UNIVERSIDADE
DE ÉVORA

Regressão e Classificação

Relatório Trabalho 1

Miguel Grilo 58387

Jorge Couto 58656

Colégio Luís António Verney

Conteúdo

Lista de Figuras

Introdução

No âmbito do desenvolvimento da nossa capacidade de resolução de problemas e procura de soluções por conta própria, foi-nos pedido pela escolha de uma base de dados a limpar, tratar e estudar como Regressão Linear Múltipla. Deveríamos estudar pelo menos 5 variáveis, 4 delas quantitativas e uma delas qualitativa, sendo uma das quantitativas a nossa variável resposta.

Apesar de saltarmos de base de dados para base de dados, retiradas do Kaggle, acabamos por escolher a base de dados *Diamonds* do ggplot2, a qual oferece amostras de mais de 50 000 diamantes quanto a preço, qualidade de corte, peso e outros.

O nosso dever, como futuros estatísticos, foi o de, com esses dados, estudar a significância das variáveis, ajustar um modelo que fosse adequado e validasse os pressupostos e, é claro, avaliar o modelo quanto à sua qualidade final, robustez e facilidade de análise. Para chegarmos a esse modelo final, no entanto, tivemos de passar por várias transformações de variável de antemão visando a validação dos pressupostos (os quais foram mais difíceis de validar do que o esperado).

Algo de pertinente a mencionar, antes da explicação do nosso processo de trabalho, têm a ver com o tamanho de amostra. Apesar de *Diamonds* ter mais de 50 000 amostras, acabamos por usar apenas 1000 para não termos dificuldades a analisar os gráficos e outputs pela quantidade de amostras elevada.

Preparação do Conjunto de Dados

Questão 1

Antes de podermos começar a trabalhar com os dados, é necessária a preparação dos mesmos. Devemos selecionar as variáveis com que trabalharemos e, no caso de existirem dados omissos, limpá-los, seja substituindo os dados omissos por outros valores ou removendo os pontos.

Entre as variáveis da nossa base de dados, escolhemos trabalhar com as variáveis Price, uma variável quantitativa que indica o preço do Diamante e, inclusive, a nossa variável resposta neste estudo, *Carat*, que indica o peso do diamante, *Depth*, a variável que indica a medida da profundidade relativa do diamante (i.e., a altura em relação à largura), *Table*, a variável que mede a proporção entre o diâmetro da "mesa" do diamante e a largura e, como variável qualitativa, Cut, que avalia a qualidade do corte do diamante em 5 níveis: *Fair*, *Good*, *Very Good*, *Premium* e *Ideal*.

Como escolhemos apenas 1000 amostras entre todas as amostras da nossa base de dados, era necessário escolhermos sempre as mesmas 1000. Se pedíssemos apenas 1000 amostras, a escolha seria randomizada, o que poderia levar a alterações nos outputs. Portanto, criamos uma semente para que as amostras possam ser reproduzidas sempre de modo igual.

Questão 2

Para avaliar as amostras quanto à presença de dados omissos verificamos quanto era a soma de amostras com valor 'NA' (dado omisso) em linha ou coluna. Dando-nos 0 nos dois outputs, podemos concluir que não temos dados omissos, o que nos permitiu continuar o estudo.

Modelos de Regressão Linear Simples

Questão 3

Guiando-nos pelas alíneas colocadas no enunciado do trabalho pela professora, precisávamos agora de desenhar as nuvens de pontos para cada par de variáveis quantitativas. Fizemos isso de dois modos, um deles mais completo que o outro.

Seguindo a abordagem simples, usando o comando *pairs* com os nomes das variáveis quantitativas na base de dados (que chamámos de dados, para facilitar) que iriamos estudar conseguimos obter automaticamente todas as nuvens de pontos de cada par de variáveis quantitativas.

