Applied Econometrics

Lecture 1: Introduction

Måns Söderbom*

Department of Economics, University of Gothenburg

1 September 2009

Note: Compared to the notes circulated in class, I have corrected a few spelling mistakes and the formula for robust variance. /ms

*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 identification strategy that enables you to answer the question.
- Mastering statistical techniques - e.g. OLS, 2SLS, GMM, 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.
- These lectures are based on the assumption that you are reasonably comfortable with the material taught in the first year courses in econometrics. In those courses you learned a lot about econometric theory. Building on this, we will stress interpretation and assumptions in this course, not derivations or theorems.
- The course is oriented towards the analysis of cross-section and panel data. Pure time series econometrics will not be covered (though the lectures on the analysis of long panels will be closely related to 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 me **one week after each computer session**, at the latest. We will follow up on these in class, plus there will be a revision class in October where you will be asked to present your solutions.
  - Oral exam (October 29-30th): 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. Also, please provide me with your name and email address, so that I can communicate with you as a group through email.

[A look at the schedule]

1.2. Two textbooks - Two points of view

- Wooldridge (2002) Econometric Analysis of Cross Section and Panel Data. Well written, fairly conventional take on econometrics (at least compared to Angrist and Pischke). Core of the conceptual framework: The population model. Model-based paradigm.

- Angrist and Pischke (2009). Mostly Harmless Econometrics. Brilliant exposition, reflects the 'new' way of thinking about econometrics, which is based on the experimentalist paradigm. Core of the conceptual framework: Potential outcomes. We will use this book intensively in the second part of the course, when discussing econometric methods for program evaluation (estimation of treatment effects).

1.3. Recommended reading for this lecture

Angrist and Pischke (2009), Chapters 1-2, 3.1-3.2.

Wooldridge (2002), Chapters 1-2, 4-5.

2. Questions about questions

Reference: Angrist & Pischke, Chapter 1. Wooldridge, 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.

2.0.1. Your causal relationship of interest

Causal relationship: tells you what will happen to some quantity of interest (expected earnings) as a result of changing the causal variable (e.g. years of schooling), holding other variables fixed. The counterfactual concept is central here - what is the counterfactual of pursuing a different educational policy, for example.

- Causality in the **experimentalist paradigm**: What might have happened to someone who was exposed to a training programme ($D_i = 1$) if that person had **not** been exposed to the programme ($D_i = 0$). In such a case where treatment is binary, the starting point for the analysis is potential outcomes:

$$\text{Potential outcome} = \left\{ \begin{array}{l} Y_{1i} \text{ if } D_i = 1 \\ Y_{0i} \text{ if } D_i = 0 \end{array} \right\}$$

  where - key! - the potential outcomes are **independent** of whether the individual actually participates in the training programme. The causal effect of treatment is defined as the difference

between Y_{1i} and Y_{0i}. Of course, only one of the potential outcomes can be observed, and so the main challenge is to come up with ways of constructing a measure of the potential outcome that we do not observe (the counterfactual). A common quantity of interest is the average treatment effect.

- Causality in the **model-based paradigm**: The causal effect of change in an 'explanatory' variable w on some outcome variable of interest, e.g. the expected value of y. In order to find the causal effect, we must hold all other relevant factors (the control variables) fixed - ceteris paribus analysis. Exactly what those other factors are is, of course, not obvious, which is why economic theory is often used to derive the estimable equation. The basic idea behind running a regression is that this enables you to condition on the control variables. Whether you are allowed a causal interpretation essentially depends on if you've managed to control for all relevant factors determining your outcome variable. .

2.0.2. 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) sample 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 a 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.

2.0.3. 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 yo 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).

2.0.4. Your mode of statistical inference

- The population you're studying.
- Your sample.
- The procedure for calculating standard errors.

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 done most of the hard work - now 'all' that remains is the statistical analysis.

3. Conditional Expectations

- Reference: Wooldridge, Chapter 2.

- Goal of most empirical studies: Find out what is the effect of a variable w on the expected value of y, holding fixed a vector of controls c. That is, we want to establish the **partial effect** of changing w on $E(y|w,c)$,holding c constant. $E(y|w,c)$ is sometimes referred to as a **structural conditional expectation**, where the word "structural" reflects the idea that theory plays an important role in determining the empirical model.

- If w is continuous, the partial effect is

$$\frac{\partial E(y|w,c)}{\partial w},$$

  while if w is a dummy variable, we would look at

$$E(y|w=1,c) - E(y|w=0,c).$$

  Other types of partial effects may be relevant too, depending on the context and the properties of w.

- Estimating **partial effects** such as these in practice is difficult, primarily because of the **unobservability problem**: typically, not all elements of the vector c is observed, and perfectly measured, in your data.

- Much of this course will be concerned with estimation and interpretation in view of precisely this problem. Using a linear regression model, we will study problems posed by omitted variables, and other sources of endogeneity bias, and discuss the leading ways by which such problems can be addressed in practice. We focus mostly on instrumental variable estimation - 2SLS and GMM - and panel data techniques

3.1. Important statistical underpinnings

- Although we will not discuss theoretical results in great detail, it is useful to keep two things in mind from now on, related to statistical theory:
  - The first relates to the **sample** and the **population**. Following Wooldridge (2002), we will usually - though not always, e.g. the sample selection model ("Heckit") - assume there is an **independent identically distributed (i.i.d)** sample drawn from the population. We assume there is a **population model**, for example

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + u,$$

    where $x_1, ..., x_K$ are explanatory variables ("regressors"), and u is a residual. Our general goal is to estimate some or all of the parameters $\beta_0, ..., \beta_K$, based on the sample.
  - The second relates to the **properties of the estimators**: Again, following Wooldridge (2002), we rely on **asymptotic** underpinnings in evaluating econometric estimators, as distinct from finite sample underpinnings. This, essentially, reflects the current state of play in econometrics: econometricians know a lot about the asymptotic properties of estimators, less about the finite sample properties. You may think this is somewhat off-putting - after all, none of us (yes?) has access to a dataset in which $N \to \infty$ (N will denote the number of observations except when we discuss panel data). As we shall see, however, how well an estimator works in practice does not exclusively depend on sample size. How informative your data are is very important too. For example, if you have a very good instrument the "small sample bias" associated with your 2SLS estimates may be negligible, whereas if the instrument is weak the bias might be severe, even in a very large sample.

