Linear Regression

Junhui Qian

October 27, 2014

Outline

- The Model
- Estimation
  - Ordinary Least Square
  - Method of Moments
  - Maximum Likelihood Estimation
- Properties of OLS Estimator
  - Unbiasedness
  - Consistency
  - Efficiency

What is linear regression?

- We want to explain an economic variable y using x, which is usually a vector.
- For example, y may be the wage of an individual, and x include factors such as experience, education, gender, and so on.
- Let $x = (1, x_1, ..., x_k)$, and let its realization for ith-individual be $x_i = (1, x_{i1}, ..., x_{ik})'$, we may write:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u. \quad (1)$$

Some Terminologies

Now we have

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u. \tag{2}$$

- y is called the "dependent variable", the "explained variable", or the "regressand"
- The elements in x are called the "independent variables", "explanatory variables", "covariates", or the "regressors".
- β's are coefficients. In particular, β_0 is usually called "intercept parameter" or simply called "constant term", and $(\beta_j, 1 \leq j \leq k)$ are usually called slope parameters.
- u is called the "error term", "residuals", or disturbances and represents factors other than x that affect y.

An Example

- We may have an econometric model of wages:

$$wage_i = \beta_0 + \beta_1 edu_i + \beta_2 expr_i + u_i$$

- edu_i denotes the education level of individual i in terms of years of schooling and $expr_i$ denotes the working experience of individual i.
- β_0 is the constant term or the intercept. It measures what a male worker would expect to get if he has zero education and zero experience.
- β_1 is the slope parameter for the explanatory variable edu. It measure the marginal increase in wage if a worker gains additional year of schooling, **holding other factors fixed**, or **controlling for other factors.**
- u_i may include the gender, the luck, or the family background of the individual, etc.

Partial Effects

- $(\beta_j, j = 1, .., k)$ can be interpreted as "partial effects".
- For example, for the wage equation, we have

$$\Delta wage = \beta_1 \Delta edu + \beta_2 \Delta expr + \Delta u.$$

  If we hold all factors other than *edu* fixed, then

$$\Delta wage = \beta_1 \Delta edu.$$

- So β_1 is the partial effect of education on wage.
- We say: With one unit of increase in *edu*, an individual's wage increases by β_1, holding other factors fixed, or controlling for other factors.

Econometric Clear Thinking

- Whenever we make comparisons or inferences, we should hold relevant factors fixed.
- This is achieved in econometrics by multiple linear regression.
- The partial effects interpretation is not without problem. It is partial equilibrium analysis.
- We may have the socalled "general equilibrium problem", which happens when a change in a variable leads to changes in the structure of regression equation.
- In most cases, however, partial effects analysis is a good approximation, or, the best alternative.

Classical Linear Regression Assumptions

(1) Linearity

(2) Random sampling $\dashrightarrow$ (x_i, y_i) are iid across i

(3) No perfect collinearity $\Leftrightarrow$ Any element in x cannot be represented by the linear combination of other elements.

(4) Zero conditional mean $\Leftrightarrow$

$$\mathbb{E}(u|x) = \mathbb{E}(u|x_1, x_2, ..., x_k) = 0.$$

(5) Homoscedasticity

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

(6) Normality

$$u|x \sim N(0, \sigma^2).$$

More on Linearity

- Linearity can be achieved by transformation.
- For example, we may have

$$\log(wage_i) = \beta_0 + \beta_1 \log(exper_i) + \beta_2 \log(educ_i) + \beta_3 female_i + u_i.$$

- Now the parameter β_1 represents the elasticity of wage with respect to changes in experiences:

$$\beta_1 = \frac{\partial \log(wage_i)}{\partial \log(exper_i)} = \frac{\partial wage_i / wage_i}{\partial exper_i / exper_i} \approx \frac{\Delta wage_i / wage_i}{\Delta exper_i / exper_i}.$$

More on No Perfect Collinearity

True or False?

