Econometric Methods (Econometrics I)

Lecture 3:

The Single-Equation Linear Model and OLS

Prof. Dr. Kai Carstensen

Kiel University

Winter Term 2023/2024

Outline of this lecture

1. Introduction
2. Assumptions and identification
3. Asymptotic properties of OLS
4. Omitted variables
5. Measurement error

Reference: Wooldridge, Chapter 4.

1. Introduction

The population model

Let us start with the workhorse model in econometrics, the linear model

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

where

- y is the dependent random variable,
- $x_1, \ldots, x_K$ are explanatory random variables,
- $y, x_1, \ldots, x_K$ are observable using a random sample of the population,
- u is the unobservable random disturbance, and
- $\beta_0, \ldots, \beta_K$ are the parameters we would like to estimate.

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

- This is fairly general, as **x** can include nonlinear functions of underlying variables, such as logarithms, squares, reciprocals, interactions...
- Example:

$$\begin{aligned} \log(wage) \quad = \quad & \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 educ \cdot exper \\ & + \beta_4 exper^2 + \beta_5 female + \beta_6 female \cdot educ + u. \end{aligned}$$

- Ultimately, we hope this equation allows us to obtain reliable partial or marginal effects on $E(wage|educ, exper, female)$.
- Violations:
  - A model nonlinear in parameters *may* be more appropriate (for example, when the range of y is restricted, such as binary, fractional, or nonnegative).
  - Perhaps the coefficients on the independent variables should also be viewed as random variables (although this does not prevent us from writing the above equation with a complicated error term).
- Later we will come back to the idea that this is simply the best linear approximation to $\log(wage)$.

Random sampling

We assume that we can collect a random sample – that is, independent and identically distributed outcomes – from the underlying population.

Given randomly sampled observations $\{(\mathbf{x}_i, y_i) : i = 1, ..., N\}$ satisfying the population model, write for the random draws

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK} + u_i, \quad i = 1, \ldots, N,$$

where N is the sample size.

There violations of random sampling, including (i) stratified sampling, (ii) missing data (sample- or self-selection?), (iii) cluster sampling. We will not discuss them in this course.

Notation

For notational convenience, absorb the intercept into the vector $\mathbf{x}$ and write

$$y = \mathbf{x}\boldsymbol{\beta} + u,$$

where $\mathbf{x}$ is $1 \times K$, with the convention that the first element x_1 is unity.

None of the main large-sample results rely on $x_1 \equiv 1$, but it is almost always true in practice.

For a random draw i we write

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + u_i.$$

With random sampling, the remaining assumptions can (and should) be stated in terms of the population.

2. Assumptions and identification

Assumption OLS.1

Assumption OLS.1 (Population orthogonality condition): In the population model $y = \mathbf{x}\boldsymbol{\beta} + u$, the condition $\mathrm{E}(\mathbf{x}'u) = \mathbf{0}$ holds.

Because $\mathbf{x}$ contains a constant, Assumption OLS.1 is equivalent to

(a) mean zero: $\mathrm{E}(u) = 0$ and

(b) uncorrelatedness: $Cov(x_j, u) = 0$, $j = 2, ..., K$.

Violations:

- An *astructural* perspective is to just use the definition of the model and assumption OLS.1 to *define* $\boldsymbol{\beta}$.
- When we begin with an underlying *structural* model, $E(\mathbf{x}'u) = \mathbf{0}$ is often violated in the presence of
  (a) omitted variables
  (b) measurement error
  (c) simultaneity

Sufficient condition: $\mathrm{E}(u|\mathbf{x}) = 0 \quad \Rightarrow \quad \mathrm{E}(\mathbf{x}'u) = \mathbf{0}$.

- $\mathrm{E}(u|\mathbf{x}) = 0$ also implies $\mathrm{E}(u) = 0$.
- Recall that $\mathrm{E}(u|\mathbf{x}) = 0$ implies

$$E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} = \beta_1 + \beta_2 x_2 + \ldots + \beta_K x_K.$$

- Very important: The difference between $E(\mathbf{x}'u) = \mathbf{0}$ and $E(u|\mathbf{x}) = 0$ is substantive.
- Under $E(u|\mathbf{x}) = \mathbf{0}$ we assume that u is uncorrelated with *any* function of $\mathbf{x}$.
- Under $E(\mathbf{x}'u) = \mathbf{0}$ we solely assume that u is uncorrelated with $\mathbf{x}$.

- A practical consequence of $E(u|\mathbf{x}) = 0$ is that all functions of the covariates affecting the population regression $E(y|\mathbf{x})$ have been accounted for in our choices of $x_2, \ldots, x_K$.
- Usually we hope to have $E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$ when we think it makes sense to condition on the explanatory variables. If we include squares, interactions, logarithms, and other nonlinear functions in $\mathbf{x}$ then we are (at least implicitly) trying to get closer to $E(y|\mathbf{x})$.
- In reality we should probably settle for $E(\mathbf{x}'u) = \mathbf{0}$ even if we put lots of nonlinear functions in $\mathbf{x}$. For example, when y has discreteness, or its range is limited in some important way — say $y \in \{0, 1\}$, or $0 \leq y \leq 1$, or $y \geq 0$ — the linear model for $E(y|\mathbf{x})$ cannot hold over a wide range of x_j.

Digression: OLS as an approximation to nonlinear models

- Often we have to concede that a model linear in $\mathbf{x}$ which we estimate by OLS is an approximation to the much more complicated reality.
- Then it is important to ask: In what sense does such a model approximate a general nonlinear regression function? Hence, what does OLS estimate?
- To answer the question, let us define the true (nonlinear) regression function $\mu(\mathbf{x}) = E(y|\mathbf{x})$ and the linear projection of y on $(1, \mathbf{x})$ (for emphasis we separate out the intercept here), denoted

$$L(y|1, \mathbf{x}) = \alpha + \mathbf{x}\boldsymbol{\beta}.$$

- In the following, we present three results:
  (1) The linear projection $L(y|1, \mathbf{x})$ provides the best linear (population) mean square error approximation to the true nonlinear regression function.
  (2) OLS estimates the linear projection parameters $\boldsymbol{\beta}$ consistently.
  (3) Under some strong assumptions, the slope parameters $\boldsymbol{\beta}$ of the linear projection approximate the average partial effects of the true nonlinear regression function.

Digression: OLS as an approximation to nonlinear models

(1) The linear projection $L(y|1,\mathbf{x})$ provides the best linear (population) mean square error approximation to the true nonlinear regression function

- Recall the linear projection $L(y|1,\mathbf{x})$ is such that the parameters solve

$$\min_{a,\mathbf{b}} \; E[(y - a - \mathbf{xb})^2].$$

- It is easily shown that α and $\boldsymbol{\beta}$ also solve

$$\min_{a,\mathbf{b}} \; E[(\mu(\mathbf{x}) - a - \mathbf{xb})^2]$$

because the difference between the two is just a constant unaffected by α and $\boldsymbol{\beta}$. (To show it, simply replace y by $\mu(\mathbf{x}) + u$ and use the properties of the linear prediction error u.)