3.2. Quantities of interest

- Let's start with the issue of functional form, ignoring the residual. Most of the time we study **parametric models**, i.e. models in which the functional form is taken to be "known" a priori.
  - Of course, the most basic parametric model is linear in variables and parameters, e.g.

$$E(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

    and so estimation can be done by means of a linear regression model. The partial effect of (say) x_1 on $E(y|x_1, x_2)$ is simply β_1 here, regardless of whether x_1 is continuous or discrete.
  - Writing the model as nonlinear in variables adds few complications to do with estimation:

$$E(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 \tag{3.1}$$

    or

$$E(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2, \tag{3.2}$$

    because we can still use a linear regression model to estimate all the parameters of the model. For sure, interpretation is a little less straightforward, but this should not be holding us back. (In the latter model, what's the partial effect of x_1, and how do you determine if this effect is significantly different from zero?)
  - However, models that are nonlinear in parameters, e.g.

$$E(y|x_1, x_2) = \Phi(\beta_0 + \beta_1 x_1 + \beta_2 x_2),$$

    where $\Phi(\cdot)$ denotes the cumulative density function for the standard normal distribution, cannot in general be estimated using the linear regression model. We will discuss estimation of such models in the second part of the course. (Incidentally, I find it interesting to note

how little Angrist & Pischke care about nonlinear models of this type. In the old days (say late 1980s and 1990s), estimating a binary choice model with OLS was widely considered a cardinal sin. Well not any more - we will see this in the first computer exercise too.)

- While the partial effect is usually the quantity of interest, sometimes we want to compute the **elasticity**, or perhaps the **semielasticity**, of the conditional expected value of y with respect to (say) x_1. Sticking to the example with two explanatory variables, we have:

$$\begin{aligned} \frac{\partial E\left(y|x_1,x_2\right)}{\partial x_1}\frac{x_1}{E\left(y|x_1,x_2\right)} &\equiv \frac{\partial \log E\left(y|x_1,x_2\right)}{\partial \log x_1} \\ &\approx \frac{\partial E\left(\log\left(y\right)|x_1,x_2\right)}{\partial \log x_1}, \end{aligned} \qquad \text{(Elasticity)}$$

$$\begin{aligned} \frac{\partial E\left(y|x_1,x_2\right)}{\partial x_1}\frac{1}{E\left(y|x_1,x_2\right)} &\equiv \frac{\partial \log E\left(y|x_1,x_2\right)}{\partial x_1} \\ &\approx \frac{\partial E\left(\log\left(y\right)|x_1,x_2\right)}{\partial x_1}. \end{aligned} \qquad \text{(Semi-Elasticity)}$$

In words, the elasticity tells us how much $E\left(y|x_1,x_2\right)$ changes, in percentage terms, in response to a 1% increase in x_1. The semi-elasticity tells us how much $E\left(y|x_1,x_2\right)$ changes, in percentage terms, in response to a one unit increase in x_1. Make sure you can define the elasticities and semi-elasticities for specifications (3.1) and (3.2) above.

4. OLS Estimation

Reference: Wooldridge, Chapter 4.

Consider a population model that is linear in parameters:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + u,$$

where $y, x_1, x_2, ..., x_K$ are observable variables, u is the unobservable random disturbance term (the residual or error term), and $\beta_0, \beta_1, ..., \beta_K$ are parameters that we wish to estimate. Whether OLS is an appropriate estimator depends on the properties of the error term. As you know, for OLS to consistently (remember: asymptotic underpinnings) estimate the β-parameters, the error term must have zero mean and be uncorrelated with the explanatory variables:

$$E(u) = 0,$$

$$Cov(x_j, u) = 0, \quad j = 0, 1, ..., K. \tag{4.1}$$

The zero mean assumption is innocuous, as the intercept β_0 would pick up a non-zero mean in u. The crucial assumption is zero covariance, (4.1). If this assumption does not hold, say because x_1 is correlated with u, we say that x_1 is **endogenous**. This terminology follows the convention in cross-section (micro) econometrics (in traditional usage, a variable is endogenous if it is determined within the context of a model). To illustrate why endogeneity is a problem, consider the simplified model

$$y = \beta_0 + \beta_1 x_1 + u.$$

To simplify the notation, rewrite this in deviations from sample mean, so that I can eliminate the intercept (not a parameter of interest here):

$$\tilde{y} = \beta_1 \tilde{x}_1 + u,$$

where $\tilde{y} = y - \bar{y}$, $\tilde{x} = x - \bar{x}$ (u is mean zero, remember). The OLS estimator is then defined

$$\hat{\beta}_1^{OLS} = \beta_1 + \frac{\sum_i \tilde{x}_{1i} u_i}{\sum_i \tilde{x}_{1i}^2},$$

(consult basic econometrics textbook if this is unclear). Hence:

$$\begin{aligned} p\lim \hat{\beta}_1^{OLS} &= \beta_1 + p\lim \frac{\sum_i \tilde{x}_{1i} u_i}{\sum_i \tilde{x}_{1i}^2}, \\ p\lim \hat{\beta}_1^{OLS} &= \beta_1 + \frac{Cov\left(x_1, u_i\right)}{var\left(x_1\right)} \neq \beta_1, \end{aligned} \tag{4.2}$$

using Slutsky's theorem (see appendix). In other words, the bias does not go away as the sample gets large, since no matter how large your sample is, the covariance between x_1 and u is nonzero.

