Universidade Federal Rural de Pernambuco

Departamento de Estatística e Informática

Bacharelado em Sistemas de Informação

Estudo de Correlação e Regressão

Rodolfo Viegas de Albuquerque

Recife

Outubro de 2020

Resumo

Este trabalho visa analisar a planilha de dados Gr6_DIAMANES.xls, que contém informações sobre: preços, quilates, profundidade, planura, cor e claridade. Neste trabalho serão analisados os atributos Preço, Quilate e Profundidade. Para estes será aplicado o índice de correlação de Pearson, verificando o grau de correlação linear entre eles. Além desse método serão aplicadas a regressão simples e múltipla e estabelecendo qual desses métodos melhor modela funcionalmente os atributos.

Palavras chaves: preço; quilate; profundidade; correlação; regressão.

1. Introdução

1.1 Apresentação e Motivação

Cientistas, engenheiros, operadores do mercado financeiro todos esses, em seus ofícios, costumam analisar os mais variados tipos de dados e como esses relacionam entre si. Tais profissionais necessitam de ferramentas que ajudem a determinar se variáveis possuem alguma relação que possa explicar fenômenos ou aplicar na produção. A estatística inferencial possui as armas, dentre essas a correlação e regressão são bastante utilizadas. Essas, de modo quantitativo, ajudam a verificar com duas grandezas relacionam-se. Qual a tendência de uma caso outra aumente? Ela diminuirá? Ou aumentará? E em quanto? Estas perguntas podem ser respondidas com o estudo da correlação e regressão entre variáveis, assim auxiliando o trabalho daqueles profissionais citados anteriormente.

Este trabalho analisará a planilha de dados Gr6_DIAMANES.xls fornecida pelo Professor Dr. Lucian Bejan. Para a análise e inferência serão utilizadas as bibliotecas Pandas, Matplotlib, Seaborn, Scikit-Learn da linguagem de programação Python as técnicas foram extraídas de Bruce, Bruce e Gedeck (2020). Os códigos serão rodados no Notebook do Google Colab, onde é oferecido de modo gratuito um espaço para profissionais e estudantes de ciência de dados.

2. Objetivos

2.1 Objetivos Gerais

Aplicar os conceitos de correlação e regressão aos dados da planilha e analisar os resultados.

2.2 Objetivos Específicos

Descrever e definir a técnicas de correlação e regressão;

Aplicar os conceitos de correlação e regressão na planilha Gr6_DIAMANES.xls.;

Analisar os resultados.

Parte A
3. Pré-Processamento

Uma parte importante da análise de dados é o pré-processamento. A planilha possui 6 colunas com 30 linhas com dados do tipo: inteiros (int), e texto (object). Este trabalho focará nas colunas PREÇO (Y), QUILATE (X1) e PROFUNDIDADE (X2), assim excluindo as demais. Para o funcionamento das bibliotecas com estes dados as colunas que são object serão primeiro modificados substituído as vírgulas por pontos. Em seguida mudando o tipo texto para números reais (float).

Figura 1 - Planilha com as 10 primeiras linhas
após seleção de colunas e modificação dos tipos

	PREÇO	QUILATE	PROFUNDIDADE
0	6958	1.00	60.5
1	5885	1.00	59.2
2	6333	1.01	62.3
3	4299	1.01	64.4
4	9589	1.02	63.9
5	6921	1.04	60.0
6	4426	1.04	62.0
7	6885	1.07	63.6
8	5826	1.07	61.6
9	3670	1.11	60.4