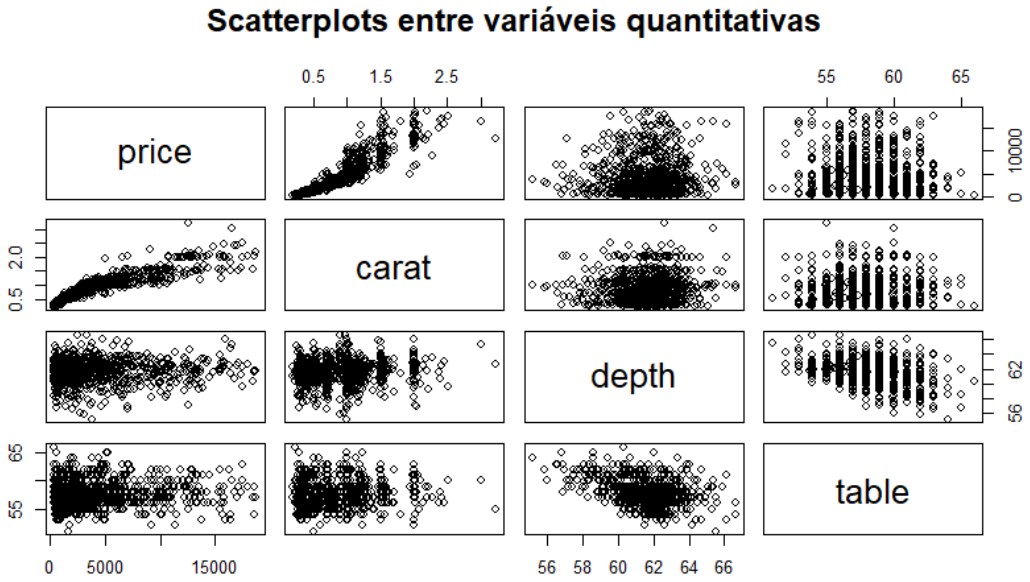

Figura 1 - Gráfico de Análise da Dispersão entre Price, Carat, Depth e Table

No entanto, seguindo uma abordagem um pouco mais complexa (isso para dizer, usando a biblioteca *GGally*), o comando *ggpairs* permite-nos obter não só as nuvens de pontos como a linha de representação da variação dos dados para cada variável, e as correlações também!

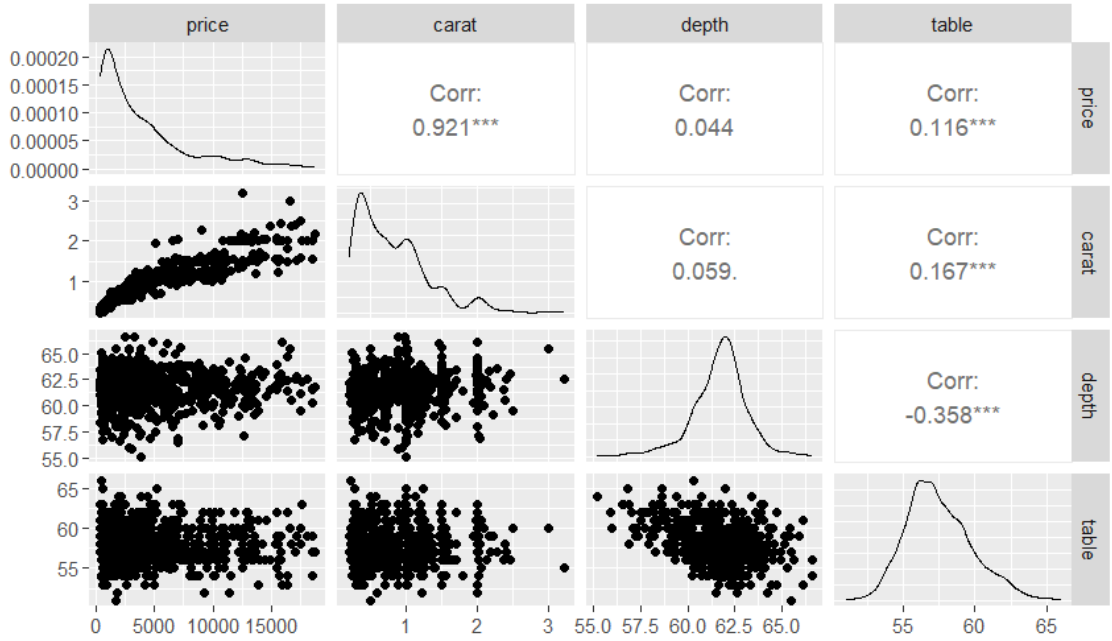

Figura 2 - Gráfico de Análise da Dispersão entre Price, Carat, Depth e Table 2

A partir desses comandos, conseguimos concluir, através da análise dos outputs:

- Existe uma relação positiva significativa muito alta entre *Price* e *Carat*. No entanto, a nuvem de pontos entre *Price* e *Carat* assume um comportamento não linear dada a curvatura que se verifica no gráfico;
- A relação entre *Price* e *Depth* é muito pequena e, pela nuvem de pontos, o impacto que *Depth* provoca no preço parece minúsculo ou inexistente;
- Existe uma relação positiva significativa baixa entre *Price* e *Table*, mas pela nuvem de pontos não parece que *Table* afete *Price*;
- A relação entre *Carat* e *Depth* é baixa e não significativa, não havendo relação entre as duas variáveis;
- *Table* apresenta uma relação positiva significativa baixa com *Carat*, implicando que diamantes de maior peso apresentam mesas um pouco maiores, e uma relação negativa significativa moderada com *Depth*, implicando que em diamantes de maior profundidade a mesa costuma ser menor.

E, para observar a relação entre *Price* e *Cut*, por ser *Cut* uma variável qualitativa, fizemos um *Boxplot* entre as duas variáveis.

