

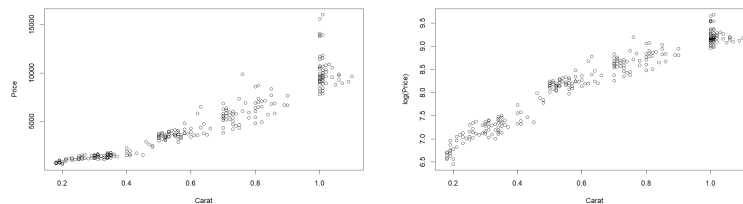
Intelligent Data Analysis - Homework 2.1

Panagiotis Michalopoulos, Javier de la Rua, Michail Gongolidis,
Ignacio Rodriguez, Daniel Minguez

December 20th, 2018

1 Plot price vs. caratage and $\log(\text{price})$ vs. caratage. Decide on which response variable is better to use.

For linear regression models, we want to make sure that there is a linear relationship between the input and output variables. Taking the log to the price makes the relationship between price and carat looks more linear. This is our main objective for linear regression, As we can see in the below graphs, plotting the log price gives a better representation of the variables.



2 Find a suitable way to include, besides caratage, the other categorical information available.

We have choose the worst level for each categorical variable: VS2 (very slightly imperfect 2) for Clarity, I for ColourPurity and HRD for certification institution as reference categories.

As we can see in the summary below, the overall model has a Multiple R-squared of 0.9723, so the model is able to predict any Y using X with accuracy of 97%.

```

Call:
lm(formula = log(Price) ~ Weight + ColourPurity + Clarity + Certifier,
    data = diamonds)

Residuals:
    Min       1Q   Median       3Q      Max
-0.31236 -0.11520  0.01613  0.10833  0.36339

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.077239    0.048091 126.369 < 2e-16 ***
Weight        2.855013    0.036968  77.230 < 2e-16 ***
ColourPurityD  0.416557    0.041382  10.066 < 2e-16 ***
ColourPurityE  0.387047    0.030824  12.557 < 2e-16 ***
ColourPurityF  0.310198    0.027479  11.288 < 2e-16 ***
ColourPurityG  0.210207    0.028359   7.412 1.32e-12 ***
ColourPurityH  0.128681    0.028523   4.511 9.31e-06 ***
ClarityIF      0.298541    0.033303   8.964 < 2e-16 ***
ClarityVS1     0.096609    0.024919   3.877 0.00013 ***
ClarityVVS1    0.297835    0.028102  10.598 < 2e-16 ***
ClarityVVS2    0.201923    0.025344   7.967 3.56e-14 ***
CertifierGIA   0.008856    0.020864   0.424 0.67155
CertifierIGI  -0.173855    0.028673  -6.063 4.07e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1382 on 295 degrees of freedom
Multiple R-squared:  0.9723,    Adjusted R-squared:  0.9712
F-statistic: 863.6 on 12 and 295 DF,  p-value: < 2.2e-16

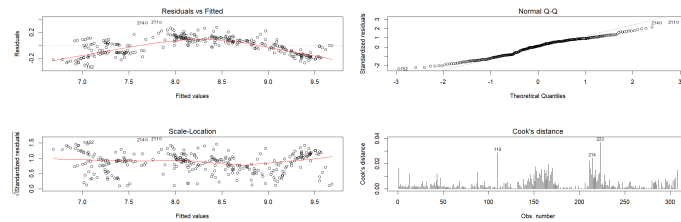
```

The test of significance for the overall model is $2.2e - 16$. We can reject the null hypothesis and say that not all coefficients are significantly equal to 0, even though CertifierGIA has low p-values and it is. We can conclude that the variables are significant for the model. We can say the following regarding the variables:

- **For the weight coefficient**, since the dependent variable is expressed in log, then we can say either:
 - In log scale, If you increase one unit of cartage , $\log(\text{price})$ will increase by 2.855
 - In Price scale, If you increase one unit of cartage the price of diamond is multiplied by a factor of $e^{2.88} = 17.37466$
- **For the color purity coefficient**, since the dependent variable is expressed in log and reference category is “I”, then we can say that the price of “ColourPurityD” is $\exp(0.41)=1.53$ the price of “ColourPurityI” , we can conclude that “ColourPurityD” is better than the reference category. Also it can be appreciated that the influence on the price is growing the better the color purity is.
- **For the clarity coefficient** we have a similar situation, we can appreciate that the influence on the price is greater the better clarity the diamond has, although there is not much difference between the two best levels (IF and VS1).

