Homework 2

Due: 12:00 pm (noon), September 29, 2020

Do not include your name on your write-up, since these will be peer-graded anonymously.

Do not include your raw R code in your write-up unless we explicitly ask for it. You will submit your R script as a separate document to the write-up itself. On Canvas, you will see *two* assignments corresponding to homework 2: one for the write-up document, and one for the R script. Your write-up is what will be peer graded, but your R script must also be submitted for the purpose of audits and ensuring compliance with course policy regarding academic integrity.

Problem 1

If an aircraft is present in a certain area, a radar correctly registers its presence with probability 0.99. If it is not present, the radar falsely registers an aircraft presence with probability 0.10. Suppose that on average across all days, an aircraft is present with probability 0.05.

The radar today just registered the presence of an aircraft. What is the probability that an aircraft is actually present? Make sure to show your work.

Problem 2

Download the data in `creatinine.csv`. Each row is a patient in a doctor's office. The variables are:

- age: patient's age in years.
- creatclear: patient's creatinine clearance rate in mL/minute, a measure of kidney health (a higher rate means better clearance, i.e., more healthy).

Use this data, together with your knowledge of linear regression, to answer three questions:

A) What creatinine clearance rate should we expect for a 55-year-old? Explain briefly (using one or two sentences and equations) how you determined this.

B) How does creatinine clearance rate change with age? (This should be a single number whose units are ml/minute per year.) Explain briefly (one or two sentences) how you determined this.

C) Whose creatinine clearance rate is healthier (higher) for their age: a 40-year-old with a rate of 135, or a 60-year-old with a rate of 112? Explain briefly (a few sentences + equations) how you determined this.

Problem 3

Epidemiologists are investigating a cluster of respiratory disease around a coal-powered electricity generating station. Of 1072 elderly residents living within 5 miles of the coal plant, 49 of them have COPD, or chronic obstructive pulmonary disease, which is an inflammatory lung disease that causes breathing problems [ref]. The baseline rate of COPD among elderly people across the country is 3.3%. Is the observed data (49 cases out of 1072) consistent with the null hypothesis that, over the long run, elderly residents within 5 miles of the power plant experience COPD at the national background rate?

Use Monte Carlo simulation (with at least 100,000 simulations) to calculate a p-value under this null hypothesis. Include the following items in your write-up:

- the null hypothesis your are testing;
- the test statistic you used to measure evidence against the null hypothesis;
- a picture of the probability distribution of the test statistic, assuming that the null hypothesis is true;
- the p-value itself;

- and a one-sentence conclusion about whether you think the null hypothesis looks plausible in light of the data.

Problem 4

An important model that finance professionals use to understand asset prices is called the Capital Asset Pricing Model (CAPM). You will learn more about the CAPM in a future finance course. But the basic assumption of the model is that the rate of return on an individual stock is linearly related to the rate of return on the overall stock market. That is, each stock's rate of return is assumed to follow a linear regression model:

$$Y_t^{(k)} = \beta_0^{(k)} + \beta_1^{(k)} X_t + e_t^{(k)} ,$$

where $Y_t^{(k)}$ is the rate of return of an individual stock (k) in some given time period t; X_t is the rate of return of the entire stock market in that same time period; and $e_t^{(k)}$ is the residual for stock k in that time period. The superscript (k)'s here are simply denoting the different stocks (Apple, Target, etc), while the subscript t's are denoting the different time periods. Note that the market rate of return (X_t) is a predictor common to all stocks. (The rate of return can be interpreted similarly to an interest rate. For example, if a stock was worth \$100 yesterday and \$102 today, then it gained 2%, for an implied daily rate of return of 0.02.)

The β_1 (slope) term in this regression model is super important to finance professionals; they just call it "beta", and they refer to the β_0 (intercept) term as "alpha." Please watch this short YouTube video to understand how beta is used to think about different stocks.

Once you've watched the video, please turn to the data in `marketmodel.csv`, which contains information on the daily returns for the S&P 500 stock index, denoted SPY, along with the returns for 6 individual stocks: Apple (AAPL), Google (GOOG), Merck (MRK), Johnson and Johnson (JNJ), Wal-Mart (WMT), and Target (TGT). (We can think of the return of the S&P 500 as a proxy for the whole market.) The data start from the beginning of 2019. The entries are interpretable as percentage returns, expressed on a 0-to-1 decimal scale—for example, if the S&P 500 gained 1.5% in value on a given day, the corresponding entry in the data frame would be 0.015.

Regress the returns for each of the 6 stocks individually on the return of S&P 500 (which is like X_t, the market return, in the equation above). Make a clean, professional looking table (e.g. in Excel) that shows the ticker symbol, intercept, slope, and R^2 for each of the 6 regressions.

In your write-up, you should include:

- a two-to-three paragraph introduction, in your own words, on what the "beta" of a stock is measuring and how it is calculated. (Watch the video and summarize it in your own words, making sure to connect it to the regression model we've written down above—this is a bridge you will have to make yourself, using what you know about regression models.) A reasonable aim for your summary is about 250 words here, but this is approximate; nobody on our end is breaking out the word counter.

- the table itself, along with an informative caption below the table, no more than 2-3 sentences in length, to give readers the information necessary to interpret the table.

- a conclusion that answers two questions: in light of your analysis, which of these six stocks has the *lowest* systematic risk? And which has the *highest* systematic risk? (Again, watch the video to understand how this is measured using the regression model.)