Financial Statistics Exam 20210322

Part 1: Time series

I will analyze International Flavors & Fragrances Inc. (IFF) weekly data from the past five years, from 22 mars 2016 to 22 mars 2021. I choose to analyze the adjusted close price since it takes into account the corporation's actions.

a. Data Description

Variable	Obs	Mean	Std. Dev.	Min	Max
date	0				
open	262	130.7944	10.97712	100.33	156.84
high	262	133.9893	10.54801	109.19	157.4
low	262	127.5085	11.46718	92.14	152.82
close	262	130.8987	11.01651	98.9	156.87
adjclose	262	123.7372	10.10694	96.99049	146.0707
volume	262	5054609	8061881	687200	1.02e+08

Table 1. Summarize of the IFF stock

From the summarize we can see that we will have 262 observations (or weeks) and that the lowest adjusted close price is 96.99 and the highest is 146.07 so we can assume that the time series will change quite a lot during these five years.

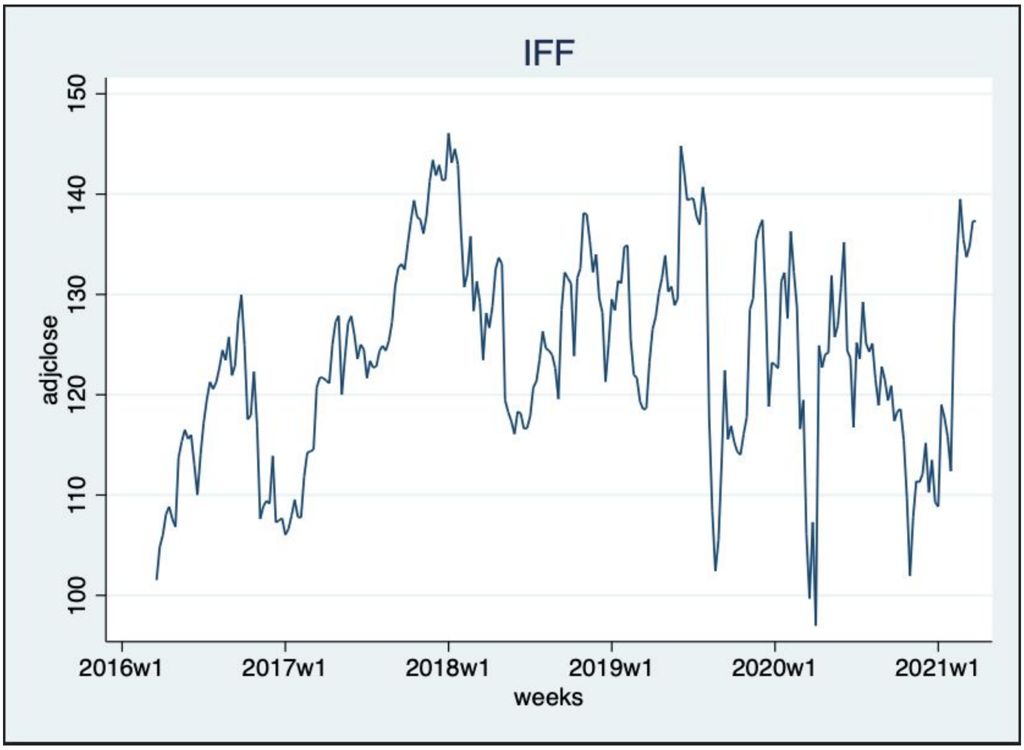


Diagram 1: Adjusted close price for IFF

From diagram 1 we can see that the stock seems to move quite a lot as assumed. The lowest adjusted price happens in the first weeks of 2020 which can be an effect of the coronavirus. At the end of each year we see an increase in the stock price and also an decrease at the end of the year (see the beginning of 2017). There is also a little upward trend and highest price of 2018 is higher than 2017. I am still assuming the corona had an effect and causes a higher difference between the highest and lowest value in 2020.

b. Stationarity

In diagram 1 we clearly can see that the time series is not stationary, so instead we try the natural log of the return.

