POLI502 Methods of Political Analysis, 2024 Fall
Instructor: Howard Liu

2024.12.9

Final Exam

Due 2024.12.9 6:00pm

Rules

- Submit your exam on your Github repository under a folder called "final502." Late submissions will not be accepted.
- Submit following files: (1) your qmd code script (e.q., POLI502-final-howardLiu.qmd). (2) a rendered pdf so I can write comments on it (POLI502-final-howardLiu.pdf). I will take 5 points off if you fail to do it.
- Make sure your script has a proper header like the my qmd samples. Also make sure your script runs without an error. I will execute your file to check if need be. I will take 5 points off if you fail to do it.
- Add comments and annotations to everything you do. Try to make your script look like mine. If your output doesn't have proper annotations, you'll lose 5 points. Don't copy and paste all the questions into your script, but do show me the question number for each question.

Notes

- The last question is probably more difficult and time consuming than others. I advise you not to spend too much time on it if you find it too difficult.

Tasks (10 points each × 10 = 100 points)

The Titanic dataset

1. Load the Titanic passenger survival dataset (titanic2.csv). You can download it [here]. Store it as an object named *td*. The dataset contains a variable named `fare`, which is the price of the ticket each passenger has. It is shown in pre-1970 GBP. (Note: $ 1 in 1911 is equivalent in purchasing power to about $ 112 in 2023.) Do female passengers tend to have a more expensive ticket compared with male passengers? Explore the relationship between the `female` variable (coded as "Female" for female passengers and "Male" for male passengers) and `fare` by a graph or a table. Interpret your exploration.
2. To more rigorously test the relationship, choose an appropriate *bivariate statistical testing* method and perform the test. Provide command(s) to perform the analysis. (Hint: I am not asking you to run a regression.) Interpret the results of the bivariate test you performed above and answer the question (do female passengers tend to have a more expensive ticket?). Comment on the observed pattern in the sample as well as the statistical significance, and draw a conclusion (i.e., answer the question posed here). Your answers must have up to three sentences.

3. Use the Titanic data set again to do the following.

   (a) Estimate a logit model of passenger survival where `survived` is the dependent variable and `fare`, `female`, and `child` are the independent variables. Estimate another logit model by including the natural log of the fare variable instead of the original fare variable. Produce a stargazer table summarizing the results from the two models estimated above.

   (b) According to the model fit statistics, which model performs the better? It goes without saying that you need to tell me the basis of your judgement as well. Your explanation here could be very brief (one sentence will do).

   (c) Produce two `effect` graphs that show the substantive effect of `fare` on passenger survival, one based on the first model and the other based on the second model, holding all the other independent variables constant at their median value.

   - Hint: Make sure that the x-axis of the second graph ranges between 0 and 500 (covering the range of the original `fare` variable), not between 0 and 6 (covering the range of the log transformed `fare` variable). If you have correctly used the `I()` function in 3-(a), you don't need to do anything special here. However, if your x-axis ranges between 0 and 6, you may want to go back to 3-(a) and correct it.

   (d) Compare the two graphs. Both graphs are somewhat non-linear, but one graph is more linear than the other (i.e., one graph is more non-linear than the other). The two graphs thus tell us different stories about the marginal effect of `fare` on passenger survival. Discuss this difference (provide your answer in your `R` file as a comment) **in five sentences** (You will lose points if you have more than five sentences or less than five sentences). Your discussion should have the following structure. Say something along the lines of: "The graph with the original `fare` variable suggests that the effect of `fare` on survival is linear / non-linear (choose one). That is, $\cdots\cdots$ (explain what a linear or non-linear relationship means in this context). On the other hand, the graph with the logged `fare` variable suggests that the effect of `fare` is linear / non-linear (choose one). That is, $\cdots\cdots$ (explain what a linear / non-linear relationship means in this context). Based on the model fit statistics, we should believe the first / second (choose one) story to be more plausible."

