Econometrics I

Lecture 3: Linear Regression

Paul T. Scott
NYU Stern

Fall 2024

Linear Regression: Introduction I

- A typical example of a linear regression model:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_K x_{K,i} + \varepsilon_i,$$

where

- $i = 1, \ldots, n$ indexes observations

Linear Regression: Introduction I

- A typical example of a linear regression model:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_K x_{K,i} + \varepsilon_i,$$

where

  - $i = 1, \ldots, n$ indexes observations
  - y_i, a scalar, is often referred to as the **dependent variable**

Linear Regression: Introduction I

- A typical example of a linear regression model:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_K x_{K,i} + \varepsilon_i,$$

where

- $i = 1, \ldots, n$ indexes observations
- y_i, a scalar, is often referred to as the **dependent variable**
- $x_{k,i}$, is the ith observation of the kth **explanatory variable**, or **independent variable** (independent of what?), or **regressor**

Linear Regression: Introduction I

- A typical example of a linear regression model:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_K x_{K,i} + \varepsilon_i,$$

where

- $i = 1, \ldots, n$ indexes observations
- y_i, a scalar, is often referred to as the **dependent variable**
- $x_{k,i}$, is the ith observation of the kth **explanatory variable**, or **independent variable** (independent of what?), or **regressor**
- The β_k terms represent the **parameters**.
  There are K parameters, one for each regressor.

Linear Regression: Introduction I

- A typical example of a linear regression model:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_K x_{K,i} + \varepsilon_i,$$

where

- $i = 1, \ldots, n$ indexes observations
- y_i, a scalar, is often referred to as the **dependent variable**
- $x_{k,i}$, is the ith observation of the kth **explanatory variable**, or **independent variable** (independent of what?), or **regressor**
- The β_k terms represent the **parameters**.
  There are K parameters, one for each regressor.
- ε_i is the **disturbance** or **error term**

Linear Regression: Introduction I

- A typical example of a linear regression model:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_K x_{K,i} + \varepsilon_i,$$

where

- $i = 1, \ldots, n$ indexes observations
- y_i, a scalar, is often referred to as the **dependent variable**
- $x_{k,i}$, is the ith observation of the kth **explanatory variable**, or **independent variable** (independent of what?), or **regressor**
- The β_k terms represent the **parameters**.
  There are K parameters, one for each regressor.
- ε_i is the **disturbance** or **error term**
- Only the y_i and $x_{k,i}$ terms are observed by the econometrician.

Linear Regression: Introduction II

- A typical example of a linear regression model:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_K x_{K,i} + \varepsilon_i,$$

- Today's questions:
  - Where does it come from?
  - What assumptions do we need to estimate β?
  - How do we estimate β?
  - How to interpret estimates?
  - What are the estimator's finite sample and asymptotic properties?

Linear Regression: Matrix Notation

- We can express the linear regression model in vector notation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

  where
  - $\mathbf{y}$ is a $n \times 1$ vector of observations of the dependent variable
  - $\mathbf{X}$ is a $n \times K$ vector of observations of the dependent variables
  - $\boldsymbol{\beta}$ is a $K \times 1$ vector of parameters
  - $\boldsymbol{\varepsilon}$ is a $n \times 1$ vector of error terms
- Note that each row of this equation corresponds to the previous equation for a single observation i:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i,$$

- Conventions: Roman symbols are observed, Greek are not, bold means vector notation, bold capitals means matrices.

Example: Mincerian Regression I

- Let's suppose we are interested in how worker's wages depend on education and experience (known as Mincerian regression or Mincer earnings function for Jacob Mincer).
- y_i could be worker i's wages, and $\mathbf{x}_i' = (1, edu_i, exp_i)$, where
  - edu_i is worker i's education (in years)
  - exp_i is worker i's work experience (in years)
  - Note that the regressors include a constant

Example: Mincerian Regression I

- y_i is i's wages, and $\mathbf{x}_i' = (1, edu_i, exp_i)$
- The data matrices for the Mincerian regression might look like this:

$$\mathbf{y} = \begin{pmatrix} 12 \\ 35 \\ 20 \\ \vdots \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & 12 & 2 \\ 1 & 16 & 5 \\ 1 & 12 & 21 \\ \vdots & \vdots & \vdots \end{pmatrix}$$

Strict Exogeneity

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

- This equation doesn't mean much until we say something about the error term ε.
- The first (and strongest) assumption on error terms we will consider is **strict exogeneity**:

$$E[\varepsilon|\mathbf{X}] = \mathbf{0}$$

- Note that the law of iterated expectations implies

$$E[\varepsilon_i] = E_{\mathbf{X}}[E[\varepsilon_i|\mathbf{X}]] = 0.$$

It also implies that the error terms are uncorrelated with the regressors: $Cov[\varepsilon_i, \mathbf{X}] = 0$ and $Cov[\varepsilon_i, \mathbf{x}_i] = 0$.

