GV300: Quantitative Political Analysis

Problem Set 7

Due Thursday, December 6, 9.45am on Faser

1. (35 marks) **Ordinary least squares estimator:**

   (a) (15 marks) Derive the ordinary least squares estimator β_1 and its sampling distribution for the population model

   $$y = b_0 + b_1 x + e.$$

   Show every step of your derivation.

   (b) (20 marks) At which steps in your derivation did you make use of the six assumptions discussed in class (see week 9 slides or your preferred text book)? Clearly indicate where you made an assumption and explain in your own words what each assumption implies. You are still encouraged to work in groups but here I really want to see your own words.

2. (30 marks) **Regression analysis, interpretation of coefficients, and model fit:**
   Input the data on campaign spending by incumbent in the 1998 U.S.Senatorial election below into your preferred statistical software.

District	Incumbent	Money	Vote Share
1	Matt Salmon	362	65
2	Ed Pastor	418	68
3	Jim Kolbe	712	52
4	Bob Stump	346	65
5	John Shadegg	426	68
6	J.D. Hayworth	1839	53

   (a) (5 marks) Draw a scatter plot with the data above and add the regression line computed from regressing **V**ote share on **M**oney. For the observation of V and M for incumbent "Bob Stump" indicate on the graph the estimated $\hat{V}_i$ ($\hat{y}_i$), the observed V_i (y_i), and the residual ϵ_i.

   (b) (10 marks) Run a regression of V on the intercept only. Show your results. What does the coefficient estimate represent? Generate a new variable "money.low" which takes on value 1 if $M < 500$ and 0 otherwise. Run a regression of V on money.low. Compute the group-wise means of V of incumbents with low campaign spending vs those with high campaign spending from the regression results. Show your computation.

   (c) (15 marks) Compute, by hand, the sum of squared residuals (SSR), the explained sum of squares (ESS), the total sum of squares (TSS), and R^2 for the regression of V on M. Show your computations. Explain what R^2 tells you about model fit for this particular regression.

3. (35 marks) **Regression analysis and causality:**

   (a) (10 marks) Say we want to know whether higher levels of education causes people to earn higher wages. Generate a 2000 observation dataset. Generate a variable "university" that equals 0 for the first 1000 observations and 1 for the second 1000 observations. This will represent half of the sample attending university. Generate a variable "income" which represents peoples' incomes. Let income = 15,000 + 5,000*university + 1,000*noise where "noise" is distributed standard normal. Regress income on university and show the regression output. What is the coefficient

estimate on university? Why should you have known before you even ran the regression what the coefficient estimate approximately will be? Is this a causal effect?

(b) (10 marks) We now further assume that education has **no** effect on earnings, but that smart people tend to both go to university and earn more money. Clear your dataset and generate a new 2000 observation dataset. Generate 2 variables with uniform distributions between 0 and 1, called "intelligence" and "luck." Generate a variable "university" which equals 1 if intelligence+luck>1 and 0 otherwise. Let income = 15,000 + 10,000*Intelligence + 1,000*noise where "noise" is distributed standard normal. Regress income on university. Show your the regression output. What's your coefficient estimate on university?

(c) (15 marks) Are the two regressions above different conceptually (that is with respect to how the regression enables us to learn something about the world)? Are the two regressions above different mechanically (that is with respect to how we try to get at an unbiased estimate of the true effect of university on income)? Speak to each question in 3-4 sentences.