Nova School of Business and Economics

NOVA SCHOOL OF
BUSINESS & ECONOMICS

EMPIRICAL METHODS IN FINANCE

First Semester, T1

2020/21

First Graded Assignment
Analysts' Earnings Forecasts

Dornigg, Thomas (41727)
Hinterhölz, Josef (45894)
Rupp, Sebastian (46093)

September 26, 2020

Contents

List of Tables

List of Figures

The following tasks in these four sections were performed with the programming language Python. The Jupyter-Notebook file and all other codes used for this assignment can be found in this Github repository.

Note: The group decided to convert the given data-files *ibes_2019.dta* and *crsp_daily_2019.dta* to csv-files in order to reduce the memory of the dataframes. Both files and the code can be found in the repository too.

1 Basic data manipulation

a) In STATA, open *ibes_2019.dta*. Briefly describe the structure of the data. What is the smallest unit of observation?

Table 1 here

Shown in Table 1 above, one can see that the dataset consists of 157.932 rows and five columns in total (note: the two date columns have been dropped in Python automatically since they are qualitative). The descriptive statistic shows the min, max, std, 25%, 50%, 75% and mean value of each feature of the dataset.

The smallest observation unit is each analyst's estimate (meaning each row) since multiple analysts provide their view on the same earnings announcements.

b) Rename *anndats_act* as date, then merge (m:1) *ibes_2019.dta* to (using) *crsp_daily_2019.dta*. Keep only the observations that are successfully merged. Compute companies' market capitalization (mktcap).

After renaming the mentioned column, we performed an inner-join on the columns *dates* and *permno* with pandas to combine both dataframes to a single one. Also, we computed descriptive statistics of the newly created dataset, as shown in Table 2 below:

Table 2 here

c) Compute the consensus forecast (consensus), defined as the median forecast across analysists, and the standard deviation of the forecasts (dispersion).

Please refer to the Jupyter Notebook file.

d) Generate a variable coverage equal to the number of analysts providing a forecast for a given earnings announcement.

Please refer to the Jupyter Notebook file.

e) Drop the following industries from the dataset: international affairs and non-op. establishments (SIC 9000-9999), foreign governments (SIC 8888), utilities (SIC 4000-4999) and agricolture, fishing and hunting (SIC 0000-0999). Create a dummy variable financials for financial firms (SIC 6000-6999).

Below, a summary shows how many rows were dropped in each iteration:

- For SICs between 9000 and 9999: 20.267 rows
- For the SIC 8888: 0 rows
- For SICs between 4000 and 4999: 14.557 rows
- For SICs between 0000 and 0999: 228 rows

For the dummy computation, we applied following scheme:

$$D = \begin{cases} 1, & \text{if } x = \text{financial firm} \\ 0, & \text{non-financial firm} \end{cases}$$

2 Summary Statistics and Plots

a) How many distinct earnings announcements events are in the data?

The total number of distinct earnings announcements per stock was counted to obtain the total number of earnings events in the data, which amounted to *2.714* events.

b) Collapse the data at earnings announcement level and keep the mean of fe, fd, coverage, mktcap, financials.

Please refer to the Jupyter Notebook file.

c) Produce a summary statistics table with the mean, standard deviation, min and max of all the variables in the dataset. Include also a correlation matrix between all the variables.

For answering all upcoming questions, the dataset generated in question 2b) was used. Below, a summary statistic was computed, depicting the row-count, mean, std, min, 25%, 50%, 75% and max value for each feature.

Table 3 here

From the correlation heatmap (Figure 1) below, one can see the pairwise correlations for each feature of the dataset. The darker the color, the more positive the features are correlated; the lighter the color, the weaker the correlation. The diagonal of the heatmap shows the correlation of each feature with itself.

Figure 1 here

In general, correlation is a measure of the strength of a relationship between two variables and is expressed numerically by the correlation coefficient, whose range is between -1 and 1. A perfect positive correlation of 1 implies that if one feature goes up, the other one follows in the same direction. On the other hand, a perfect negative correlation implies movement in the opposite direction and a correlation coefficient of 0 implies no relationship at all.

From Figure 1 above, one can see strong positive correlations for the pairs fe & fd, coverage & mktcap and slight negative, respectively, almost zero correlation between the rest of the features.

d) Do a scatter plot of fe against coverage, and of fd against coverage. Label the axes in a meaningful way. Briefly comment on the two charts.

Figure 2 here

Interpretation: scatter plot of coverage vs. fe
The scatter plot indicates a relationship between the coverage of a stock and the relative forecasting error. The graph suggests that the higher the number of analysts covering a given stock, the lower fe. Therefore, the closer analysts' earnings forecast lies to any given company's actual earnings in the data set.

