

## Notes: Lecture 1

### OLS: DIAGNOSTIC

#### 1) Model sensitivity

Model sensitivity refers to how estimates are affected by subsets of our data, that is how and if individual observations exert undue influence on the coefficients of your model. If a single observation (or small group of observations) substantially changes your results, you might want to know more about this and investigate further. Different types of influential observations:

Outliers: observation with large residual, whose dependent-variable value is unusual given its values on the independent variables. A) sample peculiarity; B) data entry error; C) other problem.

Leverage: extreme value on a predictor variable → high leverage. Leverage is a measure of how far one independent variable deviates from its mean. This can affect the estimate of coefficients.

Influence: An observation is said to be influential if removing that observation substantially changes the estimate of coefficients. It's a product of leverage & being an outlier.

use "...\\liiphart.dta", clear

***codebook ecogr709 const45 federal45 judrev45***

***reg ecogr709 const45 federal45 judrev45***

Do we see any observation with large leverage or residual?

A) outlier – plot the residuals versus fitted (predicted) values:

***rvfplot, mlabel(countryt) yline(0)***

\*You can also plot this:

***predict yhat***

***scatter ecogr709 yhat, mlabel(countryt)***

***scatter ecogr709 yhat, mlabel(countryt) || line yhat yhat***

B) leverage by the squared residuals:

***lvr2plot, mlabel(countryt)***

The lines are the means for leverage (horizontal), and for the normalized residual squared (vertical). There are three countries presenting potentially problematic values, let us take a look at them:

***list countryt ecogr709 const45 federal45 judrev45 if countryt=="JPN" / countryt=="SWI" / countryt=="IRE"***

C) measures of influence: Cook's D (combines information on the residual and the leverage)

lowest = 0; highest = most influential (cut-off point at  $4/n$ ) → in our case with 18 observations:  $4/18$

***predict d, cooks d***

***list countryt d if d > 4/18 & d != .***

D) dfbeta command will produce the DFBETAs for each predictor (variable-specific measure of influence)

***dfbeta***

***list countryt \_dfbeta\_1 \_dfbeta\_2 \_dfbeta\_3 if \_dfbeta\_1 != .***

→ AUT – federal45: increases beta by  $0.33 \times$  standard errors, i.e., .33 times the standard error of the beta of federal45 as estimated in the regression → ***di (.3382809 \* .151836)***. → can be positive or negative

***reg ecogr709 const45 federal45 judrev45 if countryt != "AUT"***

Values exceeding  $2/\sqrt{n}$  deserve further investigation: ***abs(DFBETA) > 2/sqrt(n) → di 2/(sqrt(18))***

***scatter \_dfbeta\_1 \_dfbeta\_2 \_dfbeta\_3 countryt, ylabel(-2(.5)2) yline(.47 -.47) mlabel(countryt countryt countryt)***

***list countryt \_dfbeta\_1 \_dfbeta\_2 \_dfbeta\_3 if (abs(\_dfbeta\_1) > 2/sqrt(18) & abs(\_dfbeta\_1) != .) | (abs(\_dfbeta\_2) > 2/sqrt(18) & abs(\_dfbeta\_2) != .) | (abs(\_dfbeta\_3) > 2/sqrt(18) & abs(\_dfbeta\_3) != .)***

E) avplot: added-variable plot (= partial-regression plot) → x, y after controlling for other predictors

The line represents the slope =  $\beta$ ; the axes are: residuals from regressing Y against all the independent variables except  $X_i$  & residuals from regressing  $X_i$  against the remaining independent variables.

***avplot federal45, mlabel(countryt)***

***reg ecogr709 const45 federal45 judrev45 if countryt != "IRE"***

***avplot federal45, mlabel(countryt)***

***avplots***

Note that the avplot command does not only work for the variables in the model, it also works for variables that are not in the model, which is why it is called added-variable plot. For instance, we can do an avplot on the variable effpart45.

