# Empirical Analysis Lecture Note I

Joonhwi Joo

February 14, 2015

joonhwi@uchicago.edu

This note is based on Azeem Shaikh's lecture in 2012 Fall quarter.

## Contents

**I Introduction to the Large Sample Theory 2**

**1 Motivation 2**

**2 Preliminary: Some Necessary Tools 3**
2.1 Some Basic Definitions . . . 3
2.2 Inequalities . . . 3

**3 Convergence Concepts 5**
3.1 Convergence in Probability . . . 5
3.2 $L^q$ convergence . . . 6
3.3 Convergence in Distribution . . . 7
3.4 Central Limit Theorem . . . 8
3.5 Relationships Between Modes of Convergences . . . 9
3.6 The Delta Method . . . 10
3.7 Tightness . . . 11
3.8 Stochastic Order Notations . . . 11

**4 Hypothesis Testing 13**
4.1 Basic Concepts . . . 13

**II Linear Regression 16**

**5 Conditional Expectations 16**

**6 Linear Regression 19**
6.1 Interpretations of Linear Regression . . . 19
6.2 Population Regression . . . 20
6.3 Partitioned Regression . . . 20
6.4 The Ordinary Least Squares Estimator and its Properties . . . 21
6.5 Measures of Fit . . . 27
6.6 Hypothesis Testing in Linear Regression . . . 28

# Part I
# Introduction to the Large Sample Theory

## 1 Motivation

Let $X_1, ..., X_n \in \mathbb{R}^k$ be independent and identically distributed random vectors, according to a distribution $P$. In a typical statistical problem, we want to learn some feature of $P$, denoted by $\theta(P)$. The features we might want to learn include:

- Mean $\mu(P)$
- Variance $\sigma^2(P)$
- Parameters of a linear regression model

To learn something about such features, we could

- Estimate $\theta(P)$ by a function $\hat{\theta}_n = \hat{\theta}_n(X_1, ..., X_n)$
- Test the null hypothesis $\theta(P) = \theta_0$
- Construct a confidence region for $\theta(P)$

and so on.

However, when we only deal with finite samples, we need to make strong assumptions. In order to relax those strong (and many times unrealistic) assumptions, we study the large sample theory.

# 2 Preliminary: Some Necessary Tools

In this section, we are going to present some necessary tools. The tools will include terminology, definitions, and some basic theorems and lemmas.

## 2.1 Some Basic Definitions

**Definition.** (Independence of Random Variables) Let $X$ and $Y$ be random variables. $X$ and $Y$ are independent if $P(X \leq x)P(Y \leq y) = P(X \leq x, Y \leq y)$.

**Definition.** (Identically Distributed Random Variables) Let $X_i$ and $X_j$ be random variables. $X_i$ and $X_j$ are identically distributed if $\forall x \in \mathbb{R}$, $P(X_i \leq x) = P(X_j \leq x)$.

**Definition.** (Moments) Let $X$ be a random variable. The $k$-th moment of $X$ is defined by $E[X^k]$.

**Definition.** (Central Moments) Let $X$ be a random variable. Let $E[X]$ be the first moment of $X$. Then, the $k$-th central moment of $X$ is defined by $E[(X - E[X])^k]$.

**Definition.** (Covariance) Let $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ be sequences of random variables with distribution $P$.

$$\begin{aligned} Cov(X_i, Y_i) &:= E\{[X_i - E[X_i]][Y_i - E[Y_i]]\} \\ &= E[X_iY_i] - E[X_i]E[Y_i] \end{aligned}$$

**Proposition 2.1.** *(Existence of Covariance) Let $\{(X_i, Y_i)\}_{i=1}^n$ be iid sequences of random vectors with joint distribution $P$. Suppose that $Var(X_i) < \infty$, $Var(Y_i) < \infty$. Then,*

$$Cov(X_i, Y_i) < \infty$$

**Definition.** (Correlation) Let $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ be sequences of random variables. The correlation between $X_i$ and $Y_i$,

$$Corr(X_i, Y_i) = \rho_{X_i,Y_i} = \frac{Cov(X_i, Y_i)}{\sqrt{Var(X_i)}\sqrt{Var(Y_i)}}$$

**Proposition 2.2.** *(Correlation is Bounded Above by 1) $Corr(X_i, Y_i) \leq 1$.*

*Proof.* Using the Cauchy-Schwartz inequality (in the next subsection) on $X = X_i - E[X_i]$, $Y = Y_i - E[Y_i]$, we have the desired result. □

## 2.2 Inequalities

**Lemma 2.1.** *(Cauchy-Schwartz Inequality) Let $X, Y$ be random variables such that $E[X^2] < \infty$ and $E[Y^2] < \infty$. Then, $\{E[XY]\}^2 \leq E[X^2]E[Y^2]$. The equality holds if and only if $Y = \alpha X$ for some $\alpha$.*

*Proof.* Assume $E[X^2], E[Y^2] > 0$. (Otherwise, equality is trivial.) We examine $E[(X - \alpha Y)^2] \geq 0$.

$$0 \leq E[(X - \alpha Y)^2] = E[X^2] - 2\alpha E[XY] + \alpha^2 E[Y^2] \tag{2.1}$$

Choose $\alpha$ such that $\alpha$ minimizes (2.1), and substitute back. That is, $\alpha = \frac{E[XY]}{E[Y^2]}$. Then, we get

$$\begin{aligned} 0 &\leq E[X^2] - 2\frac{\{E[XY]\}^2}{E[Y^2]} + \frac{\{E[XY]\}^2}{E[Y^2]} \\ 0 &\leq E[X^2]E[Y^2] - \{E[XY]\}^2 \end{aligned}$$