Interpretation: plot of coverage vs. fd
Similarly, the plot of fd against coverage shows that the higher the number of analysts covering a stock, the lower the dispersion of their revenue estimates in relation to the companies' actual earnings. It is inferrable that the more analysts cover a stock, the lower the dispersion of their estimates, which indicates that the analysts' forecasts converge as the number of forecasts increases.

Combining both interpretations, one can assume that the higher the coverage, the more precise the predictions, and the less variance is inhibited in analysts' forecasts.

3 OLS: estimation and interpretation of the results

a) Run a regression of fe on financials and then of fe on financials and coverage. Interpret the coefficients and the R^2 of both regressions.

First, we computed the following OLS regression,

$$fe = \beta_0 + \delta_0 financials + \epsilon \tag{1}$$

whose output is depicted in Figure 3 below. Since the computation was performed with the Python library *statsmodels*, it is important to mention that the baseline model are all non-financial firms while the financial firms are depicted in the OLS output [T.1].

Figure 3 here

From the output, one can see the intercept β_0 = 0.2021 and the dummy-variable estimator δ_0 = -0.0291. The R^2 of the model is 0.002. By definition, the R^2 is a statistical measure that represents the proportion of the variance for a dependent variable that is explained by independent variables [5]. Since the R^2 of the model is 0.002, almost none of the observed variation can be explained by the model's inputs.

The other coefficients can be interpreted as follows:

- β_0: can be interpreted as the value one would predict if the dummy variable $financials = 0$ (baseline model)
- δ_0: represents the additional value in the intercept ($\beta_0 + \delta_0$), if we control for financial firms, i.e. $financials = 1$

Next, we computed the following OLS regression,

$$fe = \beta_0 + \delta_0 financials + \beta_1 coverage + \epsilon \tag{10}$$

whose output is depicted in Figure 4 below. Again, the baseline model are all non-financial firms while the financial firms are depicted in the OLS output.

Figure 4 here

From the output in Figure 4, one can see the intercept β_0 = 0.2784, δ_0 = -0.0512 and β_1 = -0.0070. The R^2 of the model slightly increased to 0.040. That means that approximately 4.0% of the observed variation can be explained by the model's inputs which might indicates that the additional regressor increased the model performance.

The other coefficients can be interpreted as follows:

- β_0: can be interpreted as the value one would predict if the dummy variable $financials$ = 0 and the independent variable $coverage$ = 0 (baseline model)
- δ_0: represents the additional value in the intercept (β_0 + δ_0), if we control for financial firms, i.e. $financials$ = 1
- β_1: represents the difference in the predicted value of Y_i in comparison to the baseline model if all other variables are held constant (ceteris paribus). In other words, if *coverage* is increased by one unit, all else equal (ceteris paribus), one would expect Y_i to decrease approximately by 0.0070 units on average

b) Re-run the above regression, where you additionally control for mktcap. How does the interpretation of the coefficient on coverage change?

We computed the following OLS regression,

$$fe = \beta_0 + \delta_0 financials + \beta_1 coverage + \beta_2 mktcap + \epsilon \tag{25}$$

whose output is depicted in Figure 5 below. Again, the baseline model are all non-financial firms while the financial firms are depicted in the OLS output.

Figure 5 here

From the output in Figure 5, one can see the intercept β_0 = 0.2766, δ_0 = -0.0504, β_1 = -0.0068 and $\beta_2 \sim$ -0.0. The R^2 of the model is unchanged 0.040. Again, that means that approximately 4.0% of the observed variation can be explained by the model's inputs.

In comparison to the last model in 2a), there is no change in the interpretation of the variable *coverage*. If one increases coverage by one unit, all else equal (ceteris paribus), we would expect, on average, an decrease of the dependent variable fe of approx. 0.0068 units.

c) Units of measurement

i. Scale the variable mktcap such that it gives the company market capitalization in billions of dollars. Re-run the last regression. How does the coefficient change? And the t-statistic? Explain.

ii. Suppose that you are allowed to report only two decimals in your tables. Do you see a problem? What would you do?

After scaling the variable mktcap by a factor of 1.000.000 (i.e. dividing the feature by 1.000.000) we ran the following regression

$$fe = \beta_0 + \delta_0 financials + \beta_1 coverage + \beta_2 (\frac{mktcap}{1.000.000}) + \epsilon \tag{32}$$

and received the OLS output as described in Figure 6.