- In other words, the linear projection $L(y|1,\mathbf{x})$ provide the best linear (population) mean square error approximation to the true nonlinear regression function $\mu(\mathbf{x})$.
- Note that this holds even when $\mathbf{x}$ includes functions such as $exper^2$, $female \cdot educ$, and $\log(firmsize)$.

- This is a preview of a result shown later. But it is straightforward.
- Recall that the linear projection parameters are given by

$$\beta = [\mathrm{E}(\mathbf{x}'\mathbf{x})]^{-1}\,\mathrm{E}(\mathbf{x}'y).$$

and the OLS estimators are constructed as

$$\hat{\beta}_{LS} = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i'y_i\right).$$

- Under general conditions, $\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i \xrightarrow{\mathrm{p}} \mathrm{E}(\mathbf{x}'\mathbf{x})$ and $\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i'y_i \xrightarrow{\mathrm{p}} \mathrm{E}(\mathbf{x}'y)$. Hence, by the theorem of Slutzky,

$$\hat{\beta}_{LS} \xrightarrow{\mathrm{p}} [\mathrm{E}(\mathbf{x}'\mathbf{x})]^{-1}\,\mathrm{E}(\mathbf{x}'y) = \beta.$$

- In following, we only consider continuous regressors but similar results exist for discrete regressors.
- For a continuous variable x_j, its **partial effect** is

$$\frac{\partial \mu(\mathbf{x})}{\partial x_j},$$

  which is a function of $\mathbf{x}$. Hence, for general nonlinear $\mu(\mathbf{x})$, the partial effect depends on all regressors $\mathbf{x}$ and thus differs between individuals.
- To make a summary statement, economists often report the **average partial effect** by averaging the partial effect across the distribution of $\mathbf{x}$:

$$APE_j = E_{\mathbf{x}}\left[\frac{\partial \mu(\mathbf{x})}{\partial x_j}\right] \equiv \gamma_j$$

- Note that APE_j is a constant (parameter).

- Let $\boldsymbol{\beta}$ again denote the slope parameters of the linear projection.
- Under some assumptions, of which multivariate normality of $\mathbf{x}$ might be the strongest, Stoker (1986, *Econometrica*) shows that

$$\beta_j = APE_j = E_{\mathbf{x}}\left[\frac{\partial \mu(\mathbf{x})}{\partial x_j}\right], \qquad \text{for all } j.$$

- Multivariate normality is very strong and usually unrealistic, but it suggests that linear regression more generally approximates quantities of interest: the APEs.
- We can allow nonconstant partial effects in regression by using flexible functions of the covariates, so we can approximate partial effects that depend on the values of covariates.

Example:

- Suppose the nonlinear regression function is

$$\mu(\mathbf{x}) = \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2)$$

with partial effect of interest

$$\frac{\partial \mu(\mathbf{x})}{\partial x_1} = \delta_1 \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2).$$

- Then the average partial effect is

$$APE_1 = E_{\mathbf{x}}\left[\frac{\partial \mu(\mathbf{x})}{\partial x_1}\right] = \int_{x_1}\int_{x_2} \delta_1 \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2) f(x_1, x_2) dx_2 dx_1.$$

- The parameter β_1 of the linear regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

approximates APE_1.

Assumption OLS.1

- Return to the population representation

$$y = \mathbf{x}\boldsymbol{\beta} + u, \qquad E(\mathbf{x}'u) = \mathbf{0}$$

- For the following, we assume there are good reasons for wanting to estimate $\boldsymbol{\beta}$ even if $E(y|\mathbf{x}) \neq \mathbf{x}\boldsymbol{\beta}$.

Assumption OLS.2

Assumption OLS.2 (No Perfect Collinearity): In the population, there are no exact linear relationships among the covariates:

$$\text{rank } \mathrm{E}(\mathbf{x}'\mathbf{x}) = K.$$

When an intercept is included in $\mathbf{x}$, the assumption implies that the population variance-covariance matrix of the regressors is invertible.

Violations:

- None in interesting applications.
- While a high correlation (multicollinearity) among regressors often cannot be avoided, a violation of the assumption typically can.
- Sometimes perfect collinearity shows up because the model is not properly specified, e.g., if we include too many dummy variables ("dummy variables trap", see below) or mistakenly use regressors like $\log(age)$ and $\log(age^2)$.

Identification

- Under OLS.1 and OLS.2, $\boldsymbol{\beta}$ is *identified*.
- In the context of a *linear* model that means we can write $\boldsymbol{\beta}$ as a function of population moments in *observable* variables.
- To see this, multiply the population model

$$y = \mathbf{x}\boldsymbol{\beta} + u$$

by $\mathbf{x}'$ from the left, take expectations and solve for $\boldsymbol{\beta}$:

$$\begin{aligned} \mathrm{E}(\mathbf{x}'y) &= \mathrm{E}(\mathbf{x}'\mathbf{x}\boldsymbol{\beta} + \mathbf{x}'u) \\ &= \mathrm{E}(\mathbf{x}'\mathbf{x})\boldsymbol{\beta} \qquad\qquad (\text{because } \mathrm{E}(\mathbf{x}'u) = \mathbf{0}) \end{aligned}$$

$$\Rightarrow \boldsymbol{\beta} = [\mathrm{E}(\mathbf{x}'\mathbf{x})]^{-1}\,\mathrm{E}(\mathbf{x}'y)$$

- This has nothing to do with data! $\mathbf{A} \equiv \mathrm{E}(\mathbf{x}'\mathbf{x})$ is a $K \times K$ matrix of variances and covariances in the population; $\mathrm{E}(\mathbf{x}'y)$ is essentially a $K \times 1$ vector of population covariances.

Assumption OLS.3

Assumption OLS.3 (Homoskedasticity): $\mathrm{E}(u^2\mathbf{x}'\mathbf{x}) = \sigma^2\,\mathrm{E}(\mathbf{x}'\mathbf{x})$, where $\sigma^2 = \mathrm{E}(u^2)$.

- Think element-wise: Assumption OLS.3 implies that u^2 is uncorrelated with each x_j and all functions $x_j x_h$ for all j and h (including $j = h$).
- Sufficient (but stronger than OLS.3) is

$$\mathrm{E}(u^2|\mathbf{x}) = \sigma^2.$$

- If we start with $\mathrm{E}(u|\mathbf{x}) = 0$, then $\mathrm{E}(u^2|\mathbf{x}) = \sigma^2$ is the same as

$$\mathrm{Var}(u|\mathbf{x}) = \mathrm{Var}(u) \equiv \sigma^2,$$

which essentially gets us to the Gauss-Markov assumptions for cross section data.