Which is the desired inequality. □

**Lemma 2.2.** *(Markov Inequality) Let $X$ be a random variable. Let $\epsilon > 0$. Let $q > 0$. Then,*

$$P(|X| \geq \epsilon) \leq \frac{1}{\epsilon^q} E(|X|^q)$$

*Proof.* $P(|X| > \epsilon) = E[\mathbf{1}(|X| > \epsilon)]$ and $\mathbf{1}(|X| > \epsilon) \leq \frac{|X|^q}{\epsilon^q}$. ($\because$ if $|X| > \epsilon$, $\mathbf{1}(|X| > \epsilon) = 1 \leq \frac{|X|^q}{\epsilon^q}$. Otherwise $\mathbf{1}(|X| > \epsilon) = 0$) By taking expectation, we have the desired result. □

**Corollary.** *(Chebychev's Inequality) Let $X$ be a random variable. Let $g(x)$ be a nonnegative function. Then, for any $r > 0$,*

$$P(g(X) \geq r) \leq \frac{E[g(X)]}{r}$$

*Proof.* The proof is exactly the same by replacing $g(X)$ in $|X|$ of the above lemma. □

**Example 2.1.** Let $X_1, ..., X_n$ be iid. Let $P \sim Bernoulli(q)$ where $q \in (0,1)$. Let $\alpha \in (0,1)$ be given. We want to construct a nontrivial confidence interval $C_n = C_n(X_1, ..., X_n)$ such that $P(\mu(P) \in C_n) = 1 - \alpha$.

One possibility is to use the Chebychev's inequality to obtain:

$$P(|\bar{X}_n - \mu(P)| > \epsilon) = P((\bar{X}_n - \mu(P))^2 > \epsilon^2) \leq \frac{Var[\bar{X}_n]}{\epsilon^2} = \frac{Var[X_i]}{n\epsilon^2} = \frac{q(1-q)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2} \leq \alpha$$

Therefore, by setting $\epsilon = \frac{1}{2\sqrt{\alpha n}}$, we have $C_n = \left(\bar{X}_n - \frac{1}{2\sqrt{\alpha n}}, \bar{X}_n + \frac{1}{2\sqrt{\alpha n}}\right)$.

*Remark.* This is a very rough bound, and it could be developed using the characteristics of the distributions later on.

**Lemma 2.3.** *(Jensen's Inequality) Let $X$ be a random variable. Let $g : \mathbb{R} \to \mathbb{R}$ be convex. Then,*

$$E[g(X)] \geq g(E[X])$$

*Proof.* Let $\mathbb{L} = \{L : linear, \forall x, L(x) \leq g(x)\}$. Define $g$ such that

$$g(x) = \sup_{L \in \mathbb{L}} L(x)$$

Then,

$$E[g(x)] = E\left[\sup_{L \in \mathbb{L}} L(x)\right] \geq \sup_{L \in \mathbb{L}} E[L(x)] = \sup_{L \in \mathbb{L}} L(E(x)) = g(E[x])$$

□

Alternatively, we could prove this as follows:

*Proof.* Let $l(x)$ be a tangent line to $g(x)$ at the point $g(E[X])$. (Note that $E[X]$ is a constant on the real line, not a function here.) Write $l(x) = a + bx$ for some $a$ and $b$.

Because $g$ is convex, $g(x) \geq a + bx$ for all $x$. Because expectations preserve inequalities, we have the following:

$$E[g(X)] \geq E[a + bX] = a + bE[X] = l(E[X]) = g(E[X])$$

The last equality holds since $l(x)$ is set up to be tangent at the point $x = E[X]$. □

**Corollary.** *(Jensen's Inequality for Concave Functions) Let $X$ be a random variable. Let $h : \mathbb{R} \to \mathbb{R}$ be concave. Then,*

$$E[h(X)] \leq h(E[X])$$

**Corollary.** *(Existence of a Higher Moment is a Sufficient Condition for the Existence of a Lower Moment) Suppose $E[|X|^k] < \infty$. Let $0 < j < k$. Then, $E[|X|^j] < \infty$.*

*Proof.* Let $\lambda \in (0,1)$ such that $k\lambda = j$. Let $h(X) = X^\lambda$ and $h(.)$ is concave. Let $E[|X|^k] = \alpha$. Then, it follows immediately by the above corollary that

$$E[h(|X|)] = E[|X|^{k\lambda}] = E[|X|^j] \leq h(E[|X|^k]) = (E[|X|^k])^\lambda = \alpha^\lambda$$

Therefore $E[|X|^j] < \infty$. □

# 3 Convergence Concepts

## 3.1 Convergence in Probability

We begin from the following two definitions.

**Definition.** (Convergence in Probability) Let $\{X_n\}$ be a sequence of random vectors. Let $X$ be a random vector. Then, $X_n \to_p X$ if $\forall \epsilon > 0$, $P(|X_n - X| \geq \epsilon) \to 0$ as $n \to \infty$.

**Definition.** (Divergence of a Random Sequence)
Let $\{X_n\}$ be a sequence of random variables. $X_n \to_p \infty$ if $\forall c > 0$, $P(X_n > c) \to 1$.
Let $\{X_n\}$ be a sequence of random variables. $X_n \to_p -\infty$ if $\forall c < 0$, $P(X_n < c) \to 1$.

One of the most important examples of convergence in probability is the weak law of large numbers (WLLN).