Figure 6 here

Comparing the outputs of Figure 5 and Figure 6, neither did the coefficients nor the t-statistics change with exception of the scaled estimator β_2, which changed from approx. -0.0 to approx. -0.0966 (the change of the coefficient is not a real change, just a shift of the decimal places).
For both regressions, all estimators, with exception of β_2 are statistically significant for conventional thresholds (10%, 5% and 1%). The reason why the t-statistic does not change when linear transformation is applied is because, for each variable, the interpretation of the estimators and the standard error have all been scaled by the same constant, and this constant, in the above case 1.000.000 cancels out when computing the t-statistic [7].

In general, it can be noted that linear transformations do not affect the fit of a classical regression model and they do not affect predictions. However, well-chosen linear transformation can improve interpretability of coefficients an make a fitted model easier to understand [3].

Regarding the second part of the question, it appears that the topic of how many digits after the decimal point should be reported in statistical outputs is a controversial discussion [4]. For instance, when a certain model is executed and the parameter estimates are only rounded to three decimal points, it could be the case that this constitutes an insufficient precision for the predictors. When the scale of the predictor is large, the coefficient may be very small and yet significant. However, a coefficient of 0.000122, for example, will be rounded to 0.000 in the model output [6]. A possible solution to this issue is to change the display settings for regression estimates manually every time before an output is generated.

d) Under which assumptions can we interpret the coefficient on coverage as measuring the causal effect of analysts' coverage on forecast accuracy? What threats to the identification of the causal effect do you see in this case?

The causal effect of coverage on fe can be interpreted under the following assumptions, which are equal to the assumptions underlying the classical linear regression model:

1. The causal relationship between fe and coverage can be described as a linear function.
2. The data used to conduct this paper is based on a random sample.
3. The existance of variance, in this particular case the number of analysts covering a certain stock, varies across the sample.
4. The number of analysts covering a stock is independent of the unobservable factor ($\hat{u}_i$).

MLR. 1: Linearity in parameter	Not fulfilled, because of the non-linear relationship between fe and coverage which can be observed in Figure 2.
MLR. 2: Random Sampling	This condition is fulfilled.
MLR. 3: No perfect collinearity	This condition is fulfilled, because the correlations between coverage and market capitalization lies at 0.4 and is thus imperfect.
MLR. 4: Zero conditional mean	This is not fulfilled due to various omitted variables.

The breach of the MLR.1 assumption threatens the identification of cause and effect relation of *coverage* on fe. The plot in Figure 2 shows no exact linear relationship between fe and analysts coverage of stocks, which does not allow a clear statement regarding cause and effect if no non-linear transformation of *coverage* is made to reinstate MLR. 1.

Since MLR. 4 is not fulfilled, there exists the threat that one might interchange mere correlation with cause and effect due to the fact that an omitted variable is driving the market capitalization and the coverage of stocks. One example interesting to examine in this instance could be the revenue growth rates of the respective companies. One could assume that high growth rates make stocks more appealing to analysts, causing more analysts to cover the stock. Moreover, high growth rates and the narrative of growth values could drive the respective stocks' prices up, which would increase their market capitalization.

Additionally, it is a threat that the functional relationship between the regressand and the regressors might be misspecified, which can be observed in the scatter plot of Figure 2 [9]. All in all, the breaches of MLR. 1 and MLR. 4 threaten the identification of the causal relationship between fe and $coverage$.

4 OLS: assumptions and mechanics

a) After running the previous regression of fe on financials, coverage and mktcap, compute

$$\sum_{i=1}^{N} \hat{u}_i \ (\hat{y}_i \text{ - } \overline{y}).$$

What is the result of the summation? Explain.

For the computation, please refer to the Jupyter Notebook file.

When executing the above summation, the result is approximately 0, which is also in line with the algebraic construction by design. This can also be mathematically derived in the following way [10]:

1. The sum, and therefore the sample average of the residuals is zero
$$\sum_{i=1}^{N}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = \sum_{i=1}^{N} \hat{u}_i = 0$$
2. The sample covariance between the regressors X_i and the OLS residuals $\hat{u}_i$ is zero, which can be expressed by
$$\sum_{i=1}^{N}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)X_i = \sum_{i=1}^{N} X_i \hat{u}_i = 0$$
3. From the first equation of the First Order Condition can also be concluded that the point $(\overline{X}, \overline{Y})$ always lies on the sample regression line.
4. Another import property to note is that
$$\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \sum_{i=1}^{N}(\hat{Y}_i - \overline{\hat{Y}})^2 + \sum_{i=1}^{N} \hat{u}_i^2$$
which, if expressed in words can be described as SST = SSE + SSR or Sum of Squared Total = Sum of Squared Explained + Sum of Squared Residuals.

