ME104 Linear Regression Analysis: Problem Set 1

Kenneth Benoit

1. Stata revision using electoral and campaign spending data.

   For this series of questions, we will use the dataset from Benoit and Marsh (2008). After saving this file to your local folder, you can open this file from the menu File/Open, selecting the required file. We have started the file `exercise1.do` for you that we suggest you work from, using the Stata editor.

   (a) Take a look at the data using the commands `browse` and `describe`. Some variables are stored as string variables, (e.g. `district`) while other are numeric (e.g `votes1st`). The `encode` command enables you to convert string variables into numeric and vice versa using the command `encode <varname>, generate (<newvar>)` and `decode <varname>, generate (<newvar>)` . Convert `district` into a numerical variable and `party` into a string variable.

   (b) From the menu `Graphics/Histogram`, inspect the distribution of the values of the variables `votes1st` and `spend_total`, using a histogram and/or kernel density plots. How are these variables distributed?

   (c) Obtain full descriptive statistics on the variables `votes1st` and `spend_total` using the command `summarize <varlist>, detail`. Interpret the meaning of the quantile values: 1%, 25%, 50%, 75%, plus the Std. Dev., Variance, and Skewness.

   (d) Create a table of `wonseat` by `gender`, and do a χ^2 test on the resulting 2×2 table using the command `tab wonseat gender, chi2`. Discuss the results of the χ^2 test.

   (e) Perform a t-test for differences in `spend_total` by `incumbent` using the command `ttest spend_total, by (incumb)`. Discuss the results.

   (f) From the menu `Graphics/Twoway graph(scatter,line, etc.)` create a scatterplot of showing the correlation between `votes1st` on the y-axis and `spend_total` on the x-axis. Add the best fitting OLS line to this plot. Briefly discuss.

   (g) From the menu `Statistics/Linear models and related/Linear regression`, regress `spend_total` on the variables `incumb`, `senator`, and `councillor`. Briefly discuss the results.

2. Additional work

   (a) In country X, 51% of the adults are males. One adult is randomly selected for a survey involving credit card usage.

      i. Find the prior probability that the selected person is a male.

      ii. It is later learned that the selected survey subject was vegetarian. Also it is known that in this country, 11.5% of males is vegetarian, whereas 20.7% of females is vegetarian. Use this additional information to find the probability that the selected subject is a male.

(b) When S successes occur in n trials, the sample proportion $P = S/n$ is generally used as an estimator of the probability of success of π. However, sometimes there are good reasons to use an alernative estimator $P^* = (S+1)/(n+2)$. Alternatively, P^* can be written as a linear combination of the familiar estimator P:

$$P^* = \frac{nP+1}{n+2} = \left(\frac{n}{n+2}\right)P + \left(\frac{1}{n+2}\right) \tag{1}$$

i. What is the MSE of P? Is it consistent?

ii. What is the MSE of P^*? Is it consistent? (Hint: Calculate the mean and variance of P^*, in terms of the familiar mean and variance of P.)

iii. To decide which estimator is better, P or P^*, does consistency help? What criterion would help?

iv. Tabulate the efficiency of P^* relative to P, for example when $n = 10$ and $\pi = 0, .1, .2, \ldots, .9, 1.0$.

v. When might you prefer to use P^* instead of P to estimate π?