Final Exam

36-402, Advanced Data Analysis

Due at 10:30 am on 13 May 2013

This exam has two parts, involving different data and (potentially) different techniques. Points are as indicated.

You can use your notes, the textbooks, and indeed anything you find in the library or online, if that is properly acknowledged. However, **all your work must be your own**. You cannot work with classmates, friends, a tutor, or anyone else. If you are unclear about what is allowed and what is not, please check the university policy on cheating and plagiarism (`http://www.cmu.edu/policies/documents/Cheating.html`), or ask the professor.

Please include the following text in your write-up:

> I, YOUR NAME, have completed this examination honestly, without giving prohibited assistance to anyone, or receiving it from anyone.

A: Strikes (50 points total)

Finding the factors which control the frequency and severity of strikes by organized workers is an important problem in economics, sociology and political science[1]. Our data set, `http://www.stat.cmu.edu/~cshalizi/uADA/13/exams/3/strikes.csv`, kindly provided by a distinguished specialist in the field, contains information about the incidence of strikes, and several variables which are plausibly related to that, for 18 developed (OECD) countries during 1951–1985:

- Country name (not to be used unless indicated otherwise)
- Year (also not to be used unless indicated otherwise)
- Strike volume, defined as "days [of work] lost due to industrial disputes per 1000 wage salary earners"
- Unemployment rate (percentage)
- Inflation rate (consumer prices, percentage)
- "parliamentary representation of social democratic and labor parties". (For the United States, this is the fraction of Senate seats held by the Democratic Party.)
- A measure of the centralization of the leadership in that country's union movement, on a scale of 0 to 1.
- Union density, the fraction of salary earners belonging to a union (only available from 1960).

Note that some variables are missing (`NA`) for some cases. You will need to handle NAs in some sensible way, and explain what that way is.

1. (5) Use `pc()` from `pcalg` to obtain a graph, assuming all relations between variables are linear. Report the causal parents (if any) and children (if any) of every variable. If the algorithm is unable to orient one or more of the edges, report this, and in later parts of this problem, consider all the graphs which result from different possible orientations.

   Note: See `http://bactra.org/weblog/914.html` for help with installing `pcalg`. The most troublesome component is the `Rgraphviz` package. If you are unable to get `Rgraphviz` to work, you can still extract the information from the fitted model returned by `pc`: if that's `pc.fit`, then `pc.fit@graph@edgeL` is the "edge list" of the graph, listing, for each node, the nodes it has arrows *to*. With this information, you can make your own picture of the DAG.

2. (10) Linearly model each variable as a function of its parents. Report the coefficients (to reasonable precision), the standard deviation of the regression noise (ditto), and 95% confidence intervals for all of these, as determined by bootstrapping the residuals.

[1] Or it used to be, anyway.

3. (10 total) You should find that strike volume and union density are not connected, but that there is at least one directed path linking them — either density is an ancestor of strike volume, or the other way around.

   (a) (5) Find the expected change in the descendant from a one-standard-deviation increase in the ancestor above its mean value.

   (b) (5) Linearly regress the descendant on all the other variables, including the ancestor. According to this regression, what is the expected change in the descendant, when the ancestor increases one SD above its mean value and all other variables are at their mean values?

4. (15 total) Check the linearity assumption for each variable which has a parent. (Putting in interactions and/or quadratic terms is inadequate and will result in only partial credit at best.)

   (a) (5) Describe your method, and why it should work.

   (b) (5) Report the p-value for each case, to reasonable precision.

   (c) (5) What is your over-all judgment about whether it is reasonable to model each endogenous variable as linearly related to its parents? If you need more information than just p-values to reach a decision, describe it.

5. (10) Discuss the over-all adequacy of the model, on both statistical grounds (goodness-of-fit, appropriateness of modeling assumptions, etc.) and substantive, scientific ones (whether it makes sense, given what is known about the processes involved).

B: Macroeconomic Forecasting (50 points total)

The data set `macro.csv` on the class website contains five standard macroeconomic time series for the United States, from the beginning of 1948 to the beginning of 2010: total national income or GDP; value of goods consumed; investment spending; hours worked; and output per hour worked for all non-financial firms. (Some of these series are in inflation-adjusted dollars, some of them are in hours, and some of them are indexes where a particular date has been set as 100 and others are expressed relative to that.) All variables are measured "quarterly", i.e., four times a year.

Most macroeconomic forecasting models do not concern themselves directly with these values, but only with the logged fluctuations around their long-run trends.

For full credit on the modeling questions, you must use models which go beyond those available in 401, *or* you must use appropriate methods to show that linear model are justified here.

It is first necessary to remove trends; macroeconomists traditionally do this with the following function.

```
hpfilter <- function(y, w=1600){
 eye = diag(length(y))
 d = diff(eye,d=2)
 ybar = solve(eye + w*crossprod(d), y)
 yhat = log(y) - log(ybar)
 return(list(fluctuation=yhat,trend=ybar))
}
```

1. (5) Create five plots, showing each of the variables and its trend (as returned by `hpfilter`) as functions of time. Use a logged scale for the vertical axis. Report R^2, with and without logging, for each of the five trends.

2. (5) Plot the logged fluctuations around trend (as returned by `hpfilter`) for each of the five variables. Does it make sense to compare these fluctuations across variables? Do the fluctuations look stationary? — After this problem, references to the variables always mean their logged fluctuations around their trends.

3. (5) Are the variables Gaussian? (You can do better than looking at a histogram.)

4. (10) For the first four variables (GDP, consumption, investment, hours worked), fit an additive regression of each variable on the values of all four at the previous time-step. Use only data up to, but not including, 2005 ("the training period"). Report the mean squared error on the training data (to reasonable precision), and include plots of the partial response functions. Describe, in words, what the partial response functions say about the relations between these variables.

5. (10 total) Using the circular block bootstrap, with blocks of length 24, generate new time series which are as long as the training data.

   (a) (2) Write a function to calculate the mean squared errors of the fitted models from Problem 4, on a time series. (Each of the four variables should have its own MSE.) Check that it works by making sure that it gives the right answer for the training data.

   (b) (3) Report the mean MSEs, and the standard error of these means, from enough bootstrap replicates that the standard errors are no more than 10% of the means.

   (c) (5) What do you need to assume for the numbers from 5b to be good estimates of the generalization error of this model?

6. (10 total) "Real" (as opposed to "monetary") business cycle theories hold that fluctuations in macroeconomic variables are ultimately caused by exogenous "real shocks", especially changes to productivity. The productivity variable in `macro.csv` is a measurement of this variable, which, according to these theories, should be exogenous. The other variables, in such theories, are endogenous.

   (a) (5) Fit an model for each of the four endogenous variables, as an additive function of the endogenous variables in the previous quarter, and productivity for the previous four quarters. Report the MSEs and include plots of the partial response functions. Compare the plots to those in Problem 4.

   (b) (2) Describe a method which could be used to decide whether including productivity in this way really improves predictive performance. Discuss the assumptions of the method, and why you think they apply here.

   (c) (3) Implement your method. For which variables does including productivity actually help? How confident are you of this conclusion?

7. (5 total) Now consider the period 2005–2010. What are the mean squared errors, on this data, of

   (a) (2) Predicting according to the additive model from Problem 4?

   (b) (2) Predicting according to the additive model from Problem 6?

   (c) (1) Predicting the mean of each variable, as estimated from the training period?

8. (5, extra credit) Explain how what `hpfilter` does is related to spline smoothing.