Economics 403A: Project 1
Fall 2017, UCLA
Instructor: Dr. Rojas

Due Date: November 1, 2018

The document that you will submit, consists of a written report which includes answers to the questions below (including plots), and respective R source code. You only need to submit one project report per group but please make sure that every group member's name is included.

1. Regression: Least Squares, Bootstrap, and Bayesian (30%)

   For this problem we will compare regression estimates between Least Squares (LS), Bayesian, and Bootsrapping methods. Using any dataset of your choice from the AER library (you will need to install the package in R first), estimate a multiple linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$. Your data will need to have at least 2 predictors, and the LS estimates of the parameters should be statistically significant at the 5% level.

   (a) Use bootstrapping with $N = 1000$ samples to estimate the Bootstrap estimates of the model parameters (β_0, β_1, and β_2), and also the respective confidence intervals. Plot the estimates and respective intervals vs. trial number, and comment on their stability. Compare the LS estimates and confidence interval, to the overall Bootstrap estimates and confidence interval.

   (b) Using MCMC generate appropriate distributions for the model parameters and use these to perform a Bayesian regression and the respective 95% credible intervals. Plot the respective parameter posterior distributions with their credible intervals overlaid on the histograms. Compare the LS estimates and confidence interval, to the Bayesian results.

   (c) Based on your results from the 3 methods, provide a convincing argument for which method you believe is better. and why.

2. Fitting Distributions (30%)
   In this portion, you will be looking at different macroeconomic variables and determining whether or not the underlying distribution of these variables changes depending on whether or not the economy is in a state of recession. The variables you will be using will be the following:

   - S&P 500 index
   - Yield Spread (For simplicity, compute the spread as the difference between the ten year treasury bill rate and the three month treasury bill rate.)
   - Three month treasury rate
   - Japanese Central Bank's interest rates

Use monthly frequencies for the data. At a minimum, your data should span at least 50 years.

(a) For each of the four variables, plot the histogram of the data and overlay the respective density curve.

(b) Fit a distribution to each histogram in part (a).

(c) Now, subset the data to include only data points generated during recession periods. Repeat parts (a) and (b) with this subset of data.
Note: for recession periods, use the National Bureau of Economic Research's classification for when the United States is in a recession. For data corresponding to Japan, classify recessions in accordance with the Organization for Economic Cooperation and Development (OECD).

(d) Repeat part (c) with data subsetted to include only data points generated during non-recession periods.

(e) Based on your findings in parts (a) - (d), are there any noticeable differences between the estimated distributions based on how you subset the data? Explain.

(f) Confirm your answer from part (e) using the Kolmogorov-Smirnov Test.

3. Ordinary Least Squares: A lesson on CAPM (Capital Asset Pricing Model) and parameter stability/robustness. (30%)
he capital asset pricing model (CAPM) is an important model in the field of finance. It explains variations in the rate of return on a security as a function of the rate of return on a portfolio consisting of all publicly traded stocks, which is called the *market* portfolio. Generally the rate of return on any investment is measured relative to its opportunity cost, which is the return on a risk free asset. The resulting difference is called the risk premium, since it is the reward or punishment for making a risky investment. The CAPM says that the risk premium on security j is proportional to the risk premium on the market portfolio. That is,

$$r_j - r_f = \beta_j(r_m - r_f),$$

where r_j and r_f are the returns to security j and the risk-free rate, respectively, rm is the return on the market portfolio, and β_j is the jth security's "beta" value. A stocks beta is important to investors since it reveals the stock's volatility. It measures the sensitivity of security j's return to variation in the whole stock market. As such, values of beta less than 1 indicate that the stock is "defensive" since its variation is less than the market's. A beta greater than 1 indicates an "aggressive stock." Investors usually want an estimate of a stocks beta before purchasing it. The CAPM model shown above is the "economic model" in this case. The "econometric model" is obtained by including an intercept in the model (even though theory says it should be zero) and an error term,

$$r_j - r_f = \alpha_j + \beta_j(r_m - r_f) + e$$

(a) The data file capm4.dat contains monthly returns of six firms (Microsoft, GE, GM, IBM, Disney, and Mobil-Exxon), the rate of return on the market portfolio (MKT), and the rate of return on the risk free asset (RISKFREE). The 132 observations cover January 1998 to December 2008. Estimate the CAPM model for each firm, and comment on their estimated beta values. Which firm appears most aggressive? Which firm appears most defensive?

(b) Finance theory says that the intercept parameter α_j should be zero. Does this seem correct given your estimates? For the Microsoft stock, plot the fitted regression line along with the data scatter.

(c) Estimate the model (coefficients and confidence intervals) for each firm under the assumption that $\alpha_j = 0$. Do the estimates of the beta values change much?

(d) Bootstrap Sampling: Generate 1000 synthetic samples from Microsoft by sampling with replacement, estimate the model parameters for each sample, and show separate plots of your estimated β and α across each realization. In addition, define the length of the respective confidence intervals by d_α and d_β, and plot these lengths across each realization. Comment on your results

(e) Compute the mean and volatility of your estimates from part (d) for $\widehat{\beta}$, $\widehat{d_\beta}$, $\widehat{\alpha}$, and $\widehat{d_\alpha}$, and plot a histogram (including the respective density curve) for each one of your estimates.

(f) How do your results from part (e) compare against those from part (a)? Which one do you trust more and why?

4. Sequential Bayesian Learning (20%)

You will need to implement this in R using loops. Assume we are given the following facts:

- 1% of women aged 40 have breast cancer.
- A mammography test has a 99% success rate, and a 10% false alarm rate (i.e., 10% of the time, the test will return positive for having cancer when the patient does not actually have cancer).

(a) Given the above, a women aged 40 receives a positive mammography test. What is the probability that she actually has cancer? Let $C+$ = Cancer present and $T+$ = Positive mammography test, you need to find$P(C+|T+)$.

(b) Assume the same women from (a), wants to get another opinion, and then another, and so on, but that every time she gets tested, the results are the same as the first one. After how many trials, will $P(C+|T+) > 95\%$? Show a plot of $P(C+|T+)$ vs Trial Number, and comment on your finding.

(c) If on Trial 3, the test results show negative for breast cancer, but all other tests show positive, how would this affect $P(C+|T+)$ found in (b)?