Endogeneity is thus a rather serious problem, implying that we cannot rely on OLS if the goal is to estimate (causal) partial effects.

4.1. OLS and the method of moments

For reasons that will be clearer later, it is useful to derive the OLS estimator from a set of **moment conditions**, or **population orthogonality conditions**. Using matrix notation, we write the population model (now with observation subscripts explicit) as

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + u,$$

where

$$\boldsymbol{x}_i = \left[\begin{array}{ccccc} 1 & x_{1i} & x_{2i} & ... & x_{Ki} \end{array} \right]$$

is a vector of explanatory variables (the first element in $\boldsymbol{x}$ is constant at 1, reflecting the presence of an intercept in the parameter vector).[1] The zero covariance condition is now written

$$E\left(\boldsymbol{x}'u\right) = \mathbf{0},$$

which is often referred to as a set of moment conditions or orthogonality conditions (notice that $E\left(\boldsymbol{x}'u\right)$ is a $(K+1)\times 1$ column vector=. By definition this implies

$$E\left(\boldsymbol{x}'\left(y - \boldsymbol{x}\boldsymbol{\beta}\right)\right) = \mathbf{0},$$

which yields a solution for $\boldsymbol{\beta}$,

$$\boldsymbol{\beta} = E\left[\left(\boldsymbol{x}'\boldsymbol{x}\right)\right]^{-1} E\left(\boldsymbol{x}'y\right), \tag{4.3}$$

assumed that the $\boldsymbol{x}'\boldsymbol{x}$ matrix is of full rank (ruling out multicollinearity).

O course the RHS of (4.3) is expressed in terms of population moments. By the **analogy principle**,

[1] Throughout these lecture notes I will try to be strict on myself and write vectors in bold - almost certainly, I will not remember to do this all the time. In general, if x has a subscript (e.g. x_1), then it is almost certainly a scalar; if x has no subscript it is probably a vector or matrix; and if I write $\boldsymbol{x}$ it is almost certainly a vector or matrix.

however, we can construct an estimator based on sample moments rather than population moments:

$$\hat{\boldsymbol{\beta}} = \left(N^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i' \boldsymbol{x}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i' y_i \right),$$

or

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(N^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i' \boldsymbol{x}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i' u_i \right). \tag{4.4}$$

Taking plims of this yields the general version of (4.2). As long as $E(\boldsymbol{x}'u) = 0$, we have $p\lim \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$; that is, consistency of the OLS estimator.

This way of deriving the OLS estimator from the moment conditions is very general, and we will see below how various instrumental variable estimators can be derived in a similar fashion.

4.2. Variance estimation

In writing these lectures I will not spend much time deriving variance estimators. This is because I want to concentrate mainly on assumptions and derivations that are interesting from an economic, or even behavioral, point of view. I just derived the OLS estimator from an economically significant assumption, namely $E(\boldsymbol{x}'u) = 0$. If I am to justify estimating a production function by means of OLS, I have to think seriously about the economic factors making up the residual, the firm's demand for labour and capital (say), and whether it makes sense to assume that the residual is uncorrelated with the labour and capital inputs, $E(\boldsymbol{x}'u) = 0$. Thus, economic theory often helps us interpret the partial effects.

Being able to do inference is absolutely crucial for empirical research, and in order to do inference we need to estimate the covariance matrix associated with the parameter estimates. In my view, deriving the covariance matrix is usually economically less interesting than deriving the partial effects. I will therefor not go into great detail about the theoretical origins of the variance estimator. Where there is some interesting economic intuition, I will highlight it.

As you no doubt remember from your first-year econometrics course, the standard formula for the OLS variance estimator is as follows

$$\hat{Var}\left(\hat{\boldsymbol{\beta}}\right) = \hat{\sigma}^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}, \tag{4.5}$$

where $\mathbf{X}$ is the $N \times K$ data matrix of regressors with the ith row $\boldsymbol{x}_i$. Based on this we can compute standard errors, t-values, F-statistics etc. in the usual fashion.

One important assumption underlying (4.5) is that of **homoskedasticity** - i.e. that the variance of the error term u is **constant**. Often, however, this assumption is not supported by the data.

[EXAMPLE: See Section 1 in the appendix.]

Heteroskedasticity may be the result of economically interesting mechanisms (do see any economics in Figure 1?). Or it could be because the dependent variable is measured with more error at high (or low) values of the explanatory variable(s). In any case, the upshot is that if homoskedasticity does not

hold, the conventional variance formula (4.5) is no longer correct. The OLS estimator of $\boldsymbol{\beta}$, however, is still consistent.