Strict Exogeneity and Conditional Means

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$
$$E[\boldsymbol{\varepsilon}|\mathbf{X}] = 0 \tag{2}$$

- Now, we have a meaningful model.
- Note that

$$\begin{aligned} E[\mathbf{y}|\mathbf{X}] &= E[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}|\mathbf{X}] \\ &= \mathbf{X}\boldsymbol{\beta} + E[\boldsymbol{\varepsilon}|\mathbf{X}] \\ &= \mathbf{X}\boldsymbol{\beta}, \end{aligned}$$

so equations (1) and (2) already imply that the conditional mean is a linear function of $\mathbf{X}$.

Strict Exogeneity Interpretation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

$$E[\boldsymbol{\varepsilon}|\mathbf{X}] = 0 \tag{2}$$

- Strict exogeneity captures the idea that x is varied in the data without changing the mean of the unobservable factors affecting y.
- Strict exogeneity is very plausible in the context of experimental variation (especially in double blind studies)
- Unfortunately, social scientists are often unable to rely on experimental variation, and strict exogeneity is rarely plausible in the context of naturally occurring data. For this reason, in future lectures we will consider different identification assumptions, but strit exogeneit is a starting point.

Omitted Variables and Endogeneity

- Suppose we estimate the following Mincerian regression:

$$\ln(wage_i) = \beta_0 + \beta_1 edu_i + \beta_2 exp_i + \beta_3 exp_i^2 + \varepsilon_i.$$

- Furthermore suppose the true model is:

$$\ln(wage_i) = \beta_0 + \beta_1 edu_i + \beta_2 exp_i + \beta_3 exp_i^2 + \beta_4 ability_i + \varepsilon_i$$

- Then, when estimating the first equation (because ability is unobserved), the error term is effectively:

$$\widetilde{\varepsilon}_i = \beta_4 ability_i + \varepsilon_i$$

- Note that even if ε_i satisfies strict exogeneity, $\widetilde{\varepsilon}_i$ will not if, for instance, ability is correlated with education. We say x is **endogenous** if $E[\varepsilon|x] \neq 0$.

Wherefore Linearity?

- When would $E[y_i|\mathbf{x_i}] = \mathbf{x}_i'\boldsymbol{\beta}$ be true?
- We might start with the conditional mean as a general function of x

$$y_i = f(\mathbf{x}_i) + \varepsilon_i,$$

  and then what we've done is impose that f is a linear function.
- We might also think of the model as a linear approximation (using Taylor's theorem).
  - Actually, it can be a polynomial approximation, not just a linear approximation ...

Linear-in-Parameters I

- The linear regression framework does not impose that y is a linear function of any particular variable x
- The regressors in $\mathbf{x}_i$ can include squared terms, higher powers, and other functions of a variable x
- For example, $\mathbf{x}_i' = (1, edu_i, exp_i, exp_i^2)$ in the Mincerian regression would allow declining (or increasing) returns to work experience.
- The "Linear" part of "Linear Regression" really means linear-in-parameters, which is much less restrictive than being linear with respect to a particular variable.

Linear-in-Parameters II

- The dependent variable can also involve nonlinear transformations.
- For example, y_i in the Mincerian regression is typically the natural log of worker i's wage:

$$\ln(wage_i) = \beta_0 + \beta_1 edu_i + \beta_2 exp_i + \beta_3 exp_i^2 + \varepsilon_i.$$

- Models of demand (or supply) sometimes have the form

$$\ln(q_i) = \beta_0 + \beta_1 \ln(p_i) + \varepsilon_i,$$

  in which case β_1 represents the price elasticity of demand (supply) and does not depend on the units that prices and quantities are measured in (Check this). Economists love logs!
- What would it take for a model to not be linear in parameters?

Derivation from Multivariate Normal

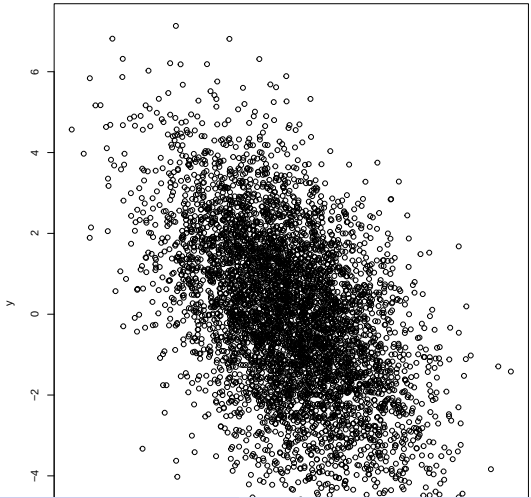
- Another way to motivate the linear regression function is from the multivariate normal distribution.
- Suppose $(y, x)' \sim \mathcal{N}(\mu, \Sigma)$, where

$$\mu = \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_y^2 & \rho\sigma_y\sigma_x \\ \rho\sigma_y\sigma_x & \sigma_x^2 \end{pmatrix}$$

Multivariate Normal Data

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 4 & -1 \\ -1 & 1 \end{pmatrix}$$



