POLS/CSSS 503:
Advanced Quantitative Political Methodology

Problem Set 4

Professor: Chris Adolph, Political Science and CSSS

Spring Quarter 2014

Due in class, 20 May 2014

General instructions for homeworks: Homework can be handwritten or typed. For any exercises done with R or other statistical packages, you should attach all code you have written and all (interesting) output. Materials should be stapled together in order by problem. The most readable and elegant format for homework answers incorporates student comments, code, output, and graphics into a seamless narrative, as one would see in a textbook.

Bonus Problem: The Effect of Measurement Error on Linear Regression

Just as omitting a control from a regression can cause bias, so to can including a control which is measured with error—even if that error is random. Read Section 6.4 of Fox carefully, and then solve the following problems:

a. Fox problem 6.10 (page 117)

b. Fox problem 6.11

c. Fox problem 6.12

d. Fox problem 6.13

Problem I: Showing Confidence

We revisit the sprinters data we considered in Problem Set 2.

a. Add confidence intervals to the plots you made for Problem Set 2, problem 1d. using the `predict()` command to generate the confidence intervals around the expected finish times.

b. Rerun the analysis and recreate the plot, adding confidence intervals, for the model:

$$\log(\mathrm{Finish_i}) = \beta_0 + \beta_1 \mathrm{Year}_i + \beta_2 \mathrm{Women}_i + \beta_3 \mathrm{Year}_i \times \mathrm{Women}_i + \epsilon_i$$

Be sure to explain in words how this specification differs from the one used in part **a.**

c. Rerun the analysis and recreate the plot, adding confidence intervals, for the model:

$$\begin{aligned} \mathrm{Finish_i} &= \beta_0 + \beta_1 \mathrm{Year}_i + \beta_2 \mathrm{Women}_i + \beta_3 \mathrm{Year}_i \times \mathrm{Women}_i \\ &\quad + \beta_4 \mathrm{Year}_i^2 + \beta_5 \mathrm{Year}_i^2 \times \mathrm{Women}_i + \epsilon_i \end{aligned}$$

Be sure to explain in words how this specification differs from the ones used in part **a.** and **b.**

d. Compare the visual fit of these models to the data within the observed period. Which do you find plausible fits?

e. Do these models have different predictions for the Olympics of 2156? (*Hint:* extending your plots to go up to 2156 is an easy way to see this.) Why or why not?

f. Now create a new variable, the ratio of men's time to women's time in each year. Logit-transform this variable and regress it on year. Plot the results, with confidence intervals, on the scale of the ratio men's time to women's time (i.e., transform it back from logit). Does this approach make any assumptions about men's times or women's times that might be problematic?

Problem 2: Model Selection: Oil & Democracy

For this problem, we will use a cleaned-up version of the dataset employed by Michael Ross in "Does Oil Hinder Democracy?" *World Politics*, 2001. In that paper, Ross estimated a time series cross-section model of Polity scores regressed on oil exports and a battery of controls. In this problem, we will focus on a single cross-section (saving the time series cross-section analysis for a later optional homework), and instead focus on model fitting and robustness to outliers. A description of the included variables follows:

Variable	Description
`regime1`	1–10 scale increasing in democracy; computed from Polity components
`oilL5`	Fuel exports as a proportion of GDP, lagged 5 years
`metalL5`	Ore and mineral exports as a proportion of GDP, lagged 5 years
`GDPpcL5`	per capita GDP in PPP dollars, lagged 5 years
`islam`	Muslims as a proportion of population, 1970 data
`oecd`	dummy for rich industrialized countries
`cty_name`	the name of the country observed
`id`	a three character abbreviation of the country name
`id1`	a numeric country code
`year`	the year of the observation (for this slice, it is always 1995)

We will use this dataset and the baseline regression from Ross (2001) to explore our tools for specifying regression models and improving fit.

a. Load the dataset `ross95.csv`, which contains a partially cleaned cross-section of replication data for the year 1995. Estimate a linear regression of `regime1` on `oilL5`, `metalL5`, `GDPpcL5`, `islam`, and `oecd`. Record the standard error of the regression, and calculate the expected change in `regime1` given a change in `oilL5` from the 50th percentile to the 95th percentile of the fully observed data, all else equal.

b. Using the residuals from the regression in part **a.**, create the following diagnostic plots: **(i)** plot the residuals against the fitted values, **(ii)** plot the residuals against each covariate, **(iii)** plot the studentized residuals against the standardized hat-values. What do these diagnostics tell you about the presence of heteroskedasticity, specification error, and outliers?

c. Rerun the regression using either log or logit transformations on any covariates you see fit. You will likely run several specifications. In each run, record the standard error of the regression, and the expected change in `regime1` given a change in `oilL5` from the 50th percentile to the 95th percentile of the fully observed data. See the appendix for some tips and warnings about transforming these data, though.

d. Rerun all your models using robust regression with an M-estimator. In each run, record the standard error of the regression, and the expected change in `regime1` given a change in `oilL5` from the 50th percentile to the 95th percentile of the fully observed data.

e. Rerun all your models using robust and resistant regression with an MM-estimator. In each run, record the standard error of the regression, and the expected change in `regime1` given a change in `oilL5` from the 50th percentile to the 95th percentile of the fully observed data.

f. How much *substantive* difference does finding the best model make? (Be specific and concrete; i.e., show what each model does. I'm asking for a more detailed answer than you usually see in articles.) How much substantive doubt is there in the result if we are not sure which of the models you fit is the "right" one?

g. Which model of those you have estimated do you trust most, and why? What other problems in the specification or estimation method remain unaddressed by our efforts?

