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:

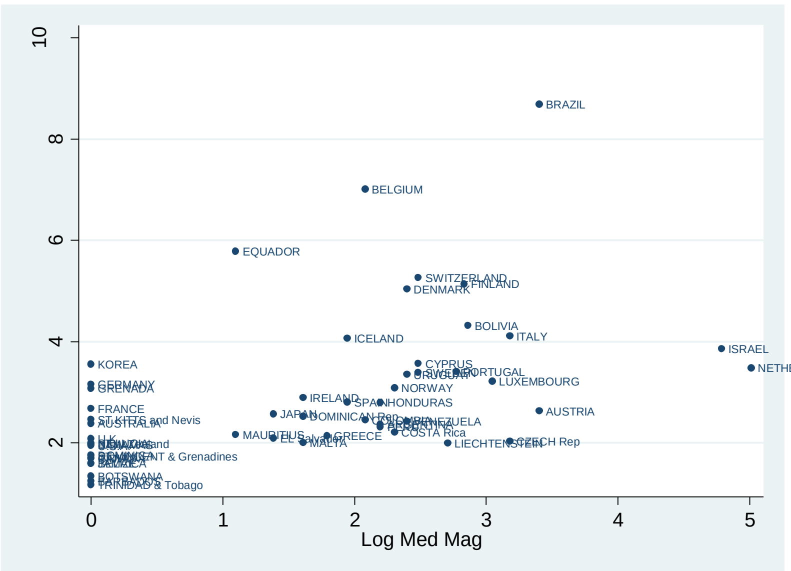
William Roberts Clark, Michael Gilligan and Matt Golder. 2006. "A Simple Multivariate Test for Asymmetric Hypotheses." *Political Analysis* 14: 311-331.

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:

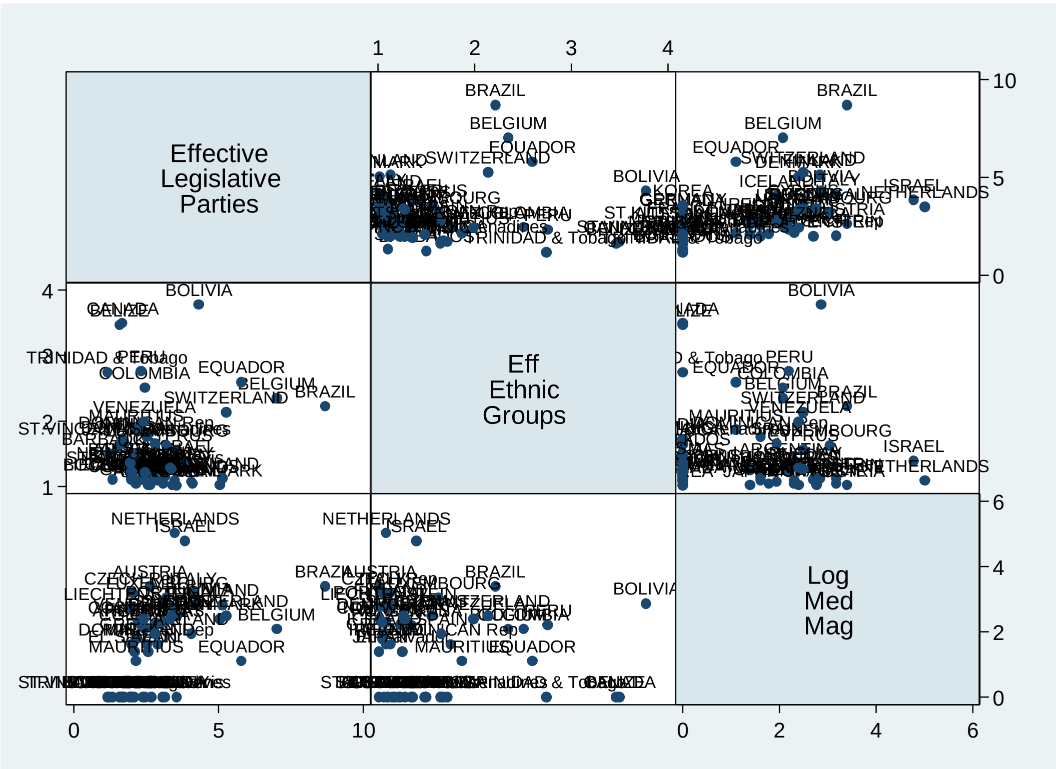
$$\text{Legislative Parties} = \beta_0 + \beta_1\text{Multimember District} + \beta_2\text{Social Heterogeneity} + \beta_3\text{Multimember District}\times\text{Social Heterogeneity} + \varepsilon$$

```
. 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

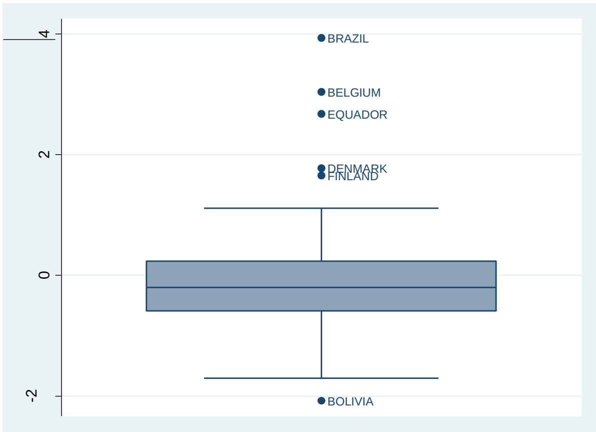
Netherlands. To explore whether these outliers may be influencing our results, we will examine the studentized residuals and their overall leverage on the regression results. Use the Stata commands to examine the studentized residuals and identify extreme values. Are our concerns regarding the three countries verified?

Answer: Based on the studentized residuals, the cases that have the largest divergence include Brazil and Bolivia, but we do not observe worrisome results for the Netherlands.