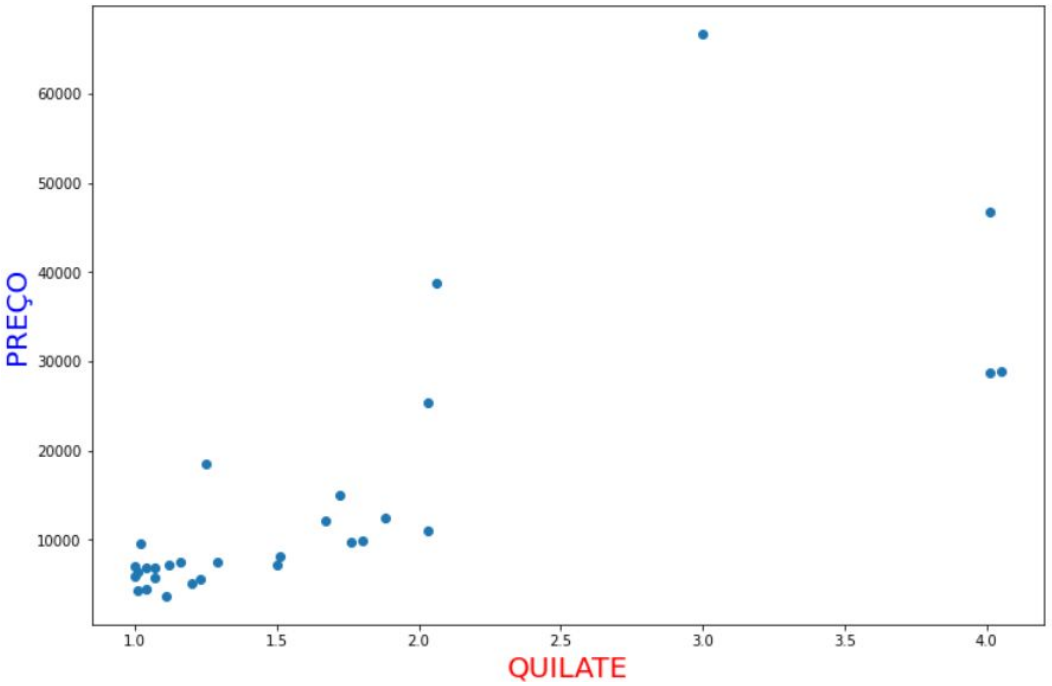
Compilação do autor

4. Correlação

Quando duas variáveis, X e Y, caminha num mesmo sentido, dize-se que elas são correlacionadas positivamente, ao passo que X aumenta Y aumenta também. Porém, se X diminui e Y aumenta seu valor dize-se agora que as variáveis são correlacionadas negativamente. Barbetta, Reis e Bornia (2010) [Referenciar] afirmam que correlação "refere-se a uma associação numérica entre duas variáveis, não implicando, necessariamente, relação de causa-e-efeito".

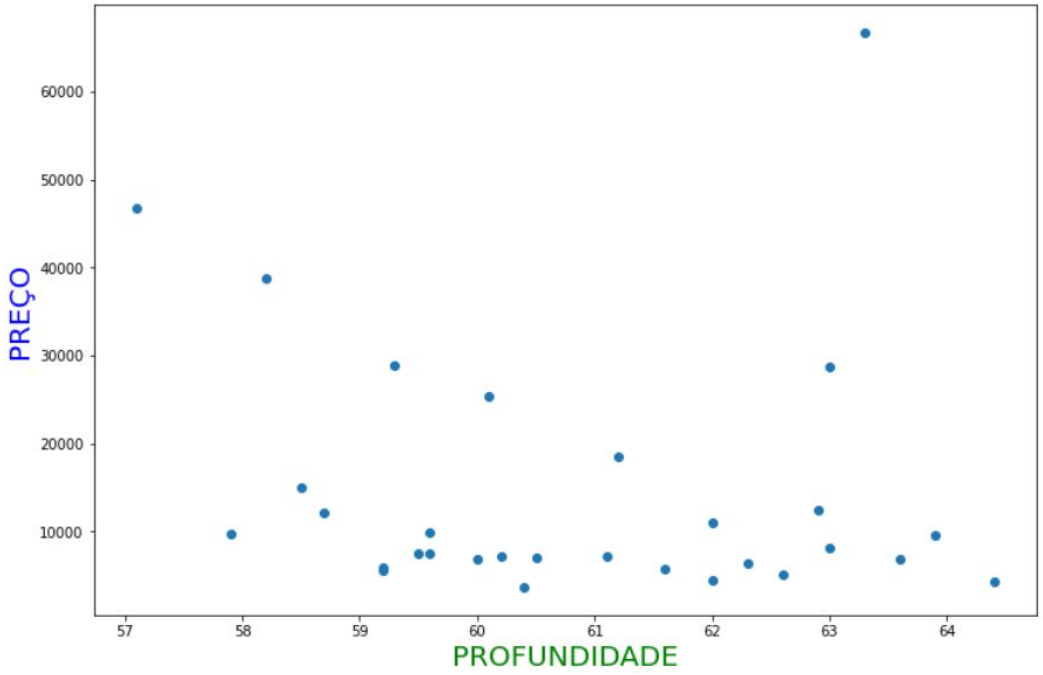
Para verificar se há alguma correlação entre duas variáveis os diagramas de dispersão são boas ferramentas. A seguir os gráficos de dispersão das variáveis preço (Y) e quilate (X1); preço (Y) e profundidade serão apresentados:

Figura 2 - Diagrama de Dispersão Preço x Quilate



Compilado pelo autor

Figura 3 - Diagrama de Dispersão Preço x Profundidade



Compilado pelo autor

Além das maneiras gráficas há uma outra maneira para buscar evidências de existência de correlação entre duas variáveis: o índice de correlação e Pearson. Representado pelo expressão:

$$r = \frac{n\sum(x_i \cdot y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2} \cdot \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

De modo manual para calcular o índice r entre duas variáveis uma boa forma é acrescentar mais colunas a tabela com os valores das variáveis dependente e independentes ao quadrado e o produto daquelas. Ao fim o somatório dos valores.

Figura 4 - Tabela com as 10 primeiras linhas e com mais colunas

	PREÇO	QUILATE	PROFUNDIDADE	PREÇO²	QUILATE²	PROFUNDIDADE²	QUILA·PREÇO	PROFU·PREÇO
0	6958	1.00	60.5	48413764	1.0000	3660.25	6958.00	420959.0
1	5885	1.00	59.2	34633225	1.0000	3504.64	5885.00	348392.0
2	6333	1.01	62.3	40106889	1.0201	3881.29	6396.33	394545.9
3	4299	1.01	64.4	18481401	1.0201	4147.36	4341.99	276855.6
4	9589	1.02	63.9	91948921	1.0404	4083.21	9780.78	612737.1

Compilação do autor

Os índice de correlação entre preço e quilate resultou em 0.7675, já preço e profundidade ficou em -0.13057.