- **For the certifiers coefficient** we can see that the HRD and GIA have significantly the same influence in the price and both of them influence a higher price than IGI.

In order to make some interpretation on the residuals, we started by plotting a graph using the plot function for the model.



From the graph we can see that:

1. The residuals don't behave nicely, there seems to be a nonlinear relationship between the outcome and predictor (seem to be following a quadratic relation not a linear one).
2. Residuals seem to come from a normal distribution, but with lighter tails.
3. In the scale location plot, we can see that the residuals seem to be spread somehow equally along with the range of the predictor.
4. In Cook's distance plot, we can see that there are 3 influential values, which are: 110, 214 and 223

We can test residuals dependency using "Durbin Watson test", the p-value for the test is p-value ($< 2.2e - 16$), so we can reject the null hypothesis which means that the residuals are dependent and have correlation (not a good interpretation, residuals should be independent).

```
Durbin-Watson test

data: modell
DW = 0.31422, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is not 0
```

Jarque-Bera test has been used to check for residuals normality and the p-value of the residuals is $p\text{-value} = 0.01775$, which means rejection of the null hypothesis, so residuals does not follow a normal distribution (not a good interpretation, residuals should be normal).

```
Jarque Bera Test

data:  modell$residuals
X-squared = 8.0626, df = 2, p-value = 0.01775
```

To check for variance equality, Breusch-Pagan test have been used, the p-value equals to $4.265e-06$, so we will reject the null hypothesis that means variances are constant (not a good interpretation, residuals should have constant variance).

```
studentized Breusch-Pagan test

data:  modell
BP = 47.223, df = 12, p-value = 4.265e-06
```

We can see that there are 6 possible values that can be considered as outliers (shown in the qqplot and cook's distance plot): 110,152,211,214 and 255.

	Weight	ColourPurity	Clarity	Certifier	Price
110	0.76		D IF	GIA	9885
152	0.18		F VVS1	IGI	823
211	0.50		G IF	IGI	3652
214	0.52		I IF	IGI	3095
223	0.71		D VS1	IGI	6160

After removing the above outliers from the model, there was no significant enhancement done and other outliers appeared, we can see that from the below

```
Call:
lm(formula = log(Price) ~ Weight + ColourPurity + Clarity + Certifier,
    data = diamondq2)

Residuals:
    Min       1Q   Median       3Q      Max
-0.30841 -0.11627  0.01916  0.10333  0.36629

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.102954    0.046837  130.301 < 2e-16 ***
Weight        2.831122    0.036090   78.446 < 2e-16 ***
ColourPurityD  0.374517    0.042199    8.875 < 2e-16 ***
ColourPurityE  0.377928    0.030029   12.586 < 2e-16 ***
ColourPurityF  0.308913    0.026624   11.603 < 2e-16 ***
ColourPurityG  0.200360    0.027561    7.270 3.36e-12 ***
ColourPurityH  0.125571    0.027601    4.550 7.91e-06 ***
ClarityIF      0.279955    0.033058    8.468 1.26e-15 ***
ClarityVS1     0.092857    0.024162    3.843 0.000149 ***
ClarityVVS1    0.302881    0.027281   11.102 < 2e-16 ***
ClarityVVS2    0.199648    0.024579    8.123 1.32e-14 ***
CertifierGIA   0.006504    0.020228    0.322 0.748044
CertifierIGI  -0.183841    0.028275   -6.502 3.46e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1337 on 290 degrees of freedom
Multiple R-squared:  0.9741,    Adjusted R-squared:  0.973
F-statistic: 908.7 on 12 and 290 DF,  p-value: < 2.2e-16
```

3 Try two different remedial actions:

3.1-3.5 Create a new categorical variable to segregate the stones according to caratage: let's say less than 0.5 carats small, 0.5 to less than 1 carat (medium) and 1 carat and over (large). Make small as the reference category. Add this new variable to the existing model as well as an interaction term between this new variable and caratage.

3.6 Include the square of carat as a new explanatory variable. It avoids the subjectivity of clusters denition.

3.1 Is this regression model satisfactory? Are the standard assumptions of linear regression validated? Are the numerical estimates sensible?

