Case 1 for February 13: Predicting crime rate

- Don't forget that each student has to upload his/her results on the first day. In case you have already a teammate you can submit twice the same job. This is stupid I know but this is not to penalized students not in teams yet.
- Note that the task is relatively light to give you the time to get used of EViews.

It shouldn't always be econometrics; there is also such a thing called criminometrics. For this first assignment, you have to make use of Crime.wf1 (already in EViews), a cross-sectional dataset with information on crime in 90 counties in North Carolina for the year 1987. It contains the following variables:

- county county identifier;
- crmrte crimes per capita;
- prbarr estimated probability of arrest, computed as the number of arrests in 1986 divided by the number of offenses in 1986;
- prbconv estimated probability of conviction given arrest, computed as the number of convictions in 1986 divided by the number of arrests in 1986;
- prbpris estimated probability of imprisonment given conviction, computed as the number of imprisonments after conviction in 1986 divided by the number of convictions in 1986;
- avgsen average sentence length, in days, in 1986;
- polpc police per capita;
- density people per square mile;
- taxpc tax revenue per capita;
- west dummy variable = 1 if county is located in western North-Carolina, 0 otherwise;
- central dummy variable = 1 if county is located in central North-Carolina, 0 otherwise;
- urban dummy variable = 1 if county is located in a metropolitan area, 0 otherwise;
- pctmin percentage of the population belonging to a minority or being nonwhite;

- mix ratio of crimes involving face-to-face contact (e.g., assault) to those that do not;
- pctymle percentage of the population being male and young (between the age of 15 and 24). 1.

a) Estimate in EViews by ordinary least squares (OLS) the following two models, in double logs (both y and x variables) and in levels:

Model 1: $\ln(\text{crmrte})_i = \beta_0 + \beta_1 \ln(\text{prbarr})_i + \beta_2 \ln(\text{prbconv})_i + \beta_3 \ln(\text{prbpris})_i + \beta_4 \ln(\text{avgsen})_i + \beta_5 \ln(\text{polpc})_i + \beta_6 \ln(\text{density})_i + \beta_7 \text{west}_i + \beta_8 \text{central}_i + \beta_9 \text{urban}_i + \beta_{10} \ln(\text{pctmin})_i + \beta_{11} \ln(\text{pctymle})_i + u_i$

Model 2: $\text{crmrte}_i = \beta_0 + \beta_1 (\text{prbarr})_i + \beta_2 (\text{prbconv})_i + \beta_3 (\text{prbpris})_i + \beta_4 (\text{avgsen})_i + \beta_5 (\text{polpc})_i + \beta_6 (\text{density})_i + \beta_7 \text{west}_i + \beta_8 \text{central}_i + \beta_9 \text{urban}_i + \beta_{10} (\text{pctmin})_i + \beta_{11} (\text{pctymle})_i + u_i$

Hint: the EViews command for the natural log ln is log.

I find that some estimated parameters are a bit weird. For instance $\hat{\beta}_5$. How can you interpret this counter intuitive result?

b) You cannot compare the two previous models based on R^2, $\bar{R}^2$ and/or information criteria. Why is that so?

c) Plot fitted values as well as residuals for both models. Do you see on graphs any sign of misspecification?

d) Use formal misspecification tests: (1) normality, (2) homoskedasticity (White without cross terms) and (3) linearity RESET tests to detect potential problems.

e) Make more parsimonious models by deleting nonsignificant variables in both models (they can be different). Do it successively using individual $t-tests$ (or $p-values$). At the end verify using a joint $F-tests$ on all individual variables that you have discarded using $t-tests$.

f) Let's call these two restricted models Model 1' and Model 2'. Repeat the mispecification tests on those new models. Use R^2, $\bar{R}^2$ and information criteria to compare respectively Model 1 and 1' and Model 2 and 2'.

g) Reestimate Model 1' and 2' on the first observations 1 to 80 and compute the prediction for counties 81 to 90 (the last 10 ones on the list). For Model 2', you can estimate it in logs and take $exp(\ln(\widehat{crmrtei}))$ to come back to the prediction $\widehat{crmrtei}$. This is a bit tedious. I would proceed differently with EViews.

When you estimate your equation with the **Quick/Estimate Equation** window, write log(crmrtei) as the dependent variable, do not generate the variables before (also for explanatory variable). So doing EViews will give you the opportunity to forecast/predict log(crmrtei) or the level crmrtei. A box proposes you to forecast crmrtei or `log(crmrtei)`. Go to crmrtei in level, give a name for both forecasts and standard errors.

h) Compute RMSE (automatically reported in EViews, we will come back to that in the lecture) of both models 1' and 2' using EViews. What model is the better one for forecasting? Be carefull that you have to compute RMSE for 10 predicions.