(1) "No Perfect Collinearity" does not allow correlation. For example, the following is perfect collinearity:

$$testscore = \beta_0 + \beta_1 eduExpend + \beta_2 familyIncome + u.$$

(2) The following model suffers from perfect collinearity:

$$cons = \beta_0 + \beta_1 income + \beta_2 income^2 + u.$$

(3) The following model suffers from perfect collinearity:

$$\log(cons) = \beta_0 + \beta_1 \log(income) + \beta_2 \log(income^2) + u.$$

(4) The following model suffers from perfect collinearity:

$$cons = \beta_0 + \beta_1 husbIncome + \beta_2 wifeIncome + \beta_3 familyIncome + u.$$

More on Zero Conditional Mean

- If $\mathbb{E}(u|x) = 0$, we call x "exogenous".
- If $\mathbb{E}(u|x) \neq 0$, we call x "endogenous".
- The notion of being "exogenous" or "endogenous" can be understood in the following model,

$$L = \alpha W + \gamma X + u,$$

where both the employment level (L) and the average wage (W) are endogenous variables, while the foreign exchange rate (X) can be considered exogenous. The residual u should contain shocks from both supply and demand sides.

Endogenous Wage

If the employment level and the average wage are determined by

$$\begin{aligned} L_s &= bW + v_s \\ L_d &= aW + cX + v_d \\ L_d &= L_s, \end{aligned}$$

Then we can solve for the equilibrium employment and wage rate:

$$\begin{aligned} W &= \frac{c}{b-a}X - \frac{v_s - v_d}{b-a} \\ L &= \frac{bc}{b-a}X - \frac{av_s - bv_d}{b-a}. \end{aligned}$$

It is obvious that $\mathrm{cov}(W, v_d) \neq 0$ and $\mathrm{cov}(W, v_s) \neq 0$. Thus W should be correlated with u, hence the endogeneity in econometric sense.

More on Zero Conditional Mean

- In econometrics, we call an explanatory variable x "endogenous" as long as $\mathbb{E}(u|x) \neq 0$.
- Usually, nonzero conditional mean is due to
  - Endogeneity
  - Missing variables (e.g., ability in wage equation)
  - Wrong functional form (e.g., missing quadratic term)

More on Homoscedasticity

- If $\mathrm{var}(u_i|x_i) = \sigma^2$, we call the model "homoscedastic". If not, we call it "heteroscedastic".
- Note that $\mathrm{var}(u_i|x_i) = \mathrm{var}(y_i|x_i)$. If $\mathrm{var}(y_i|x_i)$ is a function of some regressor, then there would be heteroscedasticity.
- Examples of heteroscedasticity
  - Income v.s. Expenditure on meals
  - Gender v.s. Wage
  - $\vdots$

Ordinary Least Square

We have

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i.$$

The OLS method is to find *beta*'s such that the sum of squared residuals (SSR) is minimized:

$$\mathrm{SSR}(\beta_0, ..., \beta_k) = \sum_{i=1}^{n} [(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})]^2.$$

- OLS minimizes *a* measure of fitting error.

First Order Conditions of OLS

To minimize SSR, we find the first-order conditions of the minimization problem:

$$
\begin{aligned}
\frac{\partial \mathrm{SSR}}{\partial \beta_0} &= 0 \\
\frac{\partial \mathrm{SSR}}{\partial \beta_1} &= 0 \\
&\vdots \\
\frac{\partial \mathrm{SSR}}{\partial \beta_k} &= 0
\end{aligned}
$$

First Order Conditions of OLS

We obtain:

$$\begin{aligned}
2\sum_{i=1}^{n}((y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik})) &= 0 \\
2\sum_{i=1}^{n}((y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}))x_{i1} &= 0 \\
&\vdots \\
2\sum_{i=1}^{n}((y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}))x_{ik} &= 0.
\end{aligned}$$

We have $(1+k)$ equations for $(1+k)$ unknowns. If there is no perfect collinearity, we can solve for these equations.

