Introduction to Econometrics
STAT-S-301
Information Session (2016/2017)
Lecturer: Yves Dominicy
Teaching Assistant: Elise Petit
Contacts

Lecturer: Yves Dominicy
Office: R42.5.222
E-mail: yves.dominicy@ulb.ac.be

Teaching Assistant: Elise Petit
Office: H.4.153
Office Hour: Thursday 10h-12h except on computer session weeks where it will be on Friday 14h-16h.
E-mail: elise.lb.petit@ulb.ac.be

Student Teaching Assistants (help for the exercise sessions):
- Abdelhakim El Allali
- Brikeno Elezi
Organisation of the Course

STAT-S-301: Introduction to econometrics (5 ECTS)

Three parts:

- Theory (Yves Dominicy)
- Theoretical exercises (Elise Petit and Students TA)
- Practical work on computer (Elise Petit)

Website: https://sites.google.com/site/yjgdominicy/teaching/stats301

Other material on: http://beta.respublicae.be (managed by students!)
Organisation of the Course - Schedule

1 - Theory

- **Friday**
  - 8h-10h
  - Room: R42.5.503

2 - Theoretical exercises (All Groups)

- **Thursday**
  - 8h-10h
  - Room: H1.309

You will find the solution of the exercises on Respublicae (the student forum).

For any question regarding exercises and computer session, please contact Elise Petit.
<table>
<thead>
<tr>
<th>Date</th>
<th>Subject (theory &amp; exercises)</th>
</tr>
</thead>
<tbody>
<tr>
<td>September</td>
<td></td>
</tr>
<tr>
<td>23 Friday</td>
<td>Introduction and Review of Statistics</td>
</tr>
<tr>
<td>29 Thursday</td>
<td>Review of Statistic/Linear Regression</td>
</tr>
<tr>
<td>30 Friday</td>
<td>Linear Regression</td>
</tr>
<tr>
<td>October</td>
<td></td>
</tr>
<tr>
<td>6 Thursday</td>
<td>Review of Statistics</td>
</tr>
<tr>
<td>7 Friday</td>
<td>Multiple Regression</td>
</tr>
<tr>
<td>13 Thursday</td>
<td>Linear Regression</td>
</tr>
<tr>
<td>14 Friday</td>
<td>Multiple Regression/ Nonlinear Regression</td>
</tr>
<tr>
<td>20 Thursday</td>
<td>Multiple Regression</td>
</tr>
<tr>
<td>21 Friday</td>
<td>Panel Data</td>
</tr>
<tr>
<td>27 Thursday</td>
<td>Nonlinear Regression</td>
</tr>
<tr>
<td>28 Friday</td>
<td>Instrumental Variables</td>
</tr>
<tr>
<td>November</td>
<td></td>
</tr>
<tr>
<td>10 Thursday</td>
<td>Panel Data</td>
</tr>
<tr>
<td>17 Thursday</td>
<td>Instrumental Variables</td>
</tr>
<tr>
<td>24 Thursday</td>
<td>Instrumental Variables</td>
</tr>
<tr>
<td>25 Friday</td>
<td>Instrumental Variables/ Times Series</td>
</tr>
<tr>
<td>December</td>
<td></td>
</tr>
<tr>
<td>1 Thursday</td>
<td>Instrumental Variables</td>
</tr>
<tr>
<td>2 Friday</td>
<td>Time Series/Quasi-experiment</td>
</tr>
<tr>
<td>15 Thursday</td>
<td>Time Series</td>
</tr>
<tr>
<td>16 Friday</td>
<td>Quasi-experiment/Summary</td>
</tr>
</tbody>
</table>
3 - **Computer Sessions (by group):**

<table>
<thead>
<tr>
<th>Week</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 3</td>
<td>STATA Introduction and Review of Statistics</td>
</tr>
<tr>
<td>Week 4</td>
<td>Linear Regression/Multiple Regression</td>
</tr>
<tr>
<td>Week 6</td>
<td>Nonlinear Regression and Panel Data</td>
</tr>
<tr>
<td>Week 12</td>
<td>Instrumental Variables and Time Series</td>
</tr>
</tbody>
</table>

- You have to register for the computer sessions next week, Wednesday from 8:00-12:00 and Thursday 8:00-12:00.
- Registration will be held in the office H 3.159 (Nicole Vanderroost’s office).
- There are four different groups (Monday (8:00-10:00 and 16:00-18:00), Wednesday (8:00-10:00) and Thursday (10:00-12:00)).
- See GeHoL for further information on the schedule and rooms.
Organisation of the Course – Final Exam

Structure of the exam:

1. multiple choice questions
2. open question on theory
3. open questions on theoretical and practical exercises

**Option 1:** Final exam (100% of the Grade)

**Option 2:** Problem sets + Mid-term exam + Final exam

*November 4: 8h-10h*

2) The weighted average between
   
   i. Problem set (25%),
   
   ii. Mid-term exam (25%),
   
   iii. Final Exam (50%).
Organisation of the Course – Problem sets

- Written Reports.