A gentle exercise until next time we meet: a) Derive the formula (4.5) under homoskedasticity; b) Show that this formula is wrong under heteroskedasticity

4.2.1. Heteroskedasticity-robust standard errors

For a long time, weighted least squares was the standard cure for heteroskedasticity. This involved transforming the observed variables $\boldsymbol{Y}, \boldsymbol{X}$ in such a way as to make the residual in the transformed regression homoskedastic, and then re-estimating the model with linear regression (i.e. run OLS again). Nowadays, a much more popular approach is to use the OLS estimates of $\boldsymbol{\beta}$ (still consistent, remember) and correct the standard errors so that they are valid in the presence of arbitrary heteroskedasticity. The formula for **heteroskedasticity-robust standard errors** is as follows:

$$\hat{Avar}\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{i=1}^{N}\hat{u}_i^2\boldsymbol{x}_i'\boldsymbol{x}_i\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}, \tag{4.6}$$

and you request this in Stata by adding 'robust' as an option to the command regress. These standard errors - usually attributed by economists to White (1980) - 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 (indeed most empirical papers now seem to favour them), in which case you can remain agnostic as to whether there is or isn't heteroskedasticity in the data. As we shall see, the formula (4.6) 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.

Once robust standard errors have been obtained, you compute t-statistics in the usual way. These t-statistics, of course, are robust to heteroskedasticity.

OLS is probably the most widely used estimator amongst applied economists. Nevertheless, the issue of endogeneity is a potentially serious problem, since, if present, we can't interpret our results causally. We now turn to this issue.

5. Sources of Endogeneity

- We said above that if $Cov(x_j, u) \neq 0$, then the variable x_j is endogenous and OLS is inconsistent. So why might a variable be endogenous? In principle, the problem of endogeneity may arise whenever economists make use of non-experimental data, because in that setting you can never be totally certain what is driving what.

- In contrast, in a perfectly clean experimental setting, where the researcher carefully and exogenously changes the values of the x-variables one by one and observes outcomes y in the process, endogeneity will not be a problem. In recent years, experiments have become very popular in certain areas of applied economics, e.g. development micro economics. However, non-experimental data are still the most common type of information underlying applied research. As we shall see later in this course, the challenge set for themselves by economists adopting the experimentalist paradigm is essentially to mimic clean experiments with their non-experimental data.

- Lots of examples in the literature. In Computer Exercise 1 we will consider the analysis by Miguel, Satyanath and Segenti (JPE, 2004). These authors estimate the impact of economic conditions on the likelihood of civil conflict in Africa during 1981-99. They argue that civil wars may impact on economic relationships and that there may be unobserved factors that impact both on the likelihood of conflict and economic conditions (e.g. governance). For this reason, the correlation between economic conditions and war incidence cannot be interpreted causally. Instrumental variables are used to address this endogeneity problem.

- In the context of non-experimental data, endogeneity typically arises in one of three ways: **omitted variables, measurement errors** and **simultaneity** (Wooldridge, Section 4.1).

5.1. Omitted variables

Omitted variables appear when we would like to - perhaps because economic theory says we should - control for one or more additional variables in our model, but, typically because we do not have the data, we cannot. For example, suppose the correct population model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i,$$

and suppose our goal is to estimate β_1. Think of y_i as log earnings, x_{1i} as years of schooling, and x_{2i} worker ability. We assume that x_1 and x_2 are uncorrelated with the residual:

$$Cov\left(x_1, u\right) = Cov\left(x_2, u\right) = 0.$$

Hence, had we observed both x_1 and x_2, OLS would have been fine.

However, suppose we observe earnings and schooling, but not ability. If we estimate the model

$$y = \gamma_0 + \gamma_1 x_1 + \varepsilon_i,$$

it must be that $\varepsilon_i = (\beta_2 x_{2i} + u_i)$. How will this affect the estimate of γ_1? In particular, is the OLS estimate of γ_1 a consistent estimate of β_1, the parameter of interest?

Modifying (4.2), we can write

$$\begin{aligned} p\lim \hat{\gamma}_1^{OLS} &= \beta_1 + p\lim \frac{\sum_i \tilde{x}_{1i}\left(\beta_2 \tilde{x}_{2i} + u_i\right)}{\sum_i \tilde{x}_{1i}^2}, \\ p\lim \hat{\gamma}_1^{OLS} &= \beta_1 + \beta_2 \frac{Cov\left(x_1, x_2\right)}{var\left(x_1\right)}, \end{aligned}$$

where $\tilde{z} = z - \bar{z}$ for any variable z denotes sample demeaning. Hence, $\hat{\gamma}_1^{OLS}$ will be a consistent estimator of β_1 if $\beta_2 = 0$ or if $Cov\left(x_1, x_2\right) = 0$. In the context of an earnings function this seems unlikely - given the model, the OLS estimate will probably be upward biased (why upward?).

5.2. Measurement Error

Thus far it has been assumed that the data used to estimate the parameters of our models are true measurements of their theoretical counterparts. In practice, this situation happens only in the best of circumstances. When we collect survey data in developing countries, for instance, we try very hard to make sure the information we get from the respondents conforms as closely as possible to the variables we have in mind for our analysis - yet it is inevitable that measurement errors creep into the data. And aggregate statistics, such as GDP, investment or size of the workforce are only estimates of their theoretical counterparts.