**Theorem 3.1.** *(Weak Law of Large Numbers) Let $X_1, X_2, ...$ be iid random variables with $E[X_i] = \mu(P)$ and $Var[X_i] = \sigma^2(P) < \infty$. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then, $\forall \epsilon > 0$,*

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \epsilon) = 1$$

*That is, $\bar{X}_n$ converges in probability to $\mu$.*[1]

*Proof.* Using the Chebychev's inequality, we have

$$P(|\bar{X}_n - \mu(P)| \geq \epsilon) = P((\bar{X}_n - \mu(P))^2 \geq \epsilon^2) \leq \frac{E(\bar{X}_n - \mu(P))^2}{\epsilon^2} = \frac{Var\bar{X}_n}{\epsilon^2} = \frac{\sigma^2(P)}{n\epsilon^2}$$

Therefore, as $n \to \infty$, $P(|\bar{X}_n - \mu(P)| \geq \epsilon) \to 0$. □

**Example 3.1.** (Sample Mean as a Consistent Estimator of Population Mean) Let $X_1, ..., X_n$ be iid, follows a distribution $P$. Suppose $E(|X_i|) < \infty$. Then, $\bar{X}_n \to_p \mu(P)$. So, $\bar{X}_n$ is a consistent estimator of $\mu(P)$.

The next theorem states that the marginal convergence in probability implies joint convergence in probability.

**Theorem 3.2.** *Let $\{X_n\}_{n=1}^\infty (\in \mathbb{R}^k)$ be a sequence of random vectors. Let $X (\in \mathbb{R}^k)$ be a random vector. Let $X_j (\in \mathbb{R})$ be the $j$-th component of $X$. Let $X_{n,j} (\in \mathbb{R})$ be the $j$-th component of $X_n$. Suppose $X_{n,j} \to_p X_j$ as $n \to \infty$. Then, $X_n \to_p X$ as $n \to \infty$.*

*Proof.* Let $\epsilon > 0$.

$$\begin{aligned} P(|X_n - X| > \epsilon) &= P\left(\sum_{j=1}^k (X_{n,j} - X_j)^2 > \epsilon^2\right) \\ &\leq P\left(\bigcup_{j=1}^k \{(X_{n,j} - X_j)^2 > \epsilon^2\}\right) \\ &\leq \sum_{j=1}^k P\left(|X_{n,j} - X_j| > \epsilon\right) \to 0 \end{aligned}$$

[2] □

When we deal with continuous functions and probability transformations, the continuous mapping theorem is used extensively. The theorem states that continuous function preserves the convergence in probability.

**Theorem 3.3.** *(Continuous Mapping for Convergence in Probability) Let $\{X_n\}_{n=1}^\infty (\in \mathbb{R}^k)$ be a sequence of random vectors. Let $X (\in \mathbb{R}^k)$ be a random vector. Let $g : \mathbb{R}^k \to \mathbb{R}^l$ be continuous on a set $C$ which $P(X \in C) = 1$. Suppose $X_n \to_p X$. Then, $g(X_n) \to_p g(X)$ as $n \to \infty$.*

[1] In a more general version of WLLN, we only have to assume that $\mu$ is finite.

[2] The last inequality is because $P(A \cup B) \leq P(A) + P(B)$. This is sometimes called the Bonferroni's inequality.

*Proof.* Let $\epsilon > 0$. $\forall \delta > 0$,

$$\begin{aligned} & P(|g(X_n) - g(X)| > \epsilon) \\ = \ & P(|g(X_n) - g(X)| > \epsilon \text{ and } |X_n - X| < \delta) + P(|g(X_n) - g(X)| > \epsilon \text{ and } |X_n - X| \geq \delta) \\ \leq \ & P(|g(X_n) - g(X)| > \epsilon \text{ and } |X_n - X| < \delta) + P(|X_n - X| \geq \delta) \\ \leq \ & P(|X_n - X| < \delta) + P(|X_n - X| \geq \delta) \end{aligned}$$

And as $n \to \infty$, $P(|X_n - X| \geq \delta) \to 0$. The first term converges to zero as $\delta \searrow 0$. □

**Example 3.2.** ($S^2$ as a consistent estimator of $\sigma^2(P)$) Let $X_1, X_2, ..., X_n$ be iid sequence of random vectors. Suppose $\sigma^2(P) < \infty$. A natural (unbiased) estimator of $\sigma^2(P)$ is

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

so that $E(S_n^2) = \sigma^2(P)$.

For consistency,

$$S_n^2 = \frac{n}{n-1} \left( \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}^2 \right) = f\left( \frac{n}{n-1}, \frac{1}{n} \sum_{i=1}^{n} X_i^2, \bar{X}_n \right) \to_p f(1, E(X^2), E(X)) = E[X^2] - E[X]^2 = \sigma^2(P) \tag{3.1}$$

because $f$ is continuous.

## 3.2 $L^q$ convergence

*Convergence in q-th moment*, or sometimes called $L^q$ convergence, is useful as well. The definition is as follows:

**Definition.** ($L^q$-convergence) Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random vectors. Let $q \geq 1$. $X_n \to_{L^q} X$ if $E(|X_n - X|^q) \to 0$.

$L^q$ convergence of $X_n$ is a sufficient condition for the convergence in probability. However, the converse may not hold. The next proposition and the example illustrate this.

**Proposition 3.1.** *Let $X_n \to_{L^q} X$. Then, $X_n \to_p X$.*

*Proof.* By Markov's inequality, we have $P(|X_n - X| > \epsilon) = P(|X_n - X|^q > \epsilon^q) \leq \frac{E(|X_n - X|^q)}{\epsilon^q} \to 0$ as $n \to \infty$. □

**Example 3.3.** (Convergence in Probability does not Imply $L^q$-convergence.) Let $q = 1$. Let $X = 0$. Let

$$X_n = \begin{cases} n & \text{with probability } \frac{1}{n} \\ 0 & \text{otherwise} \end{cases}$$

Then $X_n \to_p 0$ but $E(X_n - X) = E(X_n) = 1 \nrightarrow 0$.

