EC4051 Project and Introductory Econometrics

Dudley Cooke

Trinity College Dublin

Project Guidelines

- Each student is required to undertake an individual applied research project. This will account for 20% of the overall grade of the course.
- The topic of the research project will be of the student's own choosing, but the research should be related to the field of financial economics/economics of financial markets and requires the analysis of data.
- The research project could aim at testing a theory in financial economics, or at analyzing some type of economic behavior.
- Although the project does not strictly have to include econometrics it should, broadly-speaking, follow what others do in the literature. In any case, your aims and methodology need to be very explicit.

General Points

- The project does not strictly have to be from the range of topics covered on the course, as there are so many other topics (in that sense, the course is specialized), but if you choose to do a project on a different topic, please run it past me.
- The project should have a large portion of it devoted to what you do, i.e., your analysis, not what others do (unlike a regular term paper), although some recognition of what other people have done is important (and again, you may be drawing inspiration directly from these papers).
- The deadline for the submission of the projects will be 12 pm Friday April 10th.

Expectations

- The submitted project should include the following.

1. Title, Introduction, Background/Motivation/Literature Review.
2. Empirical approach. *If not econometrics*, what is the hypothesis you are testing? *If econometrics*, what is the dependent variable? What are the independent variables? How do you expect the independent variables to affect the dependent variable?
3. Description of the dataset. That is, the source of the data, detailed summary statistics, graphics for visual presentation of the data.
4. Empirical results. That is, tables of results (estimated coefficients, standard errors, number of observations, adjusted R^2), results of the appropriate tests and comment of the empirical results.
5. Summary/conclusions with discussion of possible extensions.

Last Year ...

- Last year, there was much worrying and speculation about the project. For any of the above:
    - My office hours are on Friday's from 1600 to 1800
    - We have classes, some of which I will run using appropriate software (the room is booked every week for your use)
    - The project average grade last year was good (mainly because the projects were good)
- Specific projects that proved popular (a risk and return trade-off applies) last year:
    - Calender effects and EMH (50% of the papers/the safest option)
    - CAPM (25%)
    - APT, Yield Curve, GARCH effects in financial markets, asset price bubbles (20%)

Planning the Project

- You should already have an idea of the project you want to do from Term 1. If not, wait a few weeks maximum.
- Again, if the topic is standard great, if not, just run it by me (there is a 95% chance it will be fine anyway).
- Think about where you can get the data and read some papers from the literature.
- Important: find one you can replicate easily and quickly.
- **Most important:** get the data (we have EcoWin and Data Stream in the 24hr PAC room, near the library)
- Run the regressions and do the write up (easier said than done).

Econometrics

- Since some people have done empirical projects and some have not, I will cover some econometrics at the beginning of the course.
- I will also demonstrate how to run regressions on XL and Microfit. But I encourage you to talk to one another (plus, the project is not marked on a curve)
- A good book is Wooldridge, J., 2006. Introductory Econometrics: A Modern Approach, Thomson (third edition). [hereafter, W]
- We'll (very quickly) cover:

1. Bivariate and multivariate models with cross-section data, W Chs. 2-3; Hypothesis (t and F) tests, Wooldridge Ch. 4.
2. Time-series data and autoregressive processes, W Chs. 10-11; Autocorrelated and heteroskedastic errors, W Ch. 12.
3. Unit roots, cointegration and GARCH, W Ch. 18.2-18.3.

Bivariate Linear Regression Model

- One of the simplest statistical models we can think of this the following.

$$y = \beta_0 + \beta_1 x + u$$

- Here, y is the dependent variable (endogenous) and x is the explanatory variable (exogenous); that is, x causes y, with a margin of error, u.
- Also, β_0 is some fixed number (the intercept) and β_1 is the slope (the extent to which a change in x affects y).
- Example: ice-cream sales (y) and temperature (x). We expect, $\beta_0 > 0$ and $\beta_1 > 0$. That is, some ice-creams are always sold, and the hotter it is, the more ice-creams are sold.

Assumptions in Linear Regression Model

- What we really want to know is the magnitude of β_1. This is an estimate, denoted $\widehat{\beta}_1$.
- This could be critical in deciding how to market ice-cream sales and/or ice-cream production, given the weather forecast.
- Out main assumption is:

$$\mathbb{E}\left(u\right) = 0$$

- We also assume the errors uncorrelated with regressors. That is,

$$Cov\left(x, u\right) = \mathbb{E}\left(xu\right) = 0 \Rightarrow E\left[x\left(y - \beta_0 - \beta_1 x\right)\right] = 0$$

Estimation - Ordinary Least Squares (OLS)

- Why do we make these assumptions? So we can find a 'good' estimate of β_1.
- Say we have $i = 1...n$ observations of ice-cream sales. Without proof, the slope estimate is,

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} \text{ where } \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- Good news: a PC, plus software (even XL) will work this out for you.

