Problem set 3: hypothesis testing and model selection

September 16, 2013

1 Introduction

This problem set is meant to accompany the undergraduate econometrics video series on Youtube; covering roughly the 86th video through to the 125th. The focus of this problem set will be hypothesis testing, and how we go about selecting between different models and arriving at an answer in econometrics.

2 2012 Presidential Election Results - practical

1. The results of US Presidential elections provide rich data to work with. Nate Silver in his book, 'The Signal and the Noise', details how he uses statistics (more specifically Bayesian statistics), to assign probabilities to the outcomes of the election results, going on to successfully predict the outcome of the US election in every state. It's well worth a read, especially since it provides some insight into how techniques we use in this class can be adapted to make accurate real world predictions.

   In this problem set we are going to try to emulate Nate Silver in predicting the outcome of the US election. Specifically, we are going to investigate which factors help predict the % of Democrat votes in each state. I have collected and combined a number of disparate data sources which should make for interesting investigation! The dataset is by no means a complete set of indicators, but I suspect it will still allow a relatively fun (and rough) investigation of the factors which affect voting habits. For a discussion of the various sources and pre-analysis done on the data, see below. To get hold of the dataset, download it from `http://www.oxbridge-tutor.co.uk/#!problem-sets/culy`

   - *Democrat* - % of state votes which were for the Democratic nominee Barack Obama. `http://www.presidency.ucsb.edu/`

- *Republican* - % state votes which were for the Republican nominee Mitt Romney. http://www.presidency.ucsb.edu/
- *Church_attendance* - % Who reported attending church, synagogue or mosque, 'at least once a week' or 'almost every week'. Gallup poll. http://tinyurl.com/2wnbfjp
- *Black* - % of state population accounted for by 'African-American Alone' population. 2010 US Census. http://tinyurl.com/82kaf2d
- *Hispanic* - % of state population accounted for by 'Hispanic and Latino population'. 2010 US Census. http://tinyurl.com/3rbhozz
- *GDP* - State GDP in $ Billion. US Government Revenue 2012. http://tinyurl.com/o4hx4of
- *Population* - Millions. US Government Revenue 2012. http://tinyurl.com/o4hx4of
- *GDP per capita* - $. US Government Revenue 2012. http://tinyurl.com/o4hx4of
- *Violence* - rates of violent crime cases reported per 100,000 of populace in 2010. See problem set 1 for a further discussion. FBI.
- *Unemployment* - % estimates of labor force employed in October (chosen as to be just before election) 2012. National Conference of State Legislators. http://tinyurl.com/oq4mg3p
- *Evangelical, Catholic, Mormon, Orthodox, Jehova, Jewish, Muslim, Buddhist, Hindu* - Pew Religious Landscape Survey 2008. % of populace affiliated with that religion. Note that when a '<0.5' is encountered in the table I rounded up to 0.5%. http://tinyurl.com/3cxdxb- Page 100.
- *Mortality* - Infant Mortality Rates (Deaths per 1,000 live births) - 2007-2009. The Henry J Kaiser Family Foundation. http://tinyurl.com/oe5gm25
- *Literacy* - 'Those lacking Basic prose literacy skills include those who scored Below Basic in prose and those who could not be tested due to language barriers.' Institute of Education Sciences 2003. http://tinyurl.com/kvb9a6
- *Med_spend* - Total Medicaid Spending FY 2011. The Henry J Kaiser Family Foundation. http://tinyurl.com/pcmhgaz
- *Med_enr* - 'Total Medicaid Enrollment FY 2011'. The Henry J Kaiser Family Foundation.http://tinyurl.com/q9pxwq2
- *Male* - State population of males. 2010 US Census. http://tinyurl.com/3spa9fp
- *Female* - State population of females. 2010 US Census. http://tinyurl.com/3spa9fp
- *Sex* - Ratio of males to females. 2010 US Census. http://tinyurl.com/3spa9fp
- *Per_*18 - % State population of under 18s. 2010 US Census. http://tinyurl.com/3spa9fp
- *Per_*45 - % State population of under 18-45 year olds. 2010 US Census. http://tinyurl.com/3spa9fp
- *Per_*65 - % State population of those over 65. 2010 US Census. http://tinyurl.com/3spa9fp
- *Med_age* - % Median age of individuals in state. 2010 US Census. http://tinyurl.com/3spa9fp
- *Debt* - Local debt. $ Million. US Government Revenue 2011 (only estimates available in 2012). http://tinyurl.com/o4hx4of