   The equation in point 4 holds, if we show that
$$\sum_{i=1}^{N} \hat{u}_i \ (\hat{y}_i \text{ - } \overline{y}) = 0.$$

   Thus, by using the formula above, we can derive a measure of goodness of fit from the fact the SST = SSE + SSR [8].

b) [Theory Question] Consider the following population model

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 x_i + \epsilon \tag{44}$$

where D_i is a dummy variable = 1 if i is a financial firm. What would be the consequence of estimating the model without D_i?

In general, dummy-variable regressors are used to incorporate qualitative explanatory variables into a linear model, substantially expanding the range of application of regression analysis [1]. The basic motivation for including a qualitative explanatory variable is the same as for including an additional quantitative explanatory variable:

- avoiding a biased assessment as a result of omitted variables
- to account more fully for the response variable, making the errors smaller and the model better

Moreover, the only thing that differs when a dummy-variable is included in the regression is that the conditional expectation, when $D_i = 1$, the population model consist of two constants β_0 and β_1 which in sum represents the intercept in that case. However, when $D_i = 0$, the conditional expectation will be given only by the constant β_0 [2]. Note, additionally, that adding a dummy-variable does not change the marginal effect of β_2 (slope) on the dependent variable Y_i.

For our population model above, this would mean that if no dummy variable is included in the regression, each firm, independent if it is a financial or non-financial firm would have the same intercept. Consequently, when a regression forecast is performed without including a dummy variable where the dependent variable represents the price forecast, the price will be either over- or underestimated on average since no distinction between financial and non-financial firms is made.

References

[1] *Econometrics.* https://ebrary.net/1013/economics/dummy_variables. Accessed on 2020-09-20.

[2] John Fox. *Lecture Notes: Dummy Variable Regression and Analysis of Variance.* https://socialsciences.mcmaster.ca/jfox/Courses/soc740/lecture-5-notes.pdf. Accessed on 2020-09-20. 2014.

[3] Andrew Gelman and Jennifer Hill. *Data Analysis using Regression and Multilevel/Hierarchical Models.* http://vulstats.ucsd.edu/pdf/Gelman.ch-04.regression-transformations.pdf. Accessed on 2020-09-20. Dec. 2006.

[4] Farrokh Habibzadeh and Parham Habibzadeh. *How much precision in reporting statistics is enough?* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4679338/. Accessed on 2020-09-20. Oct. 2015.

[5] Adam Hayes. *R-Squared Definition.* https://www.investopedia.com/terms/r/r-squared.asp. Accessed on 2020-09-20. Mar. 2020.

[6] IBM. *More decimal precision needed in Advanced model output.* https://www.ibm.com/support/pages/more-decimal-precision-needed-advanced-model-output. Accessed on 2020-09-20. June 2018.

[7] Ben van Kammen. *Multiple Regression Analysis: Further Issues.* https://rb.gy/k9apc6. Accessed on 2020-09-20.

[8] Anthony Tay. *Intermediate Econometrics / Forecasting Class Notes.* shorturl.at/blqF2. Accessed on 2020-09-20. May 2018.

[9] Jeffrey M. Wooldridge. *Introductory Econometrics - A modern approach.* Fifth Edition. Cengage Learning, 2013, 86ff.

[10] Jeffrey M. Wooldridge. *Introductory Econometrics - A modern approach.* Sixth Edition. Cengage Learning, 2016, pp. 33–34.

Appendices

A Inserted Tables

	estimator	analys	value	actual	permno
count	157932	157932	157932	157932	157932
mean	1229.52	110569.19	3.40	3.42	52751.97
std	1317.16	48965.96	4.93	5.22	32979.68
min	11.00	0.00	-9.88	-13.63	10026.00
25%	171.00	77712.00	0.83	0.86	15642.00
50%	464.00	114862.50	2.44	2.45	60943.00
75%	2394.00	147221.00	4.87	4.99	86072.00
max	4390,00	193808,00	28.86	29.94	93436,00

