

Statistics for Finance
Project
Due date: December 16th, 2020
Submission via LEARN AND Turnitin
Absolute max of words: 1500

Complete the tasks below:

1. Using Capital IQ¹, download data for at least 50 US firms for the period 2010-2019. Present a table of summary statistics for all the variables used in the project (including the components of Tobin's Q). Make sure to include: mean, standard deviation, min, max, 25% percentile, 50% percentile, 75% percentile, number of firms, number of firm-year observations. Note that it could be the case that a particular firm has data available for 10 years, whilst another firm has data available for, for example, only 5 years. As long as you have a minimum of 5 observations per firm this is ok. Label this table: Table1: Summary Statistics² and include a legend at the end of the table that defines each variable; e.g. Price denotes Day Close Price; Equity denotes Total Common Equity, etc.

[5 marks]
2. Firm size is measured as $\log(\text{Total Assets})$. Performance is measured with Tobin's Q (total assets plus market value of equity less book value of equity divided by total assets; where market value of equity equals price per share times the total number of shares outstanding)³. Choose the largest and the smallest firm for which you have 10 years of data. Is average performance statistically different between these two firms? Answer yes or no and show two different ways in which you could reach this conclusion. Make sure to show all the details of your tests and present a table for each test. Label these tables: Table2A: Differences in Performance_1 and Table2B: Differences in Performance_2 (50 words max)

[5 marks]
3. You will run a cross-sectional regression. Therefore, compute time averages for every variable for each firm. You will end up with 50 cross-sections. Run a simple regression analysis to assess whether larger firms are associated with better performance. Note that you should run this regression using 50 observations. Label this table: Table 3: Simple-regression

[2 marks]
4. Discuss whether the coefficient on size is statistically significant at the 5% level and interpret the coefficient (50 words max).

[2 marks]

¹ A separate guide to Capital IQ can be obtained from the Project folder in Learn.

² Note that sometimes zeroes denote missing values. Make sure your summary statistics look sensible. Marks will be deducted for wrong construction of variables or careless calculations.

³ Selecting the correct variables to construct Tobin's Q is part of your mark. You may want to refer to academic papers to make the right choice and include references in Appendix 2.

5. Is the beta estimator for size unbiased? Explain (100 words max) [10 marks]
6. Using SIC codes add industry dummy variables to your regression. Label this table: Table 4: Industry controls. Explain what these dummy variables are controlling for (100 words max). [10 marks]
7. Add the following variables to the regression (don't forget also to include the industry dummy variables):
- a sensible explanatory variable of your choice (you may need to look at some academic papers to make a good choice and include references in Appendix 2) [5 marks]
 - A sensible dummy variable of your choice (you may need to look at some academic papers to make a good choice and include references in Appendix 2) [5 marks]
- Include the table with results. Label this table: Table 5: Multiple Regression Model
8. Discuss whether the variable chosen in 7a. above is statistically significant at the 5% level and interpret your result (50 words max). [2 marks]
9. Discuss whether the variable chosen in 7 b. above is statistically significant at the 5% level and interpret your result. (50 words max). [2 marks]
10. Using an F-test assess whether the industry dummies are jointly significant. Show the F-statistic, critical value and discuss your results (50 words max). [5 marks]
11. Compare the coefficient of size in Table 2 vs that of Table 5. Explain why they are different (100 words max) [7 marks]
12. Compare the R-square of Table 3 to that of Table 5. Which model is better? (100 words max) [5 marks]
13. Your variable of interest is size. Even after running your multiple regression model, your model is likely to suffer from endogeneity. Give an example of an omitted variable that could lead to a positive bias and include a brief explanation. (100 words max) [10 marks]
14. Suppose you add firm's average number of employees to the multiple regression model. Would you expect your main results to change? Explain. (100 words max) [5 marks]

15. In Lecture 4 pg. 63 we used an example based on the binomial distribution to explain the Central Limit Theorem (CLT). Make up your own example and show that the CLT works (originality will be rewarded). Feel free to include graphs or figures if this adds value to your explanation. (200 words max)

[10 marks]

16. Using daily prices for 2019 for the largest of your 50 firms (you do not need to include this variable in the descriptive statistics or any of the questions above), what is the predicted price for January 1st 2020 based on the random walk model? Clearly show how you estimated this price and discuss whether this is a good prediction. (200 words max)

[10 marks]

Notes:

Your tables need to be incorporated using images, directly obtained through Stata or SAS. These tables are not included in your word count. If the instructions only asks you to include a table, you do not need to provide any further discussion for that task.

You should use Times New Roman 11 point font on A4 pages with 2.5 margins from each side. An absolute maximum of 1500 words is expected but a good project could include as little as 1000 words. Include word count in the right corner of your project.

Appendix 1 (not part of the word count) should include the names of the 50 firms you are using. References (max 250 words) should be included in Appendix 2. Only use papers published in journals ranked as 3, 4 and 4*. A list of journal rankings can be obtained from the Project folder in LEARN.

Make sure that the names you give to your variables are consistent across tables. For example, if you refer to Log(Assets) as SIZE, then your regressions should show the coefficient for SIZE and not for Log(Assets).

You will need to submit the file of commands and data that replicates your Tables. For example, if you are using STATA you need the do file and its corresponding dta file. These files are not part of the word count.