Universidade de São Paulo / Faculdade de Filosofia, Letras e Ciências Humanas

Departamento de Ciência Política

FLP-0468 & FLS-6183

2º semestre/2016

Lista – Post–Estimation Regression Diagnostics

This lab is based on follow-up of the previous lab on interactions and the following paper and corresponding replication files:

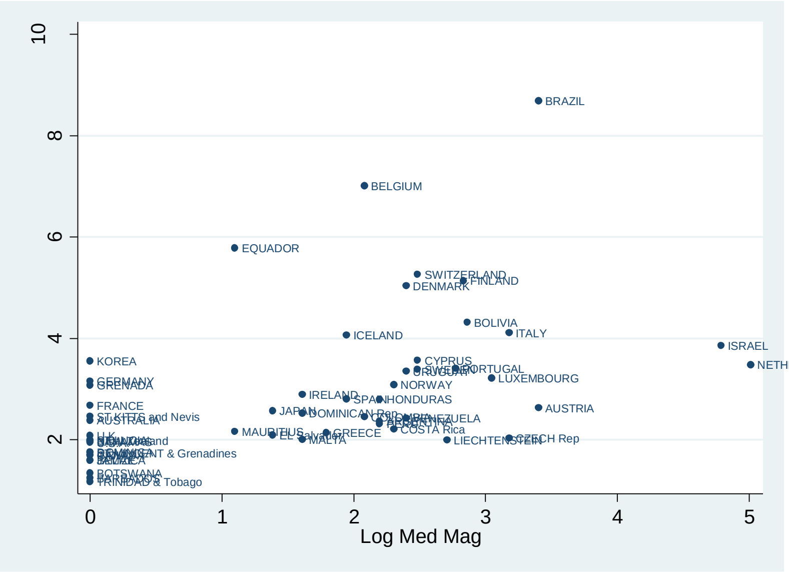
William Roberts Clark, Michael Gilligan and Matt Golder. 2006. "A Simple Multivariate Test for Asymmetric Hypotheses." *Political Analysis* 14: 311-331.

Please also review the relevant discussion in Chapter 10 of *The Fundamentals of Doing Political Science Research* and Section 13.10 of Gujarti and Porter's textbook.

As you will recall, we are interested in exploring Duverger's (1954) theory that multi-member electoral districts are necessary to produce a multiparty system (see Figure 1). We will explore this argument using the data collected and reported in:

Amorim Neto, Octavio & Gary Cox. 1997. "Electoral Institutions: Cleavage Structures and the Number of Parties." *American Journal of Political Science* 41: 149-174.

Figure 1. Number of Legislative Parties and Log Median District Magnitude



Specifically, Duverger argued that social forces are more likely to produce additional parties when countries employ multimember districts than when they do not. We tested Duverger's claims on the determinants of party system size with the following model and obtained the following regression results:

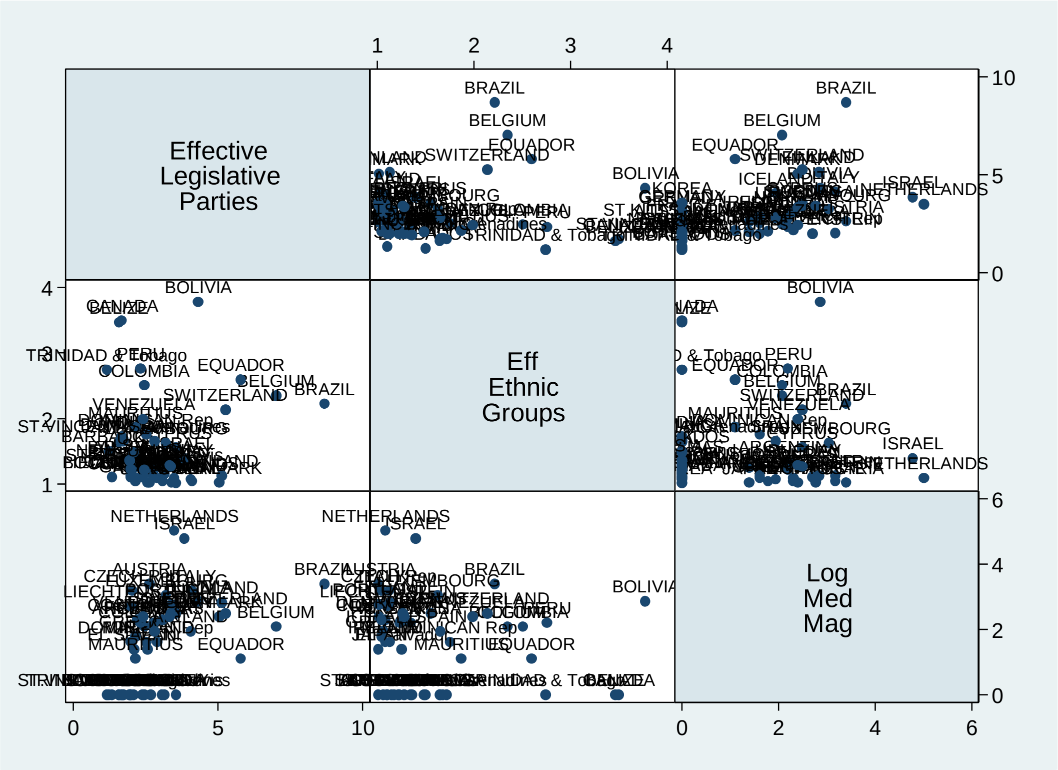
$$\text{Legislative Parties} = \beta_0 + \beta_1\text{Multimember District} + \beta_2\text{Social Heterogeneity} + \beta_3\text{Multimember District}\times\text{Social Heterogeneity} + \varepsilon$$

```
. regress  enps eneth lnml lmleneth

      Source |       SS       df       MS              Number of obs =      54
-------------+------------------------------           F(  3,    50) =    9.49
       Model |  39.7248824     3  13.2416275           Prob > F      =  0.0000
    Residual |   69.744403    50  1.39488806           R-squared     =  0.3629
-------------+------------------------------           Adj R-squared =  0.3247
       Total |  109.469285    53  2.06545822           Root MSE      =  1.1811

------------------------------------------------------------------------------
        enps |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       eneth |  -.3619712   .3486305    -1.04   0.304    -1.062216    .3382738
        lnml |  -.1911174   .2967357    -0.64   0.522    -.7871287    .4048939
    lmleneth |   .4833254   .1805094     2.68   0.010     .1207616    .8458893
       _cons |   2.671367   .6072149     4.40   0.000      1.45174    3.890994
------------------------------------------------------------------------------
```

