WARNING!!!

THIS DOCUMENT CONTAINS TYPOS. THESE TYPOS WERE CORRECTED DURING THE LECTURE, BUT THEY MAY NOT HAVE BEEN CORRECTED IN THIS DOCUMENT

1 Fundamentals of Applied Statistics

1.1 Data Generating Process (DGP)

A fundamental methodology of modern statistics is to assume that observed data are generated by some stochastic process—i.e., some probability distribution. Given this we can define a DGP.

Suppose observed economic data are realizations of a stochastic process denoted as Z. Z is defined on a suitable probability space (Ω, F, P), where Ω is the sample space, F, is a *sigma* algebra and P is a probability function.

For each t, Z_t is a $v \times 1$ random vector.

Convention:

1. Capital letters (e.g., Z_t) denote random variables or random vectors.

2. Lower case letters (e.g., z_t) denote realizations of random variables.

The probability law P gives a complete description of the stochastic process. If P where known, we would be able to know every aspect of Z, such as the conditional means, conditional variance, etc.

P is determined by the nature of the world, it is not known by the statistician. The problem of estimation and inference arises precisely because P is unknown.

If we observe a realization of the sequence Z, then we can infer some knowledge of P from this realization. In practice, observation of the entire sequence is impossible. Instead, we have a realization $z^n = (z_1, z_2, \cdots, z_n)$ of a finite history. We call z^n a sample of size n. We usually hope that this sample is *random*.

We will learn P using the information available (z^n). Note it is impossible to learn P precisely because of the limited (i.e., finite) amount of information. This is related to the Law of Small Numbers.

We can, however, learn P arbitrarily well as the sample size n goes to ∞.

Example:

Suppose z^n is a random sample from some population with mean μ and variance $\sigma^2 < \infty$. We are interested in knowing μ. For this we use the sample mean:

$$\bar{z}^n = \frac{1}{n}\sum_{t=1}^{n} z_t \tag{1}$$

Note that:

1. $V(\bar{z}^n) = \dfrac{\sigma^2}{n-1} \neq 0$ for all n.
2. $V(\bar{z}^n) \to 0$ as $n \to \infty$.

1.2 Models

Suppose $v > 1$, then we can partition $Z_t = (Y_t, X_t)'$, where Y_t is (a scalar) real-valued random variable and X_t is a $1 \times k$ random vector. Note that $k = v - 1$.

We are interested in the relationship between Y_t and X_t—i.e., in explaining the behavior of Y_t using X_t. A function of X_t, $f(X_t)$, is used to approximate Y_t. This function is called a model or a predictor for Y_t.

In practice, a linear function is most often used:

$$f(X_t) = \alpha + \beta X_t, \tag{2}$$

where $\alpha, beta \in \Re$. Obviously, any function $f(X_t)$ may be incorrect.

1.3 Loss Function

How well the model $f(X_t)$ will explain Y_t is described by a criterion function. In general, there exists a discrepancy between $f(X_t)$ and Y_t. When $f(X_t) \neq Y_t$, a "loss" will occur. This is defined as the loss function.

A loss function $l(Y_t, f(X_t))$ is a real-valued function that describes how well the model $f(X_t)$ can explain Y_t.

$$l(Y_t, f(X_t)) = (Y_t - f(X_t))^p, \tag{3}$$

where $0 \leq p \leq \infty$, is a loss function.

These least square predictor is the loss function where $p = 2$. *This is an arbitrary choice.*

The expected loss is defined as $E[l(Y_t, f(X_t))]$, where E is taken over P. When $l(Y_t, f(X_t)) = (Y_t - f(X_t))^2$, the expected loss is mean square error (MSE).

Theorem:

$$\text{MSE}(f) = E\left[(Y_t - E(Y_t|X_t))^2\right] + E\left[E(Y_t|X_t) - f(X_t)\right]^2 \tag{4}$$

$$= V(\epsilon_t) + E\left[E(Y_t|X_t) - f(X_t)\right]^2, \tag{5}$$

$$\tag{6}$$

where $\epsilon_t = Y_t - E(Y_t|X_t)$.

There are two kinds of loss: the first is determined by the DGP, and is unavoidable. The second term comes from the specification error made by the analyst.

1.4 Best Predictor

Suppose $f(X_t)$ is a predictor for Y_t from a class of function F, and $l(Y_t, f(X_t))$ is a loss function. Then the best predictor within F solves:

$$f^* = argmin E\left[l(Y_t, f(X_t)\right], \tag{7}$$

subject to $f \in F$.

Theorem:

Let f by any measurable function of X_t, then the best predictor f^* that minimizes MSE(f) is the conditional mean $E(Y_t|X_t)$—i.e., $f^*(X_t) = E(Y_t|X_t)$.

The best predictor for MSE(f) is called the best least squares predictor.

2 Linear Regression Models

Regression Function:

Suppose the stochastic sequence $Z_t = (Y_t, X_t')'$ is i.i.d. with $E(Y_t^2) \leq \infty$. This assumptions ensures that the second moment exists, which in turn ensures that the conditional mean exists. The conditional mean $E(Y_t|X_t)$ is called the "regression function" of Y_t on X_t.