I have consciously chosen to omit the 2008 results in each state as explanatory variables, since our aim here is to investigate the underlying factors which influence voter behaviour rather than minimise error of prediction (in constrast to Nate Silver). Load the dataset into Gretl, then we're ready to start!

(a) The dependent variable we are going to investigate here is *Democrat* - the % of votes for Obama in each State in 2012. A nice way to get a snapshot of how each of the variables in your dataset affects the dependent variable in question is to look at all their respective x-y scatters. There is a quick and easy way to do this rather than looking at each graph individually. Go to 'View' → 'Multiple Graphs' → 'X-Y Scatters...' Then by selecting the dependent variables as *Democrat*, you can choose the other variables you would like to see scatters against. I recommend that you look at scatters for all the variables (or in general in the future if you have many variables, for each class of variable perhaps). This allows you to see quickly whether the effect of a variable is likely particularly strong on the dependent variable, and also allows you to see quickly whether its effect is linear, or likely merits inclusion of a nonlinear term. You may need to do two runs of X-Y scatters since Gretl has a maximum size for the number of plots it can handle in one window. Which variables look the most convincing in their correlation on *Democrat*?

(b) Let's start with a basic model (it is usually best to start with models with all explanatory strong factors in them, but basic models can also be informative at the start). Create the following regression:

$$Democrat_i = \alpha + \beta_1 GDP\_pc_i + \beta_2 Mormon_i + \beta_3 Med\_enr_i + \varepsilon_i$$

This is a start, and allows us to (along with our plots from the previous question), start to test hypotheses we might have had. Strictly we shouldn't really test variables until we have a whole list of potentially relevant explanatory factors in our model, but for our purposes of learning how these tests work we will start here. Which of these variables are statistically significant at the 5% level?

(c) What is the interpretation on each of these coefficients? Compare the strength of each of these explanatory factors.

(d) Omit the factor(s) which the initial model has indicated are statistically insignificant. On the new model can you test for the presence of Heteroscedasity? What choice of tests do we have here? How are they similar/different?

(e) Is the model is functionally well-specified? How could we test whether this is the case? Carry out the test.

(f) There are often considerable benefits to taking both the log of the indepedent and dependent variables. In order to do this we first of all need to create our new logged variables. To do this we are going to use the Gretl console for the first time. This is the interactive command-line environment which allows one to carry out commands dynamically. To access it, go to 'Tools' → 'Gretl console'. This should open a new video with th Gretl console in it. There is a ? next to where you can type commands. For a start we are going to create *ldemo* which is the log of the dependent variable. To do this type 'genr ldemo=logs(Democrat) in the console, then hit return. If done correctly a message should show up saying something like, 'Generated series ldemo'. Here 'genr' is the command to generate a new variable, and logs is the command to take the natural log of a data series. So genr creates a new variable called 'ldemo', which is itself equal to the log of the share of the vote achieved by the democrats in that State. Assuming this has gone ok, go on to create the other new logged variables in the original model using the console. If you navigate back to the main Gretl variables window, you should be able to see the new variables you have just created at the bottom of the list. Note that their description is also filled in, using the equation used to create them. Let's annotate the dependent variable by filling in a description for it. To do this, right click on the *Democrat* variable, and select 'Edit attributes'. Then fill out a description for it, and click ok. It might also be an idea to save your Gretl file now for future use. To do so go to File → 'Save data', then give your file a name, and select the location you wish it to be saved to, then click 'Save'.