Figura 5 - Matriz de correlação

	PREÇO	QUILATE	PROFUNDIDADE
PREÇO	1.000000	0.767486	-0.130569
QUILATE	0.767486	1.000000	-0.196695
PROFUNDIDADE	-0.130569	-0.196695	1.000000

Compilação do autor

Olhando os gráficos de dispersão e os resultados do índices é possível afirmar há correlação positiva entre preço e quilate, ou seja, quando o valor do quilate de um diamante aumenta o preço também aumenta (mas ainda assim não é possível dizer se um causa o outro). Todavia preço e profundidade possuem correlação negativa muito fraca.

O próximo passo após é testar a hipótese de existência de correlação populacional. Barbetta, Reis e Bornia (2010) descrevem o teste como.

$H_0 : \rho = 0$ em que as variáveis X e Y não possuem correlação;
$H_0 : \rho \neq 0$ há correlação entre as variáveis.

Podendo H1 ser maior ou menor que 0, indicando se a correlação é positiva ou negativa. Com n - 2 graus de liberdade e dependendo do nível de significância, coleta-se o t crítico na tabela t-Student e verifica se t observado está na região crítica. A equação de t é a seguinte:

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$$

Para a planilha dos diamante, considerando um nível de significância de 5% temos que o t-observado entre preço e quilate é 6,557; o t-crítico, olhando a tabela t-Student é 2,048. Como lidamos com uma possível correlação positiva a região crítica localiza-se na cauda direita da curva marcado com o t-crítico os 5% de significância. O valor de 6,557, sendo maior que 2,048, está na região, então rejeita-se H_0 e aceita-se H_1. Portanto há evidências de uma correlação positiva populacional entre valor do quilate e preços.

Para a correlação entre preço e profundidade o t-observado -0,7213; já o t-crítico, com significância 5%, tem o valor de -2,048 na hipótese de uma possível correlação negativa. O valor observado é maior que o crítico, não estando na região crítica e, assim, aceitando H_0. Portanto, não há evidências de que uma correlação negativa populacional entre preços e profundidade.

Parte B

5. Regressão Linear Simples

5.1.1 Preço vs Quilate

Após constatar uma possível correlação entre variáveis o próximo passo ao analisá-las é criar um modelo que expresse uma possível relação de causa-e-efeito através de uma função. A regressão linear simples é técnica a ser utilizada.

Consideremos Y como a variável dependente ou resposta e X como a variável independente ou explicativa. Quem será quem dependerá de uma hipótese ou teoria. Esta é uma das diferenças de regressão e correlação.

A construção do modelo se dá pela hipótese de o valor esperado de Y varie em função de X:

$$E(Y) = \alpha + \beta \cdot X + \varepsilon_i$$

Os parâmetros α e β do modelo de regressão e é preciso estimá-los, um dos métodos que podem o fazer é o Método dos Mínimos Quadrados. A ideia é minimizar a soma dos erros quadráticos:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - [\alpha + \beta \cdot x_i])^2$$

Barbetta, Reis e Bornia (2010) descrevem as equações que estimam os parâmetros da equação, e minimizam os erros como:

$$b = \frac{n \cdot \sum_{i=1}^{n} (x_i y_i) - (\sum_{i=1}^{n} x_i) \cdot (\sum_{i=1}^{n} y_i)}{n \cdot \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$a = \overline{y} - b\overline{x}$$

Sendo que b e a serão os parâmetros ou estimadores para α e β e a equação da regressão linear simples ficará:

$$\widehat{y} = a + bx$$

Além dessas equações Barbetta, Reis e Bornia (2010) apresentam um importante conceito - o resíduo:

$$e_i = y_i - \widehat{y}_i$$

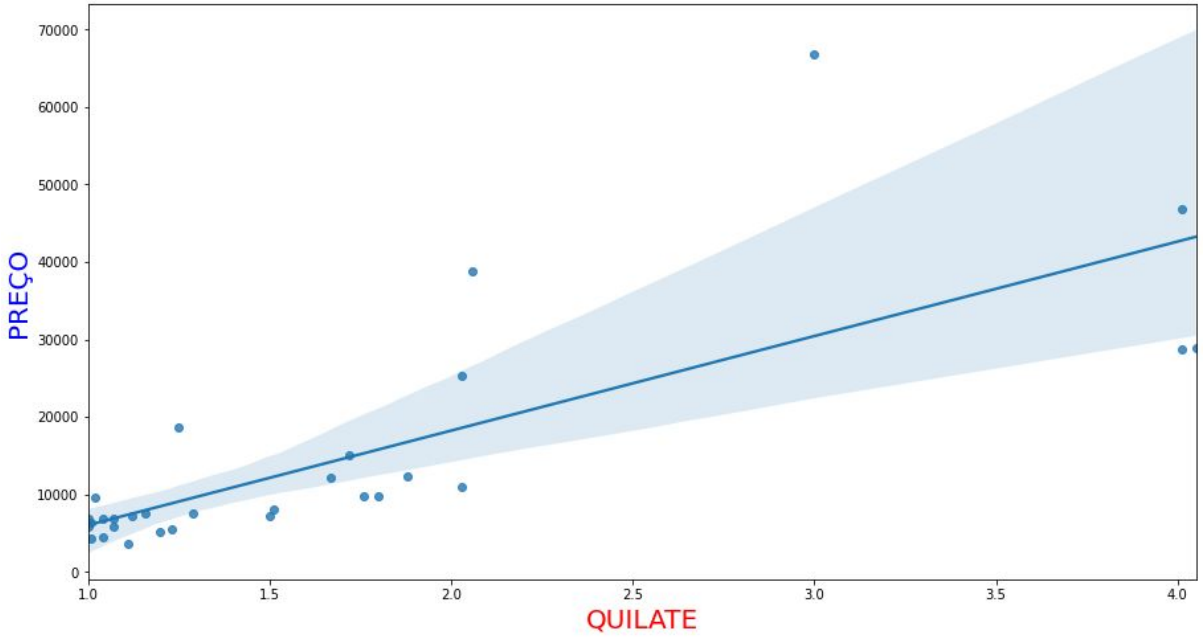
Que é a diferença entre o valor predito e o valor real podendo ser uma estimativa para o erro aleatório ε_i.