R code

```
# install.packages('MASS')
# install.packages('ggplot2')
# install.packages('reshape')
library(MASS)
library(ggplot2)
library(reshape)

mu <- c(0,0)
Sigma <- matrix(c(1,-1,-1,4),2,2)
xy <- mvrnorm(n = 5000, mu, Sigma)
plot(xy, xlab="x",ylab="y")

xy <- mvrnorm(n = 100000, mu, Sigma)
xy.df <- as.data.frame(xy)
names(xy.df) <- c("x", "y")
xy.stacked <- melt(xy.df)
ggplot(xy.stacked, aes(value, fill = variable)) + geom_density(alpha = 0.2)

xy.selected <- xy.df[xy.df$x>-.04  & xy.df$x<.04, ]
ggplot(xy.selected, aes(y)) + geom_density(alpha = 0.2)
```

Multivariate Normal and Regression Equation

- Suppose $(y, x)' \sim \mathcal{N}(\mu, \Sigma)$, where

$$\mu = \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_y^2 & \rho\sigma_y\sigma_x \\ \rho\sigma_y\sigma_x & \sigma_x^2 \end{pmatrix}$$

- It's also the case that

$$E(y|x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x)$$

and

$$y = E(y|x) + \varepsilon$$

with ε normally distributed.

- Normally distributed error terms: the first case we will consider (and easiest to analyze)

Full Rank I

- Our next assumption is that the matrix of regressors has **full rank**
  - $\mathbf{X}$ is a $n \times K$ matrix with rank K.

Full Rank I

- Our next assumption is that the matrix of regressors has **full rank**
  - $\mathbf{X}$ is a $n \times K$ matrix with rank K.
- A model of consumption that violates full rank:

$$C = \beta_1 + \beta_2 \text{Salary} + \beta_3 \text{Nonsalary income} + \beta_4 \text{Total income} + \varepsilon$$

Full Rank I

- Our next assumption is that the matrix of regressors has **full rank**
  - $\mathbf{X}$ is a $n \times K$ matrix with rank K.
- A model of consumption that violates full rank:

$$C = \beta_1 + \beta_2 \text{Salary} + \beta_3 \text{Nonsalary income} + \beta_4 \text{Total income} + \varepsilon$$

- Conditional on total income, any increase in salary must be met by a proportional decrease in nonsalary income

Full Rank II

- Consider the following model of consumption:

$$C = \beta_1 + \beta_2 \text{Salary} + \beta_3 \text{Nonsalary income} + \beta_4 \text{Total income} + \varepsilon$$

$$\text{Total income} = \text{Salary} + \text{Nonsalary income},$$

so

$$\begin{aligned} C &= \beta_1 + (\beta_2 + \beta_4)\,\text{Salary} + (\beta_3 + \beta_4)\,\text{Nonsalary Income} + \varepsilon \\ C &= \beta_1 + \widetilde{\beta}_2 \text{Salary} + \widetilde{\beta}_3 \text{Nonsalary Income} + 0 \cdot \text{Total Income} + \varepsilon \end{aligned}$$

- Assuming $\beta_4 \neq 0$ in the original equation, we have constructed an empirically equivalent equation with different parameters. That is, for this model, we have different values for β that are **observationally equivalent**.

Full Rank III

- Exercise: suppose I wanted to build a model of the approval ratings of major party nominees for US president.
- I want to include the following regressors:
  - Years holding elected office
  - Age
  - Gender
  - Indicator variable for being married to a former president
- What's the problem, assuming I have data through 2020? What about in 2025? Compare to the previous case with salaries – any difference?

Linear Independence

- Another term for the full rank assumption ($Rank\,(\mathbf{X}) = K$) is **linear independence**. **Multicollinearity** refers to a lack of linear independence. When a model has multicollinearity, we say it is **not identified**.
- Note that this is distinct from statistical independence.
- Linear independence means that one variable cannot be algebraically predicted by others.
- Statistical independence means that the variation in two random variables is unrelated. Linear independence does not imply statistical independence. Why not? Does statistical independence imply linear independence?

Assumptions so far

- The assumptions we have introduced so far:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{1}$$
$$E[\varepsilon|\mathbf{X}] = 0 \tag{2}$$
$$Rank(\mathbf{X}) = K \tag{3}$$

- Note that we have not made any assumptions on the statistical properties of $\mathbf{X}$. There is no need to do so; $\mathbf{X}$ can be fixed or random *for now*. What matters is the distribution of the error terms conditional on $\mathbf{X}$.
- Not all assumptions will be maintained throughout the course (or even this lecture). When we consider formal results, I will be explicit about which assumptions are needed.

Homoscedasticity

- Another important assumption is **homoscedasticity**:

$$E\left[\varepsilon\varepsilon'|\mathbf{X}\right] = \sigma^2\mathbf{I}$$

  where $\mathbf{I}$ is a $n \times n$ identity matrix.

- Given that $E\left[\varepsilon|\mathbf{X}\right] = 0$, this means that

$$Var\left[\varepsilon|\mathbf{X}\right] = \sigma^2\mathbf{I}$$

- In words, homoscedasticity says that each error term has the same variance; i.e., the variance of ε_i is not related to $\mathbf{x}_i$. **Heteroscedastcity** is the alternative. Can you think of some cases of heteroscedasticity?
- The assumption also rules out correlation between the error terms for different observations.

