Econometrics II

Lecture 1: Preliminaries & Fundamentals

Måns Söderbom[*]

Department of Economics, University of Gothenburg

29 March 2011

[*] mans.soderbom@economics.gu.se. www.economics.gu.se/soderbom. www.soderbom.net

1. Introduction & Organization

- The overall aim of this course is to improve
  - your understanding of empirical research in economics; and
  - your ability to apply econometric methods in your own research
- Good empirical economics: Ask an interesting research question, and find an empirical approach - an **identification strategy** - that enables you to answer the question.
- Mastering statistical techniques - e.g. OLS, 2SLS, Maximum Likelihood,.... - is only one of your tasks. Being able to program your own likelihood function in Stata is impressive, but doesn't guarantee you will be regarded as an outstanding empirical economist.
- Techniques are essentially **tools** - and if what you are 'building' is not important or interesting, it doesn't matter how rigorous your methods are.
- I would argue that the opposite applies too: You may have an important idea, but if your quantitative analysis is poor quality the research project is unlikely to be a success.
- The first part of the course - lectures 1-10 - is oriented towards the analysis of cross-section and panel data. The second part of the course - lectures 11-15 - covers time series econometrics.

1.1. Mechanics & Examination

- To get the course credits, **five** computer exercises have to be completed, plus you need to pass an oral exam.
  - Computer exercises: Feel free to work collaborate with fellow students on these (a group size of 2 or 3 students would be best). Short reports on the computer exercises - one per group & exercise - should be emailed to the person in charge of the exercise **one week after each computer session**, at the latest. We will follow up on these in class.

  - Oral exam: Details to follow, but basically each student gets assigned a 30-minute slot. During the viva, the student meets with the examiner(s) and will be requested to answer a small number of questions on the course material orally.
  - Grades (Fail, Pass or High Pass) will be based on the performance in the viva and in the computer exercises.

- The course web page will be updated continuously, especially with regards to relevant articles and research papers. So you should check for updates every now and then.

1.2. Two textbooks - Two points of view

- **Angrist and Pischke** (2009). *Mostly Harmless Econometrics.* Jazzy exposition, reflects the 'new' way of thinking about econometrics, based on the experimentalist paradigm and potential outcomes. Cuts corners sometimes and engages in slightly too much self-promotion for my taste (papers by Angrist and Krueger are cited very favorably throughout). We will use this book a lot. From now on, I'll refer to this simply as AP.

- **Greene**, W (2008). *Econometric Analysis*, 6th edition. Comprehensive. Strong on mathematical details. A bit mechanical. Not particularly good at explaining results intuitively. We will use selected chapters in this book when discussing discrete choice models, panel data, multinomial & ordered outcomes, and sample selection bias.

1.3. Recommended reading for this lecture

AP, Chapters 1-2, 3.1, 3.4.1, 3.4.3 (optional).

2. Questions about Questions

Reference: AP, Chapter 1.

This chapter emphasizes that there is more to empirical economics than just statistical techniques. The authors argue that a research agenda revolves around four questions:

- What is your **causal relationship** of interest?
- What would your **ideal experiment** look like - i.e. one that would enable you to capture the causal effect of interest?
- What is your **identification strategy**?
- What is your mode of **statistical inference**?

This really is fundamental. You could do worse than taking these four bullet points as a starting point for the introduction of the paper you are currently writing (OK, you may want to add some context & motivation after the first question). Let's discuss them briefly.

2.1. Your causal relationship of interest

Causal relationship: tells you what will happen to some quantity of interest (e.g. expected earnings) when an explanatory variable (e.g. years of schooling) changes, holding other variables fixed. The counterfactual outcome is central here - the outcome that would result from pursuing a different educational policy, for example.

2.2. Your ideal experiment

In this course we will talk a lot about the problems posed by (traditional econometrics jargon) endogeneity bias or (new econometrics jargon) selection bias. In general, if your goal is to estimate the causal effect of changing variable X on your outcome variable of interest, then the best approach is **random assignment**. In many cases this is too costly or totally impractical, and so we have no choice but to look for answers using observational (non-experimental) data. Even so, thinking hard about the 'ideal' experiment may be very useful when getting started on a research project (e.g. when you're designing a survey instrument or the survey design), and it may help you interpret the regressions you've run based on observational data.