Measurement errors may well result in (econometric) endogeneity bias. To see this, consider the classical **error-in-variables** model. Assume that the correct population model is

$$y_i = \beta_0 + \beta_1 x_{1i} + u_i.$$

Hence, with data on y and x_1, the OLS estimator would be fine. Now, suppose we do not observe x_1 - instead we observe a noisy measure of x_1, denoted x_1^{obs}, where

$$x_{1i}^{obs} = x_{1i} + v_i,$$

where v_i is a (zero mean) **measurement error** uncorrelated with the true value x_{1i}. The estimable equation is now

$$y_i = \beta_0 + \beta_1 x_{1i}^{obs} + e_i, \tag{5.1}$$

where $e_i = (u_i - \beta_1 v_i)$. Because the measurement error is a) correlated with x_{1i}^{obs} and b) enters the

residual e_i, the OLS estimate of β_1 based on (5.1) will be inconsistent:

$$\begin{aligned}
p\lim \hat{\beta}_1^{OLS} &= \beta_1 + p\lim \frac{\sum_i \tilde{x}_{1i}^{obs} e_i}{\sum_i \left(\tilde{x}_{1i}^{obs}\right)^2}, \\
p\lim \hat{\beta}_1^{OLS} &= \beta_1 + p\lim \frac{\sum_i \left(\tilde{x}_{1i} + v_i\right)\left(u_i - \beta_1 v_i\right)}{\sum_i \left(\tilde{x}_{1i} + v_i\right)^2}, \\
p\lim \hat{\beta}_1^{OLS} &= \beta_1 + \frac{-\beta_1 \sigma_v^2}{\sigma_{\tilde{x}_1}^2 + \sigma_v^2}, \\
p\lim \hat{\beta}_1^{OLS} &= \beta_1 \left(\frac{\sigma_{x_1}^2}{\sigma_{x_1}^2 + \sigma_v^2}\right)
\end{aligned}$$

where $\sigma_{x_1}^2$ is the variance of the true explanatory variable and σ_v^2 is the variance of the measurement error. Three interesting results emerge here:

- First, $p\lim \hat{\beta}_1^{OLS}$ will always be closer to zero than β_1, so long as $\sigma_v^2 > 0$, i.e. so long as there are measurement errors of the current form. This is often referred to as **attenuation bias** in econometrics ("iron law of econometrics").
- Second, the severity of the attenuation bias depends on the ratio $\sigma_{x_1}^2/\sigma_v^2$, which is known as the **signal-to-noise ratio**. If the variance of x_1 is large, relative to the variance of the measurement error, then the attenuation bias will be small, and vice versa.
- Third, the **sign** of $p\lim \hat{\beta}_1^{OLS}$ will always be the same as that of the structural parameter β_1. Hence, in this model, measurement errors will not change the sign on your coefficient (asymptotically).

The attenuation bias formula is an elegant result. Things become much more complicated when we have more than one explanatory variable. Even if only one variable is measured with error, all estimates of the model will generally be inconsistent. And if several variables are measured with error, matters become even more complex. Unfortunately, the sizes and the directions of the biases are difficult to derive, and above all difficult to interpret, in the multiple regression model.

Not all forms of measurement errors cause substantive problems however. Measurement errors in the dependent variable, for example, increase the standard errors (more noise in the residual) but do not result in inconsistency of the OLS estimator.

5.3. Simultaneity

Simultaneity arises when at least one of the explanatory variables is determined simultaneously along with the dependent variable. Consider for example the following simultaneous population model:

$$y_1 = \alpha_0 + \alpha_1 y_2 + \alpha_2 x_1 + u_1, \tag{5.2}$$

$$y_2 = \beta_0 + \beta_1 y_1 + \beta_2 x_2 + u_2, \tag{5.3}$$

where the notation is obvious. Suppose my goal is to estimate (5.3). The problem is that y_1, both determines, and depends on, y_2. More to the point, because u_2 affects y_2 in (5.3) which in turn affects y_1 through (5.2), it follows that u_2 will be correlated with y_1 in (5.3).

To see this, write down the reduced form for y_1 - you will see that it depends on u_2.

6. The Proxy Variable-OLS Solution to the Omitted Variables Problem

Serious problems thus emerge when a regressor is endogenous. All is not lost however. As we will see in this section, OLS may still provide consistent estimates of the parameters of interest if a **proxy variable** is available. Alternatively, we might be able to use instrumental variable or panel data techniques - more on this later.

Consider the following model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + \gamma q + u, \tag{6.1}$$

where q is an omitted (unobservable) variable. We want to estimate the partial effects of the observed variables, holding the other relevant determinants, including q, constant. As we've already seen in (4.2), if we simply estimate the model whilst putting q in the error term (on the grounds that it is unobserved), there will be omitted variables bias if q is correlated with one or several of the x-variables.

Now suppose a **proxy variable** z is available for the unobserved variable q. As the name suggests, this is a variable thought to be highly correlated with the unobserved variable q, and so we might be able to reduce or eliminate the bias in the estimated β_j if we include z in the set of explanatory variables.

There are two formal requirements for a proxy variable for q:

1. The proxy variable must be **redundant** in the structural equation (6.1):

$$E(y|\mathbf{x}, q, z) = E(y|\mathbf{x}, q).$$

   This is pretty obvious and uncontroversial - if z is already in the model for structural reasons, then clearly it cannot be used to proxy for an omitted variable as well.

2. The proxy variable must be such that the correlation between the omitted variable q and each x_j goes to zero, once we **condition** on q.

Let's have a closer look at the second condition. Define

$$q = \theta_0 + \theta_1 z + r,$$

where r should be thought of as variation in q not correlated with z. Note that this equation should not be given a causal interpretation. The reason is that $\theta_0 + \theta_1 z$ is defined simply as the linear projection of q on z. For z to be a good proxy for q, we then require:

$$\begin{aligned} E(r) &= 0 \text{ (holds by def.)} \\ Cov(z, r) &= 0 \text{ (holds by def.)} \\ \theta_1 &\neq 0 \text{ (if not, useless proxy)} \\ Cov(x_j, r) &= 0 \text{ (crucial!).} \end{aligned}$$

The last condition thus requires z to be closely enough related to q so that once it is included in the regression, the x_j are not partially correlated with q.

- To see that this works - under the conditions stated above - we rewrite the structural equation as

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + \gamma(\theta_0 + \theta_1 z + r) + u \\ y &= (\beta_0 + \theta_0) + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + \gamma\theta_1 z + (\gamma r + u). \end{aligned}$$

  You now see the assumption $Cov(x_j, r) = 0$ really is crucial.

- EXAMPLE: Proxying unobserved ability by IQ in an earnings regression (clearly IQ is not the same as ability, but we may reasonably suppose that IQ is **correlated** with ability). See Section 2 in the appendix.

- This all sounds rather promising, but it must be stressed that using a proxy variable can still lead to bias - in the model above this would happen if r is correlated with x_j, as already noted. In

the context of the Blackburn-Neumark regression, this could happen if, conditional on IQ, there remains a correlation between education and ability. Of course, the resulting bias may still be smaller than if we ignored the problem of omitted ability entirely.

- Hence the rule of thumb (again): *A good proxy must be such that, conditional on the proxy variable, the unobserved variable does not vary with the observed variable(s).*

- Sometimes it is helpful to have several proxy variables. Consider Tables 2.3-4 in the appendix, where we bring in KWW (a test score on the knowledge of the world of work) as an additional proxy for ability.

7. Instrumental Variables Estimation

7.1. Setting the scene

Reference: Wooldridge, Chapter 5.

The Instrumental Variables (IV) approach recognizes that the residual and the explanatory variable(s) may be correlated, and uses additional information to 'purge' the endogenous explanatory variable(s) of the part correlated with the residual in the structural equation. Consider a linear population model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + u, \tag{7.1}$$

where $E(u) = 0$, and $cov(x_j, u) = 0$, for $j = 1, 2, ..., K-1$, but where x_K might be correlated with u. Thus, while $x_1, x_2, ..., x_{K-1}$ are all exogenous, x_K is potentially endogenous.

In general, there may be several endogenous regressors, but for now it is helpful to concentrate on the case where there is at most one. Of course, x_K may be endogenous for any of the reasons discussed above, but, for the current purposes, it does not matter why x_K is endogenous.

Let's assume there is an omitted (unobserved) variable q that is a component of the residual, $u = q + e$, and also potentially correlated with x_K. You might find it helpful to think of (7.1) as an earnings equation, where x_K is years of schooling and q is unobserved ability. In any case, as we saw above, if $cov(q, x_K) \neq 0$, then OLS estimation of (7.1) generally results in inconsistent estimates of all the coefficients in the model.

The method of IV provides a general solution to problems posed by the presence of one or many endogenous explanatory variables in the model. To use this method we need an observed variable z_1, referred to as an **instrument**, that satisfies two conditions.

The **first** condition is that the instrument is exogenous, or **valid**:

$$cov(z_1, u) = 0.$$

This is often referred to as an **exclusion restriction,** on the grounds that z_1 is excluded from the

structural equation (7.1).

The **second** condition is that the instrument is **informative**, or **relevant**. This means that the instrument z_1 must be correlated with the endogenous regressor x_K, conditional on all exogenous variables in the model (i.e. $x_1, x_2, ..., x_{K-1}$). That is, if we assume there is a linear relationship between x_K and z_1 and $x_1, x_2, ..., x_{K-1}$,

$$x_K = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + r_K, \tag{7.2}$$

where r_K is mean zero and uncorrelated with all the variables on the right-hand side, we require $\theta_1 \neq 0$. Notice that if there are no exogenous variables in the structural model, this condition reduces to $cov(z, x_k) \neq 0$, which may be easier to relate to ("the instrument must be correlated with the endogenous explanatory variable"). The equation (7.2) is often referred to as the **reduced form equation** for x_K. As this name suggests, there is nothing necessarily structural about this equation. For example, if you assume that, in the earnings equation, work experience is exogenous and schooling endogenous, then the reduced form equation for schooling contains work experience as an "explanatory" variable. This does not make sense in a structural sense (the future can't determine the past...), but it is fine as a reduced form relationship.

Based on the reduced form equation for x_K, we can obtain a reduced form equation for the dependent variable of interest (i.e. y, or "earnings"), by plugging (7.2) into (7.1):

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_{K-1} x_{K-1} + \lambda_1 z_1 + v,$$

where the reduced form parameters $\alpha_0, ..., \alpha_{K-1}, \lambda_1$ are functions of the structural parameters $\beta_0, ..., \beta_K$. You can easily verify that, given the assumptions we have made above, the residual v is uncorrelated with all the explanatory variables on the right-hand side. Thus the reduced form equation for y can be estimated consistently using OLS. Estimating reduced form parameters is sometimes useful, for example if the analysis is essentially descriptive. As we shall see later, reduced form equations can also be very

useful in the context of hypothesis testing. However, if we want to pin down the causal effect of x_K on expected y, we have to estimate the parameters of (7.1).

Let's end by discussing some variables that may or may not be valid instruments for education in an earnings equation.

Which of the following variables do you think could be a good instrument for education?

- The individual's wage last year;
- Number of siblings;
- The individual's IQ;
- Mother's education.

It is vital to understand the difference between a proxy variable and an instrument. In fact, a good proxy variable typically makes a particularly bad instrument. Make sure you understand why.

