**POLITISCI4782 Lab Assignments**

**(2021, Spring)**

**Lab 1. Linear Regression Analysis**

Use the following variables from qog\_long.csv to conduct regression analysis to answer the five questions below. You need to write down your answer in a word document and submit it to this page. In this word document, whenever you mention a regression model that you run, make sure to report that regression results by a stargazer table (if there are multiple models, you can either include all of them in a table or report them separately in multiple tables). If you make a plot, make sure to include it in the document as well. At the end of the document, please also paste all the R codes that you use for producing the results. The assignment is due on January 31, by the end of the day.

Variables in qog\_long.csv that you need to use:

* iiag\_gov: overall quality of governance (outcome variable)
* mad\_gdppc: GDP per capita
* pwt\_pop: population in millions
* al\_language: language fragmentation
* bti\_ffe: free and fair election, 10 for the best
* al\_ethnic: ethnic fractionalization

Q1: Built a linear model with all the variables, estimate parameters, and briefly interpret the result (discuss the meaning of each variable’s coefficient and its significance)

Q2: How much variation of the outcome variable is explained by the model?

Q3: In light of the model result, how would the quality of governance score change if free and fair election indicator increases from 4 to 6, holding other variables at their mean values?

Q4: Among those explanatory variables with statistical significance, which one has the biggest impact?

Q5: If we define a country whose bti\_ffe > 6 as “democracy” and “autocracy” otherwise, does the effect of GDP per capita significantly differ between the two groups. Study this question first by visualization with “ggplot2” package and then by a regression model with an interactive term (hint: you need to create a binary indicator for democracy by the “ifelse()” function and use that variable instead of “bti\_ffe” in both visualization and regression parts).

Lab 2: Binary Outcome Models

Use the following variables in the “qog\_long.csv” to complete the following analyses and answer the two questions below. You need to write down your answer in a word document and submit it to this page. In this word document, whenever you mention a regression model that you run, make sure to report that regression results by a stargazer table (if there are multiple models, you can either include all of them in a table or report them separately in multiple tables). If you make a plot, make sure to include it in the document as well. At the end of the document, please also paste all the R codes that you use for producing the results. The assignment is due on February 14, by the end of the day.

* governance quality (iiag\_gov)
* GDP per capita (mad\_gdppc)
* ethnic fragmentation (al\_ethnic)
* free and fair election (bti\_ffe)

Q1: Create a binary indicator for “good governance” by dichotomize the original governance quality score, with “50” as the cutoff (a value higher than 50 will be coded as 1 in the new variable and 0 otherwise). Run a logit model and probit model by using the binary variable as the outcome variable and all other variables as explanatory variables. Report the results by a professional regression table. Which model fits the data better in terms of deviance and AIC/BIC (also briefly explain what deviance and AIC/BIC suggest)?

Q2: Do you have significant coefficients in both models? If so, interpret the effect of that variable in both logit and probit models (make sure to use both “divide-by-4” rule and odds to interpret the coefficient in logit).

**Lab 3: Count Outcome Models**

Load the dataset “smoking.dta” into R and study the number of lung cancer deaths across age groups. Use the knowledge and techniques that you learn to analyze the data as best you can and answer the two questions below. You need to write down your answer in a word document and submit it to this page. In this word document, whenever you mention a regression model that you run, make sure to report that regression results by a stargazer table (if there are multiple models, you can either include all of them in a table or report them separately in multiple tables). If you make a plot, make sure to include it in the document as well. At the end of the document, please also paste all the R codes that you use for producing the results. The assignment is due on February 21, by the end of the day.

Description of variables in the dataset:

* age: five-year age groups coded 1 to 9 for 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80+
* smoking status: coded 1 = doesn’t smoke, 2 = smokes cigars or pipe only, 3 = smokes cigarettes and cigar or pipe, and 4 = smokes cigarettes only
* population: group population (in hundreds of thousands)
* dead: number of lung cancer deaths in a year

Q1: Is smoking related to lung cancer deaths? If so, how strong is the relationship?

Q2: Is smoking cigars/pipes safer than smoking cigarettes? If so, by how much?

Hints:

1. To export a dat. format dataset, use “read.table()” function instead of “read.csv()”

2. Think carefully of what model would be the best choice to answer the two questions. And you might need a different model to answer each question.

3. Consider releveling “smoke” factor to reset the reference category (review “important technique 4” in lab 2 script if you are not sure how to do it) so that the estimated coefficients can be more helpful to answering Questions 1 and 2. Alternatively, dichotomizing smoke variable is also a good idea.

