Lab 08 - Introduction to Econometrics (2)

Learning outcomes for this lab:

• To introduce the necessary screening to test the assumptions of the linear model;

• To know how to choose the specification of the model and the estimator according to the different types of data (time-series, cross-section, panel)
OLS is BLUE if...

The OLS estimator is the **Best Linear Unbiased Estimator** if and only if the following hypotheses hold:

1. $E(u_t) = 0$  
   The statistical error has **zero mean**;
2. $\text{Var}(u_t) = \sigma^2$  
   The standard deviation of the errors is constant and of finite value (**assumption of homoscedasticity**)
3. $\text{Cov}(u_i, u_j) = 0$  
   The errors are not correlated one to each other (**assumption of no autocorrelation**)
4. $\text{Cov}(u_t, x_t) = 0$  
   The errors are not correlated with the regressors and...
5. $u_t \sim N(0, \sigma^2)$  
   The errors are distributed according to a **Gaussian (normal) distribution function**.

Once we decide to use the OLS, then, we have to: i) test assumptions 1-5; ii) understand what happens if one or more assumptions do not hold; iii) what to do to solve the problem.

As a general rule, when **some of the assumptions do not hold**, the roads to take are two:

- To use a **different specification of the model / different estimator**;
- To **transform data** in order to satisfy the assumptions.
1. If errors do not have zero mean...

If the regression includes the constant, the **mean of the residuals is always zero** (a good reason not to drop it from the model, regardless of its significance).

2. If the standard deviation of the errors is not constant (heteroscedasticity)

With **heteroscedasticity**, the OLS estimator keeps being unbiased, but the distribution of the standard errors is not reliable (i.e., **we might find as significant an estimate that is not**).

To test the existence of hetero.:  
- Test of **Goldfeld-Quandt**  
- Test of **White**  
What to do to get rid of it?  
- Use a **White-robust estimator** (reg..., robust)  
- Use a **GLS estimator** (if the variance of the error is linked to one of the regressors).  
- Transform the variables in logs.
3. If the errors are autocorrelated...

**positive** autocorrelation...

... and **negative** autocorrelation
If there is autocorrelation:

If there is autocorrelation, the OLS estimator keeps being unbiased but the distribution of standard errors is not reliable (i.e. **we might consider as significant an estimate that is not**). Therefore, the **R-squared is artificially high**.

How to test autocorrelation?
- **Durbin-Watson test**
  It can be used when the constant is included in the regression, when the regressors are not stochastic variables.

<table>
<thead>
<tr>
<th>Reject $H_0$: positive autocorrelation</th>
<th>Do not reject</th>
<th>Reject $H_0$: negative autocorrelation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inconclusive</td>
<td>$H_0$: No evidence of autocorrelation</td>
<td>Inconclusive</td>
</tr>
</tbody>
</table>

| 0 | $d_L$ | $d_u$ | 2 | $4 - d_u$ | $4 - d_L$ | 4 |

- **Breusch-Godfrey test**
If there is autocorrelation (2)

What to do?
- Try to use another estimator (i.e. the GLS, but only if the type of autocorrelation is known). If it is unknown, the solution can be worse than the problem!
- Change the model specification, building a differential model.

4. Multicollinearity

There is multicollinearity when the regressors are highly correlated; An extreme case is when the regressors are a linear combination one of each other: in such case coefficients can not be estimated! Take the Lab07 file:

\begin{verbatim}
  gen x = 1963.27*Y
  reg C Y x
\end{verbatim}

When regressors are highly correlated, R-squared is artificially high and the coefficients are unstable and easily not significant

\begin{verbatim}
  replace x = 110000 in 5
  reg C Y x
\end{verbatim}

None of the two variables (which measure, in a different way income) is now significant...
Multicollinearity (2)

Tests:
- Give a look to the correlation matrix
  \texttt{corr Y x}
- Compute the Variance Inflation Factors (VIF)
  \texttt{estat vif}

Solutions
- One of the correlated variables should be dropped from the model;
- Transform the correlated variables in just one variable;
- Collect more data!

5. The errors are not normally distributed

If the distribution of the errors is not normal, the OLS estimator is not BLUE
It is not obvious what to do, but...
... very often, the non-normality depends on a few very “strange” data (outliers)
In such a case, one can: (i) drop the observation (if one is sure that the outlier is a measurement error; (ii) transform the observation in a dummy (binary) variable, assuming value 1 in that case, 0 otherwise (if one thinks that the outlier stems from a special event... but we need to know what is the event!
An example is data on international tourism arrivals in October 2001...
Other problems: a) omitted variable bias...

If some **relevant variables are omitted from the model**, the coefficients of included variables are **biased and inconsistent** (we consider them significant although they are not reliable).
- **Theory should guide us** in deciding which variables should be included...
- and if a variable that is not considered by the theory is significant, a **new theoretical model should be developed**!

b) Inclusion of irrelevant variables

If we include irrelevant variables in the model (which **coefficients are not significant**), estimates are unbiased and consistent, but inefficient.
- We should **drop the irrelevant variables**...
- and if the variables are important in the theory, **the theory should be revisited**!

c) Stability of coefficients overtime

In the model we assume that coefficients are constant overtime. How do we test it?
- **Chow test**: split the sample in two and compare the coefficients.
- **Prediction test**: estimate the coefficients without including the last observations (**forward test**) or the early observations (**backward test**). Then, use the coefficients to estimate the missing values and compare them with the real values.
How to build the “best” econometric model

To be effective and efficient, an econometric model must:

- **pass the tests on the assumptions** on which the linear regression is built;
- be **parsimonious**;
- be **consistent** with the theory;
- have coefficients with the “right” **sign**;
- have coefficients with the “right” **size**;
- be able to explain more than what can be explained by alternative models.

Two roads to get to the best model:

- The classical road: from “**specific to general**”
- The modern road of **data mining**: from “**general to specific**”.
There is a model (and an estimator) for any data

1. **Cross-section data**: `reg`

2. **Time-series data**: firstly, tell Stata which is the time variable:
   ```
   tsset year
   ```
   There are regression models which are specifically built for time series:
   ```
   arima
   arch
   var
   ```
   and other regression commands (`help time`) – see the course of Time Series

3. **Panel data**: firstly, tell Stata that data are a panel. **Watch out**: also the country variable has to be quantitative. If it's not, write:
   ```
   egen country2 = group(country)    OR
   encode country, gen(country3)    OR
   sort country2, gen(country3)
   xtset country2 year
   ```
   Regressions for panel data:
   ```
   xtreg C Y, re
   xtreg C Y, re
   ```

4. **Probit** and **Logit**: if the dep. variable is the probability of a given event; **Tobit** if the dep. variable is censored (see the course of Tourism Microeconomics..)