As in the above example, in general convergence in probability does not imply convergence in moments. However, it is true under certain set of assumptions, such as, $\exists B > 0$ such that $P(|X_n| \leq B) = 1 \ \forall n$.

*Remark.* $L^q$-convergence $\nrightarrow$ almost sure convergence, and almost sure convergence $\nrightarrow$ $L^q$-convergence.[3]

[3]

*Definition.* (Almost Sure Convergence) A sequence of random vectors $\{X_n\}_{n=1}^{\infty}$ converges almost surely to $X$ if $P(\lim_{n\to\infty} |X_n - X| < \epsilon) = 1$.

*Remark.* This definition is more or less similar to the pointwise convergence of a sequence of functions, except that the convergence need not occur on measure zero sets. Also note that almost sure convergence implies convergence in probability, but not vice versa.

### 3.3 Convergence in Distribution

Convergence in distribution is a weaker concept than convergence in probability. Convergence in distribution is sometimes called as weak convergence.

**Definition.** (Convergence in Distribution) Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random vectors. Let $X(\in \mathbb{R}^k)$ be a random vector. $X_n \to_d X$ if

$$P(X_n \leq x) \to P(X \leq x)$$

$\forall x$ at which $P(X \leq x)$ is continuous.

*Remark.* Here, $X \leq x$ is interpreted elementwise.

Why do we require convergence only at the points $x$ at which $P(X \leq x)$ is continuous? The next example illustrates what will happen if we just require it everywhere.

**Example 3.4.** Let $X_n = \frac{1}{n}$. Let $X = 0$. $X_n \to_p X$, so we might expect that $X_n \to_d X$ as well. Indeed, for $x > 0$, $P(X_n \leq x) \to P(X \leq x)$. However, for $x = 0$, $0 = P(X_n \leq x) \nrightarrow P(X \leq x) = 1$.

The following lemma is straightforward from the definition.

**Lemma 3.1.** *Suppose $X \sim P$ and $Y \sim P$. $X_n \to_d X$ if and only if $X_n \to_d Y$.*

We present the Portmanteau's lemma below without proofs. The lemma provides equivalent formulations of convergence in distribution.

**Lemma 3.2.** *(Portmanteau) The followings are all equivalent*
*(i) $X_n \to_d X$*
*(ii) $E[f(X_n)] \to E[f(X)]$ for any continuous, bounded, real valued $f$.*
*(iii) $E[f(X_n)] \to E[f(X)]$ for any Liptshitz,[4] bounded, real valued $f$.*
*(iv) $\liminf_{n\to\infty} E[f(X_n)] \geq E[f(X)]$ for any nonnegative, continuous function $f$.*
*(v) $\liminf_{n\to\infty} P(X_n \in G) \geq P(X \in G)$ for any open $G$.*
*(vi) $\limsup_{n\to\infty} P(X_n \in H) \leq P(X \in H)$ for any closed $H$.*
*(vii) $P(X_n \in B) \to P(X \in B)$ for any Borel sets $B$ with $P(X \in \partial B) = 0$*

**Lemma 3.3.** *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random vectors. Let $X$ be a random vector such that $X_n \to_d X$. Let $\{Y_n\}_{n=1}^{\infty}$ be a sequence of random vectors such that $|X_n - Y_n| \to_p 0$. Then, $Y_n \to_d X$.*

*Proof.* We use (iii) of the Portmanteau's lemma. We want to show that $\forall f$, bounded Lipshitz real-valued function, $E[f(Y_n)] \to E[f(X)]$. Because we know that $X_n \to_d X$, it is now sufficient to show that

$$E[f(Y_n)] - E[f(X_n)] \to 0$$

Let $B > 0$, and $f \leq B$. Let $\epsilon > 0$. There exists a $L_\epsilon$ such that $|f(X_n) - f(Y_n)| \leq L_\epsilon |X_n - Y_n|$. Then,

$$\begin{aligned} |f(Y_n) - f(X_n)| &= |f(Y_n) - f(X_n)|\mathbf{1}(|X_n - Y_n| > \epsilon) + |f(Y_n) - f(X_n)|\mathbf{1}(|X_n - Y_n| \leq \epsilon) \\ &\leq 2B\mathbf{1}(|X_n - Y_n| > \epsilon) + L_\epsilon \end{aligned}$$

Take expectations, we obtain:

$$E[f(Y_n)] - E[f(X_n)] \leq 2BP(|X_n - Y_n| > \epsilon) + L_\epsilon < \delta$$

Send $n \to \infty$, we have the desired result. □

**Lemma 3.4.** *(Convergence in Probability Implies Convergence in Distribution) Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random vectors. Let $X$ be a random vector such that $X_n \to_p X$. Then, $X_n \to_d X$.*

[4] **Definition.** (Libshitz) $f$ is Libshitz continuous if $\forall \epsilon > 0$, $\exists L > 0$ such that $|f(x) - f(y)| \leq L|x - y|$.

*Proof.* Apply the above lemma with $Y_n = X_n$, $X_n = X$, $X = X$. □

Therefore we see that convergence in probability implies convergence in distribution. However, the converse is generally false, except when $X = c$, a constant.