```
. hilo r country
10 lowest and highest observations on r
```

r	country
-2.077686	BOLIVIA
-1.702701	PERU
-1.271191	COLOMBIA
-1.193153	VENEZUELA
-1.178609	CZECH Rep
-1.047931	LIECHTENSTEIN
-.805474	BOTSWANA
-.7609829	BARBADOS
-.7179551	COSTA Rica
-.7112857	MALTA

r	country
.7224833	ITALY
.7841523	GERMANY
.9920237	ICELAND
1.104914	KOREA
1.114056	SWITZERLAND
1.653789	FINLAND
1.776574	DENMARK
2.674367	EQUADOR
3.037339	BELGIUM
3.936647	BRAZIL



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?

	lev	country
2.	.0631595	AUSTRALIA
3.	.0823729	AUSTRIA
6.	.057302	BELGIUM
7.	.3455021	BELIZE
8.	.5108224	BOLIVIA
9.	.0631595	BOTSWANA
10.	.1276521	BRAZIL
11.	.3552358	CANADA
12.	.0723374	COLOMBIA
15.	.0602766	CZECH Rep
20.	.0633979	EQUADOR
22.	.0578504	FRANCE
23.	.0595895	GERMANY
25.	.0669979	GRENADA
30.	.1272372	ISRAEL
31.	.0675156	ITALY
34.	.0721723	KOREA
39.	.1938402	NETHERLANDS
40.	.0510892	NEW Zealand
42.	.1094856	PERU
43.	.0522309	PORTUGAL
45.	.0548266	ST.KITTS and Nevis
46.	.0545296	ST.LUCIA
49.	.0557068	SWITZERLAND
50.	.1563072	TRINIDAD & Tobago

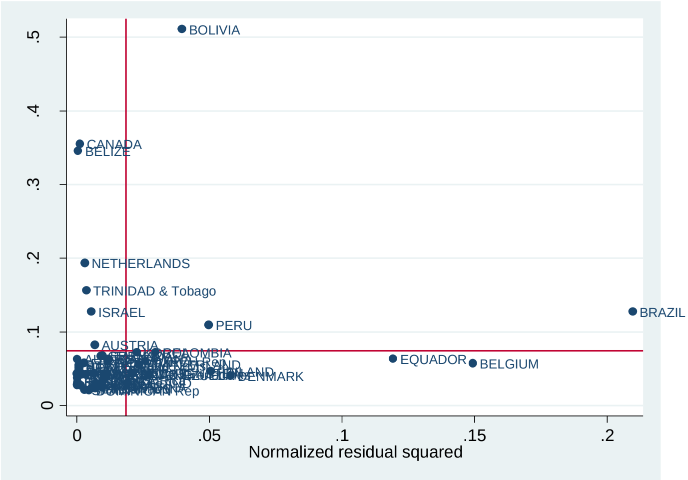
```
.
. hilo lev country, show(10) high
10 highest observations on lev
```

lev	country
.0723374	COLOMBIA
.0823729	AUSTRIA
.1094856	PERU
.1272372	ISRAEL
.1276521	BRAZIL
.1563072	TRINIDAD & Tobago
.1938402	NETHERLANDS
.3455021	BELIZE
.3552358	CANADA
.5108224	BOLIVIA

Answer: The cases that have the largest leverage include Bolivia, Canada, Belize, the Netherlands and Trinidad and Tobago. Some other cases also should concern us now. These include Brazil and Israel.

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

Answer: Based on the lvr2plot, Bolivia has the largest leverage and Brazil has the largest residual. Other important cases include Canada and Belize that have a large leverage and Ecuador and Belgium that present a large residual.



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?

Answer: Bolivia and Brazil are influential cases. The value for DFbeta for Bolivia is 1.197, which means that by being included in the analysis (as compared to being excluded), Bolivia increases the coefficient for "lnml" by 1.19 standard errors (1.19*0.297). On the other hand, the value for DFbeta for Brazil is -0.547, which means that by being included in the analysis as compared to being excluded, Brazil decreases the coefficient for "lnml" by .54 standard errors (-0.54*0.297). However, it is important to highlight that this interpretation is not quite straightforward since we are working with an interaction model.

