

## Part I (Time Series)

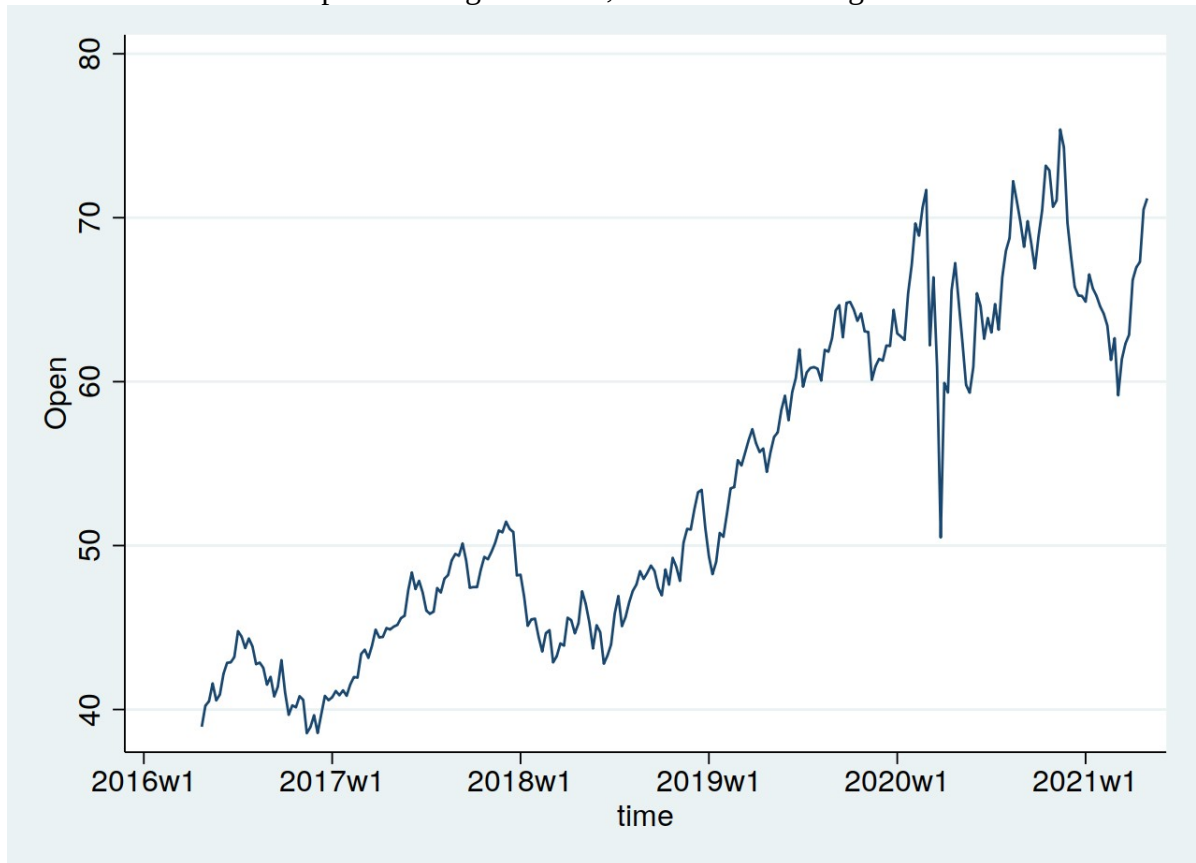
Assigned weekly data set: XEL - Xcel Energy Inc

### A) Data Description

to describe the time series we can use the summarize function in Stata.

To visualize the time series the variable was used to represent the prices of the stock for each week.

We use then those values to plot them against time, as shown in the figure below.



We can see a positive trend until around 2020Q2 when a big crash happens (most likely due to the pandemic), then the positive trend continues so that by the middle of 2020Q4 we reach an all time high. At the beginning of 2021 there is a significant drop that recovers to around pre-pandemic levels as of today.

Also the time series follows the multiplicative model

### b) Stationarity

To formally test stationarity we can use the DICKEY-FULLER UNIT ROOT TEST.

The null hypothesis ( $H_0$ ) assumes that we don't have a stationary time series, therefore if the test statistic for our time series is more negative (smaller) than the critical value for our test then we can reject our null hypothesis and assume that the alternative hypothesis ( $H_1$ ) is true for our significance level ie. the time series is stationary.

Anonymous code: 0016-DBM

We can clearly see that our original timeseries is non stationary because the expected value of the series increases with time ie. dependent on it.

To make the the time series stationary, we apply a logtransform on the data so that our multiplicative model becomes an additive model. Let us call the new variable from this transformation *logprice*. After doing the test we can see that, by applying the decision criterion described above, that our time series is still not stationary.

**Stata output:**

Dickey-Fuller test for unit root                      Number of obs =    258

----- Interpolated Dickey-Fuller -----				
Test	1% Critical	5% Critical	10% Critical	
Statistic	Value	Value	Value	
-----				
Z(t)	-1.506	-3.459	-2.880	-2.570
-----				

MacKinnon approximate p-value for Z(t) = 0.5305

We create another variable, *logreturn*, that is the difference between two logprices. Doing the test again we come to the conclusion that this time series is now stationary.

