Lab #1
Instructor: Todd BenDor

Website: http://todd.bendor.org/wuhan

Lab Preamble:

If you haven't used STATA before, please look at the three handouts on the website for this lab. Additionally, you'll find two datasets that you'll use in lieu of the datasets referenced in the handouts (since they are unavailable from the lab).

For this lab you will use data in a .csv file ('motorbus.csv') available under Lab #1.

In 2006, over 500 US transit agencies reported their passenger trip data to the Federal Transit Administration. The statistics reported includes, among many other variables, (1) the number of unlinked passenger trips taken by bus annually—*passtrips*, (2) the service area population—*svspop*, and (3) average trip fare (in dollars)—*tripfare*.

Here, you will perform three separate regression analyses, which should analyze passenger trips by bus (dependent) as explained by the service area population (independent) and the average trip fare in dollars (independent), as well as a combination of the two.

For each regression:

1. Determine the regression equation to estimate your data.
2. Compute and interpret the slope coefficients (y-intercept and b value) in terms of their type of relationship (for b, inverse/positive), their size, and whether or not they make sense under the assumed relationship.
3. Define, compute, and interpret the Pearson's correlation coefficient (r).
4. Define, compute, and interpret the coefficient of determination (R^2 value).
5. Graph the data and the estimates of passenger trips based on each independent variable (use regression equation and given values of the independent variable to determine the estimated Y values).

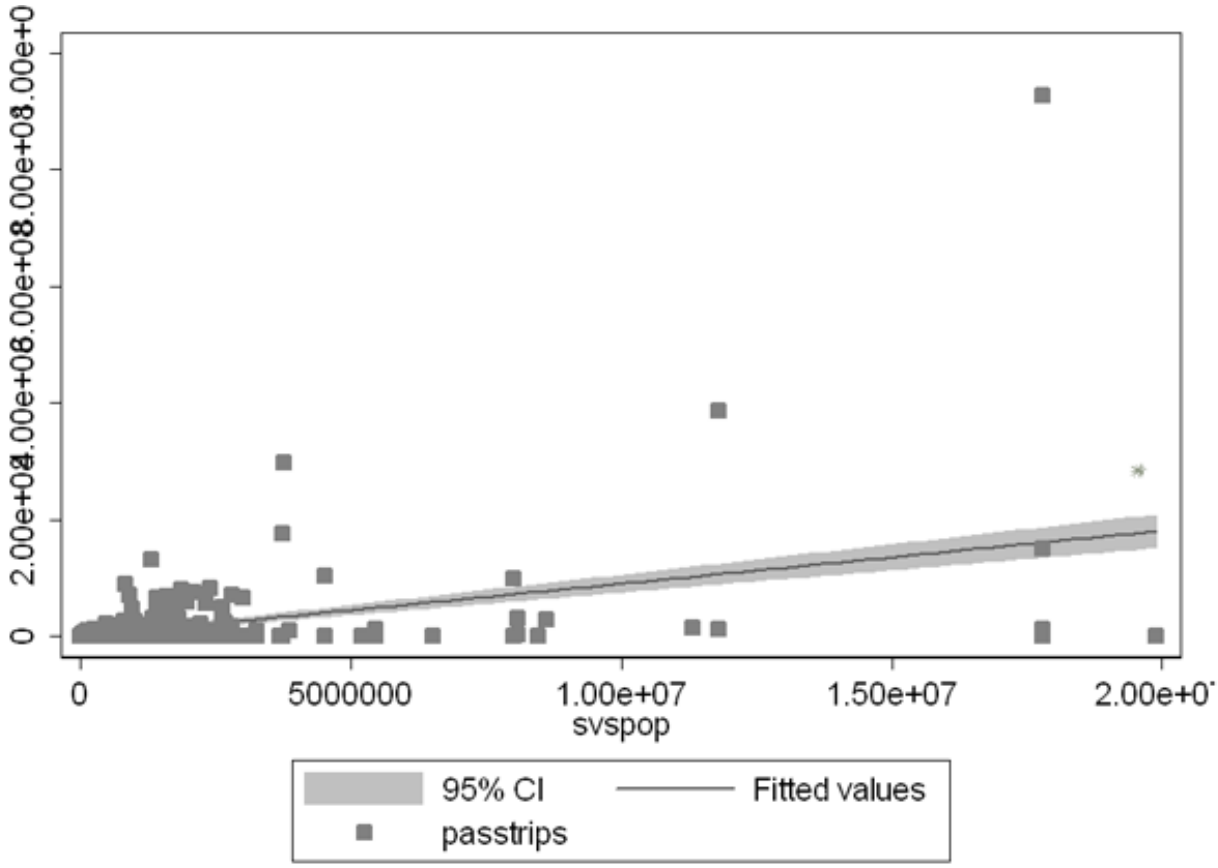
Instructions

To obtain the regression table and results, import the motorbus.csv file into Stata (File -> Import ->import ASCII Data Created by Spreadsheet). Once you've verified that the data has been correctly imported into Stata, plot both independent variables (svspop and tripfare) vs. the dependent variable (passtrips). You can do this by going to "Graphics -> Two-Way Graph." Here you can create a scatter plot and choose the variables you are interested in.

What type of relationship does each of the independent variables have with the dependent variable? Could you visualize a linear model (line) through the dataset? What would this look like? Note the approximate slope and intercept! This is your "guess" at a model. Try a "scatterplot matrix" to get an even better understanding of the distribution of your data.

Now, to run an actual regression, click on "Statistics -> Linear Models and Related -> Linear Regression." Here, you will find a dialog box asking for independent and dependent variable inputs. You should run all three of the regressions specified above.

After each regression, plot the regression fit (Choose "Graphics ->Two-way Graphs->Create->Fit Plots-> Linear Fit with Confidence Interval (CI)" or type: "twoway (lfitci yvar xvar) (scatter yvar xvar)"). Here, you'll see the fitted values (line), along with the actual values and the 95% CI.



What happens in each? What type of R^2 do you get? What is significant?

Does this data fulfill all of the assumptions for linear regression?

(Note that each of the commands you've entered in STATA via the interface can also be typed into the interface.)