INFX573–Data Science I–Final Exam

Deadline: Mon, Dec 11th, 5:30pm PST

These 100 points will give you up to 25 points of the final credit.

Instructions This is a take-home final examination. You have 53 hours and 30 minutes to solve these problems. You may use your computer, books/articles, notes, course materials, internet, etc., but all work must be your own. References must be appropriately cited. Links to solutions copied from websites, such as StackOverflow must be provided in code comments. Please explain your answers and show all work; a complete argument must be presented to obtain full credit.

All plots must be appropriately labeled, and appropriate colors/labels/font sizes must be used.

Submission You will be doing most of this exam on computer. You must submit the corresponding rmarkdown file, and the compiled pdf file. You may prefer to solve the last problem on paper, please either take picture of your handwritten solution and include in the rmarkdown as an image, or alternatively you can submit it separately. You may submit the handwritten part on paper in Ott's mailbox too (in MGH370, ask the receptionist for the exact location).

Statement of Compliance You *must* include the "signed" Statement of Compliance in your submission. The Compliance Statement is found on the last page of this exam. Failure to do so will result in your exam not being accepted.

Both Ott and Luke will be replying your questions both on email and slack, but we won't guarantee 24/7 availability!

Good luck!

Problem 1: 2016 Election Results Data (25pt)

Use the following datasets to analyze the US 2016 election results (available on canvas in files/data) *by county*.

- *US_ County_ Level_ Presidential_ Results_ 08-16.csv.bz2*
- *county_ data.csv.bz2*
- there is also an explanation file for the county data *county_ data_ variables.pdf*.

Note: the datasets can be merged by *FIPS* (Federal Information Processing Standards) codes. There are 3-digit county FIPS codes and 2-digit state FIPS codeds, some data use 5 digit FIPS instead: 2 digits for the state followed by 3 digits for the county.

1. Tidy the data. Merge these datasets, retain only more interesting variables, compute additional variables you find interesting, and consider giving these more descriptive names. Explain briefly what did you do.
2. describe the data and the more interesting variables. Which variables' relationship to the election outcomes you might want to analyze?
3. plot the percentage of votes for democrats versus the county population. What do you conclude? Use the appropriates labels/scales/colors to make the point clear.
4. Create a map of percentage of votes for democrats. Do your best to reflect the continuous percentage of votes, and the different population sizes across counties and keep county boundaries as well legible as you can. Mark state boundaries on the map.

   Explain what did you do, and what worked well, what did not work well.

   Hint: there are many ways to map data in R. You may consider function `ggplot::map_data` that includes various maps, including US administrative boundaries. However, `map_data` counties do not include FIPS code. You may rely on merging data by state name and county name, given you a) convert your names to lower case, and b) remove the word `" county"` from the end of the names. This works for most of the counties, except for Lousiana where counties are called "parish".
5. Create one more visualization regarding the election results on your choice. The plot should be informative and clear. Use appropriate colors/labels/explanations.

Problem 2: 2016 Election Model (25pt)

Use the data from the previous problem. Your task is to *estimate the probability that a county voted for democrats* in 2016 elections (ie the probability that democrats received more votes than GOP).

Note: you may want to include more/different variables than what you did in the previous problem.

1. List the variables you consider relevant, and explain why do you think these may matter for the election results.
2. Estimate a logistic regression model where you explain the probability of voting democratic as a function of the variables you considered relevant. Show the results (summary).
3. Experiment with a few different specifications and report the best one you got. Explain what did you do.

   Hint: we did not talk about choosing between different logistic regression models. You may use a pseudo-R^2 value in a similar fashion as you use R^2 for linear models. For instance, `pscl::pR2` will provide a number of different pseudo-R^2 values for estimated glm models, you may pick McFadden's version.
4. Explain the meaning of statistical significance. What does it mean that an estimated coefficient is statistically significant (at 5% confidence level)?
5. Indicate which results are statistically significant in your preferred model.
6. Interpret the results. Provide correct interpretable explanations about what the most important effect are and what do the particular numeric results mean.

   Hint: you may use either odds ratios or marginal effects.

Problem 3: Simulate the Effect of Additional Random Coefficients (25pt)
Here your task is to simulate the logit coefficients of irrelevant input variables. You may either pick your favorite model from above, or use a different specification.

1. Choose a distribution. Poisson is fine, but you may pick something else as well.
   (a) Create a vector of random numbers, exactly as long as many observations you have in your data.
   (b) Estimate the logistic regression model using your former specification, but adding the random number as an additional explanatory variable.
   (c) store the coefficient for the random variable.
   Hint: function *coef* gives you the estimated coefficients of the model. It is a named vector, you can extract the coefficient of interest as `coef(m)["varname"]` where `m` is the estimated model and `"varname"` is the name of the variable of interest.
   (d) repeat these steps a large number $R \geqslant 1000$ times. Now you have R estimates of the coefficent for pure carbage features.
2. What are the (sample) mean and (sample) standard deviation of the estimated coefficients?
3. Find the 95% confidence interval of the coefficient based on your simulations.
4. Plot the distribution of the estimates (histogram, or another density plot).
5. Assume the estimates are randomly distributed with mean and standard deviation as you found above. What are the theoretical 95% confidence intervals for the results?
6. Extra credit (2pt): run the simulations in parallel. Report how much faster did it go compared to sequential processing.

Problem 4: Coin Tossing Game (25p)
You are offered to participate in a coing-tossing game. The rules are following: the coin is tossed until tail comes up. If tail comes up in the first flip—(T), you receive \$1. If a single head pops up before the tail—(H, T), you get \$2. If two heads pop up—(H, H, T), you receive \$4. If three heads appear before the first tail—(H, H, H, T), you get \$8. And so forth, so if the realized sequence is n heads and thereafter a tail—$(\underbrace{H, H, \dots, H}_{n}, T)$, you will receive $\$2^n$. The tosses are independent.

1. Assume the coin fair.
   (a) Compute your expected payoff in this game.
   (b) How much would you actually be willing to pay for the participation? Note: we are asking your judgement/opinion here, not a computed value.
2. Assume such a game is played 10 times and the outcomes are: twice (T); three times (H, T); once (H, H, T); twice (H, H, H, T), once (H, H, H, H, T); and once (H, H, H, H, H, H, T). This is your data. Your task is to find the Maximum Likelihood estimator of p—the probability of receiving a head.
   (a) Write down the probability to receive n heads and a tail.
   (b) Write down the probability to receive all 10 outcomes listed above.
   (c) Write the log-likelihood function of this data as a function of the parameter.
   (d) Analytically solve this log-likelihood for the optimal probabilty $\hat{p}$.
   (e) Plot the log-likelihood as a function of p. Mark the ML estimator $\hat{p}$ on the figure.

Statement of Compliance

Please copy and sign the following statement. You may do it on paper (and include the image file), or add the following text with your name and date in the rmarkdown document.

I affirm that I have had no conversation regarding this exam with any persons other than the instructor or the teaching assistant. Further, I certify that the attached work represents my own thinking. Any information, concepts, or words that originate from other sources are cited in accordance with University of Washington guidelines as published in the Academic Code (available on the course website). I am aware of the serious consequences that result from improper discussions with others or from the improper citation of work that is not my own.

(signature)

(date)