**Stata output:**

Dickey-Fuller test for unit root                      Number of obs =    257

----- Interpolated Dickey-Fuller -----				
Test	1% Critical	5% Critical	10% Critical	
Statistic	Value	Value	Value	
-----				
Z(t)	-18.134	-3.459	-2.880	-2.570
-----				

MacKinnon approximate p-value for Z(t) = 0.0000

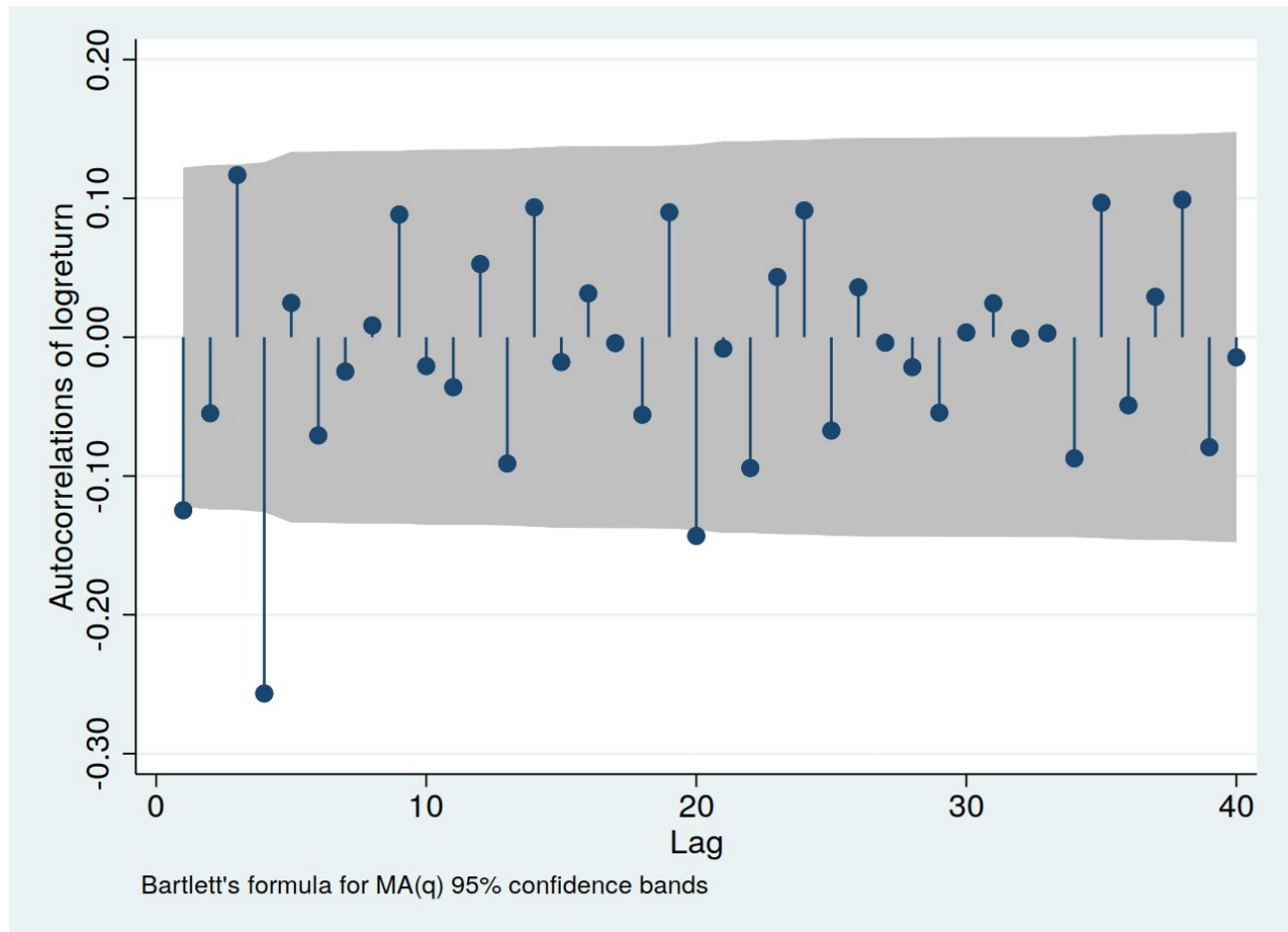
## c) ACF, PACF plots

ACF: on the figure below we can see that 1<sup>st</sup>, 4<sup>th</sup> and 20<sup>th</sup> are significant.

The Auto Correlation Function (ACF) shows us how the series autocorrelates with its lagged values. The first plot point shows how the current value is correlated with the previous value in the time series.

It is quite improbable that we have a process that its values so far back as the 20<sup>th</sup> would contribute to our current values, therefore we can assume that we have an MA(4) by looking at it. Also the 20<sup>th</sup> value is just outside the significance level which also confirms our suspicions.

Anonymous code: 0016-DBM



PACF:

PACF shows us how the current value correlates with the ones before it after only the residuals remain.

Based on the two figures above we can guess an ARMA(4,4) process.

The chosen models (with their Stata outputs) are the following:

	OPG					
D.logprice	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
logprice						
_cons	.0021012	.0019364	1.09	0.278	-.001694	.0058964

Anonymous code: 0016-DBM

```
/sigma | .0305286 .0005615 54.37 0.000 .029428 .0316292
```

---

We do The Box-Ljung test to be sure that there are non-zero lags.

Stata gives the folloing answer:

### Portmanteau test for white noise

---

Portmanteau (Q) statistic = 64.3565

Prob > chi2(40) = 0.0086

The null hypothesis can be rejected, i.e we have autocorrelation in our sample.