Multivariate Model

- Say we propose a model of ice-cream sales with more than one explanatory variable (i.e., a multivariate model). Specifically,

$$\text{ice-cream sales} = \beta_0 + \beta_1 \left(\text{temp}\right) + \beta_2 \left(\text{vans}\right) + u$$

- The interpretation is a ceteris paribus idea. That is, given the number of vans on the street, how does temperature affect ice-cream sales?
- If it turns out $\widehat{\beta}_2 \neq 0$ we made an initial mistake.
- That is, previously, $\widehat{\beta}_1$ could overstate the impact of temperature on ice-cream sales.

Multivariate Model

- In general, k factors can affect y, so,

$$y_i = \beta_0 + \sum_{k=1}^{n} \beta_k x_{ki} + u_i$$

$$E\left(u | x_1, x_2, ..., x_k\right) = 0 \text{ and } i = 1, 2, ..., n$$

- Another example:

$$\log\left(\text{salary}\right) = \beta_0 + \beta_1 \log\left(\text{sales}\right) + \beta_2 \left(\text{tenure}\right) + \beta_3 \cdot \left(\text{tenure}\right)^2$$

- Note that $\left(\text{tenure}\right)$ is linear but $\left(\text{tenure}\right)^2$ is not. However, the model is still linear (in coefficients).
- Say $\widehat{\beta}_2 > 0$, then "controlling for sales, tenure has a positive effect on salary".

How Can we Tell if Our Model is any Good?

- Is our econometric model good? R^2 tells us how well the sample regression line fits the data.
- Define the following (with $TSS = RSS + ESS$):

$$TSS = \sum_{i=1}^{n} (y_i - \overline{y})^2 \text{ and } RSS = \sum_{i=1}^{n} \widehat{u}_i^2 \text{ and } ESS = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$$

- Now,

$$R^2 = \frac{\sum_{i-1}^{n} \left[(y_i - \overline{y}) \left(\widehat{y}_i - \overline{\widehat{y}} \right) \right]^2}{\left(\sum_{i=1}^{n} (y_i - \overline{y})^2 \right) \left(\sum_{i=1}^{n} \left(\widehat{y}_i - \overline{\widehat{y}} \right)^2 \right)} = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

- As $R^2 = 1 \Rightarrow$ we get the "best" fit; if $R^2 = 0 \Rightarrow$ we get the "worst" fit. In our example, if $R^2 = 1 \Rightarrow$ temperature and vans explain 100% of ice-cream sales.

Things to Remember

1. Collinear variables. Ice-cream and vans on RHS could be bad. If temperature increased ⇒ more vans, as companies want to increase sales. Therefore, both temperature and vans explain ice-cream sales, but they must be related.
2. Irrelevant variables (model overspecification/too many RHS variables). Not too bad ⇒ can test for this.
3. Omitted variables. Suppose temperature and vans are not related and we don't include one: that affects estimated coefficient of the other.
4. Heteroskedasticity. Error, u, may have non-equal variance given any value of explanatory variables.

Hypothesis Testing

- Given the model $y_i = \beta_0 + \sum_{k=1}^{n} \beta_k x_{ki} + u_i$, we also need to test whether a variable is significant in explaining anything itself - a hypothesis test. The null is,

$$H_0 : \beta_k = 0$$

- We then,

1. Specify an alternative. These come from economic theory and are one-sided $(H_1 : \beta_k \gtrless 0)$ or - more usually -two-sided $H_1 : \beta_k \neq 0$.
2. Select a level of significance (stringency of test), which is defined as $\Pr(\text{reject } H_0 | H_0 \text{ correct})$. Usual significance levels are 5% and 10%.

- Finally, it matters how powerful our test is. That depends on "degrees of freedom" (basically number of observations vs number of explanatory variables) ⇒ rule of thumb: less observations, worse power of the test.