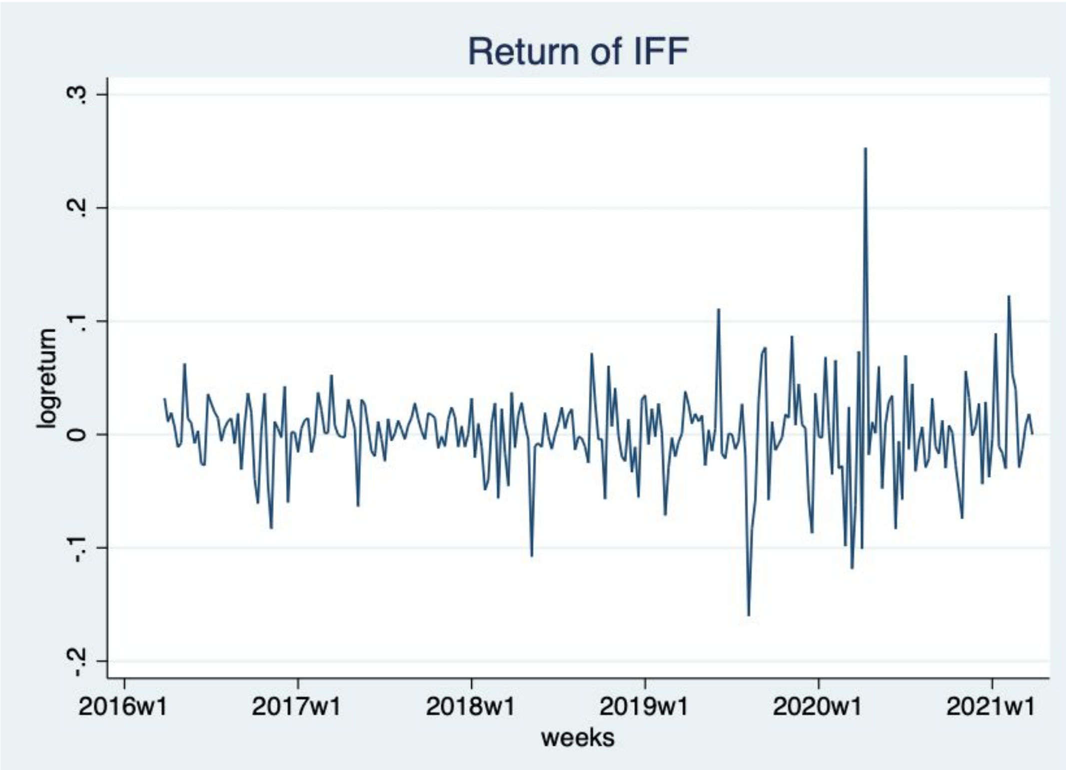


Diagram 2: Logreturn of IFF

The logged return of the IFF stock seems to be stationary from the diagram 2 compared to earlier since it fluctuates around zero. By using a Dickey fuller test to test for stationarity we can see if the data is stationary.

Start by formulating the hypothesis:

H0: The series is a random walk and therefore nonstationary

HA: The series is stationary

Significance level: 5%

Dickey-Fuller test for unit root Number of obs = 260

	Test Statistic	Interpolated Dickey-Fuller 1% Critical Value	5% Critical Value	10% Critical Value
Z(t)	-16.936	-3.459	-2.880	-2.570

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

Table 2: Output from Dickey fuller test for the logretun of IFF

We reject the null hypothesis since we have a p-value equal to zero and the test statistics -16.936< -3.459<-2.880< -2.570, so we reject the null at all significance levels.

Therefore we can conclude the logreturn of IFF is stationary.

c. ACF and PACF

The ACF and PACF plots are used to see the correlation of the data with previous values and we can use these plots to easier plot models for the time series. We need to perform the ACF and PACF plots on the logreturn since it is a stationary process. The first bar in the ACF shows if the data is correlated with the first lagged variable (previous value). ACF plot can be used to decide the moving average of an ARMA model. In our case we can not see that the first bars are significant and therefore we assume that our model is a 0 moving average process.

