FIN41660 ASSIGNMENT

Autumn Trimester 2021

UCD Michael Smurfit Graduate Business School

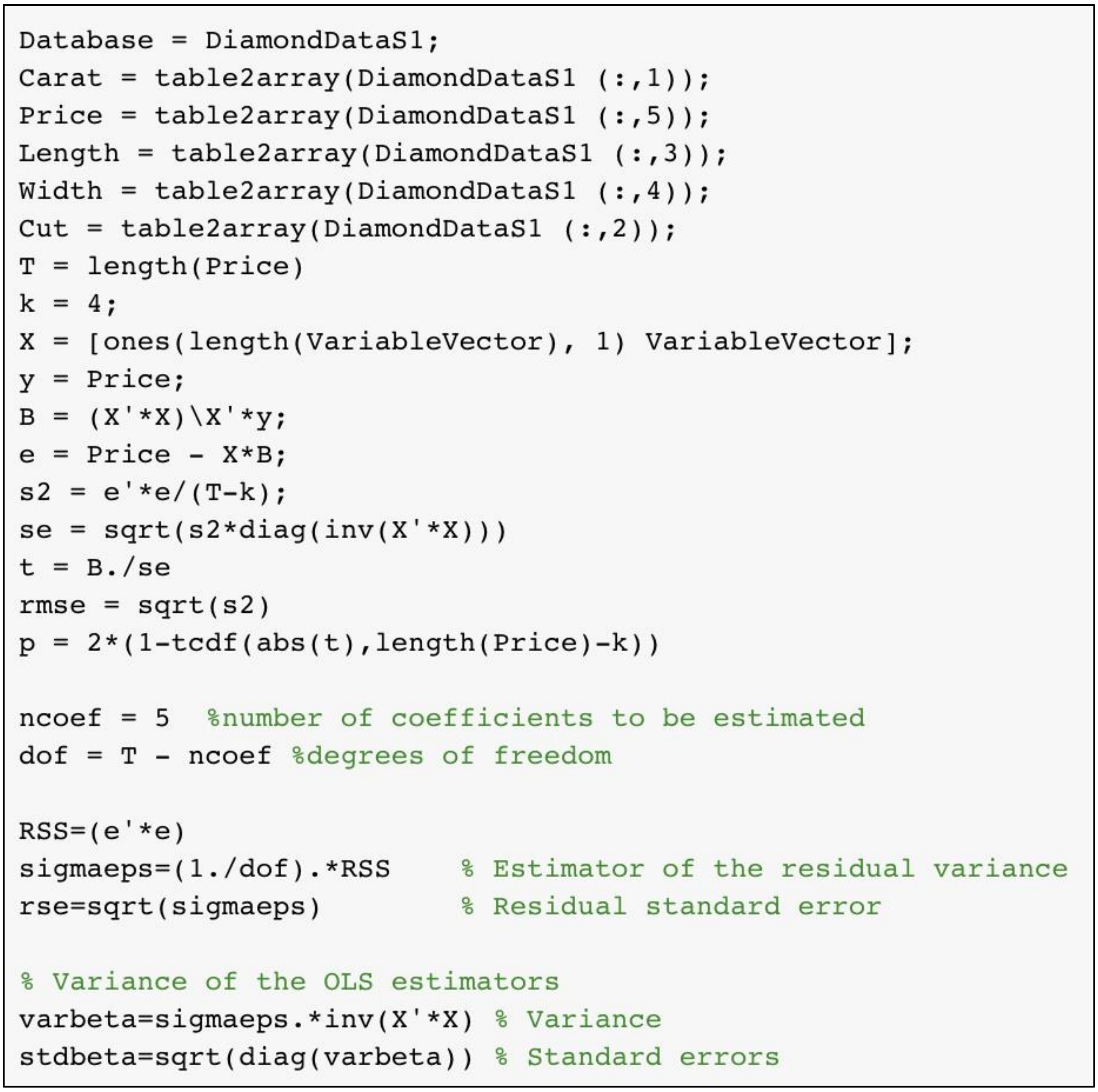
An Application of MATLAB-based Data Modelling and Regression Analysis

Joseph Collins – 21205397

Word Count: 3,349 (excl. references)

QUESTION 1 – MATLAB FUNCTIONS

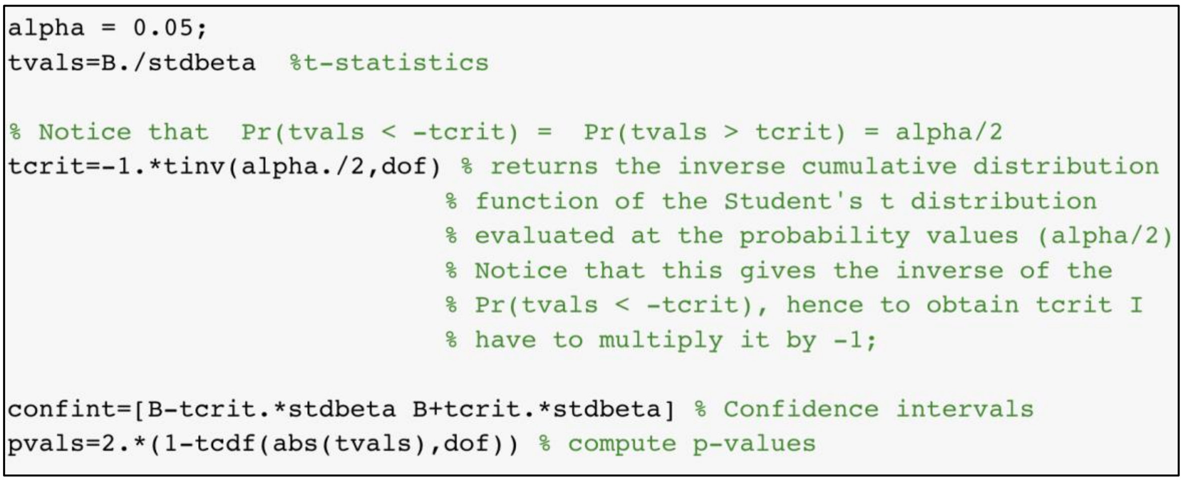
A. OLS REGRESSION



```
Database = DiamondDataS1;
Carat = table2array(DiamondDataS1 (:,1));
Price = table2array(DiamondDataS1 (:,5));
Length = table2array(DiamondDataS1 (:,3));
Width = table2array(DiamondDataS1 (:,4));
Cut = table2array(DiamondDataS1 (:,2));
T = length(Price)
k = 4;
X = [ones(length(VariableVector), 1) VariableVector];
y = Price;
B = (X'*X)\X'*y;
e = Price - X*B;
s2 = e'*e/(T-k);
se = sqrt(s2*diag(inv(X'*X)))
t = B./se
rmse = sqrt(s2)
p = 2*(1-tcdf(abs(t),length(Price)-k))

ncoef = 5  %number of coefficients to be estimated
dof = T - ncoef %degrees of freedom

RSS=(e'*e)
sigmaeps=(1./dof).*RSS     % Estimator of the residual variance
rse=sqrt(sigmaeps)         % Residual standard error

% Variance of the OLS estimators
varbeta=sigmaeps.*inv(X'*X) % Variance
stdbeta=sqrt(diag(varbeta)) % Standard errors
```

Code 1

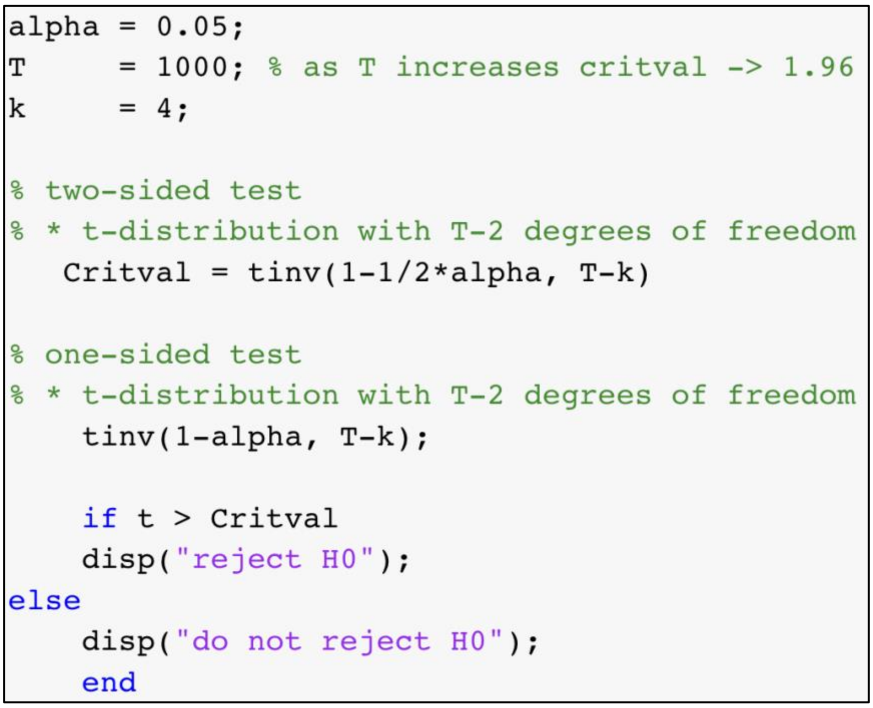
B. CONFIDENCE INTERVALS



```
alpha = 0.05;
tvals=B./stdbeta  %t-statistics

% Notice that  Pr(tvals < -tcrit) =  Pr(tvals > tcrit) = alpha/2
tcrit=-1.*tinv(alpha./2,dof) % returns the inverse cumulative distribution
                            % function of the Student's t distribution
                            % evaluated at the probability values (alpha/2)
                            % Notice that this gives the inverse of the
                            % Pr(tvals < -tcrit), hence to obtain tcrit I
                            % have to multiply it by -1;

confint=[B-tcrit.*stdbeta B+tcrit.*stdbeta] % Confidence intervals
pvals=2.*(1-tcdf(abs(tvals),dof)) % compute p-values
```