The regression model is satisfactory, even though some assumptions are not satisfactory, as the overall test of significance for the model is $2.2e-16$ and $R^2 = 0.9953$ (there is an obvious improvement in the model compared to model 1).

```
Call:
lm(formula = log(Price) ~ Weight + ColourPurity + Clarity + Certifier +
    Carat_Size + Carat_Size * Weight, data = diamonds3A)

Residuals:
    Min       1Q   Median       3Q      Max
-0.188358 -0.031815 -0.000249  0.043143  0.140535

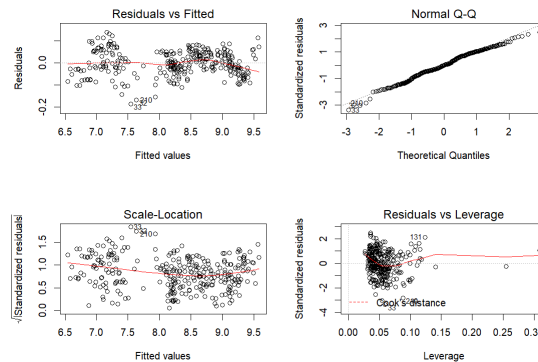
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.483149   0.029919  183.268 < 2e-16 ***
Weight         4.427061   0.069811   63.415 < 2e-16 ***
ColourPurityD   0.436261   0.017465   24.979 < 2e-16 ***
ColourPurityE   0.350912   0.012927   27.146 < 2e-16 ***
ColourPurityF   0.275010   0.011535   23.841 < 2e-16 ***
ColourPurityG   0.191449   0.011869   16.131 < 2e-16 ***
ColourPurityH   0.111067   0.011923    9.316 < 2e-16 ***
ClarityIF       0.315793   0.013935   22.662 < 2e-16 ***
ClarityVS1      0.067530   0.010498    6.432 5.15e-10 ***
ClarityVVS1     0.213448   0.011932   17.889 < 2e-16 ***
ClarityVVS2     0.132373   0.010756   12.307 < 2e-16 ***
CertifierGIA     0.005606   0.008794    0.637  0.524
CertifierIGI    -0.018082   0.012684   -1.426  0.155
Carat_SizeMedium 1.062001   0.032653   32.523 < 2e-16 ***
Carat_SizeLarge  2.340691   0.404861    5.781 1.91e-08 ***
Weight:Carat_SizeMedium -2.047162  0.074556  -27.458 < 2e-16 ***
Weight:Carat_SizeLarge -3.350469  0.399880  -8.379 2.31e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0576 on 291 degrees of freedom
Multiple R-squared:  0.9953,    Adjusted R-squared:  0.995
F-statistic: 3816 on 16 and 291 DF,  p-value: < 2.2e-16
```

Afterwards, we started plotting a graph using the plot function for the model, as in the below picture.

- **Residuals vs Fitted:** we can see that the residuals do follow somehow a linear relationship, but it does not seem to be constant for all fitted values (see between 7.5 and 8)
- **Normal Q-Q:** Residuals seems come from a normal distribution, again with differences in the tails.
- **In the scale location plot:** we can see that the residuals are divided into 2 groups along with the range of the predictor and there is a gap between those 2 groups.

- **Residuals vs Leverage:** We can see that there are some outliers like: 32, 33 and 210 that are seen far away from the concentrated data on the left



Regarding the assumptions for linear regression, we have applied the following tests:

- **Durbin Watson test:** The p-value $< 2.2e-16$, which means that the residuals have a correlation (not good interpretation for linear regression models)
- **Jarque-Bera:** The p-value = 0.172, which means accepting the null hypothesis, so residuals follow a normal distribution (good interpretation about the model)
- **Breusch-Pagan:** The p-value equals to 0.0007143, so we will reject the null hypothesis that means variance is not constant (again not good interpretation for linear regression models)

<p>Durbin-Watson test</p> <pre>data: model3a DW = 0.96989, p-value < 2.2e-16 alternative hypothesis: true autocorrelation is not 0</pre>	<p>Jarque Bera Test</p> <pre>data: model3a\$residuals X-squared = 3.517, df = 2, p-value = 0.1723</pre>
<p>studentized Breusch-Pagan test</p> <pre>data: model3a BP = 40.256, df = 16, p-value = 0.0007143</pre>	

3.2 Interpret the interaction parameter med*carat. What can we infer on the incremental pricing of caratage in the 3 clusters?

We can interpret that the interaction term for 'med*weight' is significant as p-value is less than 0.05 (5% level).