It's also a useful checkpoint: If even in an 'ideal' world you can't design the experiment you need to answer your question of interest, then chances are you won't be able to make much progress in the real world.

In short, forcing yourself to think about the mechanics of an ideal experiment highlights the forces you'd like to change, and the factors you'd like to hold constant - and you need to be clear on this to be able to say something about causality.

2.3. Your identification strategy

Recognizing that the ideal experiment is likely not practical, you have to make do with the data you've got (or can get) The term identification strategy is often used these days as a way of summarizing the manner in which you use observational data to **approximate** a real experiment. A classic example is the Angrist-Krueger (1991) QJE paper in which the authors use the interaction of compulsory attendance laws in US states and students' season of birth as a natural experiment to estimate the causal effects of finishing high school on wages.

In general, if you don't have data generated from a clean, laboratory type experiment, then using data from a natural experiment is second best (you will then likely spend most of your time at seminars arguing about whether your data really can be interpreted as having been generated by a natural experiment).

2.4. Your mode of statistical inference

- You need to be clear on the population you're studying (that's what your results are supposed to refer to)

- You need to make sure your sample is an appropriate basis for making inference about the population.

- You need to make sure the procedure for calculating standard errors is appropriate.

If you're clear on your mode of statistical inference, then you will be able to make accurate statements about what we can learn from your data analysis about mechanisms of interest in the population.

If you're clear on all these four questions (referred to as FAQ by Angrist and Pischke), you've made a good start on your project.

3. The Experimental Ideal

Reference: AP, Chapter 2.

3.1. The Selection Problem

Suppose our question of interest is as follows:

Do hospitals make people healthier?

Note that this is a question about causality. How shall we go about answering our research question?

Well one very credible approach would be random assignment, right? That is, draw say 1,000 individuals randomly from the population and assign half of them to hospital treatment and keep the other half out of hospitals (regardless of their health); and then measure everyone's health status after a suitable period of time. Great idea in theory - but totally useless in practice, of course.

Suppose we set out to answer our research question using observational (non-experimental) data. Survey data on hospital visits and health status reported by AP are summarized in the following table:

Group	Sample size	Mean health status	Std error
Hospital	7,774	3.21	.014
No hospital	90,049	3.93	.003

where health status is measured on a 1-5 scale (1=poor; 5=excellent). Clearly the average health status appears to be worse among the hospitalized (the t-value associated with H_0: health is invariant to hospital status is 58.9).

If our research question had been as follows: "Are hospitalized people healthier than nonhospitalized people?" then we would have been done by now. But this is not our question. And we cannot infer from the data above that hospitals make people less healthy, because... well because people who (choose to) go to hospital are probably less healthy to begin with.

This may strike you as a totally trivial and silly example. Indeed. And precisely because it's so obvious what the nature of the problem is, this is a nice setting in which we can explore the concept of **selection bias**.

We take our starting point the potential outcomes framework:

$$\text{Potential outcome} = \left\{ \begin{array}{c} Y_{1i} = \text{ health outcome for } i \text{ if hospitalized} \\ \\ Y_{0i} = \text{health outcome for } i \text{ if not hospitalized} \end{array} \right\}$$

For each individual (or firm, country,...; from now on, individual) there is thus a potential outcome (Y_{1i}) under **treatment**, and another potential outcome (Y_{0i}) **without treatment**. Think of Y_{1i} and Y_{0i} as outcomes in alternative states of the world. In general, we expect there to be a **distribution** of Y_{1i} and Y_{0i} in the population.

