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), $f_U(\cdot/\sigma|X = x)/\sigma$, the conditional density of Y given $X = x$ can be written as $f_U((y - x'\beta)/\sigma|X = x)/\sigma$, where $f_U(\cdot|x)$ satisfies

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

Thus the entire conditional distribution of Y given $X = x$ can be parametrized by $(\beta, f_U(\cdot/\sigma|X = x)/\sigma)$.

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 $\mathbf{X}$. 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 $\mathbf{X}$ 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.1 (unconditional 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.2: (population version) $\text{rank}E(x_i x_i') = K$.

Assumption OLS.3 (conditional homoskedasticity): $E(u_i^2|x_i) = \sigma^2$.
Assumption OLS.3: (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 OLS.4: Sampling of (x_i, y_i) is i.i.d.

In this section we will use the conditional mean version, sample version, and conditional homoskedasticity versions to study finite sample properties. For the asymptotic properties, we use the second respective versions.

4.1.1. 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 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.

4.1.2. 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

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

where $\Omega(x)$ 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.

4.1.3. 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.

4.2. 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. 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.

We will see that replacing the unknown σ^2 with the unbiased estimator of σ^2:

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

where $\hat{u}_i = y_i - x_i'\hat{\beta}$ yields $t(N - K)$ conditional on $\mathbf{X}$, under Assumptions OLS.1–OLS.5.

Note that

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

is the ratio of

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

and $\sqrt{\hat{\sigma}^2/\sigma^2}$. The proof proceeds to examine the ratio.

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

In order to not cut the flow of the argument, the proof will be provided in the last section.

Next, we turn to consider 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 could start from considering how to construct a confidence interval for a particular regression coefficient, since the logic is the same, we study constructing a confidence interval for the linear combination of the regression coefficients.

First, observe that since the distribution is derived above:

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

we can compute any pth percentile $t_p(N-K)$: 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 $1-\alpha$ (Typically $\alpha = 0.05$) confidence interval for $\sum_{j=1}^{K} c_j\beta_j$ is computed using the above result:

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

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 $1-\alpha$ 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

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

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

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

or $(.961, 1.21)$ for the coefficient on $x_1$and

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

or $(-.0045, .0479)$ for the coefficient on x_2.

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

4.4.1. idea behind the statistical hypothesis testing procedure. Sometimes we are more interested in finding out if the parameter takes a particular value or not. This inference problem is called the hypothesis test. The hypothesis to be tested is usually called the null hypothesis. We consider the null hypothesis of the form $\sum_{j=1}^{J} c_j \beta_j = a$.

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."

Usually, the null hypothesis is chosen in the way we wish to set the probability of rejecting the null hypothesis small when in fact the null hypothesis holds. This type of error is called *Type I error* and the probability of making Type I error is called the *significance level*.

Thus 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. This means we set the Type I error low, typically at 5%.

4.4.2. how to choose the rejection region. Under the null hypothesis, 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. The region is called the rejection region with the significance level α.

We reject the null hypothesis when this inequality holds. Typically α is set at 0.05 but sometimes 0.01 or 0.1 is used.

4.4.3. 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 reject the null hypothesis when in fact the hypothesis holds if we repeat this with many different data sets. This error is called the Type I error.

4.4.4. 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 minimize the probability of accepting the null hypothesis when indeed the null hypothesis does not hold (*Type II error*) especially for parameter values far away from the true parameter.

1 minus the probability of Type II error, Probability of rejecting the the null hypothesis when we should reject the null hypothesis, is called the power.

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

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

but it is not equal to α but something bigger.

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 called the non-centrality parameter.

Non-centered t-distribution is asymmetric 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.

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 or less than the null value a. Then, we only care about the power in that direction.

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.

4.4.5. One-sided alternatives: An example.

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

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.

4.4.6. 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.