**Lemma 3.5.** *(Convergence in Distribution to a Constant Implies Convergence in Probability) Let $c (\in \mathbb{R}^k)$ be a nonrandom vector. Let $\{X_n\}_{n=1}^{\infty} (\in \mathbb{R}^k)$ be a sequence of random vectors such that $X_n \to_d c$ as $n \to \infty$. Then, $X_n \to_p c$.*

*Proof.* Let $\epsilon > 0$.

$$P(|X_n - c| > \epsilon) < P(|X_n - c| \geq \epsilon) = P(X_n \in B_\epsilon(c)^c)$$

Therefore,

$$\begin{aligned} \limsup_n P(|X_n - c| > \epsilon) &\leq \limsup_n P(|X_n - c| \geq \epsilon) \\ &\leq P(c \in B_\epsilon^c(c)) \text{ by Portmanteau's lemma (vi)} \\ &= 0 \end{aligned}$$

[5] □

If $X_n \to_p X$ and $Y_n \to_p X$, then $X_n \to_p Y_n$. However, marginal convergence in distribution does not imply joint convergence in distribution. The next example illustrates the fact.

**Example 3.5.** Let

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \sim N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_n \\ \rho_n & 1 \end{pmatrix} \right)$$

Trivially the marginal distributions $X_n \to_d N(0,1)$ and $Y_n \to_d N(0,1)$. But, the joint distribution need not even settle down. For example, consider $\rho_n = (-1)^n$.

However, there is an exception, when one of the random vector converges in distribution to a constant $c$. The next lemma illustrates this exception.

**Lemma 3.6.** *Let $\{(X_n, Y_n)\}_{n=1}^{\infty}$ be a sequence of random vector such that $X_n \to_d X$ random, and $Y_n \to_d c$ constant. Then, $(X_n, Y_n) \to_d (X, c)$ jointly.*

*Proof.* By the previous lemma, $Y_n \to_p c$. Therefore,

$$|(X_n, Y_n) - (X_n, c)| \quad = \quad |Y_n - c| \text{ Euclidean difference is the same}$$

and $P(|Y_n - c| > \epsilon) \to 0$. Therefore, it is sufficient to show that $(X_n, c) \to_d (X, c)$. Equivalently, we want to show that for any continuous and bounded function $f$, $E[f(X_n, c)] \to E[f(X, c)]$.

Because $c$ is a constant, we can rewrite $f$ as a function only of $X_n$, that is, $f(X_n) = f(X_n, c)$. However, since $X_n \to_d X$, we have:

$$E[f(X_n, c)] = E[f(X_n)] \to E[f(X)] = E[f(X, c)]$$

□

## 3.4 Central Limit Theorem

We state the uni-variate CLT and Cramer-Wold lemma without proofs.

**Lemma 3.7.** *(Univariate Central Limit) Let $\{X_n\}_{n=1}^{\infty} (\in \mathbb{R})$ be an iid sequence of random variable with distribution $P$. Suppose that $\sigma^2(P) < \infty$. Then,*

$$\sqrt{n}(\bar{X}_n - \mu(P)) \to_d N(0, \sigma^2(P))$$

**Lemma 3.8.** *(Cramer-Wold) Let $\{X_n\}_{n=1}^{\infty} (\in \mathbb{R}^k)$ be a sequence of random vectors. Let $X (\in \mathbb{R}^k)$ be a random vector. Then, $X_n \to_d X$ if and only if $\forall t \in \mathbb{R}^k$, $t'X_n \to_d t'X$.*

[5] It is sufficient to prove that $\limsup_n P(|X_n - c| > \epsilon) \to 0$ since $\lim_n P(|X_n - c| > \epsilon) \leq \limsup_n P(|X_n - c| > \epsilon)$.

Now we state and prove the multivariate central limit theorem.

**Lemma 3.9.** *(Multivariate Central Limit) Let $\{X_n\}_{n=1}^{\infty}$ be an iid sequence of random vectors with distribution $P$. Suppose $\Sigma(P) < \infty$. Then,*

$$\sqrt{n}(\bar{X}_n - \mu(P)) \to_d N(0, \Sigma(P))$$

*Proof.* $\forall t \in \mathbb{R}^k$, note that $t'X_i$ has mean $t'\mu(P)$ and variance $t'\Sigma(P)t < \infty$.

$$t'\left[\sqrt{n}(\bar{X}_n - \mu(P))\right] = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} t'X_i - t'\mu(P)\right] \to_d N(0, t'\Sigma(P)t) = t'N(0, \Sigma(P))$$

by the univariate central limit theorem. □

## 3.5 Relationships Between Modes of Convergences

**Lemma 3.10.** *(Continuous Mapping for Convergence in Distribution) Let $\{X_n\}_{n=1}^{\infty}(\in \mathbb{R}^k)$ be a sequence of random vectors. Let $X(\in \mathbb{R}^k)$ be a random vector. Suppose $X_n \to_d X$. Let $g : \mathbb{R}^k \to \mathbb{R}^d$ which is continuous at each point $x \in C$ and $P(X \in C) = 1$. Then, $g(X_n) \to_d g(X)$.*

*Proof.* By Portmanteau's lemma, it is enough to show that $\forall F$ closed,

$$\limsup_{n\to\infty} P(g(X_n) \in F) \leq p(g(X) \in F)$$

or equivalently,

$$\limsup_{n\to\infty} P(X_n \in g^{-1}(F)) \leq p(X \in g^{-1}(F))$$

where $g^{-1}(F) = \{x \in \mathbb{R}^k : g(x) \in F\}$. Here, $g^{-1}(F)$ may not be closed. However, $g^{-1}(F) \subset cl(g^{-1}(F)) \subset g^{-1}(F) \cup C^c$.[6]

Therefore,

$$\begin{aligned} \limsup_{n\to\infty} P(X_n \in g^{-1}(F)) &\leq \limsup_{n\to\infty} P(X_n \in cl(g^{-1}(F))) \\ &\leq P(X \in cl(g^{-1}(F))) \text{ Because } X_n \to_d X, \text{ using Portmanteau's lemma (vi)} \\ &\leq P(X \in cl(g^{-1}(F)) \cup C^c) \\ &= P(X \in g^{-1}(F)) \end{aligned}$$