The **observed outcome** Y_i can be written in terms of potential outcomes as

$$\begin{aligned} Y_i &= \left\{ \begin{array}{c} Y_{1i} \text{ if } D_i = 1 \\ \\ Y_{0i} \text{ if } D_i = 0 \end{array} \right\} \\ &= Y_{0i} + (Y_{1i} - Y_{0i}) D_i, \end{aligned}$$

where D_i is a dummy variable equal to 1 if individual i received treatment, and 0 otherwise.

The **treatment effect** - in our example, the causal effect of hospitalization **-** is simply the difference between the two potential outcomes: $Y_{1i} - Y_{i0.}$. Note that the treatment effect is taken to be **heterogeneous** across individuals (reflected by the i subscripts).

Unfortunately, it is impossible to measure treatment effects at the individual level, as we can never observe the full set of potential outcomes in alternative states of the world - basically, because we don't have access to parallel universes. Researchers therefore focus on various forms of **average treatment effects**.

We suspect we won't be able to learn about the causal effect of hospitalization simply by comparing the average levels of health by hospitalization status. We can now formalize the problem somewhat. We begin by noting that

$$E\left[Y_i|D_i=1\right]-E\left[Y_i|D_i=0\right]=E\left[Y_{1i}|D_i=1\right]-E\left[Y_{0i}|D_i=0\right],$$

which shows that the difference in means is the difference in the average potential outcome under treatment **for treated people** minus the average potential outcome under non treatment **for non-treated people**. We can decompose this further (subtract and add $E\left[Y_{0i}|D_i=1\right]$):

$$\begin{aligned} E\left[Y_i|D_i=1\right]-E\left[Y_i|D_i=0\right] &= \\ & E\left[Y_{1i}|D_i=1\right]-E\left[Y_{0i}|D_i=1\right] \qquad \text{(ATT)} \\ & +E\left[Y_{0i}|D_i=1\right]-E\left[Y_{0i}|D_i=0\right] \quad \text{(Selection bias)}, \end{aligned}$$

where ATT = Average treatment effect on the treated is the average causal effect for the sub-group of treated individuals.

Now, ATT is an interesting quantity: it captures the difference between the averages in:

- the health of the hospitalized
- what **would** have been the health of the hospitalized had they not been hospitalized.

Note that the latter is the **counterfactual,** i.e. the outcome that didn't happen.

But as you can see, we can't infer the ATT from the difference in means, because the difference in observed means is the sum of the ATT and a selection bias term.

The selection bias is simply the difference in average health under non-treatment between those who were hospitalized and those who were not.

It seems reasonable to expect $E\left[Y_{0i}|D_i=1\right] < E\left[Y_{0i}|D_i=0\right]$ (why?), hence the selection bias is negative.

This would imply that the difference in observed means is a downward biased estimator of the ATT. And this, of course, is consistent with the numbers (worse health for the hospitalized).

3.2. Random Assignment Solves the Selection Problem

The basic problem that we encountered above was that actual treatment status D_i is not independent of potential outcomes (e.g. Y_0 likely lower for hospitalized than for non-hospitalized people). Now suppose that we could randomly assign treatment to individuals in the population. Because in this case D_i is independent of potential outcomes, the selection bias disappears. Recall, in general:

$$\begin{aligned} E\left[Y_i|D_i=1\right]-E\left[Y_i|D_i=0\right] &= \\ & E\left[Y_{1i}|D_i=1\right]-E\left[Y_{0i}|D_i=1\right] \qquad \text{(ATT)} \\ & +E\left[Y_{0i}|D_i=1\right]-E\left[Y_{0i}|D_i=0\right] \quad \text{(Selection bias)}, \end{aligned}$$

which, if D_i is independent of potential outcomes, becomes

$$E\left[Y_i|D_i=1\right]-E\left[Y_i|D_i=0\right]=E\left[Y_{1i}|D_i=1\right]-E\left[Y_{0i}|D_i=1\right] \qquad \text{(ATT)}$$

Hence, we can infer the ATT from the difference in means. In fact, this coincides with the average treatment effect in the entire population (i.e. regardless of actual treatment status) since

$$E\left[Y_{1i}|D_i=1\right]-E\left[Y_{0i}|D_i=1\right]=E\left[Y_{1i}\right]-E\left[Y_{0i}\right].$$

This is often referred to as simply as the average treatment effect (ATE).