Normal Error Terms

- A convenient assumption is that error terms are **normally distributed**:

$$\varepsilon|\mathbf{X} \sim \mathcal{N}\left(0, \sigma^2\mathbf{I}\right)$$

- Where are we going?
  - This assumption will make it easy to derive normally distributed estimators (just as normality made the LLN and CLT easy to derive)
  - Ultimately, the central limit theorem can be used to derive **asymptotic normality** of estimators (that is, estimators that will be normally distributed as the sample gets large), so in practice it's rarely necessary to assume normal error terms.

Residuals

- We used $\boldsymbol{\beta}$ to denote the true vector of parameters; let's use $\mathbf{b}$ to denote estimates of or candidates for $\boldsymbol{\beta}$.
- A residual is the fitted value given a particular potential parameter vector:

$$e_i(\mathbf{b}) = y_i - \mathbf{x}_i'\mathbf{b}$$

- A residual captures the degree to which the linear prediction $\mathbf{x}_i'\mathbf{b}$ explains the dependent variable y_i.
- Formally, we should think of residuals as being a function of candidate parameter vectors, but we will often just write e_i.
- The vector of residuals across observations is $\mathbf{e}(\mathbf{b})$

Least Squares

- It makes intuitive sense to want to find a value of $\mathbf{b}$ that makes residuals small, so that the estimated model explains the data well.
- There are many possible ways to think about making the residuals small, but by far the most popular criterion is the **sum of squared residuals**:

$$SSR(\mathbf{b}) \equiv \sum_{i=1}^{n} e_i(\mathbf{b})^2 = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i'\mathbf{b}\right)^2 = \mathbf{e}(\mathbf{b})'\mathbf{e}(\mathbf{b})$$

- The **least squares estimator** minimizes the sum of squared residuals:

$$\hat{\mathbf{b}}_{LS} \equiv \arg\min_{\mathbf{b}} SSR(\mathbf{b})$$

- For linear models with homoscedastic errors, the least squares estimate is typically called the **Ordinary Least Squares** estimator.

Regression with Single Variable I

- Let's consider a model with only a single variable regressor (and a constant):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- For example, we might have
  - y_i: Dependent Variable (e.g., test score of student i)
  - x_i: Independent Variable (e.g., class size of student i)
  - ε_i: regression error (e.g., noise in the model)

Derivation of Bivariate Linear Regression Estimator

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$SSR\left(b_0, b_1\right) = \sum_{i=1}^{n} e_i\left(b_0, b_1\right)^2 = \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_i\right)^2$$

- (On board) What are the OLS estimates of (β_0, β_1)?

OLS Estimator

- The OLS estimates are:

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{x,y}}{s_x^2}$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

where the s_x^2 refers to the sample variance of x and $s_{x,y}$ refers to the sample covariance of x and y.

Using the OLS Estimator

- Now let's go ahead and actually use the OLS estimator.
- Suppose Alice has a test score of 6 and class size of 15; Bob has a test score of 3 with a class size of 20.
- Our data is:
  - Alice: $(y_1, x_1) = (6, 15)$
  - Bob: $(y_2, x_2) = (3, 20)$
- Note that this is a ridiculously small sample size (the smallest we could have and still solve for the OLS estimator).

Implementing the OLS Estimator

- (On board) Find the OLS estimator for β_0 and β_1 for the data $(y_1, x_1) = (6, 15)$ and $(y_2, x_2) = (3, 20)$
- Formulas:

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{cov\,(x, y)}{var\,(x)}$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

- You will never have to do this algebra yourself. It becomes very cumbersome with lots of observations and with lots of regressors, which we now turn to.

Multiple Linear Regression

- Let's return to the multiple linear regression setting (multiple regressors):

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon,$$

- We can write the sum of squared residuals as follows:

$$SSR(\mathbf{b}) = \mathbf{e}_i(\mathbf{b})' \mathbf{e}_i(\mathbf{b}) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}$$

- The necessary condition for a minimum:

$$\frac{\partial SSR(\mathbf{b})}{\partial \mathbf{b}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} = 0$$

- Recalling the rules:

$$\frac{\partial \mathbf{u}'\mathbf{v}}{\partial \mathbf{v}} = \mathbf{u} \qquad \frac{\partial \mathbf{v}'\mathbf{A}\mathbf{v}}{\partial \mathbf{v}} = 2\mathbf{A}\mathbf{v}$$

The OLS Formula

- The necessary condition for a minimum:

$$\frac{\partial SSR(\mathbf{b})}{\partial \mathbf{b}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} = 0$$

- Thus, $\mathbf{b}$ must satisfy

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$$

- Assuming $\mathbf{X}'\mathbf{X}$ is invertible (be careful!), we have

$$\widehat{\mathbf{b}}_{OLS} \equiv (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- Note the similarity to the bivariate least squares estimator.