***reg ecogr709 const45 federal45 judrev45***

***avplot effparty45, mlabel(countryt)***

***reg ecogr709 const45 federal45 judrev45 effparty45***

Should we exclude influential obs.? Not really; just 1) do that as a robustness check; or 2) include a dummy/a control var to account for features in common among those obs.; 3) transform the variable (eg. District magnitude); 4) check data entry

## 2) Assumptions of OLS & regression diagnostic

### Assumptions

- Normality - the errors should be normally distributed (only the errors, not the dependent/independent variables)  
  
→ the estimator remains unbiased (no bias in the difference between the estimate and the “true” value) but the test of hypothesis would be invalid.
- Homogeneity of variance (homoscedasticity) - the error variance should be constant  
  
 $E[\varepsilon^2 | X] = \sigma^2$  (finite variance!) → plot residuals vs fitted values and check that there is constant variance across different levels of fitted values → homoscedasticity;  
  
otherwise → heteroscedasticity → Use a different specification for the model OR weighted least squares OR heteroscedasticity-consistent standard errors
- No perfect Multicollinearity - the predictors are perfectly multicollinear if one of the regressors is a perfect linear function of the other regressors
- Strict Exogeneity  
  
 $E[\varepsilon | X] = 0 \rightarrow E[\varepsilon] = 0 \rightarrow \text{cov}(X, \varepsilon) = 0 \rightarrow E[X^T \varepsilon] = 0 \rightarrow \text{independent} \rightarrow \text{Exogeneity!}$   
  
If the regressor is correlated with the error term → Endogeneity!!! → invalid OLS estimates  
  
→ need to find an “instrument” (instrumental variables) → Z (not a regressor so far) →  $\text{cov}(Z_i, X_i) \neq 0$ ;  $\text{cov}(Z_i, \varepsilon_i) = 0$
- Model specification - the model should be properly specified (including all relevant variables, and excluding irrelevant variables): no omitted variable bias!
- Independence/No autocorrelation - the errors associated with one observation are not correlated with the errors of any other observation (autocorrelation or hierarchical structure as in times series, panel data, ...) → model it OR use panel corrected standard errors, clustered standard errors OR generalized least squares
- additional: observations are independent and identically distributed (iid) ... random sample

## 2.1) Checking Normality of Residuals

The coefficients are still unbiased but the p-values of t-test and F-test might not be valid

***reg ecogr709 const45 federal45 judrev45***

***predict r, resid***

***kdensity r, normal*** (kernel density: can be thought of as a histogram with narrow bins and moving average)

\*standardized normal probability (P-P) plot (shows troubles in the middle of the distribution)

***pnorm r***

\*quantiles of a variable against the quantiles of a normal distribution (shows troubles on the tails of the distribution)

***qnorm r***

***swilk r*** ( $H_0$ : normal. If W small & test less than .05, then non-normal!)

***sktest r*** ( $H_0$ : no skew. If less than .05, then skew!)

## 2.2) Checking Homoscedasticity of Residuals

If the model is well-fitted, there should be no pattern to the residuals plotted against the fitted values. When we have a distribution of residuals that does not have a rectangular shape → Heteroscedasticity

If we have Heteroscedasticity some observations are better explained than others → there are different slopes for different observations → we didn't have correctly included in the model some variables that could explain such difference. Notice that heteroscedasticity increases standard error of  $\beta$ s

***reg ecogr709 const45 federal45 judrev45***

***rvfplot, yline(0)***

\*White test ( $H_0$ : homoscedasticity;  $H_a$  heteroscedasticity; if we accept  $H_a$  ( $<0.05$ ) → heteroscedasticity)

***estat imtest, white*** (test for non-linear heteroscedasticity)

\*Breusch-Pagan test ( $H_0$ : homoscedasticity;  $H_a$  heteroscedasticity; if we accept  $H_a$  ( $<0.05$ ) → heterosc.)