PhD Programme: Applied Econometrics
Department of Economics, University of Gothenburg
Appendix Lecture 1

Måns Söderbom

1. OLS: Illustration of heteroskedasticity

Earnings and education in Kenya
Researchers wish to estimate the effect of years of education on earnings in Kenya, using a sample of 950 individuals drawn randomly from the population of wage employees in the manufacturing sector. Data on monthly wages and years of education are available, collected in 2000. The basic earnings model is as follows:

$$lw_i = \beta_0 + \beta_1 \cdot ed_i + residual_i,$$

where lw_i is the natural logarithm of monthly wages (in USD) for individual *i*, ed_i is years of education, $residual_i$ is a residual, and β_0, β_1 are parameters to be estimated (of course, this specification is almost certainly too simplistic to be viewed as a structural model of earnings, we use it here for illustrative purposes).

Figure 1.1 Log earnings and years of education: Clear evidence of heteroskedasticity

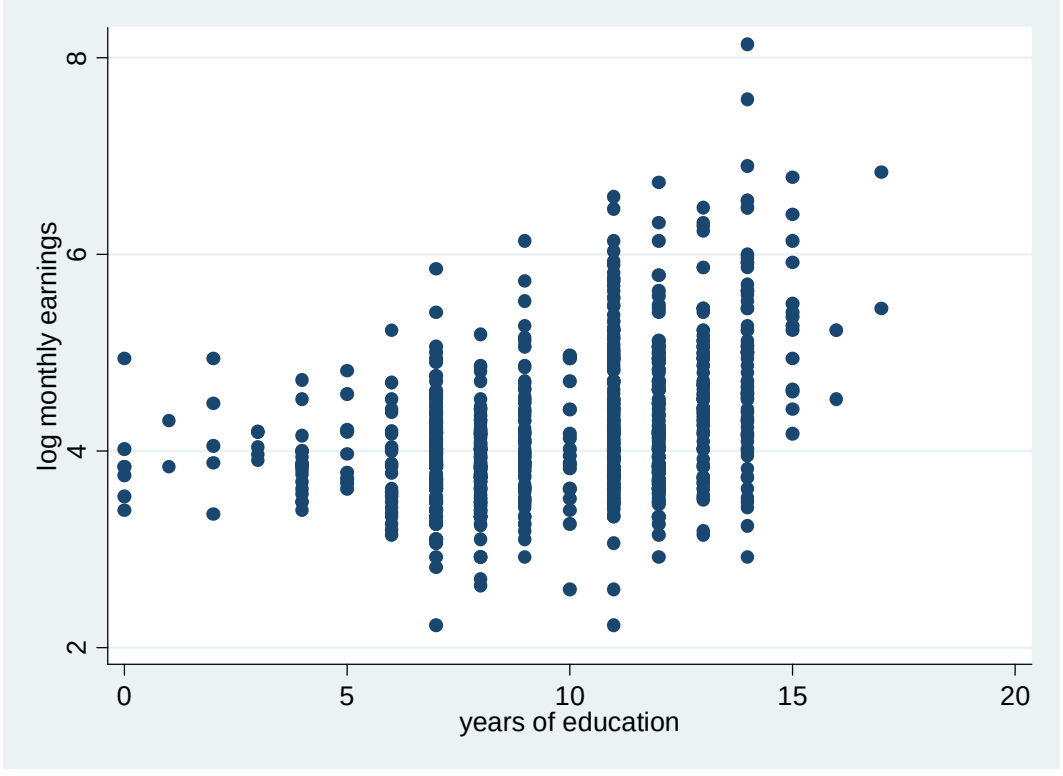

Table 1.1: OLS estimates, variance formula assumes homoskedasticity

```
. reg lw ed;

      Source |       SS       df       MS              Number of obs =     950
-------------+------------------------------           F(  1,   948) =  181.38
       Model |  84.4673729     1  84.4673729            Prob > F      =  0.0000
    Residual |   441.47884   948  .465694979           R-squared     =  0.1606
-------------+------------------------------           Adj R-squared =  0.1597
       Total |  525.946213   949  .554210973           Root MSE      =  .68242

------------------------------------------------------------------------------
          lw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ed |   .1042316   .0077394    13.47   0.000     .0890433    .1194198
       _cons |   3.169813   .0800051    39.62   0.000     3.012805     3.32682
------------------------------------------------------------------------------

/* extract predictions useful for homoskedasticity test */
. predict e, res;
. predict lwhat, xb;
```

Table 1.2: OLS estimates, heteroskedasticity-robust standard errors

```
. reg lw ed, robust;

Linear regression                                       Number of obs =     950
                                                        F(  1,   948) =  148.05
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.1606
                                                        Root MSE      =  .68242

------------------------------------------------------------------------------
             |               Robust
          lw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ed |   .1042316   .0085662    12.17   0.000     .0874206    .1210425
       _cons |   3.169813   .0808514    39.21   0.000     3.011144    3.328481
------------------------------------------------------------------------------
```

Table 1.3: Test H0: Error variance constant

```
. ge e2=e^2;
. reg e2 ed;

      Source |       SS       df       MS              Number of obs =     950
-------------+------------------------------           F(  1,   948) =   34.90
       Model |  25.2475331     1  25.2475331            Prob > F      =  0.0000
    Residual |  685.772101   948  .723388292           R-squared     =  0.0355
-------------+------------------------------           Adj R-squared =  0.0345
       Total |  711.019634   949  .749230384           Root MSE      =  .85052

------------------------------------------------------------------------------
          e2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ed |   .0569855   .0096459     5.91   0.000     .0380558    .0759152
       _cons |  -.1013612   .0997131    -1.02   0.310    -.2970452    .0943228
------------------------------------------------------------------------------
```

2. The Proxy Variable Approach

IQ as a proxy for ability in the earnings equation

This example, taken from Wooldridge (2002), p.65, illustrates the effects of using IQ as a proxy for unobserved ability in an earnings regression. The data, provided in the file NLS80, were originally used by Blackburn and Neumark (1992; *QJE*).