Violations:

- Whether OLS.3 is satisfied is always an empirical issue.
- Homoskedasticity is often violated, especially if the range of y is limited in some way (especially discrete).
- If $\mathbf{x}\boldsymbol{\beta}$ represents a linear projection and $\mathrm{E}(y|\mathbf{x}) = \mu(\mathbf{x}) \neq \mathbf{x}\boldsymbol{\beta}$, heteroskedasticity is almost certain even if $\mathrm{Var}(y|\mathbf{x})$ is constant. To see this consider the decomposition

$$\mathrm{E}(u^2|\mathbf{x}) = \mathrm{Var}(y|\mathbf{x}) + [\mu(\mathbf{x}) - \mathbf{x}\boldsymbol{\beta}]^2$$

  and note that the second term is a function of $\mathbf{x}$ if $\mu(\mathbf{x}) \neq \mathbf{x}\boldsymbol{\beta}$.
- How to find this decomposition? Write $y = \mu(\mathbf{x}) + e$ and $u = y - \mathbf{x}\boldsymbol{\beta} = e + [\mu(\mathbf{x}) - \mathbf{x}\boldsymbol{\beta}]$, square both sides, and then condition on $\mathbf{x}$. Note that $\mathrm{E}(e|\mathbf{x}) = 0$ and so e is uncorrelated with any function of $\mathbf{x}$.

3. The OLS estimator

The OLS estimator

In population, the parameters of interest are

$$\boldsymbol{\beta} = [\mathrm{E}(\mathbf{x}'\mathbf{x})]^{-1}\,\mathrm{E}(\mathbf{x}'y).$$

Analogy principle: replace the population moments $\mathrm{E}(\mathbf{x}'\mathbf{x})$ and $\mathrm{E}(\mathbf{x}'y)$ with the corresponding sample moments.

This is a method of moments estimator. It coincides with the OLS estimator, which minimizes the sum of the squared distances between the *data* and the *sample* regression line:

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

Matrix form:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

where $\mathbf{X}$ is the $N \times K$ data matrix of regressors with ith row $\mathbf{x}_i$ and $\mathbf{Y}$ is the $N \times 1$ data vector with ith element y_i.

Mean of the OLS estimator

Substitute the model into the expression for the OLS estimator:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' [\mathbf{x}_i \boldsymbol{\beta} + u_i] \right) \\ &= \boldsymbol{\beta} + \left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' u_i \right).\end{aligned}$$

Now take expectations:

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathrm{E}\left[\left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' u_i \right) \right]$$

With a minimal set of assumptions, particularly with just $\mathrm{E}(\mathbf{x}'u) = \mathbf{0}$, we cannot evaluate this expectation. Hence, in general the OLS estimator is biased.

However, assuming $\mathrm{E}(u|\mathbf{x}) = \mathbf{0}$, we can show that the OLS estimator is unbiased. First take conditional expectations:

$$\mathrm{E}(\hat{\boldsymbol{\beta}}|\mathbf{x}_1, \ldots, \mathbf{x}_N) = \boldsymbol{\beta} + \mathrm{E}\left[\left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' u_i \right) \middle| \mathbf{x}_1, \ldots, \mathbf{x}_N \right]$$

Now put the $\mathbf{x}_i$'s outside the conditional expectation operator and use $\mathrm{E}(u|\mathbf{x}) = \mathbf{0}$:

$$\mathrm{E}(\hat{\boldsymbol{\beta}}|\mathbf{x}_1, \ldots, \mathbf{x}_N) = \boldsymbol{\beta} + \left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \underbrace{\mathrm{E}(u_i|\mathbf{x}_i)}_{=0} \right) = \boldsymbol{\beta}$$

By the LIE,

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \mathrm{E}[\mathrm{E}(\hat{\boldsymbol{\beta}}|\mathbf{x}_1, \ldots, \mathbf{x}_N)] = \mathrm{E}[\boldsymbol{\beta}] = \boldsymbol{\beta}.$$

4. Asymptotic properties of OLS

Asymptotic properties of OLS

Since we cannot even show in general that the OLS estimator is unbiased in finite samples, we resort to asymptotic properties. We will show the following key results:

1. The OLS estimator is consistent. Interpretation: in large samples it is "approximately correct".
2. The OLS estimator is asymptotically normally distributed. Interpretation: in large samples we may approximate the unknown finite-sample distribution of the OLS estimator by its "asymptotic distribution".

Consistency of OLS

Key Result 1: Under Assumptions OLS.1 and OLS.2, OLS on a random sample is consistent (as $N \to \infty$) for $\boldsymbol{\beta}$: $\text{plim}\, \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$.

Proof: Start from the OLS representation and use Slutzky's theorem:

$$\text{plim}\, \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \text{plim} \left[\left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' u_i \right) \right]$$

$$= \boldsymbol{\beta} + \left(\text{plim}\, N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left(\text{plim}\, N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' u_i \right)$$

Let us consider

$$\text{plim}\, N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \qquad \text{and} \qquad \text{plim}\, N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' u_i$$

separately.

(a) plim $N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i$:

The matrix $\mathbf{x}_i' \mathbf{x}_i$ consists of elements (scalar random variables)

$$v_{gh,i} \equiv x_{g,i} x_{h,i}, \qquad g, h = 1, \ldots, K.$$

By assumption, $v_{gh,i}$ is independent and identically distributed over i.

Hence, it satisfies the WLLN:

$$\text{plim } N^{-1} \sum_{i=1}^{N} v_{gh,i} = \text{plim } N^{-1} \sum_{i=1}^{N} x_{g,i} x_{h,i} = E(v_{gh,i}) = E(x_{g,i} x_{h,i}).$$

As the WLLN applies to each element, it also applies to the whole matrix:

$$\text{plim } N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i = \text{E}(\mathbf{x}'\mathbf{x}) = \mathbf{A}.$$

(b) plim $N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' u_i$:

By the same reasoning as above,

$$\text{plim}\, N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' u_i = \text{E}(\mathbf{x}' u).$$

But according to assumption OLS.1, $\text{E}(\mathbf{x}' u) = \mathbf{0}$.

Hence,

$$\text{plim}\, N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' u_i = \mathbf{0}.$$

Taking the two steps together and using assumption OLS.2 which implies that the inverse of $\mathrm{E}(\mathbf{x}'\mathbf{x})$ exists, we obtain

$$\begin{aligned}
\operatorname{plim} \hat{\beta} &= \boldsymbol{\beta} + \left(\operatorname{plim} N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left(\operatorname{plim} N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' u_i \right) \\
&= \boldsymbol{\beta} + \left[\mathrm{E}(\mathbf{x}'\mathbf{x}) \right]^{-1} \cdot \mathbf{0} \\
&= \boldsymbol{\beta}
\end{aligned}$$

To fully appreciate the importance of this consistency result, recall that the linear projection of y on $\mathbf{x}$

$$L(y|\mathbf{x}) = L(y|1, x_2, \ldots, x_K) = \mathbf{x}\boldsymbol{\beta}$$

yields the population parameters

$$\boldsymbol{\beta} = \left[\mathrm{E}(\mathbf{x}'\mathbf{x})\right]^{-1} \mathrm{E}(\mathbf{x}'y).$$

Hence, whatever the relationship between y and $\mathbf{x}$ in the population is, OLS *always* consistently estimates the linear projection parameters.

This implies that OLS is consistent when y is discrete or even binary.

It also implies that OLS consistently estimates conditional expectations that are linear in parameters (because a conditional expectation that is linear in parameters is the same as the linear projection).

Asymptotic distribution of the OLS estimator

Key result 2: Under Assumptions OLS.1 and OLS.2 the OLS estimator is asymptotically normally distributed,

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\text{d}} \text{Normal}(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}),$$

where

$$\mathbf{A} = \text{E}(\mathbf{x}'\mathbf{x}).$$

and

$$\mathbf{B} = \text{Var}(\mathbf{x}'u) = \text{E}(u^2\mathbf{x}'\mathbf{x}).$$

Proof:

To find the limiting distribution of OLS, rearrange the OLS formula and multiply by $\sqrt{N}$:

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

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

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left(N^{-1/2} \sum_{i=1}^{N} \mathbf{x}_i' u_i \right).$$

Let us consider the asymptotic behavior of $\left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \right)^{-1}$ and $N^{-1/2} \sum_{i=1}^{N} \mathbf{x}_i' u_i$ separately.

(a) $\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}$:

We have already shown that

$$\text{plim}\, N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i = \text{E}(\mathbf{x}'\mathbf{x}) = \mathbf{A}.$$

Then, by Slutzky's theorem,

$$\text{plim}\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1} = \left(\text{plim}\, N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1} = \mathbf{A}^{-1},$$

where the invertibility is guaranteed by assumption OLS.2.

(b) $N^{-1/2} \sum_{i=1}^{N} \mathbf{x}_i' u_i$:

Note that $\mathbf{x}_i' u_i$, $i = 1, \ldots, N$ is—due to random sampling—a sequence of i.i.d. random vectors.

By assumption OLS.1, this sequence has mean zero, $\mathrm{E}(\mathbf{x}_i' u_i) = \mathbf{0}$.

We also assume that each element of $\mathbf{x}_i' u_i$ has finite variance and define the variance matrix $\mathbf{B}$,

$$\mathbf{B} = \mathrm{Var}(\mathbf{x}_i' u_i) = \mathrm{E}(u_i^2 \mathbf{x}_i' \mathbf{x}_i) = \mathrm{E}(u^2 \mathbf{x}' \mathbf{x}).$$

Then, by the central limit theorem,

$$N^{-1/2} \sum_{i=1}^{N} \mathbf{x}_i' u_i \xrightarrow{\mathrm{d}} \mathrm{Normal}(\mathbf{0}, \mathbf{B}). \tag{1}$$

We started from

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left(N^{-1/2} \sum_{i=1}^{N} \mathbf{x}_i' u_i \right)$$

and found

$$\left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \xrightarrow{\text{p}} \mathbf{A}^{-1} \qquad \text{and} \qquad N^{-1/2} \sum_{i=1}^{N} \mathbf{x}_i' u_i \xrightarrow{\text{d}} \mathbf{z},$$

where $\mathbf{z} \sim \text{Normal}(\mathbf{0}, \mathbf{B})$. Hence, by Cramer's theorem

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\text{d}} \tilde{\mathbf{z}} = \mathbf{A}^{-1} \mathbf{z}.$$

Note that

$$\text{E}(\tilde{\mathbf{z}}) = \mathbf{A}^{-1} \text{E}(\mathbf{z}) = \mathbf{0} \qquad \text{and} \qquad \text{Var}(\tilde{\mathbf{z}}) = \text{E}(\mathbf{A}^{-1} \mathbf{z}\mathbf{z}' \mathbf{A}^{-1}) = \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1}.$$

Thus

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\text{d}} \text{Normal}(\mathbf{0}, \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1}).$$

The general asymptotic variance matrix

The general asymptotic variance matrix of $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$,

$$\mathbf{V} \equiv \text{Avar}[\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1},$$

is of the *robust sandwich* form: it does not assume homoskedasticity. Hence, the asymptotic variance matrix of $\hat{\boldsymbol{\beta}}$ is (see Lecture 2)

$$\mathbf{V}_{\hat{\boldsymbol{\beta}}} \equiv \text{Avar}(\hat{\boldsymbol{\beta}}) = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}/N.$$

In practice, we act as if

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{V}_{\hat{\boldsymbol{\beta}}} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}/N,$$

which shrinks to zero at rate $1/N$, just like the variance of a sample average. Inference is thus based on the *approximate* distribution

$$\hat{\boldsymbol{\beta}} \sim \text{Normal}(\boldsymbol{\beta}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}/N).$$

The asymptotic variance matrix under homoskedasticity

Key Result 3: Under Assumptions OLS.1, OLS.2, *and* OLS.3,

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\text{d}} \text{Normal}(\mathbf{0}, \sigma^2 \mathbf{A}^{-1}).$$

Proof: If we add assumption OLS.3, then

$$\text{Var}(\mathbf{x}'u) = \mathbf{B} = \sigma^2 \, \text{E}(\mathbf{x}'\mathbf{x}) = \sigma^2 \mathbf{A}$$

and thus

$$\text{Avar}[\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1} = \sigma^2 \mathbf{A}^{-1}\mathbf{A}\mathbf{A}^{-1} = \sigma^2 \mathbf{A}^{-1}.$$

This variance form is only valid if the assumption of homoskedasticity is satisfied.

In practice, under homoskedasticity we act as if

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \mathbf{A}^{-1}/N$$

and base inference on the *approximate* distribution

$$\hat{\boldsymbol{\beta}} \sim \text{Normal}(\boldsymbol{\beta}, \sigma^2 \mathbf{A}^{-1}/N).$$

5. Estimation of the asymptotic variance matrix

The homoskedasticity-only variance matrix

We consistently estimate the variance $\text{Avar}(\hat{\boldsymbol{\beta}}) = \sigma^2 \mathbf{A}^{-1}/N$ by using the sample analogues

$$\hat{\sigma}^2 = (N-K)^{-1}\hat{\mathbf{U}}'\hat{\mathbf{U}} = (N-K)^{-1}\sum_{i=1}^{N}\hat{u}_i^2 \xrightarrow{\text{p}} \sigma^2 = \text{E}(u^2)$$

$$\hat{\mathbf{A}} = N^{-1}\mathbf{X}'\mathbf{X} = N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i \xrightarrow{\text{p}} \mathbf{A} = \text{E}(\mathbf{x}'\mathbf{x})$$

where $\hat{u}_i \equiv y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$ are the OLS residuals.

Hence, our estimate of $\text{Avar}(\hat{\boldsymbol{\beta}})$ is

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X}/N)^{-1}/N = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Comparison with the classical regression model

Remember: The classical (Gauss-Markov) assumptions imply

- $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$,
- $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, and
- $\mathrm{E}(\hat{\sigma}^2) = \sigma^2$.

Under the much weaker assumptions we have made, $\hat{\boldsymbol{\beta}}$ is not even unbiased in general. But the same formula works for estimating its asymptotic variance.

Comparison with the classical regression model

The classical regression model typically assumes normality (or conditional normality if stochastic regressors are allowed),

$$u \sim \text{Normal}(0, \sigma^2),$$

which implies $\text{E}(u) = 0$ and $\text{Var}(u) = \sigma^2$, but this is very strong.

Normality underlies exact inference: Student t distribution for t statistics and F distribution for F statistics.

But normality in practice seldom holds. Fortunately, it is *not* needed for large-sample inference:

The CLT does *not* say anything about the population distribution of u. The distribution of u in the population is fixed and has nothing to do with the size of the sample we draw. The CLT implies that standardized sample averages, such as $\sqrt{N}\bar{u} = N^{-1/2}\sum_{i=1}^{N} u_i$, have an approximate normal distribution for large N.

In typical cross-section applications, we have

- a sample size large enough ($N > 100$) to feel comfortable with the approximate normal distribution based on the CLT,
- heteroskedastic and
- non-normal disturbances.

Hence, use the CLT and make inference robust to arbitrary heteroskedasticity (drop assumption OLS.3).

The heteroscedasticity-robust covariance matrix

To estimate the asymptotic variance matrix robust to arbitrary heteroskedasticity, $\text{Avar}(\hat{\boldsymbol{\beta}}) = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}/N$, use the sample analogues of $\mathbf{A}$ and $\mathbf{B}$:

$$\hat{\mathbf{A}} = N^{-1}\mathbf{X}'\mathbf{X} = N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i \xrightarrow{\text{p}} \mathbf{A} = E(\mathbf{x}'\mathbf{x})$$

and

$$\hat{\mathbf{B}} = (N-K)^{-1}\sum_{i=1}^{N}\hat{u}_i^2\mathbf{x}_i'\mathbf{x}_i \xrightarrow{\text{p}} \mathbf{B} = E(u^2\mathbf{x}'\mathbf{x}).$$

Note: both estimators are consistent whether or not OLS.3 holds.

Now, $\text{Avar}(\hat{\boldsymbol{\beta}})$ is estimated with the *sandwich* form:

$$\begin{aligned}\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}}) &= \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}/N \\ &= \frac{N}{(N-K)}\left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\sum_{i=1}^{N}\hat{u}_i^2\mathbf{x}_i'\mathbf{x}_i\right)\left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\end{aligned}$$

The general asymptotic covariance matrix: estimation

Discussion:

- The standard errors based on the square roots of the diagonal elements of $\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}}) = \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}/N$ are often called White standard errors, Huber standard errors or Huber-White standard errors to attribute them to White (1980) and/or Huber (1967).
- The factor $N/(N-K)$ is a finite-sample correction.
- Remember, $\hat{\boldsymbol{\beta}}$ is still the OLS estimator. We are adjusting the inference for OLS.
- Note: R-squared is perfectly valid as a goodness-of-fit measure under heteroskedasticity: R^2 is a consistent estimate of $\rho^2 = 1 - \sigma_u^2/\sigma_y^2$, which is a function of the unconditional variances. Whether $Var(u|\mathbf{x})$ is constant is irrelevant for estimating ρ^2.

What about weighted least squares?

- A classical way to deal with heteroskedasticity is weighted least squares (WLS).
- WLS is feasible if we have a good estimate of $\text{Var}(y_i|\mathbf{x}_i)$.
- WLS is efficient if the model used to estimate $\text{Var}(y_i|\mathbf{x}_i)$ is correct and $\text{E}(u|\mathbf{x}) = 0$.
- However, it is inconsistent if $\text{E}(u|\mathbf{x}) \neq 0$, which is the case when we estimate linear projections of nonlinear population regression functions.
- Therefore, with large sample sizes, using OLS and heteroskedasticity-robust standard errors is generally preferred.

Example: the effect of education on wage

Blackburn and Neumark (1992, QJE)

- Blackburn and Neumark (1992, QJE) study the determinants of wages in the US. Let us concentrate of the effect of education.
- The data are taken from the Young Men's Cohort of the National Longitudinal Survey (NLS). (So what is a sensible population here?)
- Only respondents that have taken an IQ test in 1968 are included in the sample. (Can this lead to a sample selection problem?)
- The dependent variable is the log wage in 1980 (*lwage*).
- Let us start with a single explanatory variable, years of eduction (*educ*), i.e., the model

$$\mathrm{E}[\log(\mathit{wage})|\mathit{educ}] = \beta_0 + \beta_1 \mathit{educ}.$$

The Stata script to run the regression is very simple:

```
*** load data (substitute the appropriate path)
use "c:\path\nls80.dta", clear

*** run OLS using robust s.e.'s
regress lwage educ, robust
```

```
Linear regression                               Number of obs =     935
                                                F(  1,   933) =   96.89
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0974
                                                Root MSE      =  .40032
```

lwage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0598392	.0060791	9.84	0.000	.047909	.0717694
_cons	5.973063	.0822718	72.60	0.000	5.811603	6.134522

A few questions you should be able to answer:

- How many observations did Stata use?
- What is the regression R-squared?
- What is the estimated slope parameter of *educ*? Is it quantitatively relevant? Statistically significant?
- What is the 95 percent interval estimate of the slope parameter?

Let us briefly spend a few thoughts on the slope estimate of $\hat{\beta}_1 = 0.06$.

To interpret it, recall that on average (we leave out the expectation symbol for clarity of presentation)

$$\beta_1 = \frac{\partial \log(wage)}{\partial educ} = \frac{\partial wage/wage}{\partial educ}$$

and thus

$$\frac{\Delta wage}{wage} \approx \beta_1 \Delta educ.$$

Hence, another year of education ($\Delta educ = 1$) leads on average to a relative wage change of 0.06 or, in other words, of 6 percent. This is quantitatively important as, for example, can be seen by asking what is the effect of adding a four-year college degree to a highschool degree. The wage differential is predicted to be $4 \times 6 = 24$ percent which is substantial.

Do you think the result is valid?

The simple regression reported above almost certainly suffers from omitted variable bias. While we will discuss the econometric details below, the underlying concept is easy to understand.

Recall that we use observational data which do not satisfy the requirements of an experiment. If they would, people would have been assigned randomly to a specific value of *educ* and all other characteristics of the workers would differ only by random. In contrast, in the data we have, years of education is a choice variable, and high-wage workers may differ systematically from low-wage workers in many dimensions (just think of ability, socioeconomic background, and job experience).

Therefore, to compare apples and apples (and not apples and oranges), we need to control for all important dimensions in which there might be a systematic difference between high-wage and low-wage workers. The interpretation of the slope parameters is then conditional on holding these dimensions constant.

As a consequence of the preceding discussion, let us expand the structural model:

$$\log(wage) = \beta_0 + \beta_1 exper + \beta_2 tenure + \beta_3 married + \beta_4 south + \beta_5 urban + \beta_6 black + \beta_7 educ + u$$

where *educ* (years of schooling) is the regressor of interest and the additional regressors act as control variables:

- *exper* is labor market experience,
- *tenure* is number of years at work,
- *married* is a dummy variable equal to unity if married,
- *south* is a dummy variable for the southern region,
- *urban* is a dummy variable for living in a city (Standard Metropolitan Statistical Area, SMSA), and
- *black* is a race indicator.

Based on

$$\log(wage) = \beta_0 + \beta_1 exper + \beta_2 tenure + \beta_3 married + \beta_4 south + \beta_5 urban + \beta_6 black + \beta_7 educ + u$$

the regression coefficient β_7 measures the average effect of another year of schooling **holding the controls fixed**.

Why is that important?

For example, there might be structural differences in the USA, that affect both education and wages, between people living in the south and people living in other parts of the USA.

By holding *south* fixed, we only compare individuals who live in the same region and do not mix the education effect with the regional effect on wages.

Linear regression

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

lwage	Coef.	Robust 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	.121517	.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	5.173482	5.617513

Even though we added lots of controls, the point estimate did not change much (from 0.6 to 0.65).

So everything fine? Unfortunately not: we should certainly control for ability. And there might be additional threats to model validity as discussed in the next section.

6. Omitted variables

Endogeneity: why OLS may be inconsistent

Central assumption for consistency: $\mathrm{E}(\mathbf{x}'u) = 0$ (OLS.1)

If this is violated for regressor j, i.e., $\mathrm{E}(x_j u) \neq 0$, this regressor is called *endogenous*.

Potential reasons for endogeneity:

- Omitted variables $\longrightarrow$ this lecture
- Measurement error $\longrightarrow$ this lecture
- Simultaneity $\longrightarrow$ Lecture 6

Omitted variables: example

Consider a structural wage equation of the kind estimated above with unobserved ability:

$$\log(wage) = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 educ + \gamma abil + v$$

where v has the structural error property $\mathrm{E}(v|exper, educ, abil) = 0$.

Question: can we consistently estimate the structural parameters $\beta_1, \beta_2, \beta_3$ even though we do not observe ability? What do we estimate when we run a regression of

$$\log(wage) \quad \text{on} \quad 1, exper, exper^2, edu$$

leaving out *abil*?

Omitted variables: setup

Think of a structural population model with an omitted factor q,

$$\mathrm{E}(y|x_1, x_2, \ldots, x_K, q) = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K + \gamma q.$$

We are interested in the β_j, which are the partial effects of the observed explanatory variables holding the other explanatory variables constant, *including* the unobservable q.

In error form, the model is

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K + \gamma q + v$$

with the structural error v that is uncorrelated with $x_1, \ldots, x_K, q$ because

$$\mathrm{E}(v|x_1, x_2, \ldots, x_K, q) = 0.$$

Why should we not just put q into the regression disturbance and estimate

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K + u$$

where $u = \gamma q + v$?

Omitted variables: OLS is inconsistent

Problem: When we estimate

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K + u, \qquad u = \gamma q + v,$$

there might be correlation between the regressors and the error term.

From

$$\mathrm{E}[x_j u] = \mathrm{E}[x_j(\gamma q + v)] = \gamma\,\mathrm{E}[x_j q] + \mathrm{E}[x_j v] = \gamma\,\mathrm{E}[x_j q]$$

we find that $\mathrm{E}[x_j u] = 0$, $j = 1, \ldots, K$, holds only if

- either q has no partial effect on y, i.e., $\gamma = 0$,
- or q is uncorrelated with all x_j, $j = 1, \ldots, K$.

Consequence: in general assumption OLS.1 is violated—OLS is inconsistent!

Omitted variable bias

But: what do we estimate when q is omitted?

Use the linear projection of q onto the observable explanatory variables,

$$q = \delta_0 + \delta_1 x_1 + \ldots + \delta_K x_K + r$$

where, by definition of a linear projection, $\mathrm{E}(r) = 0$, $\mathrm{Cov}(x_j, r) = 0$, $j = 1, 2, \ldots, K$.

Plug it into the structural equation

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K + \gamma q + v$$

and rearrange to

$$y = (\beta_0 + \gamma\delta_0) + (\beta_1 + \gamma\delta_1) x_1 + \cdots + (\beta_K + \gamma\delta_K) x_K + v + \gamma r$$

Now, the error $v + \gamma r$ has zero mean and is uncorrelated with each regressor.

Omitted variable bias

Result: a regression of y on $1, x_1, \ldots, x_K$ estimates the population model

$$y = (\beta_0 + \gamma\delta_0) + (\beta_1 + \gamma\delta_1)\, x_1 + \cdots + (\beta_K + \gamma\delta_K)\, x_K + v + \gamma r.$$

Since it satisfies assumptions OLS.1 and OLS.2, OLS consistently estimates the coefficients of this model,

$$\text{plim}\, \hat{\beta}_j = \beta_j + \gamma\delta_j.$$

How large is the asymptotic omitted variable bias $\gamma\delta_j$?

In general, this is difficult to say because it depends on γ and the parameters of the linear projection. Remember how they are constructed:

$$\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K)' = [\text{Var}(\mathbf{x})]^{-1}\, \text{Cov}(\mathbf{x}', q).$$

Special case: q is only related to a constant and one regressor x_j:

$$q = \delta_0 + \delta_j x_j + r, \qquad \delta_i = 0, i \neq j.$$

Then: $\delta_j = [\text{Var}(x_j)]^{-1} \text{Cov}(x_j, q)$.

Hence: $\text{plim}\, \hat{\beta}_j = \beta_j + \gamma \,\text{Cov}(x_j, q)/\,\text{Var}(x_j)$.

From this expression, we can at least infer the sign of the asymptotic bias. For example, if $\gamma > 0$ and $\text{Cov}(x_j, q) > 0$, then

$$\text{plim}\, \hat{\beta}_j = \beta_j + \text{positive number} > \beta_j.$$

Omitted variables: example

Consider again the structural wage equation with unobserved ability,

$$\log(wage) = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 educ + \gamma abil + v.$$

Let us assume that ability is uncorrelated with *exper* and $exper^2$ once *educ* has been partialed out, i.e., the linear projection of *abil* on $1, exper, exper^2, educ$ is

$$abil = \delta_0 + \delta_3 educ + r,$$

where r is uncorrelated with *exper* and $exper^2$.

Then there is no asymptotic bias in $\hat{\beta}_1$ and $\hat{\beta}_2$ but

$$\text{plim}\, \hat{\beta}_3 = \beta_3 + \gamma\delta_3.$$

We would expect $\gamma > 0$ (ability has a positive direct effect on wages) and $\delta_3 > 0$ (education and ability are positively correlated), hence

$$\text{plim}\, \hat{\beta}_3 > \beta_3.$$

Policy implications?