Full Rank and Identification

- The invertibility of $\mathbf{X}'\mathbf{X}$ is not guaranteed in general, but it is implied by the full rank condition: $Rank\,(\mathbf{X}) = K$.
- When $\mathbf{X}$ does not have full rank, neither does $\mathbf{X}'\mathbf{X}$, and

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$$

  can be solved by multiple values of $\mathbf{b}$.
- Linear algebra review: when a square matrix $\mathbf{A}$ is not invertible, it has a non-trivial nullspace. This means that

$$\mathbf{A}\mathbf{b} = \mathbf{0}$$

  can be solved by multiple vectors $\mathbf{b}$. This implies that, for any $\mathbf{c}$,

$$\mathbf{A}\mathbf{b} = \mathbf{c}$$

  also has multiple solutions if it has at least one solution.

Properties of Residuals

- Let $\mathbf{e}$ denote the vector of OLS residuals, and consider

$$\mathbf{X}'\mathbf{e} = \mathbf{X}'\left(\mathbf{X}\mathbf{b}_{OLS} - \mathbf{y}\right)$$

- Substituting in the OLS formula,

$$\mathbf{X}'\mathbf{e} = \mathbf{X}'\left(\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} - \mathbf{y}\right)$$

- This simplifies to

$$\mathbf{X}'\mathbf{e} = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{y} = 0$$

- This implies that (1) the OLS residuals sum to zero, given that one of the regressors is a constant, and (2) the OLS residuals are uncorrelated with the regressors. Intuitively: there is no variation left in $\mathbf{e}$ that can be explained by $\mathbf{X}$.

Method of moments preview

- Exercise: show that $\mathbf{b}_{OLS}$ is the unique value of β satisfying

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) = 0$$

  given that $\mathbf{X}$ is full rank.

- Note that $\mathbf{X}$ is full rank $\Rightarrow \mathbf{X}'\mathbf{X}$ is full rank.

- Given this uniqueness result, we could use this condition to define the OLS estimator instead of the least squares definition. This previews **method of moments** estimators, which we will discuss later.

Multiple Regression Coefficients

- We saw that for a bivariate regression, the slope on the regressor is given by

$$\hat{b}_1 = \frac{s_{x,y}}{s_x^2}$$

- Does it follow that, with multiple regressors, the coefficient on the kth regressor is

$$\hat{b}_k = \frac{s_{x_k,y}}{s_{x_k}^2}?$$

- If not, under what conditions?

Silly Example

- Consider the following model of rectangular painting area:

$$\ln A_i = \beta_0 + \beta_1 \ln W_i + \beta_2 \ln H_i + \varepsilon_i$$

  where
  - W_i is the painting i's width
  - H_i is the painting i's width
  - A_i is the painting i's area, $A_i = W_i \cdot H_i$
  - ε_i is measurement error in painting i's area

- What should the β's be in theory (i.e., given what you know about geometry)?

Silly Example II

- Suppose $\ln W_i$ and $\ln H_i$ are correlated in the population (naturally)

$$\begin{pmatrix} \ln W_i \\ \ln H_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_w \\ \mu_h \end{pmatrix}, \begin{pmatrix} \sigma_w^2 & \sigma_{wh} \\ \sigma_{wh} & \sigma_h^2 \end{pmatrix} \right)$$

- Then, notice that

$$\begin{aligned} Cov\left(\ln A_i, \ln W_i\right) &= Cov\left(\ln W_i + \ln H_i, \ln W_i\right) \\ &= Var\left(\ln W_i\right) + Cov\left(\ln W_i, \ln H_i\right) \end{aligned}$$

- Finally,

$$\frac{Cov\left(\ln A, \ln W\right)}{Var\left(\ln W\right)} = \frac{Var\left(\ln W\right) + Cov\left(\ln H, \ln W\right)}{Var\left(\ln W\right)} = 1 + \frac{Cov\left(\ln H, \ln W\right)}{Var\left(\ln W\right)}$$

Silly Example III

- In a bivariate regression of A on W (ommiting H), recall the slope would be given by

$$\frac{Cov\left(\ln A, \ln W\right)}{Var\left(\ln W\right)} = 1 + \frac{Cov\left(\ln H, \ln W\right)}{Var\left(\ln W\right)},$$

  so the term $\frac{Cov(\ln H, \ln W)}{Var(\ln W)}$ represents bias.

- Implication: if the OLS coefficients in a multiple regression had the same formula as the coefficients in a bivariate regression, there would be bias.

- Exception: if the regressors $\ln H$ and $\ln W$ are uncorrelated, there will be no bias above, and OLS with multiple regressors delivers the same coefficients as if we ran a bivariate regression with each of the regressors separately.

- This example also makes a point about omitted variables bias: if height is unobserved, the regression of $\ln A$ on only $\ln W$ will deliver the biased coefficient above.

Silly Example IV

$$\widehat{\mathbf{b}}_{OLS} \equiv \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

- Indeed, OLS with multiple regressors does not deliver the bivariate regression coefficient $s_{x_k y}/s^2_{x_k}$ for each regressor x_k (except in the case of orthogonal regressors)

- To gain some intuition for what this formula is doing, let's consider the $\left(\mathbf{X}'\mathbf{X}\right)$ and $\mathbf{X}'\mathbf{y}$ pieces separately in the context of the silly model of painting area.

