Econ 21410 - Problem Set IV

Function Approximation and Working with Data

April 22, 2015

This homework should be done in LaTeX. The homework will be graded on correctness, but will also heavily weight clarity and professionalism. Being able to produce clean, clear, well documented write-ups and code is an important skill which will be rewarded. Its better to not do some of the harder parts than to turn in an incomprehensible document. Your R script as well as a log-file should be submitted. Alternatively, use knitr to print your code inline in your latex document.

Make sure to write code which is clear and flexible. Read the whole problem before you begin coding. Some parameters will change and the code should be written in a way to make this easy to implement. We will re-use code in this course. Flexibility and documentation now will save you headaches later in the quarter. Remember to properly indent your code!

SUBMISSION: The homework must be emailed to Oliver and myself by 9:30a.m. **Thursday**, April the 30th. The email must include a pdf with the filename lastname_pset4.pdf and R code called lastname_pset4_code.R where "lastname" should be replaced with your last name. The subject of your email should be [ECON 21410: pset4 submission].[1]

1 Function Approximation Warm up.

In this exercise you will try to predict earnings for individuals using the Current Population Survey. I will not overly specify which variables you should extract and include, but rather will leave this up to you as a researcher.

Go to ipmus.org and browse the CPS data (cps.ipums.org/) for 2004 and 2014. We will be trying to predict INCWAGE, i.e. "wage and salary income". Explore the data and download data on:

- Education
- Sex
- Race / Ethnicity
- Age
- Region of the country

[1] Note that homework this week will be due thursday, but the next pset will still be posted on Wednesday, so plan accordingly

- Any other variables that you think may be helpful in predicting wages (note the variables should exist for both 2004 and 2014)

Now, preform the following:

1. Using the CPI, adjust the two income variables to be in the same units (i.e, adjust the 2004 income for inflation so that it is in 2014 dollars). This is not in the CPS and you will need to find it and figure out how to use it.

2. Make a new varible that is "log wage income" in your data. This will be the endogenous variable we are trying to predict.

3. Construct "potential experience" which will be "Age - years of schooling - 5". Which will be the approximate number of years the individual could have potentially been in the labor force

4. Make a table (or tables) (use the stargazer or texreg package) comparing the following regressions for 2004 and 2014. Discuss and compare your results.

    - Log wages regressed on years of schooling, experience, and experience squared.

    - The same regression as above, but only for men.

    - The same regression above, but only for women.

2 Function Approximation

The goal of this section is to use the tools covered in class to write a function to predict log wages as well as possible. Using the "lm" command as well as tools like "poly()" and "bs()" in the "splines" package, produce a function that "best fits" out-of-sample mean squared error (i.e. $\frac{1}{N}\sum_i^N (Y_i - \hat{Y}_i)^2$). Make sure to also consider auxillary variables, interactions between variables, etc.. We will use only the 2014 data to begin with. Specifically:

1. Randomly select 1/5th of the observations in the 2014 sample. Estimate models on the other 4/5th of the data and try to find a model that minimizes the mean squared error in the 1/5th of the data you are not using for estimation. Discuss what your model depends on, show the code you use to run the model and to test the mean squared error. Discuss how you solved the problem and what you found. Note, you should write your own code for splitting the sample and estimating the mean squared error.

2. Come up with a model that overfits the data. Show that it has a lower mean-squared error on the 4/5ths of the data you used to estimate than the model you found above, but higher "out of sample" mean squared error on the 1/5th of the data not used to estimate the model.

3. Find an R package to perform cross-validation. Use it to peform "K-fold cross validation" where K is equal to 5. You may need to make some choices on what options to choose or which package to go with, which is a natural part of the research process. Report "how well your model fits the out-of-sample data" and see if you can improve your fit from what you found in part 1. Report your final results, and show your code. Discuss how/if your findings changed.

3 Exploring Data

This is a very open-ended assignment. Using either the CPS or the ACS (also on ipums.org), find an interesting fact in the data.[2] This could be about how a particular varible changes over time, differences between groups in some particular behavior, differences between regions in some outcome, or many many other facts.

Make a short report describing the interesting fact you found. It should involve details on how you cleaned your data, what data you used, and a discussion of your fact and why it is of economic interest. Discuss your finding and see if you can find any related research. The report should include atleast 2 figures and 3 tables. These tables and figures should have clear titles and be well labeled, well documented, and well discussed. The report should look like something you would see in a journal or a final product you would send your boss that lays out the results and explains them clearly.

For this section you will be graded based on coming up with an interesting finding, but the bulk of the grade will be clear explanation and display of this finding.

[DISCLAIMER: It is easy to get carried away here. Do not spend 20 hours fretting over if your fact is interesting enough. The goal here is to explore data and make some professional looking output – though hopefully you also find an interesting fact.]

4 Research

Next week part of the homework assignment will be to submit a research proposal. While it is fine to still be working out details, you should aim to have some sort of research question by then and some idea of what data you may use to answer it. If you are struggling with this, please come talk to Oliver and I during our office hours.

Side Projects

- Program some or all of assignment 1, 2, or 3 in C++ using Rcpp or in Julia (3 points). NOTE, YOU MAY ONLY REDO ONE PROBLEM SET PER LANGUAGE. So if you have already done one Julia pset, you must switch to C++ to do another.
- If you scored lower than 70 on one of your first 3 assignments, write a 2-3 page report documenting was wrong with your code, using the solutions fix your code so that is works properly. (up to 3 points each).
- Find a news article and lay out the economic arguments being made by the author and how they could potentially be tested with data in 1 to 2 pages (up to 2.5 points).
- Attend any economics seminar listed here http://economics.uchicago.edu/workshops/ , and write a 1-3 page summary of what the main idea was behind the paper being presented and the questions or concerns that were raised by the audience (up to 3 points).
- Old side-projects remain open unless they had a stated time-limit. ONE UPDATE, YOU MAY ONLY REDO ONE PROBLEM SET PER LANGUAGE.

[2] As a warning, the ACS is very large, so it may be best to use the CPS