Figura 3 - Boxplot entre as variáveis Price e Cut

Através desse, podemos concluir que a mediana tende a aumentar de *Fair* até *Premium*, e que os cortes *Very Good*, *Premium* e *Ideal* apresentam os melhores preços, apesar de terem muitos *outliers* associados. Algo de significativo a mencionar é que, apesar de *Ideal* ser a melhor categoria, não apresenta a maior mediana.

Questão 4

Apesar de já termos visto as correlações anteriormente, pelo *ggpairs()*, é-nos pedida na quarta alínea para calcular a matriz de correlações. Para essa, usamos *cor()* da biblioteca *dplyr* para que calcule a matriz para todos os pares de variáveis numéricas (para a variável quantitativa usamos a ANOVA). Através dessa, podemos concluir:

- Existe correlação alta entre *Price* e *Carat*, com r = 0.92085249;
- Existe correlação baixa entre *Price* e *Depth*, com r = 0.04365188;
- Existe correlação baixa entre *Price* e *Table*, com r = 0.1163144;
- Existe correlação baixa entre *Carat* e *Depth*, com r = 0.05886091;
- Existe correlação baixa entre *Carat* e *Table*, com r = 0.1669732;
- Existe correlação moderada entre *Depth* e *Table*, com r = -0.3575062;

Usando a ANOVA para obter a correlação entre a nossa variável resposta e a variável qualitativa, concluímos que *Cut* tem um efeito estatisticamente significativo sobre Price, com o r = 0.000162.

Questão 5

De seguida, é-nos pedido para ajustar os vários modelos de regressão linear simples, avaliando cada um quanto à sua significância. Para ajustarmos o modelo, usamos o *lm()*, colocando primeiro a variável resposta (*Price*) seguida da variável explicativa, separada da resposta por "~". Devemos, também, indicar de onde vêm os dados, para que o R saiba que amostras usar para o modelo.

Para o modelo *Price ~ Carat*, se *Carat* = 0 o valor do preço será negativo. Como isso não faz sentido, podemos concluir que, se o peso for negativo, o preço será nulo (ou seja, não haverá diamante). Por cada unidade de peso a mais, o preço do diamante aumentará, em média, 7539.38 unidades no preço. Como o valor p do teste é inferior a 0.0001, podemos admitir o teste como significativo e, pelo R^2, concluímos que este modelo explica 84.8% da variabilidade do preço.

$$\widehat{Price} = -2135.56 + 7539.38 \cdot Carat$$

Para o modelo *Price ~ Depth*, podemos dizer que se o *Depth* for 0 o preço também será 0 (seguindo a mesma lógica de antes), e por cada unidade no *Depth* o preço aumentará 118.49 unidades. Com um valor p = 0.525, o teste não é significativo, e pelo R^2 este modelo explica apenas 0.1905% da variabilidade do preço.

$$\widehat{Price} = -3364.91 + 118.49 \cdot Depth$$

Para *Price ~ Table*, se *Table* for 0 o preço também o será e, por cada unidade em *Table*, o preço aumentará 200.75 unidades. Pelo valor p = 0.015120 < 0.05, o teste é significativo e, através do R^2, o modelo explica 1.353% da variabilidade do preço.

$$\widehat{Price} = -7598.01 + 200.75 \cdot Table$$

Por fim, para *Price ~ Cut*, independentemente do tipo de corte o preço do diamante deverá rondar 4067.6 unidades, e existe uma tendência cúbica do preço em 893.9 unidades à medida que o corte melhora. Pelo valor p inferior a 0.0001, o teste é significativo, e de acordo com o R^2 ele explica 2.23% da variabilidade do preço.

Questão 6

O primeiro modelo foi aquele que acabou por nos oferecer o maior coeficiente de determinação (R^2 = 0.848), o que nos permite concluir que o peso do diamante é o que mais explica a variabilidade do preço.

Questão 7

Se verificarmos as nuvens de pontos anteriormente representadas podemos notar curvatura entre *Price* e *Carat*. Para resolver essa, devemos tentar transformar as variáveis do modelo de modo a conseguir a melhor relação não linear entre as duas variáveis, comparando cada transformação através do R^2 obtido.

Depois de testarmos com o logaritmo e polinómios de grau 2 e 3, o melhor modelo de regressão não linear que obtemos é entre o logaritmo de ambas as variáveis (i.e., *log(Price) ~ log(Carat)*), sendo o seu R^2 = 0.9345.

Modelos de Regressão Linear Múltipla

Questão 8

Terminados os exercícios de regressão linear simples, foi-nos pedido o uso de algoritmos sequenciais variados para selecionar as variáveis a se usar para o nosso problema de regressão linear múltipla. Afinal, queremos o mínimo de variáveis que expliquem o máximo de variabilidade da nossa variável resposta, para impedir que tenhamos modelos Hiper complexos ou, por outro lado, variáveis que digam pouco sobre o problema.