□

**Lemma 3.11.** *(Slutsky) Let $\{X_n\}_{n=1}^{\infty}(\in \mathbb{R}^k)$ and $\{Y_n\}_{n=1}^{\infty}(\in \mathbb{R}^k)$ be random vectors. Let $X_n \to_d c$ and $Y_n \to_d Y$. Then,*

*(i) $X_n + Y_n \to_d c + Y$*

*(ii) $X_n Y_n \to_d cY$*

*Proof.* We have $X_n \to_p c$ by the previous lemma. Then, use the continuous mapping theorem, since additions and multiplications are continuous mappings. □

**Example 3.6.** Let $\{X_i\}_{i=1}^{n}$ be an iid sequence of random variables on $\mathbb{R}$ with distribution $P$. Suppose $\sigma^2(P) < \infty$. Then, the central limit theorem implies

$$\sqrt{n}(\bar{X}_n - \mu(P)) \to_d N(0, \sigma^2(P))$$

and we know that $S_n^2 \to_p \sigma^2(P)$, by (3.1). If, in addition, $\sigma^2(P) < \infty$, then the continuous mapping theorem implies

$$\frac{1}{S_n^2} \to \frac{1}{\sigma^2(P)}$$

So, by Slutsky's lemma,

$$\sqrt{n}\frac{(\bar{X}_n - \mu(P))}{S_n} \to_d \frac{1}{\sigma(P)} N(0, \sigma^2(P)) = N(0, 1)$$

[6] The second inclusion: Take $x \in cl(g^{-1}(F))$. $\exists x_n \in g^{-1}(F)$ such that $x_n \to x$.
Case 1. $x \in C$. Then $g(x_n) \to g(x) \in F$ since $F$ is closed.
Case 2. $x \in C^c$.

## 3.6 The Delta Method

The delta method allows us to deduce the limiting distribution of $\phi(T_n) - \phi(\theta)$ from the limiting distribution of $T_n - \theta$.

**Theorem 3.4.** *(Delta Method) Let $\{X_n\}_{n=1}^{\infty}(\in \mathbb{R}^k)$ and $X(\in \mathbb{R}^k)$ be random vectors. Let $\tau_n$ be a sequence of real numbers such that $\tau_n \to \infty$. Let $\theta(\in \mathbb{R}^k)$ such that*

$$\tau_n(X_n - \theta) \to_d X$$

*Suppose $g : \mathbb{R}^k \to \mathbb{R}$ be differentiable at $\theta$. Then,*

$$\tau_n(g(X_n) - g(\theta)) \to_d Dg(\theta)X$$

[7]

*Proof.* Because $g$ is differentiable at $\theta$,

$$g(x) - g(\theta) = Dg(\theta) \cdot (x - \theta) + R(x - \theta)$$

and $R(h) = o(|h|)$ as $h \to 0$[8] by Taylor's theorem. Define $\frac{R(h)}{|h|} = 0$. Then we have

$$\begin{aligned} g(X_n) - g(X) &= Dg(\theta) \cdot (X_n - \theta) + R(X_n - \theta) \\ \tau_n[g(X_n) - g(X)] &= Dg(\theta) \cdot \tau_n[X_n - \theta] + \tau_n R(X_n - \theta) \end{aligned}$$

By Continuous mapping theorem,

$$Dg(\theta) \cdot \tau_n(X_n - \theta) \to_d Dg(\theta)X$$

So, it is sufficient to show that $\tau_n R(X_n - \theta) \to_p 0$. We could modify this as

$$\tau_n R(X_n - \theta) = |\tau_n R(X_n - \theta)| \frac{R(X_n - \theta)}{|X_n - \theta|} \tag{3.2}$$

(1) $|\tau_n R(X_n - \theta)| \to_d X$
(2) $\frac{R(X_n-\theta)}{|X_n-\theta|} \to_p 0$
So the proof is complete. □

*Remark.* Because $R(X_n - \theta) := 0$ when $X_n = 0$, it does not matter when the left-hand side of (3.2) is 0.

*Remark.* This proof does not require $Dg(\theta) \neq 0$.

**Corollary.** *If $X \sim N(0, \Sigma)$, then $\tau_n(g(X_n) - g(\theta)) \to_d N(0, Dg(\theta)\Sigma(Dg(\theta))')$*

**Example 3.7.** (Bernoulli Revisited) Let $X_1, ..., X_n$ iid $P \sim$ Bernoulli$(q)$, where $q \in (0, 1)$. By central limit theorem, we know that $\sqrt{n}(\bar{X}_n - q) \to_d N(0, q(1 - q))$. Let $g(q) = q(1 - q)$. We want to know the distribution of an estimator of $g(q)$, namely, $g(\bar{X}_n)$. By the delta method,

$$\begin{aligned} \sqrt{n}(g(X_n) - g(q)) &\to N(0, \{Dg(q)\}^2 g(q)) \\ &= N(0, (1 - 2q)^2 q(1 - q)) \end{aligned}$$

**Example 3.8.** (Second-order Delta Method) In the previous example, if $q = \frac{1}{2}$, the (first-order) delta method does not give any useful information. We could use the second-order delta method.

$$g(x) - g(q) = Dg(q)(x - q) + \frac{1}{2}D^2 g(q)(x - q)^2 + o(|x - q|^2)$$

Therefore, at $q = \frac{1}{2}$,

$$\begin{gathered} [g(\bar{X}_n) - g(q)] = -(\bar{X}_n - q)^2 + o(|\bar{X}_n - q|^2) \\ n[g(\bar{X}_n) - g(q)] = -n(\bar{X}_n - q)^2 + no(|\bar{X}_n - q|^2) \end{gathered}$$

where $-n(\bar{X}_n - q)^2 = -\left[\sqrt{n}(\bar{X}_n - q)\right]^2 \to -\left[N\left(0, \frac{1}{4}\right)\right]^2 = -\frac{1}{4}[N(0, 1)]^2 = -\frac{1}{4}\chi_1^2$

[7] $Dg(c)$ is the matrix of partial derivatives of $g$ evaluated at $c$. Notice that $Dg(\theta)X \in \mathbb{R}$ as well.