- STATA is the statistical software used for this course.

- Reports should be team work:
  - You can work in groups (maximum 4 people).
  - However, each student must participate actively on the report.
  - More details on the problem sets will be given during the first computer session.
Reports and Waivers

- The exam will be at the beginning of January.

- The course is considered successful if the total score is greater than or equal to 10/20.

- Students who are repeating the year are exempted from the course if and only if they have obtained at least 10/20 before.
Reference Book


<table>
<thead>
<tr>
<th>Chapter</th>
<th>Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-3</td>
<td>Review of Statistics</td>
</tr>
<tr>
<td>4-5</td>
<td><em>Linear Regression</em></td>
</tr>
<tr>
<td>6-7</td>
<td>Multiple Regression</td>
</tr>
<tr>
<td>8</td>
<td>Nonlinear Regression</td>
</tr>
<tr>
<td>9</td>
<td>Model evaluation</td>
</tr>
<tr>
<td>10</td>
<td>Panel Data</td>
</tr>
<tr>
<td>12</td>
<td>Instrumental Variable</td>
</tr>
<tr>
<td>13</td>
<td>Quasi-experiment</td>
</tr>
<tr>
<td>14</td>
<td>Time Series Regression</td>
</tr>
</tbody>
</table>

- It is recommend, but not compulsory, to buy the book.
- Do not limit your study to reading the slides!!!
Meaning of econometrics

What one can hear and read about econometrics:

- The aim in econometrics is to give an empirical context to theory.
- Broadly speaking, it is the application of statistical techniques to problems in economics.
- An econometrician can be seen as a statistician who develops the statistical techniques for solving empirical issues related to economic theory.

Wikipedia: Econometrics is the application of mathematics, statistical methods, and computer science, to economic data and is described as the branch of economics that aims to give empirical content to economic relations.

Econometrics = « Economy » + « metrics »
Some questions one might ask oneself

• What is the quantitative effect of reducing class size on a student’s achievement?

• What is the price elasticity of coffee?

• What is the effect on output growth of a one percentage point increase in interest rates by the ECB?

• How does another year of education change earnings?
This course is about using data to measure causal effects

- Ideally, we would like an experiment.
- But almost always we only have observational (non-experimental) data:
  - returns to education,
  - coffee prices,
  - monetary policy.
- Most of the course deals with difficulties arising from using observational data to estimate causal effects:
  - confounding effects (omitted factors),
  - simultaneous causality.

*Remember:* “correlation does not imply causation”!!!
In this course you will:

- Learn methods for estimating causal effects using observational data.

- Learn some tools that can be used for other purposes, for example forecasting using time series data.

- Learn to evaluate the regression analysis of others – this means you will be able to read/understand empirical economics papers in other economic courses.

- Get some hands-on experience with regression analysis in your problem sets.
A first example of an empirical problem

Class size and educational output

Policy question:
What is the effect on test scores (or some other outcome measure) of reducing class size by one student per class?
By $x$ students per class?

Answer:
We must use data to find out.
The California Test Score Data Set

Data: California school districts ($n = 420$)

The dataset contains data on test performance, school characteristics and student demographic backgrounds for school districts in California.

Variables:

1. 5$^{th}$ grade test scores (achievement test combining math and reading), district average.

2. Student-teacher ratio (STR) = number of students in the district divided by number of full-time equivalent teachers.
Initial look at the data: univariate

**TABLE 4.1** Summary of the Distribution of Student–Teacher Ratios and Fifth-Grade Test Scores for 420 K–8 Districts in California in 1998

<table>
<thead>
<tr>
<th></th>
<th>Average</th>
<th>Standard Deviation</th>
<th>10%</th>
<th>25%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
<th>75%</th>
<th>90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student–teacher ratio</td>
<td>19.6</td>
<td>1.9</td>
<td>17.3</td>
<td>18.6</td>
<td>19.3</td>
<td>19.7</td>
<td>20.1</td>
<td>20.9</td>
<td>21.9</td>
</tr>
<tr>
<td>Test score</td>
<td>665.2</td>
<td>19.1</td>
<td>630.4</td>
<td>640.0</td>
<td>649.1</td>
<td>654.5</td>
<td>659.4</td>
<td>666.7</td>
<td>679.1</td>
</tr>
</tbody>
</table>

This table does not tell us anything about the relationship between test scores (TS) and the Student Teacher Ratio (STR).
Initial look at the data: bivariate

Scatterplot of test score against student-teacher ratio

Do districts with smaller classes have higher test scores?
The empirical strategy

We *need to get some numerical evidence* on whether districts with low STRs have higher test scores – but how?