Para este problema, usamos três métodos aprendidos em aula: O *forward step*, o *backward step* e o *stepwise*. No final, compararíamos os modelos para verificar se eram iguais por cada um dos três métodos e, se não, qual deveríamos escolher. Para o primeiro método, iniciamos o nosso 'modelo 0' como um modelo constante, isso é apenas com a nossa variável resposta. Depois, corremos o método através do comando *step()* com escopo para todas as outras variáveis que poderiam vir a ser adicionadas. O modelo final resultou em um modelo de regressão linear múltipla com as variáveis *Carat* e *Cut* a explicar a nossa variável resposta, algo que, pelo R^2, permite explicar 85.12% da variabilidade do preço.

O segundo método requeria que iniciássemos o modelo completo (i.e., com todas as variáveis), pois seria a partir desse que seriam removidas as variáveis menos significativas no modelo. O modelo final deu-nos igual àquele obtido pelo primeiro método.

Por fim, usando o método *stepwise*, começando com o modelo vazio, obtemos também o mesmo modelo com *Price* de variável resposta a ser explicada pela variável quantitativa *Carat* e qualitativa *Cut*.

Questão 9

Por ter sido o modelo final seguindo por cada um dos três métodos, decidimos selecioná-lo como o modelo mais adequado.

Questão 10

Analisando melhor o nosso modelo podemos concluir, fora o que já concluímos quanto à percentagem de variabilidade do preço explicada por esse, que qualquer diamante com peso e corte nulo terá um preço de 0 (o que faz sentido, pois peso nulo implicaria um diamante inexistente), e que por cada unidade no peso o preço do diamante aumentará em 7598.87 unidades. Existe uma tendência linear para que o preço do diamante aumente 664.46 unidades à medida que o corte melhora, também, visto que a variável "L" da nossa variável qualitativa é significativa. A partir do valor p da nossa estatística F, o qual é inferior a 0.0001, podemos concluir que o modelo é globalmente significativo.

$$\widehat{Price} = -2371.88 + 7598.87 \cdot Carat + 664.46 \cdot Cut$$

Questão 11

Para validarmos os pressupostos do modelo devemos recorrer à análise gráfica dos quatro gráficos que surgem quando fazemos plot() com o modelo como variável. Vale-se notar que, fora aqueles testados pelo gráfico, existe também o pressuposto da Independência que, neste caso, considera-se já verificado por não haver dependência temporal das amostras.

Os quatro gráficos permitem-nos estudar o modelo quanto à homocedasticidade, normalidade, linearidade e presença de *outliers*.

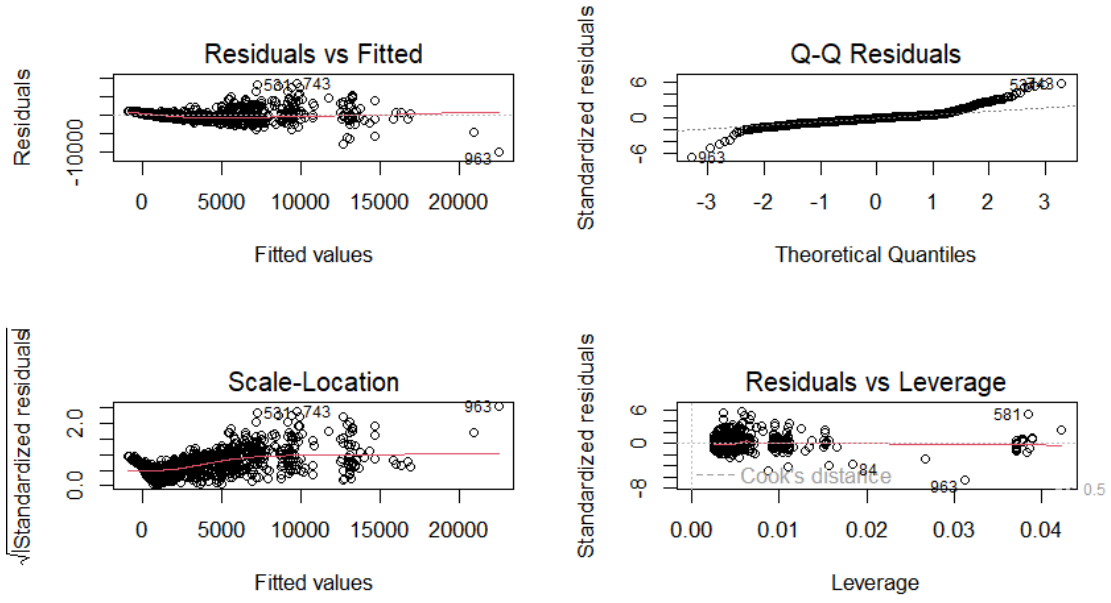

Figura 4 - Análise Gráfica dos Pressupostos do Modelo

Através do gráfico *Residuals vs Fitted* podemos estudar o modelo quanto à linearidade. Para que haja linearidade é esperado que as amostras sigam um padrão retilíneo, sem nenhum padrão de movimento para cima, para baixo ou de afunilamento. Contudo, como podemos ver no gráfico, não só as amostras parecem vir descendo como vão abrindo, mostrando óbvio desafunilamento.