Code 2

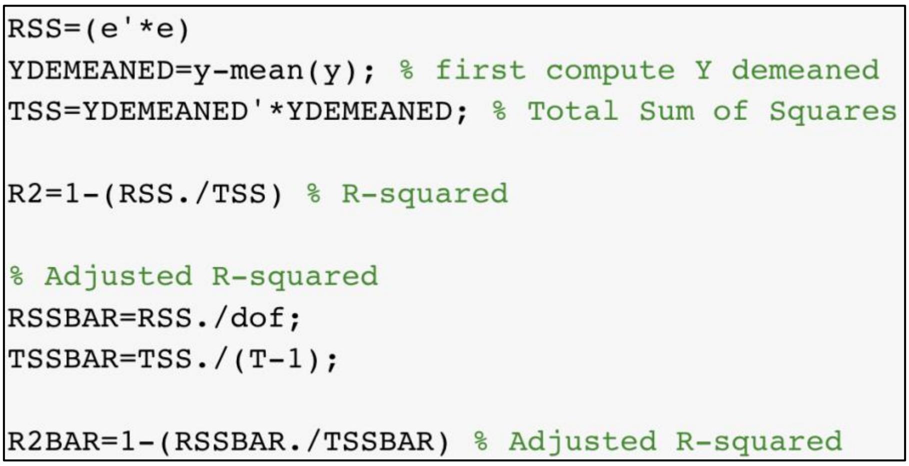
C. STATISTICAL SIGNIFICANCE



```
alpha = 0.05;
T     = 1000; % as T increases critval -> 1.96
k     = 4;

% two-sided test
% * t-distribution with T-2 degrees of freedom
   Critval = tinv(1-1/2*alpha, T-k)

% one-sided test
% * t-distribution with T-2 degrees of freedom
    tinv(1-alpha, T-k);

    if t > Critval
    disp("reject H0");
else
    disp("do not reject H0");
    end
```

Code 3

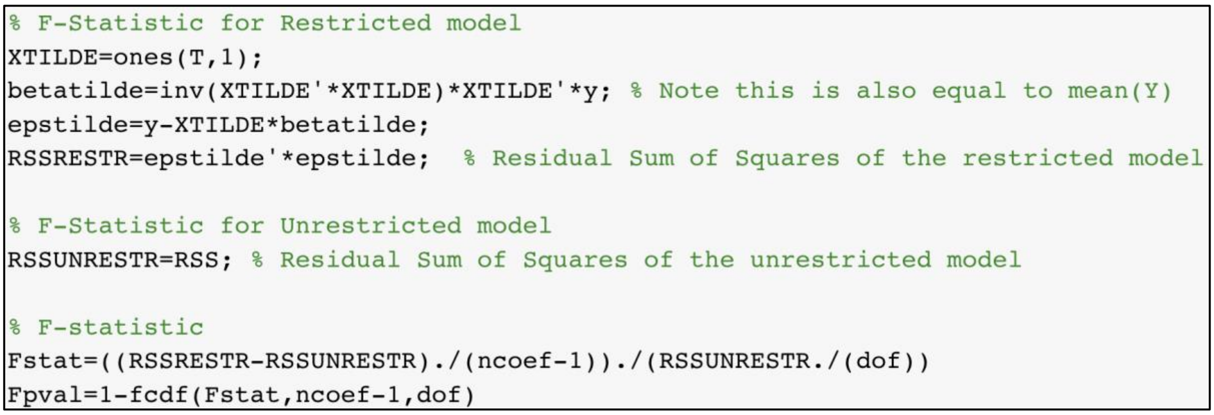
D. RSQUARED AND ADJUSTED RSQUARED



```
RSS=(e'*e)
YDEMEANED=y-mean(y); % first compute Y demeaned
TSS=YDEMEANED'*YDEMEANED; % Total Sum of Squares

R2=1-(RSS./TSS) % R-squared

% Adjusted R-squared
RSSBAR=RSS./dof;
TSSBAR=TSS./(T-1);

R2BAR=1-(RSSBAR./TSSBAR) % Adjusted R-squared
```

Code 4

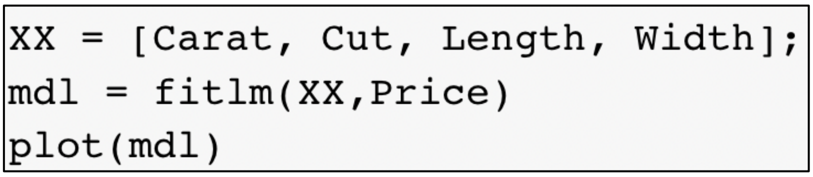
E. F-STATISTIC



```
% F-Statistic for Restricted model
XTILDE=ones(T,1);
betatilde=inv(XTILDE'*XTILDE)*XTILDE'*y; % Note this is also equal to mean(Y)
epstilde=y-XTILDE*betatilde;
RSSRESTR=epstilde'*epstilde;  % Residual Sum of Squares of the restricted model

% F-Statistic for Unrestricted model
RSSUNRESTR=RSS; % Residual Sum of Squares of the unrestricted model

% F-statistic
Fstat=((RSSRESTR-RSSUNRESTR)./(ncoef-1))./(RSSUNRESTR./(dof))
Fpval=1-fcdf(Fstat,ncoef-1,dof)
```

Code 5

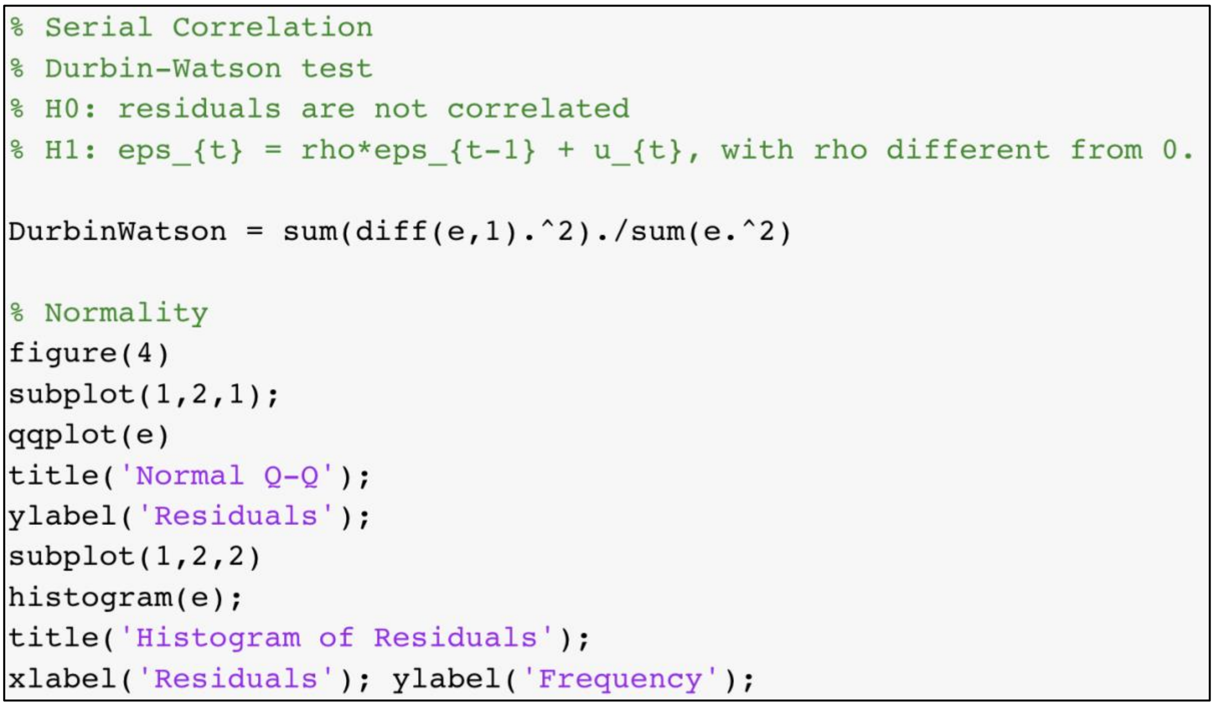
F. PLOT OF THE FITTED MODEL



```
XX = [Carat, Cut, Length, Width];
mdl = fitlm(XX,Price)
plot(mdl)
```