   (e) One way to evaluate the substantive importance of the `fare` variable would be to see how much this variable improves predictive abilities of the models in the out-of-sample setting. We can do so by comparing ROC curves with and without the logged `fare` variable. To do so, let's first estimate a logit model in the **training set** that does not include the logged `fare` variable while retaining the `female` and `child` variables. Most importantly, you need to subset the data into a training (80 % of the original data) and a test set (20 % of the original data) in order to do out-of-sample prediction

   - Hint: You need to use a subset of the data where there is **no missing value** for the `fare` variable. This is because we would like to ensure our comparison below is going to be based on the same set of observations.
   - Hint: Use the 'sample' command in R to randomly select rows into your training and test set. Something like this would help:
     - $set.seed(123)$ ←This ensures that you get the same random sample if you start with that same seed each time you run the same process
     - $train\_id = sample(1 : nrow(DATA), nrow(DATA) * 0.8)$

   (f) Produce a stargazer table that contrasts a model that includes only `female`, and `child` and another model that includes logged `fare`, `female`, and `child`.

(g) Produce a graph for ROC curves for these two models as your prediction result (i.e., the model that includes logged `fare` and the model that does not include `fare`).

- Hint 1: Try to produce one graph that shows two ROC curves in one graph.

(h) Report AUC scores for the two models. That is, write a command that gives us AUC scores for the two models.

(i) Based on the ROC curves and AUC scores you produced in 1-(i) and 1-(j), which model performs better? Your answer must be based on your interpretation of the ROC curves and/or the AUC scores (both will lead you to the same conclusion). This could be very brief (two or three sentences will do).

The Putnam data set

Load the Putnam data set (putnam.csv) from [here] and study the relationship between Institutional Performance and Civic Community:

4. Please estimate the following three regression models and produce a table that shows the regression coefficients and other statistics in multiple columns:

   (a) a simple linear regression model where Institutional Performance is the DV and Civic Community Index is the IDV;

   (b) an additive model that controls for a dummy variable that captures the North-South division. You may want to create a new dummy variable called North, rather than using the original factor variable NorthSouth for ease of interpretation;

   (c) an interactive model that allows Civic Community Index to have different slope for Northern and Southern regions.

5. Draw three graphs:

   (a) Draw a graph that shows the marginal effect of Civic Community Index based on Model (1).

   (b) Draw two graphs (one for Northern and another for Southern regions) that show the marginal effect of Civic Community Index based on Model (2).

   (c) Draw a graph that shows the marginal effect of Civic Community Index based on Model (3) for Northern and Southern regions. (Hint: your graph should have two panels, one for Northern and another for Southern regions.)

6. Based on the numerical regression results as well as the three sets of graphs you have drawn above, would you conclude that the regional distinctions make the relationship between Institutional Performance and Civic Community Index spurious? Why or why not? Your answer for this question must be written in your Rmd file and shown in your html file.

The relationship between Institutional Performance and Economic Modernization:

7. Please estimate the following three regression models and produce a table that shows the regression coefficients and other statistics in multiple columns:

   (a) a simple linear regression model where Institutional Performance is the DV and Economic Modernization as the IDV;

   (b) an additive model that controls for a dummy variable that captures the North-South division;

(c) an interactive model that allows Economic Modernization to have different slope for Northern and Southern regions.

8. Draw three graphs:

   (a) Draw a graph that shows the marginal effect of the Economic Modernization variable based on Model (1).

   (b) Draw two graphs (one for Northern and another for Southern regions) that show the marginal effect of the Economic Modernization variable based on Model (2).

   (c) Draw a graph that shows the marginal effect of the Economic Modernization variable based on Model (3) for Northern and Southern regions. (Hint: your graph should have two panels, one for Northern and another for Southern regions.)

9. Based on the numerical regression results as well as the three graphs you have drawn, would you conclude that the regional distinctions make the relationship between Institutional Performance and Economic Modernization spurious? Why or why not?

10. Derive OLS coefficient estimates, the intercept and the slope (α and β), for a simple two-variable regression regression: $Y_i = \alpha + \beta\, X_i + \mu_i$. Show the derivation process, not just the result.