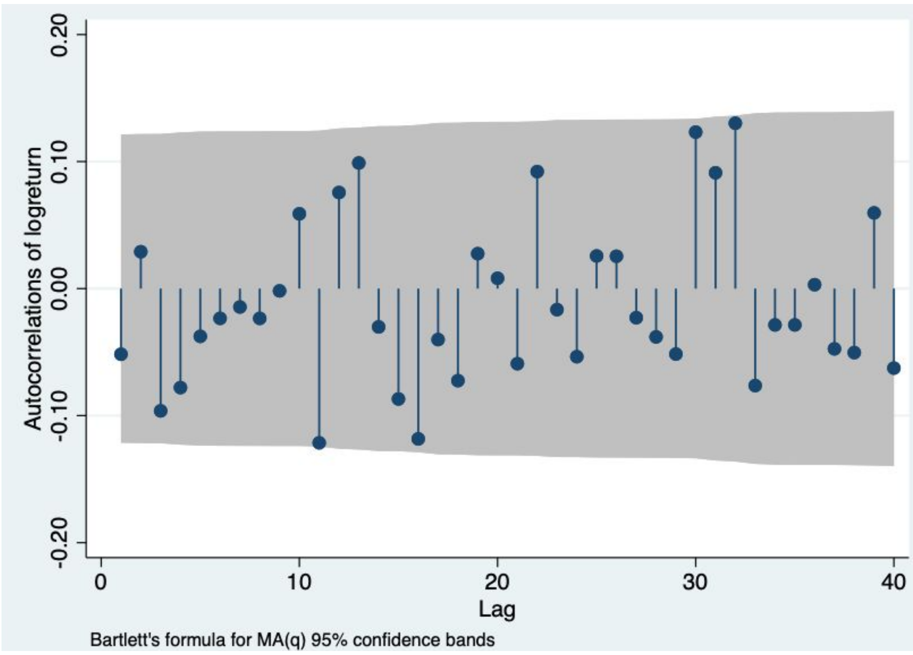


Diagram 3: ACF for the logreturn for IFF

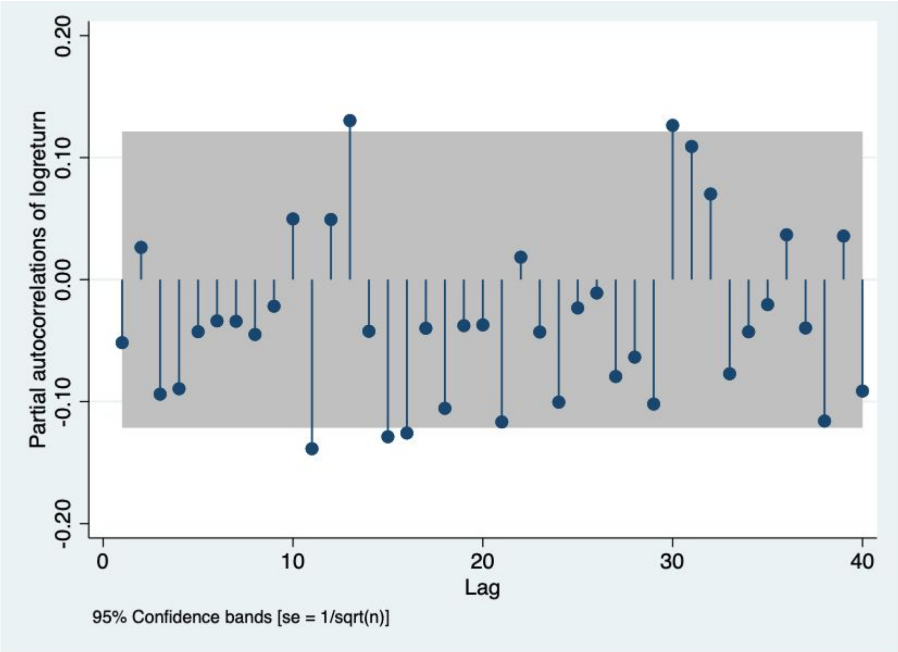


Diagram 4: **P**ACF for the logreturn for IFF

Since we cannot see a clear correlation with the first lagged variables we will test for (0,1,0) model but also (1,1,0) , (0,1,1) ,(1,1,1), (1,1,2), 2,1,1), (2,1,2).

d. ARIMA

The test will be the first difference of the logclose price since this is equivalent to the logreturn with a difference of zero. By choosing the logclose with the first difference we will estimate our model from a stationary process which we conclude in section b.

We will leave out the three last variables and perform a dynamic forecasting.

D.logclose	Coef.	OPG Std. Err.	z	P>|z|	[95% Conf.	Interval]
logclose						
_cons	.0002851	.0003402	0.84	0.402	-.0003817	.000952
ARMA						
ar						
L1.	-.053066	.1220907	-0.43	0.664	-.2923594	.1862273
L2.	.8546514	.1150229	7.43	0.000	.6292107	1.080092
ma						
L1.	-.0652855	1126.409	-0.00	1.000	-2207.786	2207.656
L2.	-.9347147	1052.813	-0.00	0.999	-2064.409	2062.54
/sigma	.0384573	21.65894	0.00	0.499	0	42.48919

Note: The test of the variance against zero is one sided, and the two-sided

Table 3. Arima (2,1,2)

D.logclose	Coef.	OPG Std. Err.	z	P>|z|	[95% Conf.	Interval]
logclose						
_cons	.0010617	.0026258	0.40	0.686	-.0040847	.006208
ARMA						
ar						
L1.	-.5041941	.8870287	-0.57	0.570	-2.242738	1.23435
L2.	.0194174	.0894549	0.22	0.828	-.1559109	.1947457
ma						
L1.	.4558774	.8748552	0.52	0.602	-1.258807	2.170562
/sigma	.0394167	.0010709	36.81	0.000	.0373178	.0415155

Note: The test of the variance against zero is one sided, and the two-sided

Table 4. Arima (2,1,1)

Sample: 2016w13 - 2021w10 Number of obs = 258
Wald chi2(2) = 1019.19
Log likelihood = 472.7874 Prob > chi2 = 0.0000

D.logclose	Coef.	OPG Std. Err.	z	P>|z|	[95% Conf.	Interval]
logclose						
_cons	.0002825	.0003293	0.86	0.391	-.0003629	.0009279
ARMA						
ar						
L1.	.8960557	.0352973	25.39	0.000	.8268741	.9652372
ma						
L1.	-.9999958	30.04451	-0.03	0.973	-59.88616	57.88616
/sigma	.0385146	.5784056	0.07	0.473	0	1.172169

Note: The test of the variance against zero is one sided, and the two-sided

Table 5: Arima (1,1,1)

Sample: 2016w13 - 2021w10 Number of obs = 258
Wald chi2(1) = 1.04
Log likelihood = 467.8671 Prob > chi2 = 0.3087

D.logclose	Coef.	OPG Std. Err.	z	P>|z|	[95% Conf.	Interval]
logclose						
_cons	.0010668	.0025447	0.42	0.675	-.0039207	.0060542
ARMA						
ar						
L1.	-.0516455	.0507339	-1.02	0.309	-.1510821	.0477912
/sigma	.0394631	.0009343	42.24	0.000	.0376319	.0412942

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

Table 6: Arima (1,1,0)

```
Sample:  2016w13 - 2021w10                      Number of obs     =        258
                                                Wald chi2(.)      =          .
Log likelihood =  467.5217                      Prob > chi2       =          .

------------------------------------------------------------------------------
             |                 OPG
  D.logclose |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
logclose     |
       _cons |    .001068   .0025032     0.43   0.670    -.0038382    .0059741
-------------+----------------------------------------------------------------
      /sigma |   .0395165   .0008206    48.15   0.000      .037908    .0411249
------------------------------------------------------------------------------
Note: The test of the variance against zero is one sided, and the two-sided
      confidence interval is truncated at zero.
```

Table 7: Arima (0,1,0)

```
ample:  2016w13 - 2021w10                       Number of obs     =        258
                                                Wald chi2(1)      =       0.93
og likelihood =  467.8513                       Prob > chi2       =     0.3347

------------------------------------------------------------------------------
             |                 OPG
  D.logclose |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ogclose      |
       _cons |   .0010651   .0025502     0.42   0.676    -.0039331    .0060634
-------------+----------------------------------------------------------------
RMA          |
          ma |
         L1. |   -.049561   .0513797    -0.96   0.335    -.1502633    .0511413
-------------+----------------------------------------------------------------
      /sigma |   .0394661   .0009202    42.89   0.000     .0376626    .0412696
------------------------------------------------------------------------------
ote: The test of the variance against zero is one sided, and the two-sided
```

Table 8: Arima (0,1,1)

By starting with the model with largest AR and MA and there after testing the next model by excluding the lag with the highest p-value and we can from the outputs already disregard ARIMA (2,1,2) (2,1,1) and (1,1,1) since the coefficients are not significant. The Arima (0,1,0) seems just from the output most promising but we can compare the AIC and BIC for the different models and by using the rule “smallest value will be the best model” we can have more evidence and thereafter decide which model is the most promising.