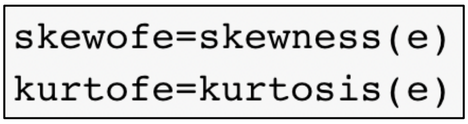
Code 6

G. DIAGNOSTIC TESTS FOR RESIDUALS (NORMALITY, HETEROSCEDASTICITY, SERIAL CORRELATION)



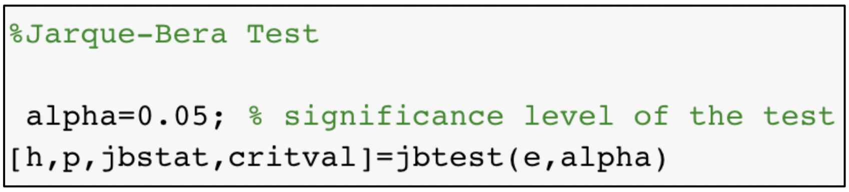
```
% Serial Correlation
% Durbin-Watson test
% H0: residuals are not correlated
% H1: eps_{t} = rho*eps_{t-1} + u_{t}, with rho different from 0.

DurbinWatson = sum(diff(e,1).^2)./sum(e.^2)

% Normality
figure(4)
subplot(1,2,1);
qqplot(e)
title('Normal Q-Q');
ylabel('Residuals');
subplot(1,2,2)
histogram(e);
title('Histogram of Residuals');
xlabel('Residuals'); ylabel('Frequency');
```

Code 7



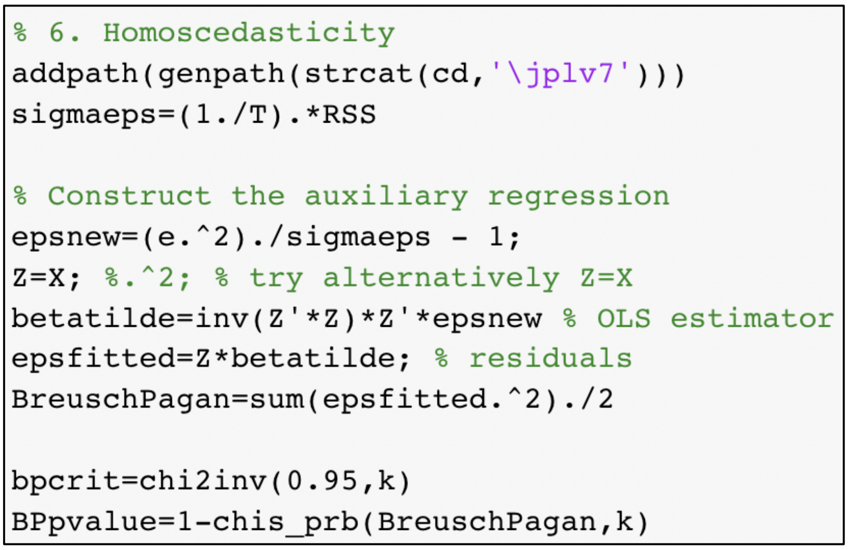
```
skewofe=skewness(e)
kurtofe=kurtosis(e)
```

Code 8



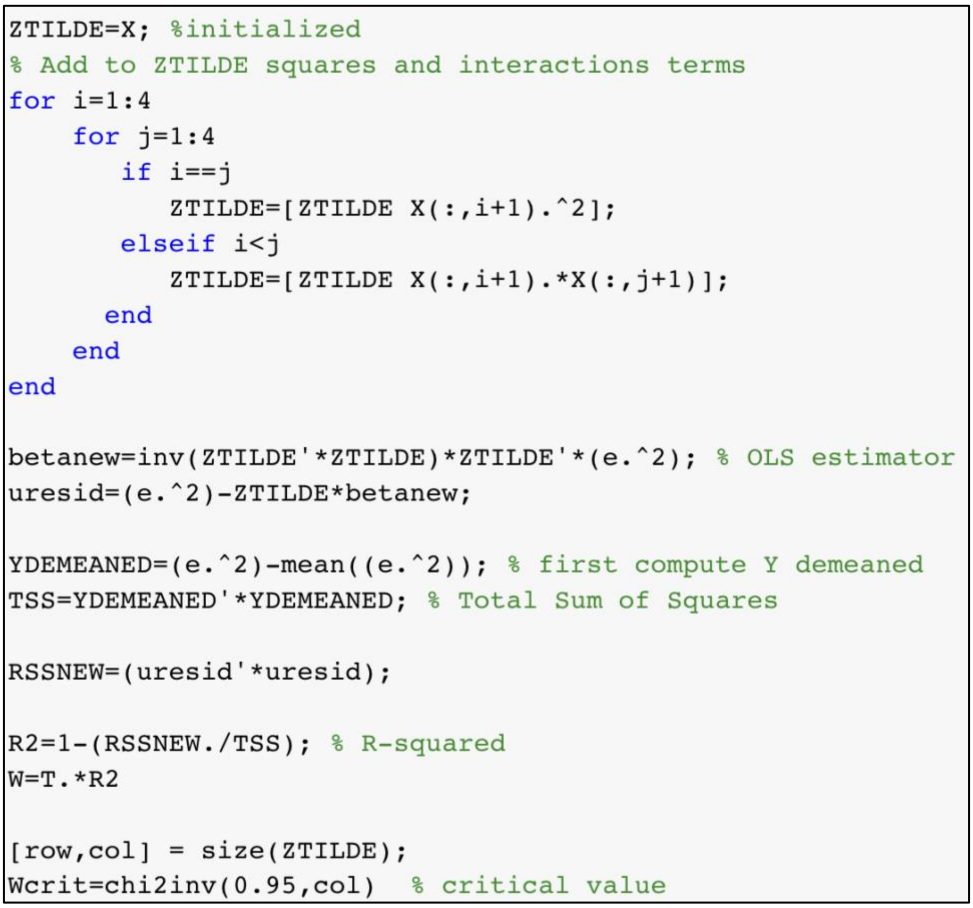
```
%Jarque-Bera Test

 alpha=0.05; % significance level of the test
[h,p,jbstat,critval]=jbtest(e,alpha)
```

Code 9



```
% 6. Homoscedasticity
addpath(genpath(strcat(cd,'\jplv7')))
sigmaeps=(1./T).*RSS

% Construct the auxiliary regression
epsnew=(e.^2)./sigmaeps - 1;
Z=X; %.^2; % try alternatively Z=X
betatilde=inv(Z'*Z)*Z'*epsnew % OLS estimator
epsfitted=Z*betatilde; % residuals
BreuschPagan=sum(epsfitted.^2)./2

bpcrit=chi2inv(0.95,k)
BPpvalue=1-chis_prb(BreuschPagan,k)
```

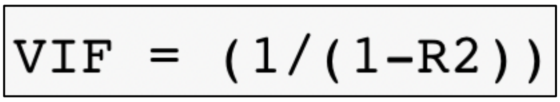
Code 10



```
ZTILDE=X; %initialized
% Add to ZTILDE squares and interactions terms
for i=1:4
    for j=1:4
       if i==j
          ZTILDE=[ZTILDE X(:,i+1).^2];
       elseif i<j
          ZTILDE=[ZTILDE X(:,i+1).*X(:,j+1)];
      end
    end
end

betanew=inv(ZTILDE'*ZTILDE)*ZTILDE'*(e.^2); % OLS estimator
uresid=(e.^2)-ZTILDE*betanew;

YDEMEANED=(e.^2)-mean((e.^2)); % first compute Y demeaned
TSS=YDEMEANED'*YDEMEANED; % Total Sum of Squares

RSSNEW=(uresid'*uresid);

R2=1-(RSSNEW./TSS); % R-squared
W=T.*R2

[row,col] = size(ZTILDE);
Wcrit=chi2inv(0.95,col)  % critical value
```

Code 11

H. MULTICOLLINEARITY

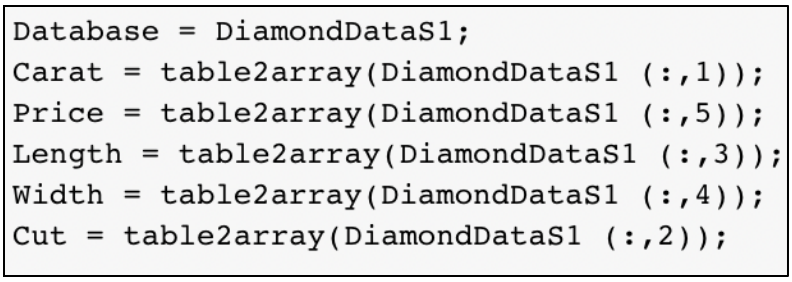


```
VIF = (1/(1-R2))
```

Code 12

QUESTION 2 – BRIEF INTRODUCTION

This regression model is based on a dataset which is concerned with the price of diamonds in relation to various quality factors of the diamond, such as: its quality of cut, its carat level, and its length and width. These factors are acting as independent variables in this regression model.



```
Database = DiamondDataS1;
Carat = table2array(DiamondDataS1 (:,1));
Price = table2array(DiamondDataS1 (:,5));
Length = table2array(DiamondDataS1 (:,3));
Width = table2array(DiamondDataS1 (:,4));
Cut = table2array(DiamondDataS1 (:,2));
```

OLS ESTIMATOR

$$Formula: y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + \beta_5 x_{5t} + \varepsilon_t$$

As per Burton (2020), the Ordinary Least Squares (OLS) regression creates a line of best fit that serves as the most accurate way of illustrating the spread of the datapoints. In its essence, the OLS creates the best possible sampling distribution of unbiased estimates in comparison to other methods of linear estimation. However, while the OLS method of estimation may be BLUE, one should keep in mind that there are several pitfalls associated with this technique that must be considered, such as the necessity of independent datasets, and the potential sensitivity of the OLS to outliers. Running *Code 1*, one can estimate the OLS model to be:

$$\boldsymbol{y_t = -824.21 + (9616.6)(x_{2t}) + (336.22)(x_{3t}) + (-714.51)(x_{4t}) + (-25.44)(x_{5t}) + \varepsilon_t}$$

	Estimate	Standard Error	t-Stat	p-value
Intercept	-824.21	666	-1.2376	0.21617
Carat	9616.6	412.34	23.322	2.55E-96
Cut	336.22	42.669	7.8796	8.58E-15
Length	-714.51	168.92	-4.2298	2.55E-05
Width	-25.44	30.278	-0.84021	0.40099

Table 1

Interpreting the above table, one observes that ‘Carat’, ‘Cut’ and ‘Length’ are all jointly statistically significant at any level, as their p-values are of negligible value. However, the variable ‘Width’ is not statistically significant at any of the accepted standard values (as per general convention), as its p-value is 0.40099. Inferring the marginal effects of each of the coefficients:

- For a one unit increase in Carat, there is a 9616.6 unit increase in price
- For a one unit increase in Cut, there is a 336.22 unit increase in price
- For a one unit increase in Length, there is a -714.51 unit decrease in price
- For a one unit increase in Width, there is a -25.44 unit decrease in price

Finally, the intercept is statistically insignificant at any level, as it maintains a p-value of 0.21617. From an economical perspective, this paper holds that the estimated marginal effects of this regression are intuitively appealing, particularly that of the ‘Carat’ variable, as one would expect such a large increase in the price of a diamond for a substantial increase in its quality.

CONFIDENCE INTERVALS

$$Formula: \overline{x} \pm z\frac{s}{\sqrt{n}}$$

In statistical analysis, the confidence intervals of a model represent the probability that a parameter of a population will fall between a set of values for an *x* proportion of times. In this paper’s regression, the conventional confidence level of 95% was implemented to estimate this probability, as illustrated by the alpha value of 0.05 in *Code 2*. One can now infer that for 95% of the time the value of Carat will lie between 0.88 and 1.0426. The confidence intervals of all model coefficients are estimated as follows:

	Lower Bound*	Upper Bound*
Intercept	-0.2131	0.0483
Carat	0.8807	1.0426
Cut	0.0252	0.042
Length	-0.1046	-0.0383
Width	-0.0085	0.0034

Table 2 (*1.0e+04)

As a final note, one should consider that if zero lies within the confidence interval, there is an additional implication that the coefficient is no longer statistically significant at the given critical value. Thus, the findings of this body of code are in tandem with the results of the p-values in the OLS regression; zero is present in the confidence interval values of the ‘Intercept’ and ‘Width’, affirming their statistical insignificance.

STATISTICAL SIGNIFICANCE

T-Statistic

$$H0\text{: } the\ difference\ in\ group\ means\ is\ zero$$

$$H1\text{: } the\ difference\ in\ group\ means\ is\ different\ from\ zero$$

Aside from the R^2, Adjusted R^2 and F-Statistic, which are discussed in detail in later sections, one can test the statistical significance of one's model by estimating the T-Statistic. Alongside this T-Statistic, one estimates a T-critical value – if the value of the T-Statistic is smaller than the critical value, then the null hypothesis is rejected, and the coefficient is deemed insignificant. The T-Statistics of the regression model are plotted in the following table:

T-Critical Value	Intercept	Carat	Cut	Length	Width
1.9624	-1.2382	23.3335	7.8836	-4.232	-0.8406

Thus, according to this statistic, one should reject the null hypothesis that the Intercept, Width and Length are statistically significant.

R^2 AND ADJUSTED R^2

The R^2

$$Formula\text{: } 1 - \frac{RSS}{TSS}$$

The R^2, also known as the coefficient of determination, represents the sum of squared deviations about the regression (Sykes 1993). In its essence, R^2 is a measure of the percentage of the dependent variable's variation which is explained by the regression model. In this paper's analysis, the R^2 was calculated to be 0.8464 – indicating that there is significant predictive power to this model. However, as this regression model is one of a multivariate nature, potential issues could present themselves with the implementation of this statistic. For example, the R^2 faces difficulty when assessing the effectiveness of additional variables, as when a variable is added to a model, the R^2 value increases regardless of the efficiency of the additional term. Furthermore, if there are too many variables present, the R^2 begins to model the data's random noise – it begins ‘overfitting the model’. Unfortunately, this results in an inflated R^2 value, which is accompanied by a diminished ability to make accurate predictions.

The Adjusted R^2

$$Formula: 1-\frac{(1-R^2)(N-1)}{N-p-1}$$

The adjusted R^2 is an ameliorated version of the R^2 – one which has been adjusted for the quantity of variables in the model. In contrast to the R^2, the adjusted R^2 increases in value solely if the additional variable improves the fit of the model. Thus, the adjusted R^2 provides an unbiased estimate of the population R^2. The adjusted R^2 of this model was estimated to be 0.8458, which further supports the ‘goodness’ of the model’s predictive ability.

	Value
R^2	0.8464
Adjusted R^2	0.8458

Table 3

As the results of this paper’s adjusted R^2 are very close to the R^2 value, one can infer that all the independent variable augment the explanatory power of the model. As a final note, while the R^2 and the Adjusted R^2 estimate of the strength of the relationship between the dependent variable and the model, it does not offer a form hypothesis test– this is instead provided by the F-statistic.

F-STATISTIC

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_1: \beta_1 \neq \beta_2 \neq \beta_3 \neq \beta_4 \neq 0$$

$$Formula: \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-(k+1))'}$$

One can implement the F-Statistic in order to test the statistical significance of models; while comparable to the T-Statistic, it can instead be used to test the joint significance of a group of variables, as opposed to testing each on an individual basis. The F-Statistic is the ratio of the two Chi-squared distributions of both the unrestricted and restricted models [SOURCE]. In a similar manner to previously mentioned statistics, the capability of the F-Statistic is assessed using a complimentary p-value, as documented in the table below. Furthermore, one could use the F-value and F-critical value can be used in conjunction with the p-value – if the former outweighs the latter, then one could reject the null hypothesis. Running *Code* 5 in this paper's model, the F-statistic is estimated to be 1.3712e^{03} and the corresponding p-value is 0, as shown in *Table 4*. As this p-value is statistically significant at any conventional level (1%, 5% or 10%), one rejects the H_0, concluding that the coefficients of this model are statistically significant.

F-statistic	P-Value
1371.2	0

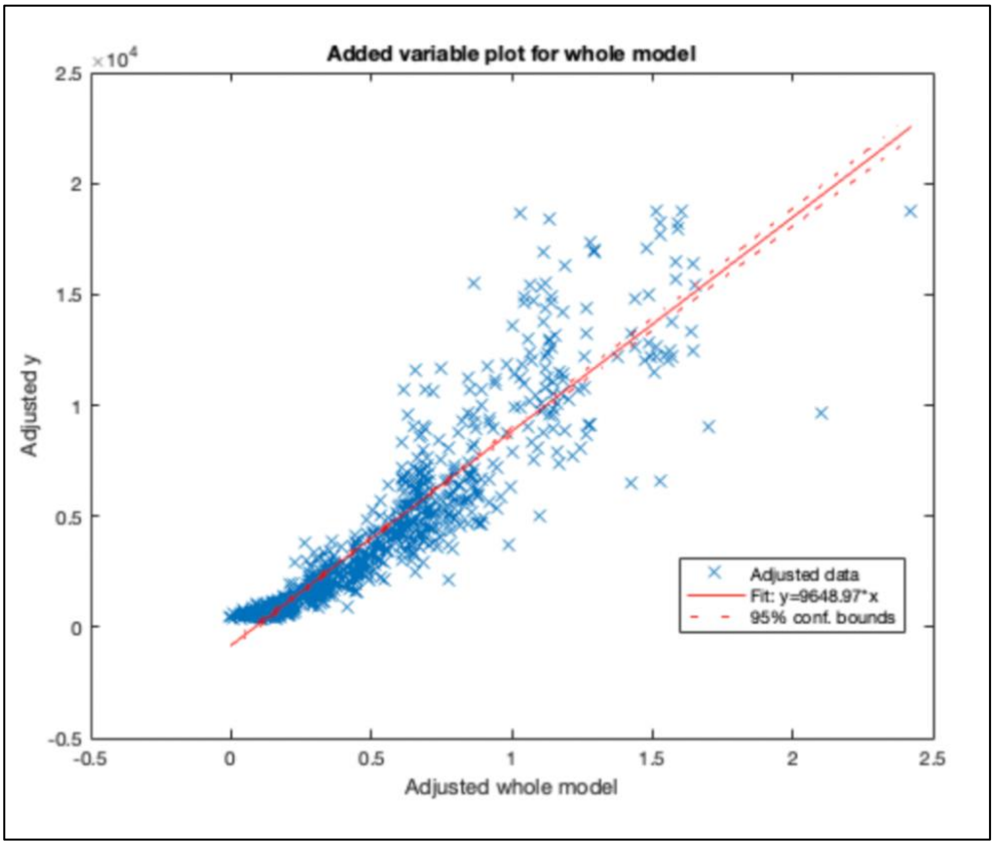
Table 4

PLOT OF THE FITTED MODEL

Using the code documented in *Code 6*, the plotted *Graph 1* was produced. From primary observations, one can see that the strong majority of data is well-behaved, and that there are very few outliers. However, there does appear to be a cluster of a datapoints on the lower left-hand side – one can postulate the this could possibly be due to:

1. the fact that the dataset focused on lower quality diamonds, or
2. the simple reasoning that high quality diamonds are simply more uncommon.

As a final note, visually one can see that while there are outliers, they are still in conjunction with the R^2 value of 0.8464, meaning that there is scope for improvement in the model.



Graph 1

DIAGNOSTIC TEST FOR RESIDUALS

NORMALITY

The concept of normality is founded in the works of Carl Gauss, in particular, his theory of errors of observations. When testing for normality, one can utilise the following techniques:

Skewness

The skewness represents the asymmetrical nature of the distribution being studied, with a skewness of zero indicating that there is perfectly symmetrical distribution. When the above-seen function in *Code 8* is implemented, one observes a value of 0.9404, indicating that the data is slightly skewed to the right. As illustrated in the histogram of residuals, this does in fact hold appear to be true.

Kurtosis

Kurtosis relates to the degree to which the distribution of a frequency is peaked or is flat. The standard normal distribution maintain a kurtosis of 3, which is labelled as 'mesokurtic' (Kallner 2018). In this paper's study, one estimates a value of 11.0013, which indicates that the residuals are extremely leptokurtic – this would argue strongly against the normality of the model.

Jarque-Bera Test

$$H0: data\ is\ normally\ distributed$$

$$H1: data\ is\ not\ normally\ distributed$$

For a test of normality, one can implement the Jarque–Bera Test statistic; a test which is founded on the aforementioned principles. It is expected to have a p-value of zero which would indicate normality. In this paper's instance, as per *Code 9,* a p-value of $9.369e^{--0.5}$ indicates that one should reject the null hypothesis, concluding that the residuals of this model are not normally distributed.

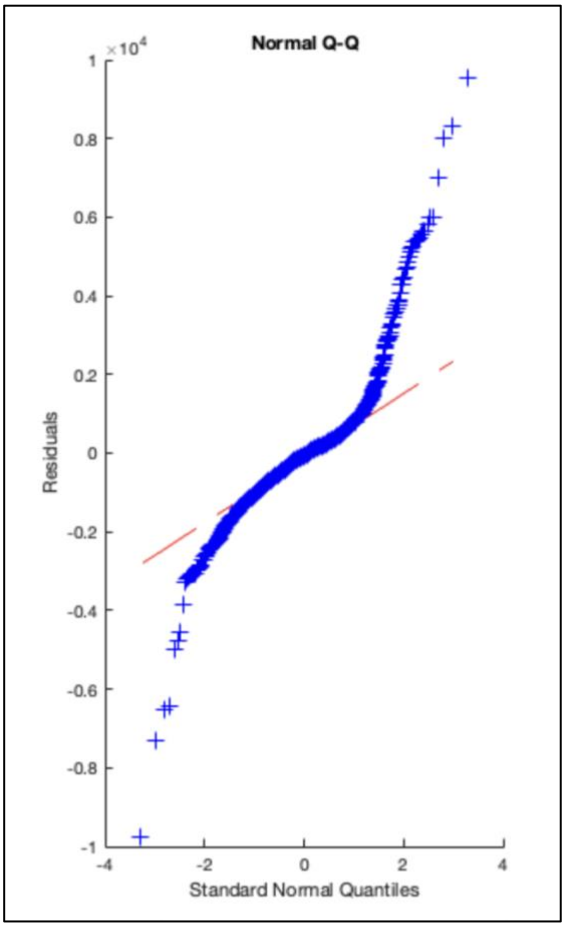
	Value
Jarque-Bera stat	2.81E+03
Chi-Square Critical	5.9282
Jarque-Bera p-value	$9.369e^{--0.5}$

Table 5

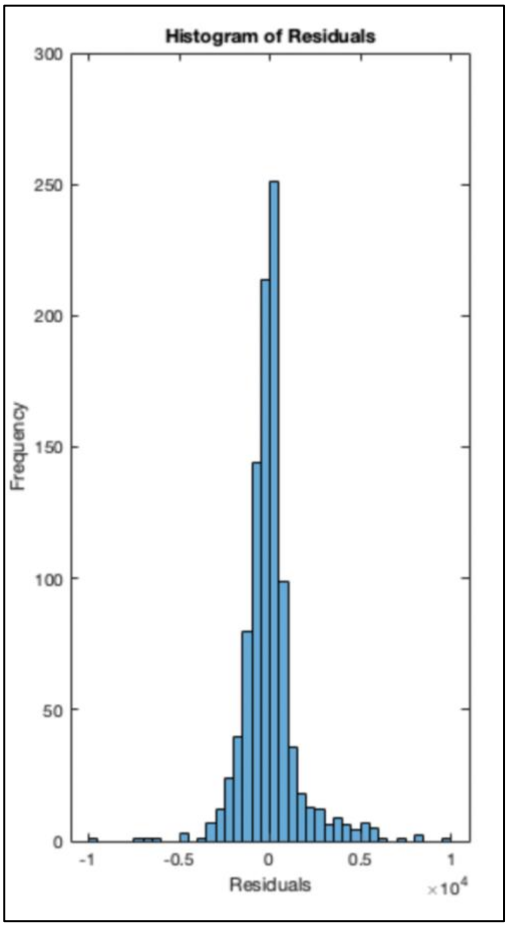
As seen in the above table, the Jarque-Bera test statistic exceeds the critical value, thus leading this paper to reject both the null hypothesis and the assumption of normality.

Plotting Normality

Using *Code 7,* a 'qqplot' and a histogram were graphed, the results of which are presented below. From first observations, one can see that there is the case made for leptokurtosis in *Graph 3,* as the residuals are heavily concentrated about the mean. Furthermore, the plot appear to be slightly asymmetric, which would support the previous claim of skewness. As for the qqplot, the plotted residuals don't really form a straight line and curve off in the extremities. Typically, one would associate such a graph with the fact that the data has more extreme values than they should if they were to be considered normal.



Graph 2



Graph 3

HETEROSCEDASTICITY

As per Gujarati and Porter (2009), the linear regression model assumes that the error term in the regression model has equal variance across the observations. However, if heteroscedasticity holds, then this variance is in fact time-varying. If this issue is present, then unfortunately one cannot use the F-statistic or t-statistic to make an inference. When testing for the presence of heteroscedasticity, one can implement either the White Test or the Breusch-Pagan Test, the results of which are shown below:

Breusch-Pagan Test

$$H0: there\ is\ no\ heteroscedasticity$$
$$H1: there\ is\ heteroscedasticity$$

When examining for the presence of heteroscedasticity in a regression analysis, one has the option of implementing a Breusch-Pagan Test. The null hypothesis of this test is that there is no heteroscedasticity in the model – failure to reject this implies homoscedasticity. The output of this test is a statistic labelled the ‘Chi-Square distributed’, and, it is supported with a corresponding p-value which gives indication as to the statistical significance of the test. However, one weakness of the Breusch-Pagan Test is that it assumes the heteroscedasticity is a linear function of the independent variables – in other words, failure to find heteroscedasticity with the Breusch-Pagan does not exclude the possibility of a non-linear relationship heteroscedasticity (Zaman 2008).

White Test

$$H0: there\ is\ no\ heteroscedasticity$$
$$H1: there\ is\ heteroscedasticity$$

The White Test is a more generalised and flexible test than the Breusch-Pagan, and is underpinned by assumptions such as no normality and no linearity. As seen above, the null hypothesis for this model is that the variances for the error terms are equal. However, by allowing for the testing of non-linear heteroscedasticity as well as the inclusion of cross terms and powers and powers in its auxiliary regression, the White Test allows for a more complete image of homoscedasticity.

	Breusch-Pagan	**White**
Critical Value	1074.7	24.9958
p-value	0.000093699	269.0703

Table 6

By running *Code 10* and *Code 11*, this paper obtains contradicting results, as seen in *Table 5*. Thus, one could possibly infer that there may be non-linear heteroscedasticity in the model that the Breusch-Pagan test is not detecting. One possible solution to this issue that econometricians can implement is the ‘Weighted Least Squares Method’.

