Homework Assignment 1

Dacheng Xiu – Chicago Booth
BUS41100 Applied Regression Analysis

Due at the beginning of class on January 15, 2020
Include the signed honor code on your solutions

Warm-up and Review (Ungraded)

Do not hand in a write-up.

0.1 Normal Distribution Probability Calculations

Suppose that $X \sim N(-10, 25)$, i.e., X has a Normal distribution with a mean of -10 and variance of 25.

(i) Compute Prob$(X > -10)$; (ii) Prob$(X < -20)$; and (iii) Prob$(X = 0)$.

(iv) Express Prob$(-22 \leq X \leq -12)$ in terms of Z, the standard normal random variable.

0.2 Functions of Random Variables

Suppose that $\mathbb{E}[X] = \mathbb{E}[Y] = 0$, $\text{var}(X) = \text{var}(Y) = 1$, and $\text{corr}(X, Y) = 0.5$.

(i) Compute $\mathbb{E}[3X - 2Y]$; and (ii) $\text{var}(3X - 2Y)$.

(iii) Compute $\mathbb{E}[X^2]$.

0.3 Summation Notation (a)

i	1	2	3	4
Z_i	2.0	-2.0	3.0	-3.0

(i) Compute $\sum_{i=1}^{4} z_i$

(ii) Compute $\sum_{i=1}^{4} (z_i - \bar{z})^2$

(iii) What is the sample variance? Assume that the z_i are i.i.d.. *Note that i.i.d. stands for "independent and identically distributed".*

0.4 Summation Notation (b)

For a general set of N numbers, $\{X_1, X_2, \ldots, X_N\}$ and $\{Y_1, Y_2, \ldots, Y_N\}$ show that

$$\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{N} (X_i - \bar{X})Y_i$$

0.5 The Sampling Distribution

Suppose we have a random sample $\{Y_i, i = 1, \ldots, N\}$, where $Y_i \overset{\text{i.i.d.}}{\sim} N(\mu, 4)$ for $i = 1, \ldots, N$.

(i) What is the variance of the sample mean?

(ii) What is the expectation of the sample mean?

(iii) What is the variance for another i.i.d. realization Y_{N+1}?

(iv) What is the standard error of $\bar{Y}$?

Data Analysis and Understanding (Graded)

Present solutions in a professional manner. Do not hand in codes.

1.1 What's Wrong With 538?

Read the following article on 538's prediction: `http://www.huffingtonpost.com/entry/whats-wrong-with-538_us_581ffe18e4b0334571e09e74`
Was 538's prediction wrong? The article was last updated on Nov 8, 2016. What would have been your view ex-ante? Comment briefly from a *statistical* point of view. (No politics please.)

1.2 Measuring Yield Curve Movements

Over the last year, the U.S. yield curve—the relationship between interest rates and maturity—has moved in interesting ways. To better understand this behavior, you decide to investigate an easier question: how are interest rates at different maturities on the yield curve correlated amongst themselves? The U.S. Treasury Department releases data on interest rates at varying maturities everyday, so you decide to use that data. The daily yields for each reported maturity in 2019 are given in the file *treasury.csv*.

(i) In some preliminary research, you learn that short-term interest rates might move while longer-term rates stay fixed and vice versa. If this is the case, explain which correlation you would expect to be larger:

  (a) The correlation between movements in the 1 month rate and the 3 month rate.

  (b) The correlation between movements in the 1 month rate and the 10 year rate.

(ii) Compute the differences in consecutive daily interest rates for each maturity. Use the following formula: if r_t is the interest rate at date t, then the change Δr_t at time t is given by

$$\Delta r_t = r_t - r_{t-1}.$$

You should get a dataframe with the same number of columns as your original dataframe, but one less row. You do not need to report anything for this part.

(iii) Compute the pairwise correlation matrix for the changes in the interest rates, and use the `corrplot` function in `R` to make a visual representation of the correlation matrix. Note that you will need to install the `corrplot` library first if you don't already have it installed. Describe any observations you notice, and explain how your observations compare with your expectations in part (i).

(iv) Now, create a scatter plot matrix using the 1-month, 2-month, 3-month, 20-year, and 30-year changes in interest rates. Describe any additional observations you notice. Do these observations agree or disagree witih your answer in part (iii)?

1.3 The AIG Stock Price

You are a smart high-frequency trader working for a Chicago-based hedge fund. On Friday, September 19, 2008, AIG's stock price starts plummeting shortly after market opening. One of your secret market-making strategies requires modeling and predicting the short-term trend of AIG's stock price.

You gather the price of AIG stock for each second between 9:35:00 and 9:36:00 in aig.csv.

(i) Ignore `timestamps`. If the `prices` are i.i.d. with mean 3.721 and variance σ^2, what would you estimate for σ^2?

(ii) Now, assume that you model `prices` as independently distributed with variance σ^2 and mean $\mathbb{E}(\texttt{price}_i) = 4 - 0.005 \cdot (\texttt{Time}_i - 93500)$. What would you estimate for σ^2? By comparison to your estimate in (i), what does this say about this model?

(iii) Find the correlation between `price` and `time` (pretending the trading time is a random variable), and use this to fit a regression line to the data. Plot the data and your line together and describe the fit.

(iv) What is the average of the residuals from your fit in (iii)?

(v) How would you use the residuals to estimate σ^2 in this fitted model? How does this estimate suggest that the fitted model compares to those in (i) and (ii)?

(vi) A senior trader, who is an alumna of Chicago Booth and has taken this course before, suggests a different model: $\mathbb{E}(\texttt{price}_i) = \texttt{price}_{i-1}$. Calculate σ^2 and compare it with those in (i), (ii) and (v). Whose strategy would your boss prefer? *Hint: you may find the* `diff()` *function useful. The last question is an open-ended one, which invites you to think outside the box.*

1.4 Rent Data Exploratory Analysis

The `rent.csv` data (available on the Chalk) contains information on *Rent* (in \$), number of *Rooms* and *Bathrooms*, the year of construction ($YearBuilt$), square footage (100 $SqFt$), and existence of Air Conditioning (AC) or *Parking* for $n = 696$ apartments in Chicago.

You are going to explore how these variables affect rent.

(i) Produce a boxplot for the marginal distribution of *Rent*, and compare this to boxplots for conditional distributions for *Rent* given each level of *AC*, and for *Rent* given the different numbers of *Rooms*. What can you say about the effect of these variables on rent?

(ii) Reconsider the questions in (i) through ANOVA rather than plots. That is, do two ANOVAs to find the SSR, SSE, and SST for *Rent* grouped either by *AC* and *Rooms*.

- In each case, what % of total variation is explained by the grouping?
- Do you think that these factors make a meaningful difference to rent?
- How does your conclusion compare to the F value and $Pr(> F)$ on the ANOVA table?
- Compare your results to the boxplots in (i).

(iii) Now investigate the effect of apartment size ($SqFt$) on rent. Calculate the correlation between $SqFt$ and *Rent* and use this to fit the regression line $Rent = b_0 + b_1 SqFt + e$. What does b_1 tell you about the influence of $SqFt$? What would you say if asked to predict at $SqFt = 0$?

(iv) Consider the results from your regression in (iii). Plot the data and regression line. Plot the residuals both as a histogram and against $SqFt$. Do you see any problems? Could you get a better model by ignoring some observations? If so, re-fit the model.