Para os valores de Preço e Quilate, utilizando os novos dados da tabela na figura 4, a reta de regressão ficará:

$$\widehat{y} = -6157,648 + 12200,50 \cdot x$$

Em que 12200,50 é o estimador b é -6157,648 o estimador a. A cada unidade de quilate calculate temos um aumento de 6042.852 no preço do diamante. O diagrama de dispersão abaixo ilustra melhor a regressão com os dados:

Figura 9 - Diagrama de dispersão com reta de regressão



Compilação do autor

O erro padrão da estimativa é definido pela equação:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

E para o exemplo da relação entre preço dos diamante e valor do quilate o erro padrão ficou em: 9479.71.

Para verificar se o quanto a regressão explica as variações em Y em função da variações de X, é preciso acrescentar 3 conceitos: variação total, variação explicada e variação não explicada.

A variação explicada se dá pela equação: $\sum(\hat{y}_i - \overline{y}_i)^2 = 3606484258{,}77$

A variação não explicada pela regressão: $\sum(y_i - \widehat{y_i})^2 = 2516216598.68$

E a variação total: $\sum(y_i - \overline{y_i})^2 = \sum(\widehat{y_i} - \overline{y_i})^2 + \sum(\widehat{y_i} - \overline{y_i})^2 = 6122700857.46$

Chama-se coeficiente de determinação (r2) o valor que mede o quanto a variação de Y em função da variação de X. A equação é:

$$r^2 = \frac{\sum(\widehat{y_i} - \overline{y_i})^2}{\sum(y_i - \overline{y_i})^2}$$

Aplicado ao exemplo de preços vs quilate temos que o r2 é igual a 58,9%.

5.1.2 Inferências sobre o Modelo

É possível fazer testes de hipótese e intervalos de confiança no modelo estimado e verificar se com uma outra amostra há a possibilidade de os resultados da regressão serem próximos. Mas para executar tais teste é preciso que o modelo esteja enquadrado em 4 suposições, são elas:

1 - A observações Yi devem ser independentes;
2 - O termo de erro deve possuir distribuição normal;
3 - Os erros devem ter média nula, ou seja, $E(\varepsilon_i) = 0$;
4 - E a variância dos erros deve ser constantes, $Var(\varepsilon_i) = \sigma^2$.

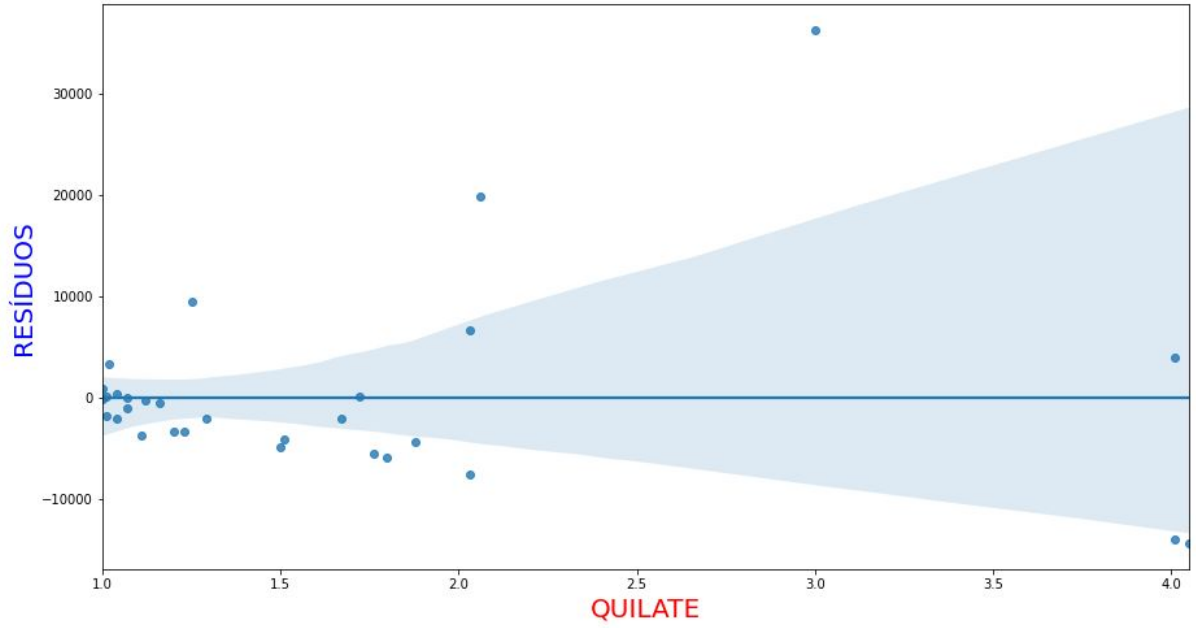
A análise gráfica dos resíduos, como explica Barbetta, Reis e Bornia (2010), é uma boa primeira evidência de que as suposições pode não estar satisfeitas, e análise da simetria das distribuições dos dados das variáveis. Porém, de antemão, a gráfico da relação entre Preço e Quilate já mostra uma possível variância inconstante, pois os pontos se distanciam uns dos outros à medida que aumentam os valores. Comecemos com os resíduos em função do Quilate.