### ARIMA (4,1,4)

Sample: 2016w18 - 2021w15                      Number of obs = 258

Wald chi2(8) = 96.11

Log likelihood = 548.9265                      Prob > chi2 = 0.0000

---

	OPG						
D.logprice	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]		
<hr/>							
logprice							
_cons	.0019574	.0013622	1.44	0.151	-.0007124	.0046272	
<hr/>							
ARMA							
ar							
L1.	.1875266	.3388256	0.55	0.580	-.4765593	.8516125	
L2.	.349967	.1815143	1.93	0.054	-.0057945	.7057285	
L3.	-.0839643	.2394052	-0.35	0.726	-.5531899	.3852612	
L4.	-.2500396	.1790949	-1.40	0.163	-.6010592	.10098	
ma							
L1.	-.2889113	.3516323	-0.82	0.411	-.9780979	.4002753	
L2.	-.4348548	.1970665	-2.21	0.027	-.821098	-.0486117	
L3.	.2465757	.2847424	0.87	0.387	-.311509	.8046605	
L4.	-.0253647	.1877675	-0.14	0.893	-.3933823	.3426528	
<hr/>							
/sigma	.0287974	.0009707	29.67	0.000	.0268949	.0306999	

---

### ARIMA (3,1,3)

Sample: 2016w18 - 2021w15                      Number of obs = 258

Wald chi2(5) = 1593.37

Log likelihood = 549.8504                      Prob > chi2 = 0.0000

Anonymous code: 0016-DBM

-----						
	OPG					
D.logprice	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
logprice						
_cons	.0022334	.0002412	9.26	0.000	.0017607	.0027061
-----+-----						
ARMA						
ar						
L1.	-.7791105	.0814294	-9.57	0.000	-.9387091	-.6195119
L2.	.6600674	.0464092	14.22	0.000	.5691072	.7510277
L3.	.8029925	.0624968	12.85	0.000	.680501	.925484
ma						
L1.	.7213404	.0781426	9.23	0.000	.5681837	.874497
L2.	-.872652	.	.	.	.	.
L3.	-.848688	.0855309	-9.92	0.000	-1.016326	-.6810504
-----+-----						
/sigma	.0285433	.0008598	33.20	0.000	.0268581	.0302285
-----						

### ARIMA (0,1,3)

Sample: 2016w18 - 2021w15      Number of obs = 258  
Wald chi2(3) = 40.97  
Log likelihood = 537.436      Prob > chi2 = 0.0000

-----						
	OPG					
D.logprice	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
logprice						
_cons	.0020968	.0020133	1.04	0.298	-.0018491	.0060427
-----+-----						
ARMA						
ma						
L1.	-.0751418	.0338398	-2.22	0.026	-.1414667	-.008817
L2.	-.1059497	.0432263	-2.45	0.014	-.1906716	-.0212277
L3.	.0965491	.0375781	2.57	0.010	.0228973	.1702008
-----+-----						
/sigma	.0301332	.0006227	48.39	0.000	.0289128	.0313537

### ARIMA (2,1,3)

Sample: 2016w18 - 2021w15      Number of obs = 258  
Wald chi2(5) = 3644.45  
Log likelihood = 546.4048      Prob > chi2 = 0.0000

-----

Anonymous code: 0016-DBM

	OPG					
D.logprice	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
logprice						
_cons	.0021047	.0019734	1.07	0.286	-.0017631	.0059726
-----+-----						
ARMA						
ar						
L1.	-1.686648	.0713662	-23.63	0.000	-1.826523	-1.546773
L2.	-.8612466	.0642386	-13.41	0.000	-.987152	-.7353412
-----+-----						
ma						
L1.	1.630913	.0707995	23.04	0.000	1.492149	1.769678
L2.	.6459686	.0820831	7.87	0.000	.4850887	.8068485
L3.	-.1137654	.041935	-2.71	0.007	-.1959565	-.0315743
-----+-----						
/sigma	.0290714	.0008452	34.40	0.000	.0274149	.030728

Note: The test of the variance against zero is one sided, and the two-sided confidence interval is truncated at zero.

We can see that all lags are significant ( $p \leq 0.05$ ) for the models apart from ARIMA (4,1,4) model, which can be discarded because of that.

The AIC scores of the models above can be summarized by the following table using Stata's output:

Akaike's information criterion and Bayesian information criterion

Model	N	ll(null)	ll(model)	df	AIC	BIC
-----+-----						
arima010	258	.	534.0997	2	-1064.199	-1057.093
arima414	258	.	548.9265	10	-1077.853	-1042.323
arima313	258	.	549.8504	7	<b>-1085.701</b>	-1060.83
arima013	258	.	537.436	5	-1064.872	-1047.107
arima213	258	.	546.4048	7	<b>-1078.81</b>	-1053.939
-----+-----						

## e) choosing 2 best ARIMA, RMSE calculation

Based on the AIC scores, the lower the score (more negative) the better thus the ARIMA (3,1,3) and ARIMA (2,1,3) models are chosen.

The RMSE values are really close and show how "accurate" the prediction is compared to our test data. The lower the RMSE the better.