Silly Example V

- Suppose that $w_i = \ln W_i - \mu_w$ and $\ln h_i = \ln H_i - \mu_h$
- Let $\mathbf{X} = [\mathbf{w} \quad \mathbf{h}]$
- It follows that

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \sum (\ln W_i - \mu_w)^2 & \sum (\ln W_i - \mu_w)(\ln H_i - \mu_h) \\ \sum (\ln W_i - \mu_w)(\ln H_i - \mu_h) & \sum (\ln H_i - \mu_h)^2 \end{pmatrix}$$

Note: if we multiply that by n^{-1}, we have the sample analog of the covariance matrix.

Silly Example VI

- Similarly,

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} \sum (\ln W_i - \mu_w)(\ln A_i - \mu_a) \\ \sum (\ln H_i - \mu_h)(\ln A_i - \mu_a) \end{pmatrix},$$

  is the sample covariance of $\mathbf{x}$ and $\mathbf{y}$ (times n).

- Therefore, we have

$$\widehat{\mathbf{b}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \left(n^{-1}\mathbf{X}'\mathbf{X}\right)^{-1}\left(n^{-1}\mathbf{X}'\mathbf{y}\right) = S_{xx}^{-1}s_{xy}$$

  where S_{xx} is the sample covariance matrix for $\mathbf{x}$ and s_{xy} is the sample covariance of $\mathbf{x}$ and y.

Silly Example VII

$$\widehat{\mathbf{b}}_{OLS} = S_{xx}^{-1} s_{xy}$$

- The population moments for the model of painting areas are

$$Var(\mathbf{x}) = \begin{pmatrix} \sigma_w^2 & \sigma_{wh} \\ \sigma_{wh} & \sigma_h^2 \end{pmatrix} \qquad Cov(\mathbf{x}, y) = \begin{pmatrix} \sigma_w^2 + \sigma_{wh} \\ \sigma_h^2 + \sigma_{wh} \end{pmatrix}$$

- Note that

$$(Var(\mathbf{x}))^{-1} = \frac{1}{\sigma_w^2 \sigma_h^2 - \sigma_{wh}^2} \begin{pmatrix} \sigma_h^2 & -\sigma_{wh} \\ -\sigma_{wh} & \sigma_w^2 \end{pmatrix}$$

- With some algebra, the population-moments version of OLS gives us the right coefficients:

$$(Var(\mathbf{x}))^{-1} Cov(\mathbf{x}, y) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

The Frisch-Waugh Theorem I

- Think about separating $\mathbf{X}$ into two sub-matrices:

$$\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2],$$

with

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

Frisch-Waugh Theorem

The OLS regression of $\mathbf{y}$ on $[\mathbf{X}_1, \mathbf{X}_2]$ yields a subvector $\mathbf{b}_2$ of coefficient estimates that is the same as the result from a regression of the residuals from a regression of $\mathbf{y}$ on $\mathbf{X}_1$ are regressed on the residuals from a regression of $\mathbf{X}_2$ on $\mathbf{X}_1$.

The Frisch-Waugh Theorem II

- In other words, let's start by regressing $\mathbf{y}$ on $\mathbf{X}_1$. Let's label the residuals from this regression $\mathbf{y}^*$.
- Let's also regress $\mathbf{X}_2$ on $\mathbf{X}_1$ (think about regressing each column of $\mathbf{X}_2$ on $\mathbf{X}_1$). Let's label the residuals from this regression $\mathbf{X}_2^*$.
- If we regress $\mathbf{y}^*$ on $\mathbf{X}_2^*$, we get the same coefficient on $\mathbf{X}_2^*$ that we would have had in the full regression of $\mathbf{y}$ on $[\mathbf{X}_1, \mathbf{X}_2]$.
- An implication is that the coefficient on each variable can be thought of as the effect of that variable after controlling for all the other variables. Thus, OLS coefficients are sometimes called **partial regression coefficients**.

The Frisch-Waugh Theorem III

- Bivariate linear regression is easy to visualize (scatter plot with a line running through). Frisch-Waugh tells us how we can visualize the effect of a single variable from a multiple regression.
- One implication of Frisch-Waugh is that, if we de-mean all the variables and then run a regression with the de-meaned variables (but leaving out the constant term in $\mathbf{X}$), we will get the same coefficients on all the variables.
  - Exercise: given the Frisch-Waugh theorem, show this formally.

Units and Coefficients

- Recalling that $E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\beta$, a natural interpretation of the coefficients is is

$$\beta_1 = \frac{d}{dX_{1i}} E[y_i|\mathbf{X}]$$

- In simple linear regression, it's easy to see that re-scaling (changing the units of x) will rescale the parameter estimates in the opposite way:

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  The same is true in the multiple regression framework (Frisch-Waugh makes this easier to see).

- Exercise: if the regressor x in a bivariate regression is the log of a variable, show that rescaling the original variable does not affect $\hat{b}_1$. What about $\hat{b}_0$?

Goodness of Fit

- The **total sum of squares**, or the **total variation** in the dependent variable:

$$SST = (\mathbf{y} - \mathbf{i}\overline{y})'(\mathbf{y} - \mathbf{i}\overline{y})$$

- The dependent variable decomposes into a prediction and a residual:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} = \widehat{\mathbf{y}} + \mathbf{e}$$

  where $\mathbf{Xb} = \widehat{\mathbf{y}}$ is the prediction or **fitted value** of y.