**Lab 4: Categorical/Duration Outcome Models**

Answer questions in Part A and B by regression analysis. You need to write down your answer in a word document and submit it to this page. In this word document, whenever you mention a regression model that you run, make sure to report that regression results by a stargazer table (if there are multiple models, you can either include all of them in a table or report them separately in multiple tables). If you make a plot, make sure to include it in the document as well. At the end of the document, please also paste all the R codes that you use for producing the results. The assignment is due on March 7, by the end of the day.

Part A (Multinomial Regression)

Load “BEPS” (British Election Panel Study from 1997-2001) dataset from “faraway” package. Using the following variables to study voters’ preferences on the three major parties:

* vote (outcome): conservative, labour, or liberal democrat
* age: in years
* gender: 0 for female and 1 for male
* economic.cond.national: assessment of current national economic conditions on a scale from 1 to 5

Q1: How do age and gender affect voters’ preferences on the Labour Party relative to the Conservative Party?

Q2: Does the inclusion of the age variable significantly improve model fit (Hint: consider running a restricted model and then doing a likelihood ratio test)?

Part B (Duration Analysis)

Download the “warchest.csv” from Carmen and load it into R. It came from a research paper by our Professor Janet Box-Steffensmeier. She studies whether a large size of money that incumbent politicians raised for their campaigns (“war chests”) can deter challengers from entering using duration analysis (See “Box-Steffensmeier 1996” on Carmen). In this analysis, the length of time that an incumbent House member ran for re-election without challengers is the duration of interest, while challenger entry is the event that terminated the duration. Use a Weibull model to study this question with the following variables:

* te: weeks of uncontested election run by the House incumbent
* censor: an indicator for challenger entry
* pvote: the incumbent’s proportion of the major party vote in the previous election
* wc: the amount of war chests
* dem: an indicator for Democrat incumbents
* south: an indicator for districts in the South

Q3: Do you think the exponential duration model can also work well in this case? Why?

Q4: Do war chests really deter challengers? By how much?

**Lab 5: Model Evaluation**

Use “Duncan” dataset from “car” package to practice out-of-sample prediction and robust standard error creation. First, perform out-of-sample prediction by randomly dividing the dataset into a training set and a test set (by a ratio of 7/3) and testing the linear model on the “prestige” variable, using all the other variables in the dataset as your explanatory variables. The final result of the out-of-sample prediction should be presented by a scatter plot of predicted values against actual values. Second, run the same regression with all the data and then calculate robust standard errors for your estimates. The final result should be a professional regression table generated by “stargazer” function.

**Lab 6: Missing Data**

Use “Duncan” dataset from “car” package to run the same linear regression model that you did in last week (regress “prestige” on the rest of the variables) for three times. In the first time, you run the model using the original dataset (no missingness); in the second time, you pick three values in the dataset, overwrite them as “NA,” and run the same model by casewise deleting the three missing values; in the third time, you use “mice” package to do multiple imputation for the overwritten dataset (with at least 5 multiples) and compute the final regression result. In the document that you submit, briefly discuss the differences between the three models and your R code and all the regression results (by either copy-pasting or screen-shotting the outputs from the R console). No need to use stargazer tables this time.

**Lab 7: Matching**

Play matching with “BEPS” dataset (British Election Panel Study) in the “car” package. The question that you want to answer is whether Conservative party identity causes the Eurosceptic sentiment. To do so, you need to create a binary treatment variable for Conservative party id based on the “vote” variable. The outcome variable is “Europe,” which is an 11-scale indicator for the degree of Eurosceptic sentiment (higher values represent greater Eurosceptic sentiment). The other covariates that you need to control for in this causal identification include: gender, age, household economic condition (economic.cond.household), knowledge of parties’ positions on European integration (political.knowledge). Match observations on these covariates as best as possible and then use a linear regression to estimate the causal effect of interest. In the document that you submit, you need to use stargazer package to report two models: (1) the linear regression with the raw data (make sure to include all the covariates) and (2) the post-matching linear regression. You also need to briefly compare the two results and give your final conclusion. Attach your R code at the end of the document.

**Lab 8: Hierarchical Model**

Use the “turnout” dataset in panelView package to build a two-way fixed-effects model (with both state and year fixed-effects) and a hierarchical model with varying intercepts by both state and year. Both models should include the following variables:

* turnout (outcome): voter turnout
* policy\_edr (treatment): a binary indicator for the introduction of the EDR
* policy\_mail\_in: a binary indicator for the mail-in registration option
* policy\_motor: a binary indicator for the motor registration option
* abb (for state-level fixed-effects): state abbreviation
* year (for year fixed-effects): a year from 1920 to 2012

Use the stargazer package to report two regression results in a table (hint: recommend to use “lfe” package for the two-way fixed-effects to get rid of fixed-effects estimates in the table), briefly compare the results, and comment on whether you find evidence that the EDR increased voter turnout. Attach your R code at the end of the document.