The formula for calculating RMSE can be found in the formula sheet therefore only the table and calculations will be provided, as shown below

Anonymous code: 0016-DBM

(Obs- predicted)^2	forecast313	forecast213
	0.02897484	0.03198732
	84	2499999
	10.4336752	10.5458016
	144	049001
	14.0600251	14.2559860
	089001	041
RMSE	<b>8.17422505</b>	8.27792497
	<b>723335</b>	716669

ARIMA (3,1,3) has lower RMSE thus is more accurate for prediction, therefore it is chosen as our best model.

### f) GARCH effect of the “best” model

We will now test the GARCH effects for our ARIMA (3,1,3) model.

Using Stata for LM test for autoregressive conditional heteroskedasticity (ARCH) we can come to the conclusion that there are garch effects present in our model.

H0: no ARCH effects

H1: there are ARCH effects

In part d it would mean that the error terms are not constant over time and therefore certain effects in the time series such as volatility clustering for example could not be modelled thus the models' predictive value (accuracy) is greatly diminished.

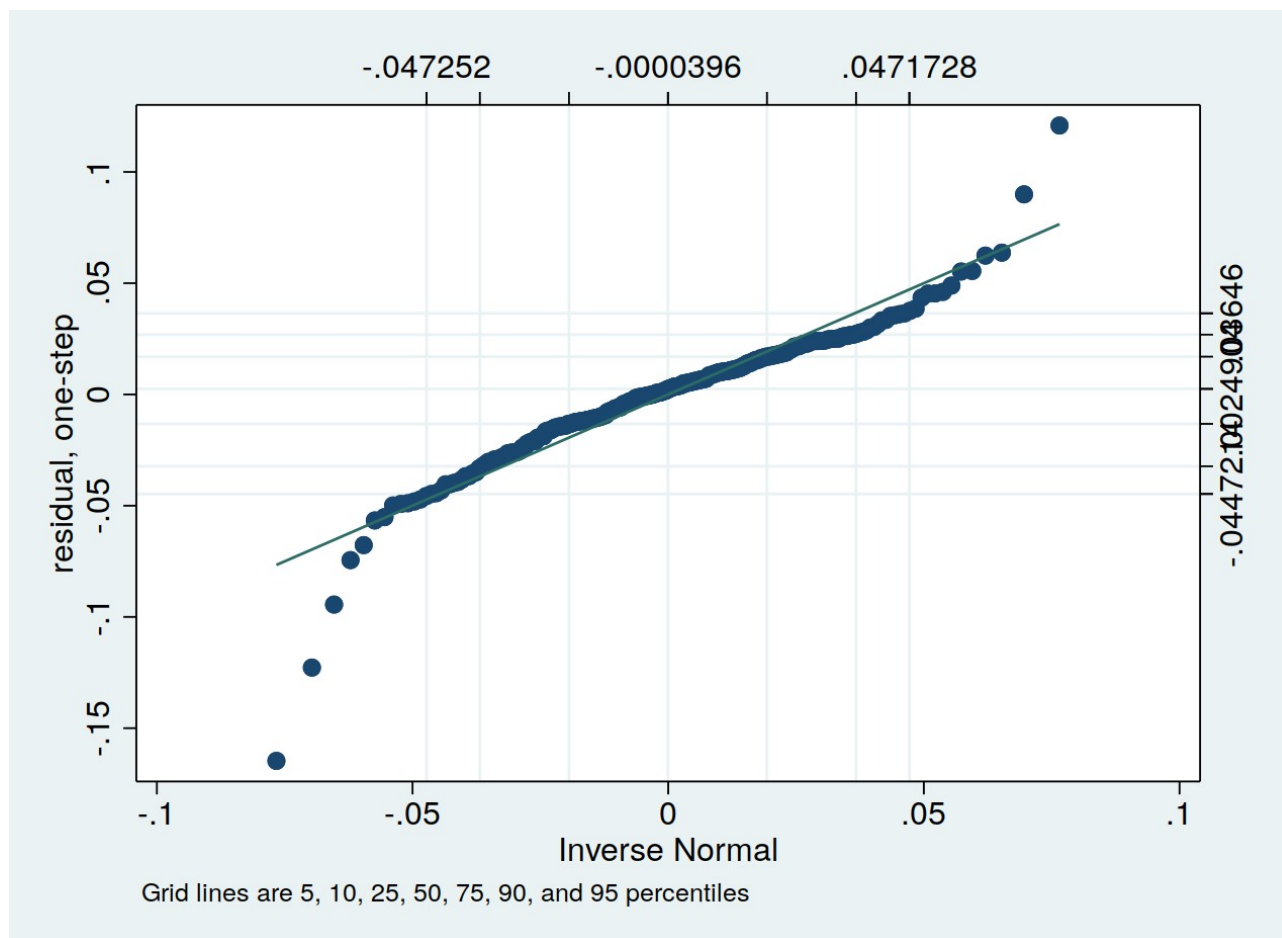
### g) Residual analysis of the best ARIMA and Conclusion

By analyzing the residuals we want to know about the normality, independence and constant variance (time independent) of the residuals.

QQ-plot for the residuals:



Anonymous code: 0016-DBM



We see that the residuals have fatter tails than the normal distribution.

#### A more formal test:

Skewness and kurtosis tests for normality

----- Joint test -----					
Variable	Obs	Pr(skewness)	Pr(kurtosis)	Adj chi2(2)	Prob>chi2
-----+-----					
residual313	261	0.0000	0.0000	51.55	0.0000

The p value is significant therefore we can reject the null hypothesis.