Proxy variables

A solution to the omitted variable bias is to use a proxy variable.

Example: use IQ test score as a proxy for ability.

To be helpful, a proxy variable z has to satisfy two conditions:

(1) it must be redundant (ignorable) in the structural equation and

(2) it must ensure that the correlation between the omitted variable q and the regressors x_j is zero once z has been partialed out.

Condition 1: redundancy

Redundancy means that z is not part of the structural model if the x_j's and q are controlled for:

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

For a linear relationship this implies

$$\mathrm{E}(y|\mathbf{x}, q, z) = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K + \gamma q.$$

Example: to explain wages, we need to know ability (and the other regressors). As long as we control for it, there is no effect of the IQ test score.

Proxy variables: eliminating the correlation between q and x_j

Condition 2: eliminating the correlation between q and x_j

Mathematically, this condition requires that a linear projection of q on the x_j and z satisfies

$$\mathrm{L}\left(q|1, x_1, \ldots, x_K, z\right) = \mathrm{L}\left(q|1, z\right).$$

In error form this means that the result of this projection is

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

where, by definition of a projection, $\mathrm{E}(r) = 0$ and $\mathrm{Cov}(z, r) = 0$ **and**

$$\mathrm{Cov}\left(x_j, r\right) = 0, \qquad j = 1, 2, \ldots, K.$$

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

Obviously, to qualify z as a reasonable proxy for q, $\theta_1 \neq 0$.

Proxy variables: consistency of OLS

If the proxy variable conditions are satisfied, we can use

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

to replace q in the structural equation

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K + \gamma q + v$$

which yields

$$y = (\beta_0 + \gamma\theta_0) + \beta_1 x_1 + \cdots + \beta_K x_K + \gamma\theta_1 z + (\gamma r + v)\,.$$

The composite error term $u \equiv \gamma r + v$ is by assumption uncorrelated with all x_j's.

Redundancy of z in the structural equation means that z is uncorrelated with v and, by definition, z is uncorrelated with r.

Result: The OLS regression of y on $1, x_1, \ldots, x_K, z$ produces consistent estimators of $\beta_0 + \gamma\theta_0, \beta_1, \ldots, \beta_K, \gamma\theta_1$.

Proxy variables: imperfect and bad proxies

- Imperfect proxy: including z *reduces* (but does not eliminate) the correlation of the regressors x_j with the error term $\longrightarrow$ may still be better to include z than to exclude it because the asymptotic bias is reduced.

- Bad proxy: z is only weakly correlated or even uncorrelated with q $\longrightarrow$ do not include z as it may even *increase* the asymptotic bias.

Example: Using IQ as a proxy for ability

Blackburn and Neumark (1992, QJE) continued

The structural model certainly includes ability:

$$\log(wage) = \beta_0 + \beta_1 exper + \beta_2 tenure + \beta_3 married + \beta_4 south + \beta_5 urban + \beta_6 black + \beta_7 educ + \gamma abil + v$$

Proxy variable for unobserved ability: IQ test score

Assumption: the linear projection of $abil$ is

$$abil = \theta_0 + \theta_1 IQ + r$$

with r being uncorrelated with IQ and all other regressors.

Prediction: OLS neglecting *abil* will lead to an overestimation of the effect of education (years of schooling) on the wage. Why?

Because *abil* should positively influence the wage ($\gamma > 0$), and *abil* and *educ* should be positively correlated $\delta_7 > 0$. Hence the asymptotic OLS bias for $\hat{\beta}_7$ is positive.

Let us compare the regression results:

OLS result neglecting *abil* (s.e. in brackets below the estimates):

$$\widehat{\log(wage)} = \underset{(0.11)}{5.40} + \underset{(.003)}{.014}\, exper + \underset{(.002)}{.012}\, tenure + \underset{(.039)}{.199}\, married - \underset{(.026)}{.091}\, south + \underset{(.027)}{.184}\, urban - \underset{(.038)}{.188}\, black + \underset{(.006)}{.065}\, educ$$

OLS result using *IQ* as a proxy for *abil*:

$$\widehat{\log(wage)} = \underset{(0.13)}{5.18} + \underset{(.003)}{.014}\, exper + \underset{(.002)}{.011}\, tenure + \underset{(.039)}{.200}\, married - \underset{(.026)}{.080}\, south + \underset{(.027)}{.182}\, urban - \underset{(.039)}{.143}\, black + \underset{(.007)}{.054}\, educ + \underset{(.0010)}{.0036}\, IQ$$

First estimate: one more year of schooling increases the wage by 6.5 percent.

Second estimate: one more year of schooling increases the wage by 5.4 percent.

7. Measurement error

Measurement error in the dependent variable

Let y^* be a variable we would like to explain.

But we observe only y, which is an imperfect measure of y^*. (For example, in the wage equation discussed previously, wage was measured by a telephone survey and respondents may have guessed their wages.)

Does it make a difference for OLS whether we use y^* or y?

To analyze the consequences of such measurement error, consider the structural regression model

$$y^* = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + v$$

which satisfies assumptions OLS.1 and OLS.2.

Now define the measurement error

$$e_0 = y - y^* \qquad \Longrightarrow \qquad y^* = y - e_0.$$

Hence, the regression model in terms of y is

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + v + e_0.$$

Does OLS using data generated by

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + v + e_0$$

produce consistent estimates for the β_j's?

- Consistency holds when assumption OLS.1 and OLS.2 are satisfied.
- OLS.2 is satisfied if it holds for the original structural equation.
- OLS.1 is satisfied if the regressors $x_1, \ldots, x_k$ are uncorrelated with the disturbance $u = v + e_0$.
- Since $x_1, \ldots, x_k$ and v are uncorrelated by construction, OLS.1 is satisfied if $x_1, \ldots, x_k$ and e_0 are uncorrelated.
- Typical assumption: measurement error independent of the x_j's. This is justified when the process of (mis)measurement is unrelated with the regressors.
- Then OLS is consistent.

Is it a good idea to assume that the measurement error is independent of the regressors?

It depends. Think of the previous example (slightly simplified)

$$\log(\textit{wage}) = \beta_0 + \beta_1 \textit{exper} + \beta_2 \textit{tenure} + \beta_3 \textit{married} + \beta_4 \textit{educ} + v$$

but assume the hourly wage is constructed as follows

$$\text{wage per hour} = \frac{\text{wage per month}}{\text{hours per month}}$$

where wage per month and hours per month are from a telephone survey. In an interview situation, respondents might not be fully sure about the number of hours worked last month (rounding, overtime, etc).

We may sensibly assume that this reporting error is typically unrelated to regressors such as *exper*, *tenure*, *married*, and *educ* so that OLS is consistent.

(However, there is the possibility that response accuracy is related to *educ* which would render OLS inconsistent.)

Summary:

- As long as the measurement error is independent of the regressors, OLS is consistent.
- Standard errors, t tests, F test etc. can be used as before.
- Of course, the measurement error increases estimation uncertainty compared to the case where we know y^*.