Akaike's information criterion and Bayesian information criterion

Model	N	ll(null)	ll(model)	df	AIC	BIC
logarima010	258	.	467.5217	2	-931.0434	-923.9375
logarima011	258	.	467.8513	3	-929.7026	-919.0437
logarima110	258	.	467.8671	3	-929.7341	-919.0752
logarima111	258	.	472.7874	4	-937.5747	-923.3629
logarima211	258	.	468.1677	5	-926.3353	-908.5705
logarima212	258	.	473.1711	6	-934.3421	-913.0244

Note: BIC uses N = number of observations. See [R] BIC note.

Table 9: Output of AIC and BIC for the estimated ARIMA models.

The ARIMA (1,1,1) has the smallest AIC and thereafter we have the ARIMA (2,1,2) but we have already excluded these models because of the coefficients not being significant. The ARIMA (0,1,0) have the smallest BIC and the smallest AIC of those models we still are considering.

The ARIMA (0,1,0) is the most promising model when considering all of our tests. The ACF plot showed a zero moving average process, the other models coefficients were not significant on a 5% level and it has the smallest BIC and AIC (only focusing on the models we have not disregarded already).

e. RMSE

The two "best models" are ARIMA (0,1,0) and (1,1,0) and by calculating the RMSE we can which model forecast better. But first we need to transform the predicted value since there are logged adjusted prices.

adjclose	real~010
134.84	135.7997
137.31	135.9448
137.31	136.0901

Table 10: Forecast against adjusted close ARIMA (0,1,0)

adjclose	real~110
134.84	135.9997
137.31	136.1453
137.31	136.2906

Table 11: Forecast against adjusted close ARIMA (1,1,0)

ARIMA 010			
Adjusted close price	forecast value	difference	sqr differenc
134,84	135,7997	-0,9597	0,92102409
137,31	135,9448	1,3652	1,86377104
137,31	136,0901	1,2199	1,48815601
MSE			1,42431705
RMSE			**1,19344755**
ARIMA 110			
Adjusted close price	**forecast value**	**difference**	**sqr differenc**
134,84	135,9997	-1,1597	1,34490409
137,31	136,1453	1,1647	1,35652609
137,31	136,2906	1,0194	1,03917636
MSE			1,24686885
RMSE			**1,11663282**

Table 12: Calculation of RMSE

The arima model (1,1,0) has a better forecast of the last observation than the (0,1,0) model but the difference in RMSE is only 0.0768. In diagram 1 we can see a clear upward going trend so it can be an advantage for ARIMA (1,1,0) since it has a lagged moving average and especially since we just tested the RMSE for the last three weeks. I still think ARIMA (0,1,0) is the model with the best fit if we take everything else into calculation .

f. GARCH

Testing for GARCH effect in stata on ARIMA (0,1,0) with the hypothesis:

H0: No arch effect

HA: There is arch effects

LM test for autoregressive conditional heteroskedasticity (ARCH)

lags(p)	chi2	df	Prob > chi2
1	116.218	1	0.0000
2	116.290	2	0.0000
3	113.812	3	0.0000
4	112.478	4	0.0000
5	113.770	5	0.0000
6	112.373	6	0.0000
7	110.703	7	0.0000
8	112.125	8	0.0000
9	113.259	9	0.0000
10	113.380	10	0.0000
11	112.952	11	0.0000
12	113.159	12	0.0000
13	113.983	13	0.0000
14	114.557	14	0.0000
15	114.171	15	0.0000

Table 13: Testing for arch effects

We test if there is an arch disturbance in 15 lags and since the p-values of all the lags is zero we reject the null and find clear evidence of arch effects. The presence of arch effects would help us pick the best model in (d) since we can by comparing the variance between the models find the model that fits my preference best. If the variance is large in a model that we think is a great estimation for our time series we maybe will find it hard to forecast our results. In financial forecasting variance is important since financial time series are affected by more things than just the previous value and therefore it is hard to estimate financial data. The variance gives us a chance to use our model based on what we can risk and also give us an answer on how much the predictions from the model can vary.

g. Residual analysis

I will continue to analyze the ARIMA (0,1,0) model and by plotting the residuals we can see if there is any strange pattern or something consering.

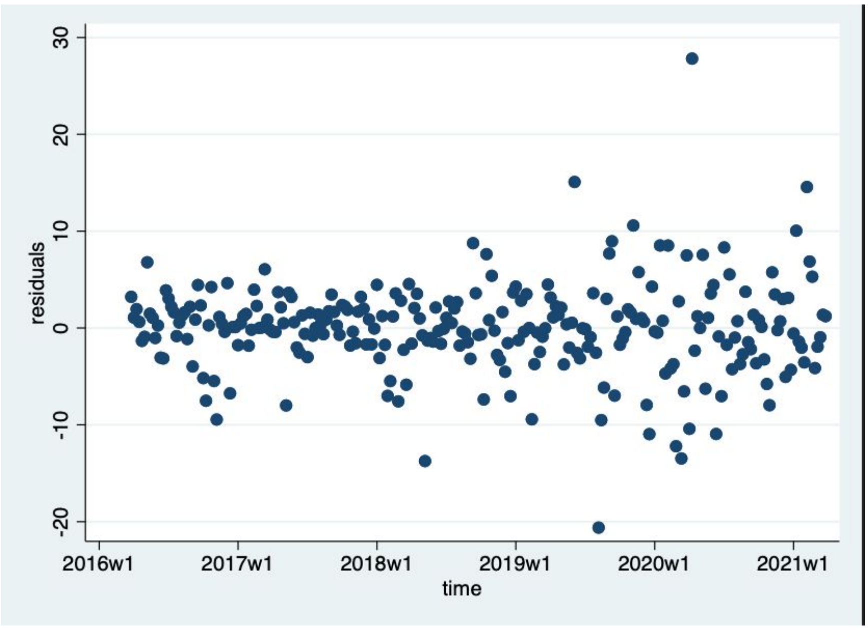


Diagram 5: Scatter of residuals over time

From the scatter plot we can see that the residuals seem to grow with time which is a sign of heteroscedasticity so we need to test for heteroscedasticity with a Breusch Pagan test.