Figura 10 - tabela com as dez primeiras observações e com nova coluna de resíduos

	PREÇO	QUILATE	RESÍDUOS(y - ŷ)
0	6958.0	1.00	915.145749
1	5885.0	1.00	-157.854251
2	6333.0	1.01	168.140726
3	4299.0	1.01	-1865.859274
4	9589.0	1.02	3302.135703
5	6921.0	1.04	390.125657
6	4426.0	1.04	-2104.874343
7	6885.0	1.07	-11.889412
8	5826.0	1.07	-1070.889412
9	3670.0	1.11	-3714.909504

Compilação do autor

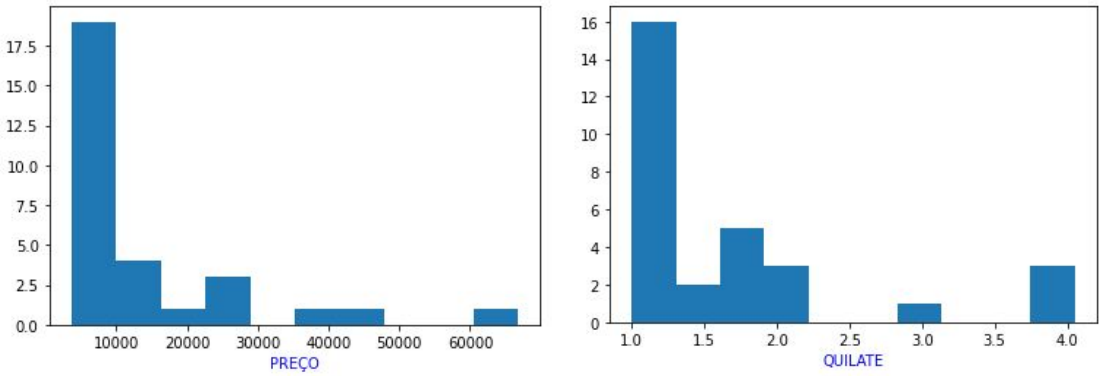
Figura 11 - Diagrama de dispersão resíduos (e) x Quilate



Compilação do autor

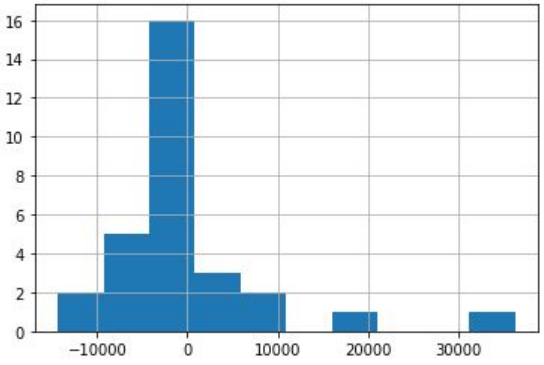
Aumentam-se as evidências quando ao plotar as distribuições de Preço e Quilate e resíduos vemos que os dados distribuem-se assimetricamente:

Figura 10 - distribuições de Preço e Quilate



Compilação do autor

Figura 11 - distribuição dos resíduos



Compilação do autor

A recomendação de Barbetta, Reis e Bornia (2010) para lidar com dados assim é usar transformação logarítmica que "aumenta as distâncias entre os valores pequenos e reduz as distÂncias entre os valores grandes, tornando distribuições com assimetria positiva (cauda mais longa à direita em distribuições aproximadamente simétricas". O modelo pode ser ajustado desta forma:

$$ln(y_i) = \alpha + \beta ln(x_i) + \varepsilon_i$$

Com os ajustes da transformação logarítmica a reta de regressão os estimadores a e b são respectivamente 8,6397 e 1,4685. Para utilizar a regressão retornando valores preditos próximos da realidade basta aplicar o resultado predito com potência da constante de Euler:

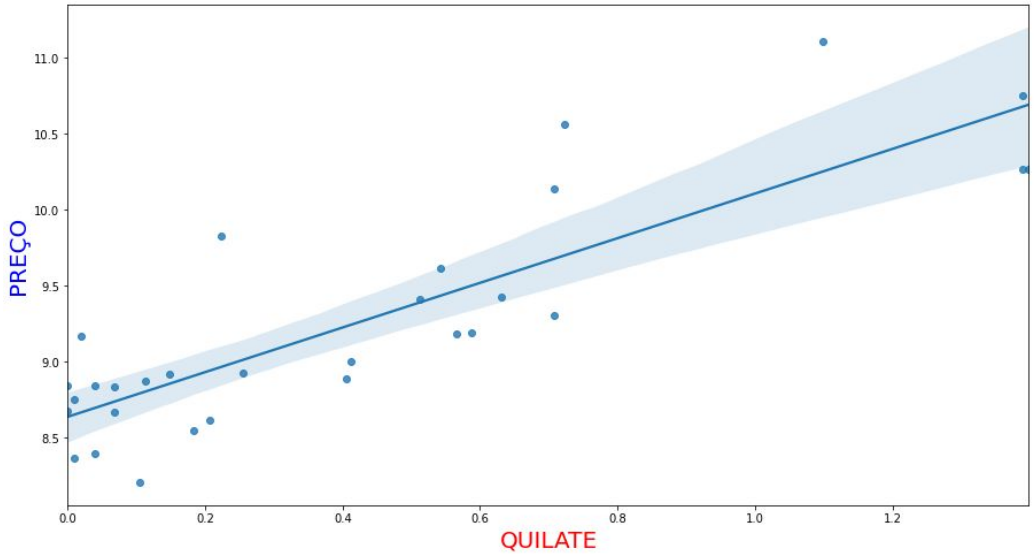
$$\widehat{y} = e^{8.6397+1.4685x}$$

Figura 12 - Tabela com as 10 primeiras linhas ajustadas e com coluna de resíduos

	PREÇO	QUILATE	RESÍDUOS(y - ŷ)
0	8.847647	0.000000	0.207897
1	8.680162	0.000000	0.040411
2	8.753529	0.009950	0.099166
3	8.366138	0.009950	-0.288225
4	9.168372	0.019803	0.499540
5	8.842316	0.039221	0.144967
6	8.395252	0.039221	-0.302097
7	8.837100	0.067659	0.097990
8	8.670086	0.067659	-0.069025
9	8.207947	0.104360	-0.585061

Compilação do autor

Figura 13 - Diagrama de dispersão ajustado com transformação logarítmica



Compilação do autor

O coeficiente de determinação, após o ajuste, ficou em:

$r^2 =$ **0,727352**.

Após os ajustes é possível testar os parâmetros da regressão, já que as suposições são válidas. Começaremos testando a hipótese existência de regressão, em que:

$$H_0 : \beta = 0$$

ou

$$H_1 : \beta > 0$$

Será calculado erro padrão de b que é a equação:

$$s_b = s_e \cdot \sqrt{\frac{n}{n\sum x_i^2 - (\sum x_i)^2}}$$

Em que s_e é o erro padrão predito já calculado anteriormente, mas agora para a nova reta ajustada é:

$$s_e = 0.39825$$

O erro padrão de b, após cálculo, resulta em 0,1699. Tendo tais valores é possível calcular o *t*-observado que é através desta expressão:

$$t = \frac{b - \beta_0}{s_b}$$

O valor do t-observado é 8,6433, já o t-crítico com *n* - 2 g.l. e 5% de significância é 2,048. Com um dentro da região crítica rejeita-se H_0 e aceita-se H_1. Ainda com esses dados é possível construir um intervalo de confiança para o estimador *b*:

$$IC = b \pm t_{cri} \cdot s_b$$

E substituindo e calculando os valores o intervalo de confiança do estimador é:

$IC = 0,3982 \pm 0.289$.

Além do teste de hipótese e intervalo de confiança para o estimador b há também aqueles para o estimado do intercepto. Comecemos com o teste de hipótese para $\alpha = 0$, isto é, se é razoável supor que o intercepto passe pela origem.

$$H_0 : \alpha = 0$$
$$H_1 : \alpha > 0$$

Para início é calculando o erro padrão do estimado a, Barbetta, Bornia e Reis (2010) definem a equação como segue:

$$s_a = s_e \cdot \sqrt{\frac{1}{n} \cdot \frac{(\sum x_i)^2}{n\sum x_i^2 - (\sum x_i)^2}}$$

O desvio padrão de a resultou em 0.39628, já o t-observado em 21.8015, com um nível de significância d e 5% e n - 2 graus de liberdade o t-crítico é 2,048 o t-observado está

na região crítica, aceitando-se a hipótese alternativa. Ou seja, há evidências de que o intercepto α populacional não passe pela origem.

Com os valores do t-crítico e o desvio padrão de a o intervalo de confiança para intercepto pode ser construído:

