CHAPTER 1

Introduction

1.1. What is econometrics?

Econometrics is a subject which study measurement problems in economics. Econometrics is the only systematic way we know now to examine the reality of the working of an economy. Science have made progress when a better way to examine reality is discovered. Telescope, radio telescope, microscope, electron microscope, X-ray, MRI, fMRI are examples. I believe the same can be said about econometrics.

You will see the usefulness of econometrics throughout the course but at the same time you will also come to understand that there are many limitations in the methods. I hope those limitations do not discourage you. Rather, I hope you will take the limitations as challenges that are inviting your contribution.

1.2. Parameters of interest and its relation to a probability model parameter

Throughout the class we refer to an object of measurement as a *parameter*. Suppose you are measuring the income distribution. Then the parameter is a distribution. Suppose you are interested in knowing the shape of the business cycle or the demand function of gasoline. Then the parameter is a function.

You as an economist determine an economic phenomenon you want to study. That should define what parameter you are interested in measuring. Once a parameter of interest is defined, econometric methods should inform you how to measure it. Typically, the parameter of interest is modeled as a part of a probability model. A systematic knowledge that aids in this measurement process is called econometrics. This course is an introduction to this subject.

1.3. Applications

Let's think about some examples of parameters we may be interested in studying. A private firm may be involved in a sex discrimination suit. Lawyers working on both sides want to know if the firm did or did not systematically paid less to a group of women in the firm. The parameter of interest in this case may be the salary difference between comparable men and women in this firm. In order to proceed with this idea, we need to formalize what we mean by "comparable men and women" more precisely and how we summarize the salary of the comparable men and women.

In order to levy property tax we need to know what the property value is for a given property. In this case the parameter of interest is the value of the property given the characteristics of the property.

In both cases, one could use the conditional mean function or the conditional median function as the parameters of interest.

1.4. Five ways economic theory and econometrics interacts

Economic theory and econometrics interact in at least five ways. First, economic theories suggests various parameters of interest. For example, a concept of demand function is created in economic theory.

Second, economic theories provide frameworks within which one can conduct measurement. In the context of demand function estimation, it is important to model how the price is determined. Economic theory provides the framework of price determination.

Third, economic theories help restrict the kind of values a parameter of interest can take and thereby help measure the parameter or they give us a set of relevant variables we should take into account. Continuing on the demand function example, economic theory tells us that a demand function is likely to slope downwards with respect to its own prices provided the income effect is not large and that the relevant variables are prices of related goods as well as its own price and variables that affect marginal utility for those goods.

Fourth, in turn, empirical results substantiate theoretical constructs. Deductive reasoning alone will not inform us the elasticity of a demand function, for example. Nor do we know, purely from deductive reasoning, how lengthening the copyright protection from 14 years to 70 years encourage creative activities by individuals. From a theoretical point of view one can guess that it is non-negative. Given that the cost of copyright protection is quite sizable, the main issue is not the direction of the effect, but the size of the effect; how much additional creative activities we expect to see when the copyright protection is extended by a proposed period.

Fifth, empirical findings may help eliminate certain type of economic models as inconsistent with observations. For example, theoretically it is possible to reduce the tax rate and raise the tax revenue as suggested some time ago. Many, based on empirical results argued that that is not plausible. In this case, empirical results did not win the debate and as a result during the 1980's the federal debt more than doubled as a percentage of GDP (from 26.8% to 44.1%).

Fifth, empirical findings may lead to new theoretical models. For a long time during 1960's many economists believed that there was a trade-off between inflation rate and the unemployment rate in an economy. After we observed the shift in the trade-off some recognized the importance of the role individual's expectation play in the trade-off and led to explicitly modeling forms of expectation in macro economic models.

Theory alone does not tell us anything about reality and we stressed the need to substantiate economic theories by empirical work. In addition to the roles of theory as suggesting some parameters of interest and an aid for measurement it is important to recognize that theory provides "explanation" of a phenomenon under study.

For example we observe higher incidences of lung cancers among smokers than among non-smokers. This is the best empirical study can provide without any theory. Although implausible, logically it is possible that people with lung cancer gene tend to like smoking so it may be the lung cancer gene that causes smoking. The empirical evidence alone cannot distinguish the two. If someone comes up with a mechanism under which smoking raises the incidence of lung cancer, then the theory may suggest some ways to distinguish the two hypotheses. In this way theories may provide a framework of "explaining" the phenomenon under study,

in this case, of smoking and lung cancer relationship. Our explanation is as good as what the current theory is. In turn the theory needs to be substantiated by empirical evidences. We need both.

In conducting empirical studies, for which this course provides tools, it is important to remember this limitation of what theories and empirical studies can each accomplish. Empirical studies in themselves do *not* provide explanation. It is also important to take advantage of whatever information theories provide in conducting the measurements.

1.5. Exercises

(1) Think about what kind of economic phenomenon you want to study and explain what the main parameter of interest is. Is it a number, a function, or a stochastic process, or something else?

CHAPTER 2

Conditional Probability Models

2.1. Probability model

In conducting an econometric measurement of a parameter of interest, like any other measurement problems, probability model is used. Thus we need to relate the parameter of interest with a particular parameter in a probability model. This is the uniquely econometric issue which we need to think carefully about.

Recall what statistics study. The parameters studied in statistics are all related to some aspects of a probability model. There is no parameter of interest outside a probability model. The uniquely econometrics problem is to formulate a parameter of interest that arises in economics to embed it in a probability model.

Since we typically do not address a completely new measurement issue, typically there is a standard framework literature uses for each of the economic measurement problem. We should make sure to think carefully whether the framework used is appropriate.

The parameter of interest which frequently arises in economics is how one variable affects another variable. As we saw earlier, the demand function relates its own price to its demand given other prices. Another example is a relationship between wage and factors affecting the human capital, such as education and experience given age and gender, industry, and occupation.

As these examples suggest, most cases in economics where we examine relationship between two variables require holding some other variables at some given values. However, if a randomized experiment is possible, we may not need to hold other variables at some given values, as we will see later in the course. Non-experimental data are usually referred to as observational data and distinguished from experimental data. Thus one may say that the requirement for holding other variables at some given values often arises because we often need to analyze observational data. More about this point later.

These relationships are studied using the concept of conditional distribution. Often, rather than studying the full conditional distribution, the conditional mean function is used to study the relationship. However, the conditional median function can be used, or more generally the conditional quantile function for different quantiles can be used. By doing so, a more complete understanding about the conditional distribution can be obtained.

2.2. Properties of conditional mean function

By far the most frequently used approach to studying a relationship between variables as a parameter of interest is the approach using the concept of conditional mean function.

There are at least four reasons why the conditional mean function is a useful way to study a relationship among variables. First, as we shall see, the inferences can be carried out relatively easily. Second, as we shall see in this section, the conditional mean function provides the best predictor and oftentimes, prediction is the purpose of study. Third, as we shall see, the classical measurement error in Y does not cause too much difficulty. Fourth, if we want to examine causal relationship, we need to focus on expectation as we shall see in this section.

As discussed earlier, however, it should be recognized that the conditional mean function is only one aspect of the relationship between two or more random variables.

Below we will discuss various properties of conditional mean function. We denote the two random variables we focus on by Y and X_1 and denote the rest of random variables we take as given by X_2, which is a vector. We denote $X' = (X_1, X_2')$, where X' denote the transpose of X. Let $m(x) = E(Y|X = x)$. We write $m(X)$ as $E(Y|X)$. Under this notation, we can show that

(1) $E(g(X)|X) = g(X)$.
(2) $E(g(X)Y|X) = g(X)E(Y|X)$.

To see the first relationship,

$$E(g(X)|X = x) = E(g(x)|X = x) = g(x)E(1|X = x) = g(x)$$

so that $E(g(X)|X) = g(X)$.

To see the second relationship,

$$E(g(X)Y|X = x) = E(g(x)Y|X = x)$$
$$= g(x)E(Y|X = x),$$

so that $E(g(X)Y|X) = g(X)E(Y|X)$.

Here we list useful relationships we often use.

(1) Law of Iterated Expectations: $E(Y) = E[E(Y|W)]$.
(2) Its conditional version: $E(Y|X) = E[E(Y|X, Z)|X]$.
(3) If Y and X_1 are independent given X_2, then $E(Y|X_1 = x_1, X_2 = x_2) = E(Y|X_2 = x_2)$.
(4) If $U \stackrel{def}{=} Y - E(Y|X)$, then $E(U|X) = 0$ so that for any function $g(\cdot)$ for which $E(|g(X)U|) < \infty$, $E(g(X)U) = 0$. In particular, $E(U) = 0$ and $Cov(g(X), U) = 0$.
(5) If $c : R \to R$ is a convex function defined on R and $E(|X|) < \infty$. Then $c(E(X|Z)) \leq E(c(X)|Z)$.
(6) $V(Y) = E(V(Y|X))+V(E(Y|X))$ where $V(Y|X) = E[(Y-E(Y|X))^2|X]$.
(7) The conditional version is

$$V(Y|X) = E(V(Y|X, Z)|X) + V(E(Y|X, Z)|X)$$

(8) When $E(Y^2) < \infty$, conditional mean function $m(X) = E(Y|X)$ can be characterized as a solution the following minimization problem

$$\min_{g(\cdot)} E[(Y - g(X))^2].$$

To see 3,

$$\begin{aligned}
&\int yf(y|X_1 = x_1, X_2 = x_2)dy \\
&= \int yf_{Y,X_1,X_2}(y, x_1, x_2)/f_{X_1,X_2}(x_1, x_2)dy \\
&= \int yf_{Y,X_1|X_2}(y, x_1|x_2)f_{X_2}(x_2)/f_{X_1,X_2}(x_1, x_2)dy \\
&= \int yf_{Y|X_2}(y|x_2)f_{X_1|X_2}(x_1|x_2)f_{X_2}(x_2)/f_{X_1,X_2}(x_1, x_2)dy \\
&= \int yf_{Y|X_2}(y|x_2)dy.
\end{aligned}$$

The last equality follows because $f_{X_1|X_2}(x_1|x_2)f_{X_2}(x_2) = f_{X_1,X_2}(x_1, x_2)$.

To see 4, note that $E(U|X) = E[Y - E(Y|X)|X] = E(Y|X) - E(Y|X) = 0$. Thus

$$\begin{aligned}
E[g(X)U] &= E\{E[g(X)U|X]\} \\
&= E\{g(X)E(U|X)\} \\
&= 0.
\end{aligned}$$

Claim 5 is the conditional version of the so called Jensen's inequality. To see this, note that when c is a convex function, one can find a linear function with slope a, $a(x - E(X|Z = z)) + c(E(X|Z = z))$, that touches the convex function $c(\cdot)$ at $(E(X|Z = z), c(E(X|Z = z)))$, but lies entirely below the convex function. Since $c(x) \geq a(x - E(X|Z = z)) + c(E(X|Z = z))$, $c(X) \geq a(X - E(X|Z = z)) + c(E(X|Z = z))$ so that

$$\begin{aligned}
E(c(X)|Z = z) &\geq a(E(X|Z = z) - E(X|Z = z)) + c(E(X|Z = z)) \\
&= c(E(X|Z = z)).
\end{aligned}$$

Thus $E(c(X)|Z) \geq c(E(X|Z))$.
This implies $E(Y^2) \geq E(Y)^2$ and $E(Y^2|X) \geq E(Y|X)^2$.

Claim 6 can be obtained using the properties of the conditional expectation operator:

$$\begin{aligned}
V(Y) &= E[(Y - E(Y))^2] \\
&= E\{[(Y - E(Y|X)) + (E(Y|X) - E(Y))]^2\} \\
&= E\{[(Y - E(Y|X))^2]\} + E\{[E(Y|X) - E(Y)]^2\} \\
&\quad 2E\{(Y - E(Y|X))(E(Y|X) - E(Y))\}.
\end{aligned}$$

The first term of the last three terms equals $E\left[E\{[(Y - E(Y|X))^2]|X\}\right]$, which equals $E[V(Y|X)]$. Since $E(Y) = E[E(Y|X)]$, the second term of the last three terms equals $V[E(Y|X)]$. Since the third of the last three terms equals two times $E\{U[E(Y|X) - E(Y)]\}$, where $U = Y - E(Y|X)$, Claim 4 implies that it equals zero. This completes the proof.

Note that $V(U|X) = E(U^2|X) = E[(Y - E(Y|X))^2|X] = V(Y|X)$.

Also, note that $V(U) = E(U^2) = E[V(Y|X)]$. The last equality follows from the iterated expectation and the result just above.

Claim 7 can be shown in the same way as for Claim 6, so I leave it as an exercise. It implies $E[V(Y|X)] \geq E\{E[V(Y|X, Z)|X]\}$ or via the law of iterated

expectations, $E[V(Y|X)] \geq E[V(Y|X,Z)]$. Conditioning on additional variables reduces the conditional variance on average. Note that for each $(X,Z) = (x,z)$, the conditional variance $V(Y|X=x, Z=z)$ may be smaller or larger than $V(Y|X = x)$.

To see 8,

$$\begin{aligned} E[(Y-g(X))^2] &= E\{[(Y-m(X)+(m(X)-g(X))]^2\} \\ &= E\{[Y-m(X)]^2\} + E\{[m(X)-g(X)]^2\} \\ &\quad + 2E\{U[m(X)-g(X)]\}, \end{aligned}$$

where $U = Y - m(X)$. Using Claim 4 above $E\{U[m(X)-g(X)]\} = 0$ so that

$$E\{[Y-g(X)]^2\} = E\{[Y-m(X)]^2\} + E\{[m(X)-g(X)]^2\}.$$

Clearly the left hand-side is minimized when $g(x)$ is chosen to be equal to $m(x)$.

2.3. Conditional mean function and the average treatment effect

Often we examine the conditional expectation function in order to study causal effect of one variable on another. Here we define the concept of causal effect and then discuss conditions under which the conditional mean function can be used to study causal effect.

In order to define the causal effect, we introduce a new notation to clearly describe the dependence of Y on X, so that

$$Y = Y(X, W).$$

The unobserved random variable W captures the randomness in Y beyond the randomness driven by X.

We define the treatment effect of X_1 on Y when X_1 changes from x_1 to x_1' and when $X_2 = x_2$ as

$$Y(x_1', x_2, W) - Y(x_1, x_2, W).$$

Suppose we observe the value of Y for which $X_1 = x_1'$ and $X_2 = x_2$. Without knowing W, we know $Y(x_1', x_2, W)$. However, because we do not know W, we would not know $Y(x_1, x_2, W)$. Analogously, by observing the value of Y for which $X_1 = x_1$ and $X_2 = x_2$, without knowing W, we know $Y(x_1, x_2, W)$ but, since we do not know W, we would not know $Y(x_1', x_2, W)$. Either way, we only know either $Y(x_1', x_2, W)$ or $Y(x_1, x_2, W)$, but not both for the same W. Therefore the treatment effect

$$Y(x_1', x_2, W) - Y(x_1, x_2, W)$$

we just defined cannot be directly obtained in data.

However, we show that the average treatment effect can be obtained under an additional assumption. To see this, we first define the average treatment effect of X_1 on Y when X_1 changes from x_1 to x_1' and when $X_2 = x_2$ as

$$E[Y(x_1', x_2, W) - Y(x_1, x_2, W)].$$

Note that

$$E[Y(x_1', x_2, W) - Y(x_1, x_2, W)] = E\{E[Y(x_1', x_2, W) - Y(x_1, x_2, W)|X_2]\},$$

one can obtain the average treatment effect if one can obtain

$$E[Y(x_1', x_2, W) - Y(x_1, x_2, W)|X_2 = x_2].$$

Since

$$E[Y(x_1', x_2, W) - Y(x_1, x_2, W)|X_2 = x_2] = E[Y(x_1', x_2, W)|X_2 = x_2] - E[Y(x_1, x_2, W)|X_2 = x_2],$$

we can obtain the average treatment effect if one can obtain

$$E[Y(x_1', x_2, W)|X_2 = x_2] \text{ and } \mathrm{E[Y(}x_1\mathrm{,}x_2\mathrm{,W)}|X_2{=}x_2\mathrm{]},$$

But, when $Y(x_1', x_2, W)$ and X_1 are independent given X_2, Claim 3 above implies that

$$\begin{aligned} E[Y(X_1, X_2, W)|X_1 = x_1', X_2 = x_2] &= E[Y(x_1', X_2, W)|X_1 = x_1', X_2 = x_2] \\ &= E[Y(x_1', X_2, W)|X_2 = x_2]. \end{aligned}$$

Analogously, when $Y(x_1, x_2, W)$ and X_1 are independent given X_2,

$$E[Y(X_1, X_2, W)|X_1 = x_1, X_2 = x_2] = E[Y(X_1, X_2, W)|X_2 = x_2].$$

Therefore

$$\begin{aligned} &E[Y(X_1, X_2, W)|X_1 = x_1', X_2 = x_2] - E[Y(X_1, X_2, W)|X_1 = x_1, X_2 = x_2] \\ &\quad = E[Y(x_1', x_2, W)|X_2 = x_2] - E[Y(x_1, x_2, W)|X_2 = x_2] \\ &\quad = E[Y(x_1', x_2, W) - Y(x_1, x_2, W)|X_2 = x_2]. \end{aligned}$$

Integrating over X_2 using the marginal distribution of X_2, we obtain the average treatment effect.

We have discussed the assumption under which the average treatment effect can be obtained using the conditional mean function. Note that it would be great if we can obtain the median or more generally a quantile of

$$Y(x_1', x_2, W) - Y(x_1, x_2, W).$$

However, that is not possible because we cannot observe

$$Y(x_1', x_2, W) - Y(x_1, x_2, W).$$

The average treatment effect is special in that we do not need a joint distribution of $Y(x_1', x_2, W)$ and $Y(x_1, x_2, W)$ due to the linearity in expectation operator:

$$E[Y(x_1', x_2, W) - Y(x_1, x_2, W)|X_2 = x_2] = E[Y(x_1', x_2, W)|X_2 = x_2] - E[Y(x_1, x_2, W)|X_2 = x_2].$$

When X_1 is a continuous random variable, one can consider taking a sequence of x_1' which converges to x_1:

$$\begin{aligned} &\frac{E[Y(X_1, X_2, W)|X_1 = x_1', X_2 = x_2] - E[Y(X_1, X_2, W)|X_1 = x_1, X_2 = x_2]}{x_1' - x_1} \\ &\quad = E\left[\frac{Y(x_1', x_2, W) - Y(x_1, x_2, W)}{x_1' - x_1}|X_2 = x_2\right] \end{aligned}$$

A sufficient conditions, in addition to the assumption that for any x_1' in the neighborhood of x_1 and for any x_2 if $Y(x_1', x_2, W)$ and X_1 are independent given X_2, for the limit on both sides of the equality to exist and equal are that (1) $E[Y(X, W)] < \infty$, (2) $Y(x, W)$ is continuously partially differentiable with respect to x_1, (3) $|\partial Y(x, W)/\partial x_1| \leq M(W)$ in a neighborhood of x with $E[M(W)] < \infty$.

Under these conditions

$$\frac{\partial E[Y|X = x]}{\partial x_1} = E\left[\frac{\partial Y}{\partial x_1}|X_2 = x_2\right].$$

2.4. Two roles of conditioning variables

We have discussed the need to consider conditional distribution when we deal with observational data. Here we discuss two specific roles conditioning achieve. First role is to define the parameter of interest. For example, demand function is a function of income, its own price, as well as prices of substitutes and complements. If we are interested in studying the demand function of females and males separately, then we need to further condition on gender. Second role is to achieve conditional independence of the two variables under focus, in order to use the conditional mean function to study causal relationship.