H0: Homoscedastic error term

HA: Heteroscedastic error term

Significance level: 5%

The critical value is 3.841.

By regressing the squared residuals and the predicted values we get an R^2 of 0.0121, so the test statistics is 3.1701(262*0.0.121), we cannot reject the null since 3.1701<3.841.

<u>Stata output for the Time series part.</u>

```
/* Part 1 of exam 20210322*/
/* I will analyze the IFF stock weekly data for the past five
years*/
clear
```

```
import delimited "/Users/Johanna/Downloads/IFF.csv",
delimiter(comma)
summarize
/*Since we will just analuze the adjusted close price we will
drop the rest variables*/
drop volume high low close open
gen time=tw(2016W12)+_n-1
format %tw time
tsset time

tsline adjclose, xtitle(weeks) ytitle(adjclose) title(IFF)

/*stationary*/
gen logclose=ln(adjclose)
gen logreturn= D1.logclose
tsline logreturn, xtitle(weeks) ytitle(logreturn) title(Return
of IFF)
dfuller logreturn, lags(0)

/*ACF and PACF*/
ac logreturn
pac logreturn

/*ARIMA*/
arima logclose in 1/259, arima(2,1,2)
estimates store logarima212
predict arima212, y dynamic(tw(2021w10))

arima logclose in 1/259, arima(2,1,1)
 estimates store logarima211
predict arima211, y dynamic(tw(2021w10))

arima logclose in 1/259, arima(1,1,1)
estimates store logarima111
predict arima111, y dynamic(tw(2021w10))

arima logclose in 1/259, arima(1,1,0)
estimates store logarima110
predict arima110, y dynamic(tw(2021w10))

arima logclose in 1/259, arima(0,1,0)
estimates store logarima010
predict arima010, y dynamic(tw(2021w10))
```

```
arima logclose in 1/259, arima(0,1,1)
estimates store logarima011
predict arima011, y dynamic(tw(2021w10))

estimates stats logarima010 logarima011 logarima110
logarima111 logarima211 logarima212

gen realarima010=exp(arima010)
gen realarima110=exp(arima110)

/*RMSE*/
list adjclose realarima010 in 260/262
list adjclose realarima110 in 260/262

/*GARCH*/
regress arima010
estat archlm, lags(1/15)

/*Residual analysis*/
gen residuals= adjclose-realarima010
scatter residuals time

gen residuals_sqr=residuals^2
regress residuals_sqr realarima010
```

Part 2. Regression.

Price: estimated in USD
Carat: Weights in carats (1 carat =200 mg)
Color_def: Dummy variable. 1= belongs to category d,e,f
Color_gh: Dummy variable. 1= belongs to category g or h
clarity_if: Dummy variable . 1=internally flawless
Clarity_vvs: Dummy variable. 1= Very very slightly included
Clarity _vs: Dummy variable. 1=very slightly included

a) **Summarize**

Variable	Obs	Mean	Std. Dev.	Min	Max
price	50	3519.14	3700.692	479	15841
carat	50	.7246	.4276142	.3	2.01
color_def	50	.42	.4985694	0	1
color_gh	50	.34	.4785181	0	1
clarity_if	50	.06	.2398979	0	1
clarity_vs	50	.36	.4848732	0	1
clarity_vvs	50	.26	.4430875	0	1

Table 1: Summarize of diamonds

The price range of the round cut diamonds is between 479 and 15 841 USD.

b. Price compared to carat

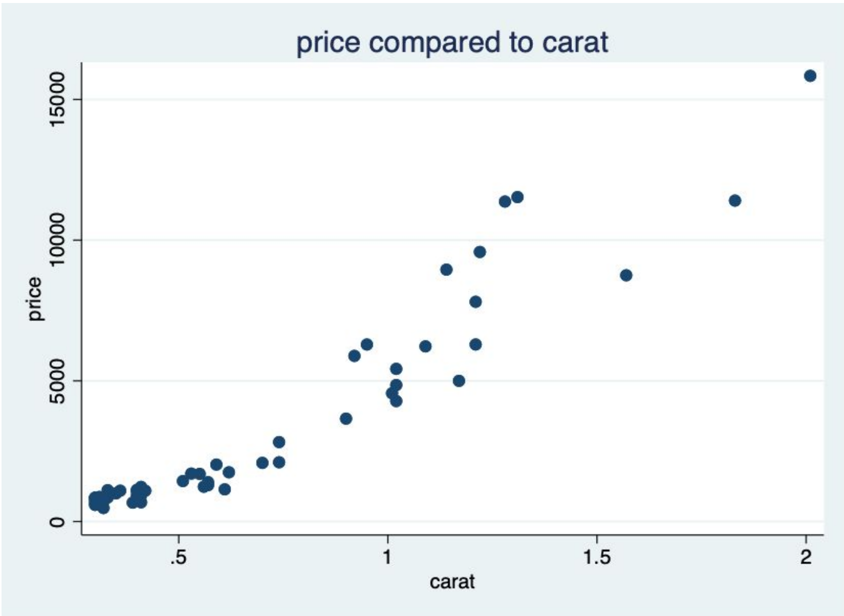


Diagram 1. Scatterplot of Price compared to carat

We can see a linear relationship between the weight of the diamond and the price of the diamond. A higher carat indicates a higher price and we can see that the highest price is reached when we have the highest carat in the sample. But we also need to consider that this is a small sample and most of the observations seem to be on a lower carat.