- We can think about decomposing the variation in y into variation in $\widehat{\mathbf{y}}$ and $\mathbf{e}$. Intuitively, we want the variation in $\widehat{\mathbf{y}}$ to account for as much as possible

Sums of Squares

- Define
$$\mathbf{M}_0 = I - n^{-1}\mathbf{i}\mathbf{i}'.$$
note that $\mathbf{M}_0$ is symmetric and **idempotent**.
- Notice that $\mathbf{y}'\mathbf{M}_0\mathbf{y}$ is the SST.
- Also, we can show that
$$\mathbf{y}'\mathbf{M}_0\mathbf{y} = \mathbf{b}'\mathbf{X}'\mathbf{M}_0\mathbf{X}\mathbf{b} + \mathbf{e}'\mathbf{e},$$
where the first term on the RHS is known as the **regression sum of squares (SSR)**, and the second term is the **error sum of squares (SSE)**.

$$SST = SSR + SSE$$

Coefficient of Determination

- The coefficient of determination is defined as the proportion of the variation in the dependent variable explained by the model:

$$R^2 = \frac{SSR}{SST} = \frac{\mathbf{b}'\mathbf{X}'\mathbf{M}_0\mathbf{X}\mathbf{b}}{\mathbf{y}'\mathbf{M}_0\mathbf{y}} = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{M}_0\mathbf{y}}$$

- R^2 is a measure of goodness of fit that always goes up as we add more regressors. It doesn't tell us whether it's "worth it" to add a new regressor to the model. (Later we will talk about why overfitting can be bad in finite samples.)
- **Adjusted** R^2 incorporates a penalty for the number of regressors:

$$\overline{R}^2 = 1 - \frac{\mathbf{e}'\mathbf{e}/(n-K)}{\mathbf{y}'\mathbf{M}_0\mathbf{y}/(n-1)}$$

Finite Sample Properties

- Given assumptions (1)-(3), homoscedasticity, and normal error terms,

$$\mathbf{b}_{OLS}|\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$$

- Proof on board
- Note: proof that it's unbiased and expression for variance do not require normally error terms, but it would be hard to say what the finite sample distribution is, exactly, without normality.

Standard Errors

$$\mathbf{b}_{OLS}|\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$$

- $\sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$ is the covariance matrix of the parameter estimates – it tells us how precise $\mathbf{b}_{OLS}$ is as an estimate of $\boldsymbol{\beta}$.
- The diagonal elements of $\sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$ are the variances of each of the parameter estimates. The square roots of those are **standard errors**. Note that standard errors are standard deviations of the distribution of an estimator.
- The off-diagonal elements of $\sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$ are also important (especially when we get to hypothesis testing). They indicate whether two parameter estimates are correlated.

Estimating Standard Errors

$$\mathbf{b}_{OLS}|\mathbf{X} \sim \mathcal{N}\left(\beta, \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$$

- $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ is just data, but σ^2 must be estimated (recall that it's the variance of the normally distributed error terms).
- Intuitively, the residuals are estimates of the error terms, so the sample variance of residuals is what we use to estimate σ^2:

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K}$$

- Similar to how the unbiased estimator of the variance of a random variable with unknown mean requires us to divide by $n-1$, here we have to divide by $n-K$, where K is the number of parameters being estimated.

Mean-squared error

- Consider a parameter vector γ and consider $\mathbf{x}'\gamma$ as a predictor of y.
- The **mean squared error** of this predictor is

$$MSE = E\left[\left(y - \mathbf{x}'\gamma\right)^2\right]$$

- This can be written

$$MSE = E\left[\left(y - E\left[y|\mathbf{x}\right]\right)^2\right] + E\left[\left(E\left[y|\mathbf{x}\right] - \mathbf{x}'\gamma\right)^2\right]$$

  (check this by expanding and applying the law of iterated expectations)
- Result: OLS is the value of γ that minimizes mean square error. This does not require normality of the error terms. (Proof on board)

Gauss-Markov Theorem

Gauss-Markov Theorem

In the linear regression model with homoscedastic errors, The OLS estimator is the **best linear unbiased estimator** (BLUE).

- "Best" means estimator with lowest variance – most precise
- Normally distributed error not assumed here
- If error terms are not homoscedastic, OLS is still unbiased given strict exogeneity, but lower-variance estimators are possible (see: Generalized Least Squares)
- Proof on board

Generalized Least Squares

- The proof that OLS is unbiased did not rely on homoskedasticity. OLS is "LUE" even without homoskedasticity. OLS is only **B**LUE (minimal variance) with homoskedasticity.
- Without homoskedasticity, but maintaining our other assumptions, **Generalized Least Squares** is BLUE. That is, suppose,

$$Var\left[\varepsilon|\mathbf{X}\right] = \mathbf{\Omega}$$

then the BLUE estimator is

$$\mathbf{b}_{GLS} = \left(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y}.$$

- To implement GLS, note that $\mathbf{\Omega}$ must be estimated (**Feasible Generalized Least Squares**). In contrast, OLS estimates the β parameters without needing to know or estimate anything about variances; σ^2 is only needed to quantify standard errors.

