POLITÉCNICA

"Ingeniamos el futuro"

CAMPUS
DE EXCELENCIA
INTERNACIONAL

Master Universitario en Ciencia de Datos

Homework 2.1

Multiple Regression Analysis

Jiayi Lin

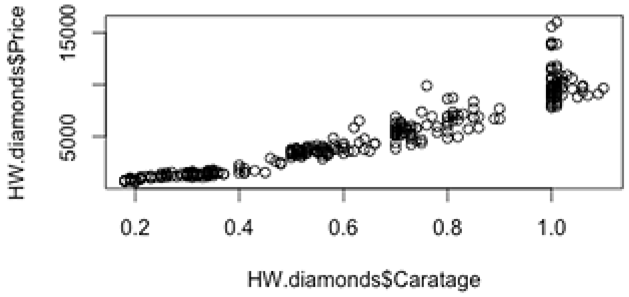
Xiao Luo

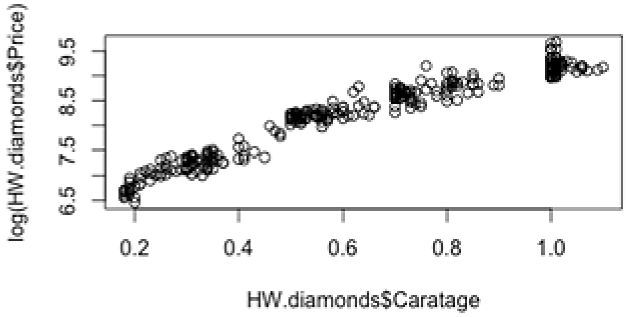
Nabil Aziz

El Abbassi Widad

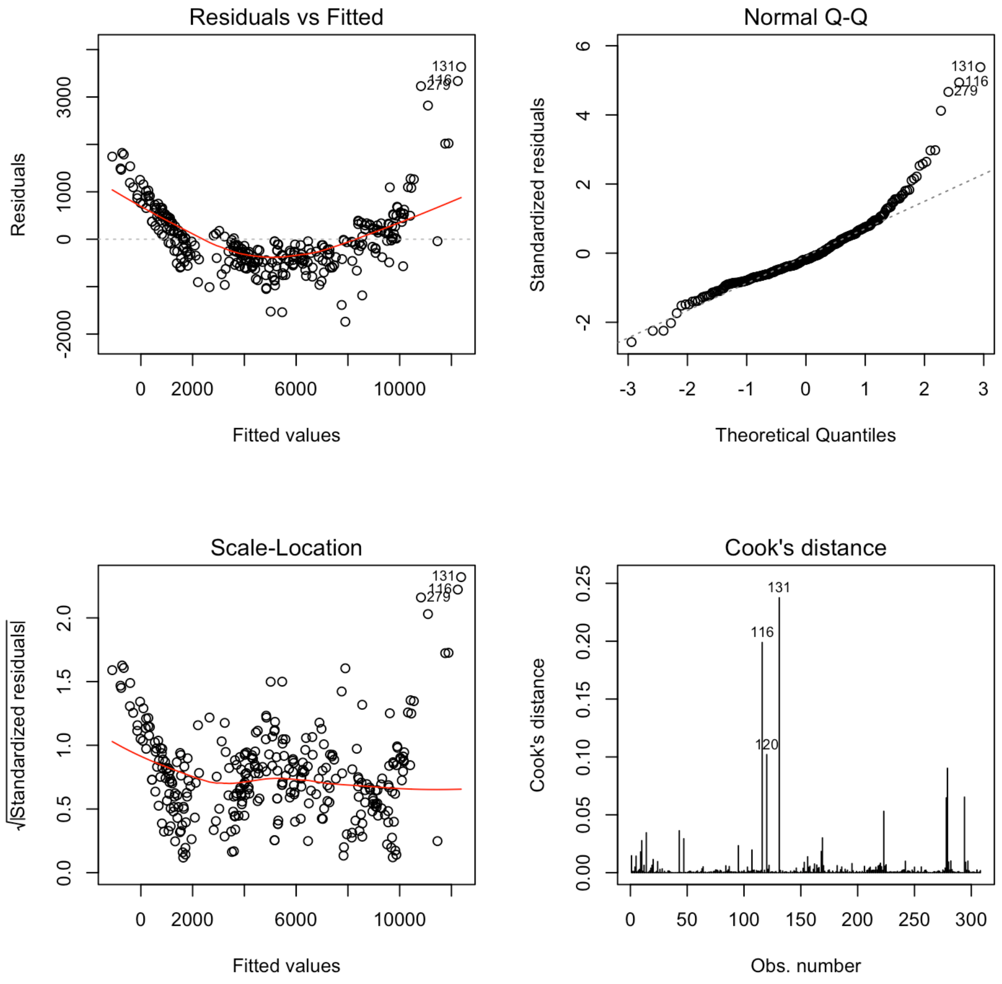
December 8th,2019

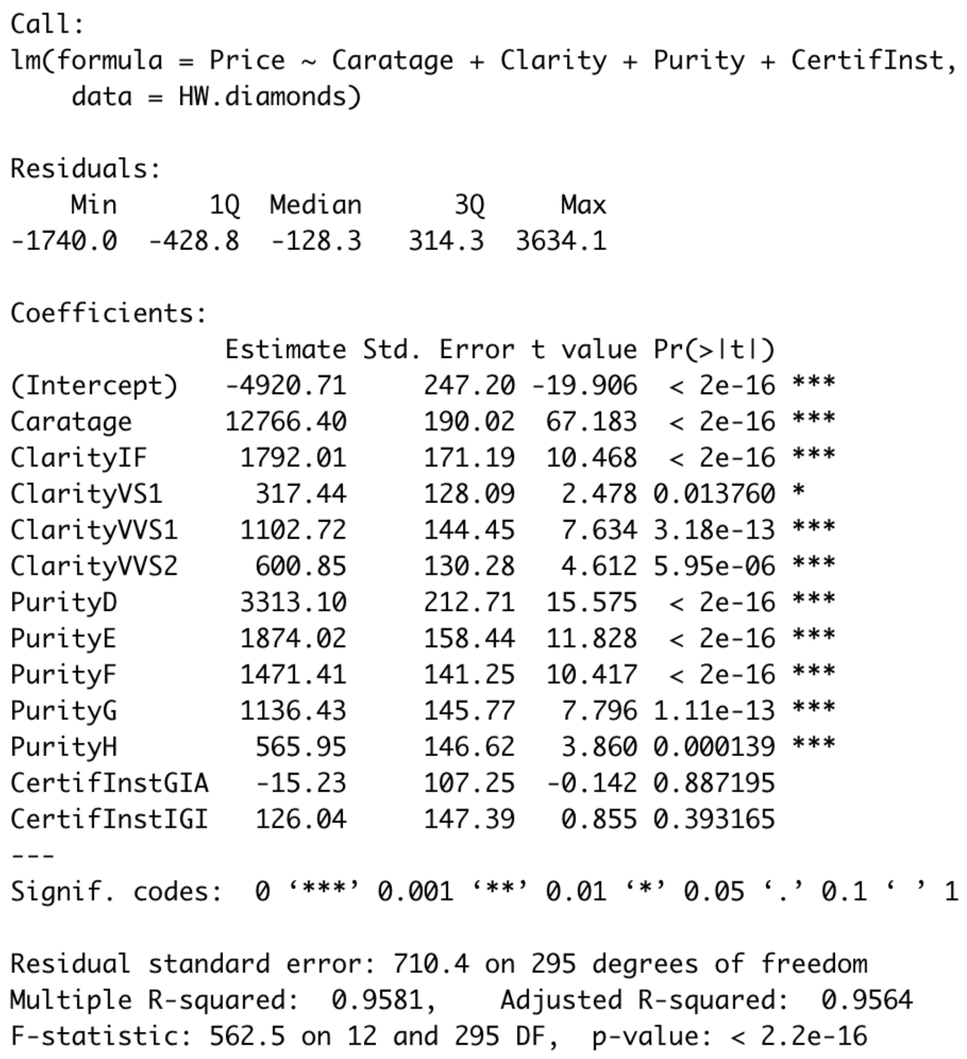
1. The plot Price vs Caratage is better because it generally shows a much higher linearity than the other, despite of the fact that at 1.0 carat the price varies much. The plot log(Price) vs Caratage behaves better at that point, but the overall model results in a curve which means much worse linearity.





2.



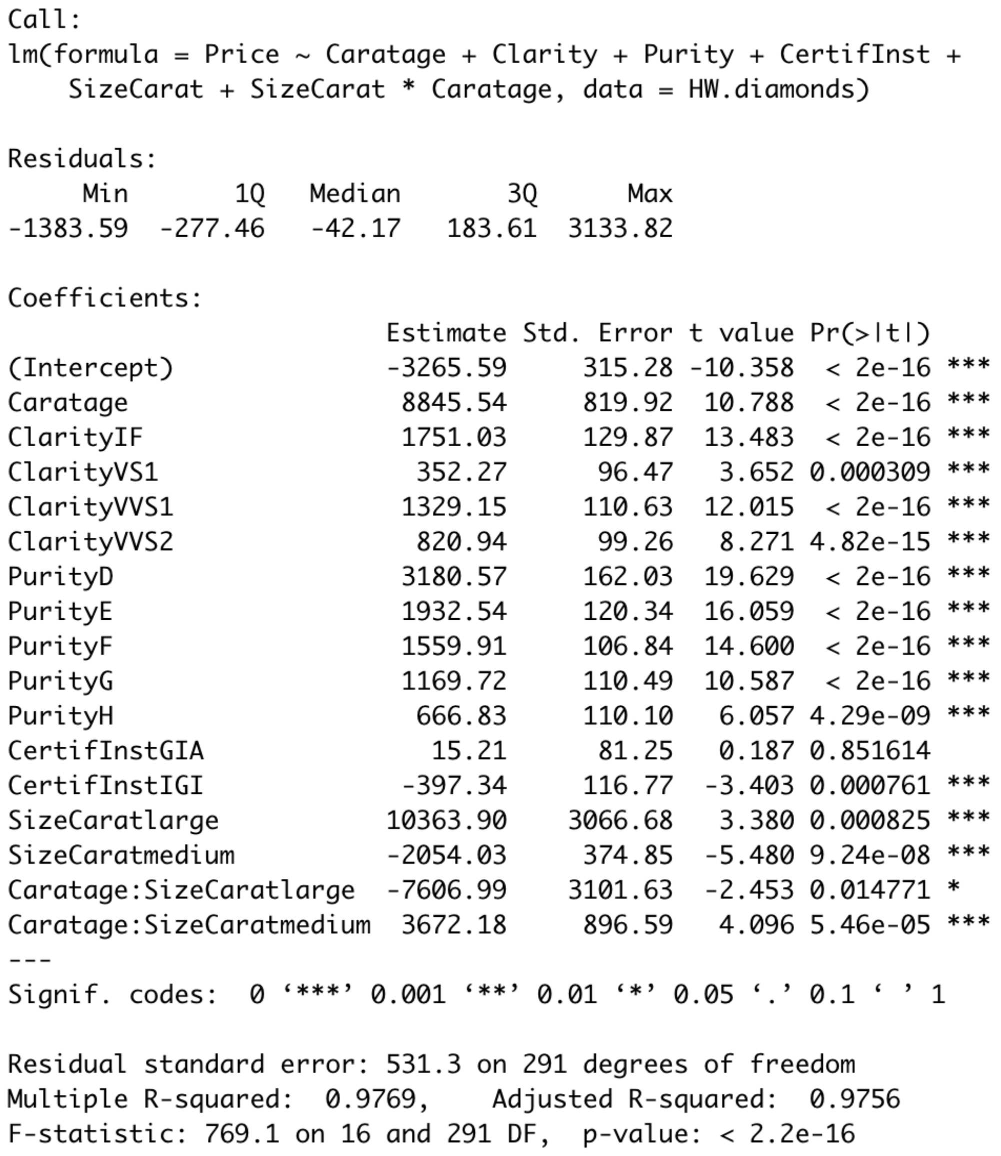


```
Call:
lm(formula = Price ~ Caratage + Clarity + Purity + CertifInst,
    data = HW.diamonds)

Residuals:
    Min      1Q  Median      3Q     Max 
-1740.0  -428.8  -128.3   314.3  3634.1 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -4920.71     247.20 -19.906  < 2e-16 ***
Caratage      12766.40     190.02  67.183  < 2e-16 ***
ClarityIF      1792.01     171.19  10.468  < 2e-16 ***
ClarityVS1      317.44     128.09   2.478 0.013760 *  
ClarityVVS1    1102.72     144.45   7.634 3.18e-13 ***
ClarityVVS2     600.85     130.28   4.612 5.95e-06 ***
PurityD        3313.10     212.71  15.575  < 2e-16 ***
PurityE        1874.02     158.44  11.828  < 2e-16 ***
PurityF        1471.41     141.25  10.417  < 2e-16 ***
PurityG        1136.43     145.77   7.796 1.11e-13 ***
PurityH         565.95     146.62   3.860 0.000139 ***
CertifInstGIA   -15.23     107.25  -0.142 0.887195    
CertifInstIGI   126.04     147.39   0.855 0.393165    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 710.4 on 295 degrees of freedom
Multiple R-squared:  0.9581,    Adjusted R-squared:  0.9564 
F-statistic: 562.5 on 12 and 295 DF,  p-value: < 2.2e-16
```