c. Regression of natural log of price

Source	SS	df	MS			
				Number of obs	=	50
				F(6, 43)	=	89.75
Model	44.5328789	6	7.42214649	Prob > F	=	0.0000
Residual	3.55621108	43	.082702583	R-squared	=	0.9260
				Adj R-squared	=	0.9157
Total	48.08909	49	.98141	Root MSE	=	.28758

logprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
carat	2.378479	.1097621	21.67	0.000	2.157122	2.599835
color_def	.2506005	.1132199	2.21	0.032	.0222708	.4789302
color_gh	-.0983266	.122126	-0.81	0.425	-.3446172	.1479641
clarity_if	.5594052	.1999006	2.80	0.008	.1562672	.9625433
clarity_vvs	.3426804	.123849	2.77	0.008	.0929151	.5924457
clarity_vs	.0846128	.1038431	0.81	0.420	-.1248068	.2940323
_cons	5.722215	.1614332	35.45	0.000	5.396654	6.047776

Table 2: Regression with Natural log as dependent variable

The regression model has a high R^2 of 0.92 but this is not a good test, we can also see that the clarity_vs and color_gh are not significant at a 5% model, their confidence intervals contain zero so these should be taken out from the model. The RMSE is 0.28758.

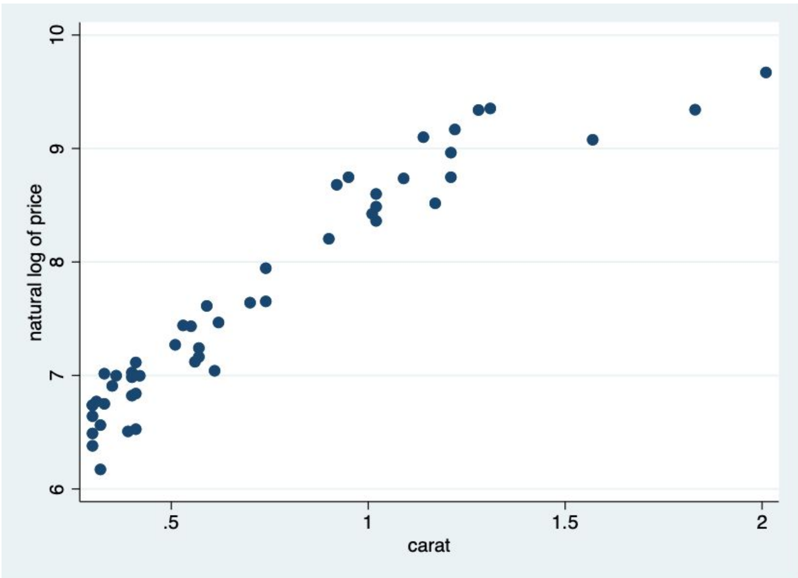


Diagram 2: Scatter of natural log compared to carat

We can still see a linear relationship between the carat and price (in this case natural log of price) but now it is easier that there is a larger change in price as the weight increases. On the lower weights we can now also see the different observations and their prices.

d. Residuals model 1

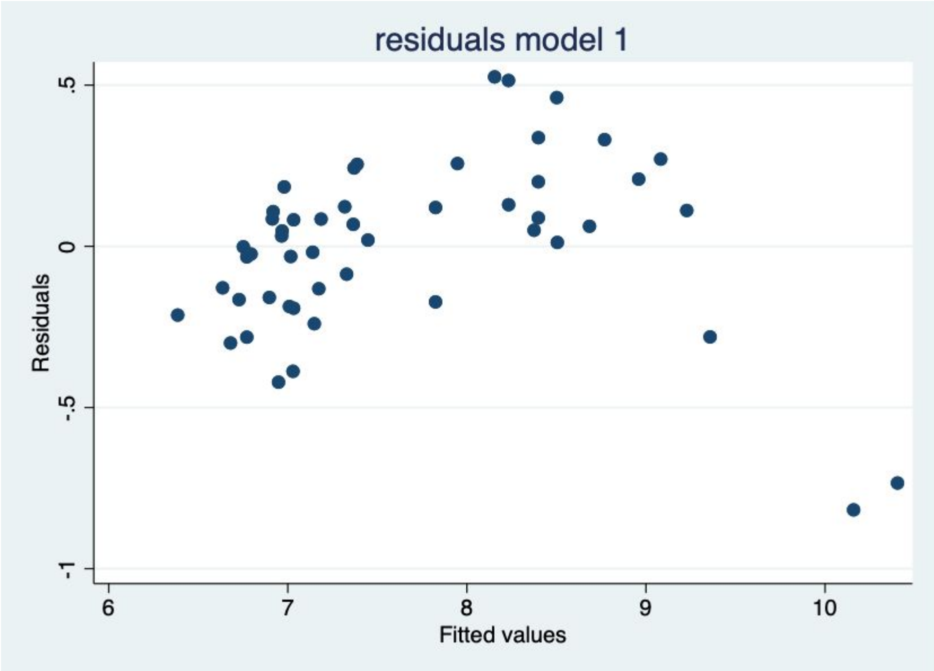


Diagram 3: Residuals model 1

From the scatterplot we can see that there is a pattern between the residuals and we can therefore assume that there are heteroscedastic this because the residuals seems to be larger with the fitted values.

e. Regression Model 2

Now we will estimate another regression model, so we add two variables (carat squared and carat cubed) .