4.4.7. 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. 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 i 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 in the Wooldridge's textbook,

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + u.$$

The null hypothesis that says batting performaces themselves do not matter can be stated as $H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$.

$$\begin{array}{rlcll} \widehat{\log(salary)} = & 11.19 & + & .0689 years & + \quad .0126 gamesyr \\ & (.29) & & (.0121) & \quad (.0026) \\ + & .00098 bavg & + & .0144 hrunsyr & + \quad .0108 rbisyr \\ & (.00110) & & (.0161) & \quad (.0072) \\ n = 353, & SSR = 183.186 & & R^2 = .6278 & \end{array}$$

In this case, we need to consider OLS estimators of β_3, β_4, and β_5 at the same time. 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 $(\hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5)'$ 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 rank. In general, the linear combinations $C\hat{\beta}$ is distributed $N(C\beta, \sigma^2 C(\mathbf{X}'\mathbf{X})^{-1}C')$. Therefore under the assumption on C,

$$[C(\mathbf{X}'\mathbf{X})^{-1}]^{-1/2}(C\hat{\beta} - C\beta) \sim N(0, I).$$

Therefore

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

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

This corresponds to examining the iso-height values of the joint normal density. 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 as if 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. Thus once we show that they are independent, then the ratio 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 the "Wald approach".

4.5.2. likelihood ratio and lagrangean multiplier approaches. While using this statistics is a natural way to construct confidence region and conduct hypothesis tests that involve 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]

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 first. In particular, one can show that, denoting the restricted and unrestricted sum of squared residuals by SSR_r and SSR_{ur}, and the number of equality to be tested as q (more generally the rank of C)

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-(k+1))} = \frac{(SSR_r - SSR_{ur})/q}{\hat{\sigma}^2},$$

is the same statistics.

In the baseball example, this is computed as the sum of squared residual obtained by running the regression with and without the restrictions.

This number tells us the increase in the sum of squared residual per restriction relative to the residual variance.

If this number is too big, it is a sign that the restriction was not consistent with the way data was generated, so we reject the null hypothesis.

$$\begin{array}{rlllll} \widehat{\log(salary)} = & 11.22 & + & .0713 years & + & .0202 gamesyr \\ & (.11) & & (.0125) & & (.0013) \\ n = 353, & SSR = 198.311 & & R^2 = .5971 & & \end{array}$$

So

$$F = \frac{(198.311 - 183.186)/3}{183.186/(353-6)} \approx 9.55$$

Critical value and the rejection region is set in an analogous way typically using 5% significance level.

In this case, the critical value for 5% significance level can be obtained by looking up the value that gives 5% tail probability for $F(3, 347)$ distribution, which is 2.60. Since F-value is above this, we reject the null hypothesis.

Clearly, as before, for a particular data set, this inequality holds or does not hold. The significance level indicates the fraction of the times we would reject the null hypothesis when in fact the hypothesis holds.

Like earlier cases, there are many other regions with the same significance level.

We choose the region considering Type II error and the logic is the same as before; we want to reject the null hypothesis when the parameter value is farther away from the null hypothesis.

There is no one-sided version to this test.

One can compute the p-value for the F-test by looking up the F-distribution tail probability above the F-statistic value.

In the baseball example, p-value is less than 1%.

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

$$\frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(N-K)} = \frac{(SST(1-R_r^2) - SST(1-R_{ur}^2))/q}{SST(1-R_{ur}^2)/(N-K)}$$
$$= \frac{(R_{ur}^2 - R_r^2)/q}{(1-R_{ur}^2)/(N-K)}.$$

A special case is to test all coefficients other than the constant term are 0. In this case,

$$\frac{(SSR_r - SSR_{ur})/(K-1))}{SSR_{ur}/(N-K)} = \frac{R_{ur}^2/(K-1)}{(1-R_{ur}^2)/(N-K)}.$$

indicate what to do when the assumptions do not hold.

Finite Sample Distribution of the OLS Estimator

Finite Sample Distribution related to the OLS Estimator

We were looking at the standardized OLS estimator:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 \sum_{j=1}^N \hat{v}_{1j}^2}}.$$

We showed that

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 \sum_{j=1}^N \hat{v}_{1j}^2}}$$

has the standard normal distribution given X and that

$$\sum_{i=1}^N (y_i - x_i'\hat{\beta})^2/\sigma^2$$

has the χ-square distribution with $N-K$ degrees of freedom.

Thus, if these two statistics are independent, then the standardized OLS estimator has the t-distribution with $N-K$ degrees of freedom.

The result holds under normality assumption on the error terms as well as under all the rest of the assumptions.

To see this, intuitively, one can observe that the OLS estimator's distribution is driven by $X'U$ and that of the $\hat{\sigma}^2$ is driven by $(I-P_X)U$. Both of them have normal distribution and that, when $E(UU'|X) = \sigma^2 I$, since

$$E(XUU'(I-P_X)|X) = 0,$$

they are uncorrelated. Intuitively, this shows independence.

The theoretical step that is missing in this argument is that $X'U$ and $(I-P_X)U$ are jointly normal.

Lecture note shows that due to a version of the spectral theorem applied to the idempotent matrices P_X and $I-P_X$, one can show that $(I-P_X)U$ depends on $N-K$ linear combinations of U and that $X'U$ depends on K linear combinations of U. Moreover, it shows that they are uncorrelated and thus independent as they are jointly normally distributed.

Next, we turn to the discussion of examining two or more linear combinations of the linear regression coefficients.

In the third lecture, we have examined

$$\frac{(C\hat{\beta} - C\beta)'[CX(X'X)^{-1}X'C']^{-1}(C\hat{\beta} - C\beta)/q}{\hat{\sigma}^2},$$

where q is the number of linear combinations (rank of C).

As we discussed, the numerator, divided by σ^2 has the χ-square distribution with q degrees of freedom, divided by q, and the denominator, divided by σ^2, has the χ-square distribution with $N-K$ degrees of freedom, divided by $N-K$.

By the same reasoning, they are independent, and thus the statistic has the F-distribution with the degrees of freedom q and $N-K$.

The approach described is the "Wald" approach.

While using this statistics is a natural way to construct confidence region and conduct hypothesis tests that involve multiple equalities, regarding hypothesis tests, there are two other approaches; the "Likelihood Ratio" approach and the "Lagrangean Multiplier" approaches.

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.

In particular, one can show that, denoting the restricted and unrestricted sum of squared residuals by SSR_r and SSR_{ur}, and the number of equality to be tested as q (more generally the rank of C)

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(N-K)} = \frac{(SSR_r - SSR_{ur})/q}{\hat{\sigma}^2},$$

is the same statistics with the Wald statistic we have examined.

In the baseball example, this is computed as the sum of squared residual obtained by running the regression with and without the restrictions.

This number tells us the increase in the sum of squared residual per restriction relative to the residual variance.

If this number is too big, it is a sign that the restriction was not consistent with the way data was generated, so we reject the null hypothesis.

$$\begin{array}{rlcll} \widehat{\log(salary)} = & 11.22 & + & .0713 years & + \quad .0202 gamesyr \\ & (.11) & & (.0125) & \quad (.0013) \\ n = 353, & R^2 = .597 & & & \\ SSR = 198.311 & & & & \end{array}$$

So

$$F = \frac{(198.311 - 183.186)/3}{183.186/(353-6)} \approx 9.55$$

Critical value and the rejection region is set in an analogous way typically using 5% significance level.

In this case, the critical value for 5% significance level can be obtained by looking up the value that gives 5% tail probability for $F(3, 347)$ distribution, which is 2.60. Since F-value is above this, we reject the null hypothesis.

Clearly, as before, for a particular data set, this inequality holds or does not hold. The significance level indicates the fraction of the times we would reject the null hypothesis when in fact the hypothesis holds.

Like earlier cases, there are many other regions with the same significance level.

We choose the region considering Type II error and the logic is the same as before; we want to reject the null hypothesis when the parameter value is farther away from the null hypothesis.

There is no one-sided version to this test.

One can compute the p-value for the F-test by looking up the F-distribution tail probability above the F-statistic value.

In the baseball example, p-value is less than 1%.

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

$$\begin{aligned}\frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(N-K)} &= \frac{(SST(1-R_r^2) - SST(1-R_{ur}^2))/q}{SST(1-R_{ur}^2)/(N-K)} \\ &= \frac{(R_{ur}^2 - R_r^2)/q}{(1-R_{ur}^2)/(N-K)}.\end{aligned}$$

A special case is to test all coefficients other than the constant term are 0. In this case,

$$\frac{(SSR_r - SSR_{ur})/(K-1))}{SSR_{ur}/(N-K)} = \frac{R_{ur}^2/(K-1)}{(1-R_{ur}^2)/(N-K)}.$$

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,

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

Karlan and Zinman (2009 Econometrica)

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.

This is the set up of Stiglitz and Weiss (1981) which implies that a higher offer rate produces riskier pool of applicants and further changes in contract rate induces different level of effort.

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.

The experiment is as follows:

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 the normal rate 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.

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

Linear Algebra

5.1. 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) When A is full 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 $trace(AB) = trace(BA)$.