Table 1: Summary statistics 1a

	estimator	analys	value	actual	permno	siccd	prc	shrout	mktcap
count	156378	156378	156378	156378	156378	156313	156377	156378	156378
mean	1229.16	110538.78	3.43	3.46	52586.19	5290.41	70.37	359033.46	2.527619e+07
std	1317.26	48968.47	4.92	5.21	32981.21	2664.61	127.26	801172.15	7.877663e+07
min	11.00	0.00	-9.88	-13.63	10026.00	250.00	-19.19	513.00	-1.888680e+05
25%	171.00	77712.00	0.84	0.87	15627.00	3334.00	18.20	52999.00	1.498716e+06
50%	464.00	114857.00	2.45	2.45	60599.00	5100.00	41.42	123259.00	5.063315e+06
75%	2394.00	147210.00	4.89	5.01	86004.00	7360.00	84.32	323943.00	1.782115e+07
max	4390.00	193808.00	28.86	29.94	93436.00	10000.00	2517.06	9814197.00	1.080869e+09

Table 2: Summary statistics 1b

	permno	mktcap	fe	fd	coverage	financials
count	2714	2713	2714	2673	2714	2714
mean	58617.07	9916172.32	0.19	0.18	10.08	0.26
std	32596.46	42969462.95	0.31	0.23	8.83	0.44
min	10026.00	-57751.45	0.00	0.00	1.00	0.00
25%	16765.75	404202.06	0.03	0.04	4.00	0.00
50%	77386.00	1556870.32	0.07	0.10	7.00	0.00
75%	87710.25	5488859.88	0.22	0.23	14.00	1.00
max	93429.00	1080868643.90	2.87	2.00	57.00	1.00