Source	SS	df	MS			
				Number of obs	=	50
				F(8, 41)	=	274.63
Model	47.2081226	8	5.90101533	Prob > F	=	0.0000
Residual	.880967412	41	.02148701	R-squared	=	0.9817
				Adj R-squared	=	0.9781
Total	48.08909	49	.98141	Root MSE	=	.14658

logprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
carat	5.544972	.8349597	6.64	0.000	3.858737	7.231208
color_def	.3655274	.0598745	6.10	0.000	.2446083	.4864464
color_gh	.1785186	.0687702	2.60	0.013	.0396344	.3174028
clarity_if	.6063287	.1065493	5.69	0.000	.391148	.8215094
clarity_vvs	.387057	.0644202	6.01	0.000	.2569578	.5171562
clarity_vs	.214462	.0564176	3.80	0.000	.1005244	.3283997
carat_cub	.2933828	.2589025	1.13	0.264	-.2294813	.816247
carat_sqr	-2.139552	.8623078	-2.48	0.017	-3.881018	-.3980865
_cons	4.481534	.2561306	17.50	0.000	3.964268	4.9988

Table 3: Regression model 2.

We can see that all coefficients except the carat_cub is significant at a 5% level and that the RMSE now is 0.14658< 0.28758 (model 1)

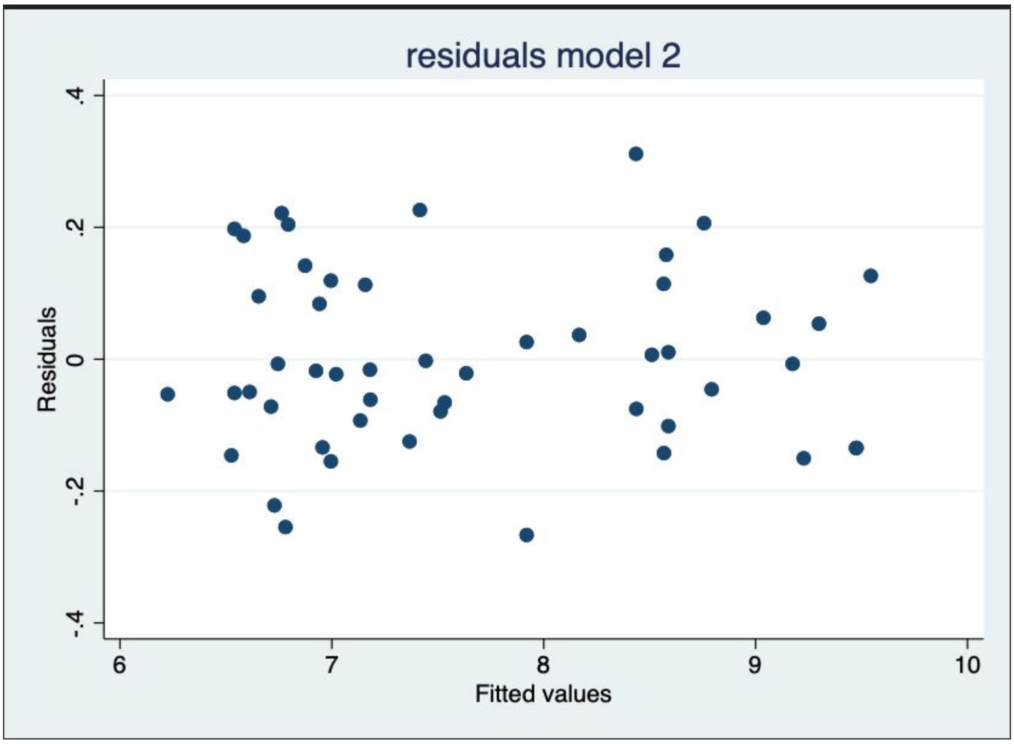


Diagram 4: Residuals model 2

Now the residuals seem to be the same and not change with the predicted value.

f. Breusch Pagan test model 2

Doing a Breusch Pagan test for model 2 to test if the error term is heteroscedastic or not.

H0: Homoscedastic error term

HA: Heteroscedastic error term.

Significance level of 5%

Then we have a critical value of 3.841.