OLS for Naive Regression

We may have the following model

$$y_i = \beta x_i + u.$$

Then the first-order condition is:

$$\sum_{i=1}^{n} (y_i - \hat{\beta} x_i) x_i = 0$$

We obtain

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

OLS for Simple Regression

The following is called a "simple regression":

$$y_i = \beta_0 + \beta_1 x_i + u_i.$$

Then the first-order conditions are:

$$\begin{aligned} \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) &= 0 \\ \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))x_i &= 0 \end{aligned}$$

OLS for Simple Regression, Continued

From the first-order conditions, we obtain

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

and $\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$, where $\bar{x} = 1/n\sum_{i=1}^{n} x_i$ and $\bar{y} = 1/n\sum_{i=1}^{n} y_i$.

More on OLS for Simple Regression

From $y = \beta_0 + \beta_1 x + u$, we have

$$\beta_1 = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)}.$$

And we have obtained

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

Hence β_1 measures the correlation between y and x.

True or False?

In the simple regression model,

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

- β_0 is the mean of y
- β_1 is the correlation coefficient between x and y

Estimated Residual

Let

$$\hat{u}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i).$$

From the first-order conditions we have

$$\bar{\hat{u}} = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i = 0,$$

and

$$\frac{1}{n} \sum_{i=1}^{n} x_i \hat{u}_i = 0.$$

Connection between Simple and Naive

We have

$$\begin{cases} y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i \\ \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} + 0. \end{cases}$$

Hence

$$y_i - \bar{y} = \hat{\beta}_1 (x_i - \bar{x}) + \hat{u}_i.$$

Using the formula for the naive regression, we obtain:

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

Regression Line

For the simple regression model,

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

We can define a “regression line”:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x.$$

It is easy to show that

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}.$$

In Matrix Language

- For models with two or more regressors, the expression for $\hat{\beta}_i$ are very complicated.
- However, we can use matrix language to obtain more beautiful and more memorable expressions. Let

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

  Then we may write the multiple regression as

$$Y = X\beta + u.$$

Some Special Vectors and Matrices

- Vector of ones, $\iota = (1, 1, ..., 1)'$. For a vector of the same length, we have

$$\sum_{i=1}^{n} x_i = x'\iota = \iota' x, \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} x_i = (\iota'\iota)^{-1}\iota' x.$$

- Vector of standard basis, $e_1 = (1, 0, 0, ..., 0)'$, $e_2 = (0, 1, 0, ..., 0)'$, etc.
- Identity matrix, I.
- Projection matrix, square matrices that satisfy $P^2 = P$.
  - If P is symmetric, it is called an orthogonal projection (e.g., $P = X(X'X)^{-1}X'$)
  - Oblique projection, e.g., $P = X(W'X)^{-1}W'$, $P = \begin{bmatrix} 0 & 0 \\ \alpha & 1 \end{bmatrix}$.
- If P is an orthogonal projection, so is $I - P$.

Range of Matrix

- The span of a set of vectors is the set of all linear combinations of the vectors.
- The range of a matrix X, $\mathcal{R}(X)$, is the span of the columns of X.
- $\mathcal{R}(X)^\perp$ is the orthogonal complement $\mathcal{R}(X)$, which contains all vectors that is orthogonal to $\mathcal{R}(X)$.
  - Two vectors, x and y, are orthogonal if $x \cdot y = x'y = 0$.
  - A vector y is orthogonal to a subspace U if for all $x \in U$, $x \cdot y = 0$.
  - $\mathcal{R}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right)$ is the x-axis.

Orthogonal Projection on $\mathcal{R}(X)$

- By definition, the orthogonal projection of y on $\mathcal{R}(X)$ can be represented by $X\beta$, where β is a vector. We denote

$$\text{proj}(y|X) \equiv P_X y = X\beta.$$

- $y - X\beta$ should be orthogonal to every element in $\mathcal{R}(X)$, which include every column of X. Then we may solve

$$X'(y - X\beta) = 0$$

  and obtain $\beta = (X'X)^{-1}X'y$. Hence $P_X = X(X'X)^{-1}X'$ is the orthogonal projection on $\mathcal{R}(X)$.
- $I - P_X$ is the orthogonal projection on $\mathcal{R}(X)^{\perp}$, or equivalently, $\mathcal{N}(X')$.

Vector Differentiation

- Let $z = (z_1, ..., z_k)$ be a vector of variables and $f(z)$ be a function of z. Then

$$\frac{\partial f}{\partial z} = \begin{pmatrix} \frac{\partial f}{\partial z_1} \\ \frac{\partial f}{\partial z_2} \\ \vdots \\ \frac{\partial f}{\partial z_k} \end{pmatrix}.$$

Vector Differentiation

- In particular, if $f(z) = a'z$, where a is a vector of constants. Then
$$\frac{\partial}{\partial z}(a'z) = a = \frac{\partial}{\partial z}(z'a)$$
- If $f(z) = Az$ is a vector-valued function, where A is a matrix, then
$$\frac{\partial}{\partial z}(Az) = A'.$$

Vector Differentiation of Quadratic Form

If $f(z) = z'Az$, where A is a matrix, then

$$\frac{\partial}{\partial z}(z'Az) = (A + A')z.$$

If A is symmetric, ie, $A = A'$, then

$$\frac{\partial}{\partial z}(z'Az) = 2Az.$$

In particular, when $A = I$, the identity matrix, then

$$\frac{\partial}{\partial z}(z'z) = 2z.$$

OLS in Matrix

- The least square problem can be written as
$$\min_{\beta}(Y - X\beta)'(Y - X\beta).$$
- The first-order condition in matrix form:
$$2X'(Y - X\hat{\beta}) = 0.$$
- Solving for β,
$$\hat{\beta} = (X'X)^{-1}X'Y.$$
- The matrix of $X'X$ is invertible since we rule out perfect collinearity.
- $X\hat{\beta}$ is nothing but the orthogonal projection of Y on $\mathcal{R}(X)$.
- If there is only one regressor and there is no constant term, X is a vector. Then the above expression reduces to the naive linear regression estimator.

An Equivalent Derivation

- The least square problem can be written as

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2,$$

where $x_i = (1, x_{i1}, ..., x_{ik})$ and $\beta = (\beta_0, \beta_1, ..., \beta_k)$.

- The first-order condition in matrix form:

$$\sum_{i=1}^{n} 2x_i(y_i - x_i'\hat{\beta}) = 0.$$

- Solving for β,

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i x_i' \right)^{-1} \left(\sum_{i=1}^{n} x_i y_i \right).$$

Equivalence

- We can check that

$$X'X = \sum_{i=1}^{n} x_i x_i', \quad \text{and} \quad X'Y = \sum_{i=1}^{n} x_i y_i.$$

- If there is only one regressor and there is no constant term, x_i is a scalar. Then the above expression reduces to the naive linear regression estimator.

The Population Moments

From the assumption $\mathbb{E}(u|x) = 0$, we have

$$\mathbb{E}(u) = 0, \quad \text{and} \quad \mathbb{E}(ux_j) = 0, j = 1, ..., k.$$

This is

$$\begin{cases} \mathbb{E}(y - (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)) = 0 \\ \mathbb{E}((y - (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k))x_1) = 0 \\ \vdots \\ \mathbb{E}((y - (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k))x_k) = 0 \end{cases}$$

The above equations are called "moment conditions".

The Sample Moments

We can estimate population moments by sample moments. For example, the sample moment of $\mathbb{E}(u)$ is

$$\frac{1}{n}\sum_{i=1}^{n} u_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})).$$

Similarly, the sample counterpart of $\mathbb{E}(ux_j) = 0$ is

$$\frac{1}{n}\sum_{i=1}^{n} u_i x_{ij} = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}))x_{ij} = 0.$$

Method of Moments (MM)

Plug the sample moments into the moment conditions, we obtain

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik})) = 0,$$

and

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}))x_{ij} = 0, j = 1, ..., k.$$

We can see that these equations are the same as those in the first-order conditions of the OLS.

