

Regression III: Homework 3

(Due 10:00AM August 11, 2011)

The goal of this homework is to get you thinking about outliers and their potential deleterious effects on model estimates.

Using the `hw3_new.dta` (which is simply the dataset you've been working with all along plus one new variable), estimate the model you proposed at the end of the last homework that best accounts for the dependence of repression on violent dissent (dummy), democracy, GDP/capita and population. This could be purely additive in the variables above, it could be multiplicative in dissent and democracy, it could be non-linear in democracy (and potentially other variables) or it could be non-linear in democracy *and* multiplicative in democracy and dissent - whatever model you think is "best". The only caveat here is that we are going to use the `lm()` command later on, so this model has to be estimable in the linear model framework. Thus, non-linearity will have to be modeled with polynomials or splines of some sort.

1. After estimating the model and presenting the results, please use the methods in class to diagnose any problems with outliers.
2. Do graphical Methods agree or disagree with the statistical methods? If we are really interested in *influence*, which of these various methods is most useful and why?
3. There is one new variables in the dataset - `region`. This variable comes from the first digit of the Correlates of War country codes and indicates "major continental region". Add this variable to the model and assess the extent to which this eliminates outliers or at least reduces their "outlyingness".
 - (a) Is the evidence here (both from the statistical model and the residual diagnostics) in favor of or against the inclusion of `region`? Explain.
4. Using MM estimation with Huber weights estimate resistant regression models for the the models above (one with `region` and one without). Do the weight variables here provide confirmatory or contradictory information about the outliers?
 - (a) Using Residual-Residual plots, assess the extent to which there are significant differences between the OLS and resistant regression models. Do a formal test of the hypothesis that the slope coefficient is equal to one.
 - (b) Confirm the findings above by presenting graphically the predictions and 95% confidence bounds for each pair of models (i.e., comparing OLS and MM estimation for the model without `region` included and comparing OLS and MM estimation'on for the model with `region` included).