Hypothesis Testing

- Say we want to test the following:

$$\begin{aligned} \text{ice-cream} &= \beta_0 + \beta_1 \left(\text{temp}\right) + \beta_2 \left(\text{vans}\right) + u \\ H_0 &: \beta_1 \neq 0 \text{ (vans explain ice-cream sales)} \\ H_0 &: \beta_1 = 0 \text{ (they don't)} \end{aligned}$$

- Note: to do this, we also need assumption on errors that $u \sim N\left(0, \sigma^2\right)$, consistent with two things;

$$\begin{aligned} E\left(u|x_1, ..., x_K\right) &= 0 = E\left(u\right) = 0 \text{ (uncorrelated errors)} \\ Var\left(u|x_1, ..., x_K\right) &= Var\left(u\right) = \sigma^2 \text{ (homoskedastic errors)} \end{aligned}$$

t-statistic

- What matters is not only the coefficient, but its standard error (s.e.). The t-statistic is then,

$$t_{\beta_2} = \frac{\widehat{\beta}_2}{s.e.\left(\widehat{\beta}_2\right)}$$

- Again, software can do this for you or see t-tables in statistics books.
- If reject, we say, "x_2 is statistically insignificant". In other words, we may get $\widehat{\beta}_2 \gtrless 0$,but statistically speaking it is zero.
- There are also p-values $\Rightarrow$ easy to interpret, as $0 \leqslant p \leqslant 1$. P-value is defined as, $\Pr\left(\underbrace{\mid T \mid}_{t_{n-k-1}} > \underbrace{\mid t \mid}_{\text{from table}}\right)$. If p is close to zero, there is evidence against the null $(t \rightarrow \infty)$ and vice-versa if p close to one.

P-values and t-statistics

- It should be clear that $\widehat{\beta}$, $s.e.\left(\widehat{\beta}\right)$, t and p-value are all related.
- The bigger the $s.e.\left(\widehat{\beta}_2\right)$ then the smaller is the t-statistic, for a given coefficient (i.e. 'null is good'). Say, $p = \Pr\left(\mid T \mid > 1.85\right) = 0.5$.
- In words "we observe a value for the t-stat as extreme as we did in 50% of all random sample when null is true" there is weak evidence against the null.

F-tests

- Suppose you want to test another possibility;

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \\
H_0 &: \beta_1 = \beta_2 \\
H_1 &= \beta_1 < \beta_2
\end{aligned}
$$

- In this case,

$$t = \frac{\widehat{\beta}_1 - \widehat{\beta}_2}{s.e.\left(\widehat{\beta}_1 - \widehat{\beta}_2\right)}$$

- This is an F-test, and is basically a combination of t-tests.

An Example of an F-test

- Example in book,

$$\widehat{\log(\text{wage})} = \underset{(0.021)}{1.472} + \underset{(0.007)}{0.067}(\text{college}) + \underset{(0.02)}{0.077}(\text{univ.}) + \underset{0.0002}{0.005}(\text{exp.})$$

- "College" and "univ" are both individually insignificant, but are they statistically different from each other?
- Well, as $\widehat{\beta_1} - \widehat{\beta_2} = -0.0162$, this suggests return to college (for wages) is 1% less than to university, and we are interested in that.
- To test, first define $Q_1 \equiv \beta_1 - \beta_2$, then,

$$\begin{aligned} \log(\text{wage}) &= \beta_0 + (Q_1 + \beta_2) \cdot (\text{college}) + \beta_2(\text{univ.}) + \beta_3(\text{exp.}) + u \\ &= \beta_0 + Q_1 \cdot (\text{college}) + \beta_2(\text{college} + \text{univ.}) + \beta_3(\text{exp.}) + u \\ H_0 &: Q_1 = 0 \Rightarrow \beta_1 = \beta_2 \\ H_1 &: Q_1 < 0 \Rightarrow \beta_1 < \beta_2 \end{aligned}$$

F-tests

- Now, suppose model is

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \qquad (*) \\
H_0 &: \beta_1 + \beta_2 = 1
\end{aligned}
$$

- How do we test H_0? We've got to run two models. Estimate (*) and a model with the restriction already imposed. That is,

$$(y - x_2) = \beta_0 + \beta_1 (x_1 - x_2) + u$$

- Compare unrestricted and restricted models,

$$F = \frac{(RSS_R - RSS_u)\frac{1}{d}}{RSSu/(n - K - 1)}$$

- Under H_0, $F \sim F_{(d,n-K-1)}$.

F-tests

- Alternatively, let we can define a new coefficient,

$$\gamma = \beta_1 + \beta_2 - 1$$

- And test a new null,

$$H_1 : \gamma = 0$$

- This gives,

$$\begin{aligned} &\Rightarrow y = \beta_0 + \beta_1 x_1 + (\gamma + 1 - \beta_1)\, x_2 + u \\ &\Leftrightarrow (y - x_2) = \beta_0 + \beta_1 (x_1 - x_2) + \gamma x_2 + u \end{aligned}$$

- Then we use the t-test under H_0, $t \sim t_{(n-K-1)}$

Roundup

- You should feel comfortable with the following:

1. Interpreting coefficients from a cross-section OLS regression.
2. Checking goodness of fit.
3. Performing simply hypothesis tests.

- Much of this needs to go into the project.