```
. list country _dfbeta_1 if abs(_dfbeta_1 ) > 2/sqrt(54)
```

	country	_dfbeta_1
8.	BOLIVIA	1.196778
10.	BRAZIL	-.5457442

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?

Answer: Brazil and Ecuador are influential cases. The value for DFbeta for Brazil is -0.27, which means that by being included in the analysis as compared to being excluded, Brazil decreases the coefficient for "eneth" by 0.27 standard errors (-0.27*0.34). On the other hand, the value for DFbeta for Ecuador is 0.46, which means that by being included in the analysis (as compared to being excluded), Ecuador increases the coefficient for "eneth" by 0.46 standard errors (0.46*0.34). However, it is important to highlight that this interpretation is not quite straightforward since we are working with an interaction model.

```
. list country _dfbeta_2 if abs(_dfbeta_2 ) > 2/sqrt(54)
```

	country	_dfbeta_2
10.	BRAZIL	-.2756748
20.	EQUADOR	.4660389

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?

Answer: Bolivia, Brazil, Belgium and Peru are important influential cases; the first two countries are the most influential observations. The value for DFbeta for Bolivia is -1.53, which means that by being included in the analysis (as compared to being excluded), Bolivia decreases the coefficient for "lmleneth" by 1.53 standard errors (-1.53*0.18). On the other hand, the value for DFbeta for Brazil is positive; by being included in the analysis as compared to being excluded, Brazil increases the coefficient for "lmleneth" by 0.96 standard errors (0.96*0.18). However, it is important to highlight that this interpretation is not quite straightforward since we are working with an interaction model.

```
. list country _dfbeta_3 if abs(_dfbeta_3 ) > 2/sqrt(54)
```

	country	_dfbeta_3
6.	BELGIUM	.3058015
8.	BOLIVIA	-1.531085
10.	BRAZIL	.9633007
42.	PERU	-.292579