Notice that, under random assignment of treatment, ATT or ATE can be obtained by running the following simple OLS regression:

$$Y_i=\beta_0+\beta_1 D_i+u_i,$$

where the estimate of β_1 is the estimated ATT and ATE. More on this below.

You see how powerful the method of randomization is. Provided you get the design of your experiment right, all you need to do is to compare mean values across the two groups ($D_i = 0, D_i = 1$). If done right, a pure randomized experiment is in many ways the most convincing method of evaluation.

It sounds easy, but, of course, life is never easy. Experiments have their own drawbacks:

- They are rare in economics, and often **expensive** to implement. 'Social experiments' carried out in the U.S. typically had very large budgets, with large teams and complex implementation. However, quite a few randomized evaluations have recently been conducted in developing countries on fairly small budgets. Indeed, this approach has become extremely popular in the development economics literature.

- They may not be amenable to **extrapolation**. That is, there may be questionmarks as to the external validity of the results of a particular experiment. The main reasons are: i) it may be very hard to replicate all components of the program elsewhere; ii) the results may be specific to the sample (you might argue this is a general problem in empirical economics - that may well be true, but typically experiments are conducted in relatively small regions, which possibly exacerbates the problem); iii) the results may be specific to the program (would a slightly different program have similar effects?).

- There may be (lots of) **practical problems** related to the implementation of experiments. Getting the design of the experiment right really is the big challenge, and as you can imagine much can go wrong in the field. Suppose you start to give free school meals randomly in 50% of the schools in a region where previously school meals were not free. One year later you plan to turn up and compare pupil performance in treated and nontreated schools. But how can you be sure parents whose kids are in nontreated schools have not reacted to your reform by changing schools? Or could treatment affect the decision as to when someone should leave school? The basic point is that you typically need time between initiating the treatment and measuring the outcome, and much can go wrong in the meantime. There may be ethical issues: why give some people treatment and not others? How justify not helping those that need it the most?

For these reasons, most empirical research is still based on non-experimental (observational) data. When we have non-experimental data, we must assume that individuals at least partly determine whether they receive treatment. As we have seen, this may lead to problems with the simple difference-in-means estimator if the individual's decision to get treatment depends on the benefits of treatment (selection bias). Addressing this problem is largely what the literature on treatment effect estimation based on non-experimental data is about. Notice that this is precisely the problem solved by randomization. Indeed, it is useful to take the position that a notional randomized trial is our benchmark.

3.3. Regression Analysis of Experiments

Regression analysis is the key tool for analyzing experimental as well as non-experimental data in applied economics. Suppose for a moment that the treatment effect is the same for everybody,

$$Y_{1i} - Y_{0i} = \rho.$$

In this case, we can re-write our expression for observed outcomes in regression form:

$$\begin{aligned} Y_i &= Y_{0i} + (Y_{1i} - Y_{0i}) D_i \\ &= E(Y_{0i}) + \rho D_i + \{Y_{0i} - E(Y_{0i})\} \\ &= \alpha + \rho D_i + \eta_{i,} \end{aligned}$$

where the residual η_i is interpretable as the random part of Y_{0i}. Now take expectations, conditional on treatment and no treatment:

$$\begin{aligned} E[Y_i|D_i = 1] &= \alpha + \rho + E[\eta_i|D_i = 1] \\ E[Y_i|D_i = 0] &= \alpha + E[\eta_i|D_i = 0], \end{aligned}$$

so that

$$\begin{aligned} E\left[Y_i|D_i=1\right]-E\left[Y_i|D_i=0\right] &= \rho & \text{(treatment effect)} \\ &\quad +E\left[\eta_i|D_i=1\right]-E\left[\eta_i|D_i=0\right] & \text{(selection bias).} \end{aligned}$$

You see how in this framework the selection bias amounts to non-zero correlation between the regression error term η_i and the regressor D_i. Since

$$E\left[\eta_i|D_i=1\right]-E\left[\eta_i|D_i=0\right]=E\left[Y_{i0}|D_i=1\right]-E\left[Y_{i0}|D_i=0\right],$$

this correlation reflects the difference in potential outcomes (under no treatment) between those who get treated and those who don't. It is also clear that, if D_i is randomly assigned, there is no selection bias so that a regression of observed outcomes Y_i on actual treatment status D_i estimates the causal effect of interest (ρ).