O *Q-Q Residuals* permite-nos estudar a normalidade do gráfico, sendo esperado que os pontos sigam um movimento aproximadamente linear, junto à linha de confiança. Contudo, as caudas do gráfico apresentam curvaturas óbvias que a afastam o suficiente do centro para que não possamos assumir a normalidade de imediato, levantando suspeitas.

O *Scale-Location* é o gráfico que nos permite estudar a Homocedasticidade. Para que esse pressuposto seja validado, os dados devem se manter aleatórios, não havendo padrão de abertura ou de fecho, algo que infelizmente se verifica com a abertura dos pontos à medida que o valor ajustado aumenta.

Por fim, o *Residuals vs Leverage* permite-nos estudar o modelo quanto à presença de *outliers*. Se um ponto passar da distância de *Cook*, que será mostrada como retas em tracejado no gráfico, podemos admitir que esse ponto é um *outlier*. Este pressuposto pode ser validado, apesar de haver pontos afastados dos outros, pois não vemos isso acontecer com nenhum dos pontos.

Para garantir que o pressuposto da normalidade é, ou não, validado, podemos criar um histograma dos resíduos estandardizados que permitirá validar o pressuposto da normalidade caso o comportamento do histograma seja semelhante ao do gráfico da normal (uma parábola ascendente com topo no centro do histograma).

E, como verificamos no gráfico abaixo, não podemos validar a normalidade. Apesar de se notar crescimento até ao centro do histograma e

perca de altura daí adiante, a diferença de altura do centro comparado aos lados não permite que validemos a normalidade.

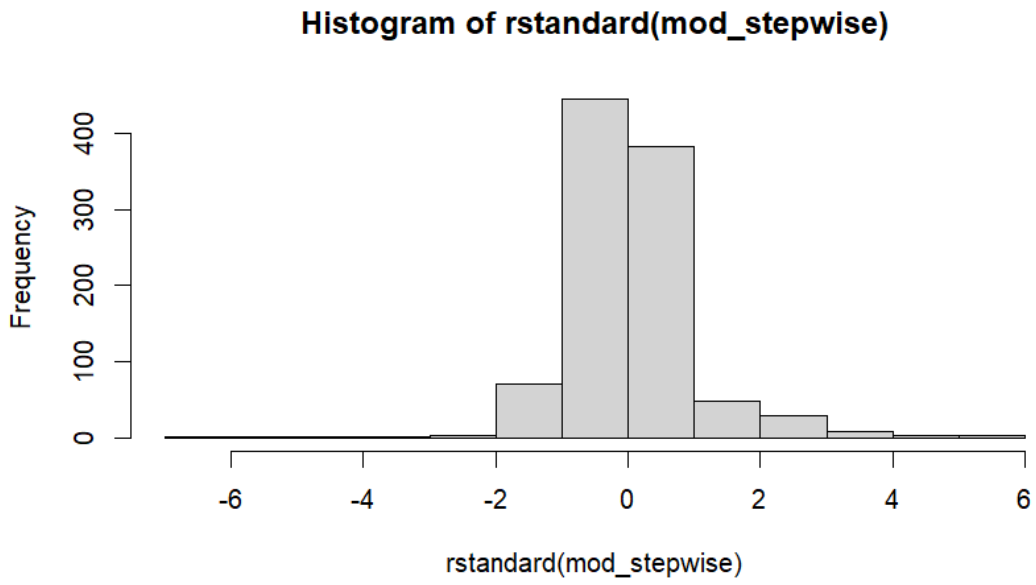

Figura 5 - Histograma dos Resíduos

Questão 12

Apesar de não termos notado *outliers* significativos pela falta de amostras fora da distância de *Cook*, no exercício anterior, devemos ainda avaliar a existência de valores atípicos. Apesar de poder ser esperado uma ou outra amostra atípica, várias amostras atípicas ou uma amostra que seja tão diferente do resto que provoque grandes alterações no modelo poderão levantar alarmes. Fazendo um plot do quarto parâmetro do nosso modelo, podemos notar que existem alguns valores muito diferentes aos demais, nomeadamente a amostra 963 e 581. Todavia, não são tão altas quanto à distância de *Cook*, chegando apenas ao 0.20. Portanto, não existem pontos influentes. Se quisermos ver quantos valores atípicos temos, no entanto (quantos pontos temos que sejam superiores a 4 / *length*, podemos calcular as distâncias de *cook* do modelo e, a partir dessas, verificar quais são os pontos cuja distância é maior. Podemos, assim, concluir que os pontos que notamos no gráfico anterior são pontos atípicos, no entanto não influentes o suficiente para levarem à violação de pressupostos.