Table 3: Summary statistics 2c

B Inserted Figures

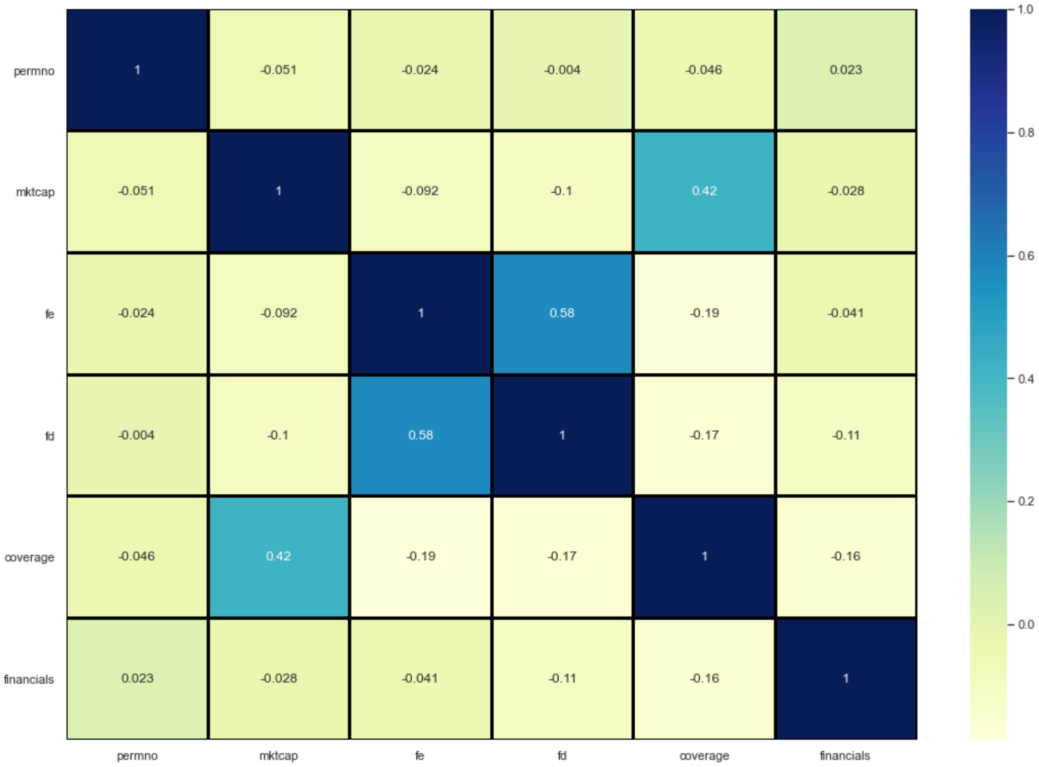


Figure 1: Correlation Heatmap

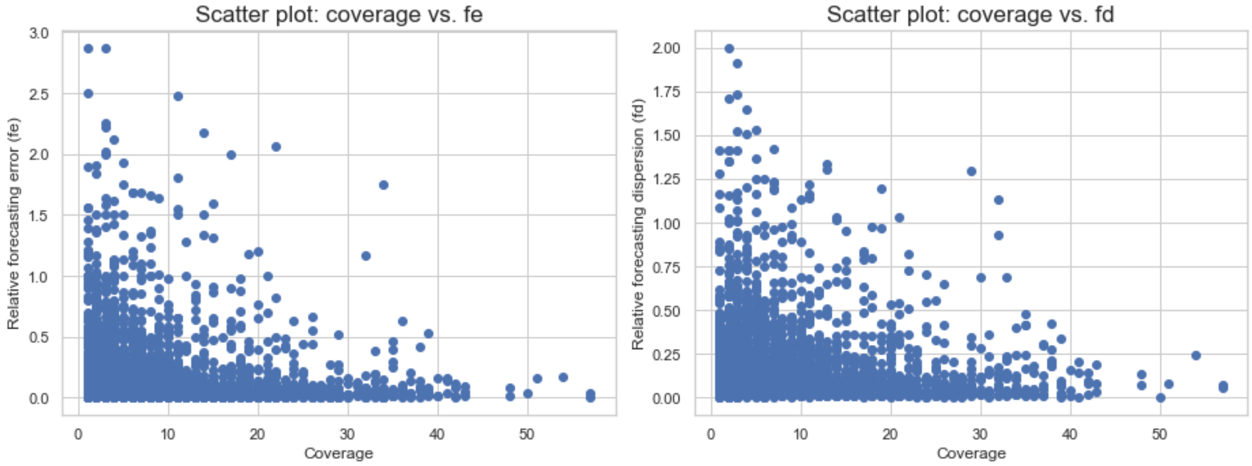
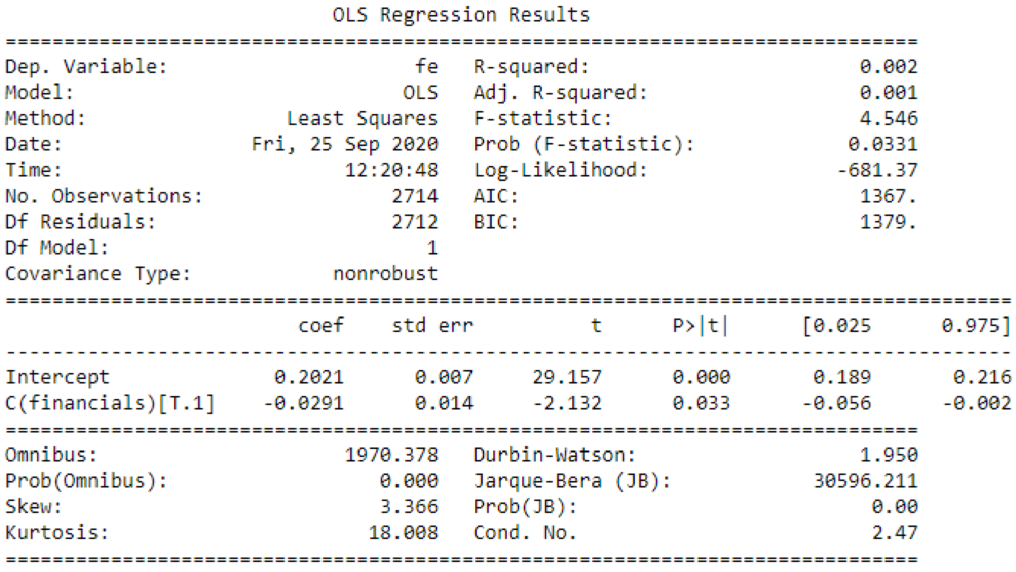
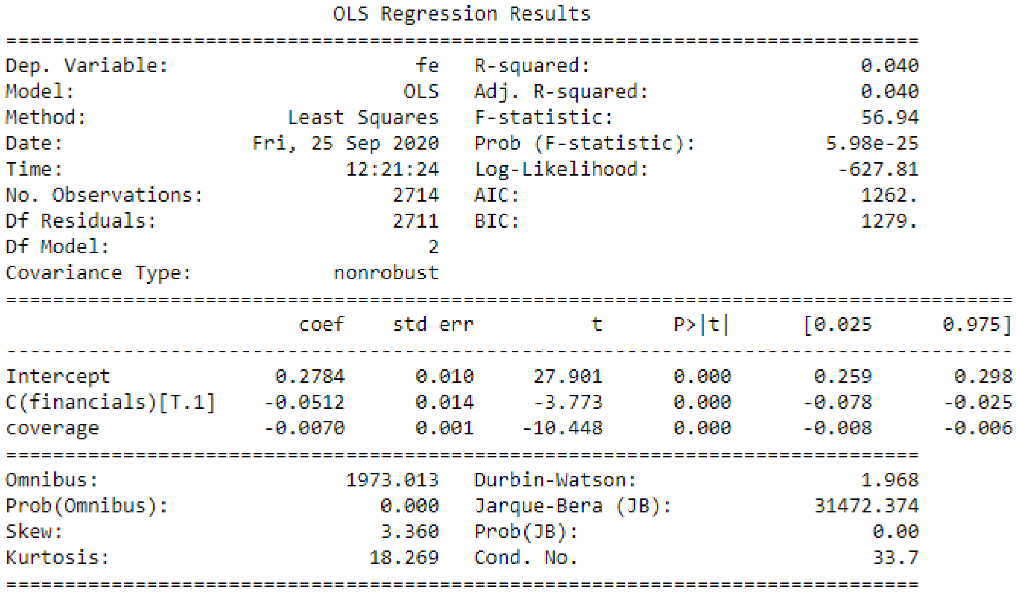