Covariates. Covariates - other explanatory variables, or 'controls' - are often included in regression specifications even if treatment was randomly assigned. There are two reasons for this.

1. The experimental design may be **conditional** random assignment. For example, suppose students are randomly assigned to classes of differing size within schools; but suppose the average class size differs across rural and urban schools. Then it becomes important to control for (in this case) rural or urban school in the regressions.

2. Inclusion of relevant control variables may increase the **precision** with which we can estimate the causal effect of interest. This is because including the control variables reduces the residual variance, which in turn lowers the standard error of the regression estimates (recall: $V\left(\widehat{\beta}_{OLS}\right)=\sigma^2\left(X'X\right)^{-1}$). Note that, if D_i is randomly assigned, adding control variables to the specification should not result in a very different estimate of ρ compared to a specification without control variables.

We will return to this and other related points later in the course.

4. Making Regression Make Sense

Reference: AP, Chapters 3.1, 3.4.1, 3.4.3 (optional).

We now turn to the first chapter of Part II: *The Core* in AP. The style of this chapter is more formal than what you encounter in chapters 1-2, focusing mostly on the link from population parameters to estimates based on finite samples, and the statistical properties of regression estimates. You may find part 3.1 quite dry. Unfortunately, it is also quite important.

4.1. Regression Fundamentals

Why are we using regression in empirical research? Let's go back to the fundamentals.

4.1.1. The Conditional Expectation Function

The Conditional Expectation Function (CEF) for a dependent variable Y_i, given a $K \times 1$ vector of covariates X_i (with elements x_{it}) is the expectation - or the *population* average - of Y_i with X_i held fixed. We write the CEF as

$$E\left[Y_i|X_i\right],$$

hence the CEF is a function of X_i. Because X_i is random, so is the CEF.

For continuous Y_i with conditional density $f_Y\left(t|X_i = x\right)$ at $Y_i = t$, the CEF is

$$E\left[Y_i|X_i = x\right] = \int t f_Y\left(t|X_i = x\right) dt,$$

whereas for discrete Y_i it is

$$E\left[Y_i|X_i = x\right] = \sum_t t P\left(Y_i = t|X_i = x\right).$$

Note that expectation is a **population** concept. In empirical research, we use samples to make inference about the population; e.g. the sample CEF is used to learn about the population CEF. But remember:

the objects of interest apply for the population, hence we start by defining these population objects.

An important complement to the CEF is the **law of iterated expectations**, which says that an unconditional expectation can be written as the unconditional average of the CEF:

$$E\left[Y_i\right] = E\left\{E\left[Y_i|X_i = x\right]\right\},$$

where the outer expectation uses the distribution of X_i. See p. 32 in AP for a proof.

Next, consider three theorems that summarize important properties of the CEF.

Theorem 3.1.1 The CEF Decomposition Property

$$Y_i = E\left[Y_i|X_i\right] + \varepsilon_i,$$

where ε_i is mean independent of X_i (i.e. $E\left[\varepsilon_i|X_i\right] = 0$), and therefore ε_i is uncorrelated with any function of X_i. See AP for a proof (it's straightforward).

Theorem 3.1.2 The CEF Prediction Property Let $m\left(X_i\right)$ be any function of X_i. The CEF solves

$$E\left[Y_i|X_i\right] = \arg\min_{m(X_i)} E\left[\left(Y_i - m\left(X_i\right)\right)^2\right],$$

hence the CEF is the **minimum mean squared error** predictor of Y_i, given X_i.