1. Compare average test scores in districts with low STRs to those with high STRs ("estimation").
2. Test the hypothesis that the mean test scores in the two types of districts are the same, against the hypothesis that they differ ("hypothesis testing").
3. Estimate an interval for the difference in the mean test scores ("confidence interval").
**Initial data analysis:**

Compare districts with “small” (STR < 20) and “large” (STR ≥ 20) class sizes:

<table>
<thead>
<tr>
<th>Class Size</th>
<th>Average score ($\bar{Y}$)</th>
<th>Standard deviation ($s_Y$)</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>657.4</td>
<td>19.4</td>
<td>238</td>
</tr>
<tr>
<td>Large</td>
<td>650.0</td>
<td>17.9</td>
<td>182</td>
</tr>
</tbody>
</table>

$\Delta = \mu_{small} - \mu_{large} = E(TS|STR < 20) - E(TS|STR \geq 20)$

1. **Estimation** of $\Delta = \text{difference between group means}$.
2. **Test the hypothesis** that $\Delta = 0$.
3. Construct a **confidence interval** for $\Delta$. 
1. Estimation

\[ \bar{Y}_{\text{small}} - \bar{Y}_{\text{large}} = \frac{1}{n_{\text{small}}} \sum_{i=1}^{n_{\text{small}}} Y_i - \frac{1}{n_{\text{large}}} \sum_{i=1}^{n_{\text{large}}} Y_i \]

\[ = 657.4 - 650.0 \]

\[ = 7.4 \]

Is this a large difference in a real-world sense?

- Difference between 60\textsuperscript{th} and 75\textsuperscript{th} percentiles of test score distribution is 667.6 – 659.4 = 8.2

- Is this a big enough difference to be important for school reform discussions, for parents, or for a school committee?
2. Hypothesis testing

Formulating hypothesis

\[ H_0: \mu_{\text{small}} - \mu_{\text{large}} = 0 \]
\[ H_1: \mu_{\text{small}} - \mu_{\text{large}} \neq 0 \]

Comparison of means for 2 populations (test at significance level \( \alpha \)).

Statistic: Compute the \( t \)-statistic:

\[
t = \frac{Y_s - Y_l}{\sqrt{\frac{s_{s}^2}{n_s} + \frac{s_{l}^2}{n_l}}} = \frac{Y_s - Y_l}{SE(Y_s - Y_l)}
\]

where \( SE(Y_s - Y_l) \) is the “standard error” of \( Y_s - Y_l \), the subscripts \( s \) and \( l \) refer to “small” and “large” STR districts, and \( s_{s}^2 = \frac{1}{n_s - 1} \sum_{i=1}^{n_s} (Y_i - \bar{Y}_s)^2 \).
2. Hypothesis testing

Compute the difference-of-means:

<table>
<thead>
<tr>
<th>Size</th>
<th>$\bar{Y}$</th>
<th>$s_Y$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>small</td>
<td>657.4</td>
<td>19.4</td>
<td>238</td>
</tr>
<tr>
<td>large</td>
<td>650.0</td>
<td>17.9</td>
<td>182</td>
</tr>
</tbody>
</table>

\[
t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{657.4 - 650.0}{\sqrt{\frac{19.4^2}{238} + \frac{17.9^2}{182}}} = \frac{7.4}{1.83} = 4.05
\]

Distribution under $H_0$: $t \approx N(0,1)$ (for $n_1, n_2 > 30$)

**Decision rule:**

$|t| > 1.96$, so reject (at the 5% significance level) the null hypothesis that the two means are the same.
3. Confidence interval

A 95% confidence interval for the difference between the means is given by

\[
CI_{95\%}(\mu_{\text{small}} - \mu_{\text{large}}) = \left[ (\overline{Y}_s - \overline{Y}_l) \pm 1.96 \times SE(\overline{Y}_s - \overline{Y}_l) \right]
\]
\[
= [7.4 \pm 1.96 \times 1.83] = [3.8, 11.0]
\]

Two equivalent statements:

1. The 95% confidence interval for \( \Delta \) does not include 0.

2. The hypothesis that \( \Delta = \mu_{\text{small}} - \mu_{\text{large}} = 0 \) is rejected at the 5% level.
What comes next...

- The mechanics of estimation, hypothesis testing, and confidence intervals should be familiar.
- These concepts extend directly to regression and its variants.
- However, before turning to regression, we will review some of the underlying theory of estimation, hypothesis testing, and confidence intervals:
  - Why do these procedures work, and why use these rather than others?