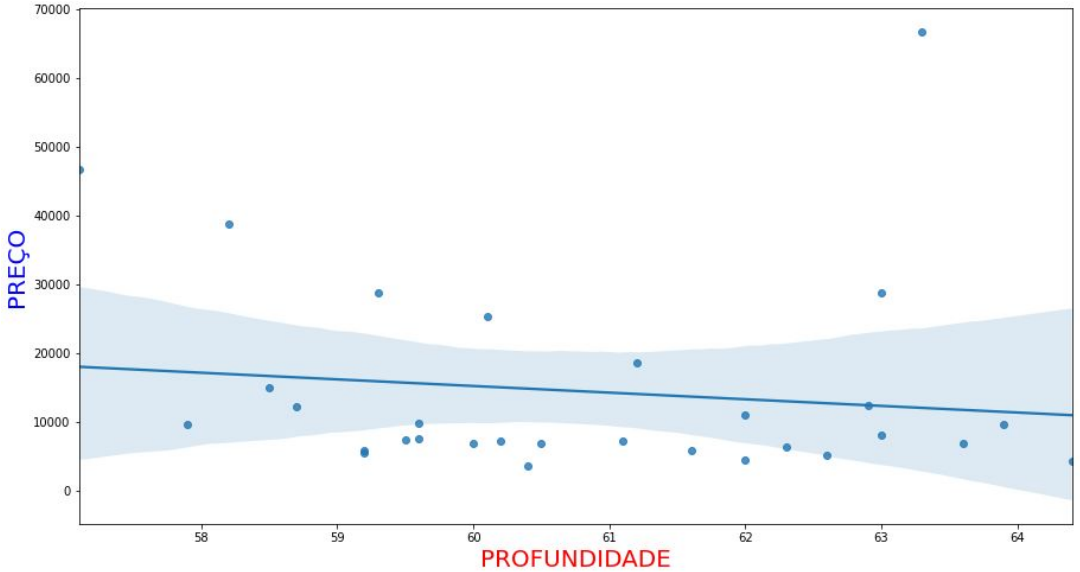
$$IC = 8.6397 \pm 0.8115$$

Encerrando assim a análise da regressão linear simples para a relação Preço e Quilate.

Parte C
5.2 Preço vs Profundidade

Os procedimentos explicados e aplicados na sessão anterior serão agora aplicados à relação entre Preço e Profundidade, verificando os resultados.

Figura 13 - Diagrama de dispersão entre Preço e Profundidade com reta de regressão



Compilação do autor

A reta de regressão ficou como:

$$y_i = 73107.009 +- 964.4277x_i$$

Figura 15 - Tabela com as 10 primeiras linhas e com a coluna dos resíduos

	PREÇO	PROFUNDIDADE	RESÍDUOS(y-ŷ)
0	6958.0	60.5	-7801.127838
1	5885.0	59.2	-10127.883966
2	6333.0	62.3	-6690.157814
3	4299.0	64.4	-6698.859452
4	9589.0	63.9	-1891.073348
5	6921.0	60.0	-8320.341733
6	4426.0	62.0	-8886.486151
7	6885.0	63.6	-4884.401685
8	5826.0	61.6	-7872.257268
9	3670.0	60.4	-11185.570617

Compilação do autor

O erro padrão estimado para esta reta $s_e = 14660,832$

$$\sum(\widehat{y_i} - \overline{y_i})^2 = 104380964.969$$

$$\sum(y_i - \widehat{y_i})^2 = 6018319892.497$$

$$\sum(y_i - \overline{y_i})^2 = \sum(\widehat{y_i} - \overline{y_i})^2 + \sum(\widehat{y_i} - \overline{y_i})^2 = 6122700857.466$$

O coeficiente de determinação (r^2) ficou em : 1,7%

5.2.1 Inferência sobre o modelo

Teste de hipótese para existência de regressão, baseado no valor do índice de correlação -0,13 será testado a existência de parâmetro populacional de relação negativa:

$$H_0 : \beta = 0$$

$$H_1 : \beta < 0$$

O desvio padrão de *b* é **1383.94**

O valor de t-observado ficou em **-0.69686**

Com nível de significância de 5% e graus de liberdade de n-2, o t-crítico de -2,048, como o t-observado está fora da região crítica aceita-se a hipótese nula. Portanto não evidências de uma possível regressão populacional.

Após testar o parâmetro de inclinação é necessário igualmente o teste no parâmetro de intercepto, para a hipótese desse passar pela origem:

$$H_0 : \alpha = 0$$

$$H_1 : \alpha > 0$$

Parte D

6.1. Análise dos Resultados

Das relações entre Preço-Quilate e Preço-Profundidade, com respectivamente índices de correlação de Pearson de **0,76** e **-0,13**, há uma clara correlação entre o primeiro par(que não necessariamente implica numa relação de causa).

Adicionalmente às métricas de correlação os modelos propostos estabelecendo na hipótese de que Preço será a variável independente a relação funcional com Quilate explica melhor as variações nos preços. O erro padrão desta é menor que quando modelado Preço-Profundidade em que os resultados dos erros são: **0.3982** (após ajuste da reta via transformação logarítmica) e **14660,832**; gerando, então, melhor predições.

Os testes de hipótese também dão informações adicionais da qualidade dos modelos, para Preço-Quilate a possível existência de um coeficiente de inclinação populacional dá alguma garantia de que com novas amostras haverá uma reta semelhante à estimada. Ao contrário do modelo Preço-Profundidade que resultou na situação de aceitar a hipótese nula, evidenciando a possibilidade de não existência de regressão.

6.2. Regressão Múltipla

Uma análise mais realista de fenômenos exige que mais que duas variáveis sejam analisadas, tornando a regressão linear simples pouco realista para aplicações em pesquisa e produção. Para suprir tal lacuna a regressão múltipla surge como técnica que incorpora uma grande quantidade de variáveis independentes que criem modelos que melhor condizem com a realidade.

O modelo é semelhante à regressão simples:

$$Y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i$$

Em que Y e X1, X2...Xk são variáveis aleatórias e yi x1i, x2i,...,xkn são observações de tais variáveis. Estas são modeladas por parâmetros de inclinação $\beta_1, \beta_2, ..., \beta_k$, um intercepto α e o erro dos fatores não observados ε_i.