(g) We are going to run an OLS regression using the console. To do this we run the command: 'ols Democrat const lgdp_pc lmormon lmed_enr'. The results of the regression should then show up in your console.

(h) What is the interpretation of the coefficient on the variable *lgdp_pc*?

(i) What, on average, would be the effect of a 10% increase in GDP per capita on the % of votes which the Democrats would receive?

(j) Suppose we had reason to believe that, in percentage terms, the effect of Medicaid dependence on voter preference was of the same magnitude, but opposite sign, to the effect of Mormon population size. How might we test this theoretically (do not carry out any tests in Gretl, just summarise the way you might test this)?

(k) Gretl allows us to test restrictions on model coefficients by using the 'restrict' command. We first need to be aware of the form of the restriction we are explictly testing. This is shown algebraically below:

$\beta_{lmormon} = -\beta_{lmed\_enr}$

The 'restrict' option relies on the above relationship being written with coefficients to the left of the equal sign, and values to the right. Writing this out we have that:

$\beta_{lmormon} + \beta_{lmed\_enr} = 0$

In order to test linear restrictions from the console, the restriction must be specified inside a 'restrict' code:

restrict
$\alpha_1 \times b[varname1] + \alpha_2 \times b[varname2] + ... + \alpha_p \times b[varname1] = number$
end restrict

This can be done from the console by entering in the code line by line to the console, but when codes are long, and required frequently it is better to create a script file. These are sections of code which contain a sequence of commands which you want to be passed to the console. Before we create a script file, we are going to change the Working Directory to a file made specifically for our

Gretl files. A Working Directory is a file location which Gretl checks each time you run a command. If the command you have run matches the name of a program, then Gretl will execute it. Create a new file for your Gretl files on the desktop (or anywhere else of your choosing), then from within Gretl go to File → 'Working Directory', and set the Working Directory to be the file which you have just created.

We are now going to create our first Gretl script file. To do so, go to File → 'Script files' → 'New script' → 'gretl script'. A new window should open up, with a blank section where you can enter your code. As a first trial, we are going to get Gretl to run the same OLS regression we have just run. Write the following into the blank window:

ols Democrat const lgdp_pc lmormon lmed_enr

Then click the save icon, then save the file within your newly created Gretl directory. To run the file either type 'run filename.inp' (replace filename with actual chosen file name) at the Gretl console prompt, or click the two cogs at the top of the script window. The results of your original regression should now be outputted in the console window. Now you should be in a position where we can test the restriction on *lmormon* and *lmed_enr* by using the code structure shown above. Go ahead and test the restriction. What do you conclude?

(l) For your information, if you are unaware of a particular command, then you can refer to the command reference pdf by going to help → Command reference. Here there are a list of all possible commands which can be entered at the Gretl command prompt.

(m) We are now going to go ahead and test for heteroscedasticity in our model using a White test. To do so, either from within a script file or at the Gretl console command prompt type, 'modtest - -white'. Here 'modtest' tells Gretl that you are going to carry out a diagnostic test on the model, and '- -white' tells Gretl that the test to be carried out is a White test for heteroscedasticity. Are the results suggestive of any heteroscedasticity? If so, along which variables (or combinations of variables)?

(n) For the variables you found in the last part (not any cross terms if you find them), check graphically for the presence of heteroscedasticity. (To do this you may need to create a series for the residual term by typing series uh = $uhat at the Gretl command prompt). How do you do this? What might be the logic behind there being heteroscedasticity across this particular variable(s)?

(o) Draw a histogram of your residuals from your regression, and test for normality using an appropriate test. What do you conclude?

(p) What are the implications for inference?

(q) Freestyle: come up with a model which best explains variation in the Democratic share of the vote in each State. You can use logged, or non-logged Democratic share as a dependent variable.

