenas a concentração de ácido não foi significativa ao nível de 5%;
- o coeficiente do intercepto aponta que quando todas as variáveis independentes são iguais a zero, o valor estimado da eficiência industrial é -39,92;
- 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;

4 Modelo de regressão linear II

- Olhando para o valor-p do teste F, apresentado na tabela ANOVA é possível notar que, a um nível de 5%, rejeitamos a hipótese de igualdade de todos os coeficientes angulares do modelo a zero.
- O coeficiente de determinação foi 0,91 e o coeficiente de determinação ajustado foi 0,91. Isso indica que 91% da variabilidade da eficiência industrial pode ser explicada por esse modelo de regressão.

5 Análise dos resíduos I

- O gráfico de resíduos contra valores preditos mostra que os resíduos não parecem ter uma variabilidade constante. É possível notar um possível *outlier*, com valor absoluto maior que três. O valor-p do teste de *goldfeld-quandt* foi 0,93, então não rejeitamos a hipótese de homocedasticidade desses resíduos a um nível de 5%. Esses resíduos não violam essa suposição.
- É possível observar esse *outlier* também no *qq-plot*, fora esse ponto, os resíduos seguem uma tendência de normalidade. O *Shapiro-Wilk* confirma isso com um valor-p de 0,82;
- Já o *durbin-watson* resultou num valor-p de aproximadamente 0,04. Ou seja, a um nível de 5% rejeitamos a hipótese nula de que não há autocorrelação serial nos resíduos. Há violação do pressuposto de independência.

6 Análise dos resíduos II

- Observando o gráfico da distância de Cook podemos notar que há um ponto influente, a observação 21. Isso é, se retirássemos essa observação do conjunto de dados e ajustássemos o modelo de regressão novamente, os coeficientes obtidos significativamente diferentes aos obtidos anteriormente;

- O gráfico do resíduo estudentizado versus pontos de alavanca também aponta essa observação além da observação 4. A 17 é um ponto de alavanca, um ponto extremo no espaço das variáveis explicativas.

7 Análise dos resíduos III e IV

- O gráficos dos DFBetas aponta novamente como ponto influente a observação 21 para a temperatura de resfriamento e a corrente de ar refrigerado;
- E a observação 4 que também apareceu anteriormente.
- A 21 também aparece novamente no gráfico dos DFFits e no COVARATIO.
- Como essa observação foi considerada não usual por diferentes métricas, é aconselhável ajustar novamente o modelo de regressão linear múltiplo sem ela e verificar o impacto da sua retirada.

8 Regressão parcial e resíduos parciais

- Esse é o gráfico da regressão parcial.
- Ele aponta a necessidade de inclusão da variável X1: corrente de ar refrigerado e X2: temperatura de resfriamento, uma vez que os pontos apresentam uma tendência de linearidade.
- Já os pontos no terceiro gráfico não apresentam nenhuma tendência clara, apontando que não talvez não seja necessário incluir a concentração de ácido no modelo, essa variável foi a que não foi significativa também.

9 Introdução

- A segunda atividade tem como objetivo avaliar a multicolinearidade no modelo de regressão linear múltiplo de dados sobre a evaporação do solo;
- A variável resposta é a evaporação do solo e as variáveis explicativas são: