Takashi Yamano

Lecture Notes on Advanced Econometrics

Lecture 4: Multivariate Regression Model in Matrix Form

In this lecture, we rewrite the multiple regression model in the matrix form. A general multiple-regression model can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i \qquad \text{for } i = 1, \dots, n.$$

In matrix form, we can rewrite this model as

$$\begin{bmatrix} y_1 \\ y_2 \\ \\ y_n \end{bmatrix} = \begin{bmatrix} 1 \; x_{11} \;\; x_{12} \; ... \; x_{1k} \\ 1 \; x_{21} \;\; x_{22} \; ... \; x_{2k} \\ \\ 1 \; x_{n1} \;\; x_{n2} \; ... \; x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \\ u_n \end{bmatrix}$$

n x 1 n x (k+1) (k+1) x 1 n x 1

$$Y = X\beta + u$$

We want to estimate β.

Least Squared Residual Approach in Matrix Form

(Please see Lecture Note A1 for details)

The strategy in the least squared residual approach is the same as in the bivariate linear regression model. First, we calculate the sum of squared residuals and, second, find a set of estimators that minimize the sum. Thus, the minimizing problem of the sum of the squared residuals in matrix form is

$$\min \quad \underset{1 \times n \;\; n \times 1}{u'u} = (Y - X\beta)'(Y - X\beta)$$

Notice here that $u'u$ is a scalar or number (such as 10,000) because u' is a 1 x n matrix and u is a n x 1 matrix and the product of these two matrices is a 1 x 1 matrix (thus a scalar). Then, we can take the first derivative of this object function in matrix form. First, we simplify the matrices:

$$u'u = (Y' - \beta'X')(Y - X\beta)$$

$$= Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta$$

$$= Y'Y - 2\beta'X'Y + \beta'X'X\beta$$

Then, by taking the first derivative with respect to β, we have:

$$\frac{\partial(u'u)}{\partial\beta} = -2X'Y + 2X'X\beta$$

From the first order condition (F.O.C.), we have

$$-2X'Y + 2X'X\hat{\beta} = 0$$

$$X'X\hat{\beta} = X'Y$$

Notice that I have replaced β with $\hat{\beta}$ because $\hat{\beta}$ satisfy the F.O.C, by definition.

Multiply the inverse matrix of $(X'X)^{-1}$ on the both sides, and we have:

$$\boxed{\hat{\beta} = (X'X)^{-1}X'Y \qquad (1)}$$

This is the least squared estimator for the multivariate regression linear model in matrix form. We call it as **the Ordinary Least Squared (OLS)** estimator.

Note that the first order conditions (4-2) can be written in matrix form as

$$X'(Y - X\hat{\beta}) = 0$$

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \\ \\ x_{1k} & x_{2k} & \dots & x_{nk} \end{bmatrix} \left(\begin{bmatrix} y_1 \\ y_2 \\ \\ y_n \end{bmatrix} - \begin{bmatrix} 1 \; x_{11} \; \dots \; x_{1k} \\ 1 \; x_{21} \; \dots x_{2k} \\ \\ 1 \; x_{n1} \; \dots \; x_{nk} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \\ \hat{\beta}_k \end{bmatrix} \right) = 0$$

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \\ \\ x_{1k} & x_{2k} & \dots & x_{nk} \end{bmatrix} \left(\begin{bmatrix} y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_{11} \dots - \hat{\beta}_k x_{1k} \\ y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_{21} \dots - \hat{\beta}_k x_{2k} \\ \\ y_n - \hat{\beta}_0 - \hat{\beta}_1 x_{n1} \dots - \hat{\beta}_k x_{nk} \end{bmatrix} \right) = 0$$

(k+1) x n n x 1

This is the same as the first order conditions, k+1 conditions, we derived in the previous lecture note (on the simple regression model):

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - b_k x_{ik}) = 0$$

$$\sum_{i=1}^{n} x_{i1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - b_k x_{ik}) = 0$$

$$\sum_{i=1}^{n} x_{ik} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - b_k x_{ik}) = 0$$

***Example 4-1* : A bivariate linear regression (k=1) in matrix form**

As an example, let's consider a bivariate model in matrix form. A bivariate model is

$y_i = \beta_0 + \beta_1 x_{i1} + u_i$ for i = 1, …, n.

In matrix form, this is

$$Y = X\beta + u$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \\ y_n \end{bmatrix} = \begin{bmatrix} 1 \; x_1 \\ 1 \; x_2 \\ \\ 1 \; x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \\ u_n \end{bmatrix}$$

From (1), we have

$$\hat{\beta} = (X'X)^{-1} X'Y \tag{2}$$

Let's consider each component in (2).

$$X'X = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \end{bmatrix} \begin{bmatrix} 1 \; x_1 \\ 1 \; x_2 \\ \\ 1 \; x_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} 1 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

This is a 2 x 2 square matrix. Thus, the inverse matrix of $X'X$ is,

$$(X'X)^{-1} = \frac{1}{n \sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix}$$

$$= \frac{1}{n \, (\sum_{i=1}^{n} x_i - \bar{x})^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix}$$

The second term is

$$X'Y = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\, y_i \end{bmatrix} = \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^{n} x_i\, y_i \end{bmatrix}$$

Thus the OLS estimators are:

$$\hat{\beta} = (X'X)^{-1} X'Y = \frac{1}{n(\sum_{i=1}^{n} x_i - \bar{x})^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^{n} x_i\, y_i \end{bmatrix}$$

$$= \frac{1}{n(\sum_{i=1}^{n} x_i - \bar{x})^2} \begin{bmatrix} n\bar{y}\sum_{i=1}^{n} x_i^2 - n\bar{x}\sum_{i=1}^{n} x_i\, y_i \\ -n\bar{x}\, n\bar{y} + n\sum_{i=1}^{n} x_i\, y_i \end{bmatrix}$$

$$= \frac{1}{(\sum_{i=1}^{n} x_i - \bar{x})^2} \begin{bmatrix} \bar{y}\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i\, y_i \\ \sum_{i=1}^{n} x_i\, y_i - n\bar{x}\bar{y} \end{bmatrix}$$

$$= \frac{1}{(\sum_{i=1}^{n} x_i - \bar{x})^2} \begin{bmatrix} \bar{y}\sum_{i=1}^{n} x_i^2 - \bar{y}\,\bar{x}^2 + \bar{y}\,\bar{x}^2 - \bar{x}\sum_{i=1}^{n} x_i\, y_i \\ \sum_{i=1}^{n} (x_i\, y_i - \bar{x}\,\bar{y}) \end{bmatrix}$$

$$= \frac{1}{(\sum_{i=1}^{n} x_i - \bar{x})^2} \begin{bmatrix} \bar{y}(\sum_{i=1}^{n} x_i^2 - \bar{x}^2) - \bar{x}(\sum_{i=1}^{n} x_i y_i - \bar{y}\,\bar{x}) \\ \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \end{bmatrix}$$

$$= \begin{bmatrix} \bar{y} - \hat{\beta}_1\,\bar{x} \\ \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(\sum_{i=1}^{n} x_i - \bar{x})^2} \end{bmatrix}$$

$$= \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$$

This is what you studied in the previous lecture note.

End of Example 4-1

Unbiasedness of OLS

In this sub-section, we show the unbiasedness of OLS under the following assumptions.

Assumptions:

E 1 (Linear in parameters): $Y = X\beta + u$ **E 2** (Zero conditional mean): $E(u \mid X) = 0$ **E 3** (No perfect collinearity): **X has rank k.**

From (2), we know the OLS estimators are

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

We can replace **y** with the population model (**E 1**),

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

$$= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u$$
$$= \beta + (X'X)^{-1}X'u$$

By taking the expectation on the both sides of the equation, we have:

$$E(\hat{\beta}) \;\; = \beta + (X'X)^{-1}E(X'u)$$

From E2, we have $E(u \mid X) = 0$. Thus,

$$E(\hat{\beta}) \;\; = \beta$$

Under the assumptions E1-E3, the OLS estimators are unbiased.

The Variance of OLS Estimators

Next, we consider the variance of the estimators.

Assumption:

E 4 (Homoskedasticity): $Var(u_i \mid X) = \sigma^2$ and $Cov(u_i, u_j) = 0$, thus $Var(u \mid X) = \sigma^2 I$.

Because of this assumption, we have

$$\underset{}{E(uu')} = E\underset{n \times 1}{\begin{bmatrix} u_1 \\ u_2 \\ \\ u_n \end{bmatrix}}\underset{1 \times n}{\begin{bmatrix} u_1 & u_2 & & u_n \end{bmatrix}} = \underset{n \times n}{\begin{bmatrix} E(u_1u_1) & E(u_1u_2) & & E(u_1u_n) \\ E(u_2u_1) & E(u_2u_2) & & E(u_2u_n) \\ \\ E(u_nu_1) & E(u_nu_2) & & E(u_nu_n) \end{bmatrix}} = \underset{n \times n}{\begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ \\ 0 & 0 & \sigma^2 \end{bmatrix}} = \underset{n \times n}{\sigma^2 I}$$

Therefore,

$$\begin{aligned} Var(\hat{\beta}) &= Var[\beta + (X'X)^{-1} X'u] \\ &= Var[(X'X)^{-1} X'u] \\ &= E[(X'X)^{-1} X'uu'X(X'X)^{-1}] \\ &= (X'X)^{-1} X'E(uu')X(X'X)^{-1} \\ &= (X'X)^{-1} X'\sigma^2 I \; X(X'X)^{-1} \qquad \text{(E4: Homoskedasticity)} \end{aligned}$$

$$Var(\hat{\beta}) = \sigma^2 (X'X)^{-1} \qquad (3)$$

GAUSS-MARKOV Theorem:

Under assumptions 1 – 4, $\hat{\beta}$ is the Best Linear Unbiased Estimator (**BLUE**).

Example 4-2: Step by Step Regression Estimation by STATA

In this sub-section, I would like to show you how the matrix calculations we have studied are used in econometrics packages. Of course, in practices you do not create matrix programs: econometrics packages already have built-in programs.

The following are matrix calculations with STATA using data called, NFIncomeUganda.dta. Here we want to estimate the following model:

$$\ln(income) y_i = \beta_0 + \beta_1 female_i + \beta_2 edu_i + \beta_3 edusq_i + u_i$$

All the variables are defined in Example 3-1. Descriptive information about the variables are here:

```
. su;

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      female |       648    .2222222    .4160609          0          1
         edu |       648    6.476852    4.198633         -8         19
       edusq |       648    59.55093    63.28897          0        361
-------------+--------------------------------------------------------
   ln_income |       648    12.81736    1.505715   7.600903   16.88356
```

First, we need to define matrices. In STATA, you can load specific variables (data) into matrices. The command is called **mkmat**. Here we create a matrix, called ***y***, containing the dependent variable, *ln_nfincome*, and a set of independent variables, called **x**, containing *female*, *educ*, *educsq*.

```
. mkmat ln_nfincome, matrix(y)

. mkmat female educ educsq, matrix(x)
```

Then, we create some components: $X'X$, $(X'X)^{-1}$, and $X'Y$ **:**

```
. matrix xx=x'*x;
. mat list xx;
symmetric xx[4,4]
         female      edu    edusq    const
female      144
   edu      878    38589
 edusq     8408   407073  4889565
 const      144     4197    38589      648
```

```
. matrix ixx=syminv(xx);
. mat list ixx;
symmetric ixx[4,4]
            female          edu        edusq        const
female    .0090144
   edu   .00021374    .00053764
 edusq  -.00001238   -.00003259    2.361e-06
 const  -.00265043    -.0015892    .00007321    .00806547
```

Here is $X'Y$:

```
. matrix xy=x'*y;
. mat list xy;
xy[4,1]
        ln_nfincome
female    1775.6364
   edu    55413.766
 edusq    519507.74
 const    8305.6492
```

Therefore the OLS estimators are $(X'X)^{-1}X'Y$:

```
. ** Estimating b hat;
. matrix bhat=ixx*xy;
```

```
. mat list bhat;
bhat[4,1]
        ln_nfincome
female   -.59366458
   edu    .04428822
 edusq    .00688388
 const    12.252496

. ** Estimating standard error for b hat;
. matrix e=y-x*bhat;
. matrix ss=(e'*e)/(648-1-3);
. matrix kk=vecdiag(ixx);

. mat list ss;
symmetric ss[1,1]
             ln_nfincome
ln_nfincome    1.8356443

. mat list kk;
kk[1,4]
       female         edu       edusq      const
r1   .0090144  .00053764  2.361e-06  .00806547
```

Let's verify what we have found.

```
. reg ln_nfincome female edu edusq;

      Source |       SS       df       MS              Number of obs =     648
-------------+------------------------------           F(  3,   644) =   51.70
       Model |  284.709551     3  94.9031835           Prob > F      =  0.0000
    Residual |  1182.15494   644  1.83564431           R-squared     =  0.1941
-------------+------------------------------           Adj R-squared =  0.1903
       Total |  1466.86449   647   2.2671785           Root MSE      =  1.3549

------------------------------------------------------------------------------
 ln_nfincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
      female |  -.5936646   .1286361    -4.62   0.000    -.8462613   -.3410678
         edu |   .0442882   .0314153     1.41   0.159    -.0174005     .105977
       edusq |   .0068839   .0020818     3.31   0.001      .002796    .0109718
       _cons |    12.2525   .1216772   100.70   0.000     12.01356    12.49143
------------------------------------------------------------------------------
end of do-file
```

Lecture 5: OLS Inference under Finite-Sample Properties

So far, we have obtained OLS estimations for $E(\hat{\beta})\, and\, Var(\hat{\beta})$. But we need to know the shape of the full sampling distribution of $\hat{\beta}$ in order to conduct statistical tests, such as *t*-tests or F-tests. The distribution of OLS estimator $\hat{\beta}$ depends on the underlying distribution of the errors. Thus, we make the following assumption (again, under finite-sample properties).

Assumption

E 5 (Normality of Errors): $u_{n\times 1} \sim N(0_{n\times 1}, \sigma^2 I_{n\times n})$

Note that $N(0_{x\times 1}, \sigma^2 I_{n\times n})$ indicates a multivariate normal distribution of *u* with mean $\mathbf{0}_{n \times 1}$ and the variance-covariance matrix $\sigma^2 I_{n\times n}$.

Remember again that only assumptions E1-3 are necessary to have unbiased OLS estimators. In addition, assumption 4 is needed to show that the OLS estimators are the best linear unbiased estimator (BLUE), the Gauss-Markov theorem. We need assumption 5 to conduct statistical tests.

Assumptions E1-5 are collectively called as **the Classical Linear Model (CLM)** assumptions. The model with all assumptions E1-5 is called the classical linear model. The OLS estimators with the CLM assumptions are the minimum variance unbiased estimators. This indicates that the OLS estimators are the most efficient estimators among all models (not only among linear models).

Normality of $\hat{\beta}$

Under the CLM assumptions (E1-5), $\hat{\beta}$ (conditional on **X**) is distributed as multivariate normal with mean B and variance-covariance matrix $\sigma^2(X'X)^{-1}$.

$$\hat{\beta} \sim N[\beta, \sigma^2(X'X)^{-1}]$$

This is a multivariate normal distribution, which means each element of $\hat{\beta}$ is normally distributed:

$$\hat{\beta}_k \sim N[\beta_k, \sigma^2(X'X)^{-1}{}_{kk}]$$

$(X'X)^{-1}{}_{kk}$ is the k-th diagonal element of $(X'X)^{-1}$. Let's denote the k-th diagonal element of $(X'X)^{-1}$ as S_{kk}. Then,

$$\sigma^2(X'X)^{-1} = \sigma^2 \begin{pmatrix} S_{11} & . & . & . & . \\ . & S_{22} & . & . & . \\ & & & & \\ . & . & . & . & S_{kk} \end{pmatrix} = \begin{pmatrix} \sigma^2 S_{11} & . & . & . & . \\ . & \sigma^2 S_{22} & . & . & . \\ & & & & \\ . & . & . & . & \sigma^2 S_{kk} \end{pmatrix}$$

This is the variance-covariance matrix of the OLS estimator. On the diagonal, there are variances of the OLS estimators. Off-the diagonal, there are covariance between the estimators. Because each OLS estimator is assumed to be normally distributed, we can obtain a standard normal distribution of an OSL estimator by subtracting the mean and dividing it by the standard deviation:

$$z_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 S_{kk}}}.$$

However, σ^2 is unknown. Thus we use an estimator of σ^2 instead. An unbiased estimator of σ^2 is

$$s^2 = \frac{\hat{u}'\hat{u}}{n-(k+1)}$$

$\hat{u}'\hat{u}$ is the sum of squared errors. (Remember $\hat{u}'\hat{u}$ is a product of a (1 x n) matrix and a (n x 1) matrix, which gives a single number.) Therefore by replacing σ^2 with s^2, we have

$$t_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{s^2 S_{kk}}} \quad .$$

This ratio has a *t*-distribution with (n-k-1) degree of freedom. It has a *t*-distribution because it is a ratio of a variable that has a standard normal distribution (the nominator in the parenthesis) and a variable that has a chi-squared distribution divided by (n-k-1).

The standard error of $\hat{\beta}_k$, se($\hat{\beta}_k$), is $\sqrt{s^2 S_{kk}}$.

Testing a Hypothesis on $\hat{\beta}_k$

In most cases we want to test the null hypothesis

H_0: $\beta_k = 0$

with the t-statistics

t-test: $(\hat{\beta}_k - 0)/\,\mathrm{se}(\hat{\beta}_k) \sim t_{\text{n-k-1}}$.

When we test the null hypothesis, the t-statistics is just a ratio of an OLS estimator over its standard error.

We may test the null hypothesis against the one-sided alternative or two-sided alternatives.

Testing a Joint Hypotheses Test on $\hat{\beta}_k's$

Suppose we have a multivariate model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + u_i$$

Sometimes we want to test to see whether a group of variables jointly has effects of y. Suppose we want to know whether independent variables x_3, x_4, and x_5 jointly have effects on y.

Thus the null hypothesis is

$$H_0: \beta_3 = \beta_4 = \beta_5 = 0.$$

The null hypothesis, therefore, poses a question whether these three variables can be excluded from the model. Thus the hypothesis is also called exclusion restrictions. A model with the exclusion is called **the restricted model**:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$$

On the other hand, the model without the exclusion is called **the unrestricted model**:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + u_i$$

We can generalize this problem by changing the number of restrictions from three to q. The joint significance of q variables is measured by how much the sum of squared residuals (SSR) increases when the q-variables are excluded. Let denote the SSR of the restricted and unrestricted models as SSR_r and SSR_{ur}, respectively. Of course the SSR_{ur} is smaller than the SSR_r because the unrestricted model has more variables than the restricted model. But the question is how much compared with the original size of SSR. The F-statistics is defined as

F-test: $$F \equiv \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}.$$

The numerator measures the change in SSR, moving from unrestricted model to restricted model, per one restriction. Like percentage, the change in SSR is divided by the size of SSR at the starting point, the SSR_{ur} standardized by the degree of freedom.

The above definition is based on how much the models cannot explain, SSR's. Instead, we can measure the contribution of a set of variables by asking how much of the explanatory power is lost by excluding a set of q variables.

The F-statistics can be re-defined as

F-test: $$F \equiv \frac{(R_{ur}^2 - R_r^2)/q}{(1-R_{ur}^2)/(n-k-1)}.$$

Again, because the unrestricted model has more variables, it has a larger R-squared than the restricted model. (Thus the numerator is always positive.) The numerator measures the loss in the explanatory power, per one restriction, when moving from the unrestricted model to the restricted model. This change is divided by the unexplained variation in y by the unrestricted model, standardized by the degree of freedom.

If the decrease in explanatory power is relatively large, then the set of q-variables is considered a jointly significant in the model. (Thus these q-variables should stay in the model.)

Lecture 6: OLS Asymptotic Properties

Consistency (instead of unbiasedness)

First, we need to define consistency. Suppose W_n is an estimator of θ on a sample of Y_1, Y_2, …, Y_n of size n. Then, W_n is a consistent estimator of θ if for every $e > 0$,

$$P(|W_n - \theta| > e) \to 0 \text{ as } n \to \infty.$$

This says that the probability that the absolute difference between W_n and θ being larger than e goes to zero as n gets bigger. Which means that this probability could be non-zero while n is not large. For instance, let's say that we are interested in finding the average income of American people and take small samples randomly. Let's assume that the small samples include Bill Gates by chance. The sample mean income is way over the population average. Thus, when sample sizes are small, the probability that the difference between the sample and population averages is larger than e, which is any positive number, can be non-zero. However, the difference between the sample and population averages would be smaller as the sample size gets bigger (as long as the sampling is properly done). As a result, as the sample size goes to infinity, the probability that the difference between the two averages is bigger than e (no matter how small e is) becomes zero.

In other words, we say that θ is the probability limit of W_n:

$$plim\,(W_n) = \theta.$$

Under the finite-sample properties, we say that W_n is <u>unbiased</u>, $E(W_n) = \theta$. Under the asymptotic properties, we say that W_n is <u>consistent</u> because W_n converges to θ as n gets larger.

The OLS estimators

From previous lectures, we know the OLS estimators can be written as

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

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

In the matrix form, we can examine the probability limit of OLS

$$p\lim \hat{\beta} \;=\; \beta \;+\left(\frac{1}{n}X'X\right)^{-1} p\lim\left(\frac{1}{n}X'u\right)$$

Here, we assume that

$$p\lim\frac{1}{n}X'X \;=\; Q\,.$$

This assumption is not a difficult one to make since the law of large numbers suggests that the each component of $\frac{1}{n}X'X$ goes to the mean values of $X'X$. And also we assume that Q^{-1} exists. From E2, we have

$$p\lim\left(\frac{1}{n}X'u\right) = 0\;.$$

Thus,

$$p\lim \hat{\beta} = \beta$$

Thus, we have shown that the OLS estimator is consistent.

Nest, we focus on the asymmetric inference of the OLS estimator. To obtain the asymptotic distribution of the OLS estimator, we first derive the limit distribution of the OLS estimators by multiplying $\sqrt{n}$ (note: we multiply $\sqrt{n}$ (scaling) on $\hat{\beta} - \beta$ to obtain non-zero yet finite variance asymptotically; see Cameron and Trivedi; also $\sqrt{n}$ will be squared later and becomes *n*, very convenient) on the OLS estimators:

$$\hat{\beta} \;=\; \beta + \left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}X'u\right)$$

$$\sqrt{n}\,(\hat{\beta} - \beta) = \left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{\sqrt{n}}X'u\right)$$

The probability limit of $\sqrt{n}\,(\hat{\beta}-\beta)$ goes to zero because of the consistency of $\hat{\beta}$. The limit variance of $\sqrt{n}\,(\hat{\beta}-\beta)$ is

$$\sqrt{n}\,(\hat{\beta}-\beta)\cdot\sqrt{n}(\hat{\beta}-\beta)' = \left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{\sqrt{n}}X'u\right)\left(\frac{1}{\sqrt{n}}X'u\right)'\left(\frac{1}{n}X'X\right)^{-1}$$

$$= \left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}X'uu'X\right)\left(\frac{1}{n}X'X\right)^{-1}$$

From E4, the probability limit of uu' goes to $\sigma^2 I$, and we assumed *plim* of $\frac{1}{n}X'X$ is Q. Thus,

$$= Q^{-1}\left(\frac{\sigma^2}{n}X'X\right)Q^{-1}$$

$$= \sigma^2 Q^{-1} Q Q^{-1}$$
$$= \sigma^2 Q^{-1}$$

Therefore, the limit distribution of the OLS estimator is

$$\sqrt{n}\,(\hat{\mathrm{B}}-\mathrm{B}) \sim^d N[0,\, \sigma^2\, Q^{-1}].$$

From this, we can obtain the asymptotically distribution of the OLS estimator by multiplying $\sqrt{n}$ and manipulating:

$$\hat{\mathrm{B}} \sim^a N[\mathrm{B},\, \sigma^2\, N^{-1} Q^{-1}].$$

***Example 6-1*: Consistency of OLS Estimators in Bivariate Linear Estimation**

A bivariate model: $y_i = \beta_0 + \beta_1 x_{i1} + u_i$ and $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

To examine the biasedness of the OLS estimator, we take the expectation

$$E(\hat{\beta}_1) = \beta_1 + E\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})\ u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

Under the assumption of zero conditional mean (SLR 3: E(u|x) = 0), we can separate the expectation of x and u:

$$E(\hat{\beta}_1) = \beta_1 + \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})\ E(u_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

Thus we need the SLR 3 to show the OLS estimator is unbiased.

Now, suppose we have a violation of SLR 3 and cannot show the unbiasedness of the OLS estimator. We consider a consistency of the OLS estimator.

$$p\lim\hat{\beta}_1 = p\lim\beta_1 + p\lim\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})\ u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

$$p\lim\hat{\beta}_1 = \beta_1 + \frac{p\lim\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})u_i\right]}{p\lim\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]}$$

$$p\lim\hat{\beta}_1 = \beta_1 + \frac{\operatorname{cov}(x,u)}{\operatorname{var}(x)}$$

$$p\lim\hat{\beta}_1 = \beta_1 \quad \textit{if} \quad \operatorname{cov}(x,u) = 0$$

Thus, as long as the covariance between x and u is zero, the OLS estimator of a bivariate model consistent.

End of Example 6-1

Lecture 7: OLS Further Issues

In this lecture, we will discuss some practical issues related to OLS estimations, such as functional forms and interpretations of several types of variables. For details, please read Wooldridge chapter 6 and 7.

Measurement Error in the Dependent Variable

Let y^* denote the variable that we would like to explain:

$$y^*_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + u_i$$

However, we can only observe y which is a measured variable of y^* with measurement errors.

$$e_0 = y - y^*$$

By replacing y^* with y and e_0, we get

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + e_0 + u_i$$

Thus, if $e_0 + u$ satisfy the OLS assumptions (such as $E(e_0 + u|X)=0$), then OLS estimators are unbiased (or consistent). But the variance of the disturbance is larger by $Var(e_0)$ with the measurement error (e_0) than without.

Note, however, that the measurement error in the dependent variable could be correlated with independent variables [Cov (x_k e_0)$\neq$0]. In that case, the estimators will be biased.

Measurement Error in an Independent Variable

Let x_k^* denote an independent variable, which could be observed in x_k with the measurement error, e_k, where $E(e_k)=0$.

$$e_k = x_k - x^*_k \tag{8-1}$$

$$y^*_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k (x_{ki} - e_k) + u_i$$

Assumption 1: Cov (x_k, e_k) = 0

Under this assumption, the error term (u - $\beta_k e_k$) has zero mean and uncorrelated with the independent variables. Thus the estimators are unbiased (consistent). The error variance, however, is bigger by ($\beta_k e_k$)^2.

Assumption 2: Cov (x^*_k, e_k) = 0

This assumption is called the Classic Errors-in-Variables (CEV) assumption. Because e_k = x_k - x^*_k, $\underline{x_k \text{ and } e_k \text{ must be correlated}}$ under the assumption 2:

$$\text{Cov}(x_k, e_k) = E(x_k e_k) = E(x^*_k e_k) + E(e^2_k) = \sigma^2_{ek}$$

Thus, we have the omitted variables problem, which gives inconsistent estimators of all independent variables.

The Attenuation Bias

For **a bivariate regression model** it is easy to show the exact bias cased by the CEV. Now, x_1 is measured with the measurement errors, instead of x_k. In a bivariate regression model, the least square estimator can be written as

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

$$= \beta_1 + \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(u_i - \beta_1 e_1)}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}$$

Thus, the probability limit of $\hat{\beta}_1$ is

$$
\begin{aligned}
p\lim(\hat{\beta}_1) &= \beta_1 + \frac{\operatorname{cov}(x_1, u - \beta_1 e_1)}{Var(x_1)} \\
&= \beta_1 + \frac{(-\beta_1 \sigma_{e1}^2)}{Var(x_1^* + e_1)} \\
&= \beta_1 \left(1 - \frac{\sigma_{e1}^2}{\sigma_{x_1^*}^2 + \sigma_{e1}^2} \right) \\
&= \beta_1 \left(\frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e1}^2} \right) < \beta_1
\end{aligned}
$$

Thus, the plim($\hat{\beta}_1$) is always closer to zero (or biased toward zero) than β_1. This is called the (famous) **attenuation bias** in OLS due to classical errors-in-variables.

For a multivariate regression model, the probability limit of $\hat{\beta}_1$ is

$$
p\lim(\hat{\beta}_1) = \beta_1 \left(\frac{\sigma_{r_1^*}^2}{\sigma_{r_1^*}^2 + \sigma_{e1}^2} \right) < \beta_1
$$

$$
\textit{where } r_1^* \textit{ is the population error in the equation}
$$

$$
x_1^* = \alpha_0 + \alpha_1 x_2 + \ldots + \alpha_{k-1} x_k + r_1^*
$$

Again the implication is the same as before. The estimated coefficient of the variable with measurement errors is biased toward zero (or less likely to reject the null hypothesis).

Data Scaling

Many variables we use have units, such as monetary units and quantity units. The bottom line is that data scaling does not change much.

Scaling up/down a dependent variable:

$$
\alpha y = (\alpha\hat{\beta}_0) + (\alpha\hat{\beta}_1)x_1 + (\alpha\hat{\beta}_2)x_2 + \ldots + (\alpha\hat{\beta}_k)x_k \;.
$$

If you scale up/down the dependent variable by ∀, the OLS estimators and standard errors will be also scaled up/down by ∀, but not *t*-statistics. Thus the significance level remains the same as before scaling.

Scaling up/down a independent variable:

$$y = \hat{\beta}_0 + (\hat{\beta}_1 / \alpha)(\alpha x_1) + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k \ .$$

If you scale up/down one independent variable, the estimated coefficient of the independent variable will be scale down/up by the same scale. Again the *t*-statistics (or significance level) does not change.

Logarithmic Forms

For a small change in *x*, a change in log(*x*) times 100, 100)log(*x*), is approximately close to a percentage change in *x*, $\Delta x / \bar{x}$. Therefore, we can interpret the following cases using percentage changes:

(1) log-log: $\log(y) = \beta_k \log(x_k) + \ldots$ $\qquad \hat{\beta}_k = \dfrac{\Delta y / \bar{y}}{\Delta x / \bar{x}}$

One percent change in x_k changes *y* by (100 $\hat{\beta}_k$) percent. $\hat{\beta}_k$ is an elasticity.

(2) log-level: $\log(y) = \beta_k x_k + \ldots$ $\qquad \hat{\beta}_k = \dfrac{\Delta y / \bar{y}}{\Delta x}$

One unit change in x_k changes *y* by (100 $\hat{\beta}_k$) percent.

(3) level-log: $y = \beta_k \log(x_k) + \ldots$ $\qquad \hat{\beta}_k = \dfrac{\Delta y}{\Delta x / \bar{x}}$

One percent change in x_k changes *y* by (100 $\hat{\beta}_k$).

(4) level-level: $y = \beta_k x_k + \ldots \qquad \hat{\beta}_k = \dfrac{\Delta y}{\Delta x}$

One unit change in x_k changes y by $\hat{\beta}_k$.

When a change in log is not small, the approximation between a change in log(x) and a change x may not be accurate. For instance, the log-level model gives us

$$\widehat{\log}(y') - \widehat{\log}(y) = \hat{\beta}_k . \qquad (8\text{-}1)$$

If the change in log is small, then there is no problem of interpreting this as "one unit of x_k changes y by (100 $\hat{\beta}_k$) percent," because $\widehat{\log}(y') - \widehat{\log}(y) \approx (y' - y)/y$. But when a change in log is not small, the approximation may not be approximate. Thus we need to transform (8-1) as:

$$\widehat{\log}(y') - \widehat{\log}(y) = \log(\widehat{y'/y}) = \hat{\beta}_k$$

$$(\widehat{y'/y}) = \exp(\hat{\beta}_k)$$

$$(\widehat{y'/y}) - 1 = \exp(\hat{\beta}_k) - 1$$

$$(\widehat{y' - y})/y = \exp(\hat{\beta}_k) - 1$$

$$\%\widehat{\Delta y} = 100[\exp(\hat{\beta}_k) - 1]$$

Thus one unit in x_k changes y by $100[\exp(\hat{\beta}_k) - 1]$ percentage.

Quadratic Form

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2 + \ldots + \hat{\beta}_k x_k$$

$$\partial y / \partial x_1 = \hat{\beta}_1 + 2\hat{\beta}_2 x_1$$

Interpretation:

$\hat{\beta}_1 > 0$ *and* $\beta_2 < 0$ "an increase in *x* increases *y* with a diminishing rate "

$\hat{\beta}_1 < 0$ *and* $\beta_2 > 0$ "an increase in *x* decreases *y* with a diminishing rate "

Turning Point:

At the turning point, the first derivative of y with respect to x_1 is zero:

$$\partial y / \partial x_1 = \hat{\beta}_1 + 2\hat{\beta}_2 x_1 = 0$$

Thus the value of x_1 at the turning this point is

$$x_1^* = -\hat{\beta}_1 / (2\hat{\beta}_2)$$

Interaction Terms

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 (x_1 x_2) + \ldots + \hat{\beta}_k x_k$$

The impact of x_1 on y is

$$\partial y / \partial x_1 = \hat{\beta}_1 + \hat{\beta}_3 x_2$$

A Dummy Variable

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k \qquad where\ x_1 = 0\ or\ 1$$

A group of observations with x_1=0 is called **a base, benchmark, or reference group**. The estimated coefficient of x_1 measures the difference in averages in $\hat{y}$ among observations for x_1=0 and x_1=1 (the base group), holding other variables constant.

$$\hat{\beta}_1 = E(y \mid x_1 = 0, x_2, \ldots, x_k) - E(y \mid x_1 = 1, x_2, \ldots, x_k)$$

"The group B, with x_1=1, has a lower or higher y than the base group, with x_1=0."

Interaction Terms with Dummies

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2 + \ ... \ + \hat{\beta}_k x_k \qquad where\ x_1 = 0\ or\ 1$$

When x_2 is a continuous variable, $\hat{\beta}_3$ measures a difference in the effect of x_2 on y between a group with x_1=0 and a group with x_1=1, or a difference in slopes of x_2:

$$\partial y / \partial x_2 = \hat{\beta}_2 \qquad when\, x_1 = 0$$
$$\partial y / \partial x_2 = \hat{\beta}_2 + \hat{\beta}_3 \qquad when\, x_1 = 1$$

Multicollinearity

From the previous lecture, we know that the variance of OLS estimators is $Var(\hat{\mathrm{B}}) = \sigma^2 (X'X)^{-1}$. The variance of an estimator, $\hat{\beta}_k$, is $\sigma^2 S_{kk}$. This can be written as

$$Var(\hat{\beta}_k) = \frac{\sigma^2}{(1 - R_k^2)\sum_{i-1}^{n}(x_{ik} - \bar{x}_k)}$$

(See Wooldridge pp94 or Greene pp57) R_k^2 is the R-squared in the regression of x_k against all other variables. In other words, R_k^2 is the proportion of the total variation in x_k that can be explained by the other independent variables.

If two variables are **perfectly correlated**, then R_k^2 will be one for both of the perfectly correlated variables, and the variance of those two variables will not be measured. Obviously, you need to drop one of the two perfectly correlated variables. In STATA, STATA drops one of perfectly correlated variables automatically. So if you see a dropped variable in STATA outputs, you should suspect that you have included perfectly correlated variables without realizing.

Even if two or more variables are not perfectly correlated, if they are highly correlated (high R_k^2), the variance of estimators will be large. This problem is called **multicollinearity**.

The problem of multicollinearity can be avoided to some extent by collecting more data and increase variance in independent variables. For instance, let's think about a sample of four individuals. All of four could be male and unmarried. In this case, a gender variable and a marital status variable will be perfectly correlated. Suppose three of them have collage education. Then an education variable will be highly correlated with the gender and marital-status variables. Of course, correlations between these variables will disappear (to some extent) as the size of sample and the variation in variables increase.

When in doubt, conduct a F-test on variables that you suspect causing multicollinearity.
A typical symptom of multicollinearity is

- A high joint significance and low individual significance
  (a high F-statistics but low t-statistics)

A simple solution is to keep them in the model. If your main focus is on a variable which is not the part of multicollinearity, then it is not a serious problem to have multicollinearity in your model. You could drop one of highly correlated variables, but by doing so may create an omitted variable problem. Remember that an omitted variable problem can cause biases in on all of estimators. This could be a more serious problem than multicollinearity.