SERIAL CORRELATION

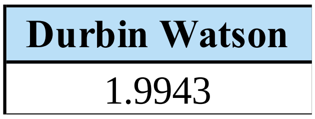
Serial correlation occurs when there is a relationship between a variable of a regression model and a lagged version of itself over various time intervals, and, is a violation of one of the OLS assumptions [SOURCE]. The issue associated with serial correlation is that it causes the estimated variances of the regression coefficients to be biased, which results in the ability to rely on the hypothesis testing [SOURCE]. One method to test for serial correlation is the implementation of the Durbin-Watson Test, as proposed in the works of Durbin and Watson (1955) on ‘Least Square Regressions’.

Durbin Watson Test

$$H0\colon \phi = 0, \mathit{there\ is\ no\ serial\ correlation\ between\ residuals}$$
$$H1\colon \phi \neq 0, \mathit{there\ is\ serial\ correlation\ present}$$

The null hypothesis of the Durbin Watson test is that there is no 1st order serial correlation between in the residuals of one’s regression analysis. It is imperative to test for such an issue, as serial correlation can lead to the underestimation of standard errors, which consequently results in one falsely presuming the significance of predictors. Ignoring inconclusive regions, the output of the Durbin Watson Test is a statistic which ranges from 0 to 4;

DW = 0 (Positive serial correlation)
DW = 2 (No serial correlation)
DW = 4 (Negative serial correlation)



Durbin Watson
1.9943

In this study’s regression, the Durbin Watson statistic was evaluated to be **1.9943** (*Code 7*), which indicates that there is no serial correlation present in this model. However, as specified earlier, the Durbin Watson can only test for serial correlation in the first lag – if one wanted to seek further confidence in the output of the model, one could opt to implement the Breusch-Godfrey Test, as this would test for serial correlation at higher orders.

MULTICOLLINEARITY

Multicollinearity creates a dynamic whereby the R^2 value becomes inflated due to a correlation between the independent variables in the model. As per Farrar and Glauber (1967), econometricians understand that multicollinearity undermines the 'best linear unbiased' element of the OLS regression model. This issue violates one of the assumptions underpinning the OLS Regression – that no independent variable is a linear function of one or more other independent variables (ibid). Multicollinearity comes in two forms: perfect and imperfect. In regard to the former, one cannot estimate the coefficients, while in the case of the latter, the coefficients estimated result in an inefficient model. One can test for the presence of the multicollinearity is the Variance-Inflation Factor (VIF), as discussed below.

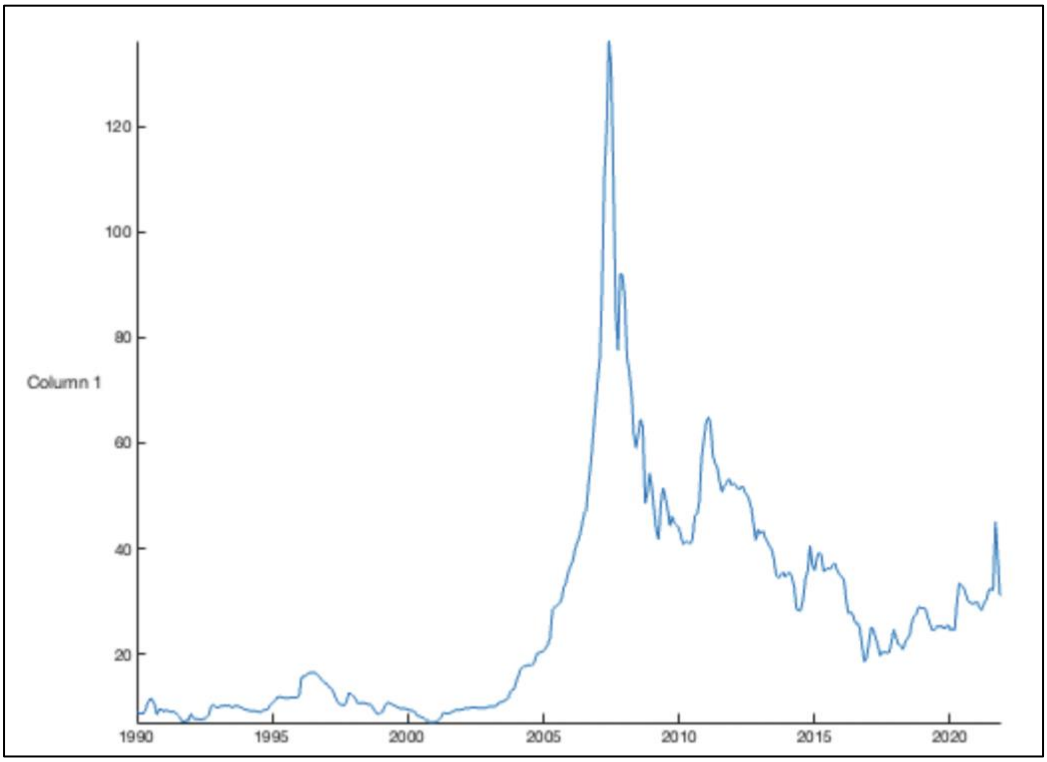
VIF

Fox (1984) describes the VIF as an indication as to the degree at which the accuracy of the model is hindered by the presence of multicollinearity. As a rule of thumb, a VIF above 10 indicates high correlation; as the estimated VIF for this paper is **6.5122** (*Code 12*), one should not be concerned. However, one should also acknowledge that conservative scholars often use a VIF of 5 – in this case, one should be prepared to support their decision to continue using such a model (Daoud 2017). Interpreting the VIF from this paper, as R^2 is above .80, and VIF is above 5, this paper should be wary of possible multicollinearity. In this instance the variables 'Length' and 'Width' may be sources of multicollinearity. To resolve this, one could consider the implementation of a Ridge Regression.

Ridge Regression

Proposed by Hoerl (1962), the Ridge regression offers an improvement to the independent coefficients by updating the insignificant coefficients to a value between 0 and 1, thereby weighting their contribution to the mode (Hoerl and Kennard 2000). While the ridge regression is not a panacea for multicollinearity, it is a very useful procedure. But one of the problems with ridge regression is the selection of the best quantity to be added to the diagonal in order to minimize bias against variance – often the solution is worse than the cause when curing multicollinearity.

QUESTION 3 – ARIMA



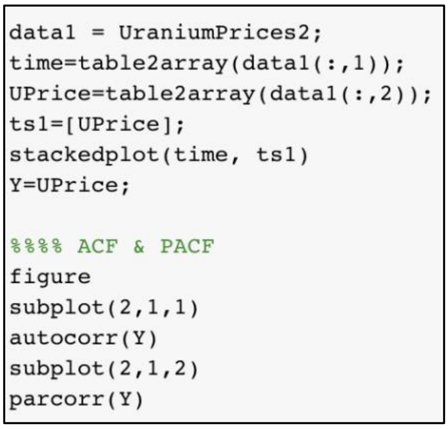
Graph 4

To examine the predictive powers of ARIMA in regard to time series data, this paper extrapolated monthly data on the dollar price per pound of uranium from the FRED database. From first observations, one can correlate several price points to important historical events:

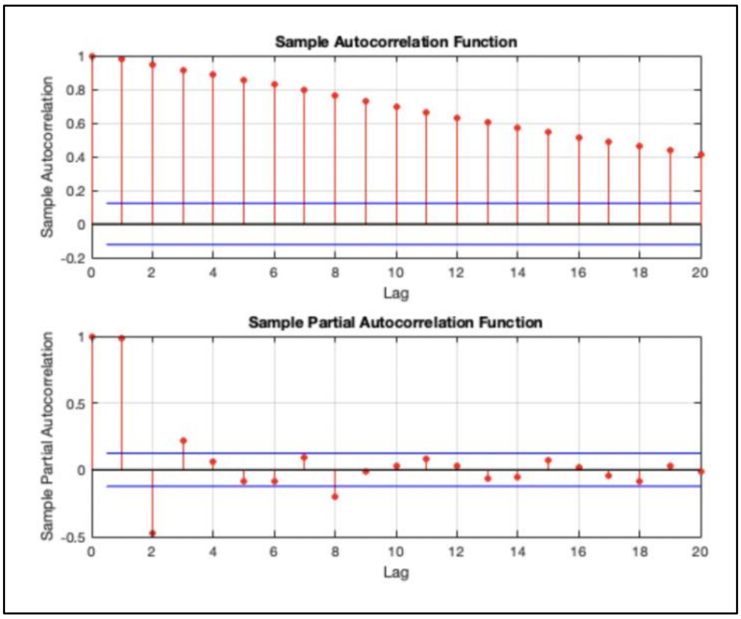
- 2006-2007: a massive supply shock erupts due to the flooding of North America's largest mining facility, causing the price per pound to skyrocket to $140
- 2010-2011: the occurrence of Fukushima sends the price of uranium into a 10-year bear market, as institutions and governments divest into other sources of energy
- 2020-onwards: as nuclear is being considered for EU taxonomy, and one observes a major number of reactors coming back online, the supply deficit (due to the previous bear market) is being felt. Thus far, the price per pound has doubled within a year, and it is the opinion of this paper that the possible price action in coming years will be one comparable to 2007. The sustained supply deficit married with both the Chinese government's intents to build 18 new reactors, as well as a multi-national movement

back into nuclear will cause an excessive amount of demand, while the supply simply is no longer there to match it.