Figure 2: Scatter plot



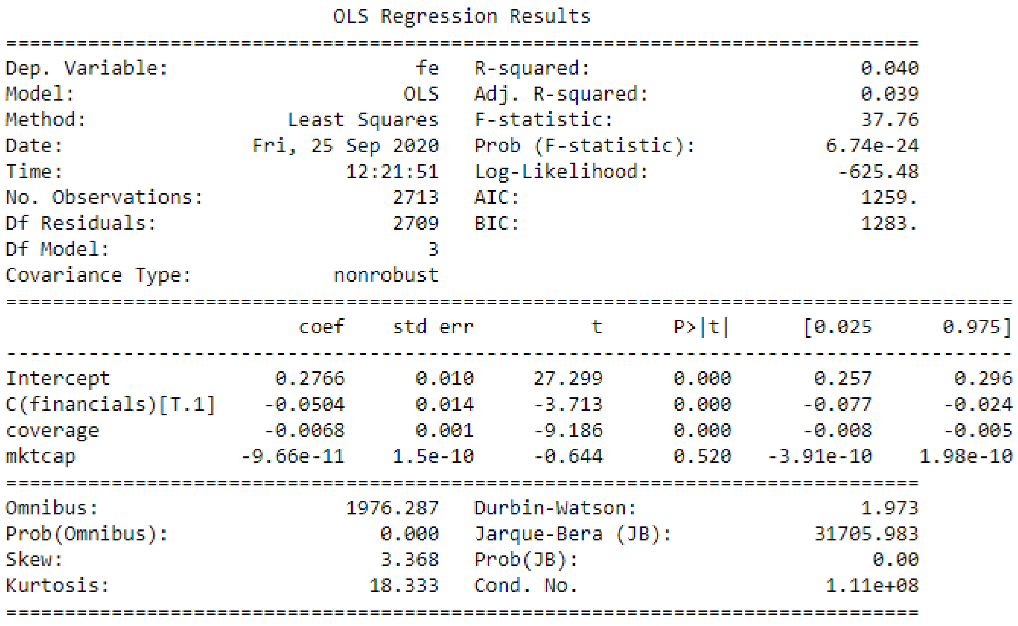
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     fe   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     4.546
Date:                Fri, 25 Sep 2020   Prob (F-statistic):             0.0331
Time:                        12:20:48   Log-Likelihood:                -681.37
No. Observations:                2714   AIC:                             1367.
Df Residuals:                    2712   BIC:                             1379.
Df Model:                           1
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept              0.2021      0.007     29.157      0.000       0.189       0.216
C(financials)[T.1]    -0.0291      0.014     -2.132      0.033      -0.056      -0.002
==============================================================================
Omnibus:                     1970.378   Durbin-Watson:                   1.950
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            30596.211
Skew:                           3.366   Prob(JB):                         0.00
Kurtosis:                      18.008   Cond. No.                         2.47
==============================================================================
```

Figure 3: Regression output 3a (1)



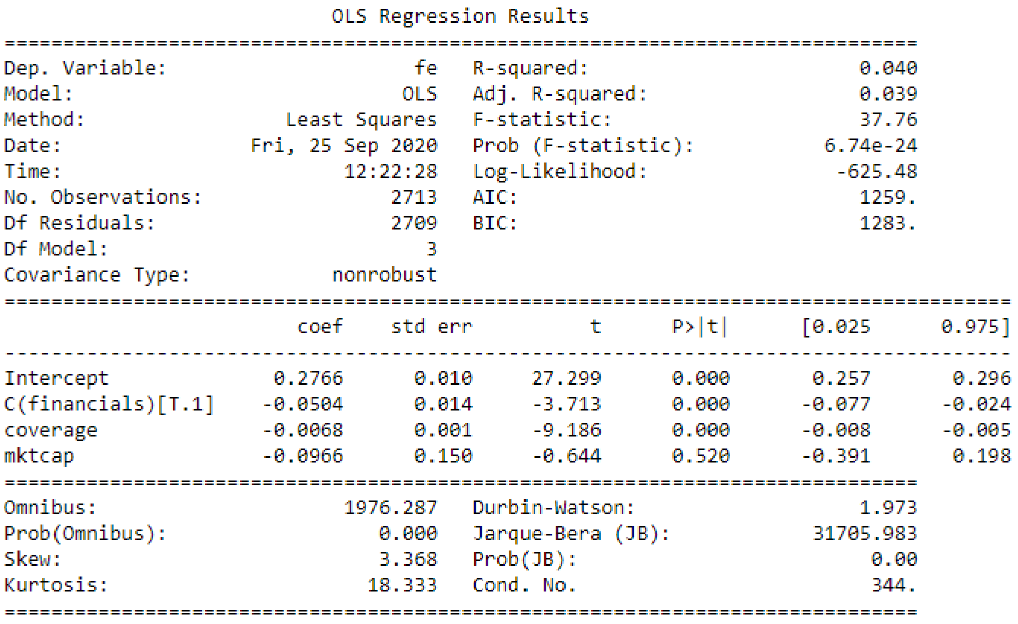
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     fe   R-squared:                       0.040
Model:                            OLS   Adj. R-squared:                  0.040
Method:                 Least Squares   F-statistic:                     56.94
Date:                Fri, 25 Sep 2020   Prob (F-statistic):           5.98e-25
Time:                        12:21:24   Log-Likelihood:                -627.81
No. Observations:                2714   AIC:                             1262.
Df Residuals:                    2711   BIC:                             1279.
Df Model:                           2
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept              0.2784      0.010     27.901      0.000       0.259       0.298
C(financials)[T.1]    -0.0512      0.014     -3.773      0.000      -0.078      -0.025
coverage              -0.0070      0.001    -10.448      0.000      -0.008      -0.006
==============================================================================
Omnibus:                     1973.013   Durbin-Watson:                   1.968
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            31472.374
Skew:                           3.360   Prob(JB):                         0.00
Kurtosis:                      18.269   Cond. No.                         33.7
==============================================================================
```