**Independence: we use the Ljung box test test.**

Portmanteau test for white noise

-----  
Portmanteau (Q) statistic = 31.2342  
Prob > chi2(40) = 0.8380

We see that the residuals are independent (null hypothesis cannot be rejected).

We know from previous part that there are garch effects therefore the variance is not constant.

Anonymous code: 0016-DBM

## **Conclusion**

We can conclude that Garch effects are present in our model, therefore our predictions would be more accurate if those were implemented.

Anonymous code: 0016-DBM

## Part II (Regression)

### A) Data summary

Variable	Obs	Mean	Std. Dev.	Min	Max
-----+-----					
price	100	2925.09	3402.705	<b>448</b>	<b>17029</b>
carat	100	.6312	.3828476	.23	1.7
color_def	100	.5	.5025189	0	1
color_gh	100	.44	.4988877	0	1
clarity_if	100	.11	.314466	0	1
-----+-----					
clarity_vs	100	.32	.4688262	0	1
clarity_vvs	100	.24	.4292347	0	1

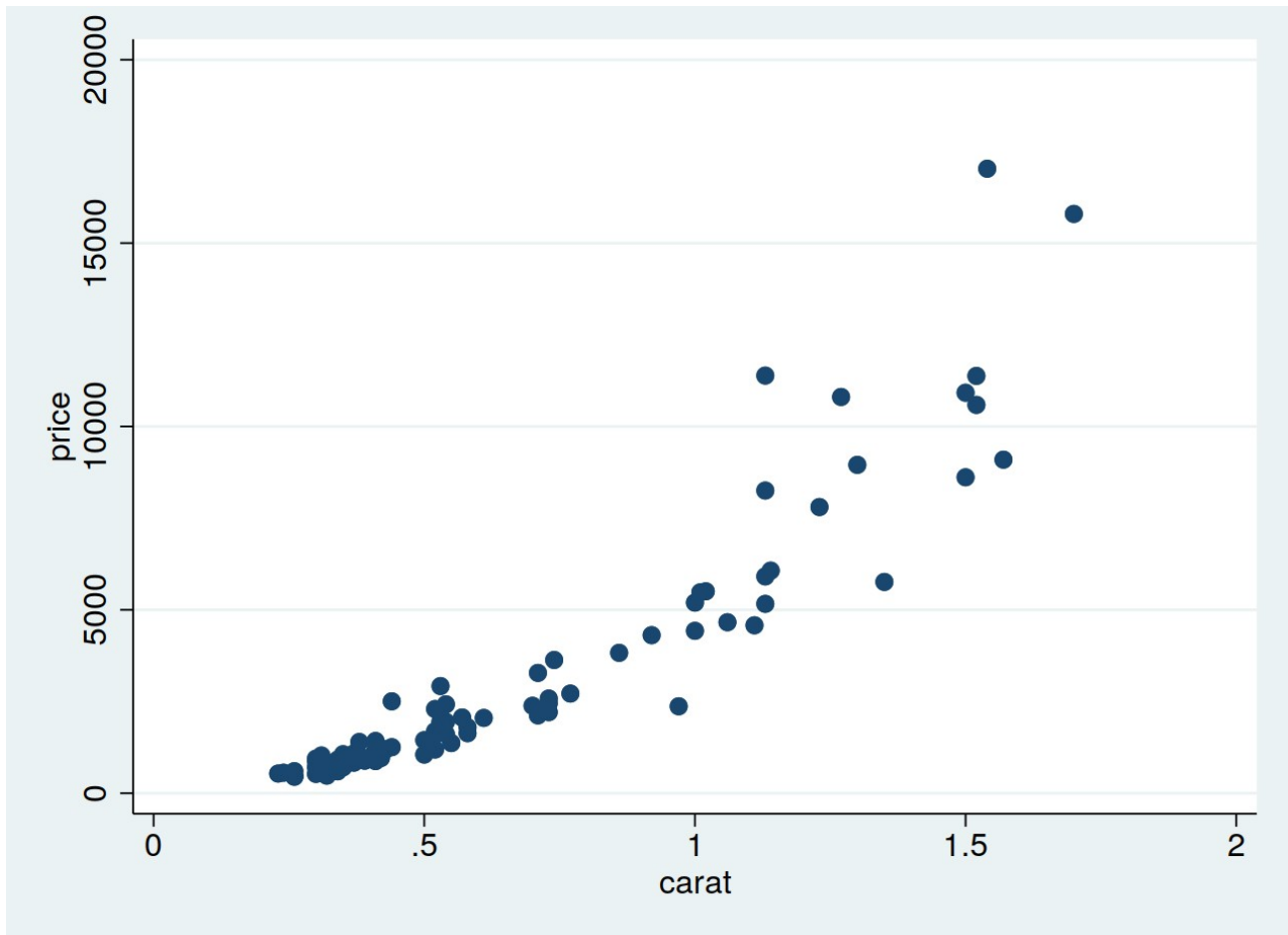
By looking at the table above we can see that the price range of the diamond is between 448 USD (min) and 17029 USD (max).

The price varies a lot depending on the carat, color and clarity. This we can see that the std. Dev is greater than the mean for the price variable. Since the the mean is not in the arithmetic middle of the range we can assume that the price distribution is non-symmetric.

### b) Scatter plot

The scatter plot between the price and carat is given in the figure below.