Plot of Autocorrelation Function and The Partial Autocorrelation Function

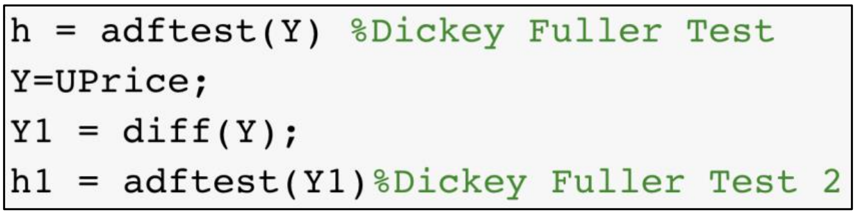


```
data1 = UraniumPrices2;
time=table2array(data1(:,1));
UPrice=table2array(data1(:,2));
ts1=[UPrice];
stackedplot(time, ts1)
Y=UPrice;

%%%% ACF & PACF
figure
subplot(2,1,1)
autocorr(Y)
subplot(2,1,2)
parcorr(Y)
```



Graph 5

Plotting the original dataset, one can observe that the data is not stationary a stationary process – to confirm this observation, a Dickey-Fueller test was implemented, which gave a value of 0. This value confirms the presence of a unit-root, implying that the data is no longer stationary. Thus, a transformation is necessary, which is performed in the following code:

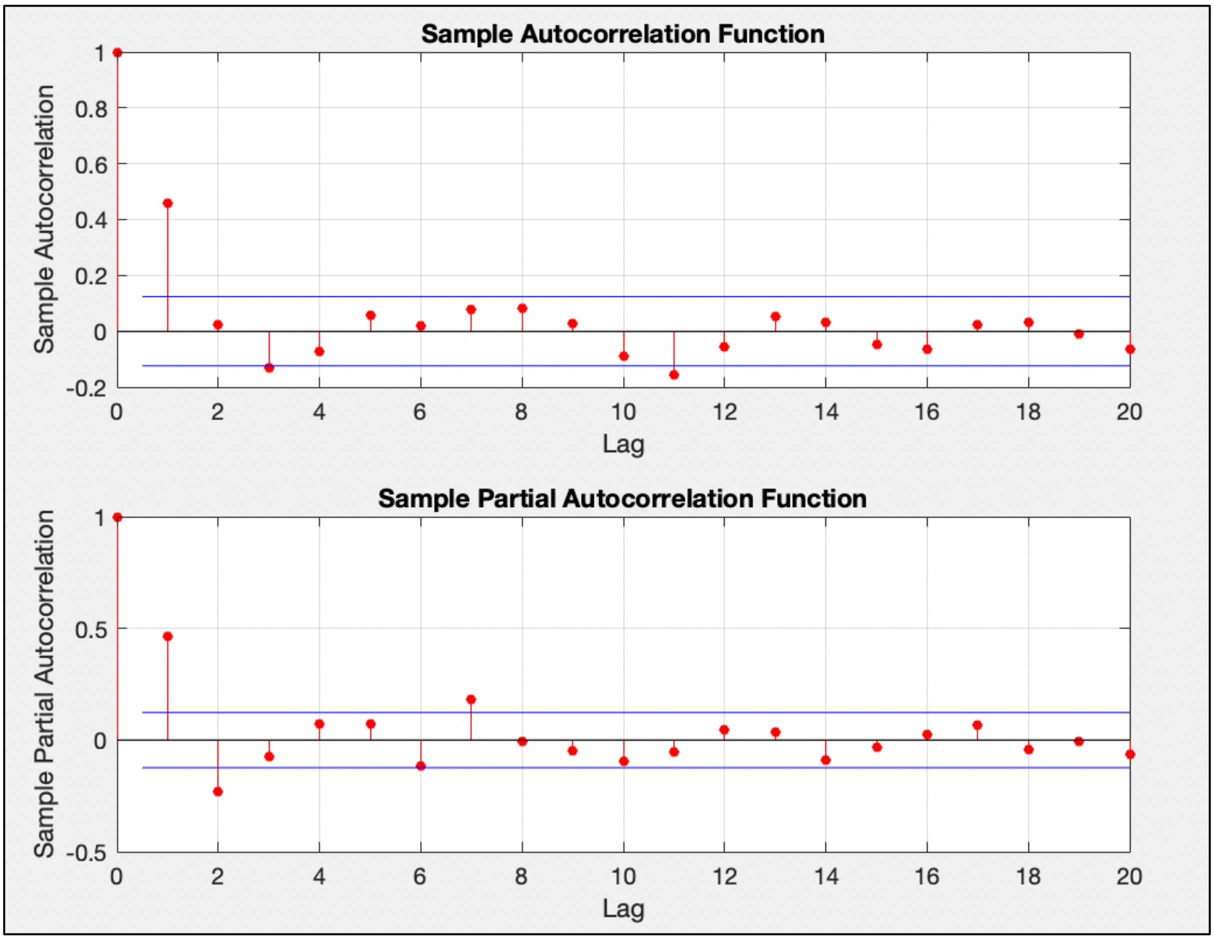


```
h = adftest(Y) %Dickey Fuller Test
Y=UPrice;
Y1 = diff(Y);
h1 = adftest(Y1)%Dickey Fuller Test 2
```

This test resulted in an output of 1, indicating that the data now rejects the presence of a unit root.

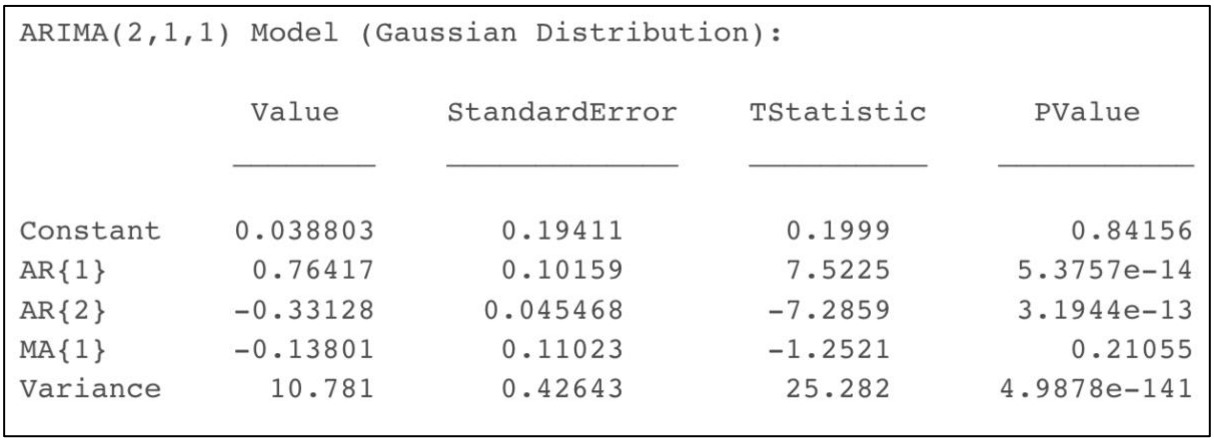
To commence the forecasting process this paper first plots the partial autocorrelation function (PACF) and the autocorrelation function (ACF), as this visualization allows:

- An observation of what point the lag exists in the dataset.
- An observation of the rate of decay of the data.