Proof: Subtract and add $E\left[Y_i|X_i\right]$ inside the brackets, then expand:

$$\begin{aligned}\left(Y_i - m\left(X_i\right)\right)^2 &= \left(\left\{Y_i - E\left[Y_i|X_i\right]\right\} + \left\{E\left[Y_i|X_i\right] - m\left(X_i\right)\right\}\right)^2 \\ &= \left\{Y_i - E\left[Y_i|X_i\right]\right\}^2 \\ &\quad +2\left\{Y_i - E\left[Y_i|X_i\right]\right\}\left\{E\left[Y_i|X_i\right] - m\left(X_i\right)\right\} \\ &\quad + \left\{E\left[Y_i|X_i\right] - m\left(X_i\right)\right\}^2.\end{aligned}$$

Note the penultimate line has expectation zero since $Y_i - E[Y_i|X_i]$ can be replaced by ε_i. Finally, note that the last term is minimized at zero when $m(X_i)$ is the CEF.

Theorem 3.1.3 The ANOVA Theorem

$$V(Y_i) = V(E[Y_i|X_i]) + E[V(Y_i|X_i)],$$

where $V(\cdot)$ denotes variance and $V(Y_i|X_i)$ is the conditional variance of Y_i, given X_i. This says that the variance of Y_i can be written a the variance of the CEF plus the variance of the residual. See AP for a proof.

4.1.2. Linear regression and the CEF

The CEF provides a natural summary of empirical relationships. For example, it shows how the conditional expectation of log earnings varies with years of education. So we consider the CEF an object of interest. Now let's think about how the CEF links to **regression**. We define the $K \times 1$ population (note!) regression coefficient vector β as the solution to the following minimization problem:

$$\beta = \arg\min_b E\left[(Y_i - X_i'b)^2\right].$$

Since there are K parameters, there will be K first-order conditions of the form

$$\begin{aligned} 2E[(Y_i - X_i'b)X_1] &= 0 \\ 2E[(Y_i - X_i'b)X_2] &= 0 \\ &(...) \\ 2E[(Y_i - X_i'b)X_K] &= 0, \end{aligned}$$

which we can write in more compact form as:

$$E\left[X_i\left(Y_i - X_i'b\right)\right] = 0$$

where the 0 is now understood to be a $K \times 1$ vector of zeros. We can now solve for β:

$$\begin{aligned} E\left[X_i\left(Y_i - X_i'\beta\right)\right] &= 0 \\ \beta &= \left(E\left[X_i X_i'\right]\right)^{-1} E\left[X_i Y_i\right]. \end{aligned}$$

Note that β are **not** estimators. These coefficients are simply features of the joint distribution of dependent and independent variables.

Key question at this point: Is the vector of population regression coefficients *of any interest*?

AP offers three reasons the vector of regression coefficients might be of interest. The premise of this discussion is that you are interested in the CEF.

Theorem 3.1.4 The Linear CEF Theorem. Regression Justification I: Suppose the CEF is linear. Then the population regression function ($X_i'\beta$) is the CEF.

Proof: So we have a linear CEF, e.g. the form is:

$$E\left[Y_i|X_i\right] = X_i'\beta^*,$$

where β^* is a $K \times 1$ vector of coefficients. The claim is that $\beta^* = \beta$. Is this true? We saw above that one property of the the CEF is

$$E\left[X_i\varepsilon_i\right] = 0,$$

hence

$$E\left[X_i\left(Y_i - E\left[Y_i|X_i\right]\right)\right] = 0.$$

Now plug in our assumed form for the CEF:

$$E\left[X_i\left(Y_i - X_i'\beta^*\right)\right] = 0,$$

and solve for β^*. We obtain

$$\beta^* = \left(E\left[X_i X_i'\right]\right)^{-1} E\left[X_i Y_i\right] = \beta.$$

Hence, if

- ....you are interested in the CEF; and
- ....you have reason to believe the CEF is linear,

then you regression is an appropriate method for estimating your object of interest.

Of course, the CEF may not be linear. In such cases, we need a different justification for using regression.