Questão 13

Como o nosso modelo não satisfaz os pressupostos necessários, é pedido no exercício 13 que tentemos encontrar um modelo alternativo que os satisfaça. Podemos tentar transformar variáveis ou adicionar interações para tal.

Primeiramente, fizemos os gráficos dos resíduos das duas variáveis no nosso modelo para descobrir se deveríamos transformar alguma delas. Se o gráfico mostrar algum padrão, é sinal de que devemos transformar essa variável.

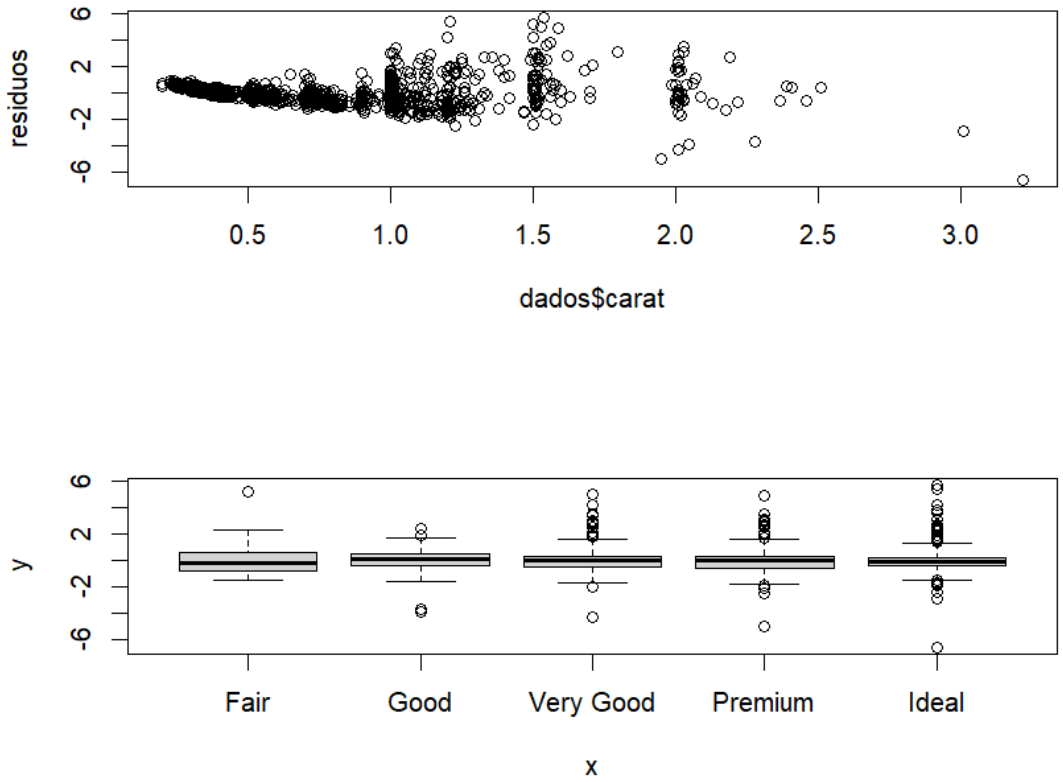

Figura 6 - Análise dos Padrões das Amostras

Como podemos verificar pela análise dos gráficos, as amostras do gráfico referente à variável *Carat* estão a fazer um padrão de abertura. Portanto, poderemos conseguir resultados se transformarmos essa variável.

Tentámos várias transformações possíveis (raíz quadrada, logaritmo, polinómio de grau 2 e inverter a variável), mas nenhuma acabou por permitir a validação de todos os pressupostos. Portanto, para não tentarmos procurar

por agulhas, pensámos em transformar a variável resposta e voltar a tentar as transformações típicas de *Carat* para uma variável resposta transformada.

A primeira transformação que tentámos foi tornar a variável resposta no logaritmo da mesma (pensando em como, na questão 6, foi esse o resultado que proporcionou o melhor erro), e voltar a tentar tanto sem transformações nas outras variáveis quanto com as transformações já antes tentadas. Acabámos por descobrir, eventualmente, um modelo que permite validar os pressupostos: log(*Price*) ~ log(*Carat*) + *Cut*.

