Homework #2
QMI First-Years

This is the first real-real homework assignment. There is going to be some assigned reading, associated question, code, and question. The expected completion time of this assignment is 2-3 hours. We know that you are busy, but we believe this is the best way to make sure you are retaining the knowledge from the previous week. We also highly recommend that you complete this assignment independently, if you need help of course reach out to your fellow first-years or us!

Finance/Statistic Questions

Read "Understanding Risk and Return, the CAPM, and the Fama-French Three-Factor Model". You can skim to the "CAPM" Section, skip the "Regression Analysis" to "Fama And French" Section, and read completely to the end.

1. A common critique of beta is that it is historical. Critics argue that because it is historical all of the information "included" in beta is already baked into the current stock price (because price(t) of a stock incorporates all publicly available information under EMH semi-strong form). What are some indications from our lecture as well as the reading that this might not completely be the case? Explain fully. [hint: see autocorrelation]

   If stock prices were completely in line with the EMH semi-strong form then prices would be a random walk, but they're not because we see that they trend towards growth. Furthermore, because stock of stock autocorrelation where lagged prices are an indicator of themselves, stock prices are not a random walk. The same idea can be applied as to why beta might not be completely historical because autocorrelation indicates that there is some sort of present information being incorporated into the stock price.

2. Offer a critique, in your own words, of the SMB factor. Provide a brief description of the factor, why it is wrong, and cite two sources not used in lecture or readings.

   The SMB factor is based off the idea that small stocks have a higher returns than large stocks. The first reason why it might not work in modern day is because the SMB premium may not have persisted over time since the publication of the three-factor model. Schwert (2002) indicates that "it seems that the small-firm anomaly has disappeared since the initial publication of the papers that discovered it". A potential reason for this is that since after the publication of a large model and paper that uses that factor, investors likely utilized the factor until the returns balanced out in an EMH-like behavior, even if the whole of EMH also doesn't hold up today as well. This can be a reason why new models instead build upon and modify the factors in the three-factor model, because changes in the market have prevented the exact same model from persisting over decades. Krch (2017) Indicates that whether or not the size premium actually still exists continues to be debated across literature. However, he finds that when liquidity is included as an adjustment in determining the size premium, the size premium tends to disappear. Thus, research into including liquidity as a factor or as a modification to the size premium may lead to conclusions against the SMB factor as well.

   Schwert, G. William William (2002). *Anomalies and Market Efficiency. SSRN Electronic Journal, (), –*
       . doi:10.2139/ssrn.338080

   Krch P. (2017) Existence of Size Premium: A Review of Literature and Suggestions for Future Research. In: Procházka D. (eds) New Trends in Finance and Accounting. Springer Proceedings in Business and Economics. Springer, Cham. https://doi.org/10.1007/978-3-319-49559-0_71

3. Define r-squared from the paper in your own words.

   R-squared indicates how much of the variation in the data can be explained by the model. In the case of finance, it's how well the model and the factors it includes predict the data. This is traditionally used as a measure to how "good" a model is because a higher r-squared value indicates how well the model fits the data. For example, an r-squared of 0.95 indicates that 95% of the variation is explained by the model, while the remaining 5% of variation is not explained.

Read "Choosing Factors". Completely read the first nine pages up to the start of the "Marginal Contributions" Section and the "Conclusion" Section.

1. Why do we use Sharpe Ratio as an optimization function under the select factor models? [if you want further explanation of optimization this starts with a good definition: https://sites.math.washington.edu/~burke/crs/515/notes/nt_1.pdf]

It helps choose the best model because using the Sharpe Ratio as an optimization function helps choose the model that provides the highest risk-adjusted returns. Comparing models in this way lets you examine specific factors within models that include or exclude those factors such as the momentum factor that the paper addresses.

2. What are the reasons for optimizing the Sharpe Ratio? What does it indicate when the Sharpe Ratio is "fully" optimized?

   Optimizing the Sharpe ratio provides the highest risk-adjusted returns, so using those models should make you the most money. When the Sharpe ratio is fully optimized the stocks that the models choose should have a minimized negative standard deviation which is the denominator and a maximized return compared to the risk free rate.

3. Define data dredging. Explain why it is problematic for model testing.

   Data dredging is when you look for any sort of statistically significant relationship in data without considering the theory or logic behind whether or not there is any sort of causation there. This is bad for model testing because data dredging can lead to models where there is a lot of factors that are included because the data indicates that they may have a good relationship that is statistically significant even when there's no understanding of why that particular factor contributes to the model. This can lead to models where there is a large excess of factors.

The previous paper uses the phrase, "multiple comparisons issue" (Fama & French 2017). Read this, article to learn more. No questions right now, but it is a very important concept for trading.

https://towardsdatascience.com/the-multiple-comparisons-problem-e5573e8b9578

Computer Science Questions/Problems

1. Google "the Euler Project" and solve the first two problems (IDs 1 and 2). Code this in Jupyter Notebook.
   https://github.com/alchemicHen/QMI/blob/master/FY_Lesson%20-%20101321/Homwork_Two/EulerProject_ID1ID2.ipynb

2. Creating a webscraper, requirements:
   a. Must use requests, BeautifulSoup, and Selenium with each other
   b. Choose your own website to scrap
   c. Code must run completely without fail
   d. Clearly comment what you are doing as you see necessary (this is down by inserting a hashtag before a line of text)
   e. The scraper should use, in a robust way, requests (easy one), BeautifulSoup, and Selenium
   f. Create a Pandas dataframe of the data that you collect from the webpage
      i. The target website cannot be overly structured (e.g. Texas Death Row Execution website)
      ii. Target should have some sort of meaningful data to collect, this threshold is low (we just do not want you creating a dataframe of every word in a paragraph)
   g. Acceptable Project: Collecting all of the names and prices of Lego City sets from their website (for this we will provide an example/walkthrough).

   https://github.com/alchemicHen/QMI/blob/master/FY_Lesson%20-%20101321/Homwork_Two/ratemyprof_scraper.ipynb

This homework is due Wednesday 20[h], 2021 at 11:59 p.m. Please submit as a PDF and direct message it to Alex on Slack.