⇨ Residuals don’t behave nicely, they show strong dependence on predicted values. They are not normally distributed: from Normal Q-Q plot we can verify that at higher theoretical quantiles, residuals fall off far away from the line. By deleting some outliers the result would be much better in term of normality. The variance is non constant, at low predicted values it’s much higher than the average. There are also several clear outliers, observations **116, 120 and 131.**

3. a.

⇨ As the results shows below, the multiple R-squared is very high **0.9769**, so the regression model is satisfactory. Standard assumptions of linear regression are not valid: residuals are not independent (because of low price of medium size caratage), they are not normal distributed unless taking off many outliers, and the model shows heteroscedasticity.



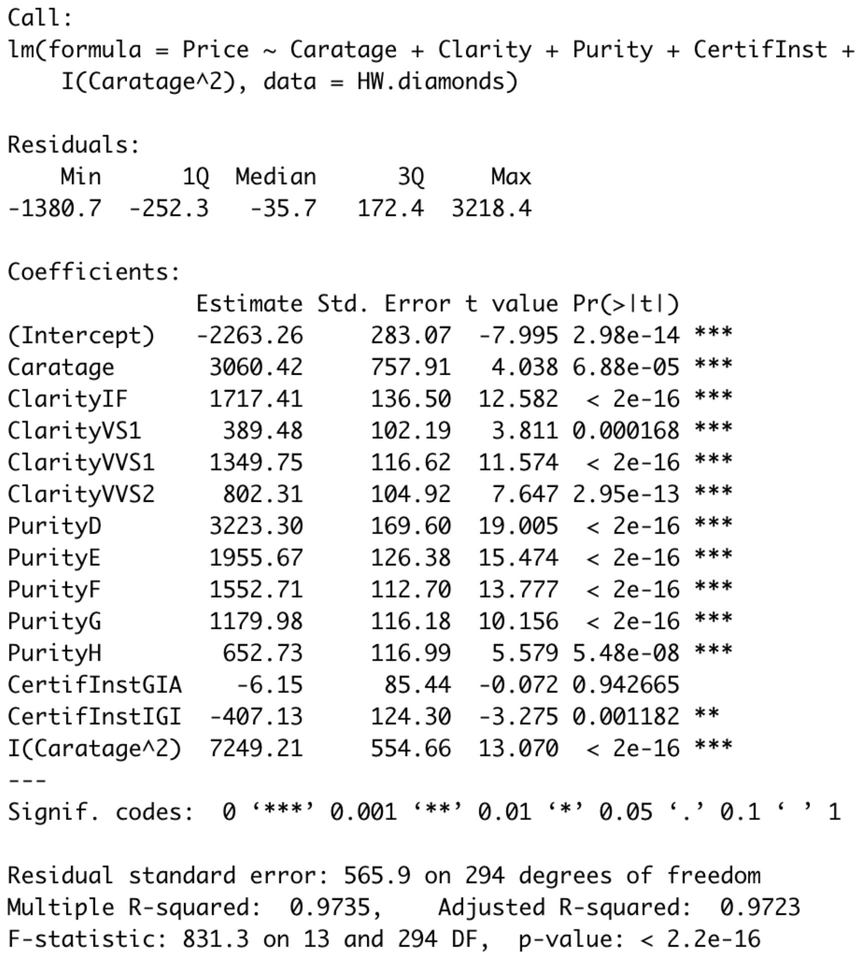
```
Call:
lm(formula = Price ~ Caratage + Clarity + Purity + CertifInst +
    SizeCarat + SizeCarat * Caratage, data = HW.diamonds)

Residuals:
     Min       1Q   Median       3Q      Max 
-1383.59  -277.46   -42.17   183.61  3133.82 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)              -3265.59     315.28 -10.358  < 2e-16 ***
Caratage                  8845.54     819.92  10.788  < 2e-16 ***
ClarityIF                 1751.03     129.87  13.483  < 2e-16 ***
ClarityVS1                 352.27      96.47   3.652 0.000309 ***
ClarityVVS1               1329.15     110.63  12.015  < 2e-16 ***
ClarityVVS2                820.94      99.26   8.271 4.82e-15 ***
PurityD                   3180.57     162.03  19.629  < 2e-16 ***
PurityE                   1932.54     120.34  16.059  < 2e-16 ***
PurityF                   1559.91     106.84  14.600  < 2e-16 ***
PurityG                   1169.72     110.49  10.587  < 2e-16 ***
PurityH                    666.83     110.10   6.057 4.29e-09 ***
CertifInstGIA               15.21      81.25   0.187 0.851614    
CertifInstIGI             -397.34     116.77  -3.403 0.000761 ***
SizeCaratlarge           10363.90    3066.68   3.380 0.000825 ***
SizeCaratmedium          -2054.03     374.85  -5.480 9.24e-08 ***
Caratage:SizeCaratlarge  -7606.99    3101.63  -2.453 0.014771 *  
Caratage:SizeCaratmedium  3672.18     896.59   4.096 5.46e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 531.3 on 291 degrees of freedom
Multiple R-squared:  0.9769,    Adjusted R-squared:  0.9756 
F-statistic: 769.1 on 16 and 291 DF,  p-value: < 2.2e-16
```