3 Theory

2. Consider the following regression conducted on a sample of 100 recent graduates from a university.

$$\widehat{\text{lwage}} = \underset{(0.54)}{1.58} + \underset{(0.23)}{1.24}\,\text{fatheduc} + \underset{(0.45)}{1.46}\,\text{iq} + \underset{(0.98)}{1.09}\,\text{books}$$

$$N = 100 \quad R^2 = 0.37 \quad F(3, 96) = 18.8$$

$$\text{(standard errors in parentheses)}$$

Where *fatheduc*, *iq* and *books* refer to an individual's father's logged education level, their logged score on a standard IQ test and the logged number of books they have read respectively.

(a) What is the interpretation on the coefficient on books?

(b) Are each of the variables significant at the 95% confidence level?

(c) Can you test whether the coefficient on *fatheduc* is significantly different from 1?

(d) One way of testing for the presence of multicollinearity is to carry out a procedure which calculates *variance inflation factors* for each of the independent variables in the regression. The way it works is to regress each of the independent variables on all of the others; reporting a statistic which is related to the R-squared from that regression. The exact form of this statistic is shown below.

$$VIF_j = \frac{1}{1 - R_j^2}$$

Where R_j^2 refers to the R-squared from the regression of the j-th independent variables on all of the others. How does this test work?

(e) The results of the variance inflation test for multicollinearity are shown below for this model.

Variable	**VIF**
fatheduc	3.12
iq	4.56
books	10.18

What does this suggest about the inclusion of each of these variables in the above regression?

(f) What would happen to the coefficient on *fatheduc* if *books* was omitted from the above regression? Why would this happen?

(g) Interpret the R-squared from this regression.

(h) The following regression is now run by the researcher.

$$\widehat{\text{lwage}} = \underset{(0.23)}{2.12} + \underset{(0.21)}{1.35}\,(\text{fatheduc+iq}) + \underset{(0.47)}{0.23}\,\text{iq} + \underset{(0.74)}{1.13}\,\text{books}$$
$$N = 100 \quad R^2 = 0.35 \quad F(3,96) = 17.7$$
$$\text{(standard errors in parentheses)}$$

What can you conclude from this regression as to the strength of the effect of father's education and iq?

3. A researcher runs an initial regression for a sample of 1,000 SME companies.

$$\widehat{\text{Sales}} = \underset{(5.34)}{70.12} + \underset{(5.62)}{15.31}\,\text{Advertising} + \underset{(3.54)}{11.76}\,\text{Consumer} + \underset{(2.11)}{10.23}\,\text{Years}$$
$$N = 1,000 \quad R^2 = 0.57 \quad F(3,96) = 18.8$$
$$\text{(standard errors in parentheses)}$$

Where *Sales* is the 2012 annual $ (in thousands) revenue, *Advertising* the monetary amount spent on advertising that year, *Consumer* is a dummy variable indicating the value 1 if the products/services are for consumers, and *Years* is the number of years since the company was founded.

(a) What is the interpretation of the coefficient on *Advertising*? Why might interpreting this coefficient as the return for every pound spent be misleading?

(b) What is the interpretation of the coefficient on *Consumer*?

An auxillary regression is carried out by researchers, using the square of the residuals from the regression shown above as a dependent variable. The results of this regression are shown below.

$$\widehat{\text{Residuals}^2} = \underset{(1.01)}{0.34} + \underset{(1.06)}{1.23}\,\text{Advertising} + \underset{(0.56)}{0.45}\,\text{Consumer} + \underset{(0.23)}{1.90}\,\text{Years}$$
$$N = 1,000 \quad R^2 = 0.14 \quad F = 4.55$$
$$\text{(standard errors in parentheses)}$$

(c) What conclusions can be made on the basis of the above results? Why might there be a theoretical reasoning for this?

(d) What are the problems posed for Ordinary Least Squares estimators, on the basis of what you found in the last part?

(e) A new regression is run with log(sales) as an independent variable. The Breusch-Pagan test for heteroscedasticity produces an R-squared very close to zero. Why might this be?