Homework 6: What Makes the Union Strong?

36-402, Advanced Data Analysis, Spring 2012

Due at 10:30 am on Tuesday, 28 February 2012

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/12/hw/06/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
- Year
- 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 Congressional 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[2].
- 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.

We will use this data set to practice the techniques we have learned so far, in preparation for the first midterm. (The midterm *may* include any topic covered in the class to date, whether or not it shows up in this assignment.)

[1] Or it used to be, anyway.

[2] This measure really should be a constant for each country over the period, but having a variable with only 8 levels is trouble for the spline smoother used in Problem 3, so a very small amount of artificial noise (± 0.005 at most) has been added to each value.

1. Estimate a linear model to predict strike volume in terms of all of the other variables, *except* country and year.

   (a) Report the coefficients, with 90% (not 95%) confidence intervals calculated according to

      i. (2) The standard formulas
      ii. (9) Resampling of the residuals
      iii. (9) Resampling of the cases

      Do not use more digits than you can justify.

   (b) (10) Describe the meaning of the coefficients *qualitatively*. (I.e., do not write "A one unit change in foo produces a change of bar units in strike volume" over and over.)

   (c) (5) Rank the predictor variables from most to least important, with "importance" measured by the magnitude of the predicted change to strike volume in response to a 1% relative change of the predictor away from its mean value.

   (d) (5) Rank the predictor variables from most to least important in terms of predicted response to a 1 standard deviation change in the variable.

   (e) (5) Do the two rankings agree? Should they? Which one seems more reasonable for this problem?

2. Some theories suggest that English-speaking countries have legal and political institutions which make strikes operate differently than in other industrialized countries. Figure out which countries in the data set are primarily English-speaking, create an indicator (dummy) variable for whether a case belongs to one of those countries, and add it to the data set.

   (a) (5) Fit a linear model in which the predictors from Problem 1 interact with the English-using variable. Report the new coefficients (to *reasonable* precision)

   (b) (5) Explain how (if at all) this model differs qualitatively from the model in Problem 1.

   (c) (5) Use five-fold cross-validation to compare this model to the model in Problem 1. Which one does better?

3. Fit an additive model for strike volume as a smooth function of all the variables except country and year.

   (a) (5) Plot all the partial response functions. Do they agree qualitatively with the conclusions you drew from the model in Problem 1?

   (b) (5) Consider increasing each of the predictor variables by 1% from its mean, leaving the other variables alone. Rank the predictors according to the magnitude of this model's predicted change in strike volume. Would the ranking be the same for a 1% decrease? *Hint:* use `predict` and a data frame with artificial data.

(c) (5) Consider increasing each of the predictor variables by one standard deviation from its mean, leaving the other variables alone. Rank the predictors according to the magnitude of this model's predicted change in strike volume.

(d) (5) Discuss the contrast (if any) between these rankings, and the corresponding ones for the linear model.

4. (10) Use the methods of Chapter 10 to test whether the linear model from Problem 1 is well-specified against an additive alternative.

5. *Continuing past the training data*

   (a) (2) What were the values of unemployment, inflation, union density, and `left.parliament` for the United States in 2009? *Hint*: You can get most of these from the last *The Statistical Abstract of the United States*.

   (b) (4) Assuming the union centralization variable for the US in 2009 was 0, what strike volume was predicted by (i) the model from problem 1, (ii) the English-is-different model from problem 2, and (iii) the additive model from problem 3?

   (c) (4) The actual strike volume for the United States in 2009 was 0.8. Is this plausible under any of the models? *Hint:* How much do you expect actual values to differ from predicted values?