Theorem 3.1.5 The Best Linear Predictor Theorem. Regression Justification II: The regression function ($X_i'\beta$) is the best **linear** predictor of Y_i given X_i, in a MMSE sense (i.e. no other vector of coefficients can generate a lower mean squared error than the regression coefficient vector β, in the class of linear functions)

Proof: This follows immediately from the definition of β (minimizes the sum of squared residuals, hence minimizes the MSE).

Theorem 3.1.6 The Regression CEF Theorem. Regression Justification III: The regression function ($X_i'\beta$) provides the MMSE linear approximation to the CEF ($E\left[Y_i|X_i\right]$); that is,

$$\beta = \arg\min_b E\left[\left(E\left(Y_i|X_i\right) - X_i'b\right)^2\right].$$

Proof: We know the definition of β:

$$\beta = \arg\min_b E\left[(Y_i - X_i'b)^2\right].$$

Now write

$$\begin{aligned}
(Y_i - X_i'b)^2 &= ([Y_i - E(Y_i|X_i)] + [E(Y_i|X_i) - X_i'b])^2 \\
(Y_i - X_i'b)^2 &= [Y_i - E(Y_i|X_i)]^2 \\
&\quad + [E(Y_i|X_i) - X_i'b]^2 \\
&\quad + 2[Y_i - E(Y_i|X_i)][E(Y_i|X_i) - X_i'b].
\end{aligned}$$

Once we take expectations on both sides, the last line disappears (expectation zero). Note that the first term does not involve b. Hence, since β minimizes $E(Y_i - X_i'b)^2$, it must be true that β minimizes $E\left[(E(Y_i|X_i) - X_i'b)^2\right]$ as well (expectations operator applied to the penultimate line). And that's what we were supposed to prove.

These may strike you as quite esoteric points. But the discussion is really fundamental. The object of interest in most empirical studies is the CEF, and regression is a very useful tool for shedding light on it.

4.1.3. Asymptotic OLS Inference

In practice, the CEF and the population regression vector are unknown. Using samples, we draw inferences about these quantities. We may, for example, want to test the hypothesis that some element β_k of the population regression vector is equal to zero. To do this, we need to know something about the sampling distribution of the estimate of β_k.

Recall the definition of the population regression vector:

$$\beta = \left(E\left[X_i X_i'\right]\right)^{-1} E\left[X_i Y_i\right].$$

As you know, the OLS estimator of β is obtained by replacing population moments with sample moments:

$$\hat{\beta} = \left[\sum_i X_i X_i'\right]^{-1} \sum_i X_i Y_i$$

a practice justified by the law of large numbers (stating that your sample moments get arbitrarily close to the population moments as the sample size increases). In order to do inference for the population vector, we need to understand the sampling distribution of $\hat{\beta}$ - think of this as the distribution of $\hat{\beta}$ that would result from repeated sampling from the population.

To derive the asymptotic sampling distribution of $\hat{\beta}$ we make use of the central limit theorem and the Slutsky theorem. These are stated in AP, page 43, and summarized here:

- Central limit theorem: Sample moments are asymptotically normally distributed after subtracting the corresponding population moment and multiplying by the square root of the sample size.
- Slutsky theorem: The asymptotic product of two random variables, one of which converges in distribution and the other converges in probability to a constant, is unaffected by replacing the one that converges to a constant by this constant.

Now write:

$$Y_i = X_i'\beta + [Y_i - X_i'\beta] \equiv X_i'\beta + e_i,$$

where e_i is a residual, uncorrelated with X_i. Use this equation to re-write the expression for $\hat{\beta}$:

$$\begin{aligned}
\hat{\beta} &= \left[\sum_i X_i X_i'\right]^{-1} \sum_i X_i Y_i \\
\hat{\beta} &= \left[\sum_i X_i X_i'\right]^{-1} \sum_i X_i X_i' \beta + \left[\sum_i X_i X_i'\right]^{-1} \sum_i X_i e_i \\
\hat{\beta} &= \beta + \left[\sum_i X_i X_i'\right]^{-1} \sum_i X_i e_i.
\end{aligned}$$

