Instructions: You must submit this homework by pushing the "hw4.Rmd" file to your team's repo. Note that that is the **only** file you will be allowed to push. Commit early and often.

Scenario

"Gross domestic product" is a standard measure of the size of an economy; it's the total value of all goods and services bought and solid in a country over the course of a year. It's not a perfect measure of prosperity[1], but it is a very common one, and many important questions in economics turn on what leads GDP to grow faster or slower.

One common idea is that poorer economies, those with lower initial GDPs, should grower faster than richer ones. The reasoning behind this "catching up" is that poor economies can copy technologies and procedures from richer ones, but already-developed countries can only grow as technology advances. A second, separate idea is that countries can boost their growth rate by under-valuing their currency, making the goods and services they export cheaper.

This data set contains the following variables:

- Country, in a three-letter code (see http://en.wikipedia.org/wiki/ISO_3166-1_alpha-3).
- Year (in five-year increments).
- Per-capita GDP, in dollars per person per year ("real" or inflation-adjusted).
- Average percentage growth rate in GDP over the next five years.
- An index of currency under-valuation[2] The index is 0 if the currency is neither over- nor under- valued, positive if under-valued, negative if it is over-valued.

Note that not all countries have data for all years. However, there are no missing values in the data table.

Load the data with

```
uv <- read.csv("http://www.stat.cmu.edu/~cshalizi/uADA/16/hw/02/uv.csv")
```

Part I: Linear models

- Linearly regress the growth rate on the under-valuation index and the log of GDP. Report the coefficients and their 95% confidence intervals (to reasonable precision). Do the coefficients support the idea of "catching up"? Do they support the idea that under-valuing a currency boosts economic growth? That is, interpret your results in the context of the problem
- Repeat the linear regression but add as covariates the country, and the year. Use `factor(year)`, not `year`, in the regression formula.
  - Report the coefficients for log GDP and undervaluation, and their 95% confidence intervals, to reasonable precision (only these 2, not all the others). Does this expanded model support the idea of catching up? Of under-valuation boosting growth?
  - Explain why it is more appropriate to use `factor(year)` in the formula than just `year`.
  - Plot the coefficients on year versus time.

[1] A standard example: if vandals break all the windows on a street, for that town, GDP goes *up* by the cost of the repairs.

[2] The idea is to compare the actual exchange rate with the US dollar to what's implied by the prices of internationally traded goods in that country — the exchange rate which would ensure "purchasing power parity". The details are in the paper this assignment is based on, which will be revealed in the solutions.

- Does adding in year and country as covariates improve the predictive ability of a linear model which includes log GDP and under-valuation?
  - What are the R^2 and the adjusted R^2 of the two models?
  - Use leave-one-out cross-validation to estimate the prediction risk mean squared of the two models. Which one actually predicts better, and by how much? **Hint:** Use the code from Chapter 3.
  - Explain why using 5-fold cross-validation would be hard here. (You don't need to do it.)
- For the best predicting model, use the bootstrap (see Chapter 6) to produce confidence intervals for log GDP and undervaluation. You should try to do both the *resample rows of the data* version and the *resample residuals* version (pp. 156-7). Can you do both? Why or why not? Compare the bootstrapped confidence intervals with the intervals you already presented. Do your conclusions change? Use $B = 250$.

Part II: Kernel smoothing

- Use kernel regression, as implemented in the `np` package, to non-parametrically regress growth on log GDP, under-valuation, country, and year (treating year as a categorical variable). **Hint:** read Chapter 4 carefully. In particular, try setting `tol` to about 10^{-3} and `ftol` to about 10^{-4} in the `npreg` command, and allow several minutes for it to run. (You should really make sure to cache this part of your code.) **Also important**: use `results="hide"` as a chunk option to keep R from printing a bunch of garbage in the knitted version.
- Give the coefficients of the kernel regression, or explain why you can't.
- Plot the predicted values of the kernel regression against the predicted values of the linear model.
- Plot the residuals of the kernel regression against its predicted values. Should these points be scattered around a flat line, if the model is right? Are they?
- The `npreg` function reports a cross-validated estimate of the mean squared error for the model it fits. What is that? Does the kernel regression predict better or worse than the linear model with the same variables?