The Distribution Assumption

- Under CLR Assumption (6), u is normally distributed with mean 0 and variance σ^2. The density function of u is given by

$$p(u;\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{u^2}{2\sigma^2}\right).$$

- Then we can estimate the linear regression model using MLE.
- More generally, we can assume other distributional function for u, t-distribution for example.

Likelihood Function

- By the Assumption (2), random sampling, the joint distribution of $(u_1, ..., u_n)$ is

$$p(u_1, ..., u_n; \theta) = p(u_1; \theta)p(u_2; \theta) \cdots p(u_n; \theta).$$

- Given observations (Y, X), the likelihood function is defined as

$$\begin{aligned} p(\beta, \theta | y, X) &= p(y_1 - x_1'\beta, ..., y_n - x_n'\beta; \theta) \\ &= p(y_1 - x_1'\beta; \theta)p(y_2 - x_2\beta; \theta) \cdots p(y_n - x_n'\beta; \theta). \end{aligned}$$

Maximum Likelihood Estimation

- MLE implicitly assumes that what happens should most likely happen.
- MLE is to solve for $\hat{\beta}$ and $\hat{\theta}$ such that the likelihood function is maximized,
$$\max_{\beta,\theta} p(\beta, \theta | y, X).$$
- In practice, we usually maximize the log likelihood function:
$$\log(p(\beta, \theta | y, X)) = \sum_{i=1}^{n} \log(p(y_i - x_i'\beta; \theta)).$$

MLE of Classical Linear Regression

- We assume $u_i \sim \text{iid } N(0, \sigma^2)$.
- The log likelihood function is

$$\log(p(\beta, \sigma | y, X)) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2.$$

- The first-order condition for β is

$$\sum_{i=1}^{n} x_i(y_i - x_i'\hat{\beta}) = 0.$$

- This yields the same $\hat{\beta}$ as in OLS.

Definition

- We call an estimator $\hat{\beta}$ is unbiased if

$$\mathbb{E}\hat{\beta} = \beta.$$

- $\hat{\beta}$ is a random variable. For example, the OLS estimator $\hat{\beta} = (X'X)^{-1}X'Y$ is random since both X and Y are sampled from a population.
- Given a sample, however, $\hat{\beta}$ is determined. So unbiasedness is NOT a measure of how good a particular estimate is, but a property of a good procedure.

The Unbiasedness of OLS Estimator

Theorem: Under Assumptions (1) through (4), we have

$$\mathbb{E}(\hat{\beta}_j) = \beta_j, \ j = 0, 1, ..., k.$$

Proof:

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}(X'X)^{-1}X'Y = \mathbb{E}(X'X)^{-1}X'(X\beta + U) = \beta + \mathbb{E}(X'X)^{-1}X'U = \beta.$$

Omitted Variable Bias

- When we, mistakenly or due to lack of data, exclude one or more relevant variables, OLS yields biased estimates. This bias is called "omitted variable bias".
- For example, suppose the wage of a worker is determined by both his education and his innate ability:

$$wage = \beta_0 + \beta_1 education + \beta_2 abililty + u.$$

The ability, however, is not observable. We may have to estimate the following model,

$$wage = \beta_0 + \beta_1 education + v,$$

where $v = \beta_2 abililty + u$.

The General Case

Suppose the true model, in matrix form, is

$$Y = X_1\beta_1 + X_2\beta_2 + U, \tag{3}$$

where β_1 is the parameter of interest. However, we omit X_2 and estimate

$$Y = X_1\beta_1 + V. \tag{4}$$

Denote the OLS estimator of β_1 in (3) as $\hat{\beta}_1$ and the OLS estimator of β_1 in (4) as $\tilde{\beta}_1$. Then

$$\mathbb{E}(\tilde{\beta}_1 | X_1, X_2) = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2.$$

Formula of Omitted-Variable Bias

Suppose we only omit one relevant variable, ie, X_2 is a vector. Then $(X_1'X_1)^{-1}X_1'X_2$ is the OLS estimator of the following regression:

$$X_2 = X_1\delta + W.$$

So we have

$$\mathbb{E}(\tilde{\beta}_1|X_1, X_2) = \beta_1 + \hat{\delta}\beta_2.$$

A Special Case

Suppose the true model is

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

But we estimated

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{v}.$$

- From the formula of omitted-variable bias,

  $$\mathbb{E}(\tilde{\beta}_1 | x_1, x_2) = \beta_1 + \hat{\delta}_1 \beta_2,$$

  where $\hat{\delta}_1$ is the OLS estimate of δ_1 in

  $$x_2 = \delta_0 + \delta_1 x_1 + w.$$

Bias Up or Down?

We have

$$\mathbb{E}(\tilde{\beta}_1 | x_1, x_2) = \beta_1 + \hat{\delta}_1 \beta_2.$$

Recall that δ_1 measures the correlation between x_1 and x_2. Hence we have

OLS Bias	$\text{corr}(x_1, x_2) > 0$	$\text{corr}(x_1, x_2) < 0$
$\beta_2 > 0$		
$\beta_2 < 0$		

Return to Education

- Back to our example, suppose the wage of a worker is determined by both his education and his innate ability:

$$wage = \beta_0 + \beta_1 education + \beta_2 abililty + u.$$

The ability, however, is not observable. We may have to estimate the following model,

$$wage = \beta_0 + \beta_1 education + v,$$

where $v = \beta_2 abililty + u$.

- Are we going to overestimate or underestimate the return to education?

Definition

- We say $\hat{\beta}$ is consistent if

$$\hat{\beta} \to \beta \quad \text{as } n \to \infty.$$

- This basically says, if we observe more and more, we can estimate our model more and more accurately till exactness.

Law of Large Number

Let $x_1, x_2, ..., x_n$ be iid random variables with mean μ. Then

$$\frac{1}{n}\sum_{i=1}^{n} x_i \rightarrow_p \mu.$$

LLN for Vectors and Matrices

- The x_i in LLN can be vectors. And if

$$\mathbb{E}x = \mathbb{E}\begin{pmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,k} \end{pmatrix} = \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix},$$

  then

$$\frac{1}{n}\sum_{i=1}^{n} x_i \to_p \mu.$$

- The same is also true for matrices.

Consistency of OLS Estimator

We have

$$\hat{\beta} = \beta + \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_i u_i \right).$$

If $\mathbb{E} x_i x_i' = Q$, and $\mathbb{E} x_i u_i = 0$, then by LLN we have

$$\hat{\beta} \to_p \beta.$$

Inconsistency of OLS Estimator

When $\mathbb{E}x_i u_i = \Delta \neq 0$, then

$$\hat{\beta} \to_p \beta + Q\Delta.$$

- Inconsistency occurs when x_i is correlated with u_i, or, x is "endogenous".
- $Q\Delta$ is called "asymptotic bias".

Relative Efficiency

- If $\hat{\theta}$ and $\tilde{\theta}$ are two unbiased estimators of θ, $\hat{\theta}$ is efficient relative to $\tilde{\theta}$ if $\mathrm{var}(\hat{\theta}) \leq \mathrm{var}(\tilde{\theta})$ for all θ, with strict inequality for at least one θ.
- Relative efficiency compares preciseness of estimators.
- Example: Suppose we want to estimate the population mean μ of an i.i.d. sample $\{x_i, i = 1, \ldots, n\}$. Both $\bar{x}$ and x_1 are unbiased estimators, however, $\bar{x}$ is more efficient since $\mathrm{var}\,(\bar{x}) = \frac{\mathrm{var}(x_1)}{n} \leq \mathrm{var}(x_1)$.
- If θ is a vector, we compare the covariance matrices of $\hat{\theta}$ and $\tilde{\theta}$ in the sense of positive definiteness.