Segundo Barbetta, Reis e Bornia (2010) as suposições são análogas à regressão simples, com o acréscimo de que não pode haver forte correlação entre as variáveis independentes (na Figura 5 a matriz de correlação mostra que há uma fraca correlação entre Quilate e Profundidade é com r = -0,19). Para este trabalho será estimado um modelo que relacione Preços função dos atributos Quilate e Profundidade.
Três hipótese foram testadas para um melhor ajuste do modelo:

1. Preço = a + $b_{1,}$·Quilate + b_2·Profundidade
2. log(Preço) = a + b_1·log(Quilate) + b_2·Profundidade
3. log(Preço) = a + b_1·log(Quilate) + b_2·log(Profundidade)

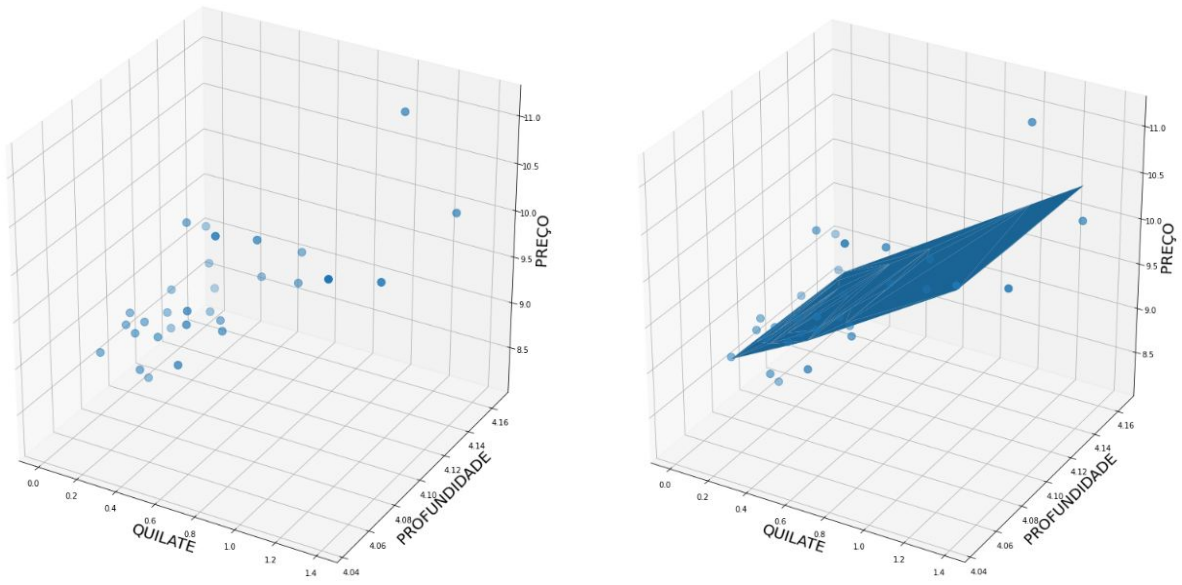
A terceira opção retornou coeficiente de determinação maior e um erro padrão estimado menor que as outras duas alternativas.

$$r^2 = 0,7278654$$
$$s_e = 0.397886$$

Estes valores foram melhores não somente em comparação com as alternativas de ajuste, mas também comparando aos modelos de regressão simples. Porém ao ser calculado o r^2 ajustado o resultado foi uma queda na qualidade:

$$r^2_{ajustado} = 0.707707$$

Figura 14 - Diagramas de dispersão com e sem o plano que representa a regressão múltipla



Compilação do autor

De acordo com o blog Minitab o r² sempre aumenta à medida que mais variáveis independentes são adicionadas, dando uma falsa impressão de melhor acurácia no modelo. "Se um modelo tem muitos preditores e polinômios de ordem superior, começa a modelar o ruído aleatório nos dados. Esta condição é conhecida com sobreajuste do modelo e produz valores de R-quadrado enganosamente altos e uma habilidade diminuída de fazer predições." Minitab Blob (2013). Para verificar se há de fato uma ganho de informação ao adicionar mais preditores o r² ajustado funciona de modo melhor.

Com valor do r² ajustado é possível dizer que o modelo piorou ao adicionar a variável Profundidade ficando com um percentual de variação explicada menor que a regressão simples de Preço vs Quilate.

7. Conclusão

A conclusão deste trabalho, após a aplicação dos métodos estatísticos à planilha com dados sobre diamantes é que o modelo que consegue melhor explicar a variação dos Preços dos diamante é o que relaciona Preços vs Quilate do diamante - via regressão linear simples. A relação funcional entre Preços e Profundidade explicam quase nada variação do atributo dependente. E ao unir as duas variáveis num modelo de regressão múltipla a variável Profundidade piora a acurácia do modelo. Portanto, para tal dataset para realizar predições de preços de diamantes fora dos conjuntos de dados a forma mais acurada é a regressão linear

simples entre Preço e Quilate. Para trabalhos futuros de análise de previsão de preços testar com os demais atributos com conjunto de dados verificando se há melhores preditores.

Referências

Bruce, P.; Bruce A.; Gedeck P. (2020). **Practical Statistics for Data Scientists**. 2ª edição, O'Reilly Media, Estados Unidos da América.

Barbetta, P.A.; Reis, M. M.; Bornia, A.C. (2010). **Estatísticas para Cursos de Engenharia e Informática**. 3ª edição, Atlas, São Paulo.

Minitab Blog (2013). **Multiple Regression Analysis: Use Adjusted R-Squared and Predicted R-Squared to Include the Correct Number of Variables.** Disponível em: https://blog.minitab.com/blog/adventures-in-statistics-2/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-include-the-correct-number-of-variables. Acessado em: 28/10/2020.