Figure 4: Regression output 3a (2)



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     fe   R-squared:                       0.040
Model:                            OLS   Adj. R-squared:                  0.039
Method:                 Least Squares   F-statistic:                     37.76
Date:                Fri, 25 Sep 2020   Prob (F-statistic):           6.74e-24
Time:                        12:21:51   Log-Likelihood:                -625.48
No. Observations:                2713   AIC:                             1259.
Df Residuals:                    2709   BIC:                             1283.
Df Model:                           3
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept              0.2766      0.010     27.299      0.000       0.257       0.296
C(financials)[T.1]    -0.0504      0.014     -3.713      0.000      -0.077      -0.024
coverage              -0.0068      0.001     -9.186      0.000      -0.008      -0.005
mktcap              -9.66e-11    1.5e-10     -0.644      0.520   -3.91e-10    1.98e-10
======================================================================================
Omnibus:                     1976.287   Durbin-Watson:                   1.973
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            31705.983
Skew:                           3.368   Prob(JB):                         0.00
Kurtosis:                      18.333   Cond. No.                     1.11e+08
==============================================================================
```

Figure 5: Regression output 3b



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     fe   R-squared:                       0.040
Model:                            OLS   Adj. R-squared:                  0.039
Method:                 Least Squares   F-statistic:                     37.76
Date:                Fri, 25 Sep 2020   Prob (F-statistic):           6.74e-24
Time:                        12:22:28   Log-Likelihood:                -625.48
No. Observations:                2713   AIC:                             1259.
Df Residuals:                    2709   BIC:                             1283.
Df Model:                           3
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept              0.2766      0.010     27.299      0.000       0.257       0.296
C(financials)[T.1]    -0.0504      0.014     -3.713      0.000      -0.077      -0.024
coverage              -0.0068      0.001     -9.186      0.000      -0.008      -0.005
mktcap                -0.0966      0.150     -0.644      0.520      -0.391       0.198
======================================================================================
Omnibus:                     1976.287   Durbin-Watson:                   1.973
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            31705.983
Skew:                           3.368   Prob(JB):                         0.00
Kurtosis:                      18.333   Cond. No.                         344.
==============================================================================
```

Figure 6: Regression output 3c