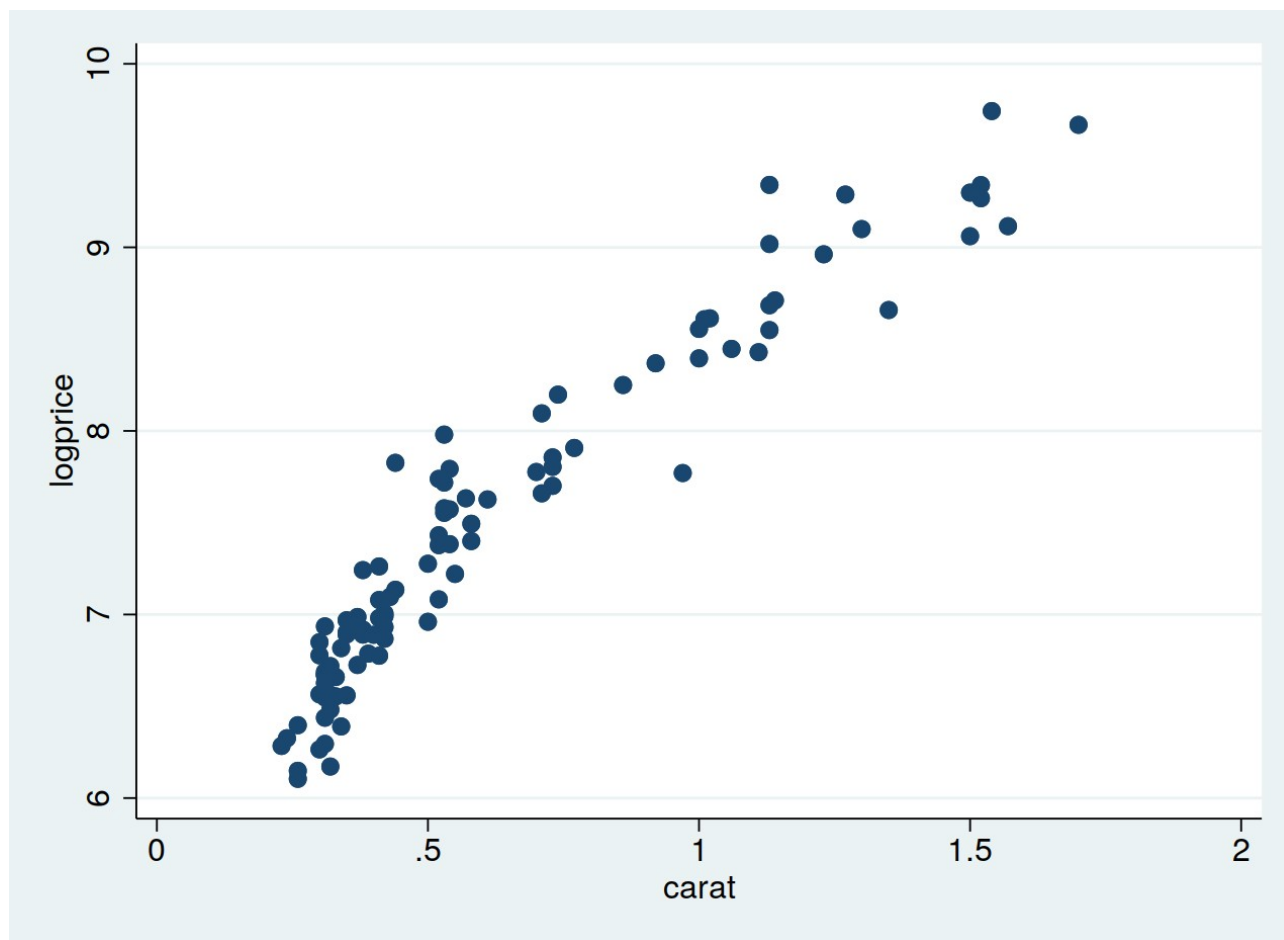
Anonymous code: 0016-DBM



We can see that the relationship is non-linear since with higher carats the price seems to increase quadratically or exponentially. Also the prices are much more together for lower carats whereas for higher carats they are more apart.

### c) Model 1

The scatter plot required in the question is given below



We can see that there is a curvature even after the log transformation on the price is applied

To analyze the residuals we have to check for the three following things:

the residuals are independent (Runs test)

the residuals are normally distributed (Jarque-Bera test)

The normal distribution has constant variance (Breusch Pagan test)

## Independence

The run tests checks whether the samples are independent from each other. The null hypothesis assumes this to be true.

Checking with Stata gives the output:

```
N(residual_m1 <= -87.03524017333984) = 50
```

```
N(residual_m1 > -87.03524017333984) = 50
```

```
obs = 100
```

```
N(runs) = 45
```

```
z = -1.21
```