[8] That is, $\lim_{h \to 0} \frac{R(h)}{|h|} = 0$.

**Example 3.9.** (Estimation of Covariance and Correlation) A consistent estimator for $Cov(X, Y)$ is

$$\hat{S}_{X_n,Y_n} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)$$

By continuous mapping theorem, we could propose a consistent estimator for $\rho_{X,Y}$ as

$$\hat{\rho}_{X,Y} = \frac{\hat{\sigma}_{X_n,Y_n}}{S_{X_n}S_{Y_n}}$$

If we further assume that $E[X_i^4], E[Y_i^4] < \infty$, then using the delta method, we could derive the distribution of $\hat{\rho}_{X_n,Y_n}$. The key insight for this is, we can write this as a smooth function of averages of the following five variables: $(X_i, Y_i, X_iY_i, X_i^2, Y_i^2)$.

## 3.7 Tightness

**Definition.** (Tightness) Let $\{X_n\}_{n=1}^{\infty}(\in \mathbb{R}^k)$ be a sequence of random vectors. $\{X_n\}_{n=1}^{\infty}$ is tight if $\forall \epsilon > 0$, $\exists B > 0$ such that $\forall n$, $\inf_n P(|X_n| \leq B) \geq 1 - \epsilon$.

*Remark.* $\{X_n\}$ is *tight*, $\{X_n\}$ is *uniformly tight*, $\{X_n\}$ is *bounded in probability* are exchangably used.

*Remark.* If $\{x_n\}$ is a converging deterministic sequence, we could always find an $N$ such that $n \geq N$ implies $|x_n| \leq B$.

**Definition.** ($\tau_n$-consistent) Let $\{\tau_n\}_{n=1}^{\infty}$ be a (deterministic) sequence such that $\tau_n \nearrow \infty$. Let $\hat{\theta}_n$ be a random vector. $\hat{\theta}_n$ is $\tau_n$-consistent for $\theta(P)$ if $\tau_n(\hat{\theta}_n - \theta(P))$ is tight.

**Exercise 3.1.** Show that $\tau_n$-consistency of $\hat{\theta}_n$ implies the consistency of $\hat{\theta}_n$.

**Example 3.10.** Let $\{X_i\}_{i=1}^{n}$ be iid. Suppose $X_i \sim P$ and $\sigma^2(P) < \infty$. Then, $\bar{X}_n$ is a $\sqrt{n}$-consistent estimator of $\mu(P)$, so $\bar{X}_n$ is a consistent estimator of $\mu(P)$.

**Exercise 3.2.** Show that if $X_n \to_d X$, then $X_n$ is tight.

The Prokhorov's theorem provides a partial converse of the above exercise. The theorem states, if $\{X_n\}_{n=1}^{\infty}$ is tight, then there exists a subsequence $\{X_{n_j}\}_{j=1}^{\infty}$ such that $X_{n_j} \to_d X$.

**Theorem 3.5.** *(Prohorov) Let $\{X_n\}_{n=1}^{\infty}(\in \mathbb{R}^k)$ be a tight sequence of random vectors. Then, there exists a subsequence $\{X_{n_j}\}_{j=1}^{\infty}$ and a random vector $X(\in \mathbb{R}^k)$ such that $X_{n_j} \to_d X$.*

## 3.8 Stochastic Order Notations

Recall the definition $o(1)$ and $O(1)$ for the deterministic sequences.

**Definition.** ($o(1)$) Let $\{x_n\}$ be a sequence of vectors. $x_n = o(1)$ if $x_n \to 0$.

**Definition.** ($O(1)$) Let $\{x_n\}$ be a sequence of vectors. $x_n = O(1)$ if $x_n$ is bounded.

By analogy, for the convergence in probability, $o_p(1)$ and $O_p(1)$ are defined similarly.

**Definition.** ($o_p(1)$) Let $\{X_n\}$ be a sequence of random vectors. $X_n = o_p(1)$ if $X_n \to_p 0$.

**Definition.** ($O_p(1)$) Let $\{X_n\}$ be a sequence of random vectors. $X_n = O_p(1)$ if $X_n$ is tight.

More generally, we could define $o_p(R_n)$ and $O_p(R_n)$ as below.

**Definition.** ($o_p(R_n)$) Let $\{X_n\}$ be a sequence of random vectors. $X_n = o_p(R_n)$ if $X_n = Y_nR_n$ where $Y_n = o_p(1)$.[9]

**Definition.** ($O_p(R_n)$) Let $\{X_n\}$ be a sequence of random vectors. $X_n = O_p(R_n)$ if $X_n = Y_nR_n$ where $Y_n = O_p(1)$.

[9] Not rigorous though, roughly we could say that $\frac{X_n}{R_n} = Y_n$ is $o_p(1)$.

**Proposition 3.2.** *The followings hold:*

*(i)* $o_p(1) + o_p(1) = o_p(1)$
*(ii)* $o_p(1) + O_p(1) = O_p(1)$
*(iii)* $o_p(1)O_p(1) = o_p(1)$
*(iv)* $\frac{1}{1+o_p(1)} = O_p(1)$
*(v)* $o_p(O_p(1)) = o_p(1)$

*Proof.* We only prove (iii) here.