Measurement error in an explanatory variable: setup

Consider the structural regression model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K^* + v$$

which satisfies assumptions OLS.1 and OLS.2. Hence, if we observe x_K^*, OLS is consistent.

However, x_K^* is measured with error as x_K, where we assume redundancy as in the proxy variable model:

$$\mathrm{E}(y|x_1, \ldots, x_{K-1}, x_K^*, x_K) = \mathrm{E}(y|x_1, \ldots, x_{K-1}, x_K^*).$$

The measurement error in the population is

$$e_K = x_K - x_K^*.$$

We assume that the average measurement error is zero: $\mathrm{E}(e_K) = 0$, which has no practical consequences because we include an intercept in the regression.

Let us assume that the measurement error is unrelated to the other explanatory variables:

$$\mathrm{E}\left(x_j, e_K\right) = 0, \quad j = 1, \ldots, K-1.$$

When we run a regression of y on $1, x_1, \ldots, x_K$, can we consistently estimate the β_j's?

Let us consider two polar cases:

(1) e_K unrelated with observed measure: $\mathrm{Cov}\left(x_K, e_K\right) = 0$.

(2) e_K unrelated with unobserved true regressor: $\mathrm{Cov}\left(x_K^*, e_K\right) = 0$.

Measurement error in an explanatory variable: case 1

Case 1: $\text{Cov}(x_K, e_K) = 0$

Write $x_K^* = x_K - e_K$. Since e_K is uncorrelated with x_K, it must be correlated with x_K^*. Sensible?

Now plug $x_K^* = x_K - e_K$ into the structural equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + (v - \beta_K e_K).$$

We have assumed that v and e_K both have zero mean and are uncorrelated with each x_j, including x_K.

Therefore, $u = v - \beta_K e_K$ has zero mean and is uncorrelated with the x_j.

It follows that OLS estimation with x_K in place of x_K^* produces consistent estimators of all of the β_j (assuming the standard rank condition Assumption OLS.2).

Compared to a world where x_K^* was known, the error variance and thus estimation imprecision has increased.

Measurement error in an explanatory variable: case 2

Case 2: $\text{Cov}(x_K^*, e_K) = 0$

This is the **classical error-in-variables (CEV)** assumption.

Write $x_K = x_K^* + e_K$. Since e_K is uncorrelated with x_K^*, it must be correlated with x_K.

Again plug $x_K^* = x_K - e_K$ into the structural equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + (v - \beta_K e_K).$$

Now there is correlation between x_K and the error term (because it contains e_K).

This renders OLS inconsistent. In general, *all* the β_j's have an asymptotic bias.

The sign and size of this bias is difficult to characterize in general. Obviously, it depends on the strength of the correlation between e_K and x_K.

To find a simple expression, consider the single-regressor model ($K = 1$), where the only regressor is measured with error.

This yields the estimable model

$$y = \beta_0 + \beta_1 x_1 + (v - \beta_1 e_1).$$

Due to the measurement error, $x_1 = x_1^* + e_1$ and there is correlation between x_1 and $u = v - \beta_1 e_1$:

$$\text{Cov}(x_1, e_1) = \text{E}(x_1 e_1) = \text{E}(x_1^* e_K) + \text{E}\left(e_1^2\right) = \sigma_{e_1}^2.$$

The OLS estimator of β_1 converges in probability as follows:

$$\text{plim}\,\hat{\beta}_1 = \text{plim}\,\frac{s_{x_1,y}}{s_{x_1}^2} = \beta_1 + \frac{\text{plim}\,s_{x_1,u}}{\text{plim}\,s_{x_1}^2} = \beta_1 + \frac{\text{Cov}(x_1, u)}{\text{Var}(x_1)}$$

Now,

$$\text{Var}(x_1) = \text{Var}(x_1^* + e_1) = \text{Var}(x_1^*) + \text{Var}(e_1) = \sigma^2_{x_1^*} + \sigma^2_{e_1}$$

and

$$\text{Cov}(x_1, u) = \text{Cov}(x_1^* + e_1, v - \beta_1 e_1) = -\beta_1 \,\text{Cov}(e_1, e_1) = -\beta_1 \,\text{Var}(e_1) = -\beta_1 \sigma^2_{e_1}.$$

Thus,

$$\text{plim}\,\hat{\beta}_1 = \beta_1 + \frac{\text{Cov}(x_1, u)}{\text{Var}(x_1)} = \beta_1 - \beta_1 \frac{\sigma^2_{e_1}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}} = \beta_1 \frac{\sigma^2_{x_1^*}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}}$$

The term multiplying β_1 is always less than unity under the CEV assumption.

This is the **attenuation bias** in OLS due to CEVs: the OLS estimator is shrinked towards zero; the partial effect of interest is underestimated (in absolute terms).

8. Practical Regression Hints

Use theory and common sense to specify the regression model

- Use logs if you think the partial effects of *relative changes* should be constant.
- Add a square to a linear regressor if you suspect diminishing marginal effects.
- Add dummies if you think certain subpopulations have different means.
- Interact dummies with regressors if you think certain subpopulations have different slopes.
- Always think hard about regressor endogeneity!
  - Is there an omitted variable that leads to omitted variable bias?
  - Is there measurement error that leads to attenuation bias?
  - Is there simultaneity/reverse causality?

Do not mechanically maximize the R-squared

Do not always attempt to maximize R-squared, adjusted R-squared, or some other goodness-of-fit measure.

- You might include regressors that are highly collinear and essentially say the same.
- Example: different measures of the same thing (check their correlation first).
- You might include regressors that should not be held fixed.
- Example: y is family demand for a product, $\mathbf{x}$ includes various product prices, income, and demographics. But we should not include the demand for a competing product because usually it does not make sense to hold a quantity demanded fixed and change the price of any good.

Always remember: It is possible to obtain a convincing estimate of a causal effect with a low R-squared. For example, under random assignment, a simple regression estimate consistently estimates the causal effect, but the treatment may not explain much of the variation in y.

But increasing the R-squared may help

Include covariates that help predict the outcome if they are uncorrelated (in the population) with the covariate(s) of interest.

So, if w is the explanatory variable of interest, and it has been randomized with respect to the response and controls, say $\mathbf{z}$, then estimate

$$y = \alpha + \beta w + \mathbf{z}\gamma + u.$$

Because $Cov(\mathbf{z}, w) = 0$, adding $\mathbf{z}$ will not cause collinearity (except slightly in any sample), but it will generally reduce the error variance and thus estimation precision. In large samples,

$$Var(\hat{\beta}) \approx \frac{\sigma_u^2}{n\sigma_w^2}.$$

As more (relevant) covariates are added to $\mathbf{z}$, σ_u^2 gets smaller. (And, of course, maximizing the variance of w in a designed experiment helps, too.)

A control that can substantially reduce the error variance is a lagged value of y.