First, I consider OLS results with ability put (implicitly) in the residual:

```
. use C:\teaching_gbg07\applied_econ07\lectures\wooldat\NLS80.dta, clear;
```

Table 2.1: Unobserved ability goes into the residual

```
. reg lwage exper tenure married south urban black educ, robust;

Linear regression                                      Number of obs =     935
                                                       F(  7,   927) =   50.83
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.2526
                                                       Root MSE      =  .36547

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |    .014043   .0032386     4.34   0.000     .0076872    .0203988
      tenure |   .0117473   .0025387     4.63   0.000      .006765    .0167295
     married |   .1994171   .0396937     5.02   0.000     .1215171    .2773171
       south |  -.0909036    .027363    -3.32   0.001    -.1446043    -.037203
       urban |   .1839121   .0271125     6.78   0.000     .1307031     .237121
       black |  -.1883499   .0367035    -5.13   0.000    -.2603816   -.1163182
        educ |   .0654307   .0064093    10.21   0.000     .0528524    .0780091
       _cons |   5.395497   .1131274    47.69   0.000     5.173481    5.617512
------------------------------------------------------------------------------
```

The implied return to education is 6.5%. Now I add IQ as an explanatory variable:

Table 2.2: Unobserved ability proxied for by IQ

```
. reg lwage exper tenure married south urban black educ iq, robust;

Linear regression                                       Number of obs =     935
                                                        F(  8,   926) =   48.51
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.2628
                                                        Root MSE      =  .36315

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0141458   .0032382     4.37   0.000     .0077908    .0205009
      tenure |   .0113951   .0025368     4.49   0.000     .0064166    .0163736
     married |   .1997644   .0390896     5.11   0.000       .12305    .2764789
       south |  -.0801695   .0277381    -2.89   0.004    -.1346063   -.0257327
       urban |   .1819463   .0267419     6.80   0.000     .1294646     .234428
       black |  -.1431253   .0376459    -3.80   0.000    -.2170064   -.0692442
        educ |   .0544106    .007273     7.48   0.000     .0401372    .0686841
          iq |   .0035591   .0009564     3.72   0.000     .0016822    .0054361
       _cons |   5.176439   .1212236    42.70   0.000     4.938534    5.414344
------------------------------------------------------------------------------
```

The coefficient on education falls to 0.054. The estimated coefficient on IQ is positive and statistically significant. Both findings are as one would expect, in this context.

Now add KWW (another test score, this time on the "knowledge of the world of work"):

Table 2.3: Unobserved ability proxied for by IQ and KWW

```
. reg lwage exper tenure married south urban black educ iq kww, robust;

Linear regression                                       Number of obs =     935
                                                        F(  9,   925) =   43.68
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.2662
                                                        Root MSE      =  .36251

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0127522   .0032973     3.87   0.000     .0062811    .0192233
      tenure |   .0109248   .0025822     4.23   0.000     .0058572    .0159924
     married |   .1921449    .038701     4.96   0.000     .1161931    .2680968
       south |  -.0820295   .0277071    -2.96   0.003    -.1364055   -.0276534
       urban |   .1758226    .026732     6.58   0.000     .1233601     .228285
       black |  -.1303995   .0391814    -3.33   0.001    -.2072942   -.0535048
        educ |   .0498375   .0078449     6.35   0.000     .0344417    .0652333
          iq |   .0031183   .0009589     3.25   0.001     .0012364    .0050001
         kww |    .003826   .0020365     1.88   0.061    -.0001707    .0078226
       _cons |   5.175643   .1209569    42.79   0.000     4.938262    5.413025
------------------------------------------------------------------------------
```

Finally, add interaction terms with education to the specification in Table 2.3:

```
. ge ediq=educ*(iq-100);

. sum kww;

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         kww |       935    35.74439    7.638788         12         56

. scalar kwwbar=r(mean);

. ge edkww=educ*(kww-kwwbar);
```

Table 2.4: Unobserved ability proxied for by IQ and KWW. Education interacted with IQ and KWW

```
. reg lwage exper tenure married south urban black educ iq kww ediq edkww,
robust;

Linear regression                                      Number of obs =     935
                                                       F( 11,   923) =   37.01
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.2728
                                                       Root MSE      =  .36127

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0121544   .0032826     3.70   0.000     .0057121    .0185966
      tenure |   .0107206   .0025668     4.18   0.000     .0056833     .015758
     married |    .197827   .0383873     5.15   0.000     .1224904    .2731636
       south |  -.0807609    .027732    -2.91   0.004    -.1351859   -.0263358
       urban |    .178431   .0267696     6.67   0.000     .1258948    .2309673
       black |  -.1381481   .0392285    -3.52   0.000    -.2151355   -.0611607
        educ |    .045241   .0079435     5.70   0.000     .0296517    .0608304
          iq |   .0048228   .0055537     0.87   0.385    -.0060766    .0157222
         kww |  -.0248007   .0106484    -2.33   0.020    -.0456986   -.0039028
        ediq |  -.0001138   .0004174    -0.27   0.785    -.0009329    .0007054
       edkww |    .002161   .0007877     2.74   0.006     .0006152    .0037068
       _cons |   6.080006   .5117145    11.88   0.000     5.075747    7.084264
------------------------------------------------------------------------------
```

Interpret these results, and you will have completed Problem 4.11 in Wooldridge (2002) in the process.