Since the dependent variable (price) is expressed in logs, then we can say that if the other variables are equal, the increase in price for a diamond of medium size is $\exp(-2.04) = 0.13$ times the increase in the price of a small diamond. This means that the price of diamond for small ones increases faster than medium ones, when you increase the carat by one.

3.3 Which is more highly valued: colour or clarity?

Colour is more highly valued since the best colour purity influence more in the price than the best clarity (and the same for the subsequent levels)

3.4 All other things being equal, what is the average price difference between a grade D diamond and another one graded (a) I (b) E?

The $\log(\text{price})$ is 0.47 higher for a D diamond than for a I diamond, and is 0.08 higher for a D diamond than for a E diamond

3.5 All other things being equal, are there price differences amongst the stones appraised by the GIA, IGI and HRD?

Given this model, all the coefficients related to the certifiers are significantly equal to 0, so we can assume that the prices are not influenced by them.

3.6 Include the square of carat as a new explanatory variable. It avoids the subjectivity of clusters denition

We have replace carat by the square of carat in order to avoid relation between variables, since if we use both carat and its square the model we obtain that there is a variable ilinearly related to the other variables. We obtain the following result:

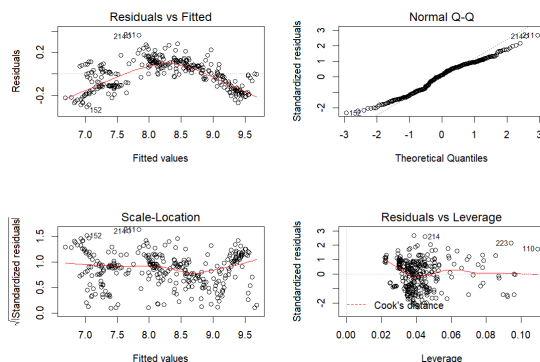
```
Call:
lm(formula = log(Price) ~ sqrt_Carat_Size + ColourPurity + Clarity +
    Certifier, data = diamonds3B)

Residuals:
    Min       1Q   Median       3Q      Max
-0.31236 -0.11520  0.01613  0.10833  0.36339

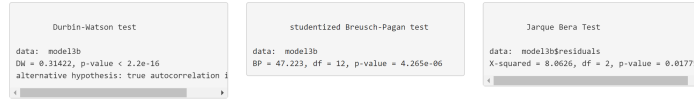
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.077239   0.048091 126.369 < 2e-16 ***
sqrt_Carat_Size 1.427506   0.018484  77.230 < 2e-16 ***
ColourPurityD   0.416557   0.041382  10.066 < 2e-16 ***
ColourPurityE   0.387047   0.030824  12.557 < 2e-16 ***
ColourPurityF   0.310198   0.027479  11.288 < 2e-16 ***
ColourPurityG   0.210207   0.028359   7.412 1.32e-12 ***
ColourPurityH   0.128681   0.028523   4.511 9.31e-06 ***
ClarityIF       0.298541   0.033303   8.964 < 2e-16 ***
ClarityVS1      0.096609   0.024919   3.877 0.00013 ***
ClarityVS1      0.297835   0.028102  10.598 < 2e-16 ***
ClarityVS2      0.201923   0.025344   7.967 3.56e-14 ***
CertifierGIA    0.008856   0.020864   0.424 0.67155
CertifierIGI    -0.173855   0.028673  -6.063 4.07e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1382 on 295 degrees of freedom
Multiple R-squared:  0.9723,    Adjusted R-squared:  0.9712
F-statistic: 863.6 on 12 and 295 DF,  p-value: < 2.2e-16
```

We attach below the plots:



The results of the test are the following:



We can see that we still have problems with the assumptions and that the R^2 in the model is not improved.

4 Which of the two remedial actions do you prefer and why? Think on terms of interpretability and validity of the assumptions.

In both models we have a good R^2 and a good p-value, regarding the assumptions:

- **Linearity of residuals** , both have the same problem.
- **Normality** , model 2 behaves better compared to model 3.
- **Variance equality**, both have the same problem.

In summary:

	Model 2	Model 3
R^2	0.9952	0.9712
Durbin-Watson test p-value (linearity)	2.2e-16	2.2e-16
Jarque Bera Test p-value	0.1723	0.01775
Breusch-Pagan test p-value	0.0007143	4.265e-06

Also the second model allow us to interpret the result based on carat size and the second not. It's more difficult to interpret the square of carat.

Therefore we conclude that the model 2 is better.