Covariance Matrix of A Random Vector

- The variance of a scalar random variable x is

$$\mathrm{var}(x) = \mathbb{E}(x - \mathbb{E}x)^2.$$

- If x is a vector with two elements,

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

then the variance of x is a 2-by-2 matrix (we call "covariance matrix"):

$$\Sigma_x = \begin{pmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1, x_2) \\ \mathrm{cov}(x_2, x_1) & \mathrm{var}(x_2) \end{pmatrix},$$

where $\mathrm{cov}(x_1, x_2)$ is the covariance between x_1 and x_2:

$$\mathrm{cov}(x_1, x_2) = \mathbb{E}(x_1 - \mathbb{E}x_1)(x_2 - \mathbb{E}x_2).$$

Covariance Matrix of A Random Vector

- More generally, if x is a vector with n elements,

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix},$$

then the covariance matrix of x is a n-by-n matrix:

$$\Sigma_x = \begin{pmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1, x_2) & \cdots & \mathrm{cov}(x_1, x_n) \\ \mathrm{cov}(x_2, x_1) & \mathrm{var}(x_2) & \cdots & \mathrm{cov}(x_2, x_n) \\ \vdots & \vdots & & \vdots \\ \mathrm{cov}(x_n, x_1) & \mathrm{cov}(x_n, x_2) & \cdots & \mathrm{var}(x_n) \end{pmatrix}.$$

Covariance Matrix of A Random Vector

- The covariance matrix is the second moment of a random vector:
$$\Sigma_x = \mathbb{E}(x - \mathbb{E}x)(x - \mathbb{E}x)'.$$
- It is obvious that Σ_x is a symmetric matrix.

The Formula of Covariance Matrix

Given random vectors x and y, if

$$y = Ax,$$

where A is a matrix. Then

$$\Sigma_y = A\Sigma_x A'.$$

Covariance Matrix of the Residual

Write the original linear regression model as

$$y_i = x_i'\beta + u_i,$$

where

$$\begin{aligned} x_i &= (1, x_{i1}, ..., x_{ik})' \\ \beta &= (\beta_0, \beta_1, ..., \beta_k). \end{aligned}$$

- $\mathbb{E}u_i = 0 \Leftarrow$ Assumption (4) zero conditional mean
- $\mathbb{E}u_i^2 = \sigma^2 \Leftarrow$ Assumption (5) homoscedasticity
- $\mathbb{E}u_i u_j = 0$ for $i \neq j \Leftarrow$ Assumption (2) random sampling

What is the covariance matrix for $U = (u_1, u_2, ..., u_n)'$?

Covariance Matrix of the OLS Estimator

- The covariance matrix for $U = (u_1, u_2, ..., u_n)'$ is

$$\Sigma_u = \sigma^2 I,$$

  where I is the identity matrix.
- We have

$$\hat{\beta} = \beta + (X'X)^{-1}X'U.$$

- The covariance matrix of $\hat{\beta}$ is then

$$\Sigma_{\hat{\beta}} = \sigma^2 (X'X)^{-1}.$$

- The diagonal elements of $\Sigma_{\hat{\beta}}$ give the standard error of $\hat{\beta}$.
- If $\tilde{\beta}$ is another unbiased estimator of β with covariance matrix $\Sigma_{\tilde{\beta}}$, we say $\hat{\beta}$ is more efficient relative to $\tilde{\beta}$ if $\Sigma_{\tilde{\beta}} - \Sigma_{\hat{\beta}}$ is semi-positive definite for all β, with strict positive definiteness for at least one β.

Simple Regression

For a simple regression,

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

We can obtain

$$\Sigma = \frac{\sigma^2}{n\sum_i (x_i - \bar{x})^2} \begin{pmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{pmatrix}.$$

- The variance of $\hat{\beta}_1$ is

$$\mathrm{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}.$$

- The less σ^2, the more accurate $\hat{\beta}_1$ is.
- The more variation in x, the more accurate $\hat{\beta}_1$ is.
- And the more sample size, the more accurate $\hat{\beta}_1$ is.

