MODEL SELECTION EXERCISE

In this exercise we try to model the employment to population ratio in the US. In the most recent recovery from recession, the employment to population ratio has remained lower than usual at this stage of a recovery. Its behavior is one reason that the Fed considers the labor market to be performing less well than might be suggested by the current unemployment rate.

The employment to population ratio normally does fall when the unemployment rate goes up, not just because people have become unemployed, but also because people who might otherwise look for work don't look for jobs when the prospects of successful job search look poor. You might also expect that when the wage is low, people might decide not to work. And finally, the US population has been aging, and older people are more likely to be retired.

The course web site has links to quarterly data on average hourly earnings (`wage`), the employment population ratio (`empratio`), the ratio of population 25 to 54 years of age to the total population 16 and over (`primeratio`) and the unemployment rate (`unrate`). The data were downloaded from FRED, the St. Louis Fed's internet database. They are available as separate files, one for each variable, in comma-separated value text (.csv) and Microsoft Excel (.xls) formats. Those files include headers describing exact data definitions and sources. These individual variable files do not include `primeratio`. That has to be constructed from data on the population 16 and over and population ages 25-54.

There is also an R data file, `psdata.RData`, that has the four series as an R multiple time series object. The file `psdata4.RData` contains the four series, plus for each series four lags of the series. The names of the series in `psdata4.RData` are, for example, `emprate`, `emprate1`, and `emprate2` for the employment to population ratio and its first two lags.

The `psdata.RData` file includes series whose time spans do not match perfectly, so it contains NA values for some dates and variables at the start and end. The `psdata4.RData` file has trimmed the series so thata all NA values are eliminated, meaning that its start and dates are not the same as for the `psdata.RData` file. You can experiment with models including different variable lists and numbers of lags while maintaining a consistent sample across models in R if you estimate models with commands like

```
lsout2 <- lm(emprat ~ wage + primeratio + unrate + emprat1, data=psdata4) .
```

The `data=psdata4` argument tells the regression program to look for variables by name within the `psdata4` multiple time series object.

[If you should want to explore longer lag lengths using R, you will want to construct an analogue of `psdata4` including more lags. A utility program that does this is on the website as `lagts.R`. It calls `trimts.R`, which is also on the web site.

Note that you are free to use any program you like to do this exercise. I just put more effort into making it straightforward with R.

Estimate [at least — you can try others if you like] the following regressions. All have `emprat` as dependent variable; the list specifies the right-hand-side variables.

(i) The three other variables, with no lags.
(ii) The same regression, but with all variables differenced once. (In R, this will mean replacing the `data=psdata4` argument to `lm()` with `data=trimts(diff(psdata))`)
(iii) The three other variables, plus two lags of `emprat`.
(iv) Just three lags of `emprat`, no other variables.
(v) Two lags of `emprat` plus the three other variables and one lag of each of those three.
(vi) Three lags of `emprat` and current and three lagged values of each of the other three.

(1) For each model (other than (iv)), use F tests for whether each of the three explanatory variables belongs in the model.
(2) For each model, check for serial correlation in residuals (`acf(lsout$residuals)` in R) and for non-normality (`hist(lsout$residuals)` and `qqnorm(lsout$residuals)` followed by `qqline()`).
(3) Compare these models, to the extent you can, using standard F-statistics. [In R, you can compare two *nested* models by applying `lm()` to estimate each, as, say, `ls1` and `ls2`, then calling `anova(ls1,ls2)`. This works also for a nested sequence longer than 2.]
(4) Compare the models using the Akaike criterion and BIC.
(5) Calculate a posterior distribution over the models. This will require putting a prior on the models and also putting a prior on parameters within the models. Further guidance on how to do this tractably will follow.
(6) Discuss the differences, if any, in the implications of the different model choice criteria for what you would use in prediction. What about if your use of the model was not prediction, but rather deciding whether the decline in the employment to population ratio was mainly non-cyclical, and therefore not a reliable indicator of labor market slack for Fed decisions?