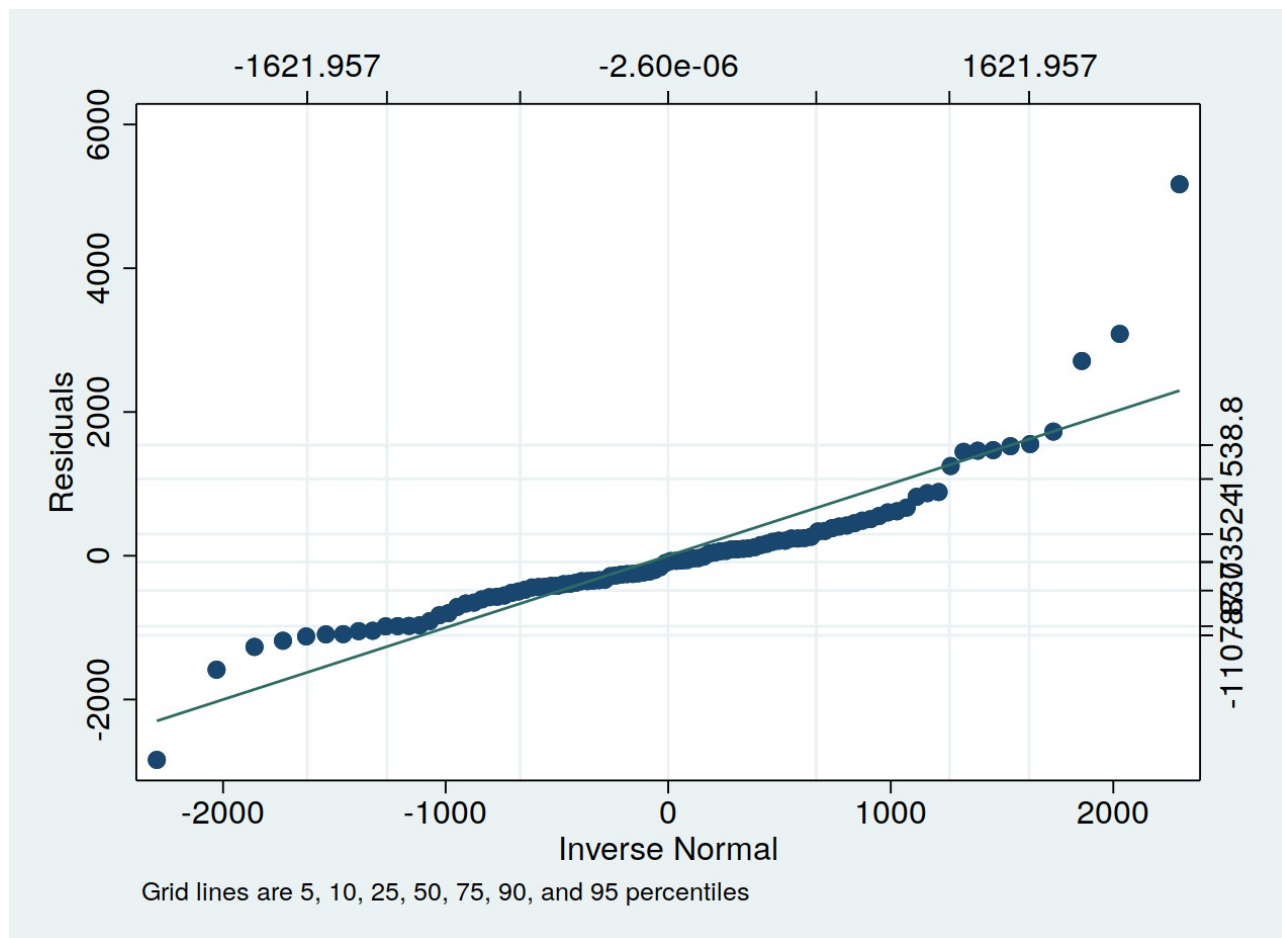
```
Prob>|z| = .23
```

Since the probability is greater than 0.05 the null hypothesis cannot be rejected at the 5% significance level, thus they are independent.

## Normality

Anonymous code: 0016-DBM

First we check the QQ-plot:



We can see that the residuals do not lie on the line.

In a more formal way we can use the Jarque-Bera test with the Stata output:

Skewness and kurtosis tests for normality

----- Joint test -----					
Variable	Obs	Pr(skewness)	Pr(kurtosis)	Adj chi2(2)	Prob>chi2
-----+-----					
residual_m1	100	0.0000	0.0000	41.52	0.0000

With  $p=0.000$  our visual interpretation of the QQ-plot is confirmed.

### Breusch Pagan test checks if the error terms have constant variance (homoscedastic)

If the null hypothesis is true then the error terms have constant variance.

H0: error term is constant

H1: error term is not constant

test variable is  $nR^2$  and it is chi-squared distributed with one degree of freedom. We use the 5% significance level.

We get from Stata:

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity  
Ho: Constant variance

Anonymous code: 0016-DBM

Variables: fitted values of price

```
chi2(1)    = 83.16
Prob > chi2 = 0.0000
```

The null hypothesis can be rejected, ie. we don't have a constant variance.

## d) Finding a better model

We could see that all assumptions apart from independence were violated for the residuals. We have to then look for ways to transform the dependent variable. We can start by taking the logarithm. We could see in the scatter plot figure in c) that it clearly isn't enough. There is a curvature that is not accommodated for in the model. Therefore we introduce a squared term and a cubed term in our regression model so that we transform it further.

We create therefore the variables `carat_cubed` and `carat_squared` that are based on the independent variable `carat` but in cubed respectively squared form.

To test our model we have to do the same procedure as for part c) residual analysis but for model 2's residuals.

We get from Stata

### Run test

Running the test in Stata gives:

```
N(residual_m2 <= -.0266800262033939) = 50
```

```
N(residual_m2 > -.0266800262033939) = 50
```

```
obs = 100
```

```
N(runs) = 45
```

```
z = -1.21
```

```
Prob>|z| = .23
```

We conclude that the residuals are independent.