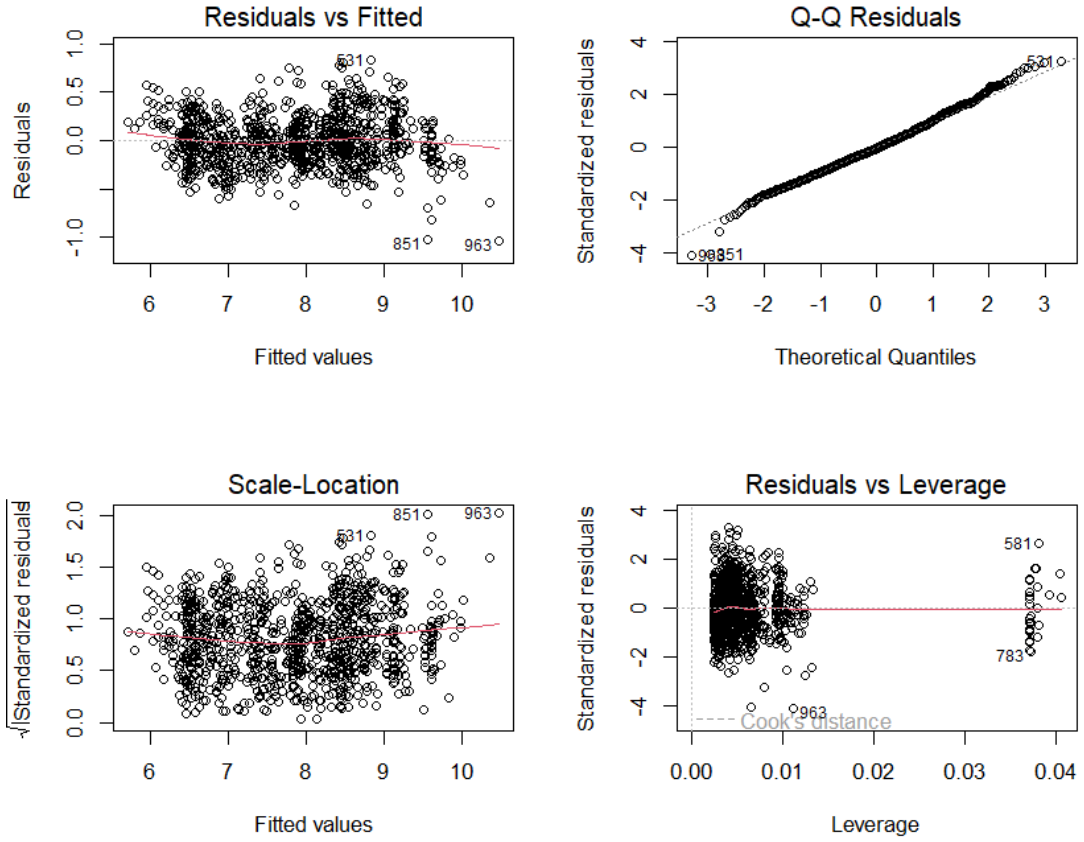

Figura 7 - Verificação dos Pressupostos

Podemos ver que cada um dos 4 pressupostos está verificado: As amostras não seguem nenhum padrão específico no primeiro gráfico e, no segundo, seguem a linha reta, permitindo assumir normalidade. A aleatoriedade ausente de padrões no terceiro gráfico permite validar a homocedasticidade, e a falta de pontos que estejam fora da distância de Cook permite garantir a falta de outliers. Usando o histograma para garantir a normalidade, pode-se também verificar que esse segue a curvatura esperada da distribuição normal.

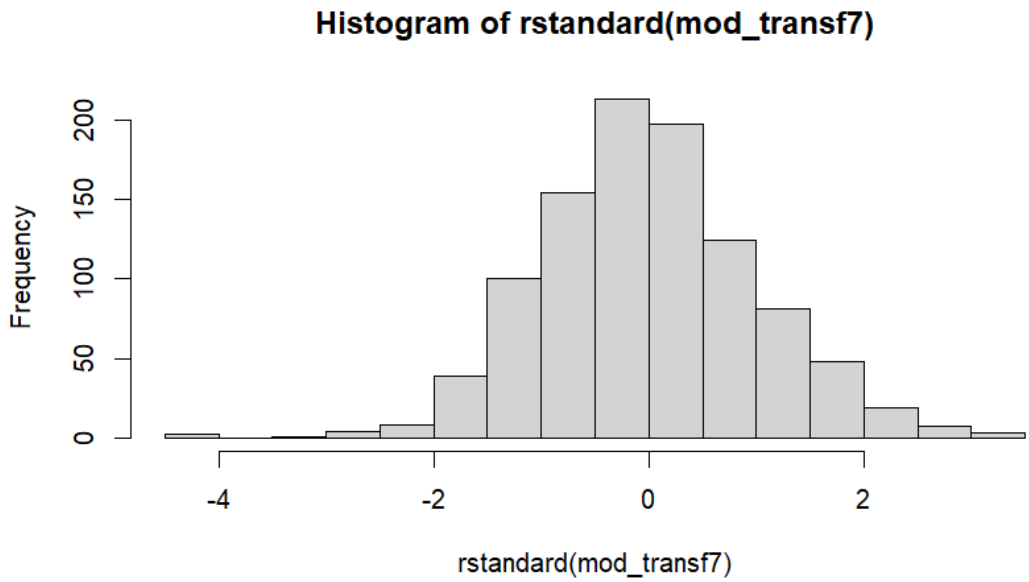

Figura 8 - Histograma dos Resíduos do Modelo Final

Portanto, ao descobrirmos um modelo que permitisse validar os pressupostos, testámos, como pedido pela alínea b da questão, adicionar a interação entre *Carat* e *Cut* para verificar se ela permite construir um modelo melhor. Contudo, a interação acaba por não mudar muito o modelo. Apesar dos pressupostos se manterem, parecem mais próximos de ser violados.