Appendix: How Do I Log a Covariate with Zeros? · *Christopher Adolph*

If you try to log or logit transform a covariate x with observed zeros, you will discover a problem: you can't log a zero! A common (but wrong) "solution" is to add a small amount to the zeros (e.g., 0.1 or 0.001, etc.). It turns out that you can introduce substantial large bias in your $\hat{\beta}$s by choosing different tiny amounts to add to your 0s: logging small numbers spreads those numbers out over a huge range. Adding 0.001 before logging a variable is not very different from subtracting 10,000 from an unlogged variable! So don't ever do this, even as a first try.

A Solution: the `logBound` and `logitBound` Transformations

A better solution that avoids arbitrary assumptions and bias is to "dummy out" the zeros before logging. This procedure treats the zero cases as *sui generis*: they are uniquely different from the rest of our cases, and we estimate the way in which they are different through a separate parameter. We end up with two variables on the right-hand side: an indicator of whether $x_i = 0$, and the log (or logit) of x_i in those cases where $x_i \neq 0$. That is, if you want to regress y on $\log(x)$ but x contains 0s, estimate this regression:

$$y_i = \beta_0 + \beta_1 I(x_i > 0) + \beta_2 \log'(x_i) + \epsilon_i \tag{1}$$

where $I(\cdot)$ is an indicator function and $\log'(\cdot)$ is defined as:

$$\log'(x) = \begin{cases} 0 & \text{if} \quad x \leq 0 \\ \log(x) & \text{if} \quad x > 0 \end{cases} \tag{2}$$

If we suppose that x_i is the number of cigarettes person i smokes per day, and y_i is i's probability of getting lung cancer, the specification makes sense: people who currently smoke even a little bit likely have a discretely higher chance of lung cancer than non-smokers, while the amount a smoker smokes may increase cancer probabilities but with diminishing marginal risk.

As you might imagine, the logic of equations 1 and 2 changes slightly if x needs to be logit transformed. Recall that the logit transformation,

$$\text{logit}(x) = \log\left(\frac{x}{1-x}\right),$$

fails if x is not between 0 and 1, so we need to dummy out $x_i \geq 1$ and $x_i \leq 0$ separately:

$$y_i = \beta_0 + \beta_1 I(x_i > 0) + \beta_2 I(x_i \geq 1) + \beta_3 \text{logit}'(x_i) + \epsilon_i \tag{3}$$

where $I(\cdot)$ is an indicator function and $\text{logit}'(\cdot)$ is defined as:

$$\text{logit}'(x) = \begin{cases} 0 & \text{if} \quad x \leq 0 \\ \log\left(\frac{x}{1-x}\right) & \text{if} \quad 0 < x < 1 \\ 0 & \text{if} \quad x \geq 1 \end{cases} \tag{4}$$

Note that we will only need all three pieces of Equation 4 if the covariate to be logit transformed contains *both* 0s and 1s; should either extreme be missing, we need only add one dummy variable to the specification.

The `logBound()` and `logitBound()` functions in the simcf package carry out these transformations for you. After loading the library, the above regression is as simple as:

```
res <- lm(y~logBound(x), data)
```

You can compare goodness of fit as usual. Moreover, you can use this technique on the right-hand side of any regression-like model, not just least squares regression.

Interpretation of results

Take special care in interpreting models in models with `logBound(x)` or `logitBound(x)` in the model formula. In setting up a hypothetical scenario for post-estimation prediction, make sure both the dummy term and the log term are set consistent with each other. For example, if the dummy is set to 0, the log must also be zero. And if the log is set to something other than 0, the dummy must be set to 1. Otherwise, you are asking the model to predict a logically impossible scenario; e.g., asking what happens when someone both smokes zero cigarettes and smokes twenty cigarettes in the same day.

I recommend either calculating the predicted values of `regime1` "by hand", or using the simcf package, as illustrated below. Our old friend `predict()` is very unlikely to return results for models including these terms, though if it does return an answer it will agree with other methods.

```
# Use simcf to predict the change in democracy given a shift from
# the 50th percentile of \% oil production to the 95th, all else equal.
# Note that this code is specifically set up for this example;
# if you wish to reuse it for other applications, you could
# rewrite it.
predOil <- function(res,formula,data,sims=10000,ci=0.95) {
  require(MASS)
  require(simcf)
  pe <- res$coefficients
  vc <- vcov(res)
  simbetas <- mvrnorm(sims,pe,vc)
  xscen <- cfMake(formula, data=data, nscen=1)
  xscen <- cfChange(xscen, ”oilL5”,
                    x = quantile(data$oilL5,probs=0.95),
                    xpre= quantile(data$oilL5,probs=0.5),
                    scen=1)
  linearsimfd(xscen, simbetas, ci=ci)
}

# Example call to this function (mdata is our listwise deleted dataframe)
m0 <- regime ~ oilL5 + metalL5 + GDPpcL5 + logitBound(islam) + oecd
res0.lm <- lm(m0, data=mdata)
PredOil0.lm <- predOil(res0.lm, m0, mdata)

# After this step, it would be a good idea to place the latest
# values of summary(res0.lm)$sigma, PredOil0.lm$pe, PredOil0$lower,
# and PredOil1$upper in vectors collecting your results across
# different runs.
```