Endogeneity

$$\mathbf{b}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

- Note that our argument for OLS being unbiased involved a step where we argue that

$$E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\right] = 0$$

  thanks to strict endogeneity.

- If one of the regressors is correlated with ε, strict exogeneity is violated, and this term will cause bias.

Omitted Variables Bias

- Suppose the econometrician only observes regressors $\mathbf{X}$, but the true model is

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{z}\gamma + \varepsilon,$$

- The OLS estimator will equal

$$\mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{z}\gamma + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon$$

- The last term is mean zero given the strict exogeneity assumption.
- Note that the second term will not be zero if $\mathbf{X}$ and $\mathbf{z}$ are correlated; i.e. if $\mathbf{X}'\mathbf{z} \neq 0$.
- Implication: correlation between omitted variables and the observed regressors makes OLS biased.

Bias with multiple variables

- Consider a model with two variables:

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

and suppose that

$$\begin{aligned} x_{1,i} &= u_{1,i} + u_{3,i} + \varepsilon_i \\ x_{2,i} &= u_{2,i} + u_{3,i} \end{aligned}$$

where the u's and ε are all independently distributed.

- We know that OLS will deliver a biased estimate of β_1, but will OLS still be consistent for β_2?

Bias with multiple variables

- Consider a model with two variables:

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

and suppose that

$$\begin{aligned} x_{1,i} &= u_{1,i} + u_{3,i} + \varepsilon_i \\ x_{2,i} &= u_{2,i} + u_{3,i} \end{aligned}$$

where the u's and ε are all independently distributed.

- We know that OLS will deliver a biased estimate of β_1, but will OLS still be consistent for β_2?
- No! The correlation between x_1 and x_2 leads to bias even in β_2.
  - Thus, endogeneity problems are a big deal not only for our variables of interest; endogenous control variables can also create problems.

Bias with multiple variables II

- Consider "controlling" for x_1 with the wrong estimate b_1 of β_1:

$$y_i - b_1 x_{1,i} = \beta_0 + \beta_2 x_{2,i} + \varepsilon_i + (\beta_1 - b_1) x_{1,i},$$

- Given that $b_1 \neq \beta_1$, we get some stuff in the error term:

$$y_i - b_1 x_{1,i} = \beta_0 + \beta_2 x_{2,i} + \varepsilon_i + (\beta_1 - b_1)(u_{1,i} + u_{3,i} + \varepsilon_i),$$

  and we now have an endogeneity problem because $x_{2,i} = u_{2,i} + u_{3,i}$ and $u_{3,i}$ is now in the error term.

- Note: we could formalize this intuition using the Frisch-Waugh Theorem.

Irrelevant Variables

- What if we include a variable in $\mathbf{X}$ that doesn't actually affect y?
- This can be accommodated in the linear regression framework as a regressor that has a coefficient $\beta_k = 0$.
- Thus, this does not create any bias in the other coefficients, and the expected coefficient on the irrelevant variable is zero.
- However, this can reduce the precision of the estimates of the other coefficients, and it's not generally a good idea to add as many variables to a regression as you can (**overfitting**). More on this later.

Finite Sample vs. Asymptotic Properties

- It's rare to have small-sample properties of an estimator
- Econometric studies typically do Monte Carlo (simulation) studies to learn about finite-sample performance of estimators.
- For most estimators, we derive asymptotic properties, i.e.,

$$\sqrt{n}\left(\mathbf{b}_{OLS} - \beta\right) \Rightarrow_d \mathcal{N}\left(0, \Sigma\right)$$

- The above statement says that the OLS estimator **converges in distribution** to a normally distributed variable.
- Part of that result is that OLS is **consistent**. Formally, consistency means that an estimator **converges in probability** to the true value.

Asymptotics: Setup

- Asymptotic analysis refers to analyzing statistics of the data as the number of observations n grows to infinity (e.g., the central limit theorem).
- To do asymptotic analysis, we specify a **data generating process**.
- A data generating process describes the distribution of a sequence of observations $(\mathbf{x}_i, \varepsilon_i)$ for $i = 1, 2, \ldots n$.
- The simplest case is assuming that observations are **independent and identically distributed** (i.i.d)
- In other settings, the observations are allowed to be correlated, but in a limited way that requires the dependence between "distant" observations to go to zero – see **stationarity** and **ergodicity** (e.g., time series, and more recently some spatial analysis)
- We also need the data to be "well-behaved", i.e. $E\left[\mathbf{x}\mathbf{x}'\right]$ must have full rank and the error terms must have finite variance.

OLS Asymptotic Distribution

- We can use the central limit theorem to show that OLS is asymptotically normally distributed:

$$\mathbf{b}_{OLS} \sim^{a} \mathcal{N}\left(\boldsymbol{\beta}, \frac{\sigma^2}{n} E\left[\mathbf{x}_i \mathbf{x}_i'\right]^{-1}\right)$$

  where $\sim^{a}$ signifies convergence in distribution.
- This is just like the finite-sample distribution we got with normally distributed error terms.