```
. list country _dfbeta_1 _dfbeta_2 _dfbeta_3
```

	country	_dfbeta_1	_dfbeta_2	_dfbeta_3
1.	ARGENTINA	-.0233911	.0152105	.0060767
2.	AUSTRALIA	-.0155597	-.0146171	.0098239
3.	AUSTRIA	-.1250779	-.0223001	.0876152
4.	BAHAMAS	.0207071	.0146819	-.0090969
5.	BARBADOS	.0575283	.0224862	-.0100158
6.	BELGIUM	-.1936575	.1240692	.3058015
7.	BELIZE	.0762741	.1283863	-.0954498
8.	BOLIVIA	1.196778	.1856355	-1.531085
9.	BOTSWANA	.1283199	.1205462	-.0810174
10.	BRAZIL	-.5457442	-.2756748	.9633007
11.	CANADA	.1235906	.2074943	-.154215
12.	COLOMBIA	.1001016	-.0658059	-.1507552
13.	COSTA Rica	-.0513025	.0211342	.0340012
14.	CYPRUS	.0012149	-.0006251	.0011612
15.	CZECH Rep	-.180451	-.0187879	.1132395
16.	DENMARK	.1598216	-.0465467	-.1131682
17.	DOMINICA	.0099765	-.0064545	.0068942
18.	DOMINICAN Rep	.0061669	-.0056552	-.0090764
19.	EL Salvador	.0114851	.0327314	-.0072106
20.	EQUADOR	.0741539	.4660389	-.1111373
21.	FINLAND	.1942716	-.0027043	-.1209966
22.	FRANCE	-.0542465	-.0483063	.0320373
23.	GERMANY	-.1174688	-.106704	.0711249
24.	GREECE	-.0132145	.0370701	.0099313
25.	GRENADA	-.1166173	-.1127302	.0762776
26.	HONDURAS	-.012128	.0073632	.0060808
27.	ICELAND	.0365711	-.050017	-.0255697
28.	INDIA	.0091408	-.0113536	.0108513
29.	IRELAND	.000092	-.0045183	-.0001268
30.	ISRAEL	-.1025648	-.0175509	.0280225
31.	ITALY	.1254592	.0150757	-.0858844
32.	JAMAICA	.0182482	-.0061239	.0078737
33.	JAPAN	.0027037	.0109922	-.0016047
34.	KOREA	-.1993724	-.1986672	.1353583
35.	LIECHTENSTEIN	-.1145586	.0080481	.073507
36.	LUXEMBOURG	-.0139993	.0078035	-.0242075
37.	MALTA	-.0002977	.0438175	.0003532
38.	MAURITIUS	.0053827	-.0222119	.0029843
39.	NETHERLANDS	-.1618759	-.0548922	.105664
40.	NEW Zealand	.0237551	.0188011	-.0120682
41.	NORWAY	.0053164	-.0021596	-.0037035
42.	PERU	.2043299	-.0901003	-.292579
43.	PORTUGAL	.0307137	-.0007345	-.0215323
44.	SPAIN	.0014937	.0013047	-.0088438
45.	ST.KITTS and Nevis	-.0272515	-.0232578	.0152529
46.	ST.LUCIA	.0274372	.0233023	-.0152618
47.	ST.VINCENT & Grenadines	.0120433	-.0050582	.0060438
48.	SWEDEN	.007541	-.0019115	-.0035102
49.	SWITZERLAND	-.071731	.000184	.1313888
50.	TRINIDAD & Tobago	-.0894446	-.1680144	.1264854
51.	U.K.	.0030508	.001348	-.0006608
52.	U.S.A.	.0211503	.0145927	-.0089553
53.	URUGUAY	.005459	-.0018537	-.0023515
54.	VENEZUELA	.047535	-.0018905	-.1022612

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.

Answer:

The VIF results are worrisome, especially with respect to log of median district magnitude and the interaction term. The inflation in the standard errors is worrisome especially for our hypothesis tests. As Kellstedt and Whitten recommend, this is a case where more data collection would be merited.

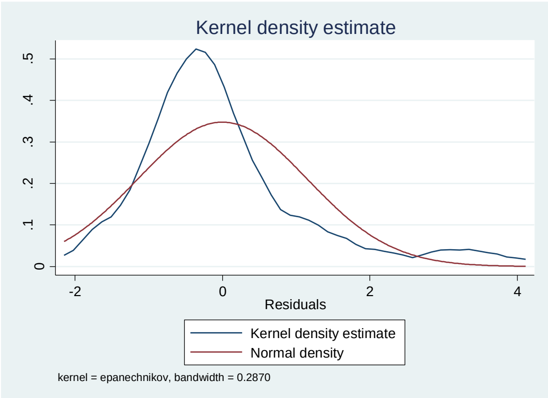
```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
    lmleneth |      6.99    0.143130
        lnml |      6.29    0.158968
       eneth |      2.15    0.465440
-------------+----------------------
    Mean VIF |      5.14
```

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?

Answer:

The graph shows that the residual distribution does not present clearly a normal distribution shape.

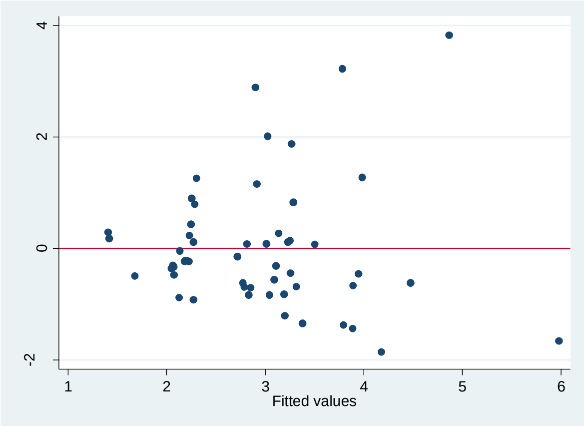


Part IV. Checking Homoscedasticity of Residuals

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.

Answer:

We see that the pattern of the data points is getting broader towards the right end, which is an indication of heteroscedasticity.



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

Answer:

In the model with robust standard errors, we see smaller standard errors. However, the difference is not significant, since all variables, which were significant before use it remain significant after using robust standard errors.

	Model 1	**Model 2** (Robust standard errors)
	b/t	b/t
eneth	-0.362	-0.362
	(-1.04)	(-1.63)
lnml	-0.191	-0.191
	(-0.64)	(-0.68)
lmleneth	0.483*	0.483*
	(2.68)	(2.19)
_cons	2.671***	2.671***
	(4.40)	(7.65)

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:

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

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

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

Answer:

In the model with the interaction term (Model 1), we are testing the hypothesis that the effect of X on Y depends on Z whereas in 2 we hypothesize that the effect of X on Y depends only on itself.