Is OLS A Good Estimator?

Define what is "good":

- Is it unbiased?
- Is it consistent?
- Does it have a small variance?

Gauss-Markov Theorem

Theorem: Under Assumption 1-5, OLS is BLUE (Best Linear Unbiased Estimator).

- Define "best": smallest variance.
- Define "linear":
$$\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i. \tag{5}$$
- And unbiasedness: $\mathbb{E}\tilde{\beta} = \beta$.
- The message:
  We need not look for alternatives that are unbiased and are in the form of (5).

Time Series Regression Assumptions

(1) Linearity

$$y_t = \beta_0 + \beta_1 x_{1t} + \cdots \beta_k x_{kt} + u_t.$$

(2) (x_t, y_t) are jointly stationary and ergodic.

(3) No perfect collinearity.

(4) Past and contemporary exogeneity $\Leftrightarrow$

$$\mathbb{E}(u_t | x_t, x_{t-1}, ...) = 0.$$

Stationarity

- Weak stationarity.

$$\mathbb{E}x_t = \mu, \quad \mathrm{cov}(x_t, x_{t-\tau}) = \gamma_\tau, \quad \tau = \ldots, -2, -1, 0, 1, 2, \ldots.$$

- Strict stationarity.

$$F(X_t, ..., X_T) = F(X_{t+\tau}, ..., X_{T+\tau}),$$

where F is the joint distribution function.

Ergodicity

- An ergodic time series (x_t) has the property that x_t and x_{t-k} are independent if k is large.
- If (x_t) is stationary and ergodic, then a law of large number holds,

$$\frac{1}{n}\sum_{t=1}^{n} x_t \to \mathbb{E}x \quad a.s. \ .$$

Exogeneity in Time Series Context

- Strict exogeneity.

$$\mathbb{E}(u_t|X) = \mathbb{E}(u_t|..., x_{t+1}, x_t, x_{t-1}, ...) = 0.$$

- Past and Contemporary exogeneity.

$$\mathbb{E}(u_t|x_t, x_{t-1}, ...) = 0.$$

Consistency of OLS

Under the Time Series Regression Assumptions (1)-(4), the OLS estimator of the time series regression is consistent.

Special Cases

- Autoregressive models (AR),

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + u_t.$$

- Autoregressive distributed lag models (ARDL)

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \gamma_1 x_{t-1} + \cdots + \gamma_q x_{t-q} + u_t.$$

- Autoregressive models with exogenous variable (ARX)

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \gamma_1 x_t + \cdots + \gamma_q x_{t-q+1} + u_t,$$

where (x_t) is past and contemporary exogenous.

Beat OLS in Efficiency

- OLS is consistent, but is not efficient in general.
- u_t may be serially correlated and/or heteroscedastic. In such cases, GLS would be a better alternative.
- A simple way to account for serial correlation is to explicitly model u_t as an ARMA process:

$$y_t = x'\beta + u_t,$$

where $u_t \sim ARMA(p, q)$. But OLS is no longer able to estimate this model. Instead, nonlinear least square or MLE should be used.

Granger Causality

- Granger causality means that if x causes y, the x is a useful predictor of y_t.
- Granger Causality Test. In the model

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \gamma_1 x_{t-1} + \cdots + \gamma_q x_{t-q} + u_t.$$

  We test:

$$\mathrm{H}_0 : \gamma_1 = \cdots = \gamma_q = 0.$$

- The above test should be more appropriately called a non-causality test. Or even more precisely, a non-predicting test.
- Example: Monetary cause of inflation.

$$\pi_t = \beta_0 + \beta_1 \pi_{t-1} + \cdots + \beta_p \pi_{t-p} + \gamma_1 M1_{t-1} + \cdots + \gamma_q M1_{t-q} + u_t.$$