## Breusch Pagan

The Stata output is the following:

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity

Ho: Constant variance

Variables: fitted values of logprice

```
chi2(1)    = 0.00
```

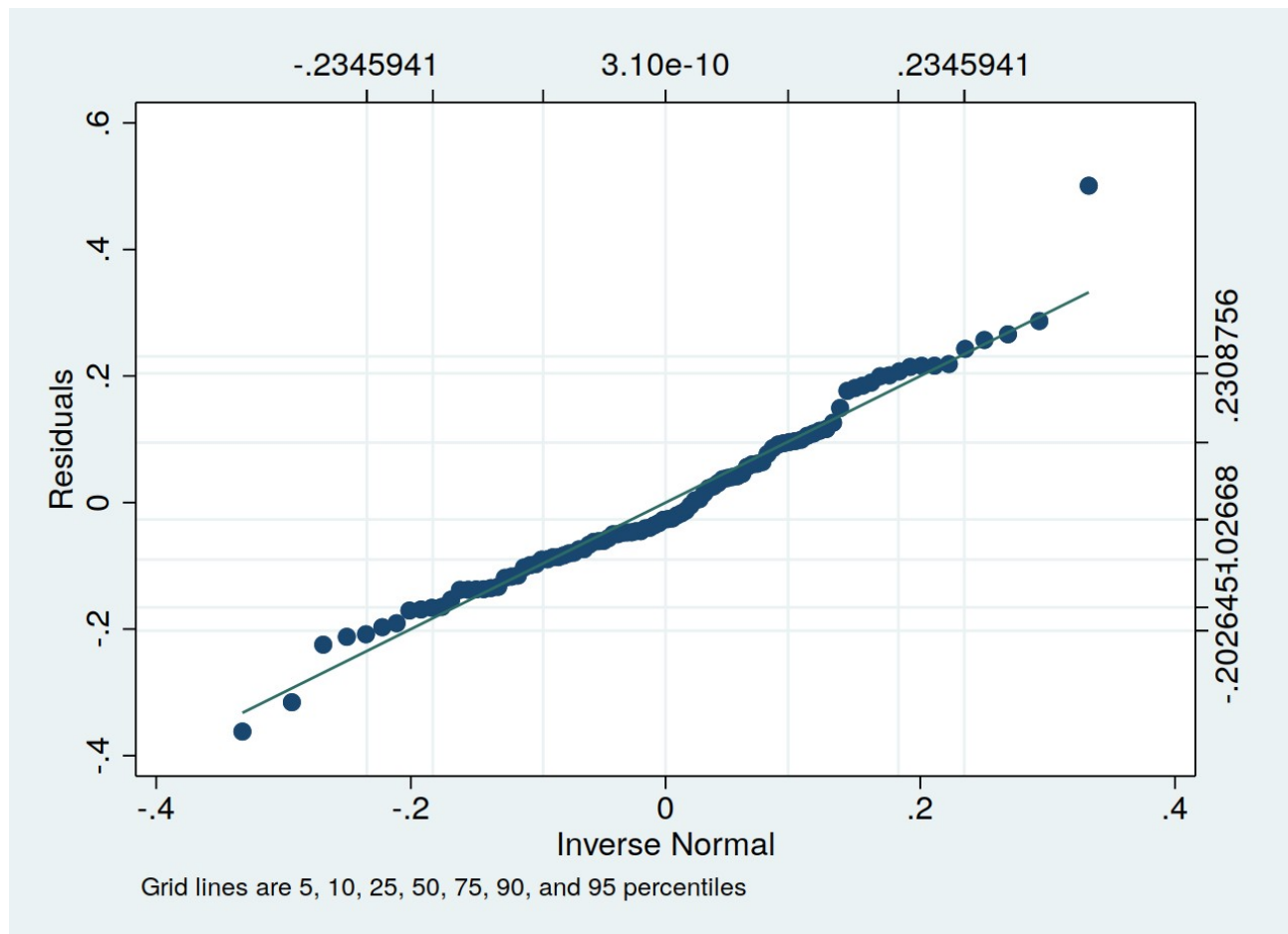
```
Prob > chi2 = 0.9800
```



Anonymous code: 0016-DBM

From the p-value we can conclude that we have constant variance.

## Jarque-Bera



The QQ-plot looks better but we still have one outlier.

Skewness and kurtosis tests for normality

----- Joint test -----

Variable	Obs	Pr(skewness)	Pr(kurtosis)	Adj chi2(2)	Prob>chi2
-----+-----					
residual_m2	100	0.0558	0.1603	5.47	0.0650

According to the test we cannot reject the null hypothesis, ie. the residuals are normally distributed

We conclude that all three requirements are fulfilled by model 2.

## e) Other tests, conclusion

We came to the conclusion that model 2 is better in part d since it fulfills all the requirements of the residuals to have a meaningful regression model. We can also look at the following values of the models: adjusted  $R^2$ , RMSE and AIC

Anonymous code: 0016-DBM

The AIC of the models are given below where the lower the value the better.

Model	N	ll(null)	ll(model)	df	AIC	BIC
model1	100	-954.6239	-830.7651	7	1675.53	1693.766
model2	100	-134.6569	53.36367	9	-88.72734	-65.28081

Model 2 wins clearly

The adjusted  $R^2$  is better for model 2 (0.9747 vs 0.9106) as well as the RMSE ( 0.14876 vs. 1017.4).