Part I. Outliers

Please review the help files in Stata to learn about the following five commands: “predict r, rstudent”, “hilo” and “predict lev, leverage”; “lvr2plot” and “DFBETA”.

Figure 2. Number of Legislative Parties, Effective Number of Ethnic Groups and Log Median District Magnitude



Exercise 1. As Figure 2 makes clear, some cases seem that may be possible outliers and may be influencing our regression results including the notable cases of Bolivia, Brazil and the Netherlands. To explore whether these outliers may be influencing our results, we will examine the studentized residuals and their overall leverage on the regression results. Use the Stata commands to examine the studentized residuals and identify extreme values. Are our concerns regarding the three countries verified?

Exercise 2. As Gujarati and Porter explain, "A data point is said to exert (high) leverage if it is disproportionately distant from the bulk of the values of a regressor(s)." Now, let's examine the high leverage cases. Let's ask Stata to report the cases that have 5% or higher leverage by executing the command "list lev country if lev >.05." What do you observe?

Exercise 3. Let's now compare the leverage-versus-residuals using the stata command lvr2plot. What can we conclude?

Exercise 4. Following a regression, we can calculate the DFBETA scores to detect the influence with and without individual cases on our regression results for each coefficient. Let's now compare the highest DFBETA scores using the stata command "dfbeta (lnml)" and then asking to see the cases with the highest cutoff values "list country _dfbeta_1 if abs(_dfbeta_1) > 2/sqrt(54)". What can we conclude regarding influential cases with respect to the log of the median district magnitude?

Exercise 5. Following a regression, we can calculate the DFBETA scores to detect the influence with and without individual cases on our results for each coefficient. Let's now compare the highest DFBETA scores using the stata command "dfbeta (eneth)" and then asking to see the cases with the highest cutoff values "list country _dfbeta_2 if abs(_dfbeta_2) > 2/sqrt(54)". What can we conclude regarding influential cases with respect to the effective number of ethnic groups?

Exercise 6. Following a regression, we can calculate the DFBETA scores to detect the influence with and without individual cases on our results for each coefficient. Let's now compare the highest DFBETA scores using the stata command "dfbeta (lmleneth)" and then asking to see the cases with the highest cutoff values "list country _dfbeta_3 if abs(_dfbeta_3) > 2/sqrt(54)". What can we conclude regarding influential cases with respect to the interaction of the log of the median district magnitude and the effective number of parties?

Part II. Multicollinearity

Exercise 7. Using the VIF command, let´s now examine if there are any specific multicollinearities that may be inflating the standard errors in our models.

Part III. Normality of Residuals

Exercise 8. Let´s now check the normality of the residuals, you already used the "predict r, resid" command to generate residuals in part I. Now use the "kdensity r, normal" command to produce a kernel density plot with the normal option requesting that a normal density be overlaid on the plot. What can we conclude regarding the normality of residuals?

Part IV. Checking Homoscedasticity of Residuals

Exercise 9. Let´s now check the homoscedasticity of residuals. One of the main assumptions for the ordinary least squares regression is the homogeneity of variance of the residuals. If the model is well-fitted, there should be no pattern to the residuals plotted against the fitted values. A graphical method for detecting heteroscedasticity is using the "rvfplot, yline(0)" command which plots the residuals versus fitted (predicted) values.

Exercise 10. Please estimate the regression results with robust standard errors and compare them to the results reported earlier. What do you conclude?

Part V. Reviewing interaction

Exercise 11. Below please find two different models and the partial effects derivatives that show how changes in each explanatory variable influence changes in the dependent variable. Please explain the difference between the following two models in terms of which interaction is being tested and concentrate your discussion only on X (Hint: draw Venn diagrams if helpful).

Model 1:

$$y = \alpha + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon$$

$$\frac{\partial y}{\partial x} = \beta_1 + \beta_3 Z$$

$$\frac{\partial y}{\partial z} = \beta_2 + \beta_3 X$$

Model 2:

$$y = \alpha + \beta_1 X + \beta_2 Z + \varepsilon$$

$$\frac{\partial y}{\partial x} = \beta_1$$

$$\frac{\partial y}{\partial z} = \beta_2$$