Theorem:

Suppose the conditions of the previous definition hold. Then,

$$Y_t = E(Y_t|X_t) + \epsilon_t, \tag{8}$$

where the disturbance has the property $E(\epsilon_t|X_t) = 0$.

Proof:

$$\epsilon_t = Y_t - E(Y_t|X_t). \tag{9}$$

$$\text{Then} \tag{10}$$

$$Y_t = E(Y_t|X_t) + \epsilon_t \tag{11}$$

$$\text{and} \tag{12}$$

$$E(\epsilon_t|X_t) = E\left[(Y_t - E(Y_t|X_t)|X_t\right] \tag{13}$$

$$= E(Y_t|X_t) - E\left[E(Y_t|X_t)|X_t\right] \tag{14}$$

$$= E(Y_t|X_t) - E(Y_t|X_t) \tag{15}$$

$$= 0 \tag{16}$$

Remarks:

1. The regression function $E(Y_t|X_t)$ is used to predict Y_t from knowledge of X_t.

2. The term ϵ_t is called the "regression disturbance." The fact $E(\epsilon_t|X_t) = 0$ implies that ϵ_t contains no systematic information of X_t in predicted Y_t. In other words, all information of X_t that is useful to predict Y_t has been summarized by $E(Y_t|X_t)$.

Theorem: Best Linear LS Predictor

Suppose:

1. $E(Y_t|X_t)$ is a linear function of X_t, α.

2. the sequence $\{Z_t\}$ is iid with $E(Y_t^2) < \infty$

3. $E(X_t X_t')$ is non-singular.

Then the best Linear LS Predictor that solves

$$argminE\left[E(Y_t - f(X_t)\right], \tag{17}$$

subject to $f \in A$ where A is the family of all measurable functions, is given by

$$f(X_t, \alpha^*) = X_t'\alpha^*, \tag{18}$$

where

$$\alpha^* = \left[E(X_t X_t')\right]^{-1} E(X_t Y_t), \tag{19}$$

where α^* is a $v \times 1$, X_t is $v \times 1$, Y_t is 1×1, and the $(X_t X_t')$ term is $v \times v$.

Proof

$$min(f \in A)E\left[Y_t - f(X_t)\right]^2 = min(\alpha \in \Re^k)E\left[Y_t - X_t'\alpha\right]^2 \tag{20}$$

This equation transforms choosing a function into choosing a parameter. From a constrained to an unconstrained optimization problem. This move is legitimate because we have shown earlier that the best predictor is $E(Y_t|X_t)$ and because of Assumption 1.

First Order Condition (F.O.C.)

Let's set the gradient to zero:

$$\nabla_\alpha E\left[(Y_t - X_t'\alpha^*)^2\right] = 0 \tag{21}$$

We can solve for the gradient by interchanging the expectation and derivative operators and using the chain rule.

$$\nabla_\alpha E\left[(Y_t - X_t'\alpha^*)^2\right] = E\left[\nabla_\alpha (Y_t - X_t'\alpha^*)^2\right] \tag{22}$$

$$= E\left[2(Y_t - X_t'\alpha)\nabla_\alpha (Y_t - X_t'\alpha)'\right] \tag{23}$$

$$= 2E\left[(Y_t - X_t'\alpha)\nabla_\alpha (-X_t'\alpha)'\right] \tag{24}$$

$$= -2E\left[(Y_t - X_t'\alpha)X_t'\right]. \tag{25}$$

Note that $\nabla_\alpha(-X_t'\alpha) = -X_t$.

Recall the Chain Rule

Let us be interested in:

$$h(x) = f\left(g(x)\right) \tag{26}$$

$$\text{then,} \tag{27}$$

$$\nabla_h(x) = \nabla_f\left(g(x)\right)\nabla_g(x) \tag{28}$$

$$\text{Here is an example} \tag{29}$$

$$h(x) = (x^2+1)^3 \tag{30}$$

$$\text{Note that} \tag{31}$$

$$f(x) = x^3 \tag{32}$$

$$g(x) = x^2 + 1 \tag{33}$$

$$\nabla_f(x) = 3x^2 \tag{34}$$

$$\nabla_g(x) = 2x \tag{35}$$

$$\text{Hence} \tag{36}$$

$$\nabla_f\left(g(x)\right) = 3(x^2+1)^2 \tag{37}$$

$$\nabla_h\left(x\right) = 3(x^2+1)^2(2x) \tag{38}$$

$$\tag{39}$$

The FOC implies:

$$E\left[X_t(Y_t - X_t'\alpha^*)\right] = 0 \tag{40}$$

where α^* is the value at the optimal point

$$E(X_tY_t) - E\left[X_tX_t'\alpha^*\right] = 0 \tag{41}$$

Since α^* is considered to be a constant vector parameter,

we can move it outside of the expectation$E(X_tY_t) = E(X_tX_t')\alpha^*$

Let us premultiply by $\left[E(X_tX_t')\right]^{-1}$

$$\left[E(X_tX_t')\right]^{-1} E(X_tY_t) = \alpha^* \tag{42}$$

Note the Slutzky's Theorem—by analogy

Suppose $X_n,\ n = 1, 2, \cdots$ are random vectors on $\Re^d$, X is a random vector on $\Re^d$, and Y, $Y_n,\ n = 1, 2, \cdots$ are random vectors on $\Re^m$. Then,

$$X_n \xrightarrow{p} X \tag{43}$$

$$X_n \xrightarrow{p} Y \text{ iff } \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{p} \begin{pmatrix} X \\ Y \end{pmatrix} \tag{44}$$

