INFO370 – Core Methods – Final Exam

Deadline: Mon, Dec 9th, 8:00pm PST

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

Instructions This is a take-home final examination. You may use your computer, books/articles, notes, course materials, internet, etc., but all work must be your own. **This is an individual exam**, no cooperation or discussion related to the exam is allowed before the deadline. 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.

You must submit both the corresponding rmarkdown file, and the compiled pdf file.

The largest tasks here are the create ML models. You can easily spend days on squeezing most out of even these relatively simple data. This is not what we expect of you. Create simple, solid models. First do all the obligatory parts. And if you still have time left over, you can tinker further with the models and improve the performance.

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.

Ott, Sharan and Vishal will be replying your questions both on email and teams, but we won't guarantee 24/7 availability!

Good luck!

1 Multiple regression (40pt)

Your task is to analyze US presidential elections 2016 (and earlier) and find evidence for certain patterns and trends.

1. Load the datasets: *us-elections-2016.csv* and *county-data.csv*. Merge these datasets.

   Hint: check out what is FIPS code. County data does not contain 5-digit FIPS but you can create it from the state fips (variable *STATE*) and county fips (variable *COUNTY*). 5-digit fips code is constructed as *SSCCC* where *SS* is the state fips code and *CCC* is the county fips code.

2. Do a consistency check: how many states do you have? Is there anything missing? Is a non-state included?

Your main task is to analyze the relationship between *percentage of votes cast for democrats*, and various state-related demographic characteristics. We know the county population (a proxy for urban areas) is a strong predictor. But are there more? Let's find it out!

3. compute the percentage of votes for democrats in 2016. Do consistency checks (minimum, maximum, mean, and missings).

   Why do we want to see minimum and maximum here?

4. Visualize the relationship between democrats' percentage and county population. A scatterplot is a good bet but you can come up with something else too.

   Hint: consider log scale.

5. What other variable you think may be related to the vote percentage? Make two more visualizations.

Now it's time to get to regression analysis. As our task is to explain the relationship, not to predict, we are not concerned about overfitting here.

6. estimate a simple regression model where you eplain the democrats' percentage by the county population. Display the summary table, and interpret all the coefficients you have.

   Are these coefficients statistically significant at 5% confidence level?

7. re-estimate the model, but instead of using population, use log(population) as your explanatory variable.

   Suggestion: create a new variable log(population).

   Note: usually, when doing log transform, one uses natural logs, not decimal logs.

8. Compare the two models. Which one do you prefer? Why?

9. Now estimate a multiple regression model. Include all the variables you think are relevant in explaining the democrats' share. You may use log transform (or other transforms) if you prefer. Interpret at least 5 coefficients.

2 Theoretical questions (20pt)

Please answer the questions below by writing a short response.

1. Assume you run a regression model and get the following results for a particular parameter: $\beta = 1.23$ and it's standard deviation $\text{sd}\,\beta = 0.3$. What is the t value?

2. Is this estimate statistically significant (at 5% confidence level)?

3. What does it mean: a parameter is statistically significant at 5% confidence level?

4. You are developing a medical test for a rare disease. Only 1% of those who take the test actually have the disease. The test is pretty good (high recall): if a person with the disease takes it, it is positive in 99% of cases. But it is imperfect (not perfect precision): if a healty person takes the test, it is still positive in 1% of cases.

   What is the probability that one has the disease if the test turns out positive?

3 Classification (40pt)

In this problem you work with credit card application data. The dataset originates from UCI machine learning repository `https://archive.ics.uci.edu/ml/datasets/credit+approval`. As a way of protecting privacy of the involved, all the values are changed to meaningless labels, and the variables names are just *A1–A16*. Your task is to predict *A16* being "+".

1. Load the data. Describe it briefly, in terms of observations, missings, and such. See also the associated metadata.

2. Next, ensure the variables are of appropriate type, in particular, check if the data type of the numeric variables is numeric.

3. Now describe the variables: for the categorical variables print the possible categories, for the numeric variables compute the means and range.

Now we are done with the preparatory work. The next task is to predict A16 being "+", and to develop the best model for this task. Follow these broad steps: split your data into training/testing; develop 3 different linear probability models (just linear regressions predicting the probability of "+"). Complement these with similar logistic regression models, i.e. logistic regression models that use exactly the same explanatory variables (or exactly the same design matrix X if you prefer to model using matrices). In each case compute the accuracy on the testing data, and finally report your best model.

4. Split your data into training/testing part. Depending on the libraries you are using you may or may not want to separate A16 as your outcome variable. Keep this training/testing split through the rest of the problem.

5. Create a LMP (linear probability model, just linear regression) predicting A16 being positive. Let your first model be a simple one, containing no more than a few explanatory variables. Make sure you use only training data to fit this model!

   Hint: you may want to transform A16 into a numeric variable. You may also start with a model that just contains a constant term.

6. Compute the prediction accuracy on testing data using this model.

7. Repeat these two steps with a similar logistic regression model.

   Note: do not re-split the data!

8. Next, develop two more complex models adding more variables from the data. Do both the LPM and logistic regression versions of both models. Attempt to include as many variables as you can, but avoid overfitting (and numerical problems you may run into).

9. Finally, present your best model (in terms of accuracy on test data). In your approach–did the LMP or logistic perform better?

4 Extra credit (5 extra credit points)

How does the confidence interval depend on the number of observations for different distributions?

Your task is to simulate empirical confidence interval for two types of data: a) the democrats' vote share you calculated above, and b) citation counts for academic papers.

1. Load the datasets. Vote share you already did above, the other data is in file *mag-in-citations.csv*, and the variable of interest is *citations*. Display the main characteristics (min, median, mean, max) for both datasets.
2. plot histograms of both datasets. Discuss the differences.
3. Choose $\mathsf{N} = 10$ and $\mathsf{R} \geqslant 1000$.
4. Do the following with the votes data:
   (a) sample randomly (with replacement) N vote shares from the dataset. Compute your sample mean.
   (b) Repeat this process R times. You will have R means.
   (c) Find mean-of-the means, and it's 95% empirical confidence intervals (i.e. 0.025 and 0.975 quantiles). Output the width of the confidence interval (i.e. 0.975 quantile − 0.025 quantile).
5. Repeat the above with citation data.
6. Repeat the steps for both salaries and citations with $\mathsf{N} = 30$ and $\mathsf{N} = 100$.
7. Describe how do the confidence interval get narrower as your N increases. How is this related to the data distribution?

Finally...

please tell how much time (hours) did you spend on this exam!

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)