- Numerical estimates are all sensible except for GIA as certificate institution because of its high p-value.
  - The interaction parameter *med*carat* indicates that for those classified as medium size, the price increases 3672.18 Singapore dollars by each additional carat.
  - At the same classified size, the price variation, in descendent order, would be *Medium* > *Small* > *Large*.
  - Colour purity and clarity are both highly valued because of very low p-value.

All other things being equal, the average price of a grade **D** diamond is *3180.57 Singapore dollars* higher than a grade **I** one, and the same (grade **D**) is *1248.03 Singapore dollars* higher than a grade **E** one.

All other things being equal, there is no price difference between stones appraised by HRD and GIA because of its low value and high p-value, but 397.34 Singapore dollars less for stones appraised by IGI, with significant p-value.

3.b



```
Call:
lm(formula = Price ~ Caratage + Clarity + Purity + CertifInst +
    I(Caratage^2), data = HW.diamonds)

Residuals:
    Min      1Q  Median      3Q     Max 
-1380.7  -252.3   -35.7   172.4  3218.4 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -2263.26     283.07  -7.995 2.98e-14 ***
Caratage       3060.42     757.91   4.038 6.88e-05 ***
ClarityIF      1717.41     136.50  12.582  < 2e-16 ***
ClarityVS1      389.48     102.19   3.811 0.000168 ***
ClarityVVS1    1349.75     116.62  11.574  < 2e-16 ***
ClarityVVS2     802.31     104.92   7.647 2.95e-13 ***
PurityD        3223.30     169.60  19.005  < 2e-16 ***
PurityE        1955.67     126.38  15.474  < 2e-16 ***
PurityF        1552.71     112.70  13.777  < 2e-16 ***
PurityG        1179.98     116.18  10.156  < 2e-16 ***
PurityH         652.73     116.99   5.579 5.48e-08 ***
CertifInstGIA    -6.15      85.44  -0.072 0.942665    
CertifInstIGI  -407.13     124.30  -3.275 0.001182 ** 
I(Caratage^2)  7249.21     554.66  13.070  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 565.9 on 294 degrees of freedom
Multiple R-squared:  0.9735,    Adjusted R-squared:  0.9723 
F-statistic: 831.3 on 13 and 294 DF,  p-value: < 2.2e-16
```

4.

The first remedial action is preferable, not just because its multiple R-squared is higher meaning the model explains more variability than the second model, but it interprets the price with higher precision through statistically significant explanatory variables.