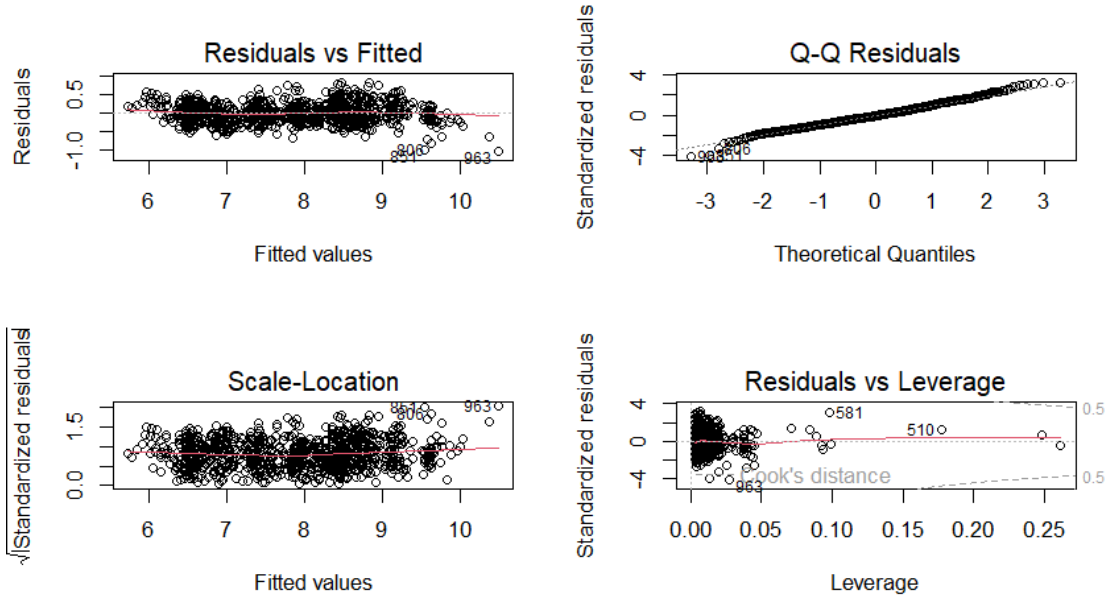

Figura 9 - Verificação dos Pressupostos 2

Como podemos verificar, o padrão de descida do primeiro gráfico é mais visível (não o suficiente para rejeitar a normalidade, todavia), e existem várias amostras no quarto gráfico mais próximas da distância de Cook. Pela interação acabar por não mudar muito no nosso modelo e apenas parecer agravar os pressupostos, optámos por retirá-la do modelo final.

Questão 14

Com tudo feito, o nosso modelo final acabou por *ficar log(Price) ~ log(Carat) + Cut*, não havendo interação. O modelo permite explicar 93.7% da variabilidade do preço e, pelo valor p da estatística F, é globalmente significativo. Através dos intervalos de confiança a 95%, podemos concluir que, com 95% de confiança, o preço em logaritmo de um diamante qualquer varia entre 8.387 e 8.439 e que, por unidade de peso em logaritmo, o preço é influenciado positivamente entre 1.659 e 1.715 unidades. Como o corte linear é o mais significativo entre os quatro, mas o corte de grau 4 também é um pouco significativo, podemos concluir que o preço aumenta de modo aproximadamente linear com base no aumento do corte, no entanto essa linha apresenta uma pequena curvatura para cima. Verificando os intervalos de 95% de confiança, a variação linear de que falamos pode variar entre os 0.045 e 0.177 e a variação de grau 4 (a curvatura) pode variar entre os 0.008 e os 0.085, sempre positivos, ou seja o preço sempre aumentando com base no corte.

Interpretando o modelo final, podemos dizer, por ser um modelo de logaritmo com logaritmo, que um aumento de 1% no peso leva ao aumento do preço do diamante em aproximadamente 1.69%. Para diferentes níveis de corte o preço do diamante parece aumentar linearmente em 11.7% (exp(0.111) ~ 1.117), havendo pequena curvatura em quanto o aumento do preço é com base no corte.

$$\log(\widehat{Price}) = 8.41299 + 1.68705 \cdot \log(Carat) + 0.11113 \cdot Cut$$