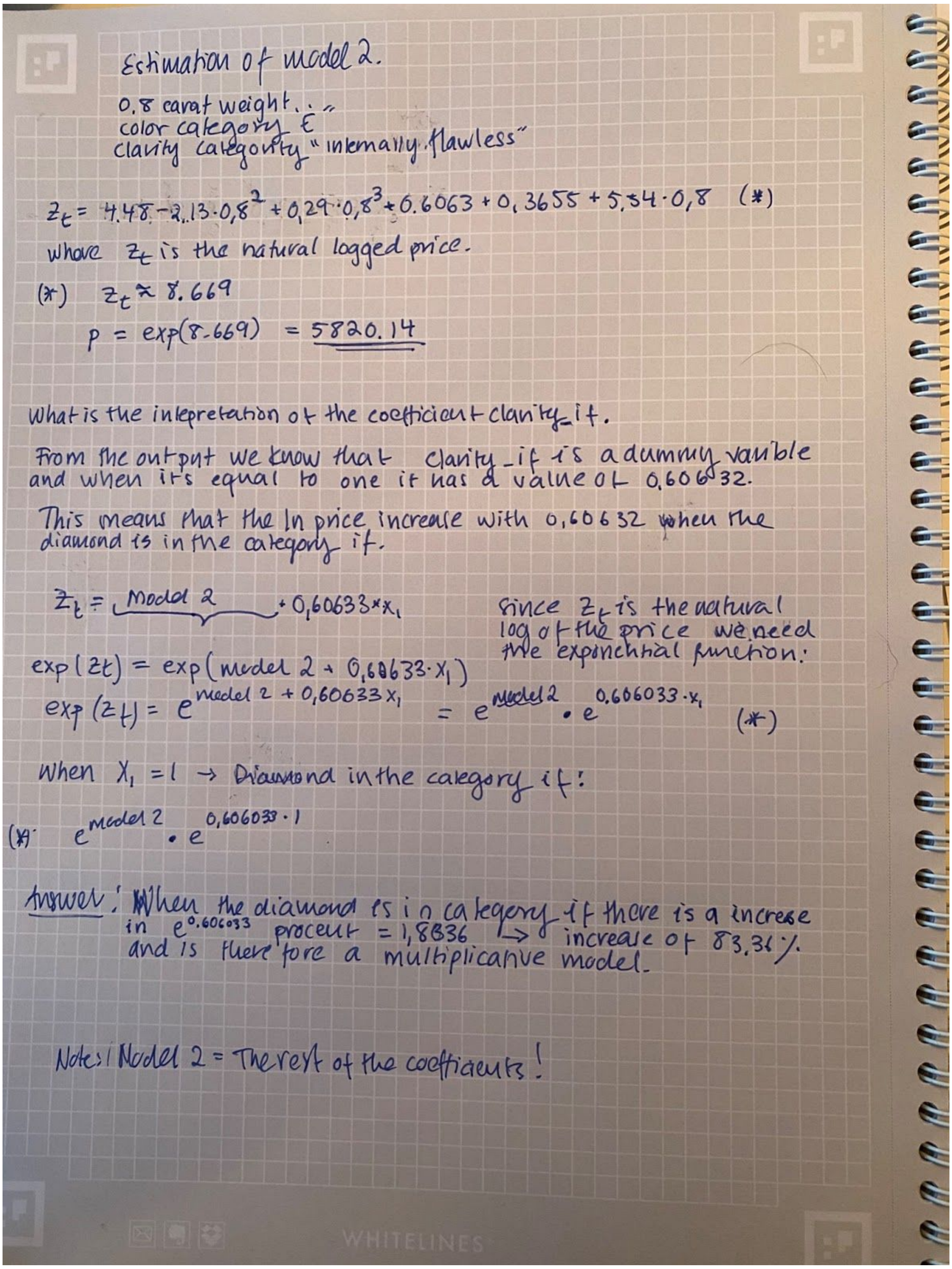
```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of logprice

         chi2(1)      =     0.12
         Prob > chi2  =   0.7254
```

Table 4: Breusch Pagan test for model 2

Since the p-value is larger than 0,05 and the test statistics 0.12< 3.841 we cannot reject the null and therefore the error term is homoscedastic for model 2.

g. Estimation

Here is my calculations:



Estimation of model 2.

0.8 carat weight . . "
color category E "
clarity categority "internally flawless"

$$z_t = 4.48 - 2.13 \cdot 0{,}8^2 + 0{,}29 \cdot 0{,}8^3 + 6.6063 + 0{,}3655 + 5{,}34 \cdot 0{,}8 \quad (*)$$

where z_t is the natural logged price.

$(*)$ $z_t \approx 8.669$

$$p = \exp(8.669) = \underline{\underline{5820.14}}$$

What is the interpretation of the coefficient clarity_if.

From the output we know that clarity_if is a dummy varible and when it's equal to one it has a value of 0,60632.

This means that the ln price increase with 0,60632 when the diamond is in the category if.

$$z_t = \underbrace{\text{Model 2}} + 0{,}60633 * x_1$$

Since z_t is the natural log of the price we need the exponential function:

$$\exp(zt) = \exp(\text{model } 2 + 0{,}60633 \cdot x_1)$$

$$\exp(zt) = e^{\text{model } 2 + 0{,}60633 x_1} = e^{\text{model } 2} \cdot e^{0.606033 \cdot x_1} \quad (*)$$

When $x_1 = 1 \rightarrow$ Diamond in the category if:

$$(*) \quad e^{\text{model } 2} \cdot e^{0{,}606033 \cdot 1}$$

Answer: When the diamond is in category if there is a increse in $e^{0.606033}$ procent $= 1{,}8336 \rightarrow$ increase of 83,36 %. and is therefore a multiplicative model.

Note: Model 2 = The rest of the coefficients!

Stata output for regression part :

```
/*regression part of exam*/
clear
 use
"/Users/Johanna/Library/Containers/com.apple.mail/Data/Library
/Mail
Downloads/99BBB6AB-EA8B-40A6-B40D-388B5F463C16/diamonds.dta"
summarize
scatter price carat, title( price compared to carat)

/* regression on natural log of price*/
gen logprice=ln(price)
regress logprice carat color_def color_gh clarity_if
clarity_vvs clarity_vs
scatter logprice carat, xtitle(carat) ytitle(natural log of
price)

predict price_hat_m1
predict residual_m1,residual
scatter residual_m1 price_hat_m1,title(residuals model 1)

/*regression model 2*/
gen carat_sqr=carat^2
gen carat_cub=carat^3

regress logprice carat color_def color_gh clarity_if
clarity_vvs clarity_vs carat_cub carat_sqr

predict price_hat_m2
predict residual_m2,residual
scatter residual_m2 price_hat_m2,title(residuals model 2)

hettest
```