***estat hettest*** (test for linear heteroscedasticity)

***reg secdim71 const71 numiss***

***reg secdim71 const71 mincab71***

### 2.3) Robust standard errors

With the robust option, the point estimates of the coefficients are exactly the same as in ordinary OLS, but the standard errors take into account issues concerning heterogeneity and lack of normality.

```
reg ecogr709 const45 federal45 judrev45 effparty45
```

```
reg ecogr709 const45 federal45 judrev45 effparty45, r
```

```
reg ecogr709 const45 federal45 judrev45 effparty45, vce(r)
```

\*with small samples:

```
reg ecogr709 const45 federal45 judrev45 effparty45, vce(bootstrap, rep(1000))
```

**HINT:** To use or not to use robust standard errors? Two schools of thoughts.

1) Stock and Watson: compare normal (or homoskedasticity) standard errors and robust standard errors, and if there are some differences, you should use the more reliable ones.

2) Compare OLS with and without robust standard errors. If there are large difference it means that you have a problem with your model specification. You should modify the model: standard errors just as a diagnostic tool. See for instance: <http://gking.harvard.edu/publications/how-robust-standard-errors-expose-methodological-problems-they-do-not-fix>

When you will work as a reviewer, if someone displays robust standard errors, this is a “yellow light” meaning that “the model is misspecified”. That is, if robust and classical standard error estimates differ (as we could argue, otherwise why to show only robust standard errors?), then we know that the statistical model being estimated is misspecified and so at least some estimates drawn from it are biased—often in a way that can be fixed but have not been.

### 2.4) Checking for Multicollinearity

Collinearity: two variables are an almost perfect linear combination of each other. Perfect Collinearity: estimates for a regression model cannot be uniquely computed. With more than two variables this is called multicollinearity. A minimum change in data may affect completely the betas coefficients; coefficients become unstable and the standard errors for the coefficients can get wildly inflated (also  $R^2$  can be inflated).

To check for multicollinearity: **vif** (variance inflation factor); vif should be  $< 10$ . Type vif after the regression.

```
reg ecogr709 const45 federal45 judrev45 effparty45
```

```
reg ecogr709 firstdim45 firstdim71
```

```
reg toxic csat msat
```

```
reg toxic csat vsat
```

## 2.5) Model Specification

A model specification error can occur when one or more relevant variables are omitted from the model or one or more irrelevant variables are included in the model. The most relevant case is when one omitted variable is correlated with the regressors included in the analysis and such omitted variable is a determinant of the DV.

→ omitted variable bias (can also be a cause of endogeneity) --- when variable Z, omitted from the model!, is correlated with both Y and X; being omitted, Z is absorbed into the error term, then X will be correlated with the error term [remember: other sources of endogeneity are 1) measurement error; 2) simultaneity]

→ check the model specification (1° consider the variables deemed relevant in the existing literature; 2° consider the variables deemed relevant by your theory; 3° use technical methods to detect any additional concern)

Test: If non-linear combinations of the explanatory variables have any power in explaining the response variable, the model is misspecified in the sense that the data generating process might be better approximated by a polynomial or another non-linear functional form.

***xi: reg kerry\_therm i.gender i.partyid3***

***linktest***

If a regression is properly specified, in principle we should not be able to find any additional independent variable that is significant (except by chance!) → the linktest tests what happens with 2 new variables:

\_hat (prediction), and \_hatsq (squared prediction).

\_hat should be significant since it is the predicted value. On the other hand, \_hatsq shouldn't, because if our model is specified correctly, the squared predictions should not have much explanatory power.

***ovtest***

Ovest is regression specification error test (RESET) for omitted variables; new variables are created based on the predicted values (squared, cubic, and power of four) and the ovtest refits the model using them.

***xi: reg kerry\_therm i.gender iraq\_approve***

***linktest***

***xi: reg kerry\_therm i.gender iraq\_approve i.partyid3***

***linktest***

***xi: reg kerry\_therm i.gender age***

***ovtest***

***xi: reg kerry\_therm i.gender age i.partyid3***

***ovtest***