It follows that

$$\sqrt{N}\left(\hat{\beta} - \beta\right) = N \left[\sum_i X_i X_i'\right]^{-1} \frac{1}{\sqrt{N}} \sum_i X_i e_i. \tag{4.1}$$

It then follows from the Slutsky theorem - which states that the asymptotic product of two random variables, one of which converges in distribution and the other converges in probability to a constant, is unaffected by replacing the one that converges to a constant by this constant - that (4.1) has the same asymptotic distribution as

$$E\left[X_i X_i'\right]^{-1} \frac{1}{\sqrt{N}} \sum_i X_i e_i.$$

The central limit theorem implies that $\frac{1}{\sqrt{N}} \sum_i X_i e_i$ is asymptotically normally distributed with mean zero and covariance matrix $E\left(X_i X_i' e_i^2\right)$, since this matrix is the covariance matrix of $X_i e_i$.

Therefore, $\hat{\beta}$ has an asymptotic normal distribution with probability limit β and covariance matrix

$$E\left[X_i X_i'\right]^{-1} E\left(X_i X_i' e_i^2\right) E\left[X_i X_i'\right]^{-1}. \tag{4.2}$$

The theoretical standard errors used to construct t-statistics are the square roots of the diagonal elements of this matrix. In practice, these standard errors are estimated by using sums for expectations and estimated residuals. Standard errors computed in this way are known as heteroskedasticity-consistent standard errors (White, 1980). In Stata, you can get such standard errors by adding "robust" to the regression options. These are asymptotically valid in the presence of any kind of heteroskedasticity, including homoskedasticity. Therefore, it would seem you might as well always use robust standard errors, in which case you can remain agnostic as to whether there is or isn't heteroskedasticity in the data. As we shall see later on in the course, the formula (4.2) can be tweaked depending on the nature of the problem, e.g. to take into account arbitrary serial correlation in panel data or intra-cluster correlation of the residual in survey data.

Default standard errors are derived under a homoskedasticity assumption:

$$E\left[e_i^2|X_i\right] = \sigma^2,$$

which is a constant. This implies the covariance matrix reduces to

$$\sigma^2 E\left[X_i X_i'\right]^{-1}.$$

I think it's safe to say that, these days, most empirical papers report robust standard errors.

4.1.4. Saturated models, main effects and other regression talk

Saturated models are regression models with discrete explanatory variables, where the model includes a separate parameter for all possible values taken on by the explanatory variables. For example, if you're modelling log wage as a function of education, you may construct dummy variables for every level of education:

$$Y_i = \alpha + \beta_1 d_{1i} + \beta_2 d_{2i} + ... + +\beta_\tau d_{\tau i} + \varepsilon_i,$$

where d_{ji} is a dummy for schooling level j and β_j is the jth-level schooling effects. Note that a saturated regression model fits the CEF perfectly, since the CEF is a linear function of the dummy regressors used to saturate (i.e. the CEF takes on τ values only).

If there are two explanatory variables - e.g. one dummy for college education and one dummy for sex, the model is saturated by including these two dummies, their product (an interaction term) and a constant. See AP, pp.50-51 for a nice example and further discussion.

4.1.5. Weighting regression

The premise so far in the discussion has been that the sample is drawn randomly from the population. In fact, in many datasets this is not true by design, since certain types of individuals are 'oversampled'. If sampling weights w_i are available, defined as the inverse of the probability of being included in the sample, we can use weighted regression to mimic a random sample (and, hopefully, get closer to the population object of interest e.g. β). This is easily done in Stata (e.g. by using pweights).

Weighting is sometimes used to solve quite a different problem, namely one posed by heteroskedasticity. The idea here is to transform the dependent and independent variable by means of some suitably chosen weight so as to make the residual variance constant, and then do regression - e.g. weighted least squares. You don't often see this kind of approach these days - perhaps mainly because robust standard errors take care of the main problem posed by heteroskedasticity.