A proof will be provided later. For now it is enough to note that if we assume $X_n \xrightarrow{p} X$, $Y_n \xrightarrow{p} Y$, and $Z_n \xrightarrow{p} Z$, we can show

1. $X_n + Y_n \xrightarrow{p} Y$,

2. $X_n' Y_n \xrightarrow{p} X'$,

3. $\dfrac{1}{Z_n} X_n \xrightarrow{p} \dfrac{1}{Z} X$

We shall ignore the *Second Order Condition* (SOC).

Remarks:

1. α^* is the best linear LS coefficient, we have not yet obtained $\hat{\alpha}_{OLS}$.

2. The condition $E(Y_t^2) < \infty$ ensures the existence of $E(Y_t|X_t)$ and $E\left[(Y_t - X_t')^2\right]$

3. Non-singularity of $E(X_t X_t')$ ensures that

   (a) α^* is unique

   (b) there is a global minimum

4. It **must** be remembered that, in general,

   $E(Y_t|X_t) \neq X_t'\alpha^*$

Theorem Suppose the conditions of the previous theorem hold. Let $Y_t = X_t'\alpha + U_t$, then $\hat{\alpha} = \alpha^*$ if and only if $E(X_t U_t) = 0$. This is called the *orthogonality condition*. It must be remembered that the $E(\epsilon_t|X_t) = 0$ condition is fundamentally different from this one.

Note that the use of U_t denotes just a model *not* a data generating processes.

Remarks:

1. Necessary part says that if we choose α to minimize the MSE, $E[Y_t - X_t'\alpha]^2$, then the disturbance $U_t = Y_t - X_t'\alpha^*$ is automatically orthogonal to X_t. The orthogonality condition is not really a condition but the consequence of the LS estimator.

2. Sufficiency part says that if $U_t = Y_t - X_t'\alpha$ is orthogonal to X_t, then α must be the best linear LS coefficient α^*.

Proof:

First we shall prove necessity. show if $\alpha = \alpha^*$, then $E(U_t U_t) = 0$. When $\alpha = \alpha^*$ we have:

$$E(X_t U_t) = E\left[X_t(Y_t - X_t'\alpha^*)\right] = 0. \tag{45}$$

This is true by the FOC.

We shall now prove sufficiency. If $E[X_t U_t] = 0$, then $\alpha = \alpha^*$. Because $U_t = Y_t - X_t'\alpha$, $E(X_t U_t) = 0$ implies:

$$E\left[X_t(Y_t - X_t'\alpha)\right] = 0 \tag{46}$$

$$E\left[X_t Y_t\right] - E\left[X_t X_t'\right]\alpha = 0 \tag{47}$$

$$\left[E(X_t X_t')\right]^{-1} E(X_t Y_t) = \alpha \tag{48}$$

$$= \alpha^* \text{ by definition} \tag{49}$$

2.1 Classical Assumptions

Assumption A1. $Y_t = X_t'\alpha + U_t, \;\; t = 1, 2, 3, \cdots, n$

Assumption A2. X_t is a nonstochastic $k \times 1$ vector, $t = 1, 2, 3, \cdots, n$. X_t is a fixed constant, there is no disturbance associated with X_t. Therefore, X_t may always be moved outside the expectation.

Assumption A3. The $k \times k$ matrix $X_t X_t'$ is non-singular for every $n \geq 1$.

Assumption A4. $E[U_t] = 0$, $t = 1, 2, 3, \cdots, n$. Since X_t is assumed to be nonstochastic (A2), (A4) implies that $E[X_t U_t] = 0$. (A4) *always* holds if there is an intercept.

Assumption A5. $E\left[UU'\right] = \sigma^2 I$, where U is an $n \times 1$ matrix and I is an $n \times n$ identity matrix.

Remarks:

(a) It follows that $E[U_t^2] = \sigma^2$, $t = 1, 2, 3, \cdots, n$. But the $4^t h$ moment may vary with t. (A5) is, therefore, weaker than the i.i.d. assumption.

(b) $cov(U_t U_\tau) = 0$ for all $t \neq \tau$. iid implies these two items, but they do not imply iid.

(c) (A5) is the homoscedasticity assumption. It is similar to $E(U_t^2 | X_t) = \sigma^2$.

Theorem: Existence

Suppose Assumptions (A1) to (A3) hold, then the OLS estimator $\hat{\alpha}$ exists. And

$$\hat{\alpha} = \left[X_t X_t'\right]^{-1} X_t Y_t \tag{50}$$

The proof for this follows easily from the foregoing. It is analogous to the proof for α^*.