Under the assumption we have seen that the difference in the conditional mean function equals

$$E[Y(x_1', x_2, W) - Y(x_1, x_2, W)|X_2 = x_2].$$

However, we may not be interested in examining the average treatment effect of the groups defined by the conditioning variable, which is chosen to achieve the conditional independence assumption. In this case, after obtaining the average treatment effect given X_2, we need to integrate out with respect to the subvector of X_2, which does not correspond to the parameter of interest. Let $X_2' = (X_{21}, X_{22})'$, where X_{21} is the conditioning random vector corresponding to the parameter of interest, and X_{22} is the conditioning random variable which does not correspond to the parameter of interest. In this case, we can integrate out X_{21} conditioning on X_{11}, i.e.

$$E[Y(x_1', x_2, W) - Y(x_1, x_2, W)|X_{21} = x_{21}] = E\left\{E[Y(x_1', x_2, W) - Y(x_1, x_2, W)|X_{21}, X_{22}]|X_{21} = x_{21}\right\}$$

2.5. Exercises

(1) For a random variable Y and a random vector X of length K, show that $V(Y|X) = E[V(Y|X, Z)|X] + V[E(Y|X, Z)|X]$.
(2) For a random variable Y and a random vector X of length K, write out the elements of $E(XX')$, $V(X)$, and $Cov(X, Y)$ using elements of X such as X_k for $k = 1, \ldots, K$.
(3) Consider the discrete random vector (Y, X) where Y and X takes on values y_j, $j = 1, \ldots, J$ and x_k, $k = 1, \ldots, K$, respectively with $(Y, X) = (y_j, x_k)$ with probability p_{jk} for $j = 1, \ldots, J$ and $k = 1, \ldots, K$.
   (a) What is the (marginal) distributions of Y and X, respectively?
   (b) Describe the random variable $E(Y|X)$.
   (c) In this example, show that $E(Y) = E(E(Y|X))$.

CHAPTER 3

Linear Regression Model and Its Parameter Estimation

3.1. Linear Regression Model

So far we have discussed the conditional mean function. We now discuss how we estimate it. In this chapter, we discuss the linear regression model of Y given $X = x$, which specifies the conditional mean function of Y given $X = x$ parametrically as a linear in coefficient model

$$m(x) = \beta_0 + \beta_1 r_1(x) + \cdots + \beta_K r_K(x) = r(x)'\beta, \tag{3.1.1}$$

where $\beta' = (\beta_0, \beta_1, \ldots, \beta_K)$ and $r(x)' = (1, r_1(x), \ldots, r_K(x))$ is a known vector valued function of x.

From the properties of the conditional mean function, we can define $U = Y - m(X)$ and write

$$Y = r(X)'\beta + U,$$

where $E(U|X) = 0$. Note that $E(U|X = x) = 0$ implies $E(Y|X = x) = r(x)'\beta$ so that $E(U|X = x) = 0$ if and only if $E(Y|X = x) = r(x)'\beta$.

The linear regression model requires us to specify the functional form of the conditional mean function using *the linear in parameter specification.*

For example,

$$m(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2 + \beta_5 x_1 x_2.$$

We write $r(x_1, x_2) = (1, x_1, x_1^2, x_2, x_2^2, x_1 x_2)'$ and $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)'$ and write $m(x_1, x_2) = r(x_1, x_2)'\beta$.

Note that there can be a difference between the number of conditioning variables and the number of regressors.

Important restrictions of the linear regression model are that which variables enter is known, the functional form of $r(x)$ is known, and that it is linear in parameters.

Linear in parameter specification is restrictive but fairly general function can be included because the variables can enter nonlinearly although in a known way via $r(x)$.

There are cases in which linearity is not a binding constraint. This is sometimes referred to as a satiated model. For example, let x_1 and $x_2$be binary variables, both taking values 0 and 1. In this case,

$$m(x_1, x_2) = m(0, 0) + m(1, 0)x_1(1 - x_2) + m(0, 1)(1 - x_1)x_2 + m(1, 1)x_1 x_2.$$

In general, however, specifying the conditional mean function to this extent a priori, is demanding.

We shall study later nonparametric methods, which do not require the specification of a linear regression model. Note that nonparametric methods still require

us to specify which variables to condition on, although we do not need to know what the functional form of them is.

We shall see that the nonparametric approaches also use the linear regression model as a base. Thus the understanding of the linear regression model provides a base for the nonparametric analysis.

As we shall see, even if the linear regression model specification is incorrect, there is a sense in which an approximation to the conditional mean function is obtained.

Often, the linear regression model is written as

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K + U,$$

where for any values x_1, ..., x_K, $E(U|X_1 = x_1, \ldots, X_K = x_K) = 0$. In this notation, as we discussed, X_2 may be a square of X_1 but the expression in the conditional mean function does not fully reflect this because in this case, conditioning on X_2 is redundant. This is the reason for adopting a slightly cumbersome notation using $r(x)$. However, from now on, we will use the standard notation and the constant term is also absorbed into x so that now

$$x = (1, x_1, \ldots, x_K)'.$$

Since it is more concise to say that the number of regressors is K rather than $K+1$, we absorb the constant term in x_1 if there is a constant term so that from now on $x \in R^K$.

In the linear regression model, Y is called a dependent variable and X_j for $j = 1, \ldots K$ is called an independent variable. Other ways to refer to Y and X are explained and explanatory variables or a response variable and a control variable or a predicted variable and a predictor variable or a regressand and a regressor.

3.2. Homoskedasticity and Heteroskedasticiy

In addition to the linear regression specification, sometimes the so called homoskedasticity assumption

$$V(Y|X = x) = \sigma^2 \tag{3.2.1}$$

is maintained.

Homoskedasticity assumes that the conditional variance of the dependent variable does not depend on regressors. Under this assumption, the regressor may shift the conditional mean but not the conditional variance. Since this is unreasonable we typically do not rely on this assumption and leave the conditional variance as a general function of x, as in $\sigma^2(x)$. This is referred to as the case of heteroskedasticity.

3.3. Parameters in the linear regression model

In the homoskedastic linear regression model β and σ^2 are the unknown parameters to be estimated. Clearly the conditional mean function and the conditional variance function generally do not specify the entire distribution of Y given $X = x$.

Using the conditional density of U given $X = x$ (assuming that it exists), $f_{U|X}(\cdot/\sigma|x)/\sigma$, the conditional density of Y given $X = x$ can be written as $f_{U|X}((y-$

$x'\beta)/\sigma|x)/\sigma$, where $f_{U|X}(\cdot|x)$ satisfies

$$\int_{-\infty}^{\infty} f_{U|X}(u|x)du = 1, \quad \int_{-\infty}^{\infty} u f_{U|X}(u|x)du = 0, \text{ and } \int_{-\infty}^{\infty} u^2 f_{U|X}(u|x)du = 1.$$

Thus the entire conditional distribution of Y given $X = x$ can be parametrized by $(\beta, \sigma, f_{U|X}(\cdot/\sigma|x)/\sigma)$. Under heteroskedasticiy, the conditional distribution of Y given $X = x$ can be parametrized by $(\beta, \sigma(x), f_{U|X}(\cdot/\sigma(x)|x)/\sigma(x)$.

The linear regression model can be thought of as a generalization of estimating the mean $E(Y) = \mu$. We estimate the conditional mean function rather than unconditional mean and study how the conditional mean changes with respect to the value of the conditioning variables X.

3.4. Constant versus random regressors

The conditional mean notation assumes that the regressors are random variables. When we study finite sample properties of OLS estimator, whether regressors are stochastic or constant make little difference. When we study asymptotic properties, some differences arise as we shall see.

There are three cases where regressors can be regarded as constants: First, when they are iterally constant, for example, in the context of experiments. Second, when we have observations from stratified sampling, Third, when we condition on them.

3.5. Ordinary Least Squares (OLS) Estimator

3.5.1. Definition of the OLS estimator. We now turn to the discussion of how we estimate β by the Ordinary Least Squares method assuming a random sample of (X, Y) of size N, which we denote by $\{(x_i, y_i)\}_{i=1}^N$.

Recall that the conditional mean function can be characterized as a function which minimizes the prediction error using the mean squared error as the criterion; the solution to the following problem

$$\min_{g(\cdot)} E[(Y - g(X))^2]$$

is the conditional mean function. Since the conditional mean function is parameterized as $x'\beta$, a natural analog is to consider a class of functions $x'b$ for $b \in R^{K+1}$ and find the solution to the following problem

$$\min_{b} E[(Y - X'b))^2].$$

Since we cannot calculate the expectation, we approximate it by the sample analog and to define an estimator of β by the solution to the following problem

$$\min_{b} N^{-1} \sum_{i=1}^{N} (y_i - x_i'b)^2.$$

The estimator of β defined above is called the Ordinary Least Squares (OLS) estimator. Since the objective function is a quadratic function of arguments in b, the solution, say $\hat{\beta}$, can be computed explicitly. Denoting $\mathbf{Y} = (y_1, \ldots, y_N)'$ and $\mathbf{X} = (x_1, \ldots, x_N)'$, the objective function to be minimized can be written as

$$(\mathbf{Y} - \mathbf{X}b)'(\mathbf{Y} - \mathbf{X}b).$$

The first order condition with respect to b is

$$-\mathbf{X}'(\mathbf{Y}-\mathbf{X}b)=0,$$

which yields the solution: $\hat{\beta}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ when $\mathbf{X}$ is full rank so that $\mathbf{X}'\mathbf{X}$ is invertible.

An alternative way to motivate the OLS estimator is to view it as a method of moment estimator. To see this, note that $E(U|X)=0$implies $E(XU)=0$. Since $U=Y-X'\beta$, one can mimic the condition by a hypothetical value, b, in sample:

$$\frac{1}{N}\sum_{i=1}^{N}x_i(y_i-x_i'b)=0.$$

In matrix notation this is

$$\frac{1}{N}\mathbf{X}'(\mathbf{Y}-\mathbf{X}b)=0,$$

which corresponds (aside from the difference in the scalar multiples (-1 versus N^{-1}), which does not affect the solution) to the first order condition.

By inspecting the mean square objective function it is straight-forward to verified that (1) if we multiply y_i by c, then all of the OLS estimate is multiplied by c as well, (2) if we multiply the jth conditional variable by $c\neq 0$, then the jth OLS coefficient will be multiplied by $1/c$ without changing any other OLS estimates, and (3) if we add a constant to any variable, it shifts the constant term alone.

Because we use the squared loss function we recover the conditional mean function. If we use some other loss function, e.g. absolute deviation loss, then we will in general estimate a different conditional function.

For example, if we use the absolute deviation loss, then we estimate the conditional median function.

When there is only one regressor and a constant term, the model is called the simple regression model.

3.6. A geometry of the OLS estimator

3.6.1. OLS estimator as a decomposition of the projection of Y on the linear space spanned by column vectors of X. In order to graph the situation, we consider the case with 2 conditioning variables X_1 and X_2 with three data points so that

$$\mathbf{Y}=(y_1,y_2,y_3)',\quad \mathbf{X_1}=(x_{11},x_{12},x_{13})',\quad \mathbf{X_2}=(x_{21},x_{22},x_{23})',\quad \mathbf{U}=(u_1,u_2,u_3)'.$$

OLS picks b_1 and b_2 so that the (Euclidean) distance between $\mathbf{Y}$ and $b_1\mathbf{X_1}+b_2\mathbf{X_2}$ is minimized. The solution can be obtained by finding the projection of $\mathbf{Y}$ onto the space spanned by $\mathbf{X_1}$ and $\mathbf{X_2}$. The OLS estimate can be obtained by finding the linear combination of $\mathbf{X_1}$ and $\mathbf{X_2}$ that gives the point.

Note that, in order to find the unique combination, $\mathbf{X_1}$ and $\mathbf{X_2}$ should not lie on the same line; i.e. $\mathbf{X_1}$ and $\mathbf{X_2}$ should span a plane. This correspond to the rank condition on $\mathbf{X}$.

Although the point closest to Y on the space spanned by $\mathbf{X_1}$ and $\mathbf{X_2}$ is determined generally, there may not be a unique linear combination of $\mathbf{X_1}$ and $\mathbf{X_2}$.

In order for the OLS estimate to exist in the two independent variables case, it is necessary and sufficient for $\mathbf{X_1}$ and $\mathbf{X_2}$ to be linearly independent. (which is a different concept from the independence of two random variables.)

Denoting the OLS estimate by $(\hat{\beta}_1, \hat{\beta}_2)$ and defining $\hat{\mathbf{U}} = Y - \hat{\beta}_1 \mathbf{X_1} - \hat{\beta}_2 \mathbf{X_2}$, from the observation above, $\hat{\mathbf{U}}$ is orthogonal to the space spanned by $\mathbf{X_1}$ and $\mathbf{X_2}$ so that in particular $\hat{\mathbf{U}}$ and $\mathbf{X_1}$ are orthogonal and $\hat{\mathbf{U}}$ and $\mathbf{X_2}$ are orthogonal as well.

In a sense, OLS estimate is defined to have these properties. Recall the first order condition or the moment conditions; writing

$$\hat{\mathbf{U}} = \mathbf{Y} - \mathbf{X}\hat{\beta},$$

the OLS estimator satisfying either conditions implies $\mathbf{X}'\hat{\mathbf{U}} = 0$.

One can verify this mathematically, since

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

so that $\mathbf{X}'\hat{\mathbf{U}} = 0$. Note that this result holds regardless of what the true model is. It holds by the construction of the OLS estimate.

3.6.2. auxiliary regression. Using this observation, one can obtain more explicit expression for the OLS estimator.

First note that

$$\mathbf{Y} = \hat{\beta}_1 \mathbf{X_1} + \cdots + \hat{\beta}_K \mathbf{X_K} + \hat{\mathbf{U}}. \tag{3.6.1}$$

Consider a regression of X_1 on all other regressors $X_2, \ldots, X_K$ and denote the result similarly as above: this is called an *auxiliary regression* of X_1 on all other regressors.

$$\mathbf{X_1} = \hat{\alpha}_2 \mathbf{X_2} + \cdots + \hat{\alpha}_K \mathbf{X_K} + \hat{\mathbf{V}}_\mathbf{1}.$$

Denote

$$\hat{\mathbf{X}}_\mathbf{1} = \hat{\alpha}_2 \mathbf{X_2} + \cdots + \hat{\alpha}_K \mathbf{X_K}$$

so that $\mathbf{X_1} = \hat{\mathbf{X}}_\mathbf{1} + \hat{\mathbf{V}}_\mathbf{1}$.

Note that $\mathbf{X_j}'\hat{\mathbf{V}}_\mathbf{1} = 0$ for $j = 2, \ldots, K$ because of the same reasons we discussed to show $\mathbf{X}'\hat{\mathbf{U}} = 0$, only applying to the regression of X_1 on X_2 through X_K. Since $\hat{\mathbf{X}}_\mathbf{1}$ is a linear combinations of $\mathbf{X_j}$for $j = 2, \ldots, K$,

$$\hat{\mathbf{V}}_\mathbf{1}'\mathbf{X_1} = \hat{\mathbf{V}}_\mathbf{1}'\hat{\mathbf{X}}_\mathbf{1} + \hat{\mathbf{V}}_\mathbf{1}'\hat{\mathbf{V}}_\mathbf{1} = \hat{\mathbf{V}}_\mathbf{1}'\hat{\mathbf{V}}_\mathbf{1}$$

Thus, multiplying both sides of equation (3.6.1) by $\hat{\mathbf{V}}_\mathbf{1}'$, we obtain

$$\hat{\mathbf{V}}_\mathbf{1}'\mathbf{Y} = \hat{\beta}_1 \hat{\mathbf{V}}_\mathbf{1}'\hat{\mathbf{V}}_\mathbf{1} + \hat{\mathbf{V}}_\mathbf{1}'\hat{\mathbf{U}} = \hat{\beta}_1 \hat{\mathbf{V}}_\mathbf{1}'\hat{\mathbf{V}}_\mathbf{1}.$$

The last equality follows because $\hat{\mathbf{V}}_\mathbf{1} = \mathbf{X_1} - \hat{\mathbf{X}}_\mathbf{1}$ so that it is a linear combination of $\mathbf{X_j}$for $j = 1, \ldots, K$ and $\hat{\mathbf{U}}$ is orthogonal to all of them. Thus, when $\hat{\mathbf{V}}_\mathbf{1}'\hat{\mathbf{V}}_\mathbf{1} \neq 0$, $\hat{\beta}_1 = \hat{\mathbf{V}}_\mathbf{1}'\mathbf{Y}/\hat{\mathbf{V}}_\mathbf{1}'\hat{\mathbf{V}}_\mathbf{1}$.

If we regress Y on $X_2, \ldots, X_K$ and define the residual from this auxiliary regression to be $\hat{\mathbf{V}}_\mathbf{Y}$, then $\mathbf{Y} = \hat{\mathbf{Y}} + \hat{\mathbf{V}}_\mathbf{Y}$ where $\hat{\mathbf{Y}}$ is a linear combination of $\mathbf{X_j}$for $j = 2, \ldots, K$ so that $\hat{\mathbf{Y}}$ is orthogonal to $\hat{\mathbf{V}}_\mathbf{1}$. This implies that

$$\hat{\beta}_1 = \hat{\mathbf{V}}_\mathbf{1}'\mathbf{Y}/\hat{\mathbf{V}}_\mathbf{1}'\hat{\mathbf{V}}_\mathbf{1} = \hat{\mathbf{V}}_\mathbf{1}'(\hat{\mathbf{Y}} + \hat{V}_Y)/\hat{\mathbf{V}}_\mathbf{1}'\hat{\mathbf{V}}_\mathbf{1} = \hat{\mathbf{V}}_\mathbf{1}'\hat{\mathbf{V}}_\mathbf{Y}/\hat{\mathbf{V}}_\mathbf{1}'\hat{\mathbf{V}}_\mathbf{1}.$$

Note that an analogous result should hold for any other coefficient.

The result indicates that the OLS estimator of the jth coefficient can be obtained by regressing Y on the residual of the jth regressor's auxiliary regression on all other regressors.

The OLS estimator of the jth coefficient exists if the auxiliary regression residual is not identically 0. This is a weaker condition than the full rank condition on

X. It is possible that the OLS estimate corresponding to a subvector of β is well defined, whereas that for the complement subvector of β is not.

3.7. What does the OLS estimator estimate in general?

While the linear regression model can be flexible, it is hard to imagine that the conditional mean function is exactly correctly specified by the linear in coefficient model. Even when the linear regression model is misspecified, so that $E(Y|X = x) \neq x'\beta$, we show that the OLS estimator "approximates" the conditional mean function in the sense discussed below.

As we observed the least square objective function can be interpreted to seek a function $g(\cdot)$ that minimizes $E[(Y - g(X))^2]$.

In particular, we observed that

$$E[(Y - g(X))^2] = E[(Y - E(Y|X))^2] + E[(E(Y|X) - g(X))^2].$$

This implies that even if the function $g(X)$ is restricted to a class of functions which does not include $E(Y|X)$, it is still best to find a solution that minimizes $E[(Y - g(X))^2]$ because it corresponds to minimizing $E[(E(Y|X) - g(X))^2]$ within the class.

In the case of the linear regression model, $g(x) = x'b$ for $b \in R^K$, so that the objective function is a constant term plus $E[(E(Y|X) - g(X))^2]$ and the solution is $E(XX')^{-1}E(XE(Y|X))$ which equals $E(XX')^{-1}E(XY)$.

Denoting the set of elements in the support of X by $\text{Supp}(X)$, since

$$E[(E(Y|X) - g(X))^2] = \int_{x \in \text{Supp}(X)} (E(Y|X = x) - g(x))^2 f_X(x)dx,$$

this "approximation" depends on the distribution of X.

In particular the area in which there is no X data will be ignored so that it is *not* approximated at all.

If the $\text{Supp}(X)$ is very narrow, the approximation becomes fragile in some directions in a finite set of observations. This means that even if **X** has full rank, so that the OLS estimate is well defined, the result may be unstable in a sense that one data points away from the observed point may affect the estimate very much.

Note: Theil's textbook and Goldberger's Harvard U Press textbook is a good source for this part of the course. In addition, Theil's linear algebra statements in the textbook provide good exercises.

3.8. Goodness of fit measure

We have introduced the OLS estimator $\hat{\beta}$ of β in the linear regression model:

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

When x_i and u_i are not correlated,

$$V(y_i) = V(x_i\beta) + V(u_i)$$

so that assuming $V(y_i) > 0$,

$$\frac{V(x_i\beta)}{V(y_i)} + \frac{V(u_i)}{V(y_i)} = 1.$$

The fraction of the total variance of y_i explained by the observed component x_i, i.e. $Var(x_i\beta)/Var(y_i)$ is a goodness of fit measure often used.

It is estimated by the sample analog called the R^2:

$$R^2 = \frac{N^{-1}\sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2}{N^{-1}\sum_{i=1}^N [y_i - \bar{y}]^2},$$

where $\hat{y}_i = x_i'\hat{\beta}$ and $\bar{\hat{y}} = N^{-1}\sum_{i=1}^N \hat{y}_i$.

If there is a constant term among the regressors, which is usually the case, then $\bar{\hat{y}} = \bar{y}$ so that an alternative form often used for R^2 results:

$$R^2 = \frac{N^{-1}\sum_{i=1}^N (\hat{y}_i - \bar{y})^2}{N^{-1}\sum_{i=1}^N [y_i - \bar{y}]^2}.$$

Note that the first expression is more robust as it is a valid measure even if there is no constant term among the regressor.

To see that $\bar{\hat{y}} = \bar{y}$ when there is a constant term among the regressors, note that $y_i = \hat{y}_i + \hat{u}_i$ and that $\sum_{i=1}^N \hat{u}_i = 0$ if there is a constant term among the regressor. This implies that $\sum_{i=1}^N y_i = \sum_{i=1}^N \hat{y}_i$ so that $N^{-1}\sum_{i=1}^N y_i = N^{-1}\sum_{i=1}^N \hat{y}_i$.

Sometimes the quality of the regression results are dicussed based on the values of R^2, but that is misleading for at least three reasons. First, often the objective of an empirical study is to learn β and therefore the issue is whether we can estimate β accurately. Whether R^2 is large or not only tangentially related to this purpose if at all as we shall see. Second, high R^2 is not a part of the assumptions we need to maintain to derive the desirable properties of the OLS estimator as we shall see. For example, the R^2 may be high because regressors are correlated with the residual (a violation of the assumption we will maintain) term and thus there is very little "unobserved" part left. Third, there is a mechanical relationship that R^2 is higher if we include more regressors regardless of whether the model is correctly specified.

To address the third issue, sometimes "adjusted R^2" or "corrected R^2", typically denoted $\bar{R}^2$, is used. It is defined as

$$\bar{R}^2 = 1 - \frac{(N-K)^{-1}\sum_{i=1}^N \hat{u}_i^2}{(N-1)^{-1}\sum_{i=1}^N (y_i - \bar{y})^2},$$

where K denotes the number of regressors including the constant term. We will motivate this formula later.

However, if the objective of the study is to see how much a theory as captured by the observable component explains the dependent variable, then low R^2 or low $\bar{R}^2$ may be an issue. For example, wage regression based on human capital theory typically has R^2 of 0.1 to 0.3. It mean a large fraction of wage variation is left unexplained by the theory.

3.9. Exercises

(1) Suppose X_1 denotes years of education and X_2 denotes gender, where $X_2 = 1$ denotes males and $X_2 = 0$ denotes females. Y is wage.
   (a) Please specify a linear regression model which allows years of education to affect wage differently for males and females and also for whether one is in manufacturing sector or not.
   (b) If Y is log-wage, how does the interpretation of the coefficients change?

(2) In addition to X_1 and X_2 above, assume that X_3 denotes a manufacturing industry dummy variable where $X_3 = 1$ if the person works in a manufacture industry and $X_3 = 0$ if not.

(a) Specify the most general model in which X_1 enters linearly.
(b) Specify the model in which all the variables enter linearly and explain the restrictions imposed in the less general model.

(3) Let X_1 and X_2 be both binary random variables taking values x_{10} and x_{11}, and also x_{20} and x_{21}, respectively. Using the indicator variables $1\{x_1 = x_{10}\}$ and $1\{x_2 = x_{20}\}$, describe how one can use a linear regression model to formulate a satiated model.
(4) Show that if we multiply the dependent variables by c, then all of the OLS estimate is multiplied by c as well.
(5) Show that if we multiply the jth independent variable by $c \neq 0$, then the jth OLS coefficient will be multiplied by $1/c$ without changing any other OLS estimates.
(6) Show that if we add a constant to any variable, it shifts the constant term alone. Show how the constant term shifts.
(7) If we include an additional regressor, R^2 becomes always strictly larger if the additional regressor is linearly independent from the rest of the regressors and the OLS estimate of the additional coefficient is not zero. To show this consider the regression model $y_i = x_i'\beta + \alpha z_i + \epsilon_i$ and compare the R^2 for this model with the R^2 for the model $y_i = x_i'\beta + u_i$. Consider the auxiliary regression of z_i on x_i and a constant term, if its already not included among regressors in x_i and define the predicted value as $\hat{z}_i = x_i'\hat{\pi} + \hat{\pi}_0$ and the residual as $\hat{v}_{zi}$. Let $\hat{\beta}$ be the OLS estimate of the smaller model and $\hat{\beta}_1$ and $\hat{\alpha}$ be the OLS estimates of the larger model. Also let $\hat{u}_i$ denote the OLS residual from the smaller model and denote the OLS residual from the larger model $\hat{\epsilon}_i$, so that

$$y_i = x_i'\hat{\beta} + \hat{u}_i, \qquad y_i = x_i'\hat{\beta}_1 + \hat{\alpha} z_i + \hat{\epsilon}_i$$

(a) Show that $\hat{\beta} = \hat{\beta}_1 + \hat{\alpha}\hat{\pi}$ and that $y_i = x_i'\hat{\beta} + \hat{\alpha}\hat{v}_{zi} + \hat{\alpha}\hat{\pi}_0 + \hat{\epsilon}_i$, where $\hat{u}_i = \hat{\alpha}\hat{v}_{zi} + \hat{\alpha}\hat{\pi}_0 + \hat{\epsilon}_i$.
(b) Let $\hat{y}_i = x_i'\hat{\beta}_1 + \hat{\alpha} z_i$ and $\hat{\hat{y}}_i = x_i'\hat{\beta}$. Show that $\hat{y}_i = x_i'\hat{\beta} + \hat{\alpha}\hat{v}_{zi} + \hat{\alpha}\hat{\pi}_0$ so that

$$\hat{y}_i = \hat{\hat{y}}_i + \hat{\alpha}\hat{v}_{zi} + \hat{\alpha}\hat{\pi}_0.$$

(c) Show that $\bar{\hat{y}} = \bar{\hat{\hat{y}}} + \hat{\alpha}\hat{\pi}_0$ and that $\hat{y}_i - \bar{\hat{y}} = \hat{\hat{y}}_i - \bar{\hat{\hat{y}}} + \hat{\alpha}\hat{v}_{zi}$.
(d) Prove the main statement.
(e) Show, by inspecting the objective function, that R^2 becomes always weakly larger if a regressor is added.

CHAPTER 4

Properties of the OLS Estimator

4.1. Finite sample properties of the OLS estimator

We examine the properties of the OLS estimator under the following assumptions:

Assumption OLS.1 (conditional mean version): $y_i = x_i'\beta + u_i$ and $E(u_i|x_i) = 0$.

Assumption OLS.2: (sample version) $\text{rank}(\mathbf{X}) = K$.

Assumption OLS.3 (conditional homoskedasticity): $E(u_i^2|x_i) = \sigma^2$.

Assumption OLS.4: Sampling of (x_i, y_i) is i.i.d.

4.1.1. expectation of a random vector or a random matrix. In this section we use the conditional expectation notation on a random vectors or a random matrices. For a random vector $\mathbf{Z} = (Z_1, \ldots, Z_m)'$

$$E(\mathbf{Z}|\mathbf{X}) = (E(Z_1|\mathbf{X}), \ldots, E(Z_m|\mathbf{X}))'.$$

For a random matrix, the notation is analogous. Note that for a matrix $\mathbf{A}$ of size $n \times m$, for which each element is a function of $\mathbf{X}$ only,

$$E(\mathbf{AZ}|\mathbf{X}) = \mathbf{A}E(\mathbf{Z}|\mathbf{X}).$$

This can be verified by inspecting each term on the left and the right using the notation above and exploiting the linearity property of the expectation. For example, writing the (i, j) element of $\mathbf{A}$ by a_{ij}, the first element of the left-hand side is

$$E(a_{11}Z_1 + a_{12}Z_2 + \cdots + a_{1m}Z_m|\mathbf{X}) = a_{11}E(Z_1|\mathbf{X}) + a_{12}E(Z_2|\mathbf{X}) + \cdots + a_{1m}E(Z_m|\mathbf{X})$$

which equals the first element on the right-hand side.

When $\mathbf{Z}$ is an $m \times \ell$ matrix, one can apply the above results on columns of $\mathbf{AZ}$, because writing

$$\mathbf{Z} = (\mathbf{Z_1}, \ldots, \mathbf{Z}_\ell)$$
$$\mathbf{AZ} = (\mathbf{AZ_1}, \ldots, \mathbf{AZ}_\ell)$$

so that

$$\begin{aligned} E(\mathbf{AZ}|\mathbf{X}) &= (E(\mathbf{AZ_1}|\mathbf{X}), \ldots, E(\mathbf{AZ}_\ell|\mathbf{X})) \\ &= (\mathbf{A}E(\mathbf{Z_1}|\mathbf{X}), \ldots, \mathbf{A}E(\mathbf{Z}_\ell|\mathbf{X})) = \mathbf{A}E(\mathbf{Z}|\mathbf{X}). \end{aligned}$$

Cases with post-multiplication follows a similar rule. This can be verified by observing that the transposing a vectors or matrices on both sides of the equality keeps the equality.

4.1.2. conditional unbiasedness of the OLS estimator. The OLS estimator is a rare example for which we can compute the conditional mean and the conditional variance of the estimator.

We first show that under Assumptions OLS.1 (conditional mean version), OLS.2 (sample version), and OLS.4, the OLS estimator is conditionally unbiased by showing that the conditional mean of the OLS estimator is β.

Recall that the OLS estimator is $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Note that

$$\begin{aligned}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{U}) \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}.\end{aligned}$$

From this, we obtain

$$\begin{aligned}E(\hat{\beta}|\mathbf{X}) &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{U}|\mathbf{X}) \\ &= \beta.\end{aligned}$$

The last equality follows because $E(u_i|\mathbf{X}) = E(u_i|x_i) = 0$ by the random sampling assumption and then by OLS.1 (conditional mean version).

Note that the conditional homoskedasticity assumption is not needed for the conditional unbiasedness result.

Note that $\mathbf{X}$needs to be full rank for the OLS estimator to be well defined. It means that the conditional mean unbiasedness of the OLS is well defined only for those observations for which $\mathbf{X}$has full rank. Therefore unconditional unbiasedness of the OLS estimator does not follow from the conditional unbiasedness unless the probability that $\mathbf{X}$has full rank is one. This only holds if there is a continuous random variable.

For example, consider a simple regression with a dummy regressor taking value zero or one. For any finite sample size, the probability that dummy variable realizations are all zeros or all ones is positive.

4.1.3. conditional variance of the OLS estimator. Next we show that under Assumptions OLS.1 (conditional mean version), OLS.2 (sample version), OLS.3 (conditional homoskedasticity) and OLS.4, $V(\hat{\beta}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

To see this, recall that $\hat{\beta} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}$ and the conditional mean of $\hat{\beta}$ is β, so that the conditional variance of $\hat{\beta}$ is

$$\begin{aligned}E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}\mathbf{U}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{U}\mathbf{U}'|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

This result obtains the conditional variance and the conditional covariance at the same time.

When the conditional homoskedasticity condition does not hold, we can still obtain the conditional variance as

$$\begin{aligned}E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}\mathbf{U}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{U}\mathbf{U}'|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega(\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

where $\Omega(\mathbf{X})$ is a diagonal matrix with the ith element being $\sigma^2(x_i)$. In vector notation, the last expression is

$$(\sum_{i=1}^{N} x_i x_i')^{-1} \sum_{i=1}^{N} x_i x_i' \sigma^2(x_i) (\sum_{i=1}^{N} x_i x_i')^{-1}.$$

From this expression, it is not immediately clear what factors determine the conditional variance. We examine this next using the explicit expression for the jth argument of the OLS estimator earlier.

Recall that

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\hat{\mathbf{V}}_\mathbf{1}' \mathbf{Y}}{\hat{\mathbf{V}}_\mathbf{1}' \hat{\mathbf{V}}_\mathbf{1}} \\
&= \frac{\hat{\mathbf{V}}_\mathbf{1}' (\beta_\mathbf{1} \mathbf{X}_\mathbf{1} + \cdots + \beta_\mathbf{K} \mathbf{X}_\mathbf{K} + \mathbf{U})}{\hat{\mathbf{V}}_\mathbf{1}' \hat{\mathbf{V}}_\mathbf{1}} \\
&= \frac{\beta_1 \hat{\mathbf{V}}_\mathbf{1}' \hat{\mathbf{V}}_\mathbf{1} + \mathbf{U}}{\hat{\mathbf{V}}_\mathbf{1}' \hat{\mathbf{V}}_\mathbf{1}}.
\end{aligned}$$

The last equality follows because $\mathbf{X_1} = \hat{\mathbf{X}}_\mathbf{1} + \hat{\mathbf{V}}_\mathbf{1}$, where $\hat{\mathbf{X}}_\mathbf{1}$ is a linear combination of $\mathbf{X_j}$ for $j = 2, \ldots, K$, and that $\hat{\mathbf{V}}_\mathbf{1}$ is orthogonal to all of $\mathbf{X_j}$ for $j = 2, \ldots, K$. Thus

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{N} \hat{v}_{1i} u_i}{\sum_{i=1}^{N} \hat{v}_{1i}^2},$$

where $\hat{v}_{1i}$ is the ith element of $\hat{\mathbf{V}}_\mathbf{1}$. Observing that $\hat{v}_{1i}$ for all $j = 1, \ldots, K$ are all computed using regressors only, they are constants given $\mathbf{X}$.

This implies that

$$\begin{aligned}
V(\hat{\beta}_1 | \mathbf{X}) &= \frac{\sum_{i=1}^{N} \hat{v}_{1i}^2 E(u_i^2 | \mathbf{X})}{[\sum_{i=1}^{N} \hat{v}_{1i}^2]^2} \\
&= \frac{\sum_{i=1}^{N} \hat{v}_{1i}^2 E(u_i^2 | x_i)}{[\sum_{i=1}^{N} \hat{v}_{1i}^2]^2} \\
&= \frac{\sigma^2}{\sum_{i=1}^{N} \hat{v}_{1i}^2} \\
&= \frac{\sigma^2}{\frac{\sum_{i=1}^{N} \hat{v}_{1i}^2}{\sum_{i=1}^{N} (x_{1i} - \bar{x}_1)^2} \sum_{i=1}^{N} (x_{1i} - \bar{x}_1)^2} \\
&= \frac{\sigma^2 / N}{(1 - R_1^2) \hat{V}(x_1)}.
\end{aligned}$$

The second equality follows from the random sampling assumption. The last equality follows from the definition of the R^2 applied to the auxiliary regression of X_1 on X_j for $j = 2, \ldots, K$, which we denote by R_1^2.

From the third equality it follows that if X_1 can be predicted by the rest of the regressors well in the sense that $\sum_{i=1}^{N} \hat{v}_{1i}^2$ is small, then the conditional variance of $\hat{\beta}_1$ is large.

This can be decomposed into 4 factors:

(1) Conditional variance of the dependent variable σ^2. (conditional variance of $\hat{\beta}_1$ is larger as it is larger)
(2) Variance of the regressor under consideration. (conditional variance of $\hat{\beta}_1$ is smaller as it is larger)
(3) R^2 of the auxiliary regression. (conditional variance of $\hat{\beta}_1$ is larger as it is larger)
(4) The sample size. (conditional variance of $\hat{\beta}_1$ is smaller as it is larger)

The result clearly applies to any other OLS estimator of the rest of the coefficients.

If the conditional homoskedasticity assumption does not hold, then

$$\begin{aligned} V(\hat{\beta}_1|\mathbf{X}) &= \frac{\sum_{i=1}^N \hat{v}_{1i}^2 \sigma^2(x_i)}{[\sum_{i=1}^N \hat{v}_{1i}^2]^2} \\ &= \frac{\sum_{i=1}^N \hat{v}_{1i}^2 \sigma^2(x_i)}{\sum_{i=1}^N \hat{v}_{1i}^2} \frac{1}{\sum_{i=1}^N \hat{v}_{1i}^2} \\ &= \frac{N^{-1}\sum_{i=1}^N \hat{v}_{1i}^2 \sigma^2(x_i)}{N^{-1}\sum_{i=1}^N \hat{v}_{1i}^2 N^{-1}\sum_{i=1}^N \sigma^2(x_i)} \frac{N^{-1}\sum_{i=1}^N \sigma^2(x_i)}{\sum_{i=1}^N \hat{v}_{1i}^2}. \end{aligned}$$

The second term of the last expression corresponds to the homoskedastic case above. The first term is one if the sample covariance of $\hat{v}_{1i}^2$ and $\sigma^2(x_i)$ for $i = 1, \ldots, N$ is zero, greater than one if it is positive, and less than 1 if it is negative. Recall that $\hat{v}_{1i}$ is the OLS residual of regressing the first regressor on the rest of the regressors and this is solely related to relationship among regressors. Thus there is no reason why the size of the residual is positively or negatively correlated with the conditional variance of Y_i given X_i.

[Graphical explanation about why positive correlation between $\hat{v}_{1i}^2$ and $\sigma(x_i)^2$ will lead to a larger variance.]

4.1.4. estimation of σ^2. The conditional variance of the OLS estimator depends on σ^2, which is an unknown parameter of the linear regression model. If we observe u_i, then the method of moment estimator is

$$\frac{1}{N}\sum_{i=1}^N u_i^2.$$

Since u_i is not observed, a natural estimator of σ^2 can be constructed using its estimate $\hat{u}_i = y_i - x_i'\hat{\beta}$,

$$\frac{1}{N}\sum_{i=1}^N \hat{u}_i^2.$$

It turned out this estimator is biased, so we will consider the following estimator

$$\hat{\sigma}^2 = \frac{1}{N-K}\sum_{i=1}^N \hat{u}_i^2.$$

In matrix notation let $\hat{\mathbf{U}} = \mathbf{Y} - \mathbf{X}\hat{\beta} = [I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} = [I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{U}$ so that

$$\hat{\sigma}^2 = \frac{1}{N-K}\mathbf{U}'[I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{U}.$$

Equality follows because $I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is an idempotent matrix.

We compute the conditional mean of $\hat{\sigma}^2$:

$$\begin{aligned}
E(\hat{\sigma}^2|\mathbf{X}) &= \frac{1}{N-K}E\{\mathbf{U}'[I-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{U}|\mathbf{X}\} \\
&= \frac{1}{N-K}E\{trace\mathbf{U}'[I-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{U}|\mathbf{X}\} \\
&= \frac{1}{N-K}E\{trace[I-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{U}\mathbf{U}'|\mathbf{X}\} \\
&= \frac{1}{N-K}traceE\{[I-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{U}\mathbf{U}'|\mathbf{X}\} \\
&= \frac{1}{N-K}trace[I-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']E\{\mathbf{U}\mathbf{U}'|\mathbf{X}\} \\
&= \frac{\sigma^2}{N-K}trace[I-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \\
&= \frac{\sigma^2}{N-K}\{trace(I)-trace[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\} \\
&= \frac{\sigma^2}{N-K}\{N-trace[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}]\} \\
&= \frac{\sigma^2(N-K)}{N-K} = \sigma^2.
\end{aligned}$$

Conditional variance of $\hat{\sigma}^2$ can be shown to be

$$2\sigma^4/(N-K) + (\mu_4 - 3\sigma^4)\sum_{i=1}^{N}[1-x_i'(X'X)^{-1}x_i]^2/(N-K)^2.$$

Under conditional normality of the residual terms, $\mu_4 = 3\sigma^4$ so that the conditional variance simplifies to just the first term. One can show that the expression can be shown to be

$$\frac{\mu_4-\sigma^4}{N-K} + O(N^{-2}),$$

so that the deviation from the normal case is of lower order. A proof is given in the last section.

4.1.5. OLS is the Best Linear Unbiased Estimator (BLUE). The OLS estimator has the property that it has the smallest conditional variance among all the estimators that are obtained as a linear combination of the dependent variables and also are conditionally unbiased. Having this property, the OLS estimator is BLUE.

To see this, we consider an estimator of $c'\beta$, for a given constant vector $c \in R^K$. Since the estimator needs to be a linear combination of the dependent variable, we take a vector $a \in R^N$ and write the estimator as $a'\mathbf{Y}$.In order for the linear estimator to be conditionally unbiased,

$$E(a'\mathbf{Y}|\mathbf{X}) = c'\beta$$

should hold for any $\beta \in R^K$. The left hand side equals $a'\mathbf{X}\beta$ since $E(\mathbf{U}|\mathbf{X}) = 0$ so that

$$a'\mathbf{X}\beta = c'\beta$$

should hold for any $\beta \in R^K$. In particular, by taking $\beta = e_j$, for $j = 1, \ldots, N$, where e_j denotes a vector with all elements except the jth element to be zero and

the jth element equals one, we see that $a'\mathbf{X} = c'$. On the other hand, if $a'\mathbf{X} = c'$, then

$$a'\mathbf{X}\beta = c'\beta$$

should hold for any $\beta \in R^K$ so that these two conditions are equivalent. Thus the conditional unbiasedness requirement of the estimator can be stated as $a'\mathbf{X} = c'$ or by transposing it, $\mathbf{X}'a = c$.

Next consider the conditional variance of $a'\mathbf{Y}$.Since conditional variance equals that of $a'\mathbf{U}$, which equals

$$E[(a'\mathbf{U})^2|\mathbf{X}] = E[(a'\mathbf{U})(\mathbf{U}'a)|\mathbf{X}] = a'E[\mathbf{U}\mathbf{U}'|\mathbf{X}]a = \sigma^2 a'a.$$

We want to choose a to minimize this under the constraint that $\mathbf{X}'a = c$, so that the unbiasedness holds. The problem is equivalent to the following minimization problem:

$$\min_{a \in \{a \in R^N; \mathbf{X}'a = c\}} \frac{1}{2}a'a.$$

The Lagrangian problem for this constrained optimization problem with the Lagrangian multiplier $\lambda \in R^K$ is

$$\mathcal{L} = \frac{1}{2}a'a - \lambda'(\mathbf{X}'a - c).$$

The first order conditions are

$$a - \mathbf{X}\lambda = 0$$
$$\mathbf{X}'a - c = 0.$$

Substituting the a in the first equation in place of the second equation we have $\mathbf{X}'\mathbf{X}\lambda = c$. Thus assuming that $\mathbf{X}$has full rank, $\lambda = (\mathbf{X}'\mathbf{X})^{-1}c$. Substituting this expression in place of λ in the first equation we obtain

$$a = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}c.$$

That is, the best linear estimator of $c'\beta$ is $c'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{Y}$. So the best estimator must coincide with that using the OLS estimator on the linear combination.

Thus regardless of what c is, it is best to use the OLS estimator to estimate $c'\beta$.

Note that if we consider a conditionally unbiased estimator of β by $A\mathbf{Y}$ for any $K \times N$ matrix A, by the same reasoning as above, we should have $A\mathbf{X} = I_K$. With this restriction on A, the conditional variance-covariance matrix of $A\mathbf{Y}$ is

$$E(A\mathbf{U}\mathbf{U}'A'|\mathbf{X}) = AE(\mathbf{U}\mathbf{U}'|\mathbf{X})A' = \sigma^2 AA'.$$

If A satisfies $A\mathbf{X} = I_K$, because $c'A\mathbf{Y}$ is a linear in $\mathbf{Y}$ estimator, which is conditionally unbiased for $c'\beta$, for any $c \in R^K$,

$$c'AA'c \geq c'(\mathbf{X}'\mathbf{X})^{-1}c$$

should hold from the discussion above. This implies that $AA' - (\mathbf{X}'\mathbf{X})^{-1}$ is a positive semi-definite matrix if $A\mathbf{X} = I_K$. This can be verified directly by observing that

$$\begin{aligned}[A - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][A - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\ = AA' - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'A' - A\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1} \\ = AA' - (\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

Since the first expression is positive semidefinite, the last expression is also.

Note that we have assumed homoskedasticity in the discussion. The OLS estimator is not BLUE when homoskedasticity assumption does not hold. We will discuss this point when we discuss deviations from the OLS assumptions.

4.2. Finite Sample Distribution of the OLS Estimator

We have shown that the OLS estimator is a conditionally unbiased estimator and also computed its conditional variance. However the distribution of it is not known.

In fact, as we discussed, the conditional distribution of U given X is unspecified except that it has mean zero and a constant variance for the case of homoskedasticity. For the case of heteroskedasticity, the conditional variance is allowed to be a general function of $X = x$. Without a specific distribution of U, we cannot derive the distribution of the OLS estimator.

Without knowing the distribution of the OLS estimator, we cannot evaluate the probability statements about the OLS estimator.

In order to derive the distribution of the OLS estimator, we assume the distribution of U given X.

Assumption OLS.5 (conditional normality): $U|X = x \sim N(0, \sigma^2)$.

Assumption OLS.5 includes Assumptions OLS.1 and OLS.3 and in addition, assume that the conditional distribution is normal. This assumption is equivalent to $Y|X = x \sim N(x'\beta, \sigma^2)$.

Normality assumption itself is hard to justify but all the results we discuss below hold asymptotically without the normality assumption. That is why the normality assumption is useful.

Recall that the OLS estimator can be written as (without losing generality we continue to focus on $\hat{\beta}_1$)

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^N \hat{v}_{1i} u_i}{\sum_{i=1}^N \hat{v}_{1i}^2},$$

Since a linear combination of the jointly normal random variables has the normal distribution, we have

$$\hat{\beta}_1 | \mathbf{X} \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^N \hat{v}_{1i}^2}\right).$$

From this, we have

$$\frac{\hat{\beta}_1 - \beta_1}{[\sigma^2 / \sum_{i=1}^N \hat{v}_{1i}^2]^{1/2}} | \mathbf{X} \sim N(0, 1).$$

Although the transformed random variable is the standard normal random variable, we cannot use the result to compute the confidence interval for β_1 or to conduct hypothesis test about β_1 because σ^2 is not known.

However, by dividing this by $\sqrt{\hat{\sigma}^2/\sigma^2}$ eliminates the unknown σ and obtain the expression just like the one above, except that the unknown σ^2 is replaced by its unbiased estimator $\hat{\sigma}^2$,

$$\hat{\sigma}^2 = \frac{1}{N-K} \sum_{i=1}^N \hat{u}_i^2,$$

where $\hat{u}_i = y_i - x_i'\hat{\beta}$. We show that the ratio conditional on $\mathbf{X}$has $t(N-K)$ distribution, under Assumptions OLS.1–OLS.5.

Recall that the t-distribution with m degrees of freedom results when we have the ratio of two independent random variables, where the numerator is the standard normal random variable and the denominator is the square root of the χ^2 random variable with m degrees of freedom divided by m; thus

$$t(m) = \frac{N(0,1)}{\sqrt{\chi^2(m)/m}}$$

has the t-distribution with m degrees of freedom. We have seen already that the numerator

$$\frac{\hat{\beta}_1 - \beta_1}{[\sigma^2/\sum_{i=1}^N \hat{v}_{1i}^2]^{1/2}}|\mathbf{X}$$

has the standard normal distribution. Below we show that $(N-K)\hat{\sigma}^2/\sigma^2|\mathbf{X}$ has the $\chi^2(N-K)$ distribution and it is independent from the numerator.

Note that

$$\begin{aligned}\frac{(N-K)\hat{\sigma}^2}{\sigma^2} &= \frac{\hat{\mathbf{U}}'\hat{\mathbf{U}}}{\sigma^2} = \frac{(\mathbf{Y}-\mathbf{X}\hat{\beta})'(\mathbf{Y}-\mathbf{X}\hat{\beta})}{\sigma^2} \\ &= \frac{(\mathbf{Y}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y})'(\mathbf{Y}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y})}{\sigma^2} \\ &= \frac{\mathbf{Y}'[I-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}}{\sigma^2} = \frac{\mathbf{U}'}{\sigma}[I-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\frac{\mathbf{U}}{\sigma}.\end{aligned}$$

Conditional on $\mathbf{X}$,$\mathbf{U}/\sigma$ is a vector of independent standard normal random variables and $I-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is an idempotent matrix.

If V is a vector of independent standard normal random variables, then for any symmetric idempotent matrix A, $V'AV$ has a chi-square distribution with the degrees freedom equals to the rank of A. A proof is given in the Appendix.

Therefore, $\hat{\sigma}^2/\sigma^2$ is a chi-square random variable divided by $N-K$.

To see that $\hat{\beta}$ and $\hat{\sigma}^2$ are independent, given $\mathbf{X}$,observe that $\hat{\beta}$ is a function of $\mathbf{X}'\mathbf{U}$ and $\hat{\sigma}^2$ is a function of $\hat{\mathbf{U}} = [I-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{U}$. So if we can show that $\mathbf{X}'\mathbf{U}$ and $[I-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{U}$ are independent, $\hat{\beta}$ and $\hat{\sigma}^2$ are independent. Since $[I-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{U}$ is a random vector which lies in the space spanned by the independent column vectors of $I-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, we can find $N-K$ orthogonal vectors h_j, for $j=1,\ldots,N-K$ such that they lie in the space spanned by the column vectors of $I-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ such that

$$\hat{\mathbf{U}} = (h_1'\mathbf{U},\ldots,h_{N-K}'\mathbf{U},\tilde{h}_1'\mathbf{U},\ldots,\tilde{h}_K'\mathbf{U})'$$

where each of $\tilde{h}_j$ for $j=1,\ldots,K$ can be written as a linear combination of h_j, for $j=1,\ldots,N-K$. This implies that $\hat{\mathbf{U}}$ is a function of $H\mathbf{U}$, where $H = (h_1,\ldots,h_{N-K})'$. Now consider an $N\times N$ matrix

$$\begin{pmatrix} h_1' \\ \vdots \\ h_{N-K}' \\ \mathbf{X}' \end{pmatrix} = \begin{pmatrix} H \\ \mathbf{X}' \end{pmatrix}.$$

Note that h_j for any $j=1,\ldots,N-K$ are mutually independent, $\mathbf{X}'h_j = 0$ for any $j=1,\ldots,N-K$ so that columns of $\mathbf{X}$are independent from h_j for any $j=$

$1, \ldots, N-K$, and that $\mathbf{X}$ has rank K so that each column of $\mathbf{X}$ are independent, so that the matrix has full rank. Thus

$$\begin{pmatrix} H \\ \mathbf{X}' \end{pmatrix} \mathbf{U}$$

is a multivariate normal random vector and that covariance of $H\mathbf{U}$ and $\mathbf{X}'\mathbf{U}$is, under homoskedasticity

$$E(H\mathbf{U}\mathbf{U}'\mathbf{X}|\mathbf{X}) = HE(\mathbf{U}\mathbf{U}'|\mathbf{X})\mathbf{X} = \sigma^2 H\mathbf{X} = 0.$$

Thus $H\mathbf{U}$ and $\mathbf{X}'\mathbf{U}$ are independent.

Next, we turn to considering the distribution of a linear combination of the OLS estimators of the coefficients in the linear regression model.

Using the analogous notations as above, for a given constants c_j for $j = 1, \ldots, K$, a linear combination of the OLS estimator of the coefficients in the linear regression model can be rewritten as

$$\begin{aligned} \sum_{j=1}^{K} c_j \hat{\beta}_j &= \sum_{j=1}^{K} c_j \beta_j + \sum_{j=1}^{K} c_j \frac{\sum_{i=1}^{N} \hat{v}_{ji} u_i}{\sum_{i=1}^{N} \hat{v}_{ji}^2} \\ &= \sum_{j=1}^{K} c_j \beta_j + \sum_{i=1}^{N} u_i \left[\sum_{j=1}^{K} c_j \frac{\hat{v}_{ji}}{\sum_{i=1}^{N} \hat{v}_{ji}^2} \right]. \end{aligned}$$

By the same reasoning as that for the single coefficient case, writing

$$\hat{A}_i = \sum_{j=1}^{K} c_j \frac{\hat{v}_{ji}}{\sum_{i=1}^{N} \hat{v}_{ji}^2},$$

and $\hat{V}_c = \sum_{i=1}^{N} \hat{A}_i^2$ we obtain,

$$\sum_{j=1}^{K} c_j \hat{\beta}_j | \mathbf{X} \sim N\left(\sum_{j=1}^{K} c_j \beta_j, \sigma^2 \hat{V}_c \right)$$

so that

$$\frac{\sum_{j=1}^{K} c_j \hat{\beta}_j - \sum_{j=1}^{K} c_j \beta_j}{\sigma \sqrt{\hat{V}_c}} | \mathbf{X} \sim N(0,1).$$

Again, substituting σ with $\hat{\sigma}$, under Assumptions OLS.1–OLS.5, yields

$$\frac{\sum_{j=1}^{K} c_j \hat{\beta}_j - \sum_{j=1}^{K} c_j \beta_j}{\hat{\sigma} \sqrt{\hat{V}_c}} | \mathbf{X} \sim t(N-K).$$

We use these results to construct confidence interval and conduct a hypothesis test on a linear combination of the linear regression coefficients.

4.3. Confidence interval for a linear combination of the linear regression coefficients

4.3.1. how to choose a confidence interval. OLS estimator provides a point estimate of the coefficient in the linear regression model given a particular data set. Here we consider interval estimation. We study constructing a confidence interval for the linear combination of the regression coefficients.

First, observe that the distribution is as derived above:

$$\frac{\sum_{j=1}^K c_j\hat{\beta}_j - \sum_{j=1}^K c_j\beta_j}{\sqrt{\hat{\sigma}^2\hat{V}_c}}|\mathbf{X} \sim t(N-K).$$

Thus for any pth percentile, we can compute $t_p(N-K)$: i.e.

$$\Pr\left\{\frac{\sum_{j=1}^K c_j\hat{\beta}_j - \sum_{j=1}^K c_j\beta_j}{\sqrt{\hat{\sigma}^2\hat{V}_c}} < t_p(N-K) \mid X\right\} = p.$$

Using this, and observing that the t-distribution is symmetric, one can compute a number $t_{\alpha/2}(N-K)$ such that (assume $\alpha < 0.5$)

$$\Pr\left\{t_{\alpha/2}(N-K) < \frac{\sum_{j=1}^K c_j\hat{\beta}_j - \sum_{j=1}^K c_j\beta_j}{\sqrt{\hat{\sigma}^2\hat{V}_c}} < |t_{\alpha/2}(n-k)| \mid X\right\} = 1-\alpha.$$

So called the *level $1-\alpha$ confidence interval* for $\sum_{j=1}^K c_j\beta_j$ is computed using the above result:

$$[\sum_{j=1}^K c_j\hat{\beta}_j - \sqrt{\hat{\sigma}^2\hat{V}_c}|t_{\alpha/2}(N-K)|, \sum_{j=1}^K c_j\hat{\beta}_j + \sqrt{\hat{\sigma}^2\hat{V}_c}|t_{\alpha/2}(N-K)|].$$

Typically α is set at 0.05.

4.3.2. meaning of a confidence level. For any particular data set, this interval either includes $\sum_{j=1}^K c_j\beta_j$ or does not include it. We say this is level $1-\alpha$ confidence interval in the sense if we keep using this interval for different data sets, holding $\mathbf{X}$constant, then about $1-\alpha$ of the times, the interval would include $\sum_{j=1}^K c_j\beta_j$.

4.3.3. why we choose the interval in the middle? Note that there are many intervals with the above properties. One way to justify the interval is to seek the shortest interval with the above property. Since the normal distribution has its peak at the mean, in order to make the 95% interval the shortest, we definitely need to include the area around the peak. The same consideration leads to the symmetric region around the mean.

4.3.4. An example. One can use the standard statistical package to conduct inference about the linear combination of the coefficients.

An example of OLS regression output

$$\begin{array}{rcccccc} \hat{y} = & -4.38 & + & 1.084x_1 & + & .0217x_2 \\ & (.47) & & (.060) & & (.0128) \\ N = 32, & & & R^2 = .218 & & \end{array}$$

97.5 percentile for t-distribution with 29 degrees of freedom is 2.045.
So the 95% confidence interval for the coefficient on x_1 can be computed by

$$1.084 \pm .060 \times 2.045$$

or $(.961, 1.21)$ and for the coefficient on x_2

$$.0217 \pm .0128 \times 2.045$$

or $(-.0045, .0479)$.

4.4. Hypothesis test about a linear combination of the linear regression coefficients

4.4.1. review of the statistical hypothesis testing procedure. Sometimes we are more interested in finding out if the parameter takes a particular value or a set of values or not. This inference problem is called the *hypothesis test*. Let T be a random vector used to conduct a hypothesis test and its probability model be indexed by θ. For a set A, the probability of $T \in A$ is thus denoted as $P_\theta(T \in A)$. Typically T is a function of underlying data rather than data themselves. Such a random vector is called a *test statistic*.

The hypothesis to be tested is called the *null hypothesis* and stated as $H_0 : \theta \in \Theta_0$. The set that is assumed to hold when the null hypothesis does not hold is called the *alternative hypothsis* and stated as $H_1 : \theta \in \Theta_1$ or $H_A : \theta \in \Theta_1$. We define the entirety of the parameter space under consideration as $\Theta = \Theta_0 \cup \Theta_1$ or $\Theta = \Theta_0 \cup \Theta_A$. We use Θ_1 to denote the set of alternative values.

We define a subset of the support of T, a *region of rejection*, R, where we reject the null hypothesis if and only if $T \in R$. The region R^c, may be called the region of acceptance, or the region of non-rejection.

When conducting a hypothesis test, we may make two types of errors. One is when in fact $\theta \in \Theta_0$ and infer that $\theta \in \Theta_1$ and the other is when in fact $\theta \in \Theta_1$ and infer that $\theta \in \Theta_0$. The first is called *Type I error* and the second is called *Type II error*.

The probability of Type I error is $P_\theta(T \in R)$ for $\theta \in \Theta_0$. The probability of Type II error is $P_\theta(T \in R^c)$ for $\theta \in \Theta_1$. Note that for $\theta \in \Theta_1$, $P_\theta(T \in R^c) = 1 - P_\theta(T \in R)$ so that the same function with which we studied Type I error, now defined over a different region, Θ_1, $P_\theta(T \in R)$ has the same information as that with the probability of Type II error. We call $P_\theta(T \in R)$ defined over Θ_1, the *power function*. It equals one minus the probability of Type II error.

Thus we can examine $P_\theta(T \in R)$ over Θ. When $\theta \in \Theta_0$,it gives the Type I error, and when $\theta \in \Theta_1$, it gives the power.

When Θ_0 is a singleton, the probability of Type I error, $\Pr_\theta(Z \in R)$ is a single value. But when Θ_0 is not a singleton, then $\Pr_\theta(Z \in R)$ generally takes on multiple values. When the probability of Type I error is always less than or equal to α, the test is called *significance level α test*. If a test is a level α test, it is also a level α' test for any α' which is bigger than α. The supremum of the Type I error probabilities over $\theta \in \Theta_0$ is called the *size* of the test.

Ideally, over $\theta \in \Theta_0$, we want to choose T and R such that $P_\theta(T \in R) = 0$ and over $\theta \in \Theta_1$, $P_\theta(T \in R) = 1$. This is not possible in general.

Current practice is to find a statistic T and a rejection region R to construct size α test, such as $\alpha = 0.05$, and have highest power over Θ_1 given the significance level.

This practice may be justified when the null hypothesis is very likely to hold. If not, this practice may not be reasonable.

To see this point, consider a decision theoretic framework. Let the utility of correctly accepting the null be u_{00}, incorrectly rejecting the null be u_{0A}, incorrectly accepting the null be u_{A0}, and correctly rejecting the null be u_{AA}. The decision

variable is $\delta(T) = 1\{T \in R^c\}$. Let the utility function $U(\delta, \theta)$ be

$$U(\delta,\theta) = \begin{cases} u_{00} & \text{if } \delta = 1 \text{ and } \theta \in \Theta_0 \\ u_{0A} & \text{if } \delta = 0 \text{ and } \theta \in \Theta_0 \\ u_{A0} & \text{if } \delta = 1 \text{ and } \theta \in \Theta_1 \\ u_{AA} & \text{if } \delta = 0 \text{ and } \theta \in \Theta_1. \end{cases}$$

Note that this formulation does not account for the distance of $\theta \in \Theta_1$ to Θ_0.

Then the expected utility is

$$\begin{cases} u_{00}P_\theta(T \in R^c) + u_{0A}P_\theta(T \in R) & \text{if } \theta \in \Theta_0 \\ u_{A0}P_\theta(T \in R^c) + u_{AA}P_\theta(T \in R) & \text{if } \theta \in \Theta_1. \end{cases}$$

Rewriting this in terms of the Type I error and the power, we have

$$\begin{cases} u_{00} + (u_{0A} - u_{00})P_\theta(T \in R) & \text{if } \theta \in \Theta_0 \\ u_{A0} + (u_{AA} - u_{A0})P_\theta(T \in R) & \text{if } \theta \in \Theta_1. \end{cases}$$

Observe that since it is usually better to be correct than wrong, $u_{0A} - u_{00} < 0$ and $u_{AA} - u_{A0} > 0$ are reasonable assumptions to make. Under these assumptions, we want to choose the Type I probability $P_\theta(T \in R)$ as low as possible over $\theta \in \Theta_0$ and power $P_\theta(T \in R)$ as high as possible over $\theta \in \Theta_1$.

Generally we cannot set $P_\theta(T \in R) = 0$ over $\theta \in \Theta_0$ and $P_\theta(T \in R) = 1$ over $\theta \in \Theta_1$, nor can we find the common test statistic and the rejection region for the two optimization problem, so we need to consider the trade-off.

In order to consider the trade-off, let $\pi(\theta)$ be the prior probability density over Θ. Then the expected utility for T and R is

$$\begin{aligned} u_{00} + (u_{0A} - u_{00}) &\int_{\theta\in\Theta_0} P_\theta(T \in R)\pi(\theta)d\theta \\ &+ u_{A0} + (u_{AA} - u_{A0}) \int_{\theta\in\Theta_1} P_\theta(T \in R)\pi(\theta)d\theta \\ = u_{00} + (u_{0A} - u_{00}) &\int_{\theta\in\Theta_0} \pi(\theta)d\theta \int_{\theta\in\Theta_0} P_\theta(T \in R)\frac{\pi(\theta)}{\int_{\theta\in\Theta_0}\pi(\theta)d\theta}d\theta \\ + u_{A0} + (u_{AA} - u_{A0}) &\int_{\theta\in\Theta_1} \pi(\theta)d\theta \int_{\theta\in\Theta_1} P_\theta(T \in R)\frac{\pi(\theta)}{\int_{\theta\in\Theta_1}\pi(\theta)d\theta}d\theta. \end{aligned}$$

If

$$\frac{(u_{0A} - u_{00})\int_{\theta\in\Theta_0}\pi(\theta)d\theta}{(u_{AA} - u_{A0})\int_{\theta\in\Theta_1}\pi(\theta)d\theta}$$

is large negative value, then we will emphasize more on choosing smaller Type I error than power. This will be the case if $\int_{\theta\in\Theta_0}\pi(\theta)d\theta$ is close to one. This is especially so if the loss of making Type I error is much bigger than the gain from not making Type II error so that

$$\frac{u_{0A} - u_{00}}{u_{AA} - u_{A0}}$$

is much smaller than minus one. On the other other hand, if

$$\frac{(u_{0A} - u_{00})\int_{\theta\in\Theta_0}\pi(\theta)d\theta}{(u_{AA} - u_{A0})\int_{\theta\in\Theta_1}\pi(\theta)d\theta}$$

is close to zero, then there is no reason we should make the Type I error so small if power can be made large. For example, in testing the gender equality in wage, our prior probability that the null hypothesis of gender equality actually holds may be close to zero and also it may not be so unreasonable to assume that $u_{00} - u_{0A}$ is about the same with $u_{AA} - u_{A0}$.

Generally we should adjust the significance level and the power of a test, taking into accout the cost of making Type I and Type II errors and the prior belief about the parameter values. However, this is typically not done.

4.4.2. hypothesis test about a linear combination of parameters in the linear regression model. In the linear regression model, under Assumptions OLS.1–OLS.5, $\theta = (\beta, \sigma)$ and $\Theta = R^K \times R_+$, where R_+ is a strictly positive real line.

We consider the null hypothesis of the form $\sum_{j=1}^K c_j \beta_j = a$. The null hypothesis is

$$H_0 : \sum_{j=1}^K c_j \beta_j = a.$$

Along with the null hypothesis, we need to consider the alternative hypothesis when the null hypothesis does not hold. In the context of testing the equality of a linear combination of the regression coefficients to be equal to a given value, an alternative assumption can be $\sum_{j=1}^K c_j \beta_j \neq a$ or $\sum_{j=1}^K c_j \beta_j < a$ or $\sum_{j=1}^K c_j \beta_j > a$. The first is called two sided alternative and the latter two are called one sided alternatives. They are written as

$$H_1 : \sum_{j=1}^K c_j \beta_j \neq a$$

or

$$H_1 : \sum_{j=1}^K c_j \beta_j < a$$

or

$$H_1 : \sum_{j=1}^K c_j \beta_j > a.$$

The idea of the statistical hypothesis testing procedure is analogous to the proof by a contradiction. We suppose the null hypothesis to hold and examine if the test statistic takes on a "value cosistent with the null hypothesis." If it takes on an "unlikely value under the null hypothesis", (lies in the "rejection region") then the null hypothesis is rejected as it is "contradicting the null hypothesis."

The null hypothesis is usually a conventional wisdom or the effect of a new procedure to be 0, so that unless there is a strong evidence against it, we don't reject it. That is, we do not wish to reject the null hypothesis when in fact the null hypothesis holds.

4.4.3. how to choose the rejection region. Under the null hypothesis $\sum_{j=1}^K c_j \beta_j = a$, our earlier result implies,

$$\Pr\left\{ t_{\alpha/2}(N-K) < \frac{\sum_{j=1}^K c_j \hat{\beta}_j - a}{\sqrt{\hat{\sigma}^2 \hat{V}_c}} < |t_{\alpha/2}(N-K)| \mid X \right\}$$

equals $1-\alpha$. So the probability that

$$\left|\frac{\sum_{j=1}^{K} c_j\hat{\beta}_j - a}{\sqrt{\hat{\sigma}^2\hat{V}_c}}\right| > |t_{\alpha/2}(N-K)|$$

occur is α, a small probability. The region is called the rejection region against $H_1 : \sum_{j=1}^{K} c_j\beta_j \neq a$ with the significance level α.

We reject the null hypothesis when this inequality holds. As discussed above, typically α is set at 0.05 but sometimes 0.01 or 0.1 is used.

4.4.4. meaning of a significance level. Clearly, for a particular data set, this inequality holds or does not hold. The significance level indicates the fraction of the times we would make the Type I eror, rejecting the null hypothesis when in fact the hypothesis holds, if we repeat this with many different data sets.

4.4.5. how the shape of the rejection region is chosen? There are many other intervals or sets of intervals with the same significance level. The rejection region is chosen in order to maximize the power of the test, especially for parameter values far away from the true parameter.

For example for the rejection region we just studied, the probability of rejecting the null hypothesis when $\sum_{j=1}^{K} c_j\beta_j = a'$, where $a' \neq a$, equals

$$\Pr\left\{\left|\frac{\sum_{j=1}^{K} c_j\hat{\beta}_j - a}{\sqrt{\hat{\sigma}^2\hat{V}_c}}\right| > |t_{\alpha/2}(N-K)| \mid X\right\}.$$

It is not equal to α but something bigger when $\sum_{j=1}^{K} c_j\beta_j = a'$, where $a' \neq a$. In order to compute the power of a test in this context, we use the non-central t-distribution explained below.

4.4.6. noncentral t-distribution with degrees of freedom m and non-central parameter δ. Recall that

$$t'(m,\delta) = \frac{N(0,1)-\delta}{\sqrt{\chi^2(m)/m}},$$

where $N(0,1)$ and $\chi^2(m)$ are independent, then $t'(m,\delta)$ has the non-central t-distribution with degrees of freedom m and the noncentrality parameter δ. Non-centered t-distribution is asymmetric unless $\delta = 0$, with heavier tail toward the direction of the non-centrality parameter. For the same non-centrality parameter, the skewness is mitigated when the degrees of freedom is bigger. We use the non-central t-distribution to examine the power function of the test statistic.

When $\sum_{j=1}^{K} c_j\beta_j = a'$, the t-distribution is not centered at 0. In this case $(a'-a)/\sqrt{\sigma^2\hat{V}_c}$ is the non-centrality parameter.

As one can see, the power is a function of the true parameter. As a function of the true parameter, it is called the power function.

The rejection region is set up so that the power is higher for the parameter values farther away from the null hypothesis.

In particular, if we believe that when the null hypothesis does not hold, the parameter is greater than the null value a. Then, we only care about the power in that direction. The same is true if we think the parameter is less than the null value if the null hypothesis does not hold.

This yields the one-sided tests. For example if the alternative to the null hypothesis $\sum_{j=1}^{K} c_j\beta_j = a$ is $\sum_{j=1}^{K} c_j\beta_j > a$, then the rejection region is (for $\alpha < 0.5$)

$$\frac{\sum_{j=1}^{K} c_j\hat{\beta}_j - a}{\sqrt{\hat{\sigma}^2\hat{V}_c}} > |t_\alpha(N-K)|$$

and when the alternative to the null hypothesis $\sum_{j=1}^{K} c_j\beta_j = a$ is $\sum_{j=1}^{K} c_j\beta_j < a$, then the rejection region is

$$\frac{\sum_{j=1}^{k} c_j\hat{\beta}_j - a}{\sqrt{\hat{\sigma}^2\hat{V}_c}} < t_\alpha(N-K).$$

The tests using one-sided rejection region are called *one-sided test* and the tests using the two-sided rejection region are called *two-sided test*.

Instead of conducting a hypothesis test using a particular significance level, researchers may prefer to report the so called *p-value*. It is the smallest significance level at which the hypothesis is rejected with the current data. Thus a smaller p-value is interpreted to correspond to a stronger evidence against the null hypothesis under consideration. The p-value allows the result to be conveyed without a reporter committing to a particular significance level a priori.

4.4.7. One-sided alternatives: An example.

$$\begin{array}{lllllllll} \widehat{\log(wage)} = & .284 & + & .092educ & + & .0041exper & + & .022tenure \\ & (.104) & & (.007) & & (.0017) & & (.003) \\ n = 526, & & & R^2 = .316 & & & & \end{array}$$

H_0: $\beta_{exper} = 0$ versus H_1: $\beta_{exper} > 0$.

$$t_{exper} = .0041/.0017 \approx 2.41.$$

5% critical value is 1.645, 1% critical value is 2.326. Thus the p-value is lower than 1%.

4.4.8. statistical significance and significance in reality. One needs to be careful to distinguish statistical significance and significance in reality. If the standard error is very small, then we would not reject the 0 coefficient hypothesis even if the OLS estimate itself is very small from economic perspective because t-value is computed using

$$\hat{\beta}_j/[\text{standard error of } \hat{\beta}_j]$$

and the magnitude of this is compared against the critical value. If we forget to consider the magnitude of the effect under discussion, we may misunderstand an empirical evidence.

4.4.9. computing standard error of a linear combination of coefficients. Here we describe how to use the standard statistical package to compute standard error of a linear combination of the OLS estimator of the linear regression model's coefficients.

This is useful in conducting hypothesis test as well. A modern statistical package, such as Stata, allows you to compute the test statistics directly, so that this is not useful once you know how to use a modern statistical package. However,

following the logic below helps you enhance your understanding of the working of the OLS estimation, generally.

The idea is to rearrage the variables so that the linear combination we want to examine appear as the coefficient on one variable.

For example, consider $\hat{\beta}_1 + \hat{\beta}_2$ in the linear regression model:

$$y_i = \hat{\alpha} + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{u}_i.$$

Then

$$\begin{aligned} y_i &= \hat{\alpha} + (\hat{\beta}_1 + \hat{\beta}_2 - \hat{\beta}_2) x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{u}_i \\ &= \hat{\alpha} + (\hat{\beta}_1 + \hat{\beta}_2) x_{i1} + \hat{\beta}_2 (x_{i2} - x_{i1}) + \hat{\beta}_3 x_{i3} + \hat{u}_i \end{aligned}$$

and that clearly $\hat{u}_i$ and 1, x_{i1} and $x_{i2} - x_{i1}$ are orthogonal so that so that the OLS estimate of the coefficient on x_{i1} one obtains by regressing y_i on a constant term, x_{i1}, $x_{i2} - x_{i1}$, and x_{i3} and $\hat{\beta}_1 + \hat{\beta}_2$ are the same.

For another example, consider $2\hat{\beta}_1 + \hat{\beta}_2$ in the same model as above. Then

$$\begin{aligned} y_i &= \hat{\alpha} + (2\hat{\beta}_1 + \hat{\beta}_2 - \hat{\beta}_2) x_{i1}/2 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{u}_i \\ &= \hat{\alpha} + (2\hat{\beta}_1 + \hat{\beta}_2) x_{i1}/2 + \hat{\beta}_2 (x_{i2} - x_{i1}/2) + \hat{\beta}_3 x_{i3} + \hat{u}_i \end{aligned}$$

so that the OLS estimate of the coefficient on x_{i1} one obtains by regressing y_i on a constant term, x_{i1}, $x_{i2} - x_{i1}/2$, and x_{i3} gives the result.

4.5. Confidence region and Hypothesis test about multiple linear combinations of the regression coefficients

In some cases, we want to test multiple linear combinations of the regression coefficients. For example, in order to test if "non-cognitive skills" do not affect test scores, all coefficients of variables capturing the "non-cognitive skills" need to be tested to be zero at once. Even when only one variable is involved, if there are non-linear terms included in a regression in addition to a linear term, all coefficients involving the variables need to be tested to be zero at once.

For example consider a wage regression:

$$\begin{aligned} \log w_i = \beta_1 + \beta_2 education_i + \beta_3 education_i^2 + \beta_4 experience_i \\ + \beta_5 experience_i^2 + \beta_6 education_i \cdot experience_i + \cdots + u_i \end{aligned}$$

In this case, we want to test if experience does not affect wage, then need to consider OLS estimators of β_4, β_5, and β_6 at the same time and test if all are simultaneously zero. We have seen that generally, $\hat{\beta}$ given $\mathbf{X}$is distributed $N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$ so that the three dimensional sub-vector $(\hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6)'$ given $\mathbf{X}$, has a joint distribution of the three dimensional normal random vector.

4.5.1. Wald approach. Consider an $r \times K$ constant matrix C, where $r \leq K$ and that C has full row rank. The assumption that C has full row rank means there is no redundant linear combination that is implied by other linear combinations. In general, the linear combinations $C\hat{\beta}$ is distributed $N(C\beta, \sigma^2 C(\mathbf{X}'\mathbf{X})^{-1}C')$.

Recall that when we study a single linear combination $c'\hat{\beta}$, we divided by the square root of the conditional variance of $c'\hat{\beta}$ to show that

$$\frac{c'\hat{\beta} - c'\beta}{\sqrt{\sigma^2 c'(\mathbf{X}'\mathbf{X})^{-1}c}}$$

is the standard normal random variable. To obtain an analogous result, we consider a square root of a matrix $\sigma^2 C(\mathbf{X}'\mathbf{X})^{-1}C'$. This is provided by the Cholesky decomposition.

THEOREM 4.1. *If a matrix Ω is positive semi-definite, then there is a lower triangular matrix Γ, such that $\Omega = \Gamma\Gamma'$.*

This decomposition is called Cholesky decomposition and Γ is called the Cholesky factor of Ω. Moreover, one can show that for any invertible matrix A, $(A^{-1})' = (A')^{-1}$. We use these results below.

Let Γ be the Cholesky factor of $\sigma^2 C(\mathbf{X}'\mathbf{X})^{-1}C'$. Then

$$\Gamma^{-1}(C\hat{\beta} - C\beta) \sim N(0, \Gamma^{-1}(\Gamma\Gamma')(\Gamma^{-1})') = N(0, \Gamma'(\Gamma')^{-1}) = N(0, I).$$

Therefore

$$\begin{aligned}
&[\Gamma^{-1}(C\hat{\beta} - C\beta)]'[\Gamma^{-1}(C\hat{\beta} - C\beta)] \\
&\quad = (C\hat{\beta} - C\beta)'(\Gamma^{-1})'\Gamma^{-1}(C\hat{\beta} - C\beta) = (C\hat{\beta} - C\beta)'(\Gamma')^{-1}\Gamma^{-1}(C\hat{\beta} - C\beta) \\
&= (C\hat{\beta} - C\beta)'(\Gamma\Gamma')^{-1}(C\hat{\beta} - C\beta) = (C\hat{\beta} - C\beta)'[\sigma^2 C(\mathbf{X}'\mathbf{X})^{-1}C']^{-1}(C\hat{\beta} - C\beta).
\end{aligned}$$

Being the sum of squares of r independent standard nornal random variables, this has the Chi-square distribution with degrees of freedom r.

Recall that the density of the multivariate normal random vector Z of length $d$with mean μ and variance Σ, evaluated at z is

$$\frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}} \exp[-(z-\mu)'\Sigma^{-1}(z-\mu)/2].$$

Thus, examining $(C\hat{\beta} - C\beta)'[\sigma^2 C(\mathbf{X}'\mathbf{X})^{-1}C']^{-1}(C\hat{\beta} - C\beta)$ corresponds to examining the iso-height values of the joint normal density of $C\hat{\beta}$. This observation can be used to think about why the confidence region or the rejection region are chosen in certain ways.

Just like the single equality case, the transformation is not completely known because σ^2 is not known. Like in that case, we consider taking the ratio of the term just examined divided by the degrees of freedom and $\hat{\sigma}^2/\sigma^2$. Taking the ratio, σ^2 is cancelled and results in an object which looks like the original term, except that σ^2 is replaced by $\hat{\sigma}^2$. Recall that $\hat{\sigma}^2/\sigma^2|\mathbf{X}$ is the chi-square random variable with $N-K$ degrees of freedom divided by its degrees of freedom. As we saw, these two terms are independent. Thus

$$(C\hat{\beta} - C\beta)'[\hat{\sigma}^2 C(\mathbf{X}'\mathbf{X})^{-1})C']^{-1}(C\hat{\beta} - C\beta)/r$$

conditional on $\mathbf{X}$, has the F-distribution with degrees of freedom equal to the number of equalities under study and $N-K$. The approach described is usually referred to as the "Wald approach".

4.5.2. likelihood ratio and lagrangean multiplier approaches. While focusing on the difference of the estimate and the true parameter is a natural way to construct confidence region and conduct hypothesis tests that involve single or multiple equalities, there are two other approaches for conducting hypothesis tests; the "Likelihood Ratio (LR)" approach and the "Lagrangean Multiplier (LM)" approaches.

[graph of the three tests]

Graphically, the Wald approach compares the estimate with the null value and asks if the distance is small statistically. The LR approach compares the value of the objective function at the estimate with that at the null value and asks if the difference in the values is small statistically. The LM approach examines the slope of the objective function under the null hypothesis and asks if the slope is statistically different from zero.

For the null hypothesis on the coefficients in the linear regression model, these three tests coincide when appropriate way to estimate σ^2 is used.

We describe the LR approach in the context of testing the coefficients in the linear regression model first. In order to define the LR approach we consider the least square estimate under the null hypothesis:

$$\min_{b\in\{b\in R^K|Cb=A\}} \sum_{i=1}^{N}(y_i - x_i'b)^2$$

Let $\hat{\beta}_R$ be the solution to this problem and define $\hat{\mathbf{U}}_R = \mathbf{Y} - \mathbf{X}\hat{\beta}_R$ and

$$SSR_R = \hat{\mathbf{U}}_R'\hat{\mathbf{U}}_R \qquad SSR_{UR} = \hat{\mathbf{U}}'\hat{\mathbf{U}},$$

where $\hat{\mathbf{U}}$is the OLS residual vector. Then the LR-test is to compare how much increase in the sum of squared residuals we would see per restriction under the null hypothesis relative to the overall variance:

$$F = \frac{(SSR_R - SSR_{UR})/r}{SSR_{UR}/(N-K)} = \frac{(SSR_R - SSR_{UR})/r}{\hat{\sigma}^2}.$$

It turns out that this is the same statistics with the Wald statistic.

To see this, we derive the restricted OLS estimator. Setting up the Lagrangian problem,

$$\mathcal{L} = \frac{1}{2}(\mathbf{Y} - \mathbf{X}b)'(\mathbf{Y} - \mathbf{X}b) - \lambda'(Cb - A),$$

the first order conditions are

$$\begin{aligned}\mathbf{X}'(\mathbf{Y} - \mathbf{X}b) - C'\lambda &= 0,\\ Cb &= A.\end{aligned}$$

Solving for b in terms of λ from the first set of equations and substituting into the second set of equations, we obtain

$$\hat{\lambda} = [C(\mathbf{X}'\mathbf{X})^{-1}C']^{-1}(C\hat{\beta} - A),$$

from which we obtain

$$\hat{\beta}_R = \hat{\beta} - (\mathbf{X}'\mathbf{X})^{-1}C'\hat{\lambda}.$$

This implies that

$$\begin{aligned}\hat{\mathbf{U}}_R &= \mathbf{Y} - \mathbf{X}\hat{\beta}_R\\ &= \mathbf{Y} - \mathbf{X}\hat{\beta} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}C'\hat{\lambda}\\ &= \hat{\mathbf{U}} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}C'\hat{\lambda}\end{aligned}$$

and since $\mathbf{X}'\hat{\mathbf{U}} = 0$,

$$\begin{aligned}\hat{\mathbf{U}}_R'\hat{\mathbf{U}}_R &= \hat{\mathbf{U}}'\hat{\mathbf{U}} + \hat{\lambda}'C(\mathbf{X}'\mathbf{X})^{-1}C'\hat{\lambda}\\ &= \hat{\mathbf{U}}'\hat{\mathbf{U}} + (C\hat{\beta} - A)'[C(\mathbf{X}'\mathbf{X})^{-1}C']^{-1}(C\hat{\beta} - A).\end{aligned}$$

Therefore, using the first equality, we see that

$$\frac{(\hat{\mathbf{U}}_R'\hat{\mathbf{U}}_R - \hat{\mathbf{U}}'\hat{\mathbf{U}})/r}{\hat{\mathbf{U}}'\hat{\mathbf{U}}/(N-K)} = \frac{\hat{\lambda}'C(\mathbf{X}'\mathbf{X})^{-1}C'\hat{\lambda}/r}{\hat{\sigma}^2},$$

and using the second equality, we see that

$$\frac{(\hat{\mathbf{U}}_R'\hat{\mathbf{U}}_R - \hat{\mathbf{U}}'\hat{\mathbf{U}})/r}{\hat{\mathbf{U}}'\hat{\mathbf{U}}/(N-K)} = \frac{(C\hat{\beta}-A)'[C(\mathbf{X}'\mathbf{X})^{-1}C']^{-1}(C\hat{\beta}-A)/r}{\hat{\sigma}^2}.$$

Therefore LM-test, Wald-test, and the LR-test, in this context, if we use $\hat{\sigma}^2$ in place of σ^2, will be numerically the same.

Note that LM-test can be carried out by just using the restricted estimator whereas

rather than

$$\tilde{\sigma}^2 = \frac{1}{N-K+r}\sum_{i=1}^{N}(y_i - x_i'\hat{\beta}_R)^2,$$

which is the estimator of σ^2 under the null hypothesis.

4.5.3. using R^2 for the unrestriced and restricted regressions to compute the F-statistic. One can use R^2 from the restricted and unrestricted regressions to compute the F statistics.

To see this, note that $R^2 = 1 - SSR/SST$ so that $SSR_R = SST(1-R_R^2)$ and $SSR_{UR} = SST(1-R_{UR}^2)$.

Substituting this into the formula gives

$$\begin{aligned}\frac{(SSR_R - SSR_{UR})/r}{SSR_{UR}/(N-K)} &= \frac{(SST(1-R_R^2) - SST(1-R_{UR}^2))/r}{SST(1-R_{UR}^2)/(N-K)} \\ &= \frac{(R_{UR}^2 - R_R^2)/r}{(1-R_{UR}^2)/(N-K)}.\end{aligned}$$

A special case is to test all coefficients other than the constant term are 0. In this case, since $R_R^2 = 0$,

$$\frac{(SSR_R - SSR_{UR})/r}{SSR_{UR}/(N-K)} = \frac{R_{UR}^2/(K-1)}{(1-R_{UR}^2)/(N-K)}.$$

indicate what to do when the assumptions do not hold.

Other points on the finite sample properties related to the OLS estimator

Other finite sample properties of statistics related to the OLS estimator

OLS estimator is the Best Linear Unbiased Estimator (BLUE) under all the assumptions except for the normality assumption.

The claim is that any linear combination of the regression coefficient can be estimated with the smallest conditional variance by the same linear combination of the OLS estimator if we restrict the estimators in the class of conditionally unbiased estimators that are linear combination of the dependent variables.

Here, "best" refers to having the smallest conditional variance.

OLS estimator can be interpreted as a Maximum Likelihood Estimator under normality assumption.

We have used $\hat{\sigma}^2$ as the estimator of σ^2, but have not investigated it's properties.

One can show that it is an unbiased estimator given X, and its conditional variance can be computed also, as shown in the lecture note.

The influence of the i-th observation can be computed exactly for the linear regression estimator:

$$\hat{\beta} - \hat{\beta}_{(i)} = (X'X)^{-1}x_i\hat{u}_i/(1 - x_i(X'X)^{-1}x_i).$$

This result is obtained using the following result: when X is full column rank and A is invertible,

$$(A - XX')^{-1} = A^{-1} + A^{-1}X(I - X'A^{-1}X)X'A^{-1}.$$

We apply this to $A = \mathbf{X}'\mathbf{X}$ and $X = x_i$. Then, denoting the regressor matrix excluding the ith observation by $\mathbf{X}_{(i)}$,

$$(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1} = (\mathbf{X}'\mathbf{X} - x_ix_i')^{-1} = (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}x_ix_i'(\mathbf{X}'\mathbf{X})^{-1}}{1 - x_i'(\mathbf{X}'\mathbf{X})^{-1}x_i}.$$

From this we obtain

$$\begin{aligned}\hat{\beta}_{(i)} &= \hat{\beta} - (\mathbf{X}'\mathbf{X})^{-1}x_iy_i + \frac{(\mathbf{X}'\mathbf{X})^{-1}x_ix_i'[\hat{\beta} - (\mathbf{X}'\mathbf{X})^{-1}x_iy_i]}{1 - x_i'(\mathbf{X}'\mathbf{X})^{-1}x_i} \\ &= \hat{\beta} - \frac{(\mathbf{X}'\mathbf{X})^{-1}x_i(y_i - x_i'\hat{\beta})}{1 - x_i'(\mathbf{X}'\mathbf{X})^{-1}x_i} = \hat{\beta} - \frac{(\mathbf{X}'\mathbf{X})^{-1}x_i\hat{u}_i}{1 - x_i'(\mathbf{X}'\mathbf{X})^{-1}x_i}.\end{aligned}$$

This result is useful to compute $\hat{\beta}_{(i)}$ because we do not have to invert $\mathbf{X}'_{(i)}\mathbf{X}_{(i)}$ for different $i = 1, \ldots, N$. On the other hand, the formula $\mathbf{X}'\mathbf{X}$ contains the ith observation, so that the effect of the ith observation is not completely separated from the rest of the observations.

To obtain the formula which separates the effect of the ith observation from the rest of the observations, we use

$$(A + XX')^{-1} = A^{-1} - A^{-1}X(I + X'A^{-1}X)^{-1}X'A^{-1}.$$

We apply this to $A = \mathbf{X}'_{(i)}\mathbf{X}_{(i)}$ and $X = x_i$. Then

$$(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{X}'_{(i)}\mathbf{X}_{(i)} + x_ix_i')^{-1} = (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1} - \frac{(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}x_ix_i'(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}}{1 + x_i'(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}x_i}$$

so that

$$\hat{\beta} = (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}\mathbf{Y} - (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\frac{x_ix_i'(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}\mathbf{Y}}{1 + x_i'(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}x_i}.$$

Let the OLS estimator obtained using all data but the ith observation be $\hat{\beta}_{(i)}$ and the dependent variable data vector exluding the ith observation be $\mathbf{Y}_{(i)}$. Then $\mathbf{X}\mathbf{Y} = \mathbf{X}_{(i)}\mathbf{Y}_{(i)} + x_iy_i$ and

$$\begin{aligned}\hat{\beta} &= \hat{\beta}_{(i)} + (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}x_iy_i - (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\frac{x_ix_i'[\hat{\beta}_{(i)} + (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}x_iy_i]}{1 + x_i'(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}x_i} \\ &= \hat{\beta}_{(i)} + (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}x_iy_i - (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\frac{x_i[x_i'\hat{\beta}_{(i)} + x_i'(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}x_iy_i]}{1 + x_i'(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}x_i} \\ &= \hat{\beta}_{(i)} + (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}x_i\frac{y_i - x_i'\hat{\beta}_{(i)}}{1 + x_i'(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}x_i}.\end{aligned}$$

This implies

$$x_i\hat{\beta} - x_i'\hat{\beta}_{(i)} = \frac{x_i'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}x_i}{1 + x_i'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}x_i}(y_i - x_i'\hat{\beta}_{(i)})$$

The ith observation has a large effect on the predicted value $x_i'\hat{\beta}$, if the predicted value of y_i using the rest of the data and x_i, by $x_i'\hat{\beta}_{(i)}$, differs from the realized y_i and especially so when $x_i'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}x_i$ is large. This happens when x_i is proportional to the eigen-vector corresponding to the largest eigen-value of $(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}$. And it corresponds to the situation wheren $x_i'\beta$ is estimated with high variance by the OLS using the rest of the observations, holding the conditional variance of the dependent variable constant.

4.6. Karlan and Zinman (2009 Econometrica)

4.6.1. background. Karlan and Zinman (Econometrica 2009) conducted a randomized experiment in South Africa using 57,533 former clients with good repayment histories from a loan company.

Recall that there are two types of problems with informational asymmetry in the credit market: adverse selection and moral hazard. Adverse selection arises when a higher interest rate leads to a riskier pool of borrowers. Moral hazard arises when, for the same risk types, less effort is put in from preventing default.

Karlan and Zinman examines, by the randomized experiment, importance of informational asymmetry in the small (median loan size is \$150, (32% of borrowers' median gross monthly income)), high interest (7.75 to 11.75% for 4 months (typical cash lenders charge 30% per month for observably highest-risk group and 3% per month for observably lower-risk group)), short term (90% is 4 months loan), uncollateralized consumer credit with a fixed monthly repayment schedule to a working poor population in South Africa. Average default rate of repeat and first time borrowers are 15% and 30%, respectively.

4.6.2. experiment. The individual is sent an offer with rate r^o (randomly given) with the repeat rate specified as r.

He/She decides whether to borrow at the offer rate r^o assuming the repeat rate will be r. Because the offered rate is randomly chosen, those faced with different offered rate should be basically the same population. If persons who are offered with higher rates default with higher rates, it is due to the adverse selection.

After the borrower agreed to the offered rate, the offer rate is lowered randomly to $r^c < r^o$ for some borrowers and for some borrowers the repeat rate is also lowered to $r^f = r^c < r$. Because the offered rate and contract rate are randomly chosen, those faced with different offered rates and contract rates/repeat rates should be basically the same population.

Given the contract rate and the repeat rate, borrowers decide how much effort to put in. Lower future repeat rate provides an additional incentive not to default.

Project return is realized and borrowers decide whether to repay the loan.

By looking at applicants for different offered rates and subsequent differences in their default rates reveal the extent of adverse selection.

By examining the applicants for the same offered rate but with different contract rates and repeat rates reveal the extent of moral hazard.

4.6.3. empirical results. The equation used to estimate the effect is

$$Y_i = \alpha + \beta_o r_i^o + \beta_c r_i^c + \beta_b C_i + X_i\gamma + \epsilon_i.$$

C_i is a binary variable taking value 1 if offered a better repeat rate on future loans conditional on repayment and 0 otherwise, and X_i is a vector of observables such as risk-type, bank branch, and the month in which the solicitation letter was sent.

The results are summarized in Table I:

TABLE I

EMPIRICAL TESTS OF HIDDEN INFORMATION AND HIDDEN ACTION: FULL SAMPLE

	OLS							
Dependent Variable:	*Monthly Average Proportion Past Due*		*Proportion of Months in Arrears*		*Account in Collection Status*		*Standardized Index of Three Default Measures*	
Mean of Dependent Variable:	0.09	0.09	0.22	0.22	0.12	0.12	0	0
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Contract rate (Hidden Action Effect 1)	0.005	0.002	0.006*	0.002	0.001	−0.001	0.014	0.004
	(0.003)	(0.004)	(0.003)	(0.004)	(0.005)	(0.005)	(0.011)	(0.013)
Dynamic repayment incentive dummy (Hidden Action Effect 2)	−0.019*	−0.000	−0.028**	0.004	−0.025**	−0.004	−0.080**	−0.000
	(0.010)	(0.017)	(0.011)	(0.021)	(0.012)	(0.020)	(0.032)	(0.057)
Dynamic repayment incentive size		−0.005		−0.009**		−0.006		−0.023*
		(0.004)		(0.004)		(0.005)		(0.013)
Offer rate (Hidden Information Effect)	0.005	0.004	0.002	0.002	0.007	0.007	0.015	0.015
	(0.003)	(0.003)	(0.003)	(0.004)	(0.005)	(0.005)	(0.011)	(0.012)
Observations	4348	4348	4348	4348	4348	4348	4348	4348
Adjusted *R*-squared	0.08	0.08	0.14	0.15	0.06	0.06	0.10	0.11
Probability(both dynamic incentive variables = 0)		0.06		0.00		0.06		0.01
Probability(all 3 or 4 interest rate variables = 0)	0.0004	0.0005	0.0003	0.0012	0.0006	0.0016	0.0000	0.0001

*significant at 10%; **significant at 5%; ***significant at 1%. Each column presents results from a single OLS model with the RHS variables shown and controls for the randomization condtions: observable risk, month of offer letter, and branch. Adding loan size and maturity as additional controls does not change the results. Robust standard errors in parentheses are corrected for clustering at the branch level. "Offer rate" and "Contract rate" are in monthly percentage point units (7.00% interest per month is coded as 7.00). "Dynamic repayment incentive" is an indicator variable equal to one if the contract interest rate is valid for one year (rather than just one loan) before reverting back to the normal (higher) interest rates. "Dynamic repayment incentive size" interacts the above indicator variable with the difference between the lender's normal rate for that individual's risk category and the experimentally assigned contract interest rate. A positive coefficient on the Offer Rate variable indicates hidden information, a positive coefficient on the Contract Rate or Dynamic Repayment Incentive variables indicates hidden action (moral hazard).
The dependent variable in columns (7) and (8) is a summary index of the three dependent variables used in columns (1)–(6). The summary index is the mean of the standardized value for each of the three measures of default.

Since people who applied for loan with different offered rates are potentially in different risk groups via adverse selection, it is more desirable to examine the moral hazard issues for each of the different offered rates.

This amounts to studying $\varphi(r_i^o, r_i^c, C_i, type_i)$, where $type_i$ denotes observable risk types.

This formulation ignores bank branch information which is related with the local economic condition. Perhaps one can justify using dummy variables for each of the branches and waves for this reason.

Denoting a set of such dummy variables by x_i, we have:

$$y_i = x_i'\beta + \varphi(r_i^o, r_i^c, C_i, type_i) + u_i.$$

CHAPTER 5

Asymptotic Analysis I

The objective of this chapter is to review elements of large sample theories. Useful references are

- *Amemiya, T. (1985): Advanced Econometrics, Harvard University Press. Chapters 3.
- R. Serfling (1980): Approximation Theorems of Mathematical Statistics, Wiley. Chapters 1 and 5.
- Newey, W. K. and D. L. McFadden (1994) "Large sample estimation and hypothesis testing, in the Handbook of Econometrics," Vol. 4, ed. by R.F. Engle and D.L. McFadden. Amsterdam: North-Holland.
- Van der Vaart, A. W. (1998) Asymptotic Statistics, Cambridge University Press. Chapters 4, 5, 10, and section 25.10
- Van der Vaart, A. W. and Jon A. Wellner (1996) Weak Convergence and Empirical Processes, Springer. Sections 3.1–3.4.

I recommend studying Amemiya's textbook chapter 3 as you read this note and solving all the questions at the end of the chapter. Other references include more advanced topics than this lecture note.

The three leading uses of the asymptotic results are: (1) Understand general conditions under which an estimator has desirable properties such as consistency and asymptotic normality. (2) Make approximate statistical inferences or conduct hypothesis testing. For the standard case of asymptotic normality, a scalar case uses the normal random variable as an approximation, a vector case uses chi-squared distribution with an appropriate degrees of freedom as an approximation. Make sure you can construct confidence intervals/confidence regions and conduct hypothesis testing using these asymptotic results. (3) Compare different estimators.

5.1. Elements of large sample theories

We first review some key concepts.

5.1.1. Random vector.

DEFINITION 5.1 (σ-algebra). A class of subsets of Ω, denoted $\mathcal{S}$, is called a σ-algebra, if (i) Ω and $\varnothing \in \mathcal{S}$, (ii) whenever $E \in \mathcal{S}$, $E^c \in \mathcal{S}$, and (iii) whenever $E_1, \ldots, E_n, \ldots \in \mathcal{S}$, then $\cup_{j=1}^{\infty} E_j \in \mathcal{S}$.

REMARK 5.1. We will treat Ω as the totality of what could happen (sample space) and an element of $\mathcal{S}$, as an event for which we will define probability.

DEFINITION 5.2 (Probability Space). Probability space is a triplet $(\Omega, \mathcal{S}, \mathbb{P})$ where Ω is a sample space, $\mathcal{S}$ is a σ-algebra, and $\mathbb{P}$ is a mapping from $\mathcal{S}$ into $[0, 1]$ which satisfies

(1) $\mathbb{P}(A) \geq 0$ for any $A \in \mathcal{S}$,
(2) $\mathbb{P}(\Omega) = 1$,
(3) for any disjoint sets $E_1, ..., E_n$, $\mathbb{P}\left(\cup_{j=1}^{n} E_j\right) = \sum_{j=1}^{n} \mathbb{P}(E_j)$,
(4) $\lim_{n\to\infty} \mathbb{P}(E_n) = 0$ for any sequence of sets $\{E_j\}_{j=1}^{\infty}$ with $E_1 \supset E_2 \supset \cdots \supset E_n \supset \cdots$ and $\cap_{j=1}^{\infty} E_j = \varnothing$.

REMARK 5.2. The only unintuitive property is the fourth condition. It can be understood as a continuity requirement on $\mathbb{P}$. To see this, suppose that a sequence of sets converges to a set A in the sense that $E_j \supset A$ for any j and that $E_1 \supset E_2 \supset \cdots \supset E_n \supset \cdots$ and $\cap_{j=1}^{\infty} E_j = A$. In this case condition 4 implies that $\lim_{n\to\infty} \mathbb{P}(E_n \setminus A) = 0$. Since $\mathbb{P}(E_n) = \mathbb{P}(E_n \setminus A) + \mathbb{P}(A)$ this implies that $\lim_{n\to\infty} \mathbb{P}(E_n) = \mathbb{P}(A)$. So when E_n converges to A, the corresponding measure converges to that of A as well.

REMARK 5.3. When condition 4 holds, analogous result for any monotonically increasing sets also holds.

REMARK 5.4. If we define the probability function $F(t) = \mathbb{P}(X \leq t)$ for a random variable X, then one can show that it is right-continuous. Note that for any $t_n > t$,

$$F(t_n) = \mathbb{P}(X \leq t_n) = \mathbb{P}(X \leq t) + \mathbb{P}(t < X \leq t_n).$$

Since the set in the second term in the last expression converges to an empty set monotonically, when t_n declines to t, condition 4 implies that the probability converges to 0. Thus $F(t_n) \to F(t)$.

REMARK 5.5. On the other hand, if we define the probability function $F(t) = \mathbb{P}(X < t)$, then one can show that it is left-continuous. Note that for any $t_n < t$,

$$F(t_n) = \mathbb{P}(X < t_n) = \mathbb{P}(X < t) - \mathbb{P}(t_n \leq X < t).$$

Since the set in the second term on the last expression converges to an empty set monotonically, when t_n increases to t, condition 4 implies that the probability converges to 0. Thus $F(t_n) \to F(t)$.

DEFINITION 5.3 (Random Vector). A function X from Ω into $\mathbb{R}^k$ is a random vector in the probability space $(\Omega, \mathcal{S}, \mathbb{P})$ if for any Borel set B, $X^{-1}(B) \in \mathcal{S}$, i.e. if X is a Borel-measurable function.

REMARK 5.6. The smallest σ-algebra constructed based on open sets of a topological space is called the Borel sets. This particular σ-algebra is used to define measurability because typically we want any continuous functions to be measurable.

5.1.2. Four modes of convergence. Four modes of convergence we consider are convergence in probability, almost sure convergence, convergence in rth mean, and convergence in distribution.

We will use the modes of convergence to study asymptotic properties of various estimators. For this purpose it is useful to regard X_n as an estimator and ω as a data set. Thus $X_n(\omega)$ denotes a value of an estimator using n observations from data set ω.

Recall that a random vector is a function. So when we define a convergence of a sequence of random vectors, we need to define a convergence notion of a sequence of functions. Let $X_1, ..., X_n, ...$and X_∞ be random vectors defined on $(\Omega, \mathcal{S}, \mathbb{P})$ in the first three definitions below.

The first concept uses the probability that the converging function and the limit function are different by more than a given number as the measure of discrepancy between the two functions.

DEFINITION 5.4 (Convergence in Probability). $\{X_n\}$ converges in probability to X_∞ if for any $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\left\{\omega \in \Omega; \left\|X_n(\omega) - X_\infty(\omega)\right\| < \varepsilon\right\} = 1.$$

REMARK 5.7. We write $X_n \xrightarrow{p} X_\infty$. Note that by the way it is defined, $X_n \xrightarrow{p} X_\infty$ is equivalent to $X_n - X_\infty \xrightarrow{p} 0$.

REMARK 5.8. An ω corresponds to a data set. If $\{X_n\}$ converges in probability to X_∞, for a large sample size, all but for a small fraction of data sets, the estimator under consideration does not differ from a random vector X_∞ more than ϵ for any fixed $\epsilon > 0$.

REMARK 5.9. In most applications, X_∞ is a constant vector. In this case convergence in probability implies that the estimator under consideration "piles up" near a point.

The second concept, almost sure convergence, or convergence with probability one, corresponds to the point-wise convergence of a sequence of functions to a function. As we shall see later, convergence in this concept implies convergence in probability.

DEFINITION 5.5 (Almost Sure Convergence). $\{X_n\}$ converges almost surely to X_∞ if

$$\mathbb{P}\left\{\omega \in \Omega; \lim_{n\to\infty} X_n(\omega) = X_\infty(\omega)\right\} = 1.$$

REMARK 5.10. Note that

$$\begin{aligned}&\left\{\omega \in \Omega; \lim_{n\to\infty} X_n(\omega) = X_\infty(\omega)\right\}\\ &\qquad = \cap_{k=1}^{\infty} \cup_{j=1}^{\infty} \left\{\omega \in \Omega; \left\|X_m(\omega) - X_\infty(\omega)\right\| < 1/k \text{ for all } m \geq j\right\}\end{aligned}$$

and since the right-hand side is measurable, the left-hand side is measurable.

REMARK 5.11. We write $X_n \xrightarrow{a.s.} X_\infty$. Note that by the way it is defined, $X_n \xrightarrow{a.s.} X_\infty$ is equivalent to $X_n - X_\infty \xrightarrow{a.s.} 0$.

REMARK 5.12. As before, an ω corresponds to a data set. If $\{X_n\}$ converges almost surely to X_∞, for any data set, and for any $\epsilon > 0$ there is a large sample size such that the estimator under consideration does not differ from a random vector X_∞ more than ϵ.

The third concept of convergence uses the area between the converging function and the limit function as the measure of discrepancy between the two functions.

DEFINITION 5.6 (Convergence in rth Mean). Let $E\left\|X_n\right\|^r$ and $E\left\|X_\infty\right\|^r$ be finite for a given $r > 0$. $\{X_n\}$ converges in rth mean to X_∞ if

$$\lim_{n\to\infty} E\left\|X_n - X_\infty\right\|^r = 0.$$

REMARK 5.13. We write $X_n \xrightarrow{r} X_\infty$. Note that by the way it is defined $X_n \xrightarrow{r} X_\infty$ is equivalent to $X_n - X_\infty \xrightarrow{r} 0$.

The fourth concept examines the X_n as a random vector and claims convergence so long as its distribution converges to a fixed distribution. Note that this does not require that X_n and X_∞ be defined on the same probability space. Above three concepts of convergence presumed that X_n and X_∞ are defined on the same probability space.

Let $F(t) = \Pr\{X \leq t\}$ be the cumulative distribution function (CDF).

REMARK 5.14. If X is a random variable, then $F(t)$ is continuous from right, $F(\infty) = 1$ and $F(-\infty) = 0$ and non-decreasing.

REMARK 5.15. If X is a random vector, then $F(t)$ needs to satisfy more conditions for it to qualify as a CDF. For example, for two dimensional case

$$F(t_1 + \Delta_1, t_2 + \Delta_2) - F(t_1, t_2 + \Delta_2) - F(t_1 + \Delta_1, t_2) + F(t_1, t_2) \geq 0$$

for any t_j and Δ_j for $j = 1$ and 2.

DEFINITION 5.7 (Convergence in Distribution). $\{X_n\}$ converges in distribution to X_∞ if the CDF of $X_1, X_2, \ldots$ denoted by $F_1, F_2, \ldots$ converge to the cumulative distribution function of X_∞, denoted F_∞ at each continuity point of F_∞.

REMARK 5.16. The convergence is required only at the continuity point. If the convergence at the continuity points hold, then the probability assignments for the rest of the points can be inferred from these points by the properties of the CDF. The definition allows a sequence of the CDFs to converge to a function which is not a CDF. So long as it coincides with a CDF F_∞ at the continuity points of F_∞, we claim the convergence in distribution.

The definition relies on the concept of CDF which is not well defined for the infinite dimensional case. Even for the finite dimensional case the above complication arises. Therefore the following definition is more often used:

DEFINITION 5.8 (Convergence in Distribution). $\{X_n\}$ converges in distribution to X_∞ if for any bounded continuous functions f,

$$\lim_{n\to\infty} Ef(X_n) = Ef(X_\infty).$$

REMARK 5.17. Since sine and cosine functions are bounded continuous functions, if the condition holds, the characteristic function of X_n converges to that of X_∞.

REMARK 5.18. That the two definitions are equivalent requires a proof as a step function used to define the CDF is not a continuous function, for example. See Van der Vaart's textbook p.6.

REMARK 5.19. We write $X_n \xrightarrow{d} X_\infty$. Note that this is NOT equivalent to $X_n - X_\infty \xrightarrow{d} 0$.

REMARK 5.20. We will see that, $X_n - X_\infty \xrightarrow{d} 0$ implies $X_n \xrightarrow{d} X_\infty$ but the other direction does not necessarily hold. See the next remark for a counter-example.

REMARK 5.21. Let ω be distributed uniformly over $[0, 1]$ and define $X_n(\omega) = \omega$ and $X_\infty(\omega) = 1 - \omega$. In this case the CDFs of X_n and X_∞ are the same so that $X_n \xrightarrow{d} X_\infty$ holds. But clearly the CDF of $X_n - X_\infty$ is not concentrated at 0 so that $X_n - X_\infty \xrightarrow{d} 0$ does not hold.

REMARK 5.22. When $V_n^{-1/2}(X_n - \mu_n) \xrightarrow{d} N(0, I)$ we write $X_n = AN(\mu_n, V_n)$.

5.1.2.1. *Exercises.*

(1) Verify remarks 5.10 and 5.21.

5.1.3. Relationships among different modes of convergence. The following relationships holds:

$$\begin{array}{ll} X_n \overset{a.s.}{\to} X_\infty & \searrow \\ X_n \overset{r}{\to} X_\infty & \nearrow \end{array} X_n \overset{p}{\to} X_\infty \longrightarrow X_n \overset{d}{\to} X_\infty.$$

THEOREM 5.1. *$X_n \overset{a.s.}{\to} X_\infty$ implies $X_n \overset{p}{\to} X_\infty$.*

PROOF. The result is a corollary to the following Lemma. □

LEMMA 5.1. *$X_n \overset{a.s.}{\to} X_\infty$ if and only if for all $\varepsilon > 0$*

$$\lim_{j\to\infty} \mathbb{P}\left\{\omega \in \Omega; \|X_m(\omega) - X_\infty(\omega)\| < \varepsilon \text{ for all } m \geq j\right\}.$$

PROOF. Note that

$$\begin{aligned} 1 &= \mathbb{P}\left\{\cap_{k=1}^{\infty} \cup_{j=1}^{\infty} \left\{\omega \in \Omega; \|X_m(\omega) - X_\infty(\omega)\| < 1/k \text{ for all } m \geq j\right\}\right\} \\ &= \lim_{k\to\infty} \mathbb{P}\left\{\cup_{j=1}^{\infty} \left\{\omega \in \Omega; \|X_m(\omega) - X_\infty(\omega)\| < 1/k \text{ for all } m \geq j\right\}\right\} \\ &= \lim_{k\to\infty} \lim_{j\to\infty} \mathbb{P}\left\{\omega \in \Omega; \|X_m(\omega) - X_\infty(\omega)\| < 1/k \text{ for all } m \geq j\right\}. \end{aligned}$$

The first equality holds by remark 5.10. The second and the third equalities hold by the continuity of $\mathbb{P}$. Thus for any $\varepsilon > 0$,

$$\begin{aligned} &\lim_{j\to\infty} \mathbb{P}\left\{\omega \in \Omega; \|X_m(\omega) - X_\infty(\omega)\| < \varepsilon \text{ for all } m \geq j\right\} \\ \geq\ &\lim_{k\to\infty} \lim_{j\to\infty} \mathbb{P}\left\{\omega \in \Omega; \|X_m(\omega) - X_\infty(\omega)\| < 1/k \text{ for all } m \geq j\right\} = 1. \end{aligned}$$

This shows one direction. If on the other hand for any $\varepsilon > 0$,

$$\lim_{j\to\infty} \mathbb{P}\left\{\omega \in \Omega; \|X_m(\omega) - X_\infty(\omega)\| < \varepsilon \text{ for all } m \geq j\right\} = 1.$$

Then for any k

$$\begin{aligned} 1 &= \lim_{j\to\infty} \mathbb{P}\left\{\omega \in \Omega; \|X_m(\omega) - X_\infty(\omega)\| < 1/k \text{ for all } m \geq j\right\} \\ &= \mathbb{P}\left\{\cup_{j=1}^{\infty} \left\{\omega \in \Omega; \|X_m(\omega) - X_\infty(\omega)\| < 1/k \text{ for all } m \geq j\right\}\right\} \\ &= \mathbb{P}\left\{\cap_{s=1}^{k} \cup_{j=1}^{\infty} \left\{\omega \in \Omega; \|X_m(\omega) - X_\infty(\omega)\| < 1/s \text{ for all } m \geq j\right\}\right\}. \end{aligned}$$

Take $k \to \infty$ and noting that by continuity of $\mathbb{P}$, the last expression is the probability of

$$\cap_{k=1}^{\infty} \cup_{j=1}^{\infty} \left\{\omega \in \Omega; \|X_m(\omega) - X_\infty(\omega)\| < 1/k \text{ for all } m \geq j\right\}.$$

□

THEOREM 5.2. *$X_n \overset{r}{\to} X_\infty$ implies $X_n \overset{p}{\to} X_\infty$.*

PROOF. Note that for any $\varepsilon > 0$,

$$\begin{aligned} E\|X_n - X_\infty\|^r &\geq E\left\{\|X_n - X_\infty\|^r 1\left\{\|X_n - X_\infty\| > \varepsilon\right\}\right\} \\ &\geq \varepsilon^r E 1\left\{\|X_n - X_\infty\| > \varepsilon\right\} \\ &= \varepsilon^r \mathbb{P}\left\{\|X_n - X_\infty\| > \varepsilon\right\}. \end{aligned}$$

Thus

$$\mathbb{P}\left\{\|X_n - X_\infty\| > \varepsilon\right\} \leq E\|X_n - X_\infty\|^r / \varepsilon^r \to 0.$$

□

THEOREM 5.3. *$X_n \xrightarrow{r} X_\infty$ implies $X_n \xrightarrow{s} X_\infty$ for any $s > 0$ and $s \leq r$.*

PROOF. For any $\varepsilon > 0$

$$\begin{aligned} E\left\|X_n - X_\infty\right\|^s &= E\left\|X_n - X_\infty\right\|^s 1\left\{\left\|X_n - X_\infty\right\| > 1\right\} \\ &\quad + E\left\|X_n - X_\infty\right\|^s 1\left\{1 \geq \left\|X_n - X_\infty\right\| > \varepsilon\right\} \\ &\quad + E\left\|X_n - X_\infty\right\|^s 1\left\{\varepsilon \geq \left\|X_n - X_\infty\right\|\right\}. \\ &\leq E\left\|X_n - X_\infty\right\|^r + \mathbb{P}\left\{\left\|X_n - X_\infty\right\| > \varepsilon\right\} + \varepsilon^s. \end{aligned}$$

The first two terms converge to zero and the last term can be made arbitrarily small by taking ε small. □

THEOREM 5.4. *$X_n \xrightarrow{p} X_\infty$ implies $X_n \xrightarrow{d} X_\infty$.*

PROOF. Note that for any function f, $|Ef(X_n) - Ef(X_\infty)|$ is bounded above by $E|f(X_n) - f(X_\infty)|$ and in turn it equals the sum of three terms:

$$\begin{aligned} &E|f(X_n) - f(X_\infty)|1\left\{\left\|X_n - X_\infty\right\| \leq \varepsilon \text{ and } \left\|X_\infty\right\| \leq M\right\}, \\ &E|f(X_n) - f(X_\infty)|1\left\{\left\|X_n - X_\infty\right\| > \varepsilon \text{ and } \left\|X_\infty\right\| \leq M\right\}, \\ &\text{and } E|f(X_n) - f(X_\infty)|1\left\{\left\|X_\infty\right\| > M\right\}. \end{aligned}$$

Since f restricted to $\|X_n - X_\infty\| \leq \varepsilon$ and $\|X_\infty\| \leq M$ is uniformly continuous if f is continuous and thus for any $\delta > 0$ there is a $\varepsilon_\delta > 0$ such that the first term is less than δ. When f is bounded and $X_n \xrightarrow{p} X_\infty$, the second term converges to zero. Finally, when f is bounded by taking M large, the last expression can be made arbitrarily small. Thus for any bounded and continuous function $|Ef(X_n) - Ef(X_\infty)|$ converges to zero. □

5.1.3.1. *Examples that show the converses do not hold.* Before reading on, if you really want to understand, then you should try on your own to construct examples.

EXAMPLE 5.1. $X_n \xrightarrow{d} X_\infty$ but not $X_n \xrightarrow{p} X_\infty$:
Take $X_n = N(0,1)$ and $X_\infty = -N(0,1)$.

EXAMPLE 5.2. $X_n \xrightarrow{a.s.} X_\infty$ but not $X_n \xrightarrow{r} X_\infty$:
Let $X_n = e^n$ with probability $1/n$ and 0 with probability $1 - 1/n$ and $X_\infty = 0$.

EXAMPLE 5.3. $X_n \xrightarrow{r} X_\infty$ but not $X_n \xrightarrow{a.s.} X_\infty$:
Recall that the convergence in the rth mean uses the measure of the area between the two functions as the criteria of convergence. Since the criteria is met so long as the area becomes small the following construction satisfies this but not the almost sure convergence; Let an initial partition of $[0,1]$ into intervals A_{1j} $j \in J_1$. Now we consider a finer partition of $[0,1]$ into intervals A_{2j} $j \in J_2$. We continue these rounds and make sure that the length of the largest interval converges to zero. For each round, define a sequence of functions that is zero except for one interval where the function's value is one. Now consider a sequence of functions linking up rounds. This sequence of functions does not converge because the function value oscillates between zero and one at any point. But it does converge in the rth mean to zero because the area between the function and zero converges to zero.

EXAMPLE 5.4. Since $X_n \xrightarrow{r} X_\infty$ implies $X_n \xrightarrow{p} X_\infty$, this example also shows that $X_n \xrightarrow{p} X_\infty$ does not imply $X_n \xrightarrow{a.s.} X_\infty$.

To see sufficient conditions for converses to hold, see Serfling's textbook. Here is a useful result:

DEFINITION 5.9. A sequence of random vector $\{Y_n\}$ is asymptotically uniformly integrable if

$$\lim_{M\to\infty} \sup_n E\|Y_n\| 1\{\|Y_n\| > M\} = 0.$$

REMARK 5.23. If $\sup_n E\|Y_n\|^{1+\varepsilon} < \infty$ for some $\varepsilon > 0$, then $\{Y_n\}$ is uniformly integrable.

THEOREM 5.5. *Let $f : \mathbb{R}^k \to \mathbb{R}$ be measurable and continuous at every point in a set C. Let $X_n \xrightarrow{d} X_\infty$ where X_∞ takes values in C. Then $Ef(X_n) \to Ef(X_\infty)$ if and only if $\{f(X_n)\}$ is asymptotically uniformly integrable.*

THEOREM 5.6 (Dominated Convergence). *Let $X_n \xrightarrow{a.s.} X_\infty$ and assume $|X_n| < X^*$ and $EX^* < \infty$. Then $E|X_\infty| < \infty$, $E|X_n| \to E|X_\infty|$, and*

$$E|X_n - X_\infty| \to 0 \text{ as } n \to \infty.$$

EXERCISE 5.1. In view of the dominated convergence, what is the sufficient condition for the a.s. convergence to imply the 1st mean convergence?

EXERCISE 5.2. Discuss which assumption in the dominated convergence theorem is violated in example 5.2.

5.1.4. Other useful results. The following results are useful in establishing asymptotic properties of estimators.

THEOREM 5.7 (Continuous Mapping Theorem). *Let X_n be a $\mathbb{R}^k$-valued random vector, g is $\mathbb{R}^m$-valued and continuous on C with $\Pr\{X_\infty \in C\} = 1$. Then*

(1) $X_n \xrightarrow{a.s.} X_\infty$ implies $g(X_n) \xrightarrow{a.s.} g(X_\infty)$,
(2) $X_n \xrightarrow{p} X_\infty$ implies $g(X_n) \xrightarrow{p} g(X_\infty)$,
(3) $X_n \xrightarrow{d} X_\infty$ implies $g(X_n) \xrightarrow{d} g(X_\infty)$.

EXAMPLE 5.5. If $X_n \xrightarrow{d} N(0, I_k)$ then $X_n' X_n \xrightarrow{d} \chi^2(k)$.

The following results often referred to as Slutsky's lemma follows from the continuous mapping theorem.

THEOREM 5.8 (Slutsky). *Let $X_n \xrightarrow{d} X_\infty$ and $Y_n \xrightarrow{d} c$ for some constant $|c| < \infty$.*

(1) $X_n + Y_n \xrightarrow{d} X_\infty + c$
(2) $X_n \cdot Y_n \xrightarrow{d} c \cdot X_\infty$
(3) $X_n / Y_n \xrightarrow{d} X_\infty / c$ if $c \neq 0$.

REMARK 5.24. $Y_n \xrightarrow{d} c$ if and only if $Y_n \xrightarrow{p} c$.

The following result is often referred to as the Cramér-Wold device and used to reduce the problem of showing convergence in distribution in high dimension to that in one dimension.

THEOREM 5.9 (Cramér and Wold). *Let* $X_n = (X_{1n}, ..., X_{mn})'$. *Then* $X_n \xrightarrow{d} X_\infty$ *if and only if for any* $\lambda \in \mathbb{R}^m$

$$\lambda_1 X_{1n} + \cdots + \lambda_m X_{mn} \xrightarrow{d} \lambda_1 X_{1\infty} + \cdots + \lambda_m X_{m\infty}.$$

REMARK 5.25. The result states that if the asymptotic distribution of all linear combination of a sequence of random vectors coincide with the distribution of the same linear combination of a given random vector, then the sequence of random vectors converges to the given random vector.

5.2. Law of Large Numbers

So far we defined some concepts and examined the relationship among them. Next we turn to some results which can be used to establish some convergence results. Results to show convergence in probability, almost sure convergence and convergence in rth mean are referred to as the law of large numbers (LLN).

Chebyshev's inequality is a useful result to show a LLN as we will see.

THEOREM 5.10 (Chebyshev's inequality). *If* ϕ *is strictly positive and increasing function on* $(0, \infty)$ *and* $\phi(t) = \phi(-t)$ *for any* t *and* X *is a random variable such that* $E\phi(X) < \infty$, *then for any* $\varepsilon > 0$,

$$\Pr\{|X| \geq \varepsilon\} \leq E\phi(X)/\phi(\varepsilon).$$

PROOF. $E\phi(X) \geq E\phi(X)\,1\{|X| \geq \varepsilon\} \geq \phi(\varepsilon)\Pr\{|X| \geq \varepsilon\}$. □

Convergence in probability is sometimes called a weak LLN (WLLN) and convergence almost surely is sometimes called a strong LLN (SLLN).

THEOREM 5.11 (Chebyshev's WLLN). *Let* $E(X_i) = \mu_i$. *For any* $\varepsilon > 0$,

$$\Pr\left\{\left|n^{-1}\sum_{i=1}^{n}(X_i - \mu_i)\right| > \varepsilon\right\} \to 0$$

if $E\left\{\left(n^{-1}\sum_{i=1}^{n}(X_i - \mu_i)\right)^2\right\} \to 0$.

PROOF. A direct consequence of the Chebyshev's inequality applied to $\phi(u) = u^2$. □

REMARK 5.26. The theorem can allow very general correlation and X_i can depend on n although the statement does not make that explicit.

REMARK 5.27. If the sequence are independent and the variances are uniformly bounded, then almost sure convergence holds as the next result implies.

THEOREM 5.12 (Kolmogorov's SLLN). *If* $\{X_i\}$ *is an independent sequence,* $E(X_i) = \mu_i$, $Var(X_i) = \sigma_i^2$ *and* $\sum_{i=1}^{\infty}\sigma_i^2/i^2 < \infty$, *then*

$$n^{-1}\sum_{i=1}^{n}(X_i - \mu_i) \xrightarrow{a.s.} 0$$

as $n \to \infty$.

REMARK 5.28. Kronecker's lemma states that if $\sum_{i=1}^{\infty} x_i = s$ for some finite s, then for any nondecreasing positive sequence $\{b_i\}$ which diverges, $\lim_{n\to\infty}(\sum_{i=1}^{n} b_i x_i)/b_n = 0$. Using the lemma one can show that the condition in Kolmogorov's theorem implies the Chebyshev's WLLN for the case of independent sequence. This result shows the almost sure convergence holds so the consequence is stronger.

5.3. Central Limit Theorem

THEOREM 5.13 (Lindeberg). *Let $\{X_{ni}\}$ be independent with means $\{\mu_{ni}\}$ and finite variance $\left\{\sigma_{ni}^2\right\}$. Let $c_n = \left(\sum_{i=1}^{k_n} \sigma_{ni}^2\right)^{1/2}$. If for any $\eta > 0$*

$$\lim_{n\to\infty} \frac{1}{c_n^2} \sum_{i=1}^{k_n} E\left\{\left(X_{ni} - \mu_{ni}\right)^2 1\left\{\left|X_{ni} - \mu_{ni}\right| > \eta \cdot c_n\right\}\right\} = 0,$$

then

$$\frac{1}{c_n} \sum_{i=1}^{k_n} \left(X_{ni} - \mu_{ni}\right) \xrightarrow{d} N\left(0,1\right).$$

REMARK 5.29. Suppose $\{X_i\}$ is iid with mean μ and variance σ^2. Then the Lindeberg's condition holds with $k_n = n$ and thus

$$\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} \left(X_i - \mu\right) \xrightarrow{d} N\left(0,1\right).$$

EXERCISE 5.3. Show the assertion that under iid and finite variance, the Lindeberg's condition holds with $k_n = n$. (Use the dominated convergence.)

5.4. Big O_P and little o_p

In obtaining asymptotic properties of an estimator, O_P, o_p notations are useful in eliminating messy calculations. After defining some notations we show some valid calculations using them.

DEFINITION 5.10 (little o_p). When $X_n - X_\infty \xrightarrow{p} 0$ we write $X_n = X_\infty + o_p\left(1\right)$. When $n^\alpha\left(X_n - X_\infty\right) \xrightarrow{p} 0$ for some α, we write $X_n = X_\infty + o_p\left(n^{-\alpha}\right)$.

We can think of $o_p\left(1\right)$ or $o_p\left(n^{-\alpha}\right)$ for any $\alpha \geq 0$ as an approximation error which converges to zero in probability.

Big O_P corresponds to the boundedness of a sequence of numbers. Since it is about a sequence of random vectors, the concept requires that the support of the random vectors do not drift off from a finite region.

DEFINITION 5.11 (Big O_P). If for any $\varepsilon > 0$, there exists M_ε (not depending on n) such that $\Pr\left\{\|X_n\| > M_\varepsilon\right\} < \varepsilon$ then we write $X_n = O_P\left(1\right)$. If the same holds for $n^\alpha X_n$ then we write $X_n = O_P\left(n^{-\alpha}\right)$.

REMARK 5.30. This is sometimes called stochastic boundedness or tightness.

REMARK 5.31. Any random variable is $O_P\left(1\right)$.

REMARK 5.32. $X_n = O_P\left(1\right)$ is a weaker requirement than $X_n \xrightarrow{d} X_\infty$.

5.4.1. Useful facts for computations.

(1) If $X_n \xrightarrow{d} X_\infty$ then $X_n = O_P\left(1\right)$.
(2) If $X_n = O_P\left(1\right)$ and $Y_n = o_p\left(1\right)$, then $X_n \cdot Y_n = o_p\left(1\right)$ and $X_n + Y_n = O_P\left(1\right)$.
(3) If $X_n = O_P\left(1\right)$ and $Y_n = O_p\left(1\right)$, then $X_n \cdot Y_n = O_p\left(1\right)$ and $X_n + Y_n = O_P\left(1\right)$.
(4) If $X_n = o_p\left(1\right)$ and $Y_n = o_p\left(1\right)$, then $X_n \cdot Y_n = o_p\left(1\right)$ and $X_n + Y_n = o_p\left(1\right)$.

EXERCISE 5.4. Show all the results above.

EXERCISE 5.5. Show that when $\hat{\sigma} \xrightarrow{p} \sigma$ and $\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} N(0,\sigma)$, $\sqrt{n}\left(\hat{\theta}-\theta_0\right)/\hat{\sigma} \xrightarrow{d} N(0,1)$.

5.5. Desirable properties of an estimator

We consider an estimator $\hat{T}_n$ defined on a probability model parametrized by $\theta \in \Theta$ where Θ is the parameter space. An estimator is a function of data which does not depend on θ. An aspect of the parameter θ expressed by $g(\theta)$ is the parameter $\hat{T}_n$ estimates.

5.5.1. Asymptotic unbiasedness. When for any $\theta \in \Theta$, $\lim_{n\to\infty} E_\theta \hat{T}_n = g(\theta)$, $\hat{T}_n$ is called asymptotically unbiased. The expectation is taken using the distribution that correspond to θ, which is denoted by E_θ. We require the condition to hold for any $\theta \in \Theta$. If the condition holds for a particular value of θ but not others, then the estimator is not very attractive as we do not know which θ generates data in a particular case.

REMARK 5.33. Unbiasedness requires $E_\theta \hat{T}_n = g(\theta)$ for all $\theta \in \Theta$.

REMARK 5.34. This condition only means that the location of the estimator under consideration is about right.

5.5.2. Consistency. $\hat{T}_n$ is weakly consistent when $\hat{T}_n \xrightarrow{p} g(\theta)$ for all $\theta \in \Theta$. $\hat{T}_n$ is strongly consistent when $\hat{T}_n \xrightarrow{a.s.} g(\theta)$ for all $\theta \in \Theta$. $\hat{T}_n$ is consistent in rth mean when $\hat{T}_n \xrightarrow{r} g(\theta)$ for all $\theta \in \Theta$.

Since $g(\theta)$ is constant all definitions of consistency imply that the probability piles up at $g(\theta)$. This is usually taken as a minimum requirement for a valid estimation procedure.

5.5.3. Rate of convergence. Consistency does not inform us how fast the convergence is. A common way to get some idea about this is to obtain the maximal rate at which we can blow up the difference $\hat{T}_n - g(\theta)$ and still make the difference stochastically bounded. This rate is called the convergence rate. Typically $\sqrt{n}\left(\hat{T}_n - g(\theta)\right) = O_P(1)$ can be shown but other rates are known particularly for nonparametric and time series cases.

5.5.4. Asymptotic distribution theory. Consistency implies $\hat{T}_n \xrightarrow{d} g(\theta)$ for all $\theta \in \Theta$. (Can you tell why?) However this or the knowledge about the convergence rate does not provide a way to assess inaccuracy of the estimator $\hat{T}_n$ about $g(\theta)$.

Often by recentering and rescaling $\hat{T}_n$, $\tilde{T}_n = r_n\left(\hat{T}_n - b_n\right)$ one can obtain a way to measure the inaccuracy. Typically $b_n = g(\theta)$ and $r_n = \sqrt{n}$.

5.5.5. Asymptotic relative efficiency. Consider two estimators of the same parameter $g(\theta)$. When the two estimators have inaccuracy measures A_{1n} and A_{2n}, A_{1n}/A_{2n} is called the asymptotic relative efficiency of the second estimator with respect to the first estimator. Usually variances are used. In the case of the variances typically one can translate the result to asymptotic sample sizes because the convergence rate is typically proportional to the square root of the sample size,

so that the variance will go down inverse proportionally to the sample size. If the ratio is 2 for example, then the one with smaller asymptotic variance, in effect behaves as if it has twice the data.

CHAPTER 6

Asymptotic Analysis of the Estimators of the Parameters in the Linear Regression Model

6.1. Consistency

We show that the OLS estimator is consistent. We have seen that, in order to interpret the coefficient on the linear regression model as a meaningful parameter, we need to maintain the conditional mean under Assumptions OLS.1, 2, and 4.

ASSUMPTION 6.1 (**conditional mean version**). *$y_i = x_i'\beta + u_i$ and $E(u_i|x_i) = 0$.*

ASSUMPTION 6.2 (**unconditional version**). *$y_i = x_i'\beta + u_i$ and $E(u_i x_i) = 0$.*

ASSUMPTION 6.3 (**sample version**). *$rank(\mathbf{X}) = K$.*

ASSUMPTION 6.4 (**population version**). *$rank E(x_i x_i') = K$.*

ASSUMPTION 6.5 (**conditional homoskedasticity**). *$E(u_i^2|x_i) = \sigma^2$.*

ASSUMPTION 6.6 (**no correlation of u_i^2 and any cross terms of x_i**). *$E(u_i^2 x_i x_i') = \sigma^2 E(x_i x_i')$, where $\sigma^2 = E(u_i^2)$.*

ASSUMPTION 6.7. *Sampling of (x_i, y_i) is i.i.d.*

To see this, note that

$$\begin{aligned}\hat{\beta} &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U} \\ &= \beta + (N^{-1}\mathbf{X}'\mathbf{X})^{-1}N^{-1}\mathbf{X}'\mathbf{U}\end{aligned}$$

and that $N^{-1}\mathbf{X}'\mathbf{X} = N^{-1}\sum_{i=1}^N x_i x_i' \xrightarrow{p} E(x_i x_i')$ and
$N^{-1}\mathbf{X}'\mathbf{U} = N^{-1}\sum_{i=1}^N x_i u_i \xrightarrow{p} E(x_i u_i) = 0$.

Under OLS.2, the continuous mapping theorem is applicable and hence consistency of the OLS estimator is proved.

6.2. Asymptotic Normality

To see the asymptotic normality note that under Assumptions OLS.1, 2, and 4,

$$\begin{aligned}\sqrt{N}(\hat{\beta} - \beta) &= \sqrt{N}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U} \\ &= (N^{-1}\mathbf{X}'\mathbf{X})^{-1}N^{-1/2}\mathbf{X}'\mathbf{U}\end{aligned}$$

and that $N^{-1/2}\mathbf{X}'\mathbf{U} = N^{-1/2}\sum_{i=1}^N x_i u_i \xrightarrow{d} \mathcal{N}(0, Var(x_i u_i))$.

Continuous mapping theorem implies that $\sqrt{N}(\hat{\beta}-\beta) \xrightarrow{d} \mathcal{N}(0, E(x_i x_i')^{-1} Var(x_i u_i) E(x_i x_i')^{-1})$.

Note that $Var(x_i u_i) = E(u_i^2 x_i x_i')$ and that under Assumption OLS.3, $E(u_i^2 x_i x_i') = \sigma^2 E(x_i x_i')$ so that under Assumptions OLS.1–4, the asymptotic distribution of $\sqrt{N}(\hat{\beta} - \beta)$ is $\mathcal{N}(0, \sigma^2 E(x_i x_i')^{-1})$.

The result tells us that the OLS estimator converges at rate $1/\sqrt{N}$ and that the asymptotic distribution is normal. We know the exact formula of the distribution but the variance-covariance matrix is not known. We consider how we estimate it next.

6.3. Estimation of the Variance-Covariance Matrix

We can consistently estimate $E(x_i x_i')^{-1}$ by $(N^{-1}\sum_{i=1}^N x_i x_i')^{-1}$ under Assumption OLS.2 and 4 as we have seen earlier.

In order to estimate σ^2, we use the OLS residual $\hat{u}_i$.

A natural estimator is $N^{-1}\sum_{i=1}^N \hat{u}_i^2$.it next.

Recall that $\hat{u}_i = y_i - x_i'\hat{\beta}$ so that $\hat{u}_i = u_i - x_i'(\hat{\beta} - \beta)$. Thus

$$\hat{u}_i^2 = u_i^2 + (\hat{\beta} - \beta)' x_i x_i' (\hat{\beta} - \beta) - 2u_i x_i'(\hat{\beta} - \beta)$$

so that

$$N^{-1}\sum_{i=1}^N \hat{u}_i^2 = N^{-1}\sum_{i=1}^N u_i^2 + (\hat{\beta} - \beta)' N^{-1}\sum_{i=1}^N x_i x_i' (\hat{\beta} - \beta) - 2N^{-1}\sum_{i=1}^N u_i x_i'(\hat{\beta} - \beta).$$

Without making assumption OLS.3, we still know that the OLS estimator is asymptoically normal with the asymptotic variance-covariance matrix $E(x_i x_i')^{-1} Var(x_i u_i) E(x_i x_i')^{-1}$.

We can show that $Var(x_i u_i)$ can be estimated consistently by $N^{-1}\sum_{i=1}^N \hat{u}_i^2 x_i x_i'$ under the existence of the 4th moments of regressors. Thus the asymptotic variance-covariance matrix can be consistently estimated by

$$(N^{-1}\sum_{i=1}^N x_i x_i')^{-1} N^{-1}\sum_{i=1}^N \hat{u}_i^2 x_i x_i' (N^{-1}\sum_{i=1}^N x_i x_i')^{-1}.$$

with the additional assumption that the 4th moments of the regressors exist.

6.4. Inference Using Asymptotic Results

We examine how the confidence interval for a linear combination of the parameters in the linear regression model, $c'\beta$ can be constructed using the OLS estimator, by appealing to the asymptotic theory.

Note that by the so called Delta method, a non-linear combination of parameters $g(\beta)$ can be dealt with by setting $c = \partial g(\beta)/\partial\beta$ so that studying the linear case is sufficient to the first order asymptotic analysis.

Theorem: Delta Method Suppose $g(z) : R^k \to R^m$ is continuously differentiable. If length k vector-valued estimator $\hat{\beta}$ of β is such that $\sqrt{N}(\hat{\beta} - \beta)$ converges in distribution to the normal random vector with mean 0 and the variance-covariance matrix V, then $\sqrt{N}(g(\hat{\beta}) - g(\beta))$ converges in distribution to the normal random vector with mean 0 and the variance-covariance matrix CVC', where $m \times k$ matrix C is defined as $C = \partial g(\beta)/\partial z'$.

CHAPTER 7

Linear Algebra

7.1. Exercises

(1) Denote the transpose of a matrix or a vector by the prime so that the transpose of a vector x is x' and the transpose of a matrix A is A'. Show that $A = (A')'$.
(2) Let A be an $m \times n$ matrix and its jth column is denoted as a_j. Let B be an $n \times m$ matrix and its jth row is denoted as b'_j. Show that $AB = \sum_{j=1}^{n} a_j b'_j$ and that

$$BA = \begin{pmatrix} b'_1 a_1 & b'_1 a_2 & \cdots & b'_1 a_n \\ b'_2 a_1 & b'_2 a_2 & \cdots & b'_2 a_n \\ \vdots & \vdots & \ddots & \vdots \\ b'_n a_1 & b'_n a_2 & \cdots & b'_n a_n \end{pmatrix}.$$

(3) For the same matrices A and B as above, show that $(AB)' = B'A'$.
(4) If $A'A = 0$, then $A = 0$.
(5) Define what the row rank and the column rank of an $m \times n$ matrix A are. Show that they are equal when either of the rank is 1. (Can you show that they are the same more generally? This result allows us to talk about the rank of a matrix.)
(6) Matrix A has full column rank if and only if $A'A$ is invertible.
(7) Let A and B be an $m \times n$ matrix and an $n \times m$ matrix, respectively.
   (a) Express column vectors of matrix AB in terms of column vectors of A to show that the column rank of AB does not exceed the column rank of A.
   (b) Similarly, express the row vectors of matrix of AB in terms of row vectors of B to show that the row rank of AB does not exceed the row rank of B.
   (c) Use the above results and the fact the row rank and the column rank are the same (Question 5), to prove that the rank of AB does not exceed the rank of A or the rank of B.
(8) Show that $trace(AB) = trace(BA)$.
(9) Partition Inverse: Assume that A and D are invertible and that the matrix

$$\begin{pmatrix} A & C \\ B' & D \end{pmatrix}$$

is invertible.

$$\begin{pmatrix} A & C \\ B' & D \end{pmatrix}\begin{pmatrix} W & Y \\ X' & Z \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}$$
$$\begin{pmatrix} AW + CX' & AY + CZ \\ B'W + DX' & B'Y + DZ \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}.$$

Thus $AY = -CZ$ so that $Y = -A^{-1}CZ$ and $DX' = -B'W$ so that $X' = -D^{-1}B'W$. Substituting these into the diagonal matrices, we obtain

$$AW - CD^{-1}B'W = I, \qquad -B'A^{-1}CZ + DZ = I.$$

Note that if $(A - CD^{-1}B')$ is not invertible, then there exists a non-zero vector W such that

$$AW - CD^{-1}B'W = 0.$$

Then using the same W and setting $X' = -D^{-1}B'W$,

$$\begin{pmatrix} A & C \\ B' & D \end{pmatrix}\begin{pmatrix} W \\ -D^{-1}B'W \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

But this contradicts the assumption that the matrix is invertible. So that these imply

$$W = (A - CD^{-1}B')^{-1}, \qquad Z = (D - B'A^{-1}C)^{-1}.$$

Also,

$$X' = -D^{-1}B'(A - CD^{-1}B')^{-1}, \qquad Y = -A^{-1}C(D - B'A^{-1}C)^{-1}.$$

Similarly,

$$\begin{pmatrix} W & Y \\ X' & Z \end{pmatrix}\begin{pmatrix} A & C \\ B' & D \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}$$
$$\begin{pmatrix} WA + YB' & WC + YD \\ X'A + ZB' & X'C + ZD \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}.$$

From the diagonal equalities we have

$$W = A^{-1} - YB'A^{-1}, \qquad Z = D^{-1} - X'CD^{-1}.$$

These imply

$$W = A^{-1} + A^{-1}C(D - B'A^{-1}C)^{-1}B'A^{-1}$$
$$Z = D^{-1} + D^{-1}B'(A - CD^{-1}B')^{-1}CD^{-1}.$$

Since these should coincide

$$(A - CD^{-1}B')^{-1} = A^{-1} + A^{-1}C(D - B'A^{-1}C)^{-1}B'A^{-1}.$$

CHAPTER 8

Probability Theory

(1) Let V be a vector of independent standard normal random variables of length n and A be an $n \times n$ symmetric idempotent matrix. Then $V'AV$ is the chi-square random variable with degrees of freedom equal to the rank of A.

Since A is idempotent, using the spectral decomposition, $A = H\Lambda H'$, where Λ is the diagonal matrix with the eigen-values of A as the diagonal elements. Matrix H is the corresponding eigen-vectors ordered according to the eigen-values. Since A is idempotent, each of the eigen-values is either zero or one. Also $H'H = I$. Thus,

$$V'AV = (V'H)\Lambda(H'V)$$

and that $H'V$ is a vector of a normal random variable with mean zero and variance-covariance matrix $H'H = I$. Thus elements of $H'V$ are mutually independent, being jointly normal and having covariance zero. Thus $(V'H)\Lambda(H'V)$ is a sum of squares of the standard normal random variables. The summation is over the elements corresponding to the eigen-values of one. Since the number of elements with eigen-values being one, equals to the rank of A, the distribution is chi-square with the degrees of freedom equal to the rank of A.