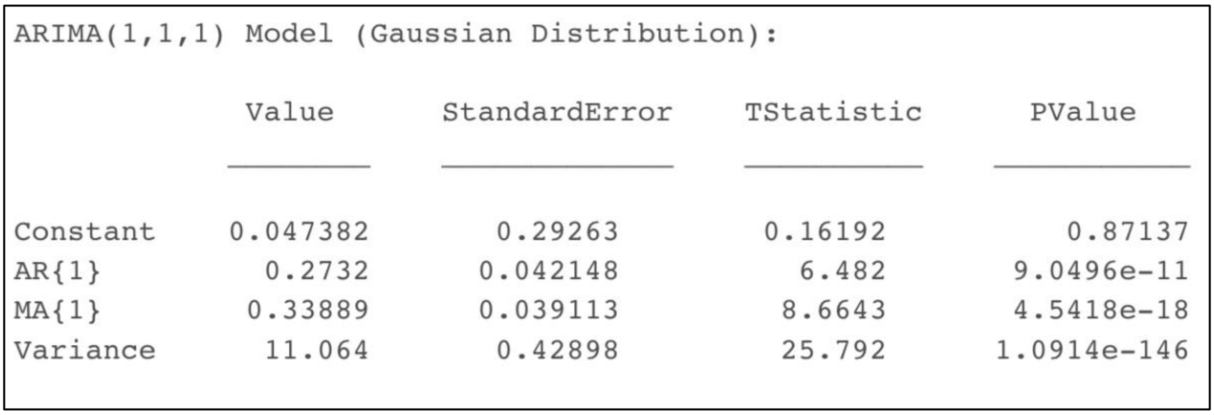
Graph 6

In both instances, it can be seen that the residuals are mean reverting – in economic terms, this means that the price of uranium reverts back to its long-term mean. As seen in the correlogram, the differentiation performed on the data means that the series is now a time stationary one, which enables the implementation of the ARIMA. This paper ran four different loops to estimate the most suitable ARIMA model – from observing the resulting BIC (Bayesian Information Criterion) output, the model ARIMA(2,1,1) was selected. To estimate the efficiency of this ARIMA (2,1,1) model, it is compared to an arbitrarily selected ARIMA (1,1,1) model – and is compared on the grounds of RMSFE and Maximum Likelihood.



```
ARIMA(2,1,1) Model (Gaussian Distribution):

                  Value      StandardError    TStatistic      PValue   
                ________    _____________    __________    ___________

    Constant    0.038803       0.19411         0.1999        0.84156
    AR{1}        0.76417       0.10159         7.5225     5.3757e-14
    AR{2}       -0.33128      0.045468        -7.2859     3.1944e-13
    MA{1}       -0.13801       0.11023        -1.2521        0.21055
    Variance      10.781       0.42643         25.282    4.9878e-141
```

Graph 7 – ARIMA (2,1,1)

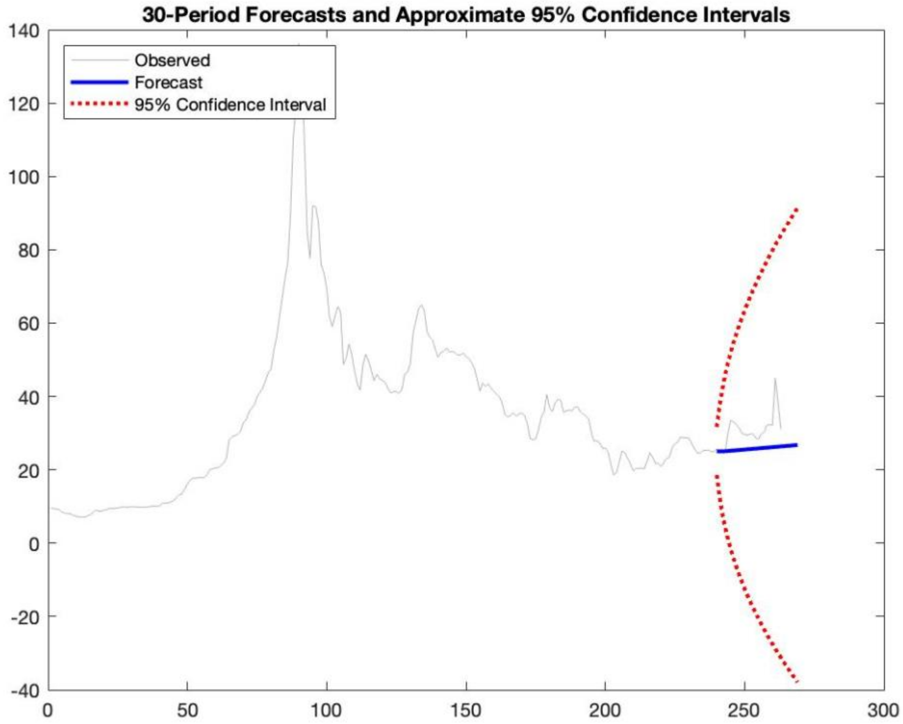


```
ARIMA(1,1,1) Model (Gaussian Distribution):

                  Value      StandardError    TStatistic      PValue   
                ________    _____________    __________    ___________

    Constant    0.047382       0.29263        0.16192        0.87137
    AR{1}         0.2732      0.042148          6.482     9.0496e-11
    MA{1}        0.33889      0.039113         8.6643     4.5418e-18
    Variance      11.064       0.42898         25.792    1.0914e-146
```

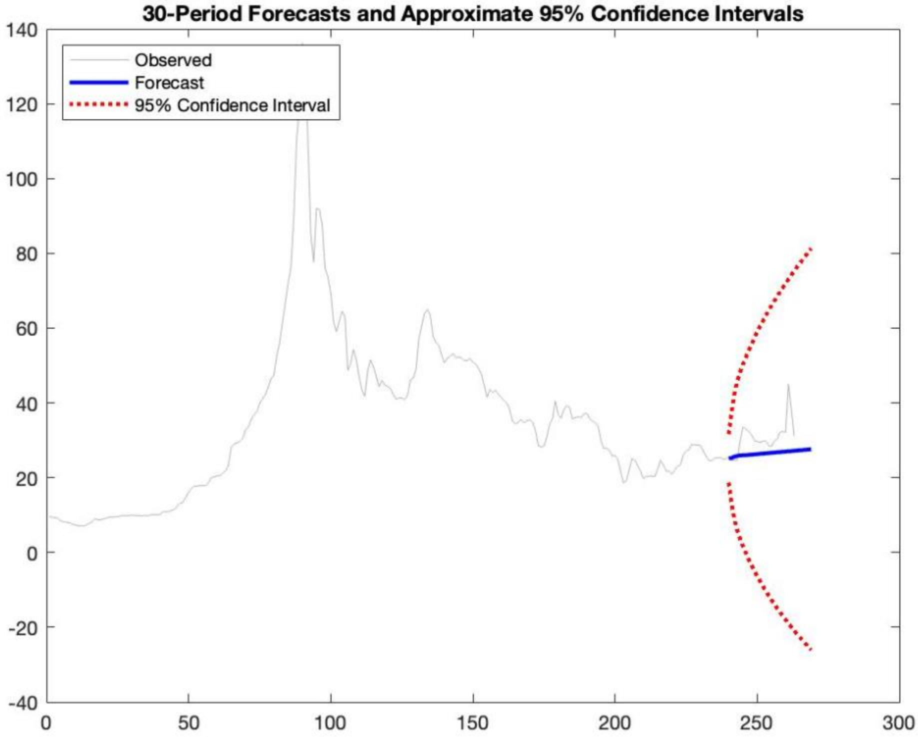
Graph 8 – ARIMA(1,1,1)

	RMSFE Value
ARIMA(2,1,1)	24.3528
ARIMA(1,1,1)	29.6011

The new RMSFE of the ARIMA(2,1,1) is estimated to be marginally smaller than that of ARIMA(1,1,1) meaning that is a better estimator to use.



Graph 9 – Plotted ARIMA (2,1,1)



Graph 10 – Plotted ARIMA (1,1,1)

In *Graph 9* and *Graph 10*, the 30-month forecast of the price of uranium per pound is plotted alongside the upper and lower thresholds of the 95% Confidence Intervals. As indicated by both graphs, there is a potential for a large possible movement in the price – a suggestion which would reinforce my process of thought about a looming supply shock.

Diebold Mariano Test

The Diebold Mariano Test is also implemented to verify if the aforementioned ARIMA models are statistically significant from each other. In this paper's analysis, the p-value of this test was estimated to be 0.1877, indicating that we fail to reject the null hypothesis of this test. One can infer from this result that the models do not have different estimation accuracies. To verify this, one option is to consult both *Graph 9* and *Graph 10* – the are quite comparable and essentially indistinguishable. Furthermore, this rejection of the null hypothesis is supported by the RMSFE values, which were are as follows:

	RMSFE Value
ARIMA(2,1,1)	24.3528
ARIMA(1,1,1)	29.6011

From a quick observation, one can see that ARIMA(2,1,1) only performs marginal better than ARIMA(1,1,1).

BIBLIOGRAPHY

1. Burton, A. (2020) ‘OLS Linear Regression, The encyclopaedia of research methods in criminology and criminal justice (pp.509 - 514), Chapter: 104, available: https://www.researchgate.net/publication/339675576_OLS_Linear_Regression [accessed 18 Dec 2021].

2. Daoud, J. (2017) Multicollinearity and Regression Analysis, Journal of Physics, Conf. Ser. 949, available: https://iopscience.iop.org/article/10.1088/1742-6596/949/1/012009/pdf [accessed 19 Dec 2021].

3. Fox, J. (1984) *Linear Statistical models and Related Methods: with applications to social research, New York: John Wiley*.

4. Sykes, A. (1993) An introduction to regression analysis, (Coase-Sandor Institute for Law & Economics Working Paper, 20, available: https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1050&context=law_and_economics [accessed 18 Dec 2021].

5. Gujarati, D. N. and Porter, D. C. (2009) Basic econometrics. Boston, Mass, McGraw-Hill.

6. Watson, G. S. (1955) ‘Serial Correlation in Regression Analysis’, *Biometrika*, 42(4), 327–341, available: https://doi.org/10.2307/2333382

7. Hoerl, A. E. and Kennard, R. W. (2000) ‘Ridge Regression: Biased Estimation for Nonorthogonal Problems’, *Technometrics*, 42(1), 80–86, available: https://doi.org/10.2307/1271436