Let $X_n = o_p(1)$ and $Y_n = O_p(1)$. Suppose $X_nY_n \nrightarrow_p 0$. Then, $\exists\epsilon > 0$, $\exists\delta > 0$ such that $\forall N > 0$, $n > N$ and $P(|X_nY_n| > \epsilon) \geq \delta$. Because $Y_{n_j}$ is tight, by Prokhorov's theorem, there exists an index of subsequence $n_j$ of $Y_n$ and a random vector $Y$ such that $Y_{n_j} \rightarrow_d Y$ and $X_{n_j} \rightarrow_p 0$. Using the Slutsky's lemma, $X_{n_j}Y_{n_j} \rightarrow_d 0 \cdot Y = 0$. Therefore, we have $X_{n_j}Y_{n_j} \rightarrow_p 0$, which is contradictory. □

# 4 Hypothesis Testing

## 4.1 Basic Concepts

Consider we want to test $H_0 : \mu(P) = 0$ versus $H_1 : \mu(P) > 0$. $H_0$ is called the null hypothesis, while $H_1$ is called the alternative hypothesis. When testing, two types of mistakes could be involved.

Type I error: Reject $H_0$ when $H_0$ is true.

$$\begin{aligned} P(\text{Type I error}) &= \text{size of a test} \\ &\leq \text{significance level of a test} \\ &= \alpha \end{aligned}$$

Here, significance level of a test is the highest probability of Type I error which could be tolerated. Because we want to deal with the large samples,

$$\limsup_{n\to\infty} P(\text{Type I error}) \leq \alpha$$

Type II error: Do not reject $H_1$ when $H_1$ is false.

$$P(\text{Type II error}) = \beta$$

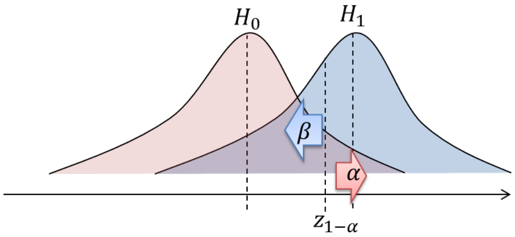
and

$$\begin{aligned} 1 - P(\text{Type II error}) &= P(\text{Reject } H_0 \text{ when } H_0 \text{ is false}) \\ &= \text{Power of a test} \\ &= 1 - \beta \end{aligned}$$

To summarize,

| | $H_0$ True | $H_0$ False |
|---|---|---|
| Don't reject $H_0$ | $1-\alpha$ | $\beta$ |
| Reject $H_0$ | $\alpha$ | $1-\beta$ |

And there are tradeoffs between $\alpha$ and $\beta$. The next figure illustrates this fact.



We want to restrict our attention to test $\phi_n = \phi_n(X_1, ..., X_n)$ of the form $\phi_n = I(T_n > c_n)$, where $T_n$ is the test statistic and $c_n$ is the critical value. The test statistic $T_n$ is a function of samples which provides evidence against $H_0$.

We want to introduce a concept which summarizes the data without any dependence for any particular $\alpha$. This is called the $p$-value.

**Definition.** ($p$-value) $p$-value is the smallest value of $\alpha$ for which one reject $H_0$

**Example 4.1.** We want to test $H_0 : \mu(P) = 0$ versus $H_1 : \mu(P) > 0$. A suitable choice of $T_n$ is:

$$T_n = \sqrt{n}\frac{\bar{X}_n}{S_n}$$

and choice of $c_n = \Phi^{-1}(1-\alpha) = z_{1-\alpha}$, which is the $1-\alpha$th quantile of standard normal distribution.[10] The test then reject $H_0$ when $T_n \to c_n$ is consistent in level. That is, $\limsup_{n\to\infty} P(T_n < c_n) \leq \alpha$ where $H_0$ is true. We have:

$$\begin{aligned}\limsup_{n\to\infty} P(T_n > c_n) &= \limsup_{n\to\infty} P\left(\sqrt{n}\frac{\bar{X}_n}{S_n} > z_{1-\alpha}\right)\\ &= \limsup_{n\to\infty} P\left(\sqrt{n}\frac{\bar{X}_n-\mu(P)}{S_n} + \sqrt{n}\frac{\mu(P)}{S_n} > z_{1-\alpha}\right) \text{ if } H_0 \text{is true, } \sqrt{n}\frac{\mu(P)}{S_n} < 0\\ &\leq \limsup_{n\to\infty} P\left(\sqrt{n}\frac{\bar{X}_n-\mu(P)}{S_n} > z_{1-\alpha}\right) = \alpha\end{aligned}$$

since $\sqrt{n}\frac{\bar{X}_n-\mu(P)}{S_n} \to N(0,1)$.

We reject

$$\hat{p}_n = \inf\left\{\alpha \in (0,1) : \sqrt{n}\frac{\bar{X}_n}{S_n} > \Phi^{-1}(1-\alpha)\right\}$$

which is equivalent to

$$\Phi\left(\frac{\sqrt{n}\bar{X}_n}{S_n}\right) > 1-\alpha$$

so that $\alpha > 1 - \Phi\left(\frac{\sqrt{n}\bar{X}_n}{S_n}\right)$.

**Example 4.2.** (Binomial Testing) Let $X_1, ..., X_n$ be iid with distribution $P$. Let $P = \text{Bernoulli}(q)$, $0 \leq q \leq 1$. Consider, for $\alpha \in (0,1)$, constructing a confidence region $C_n$ such that

$$P(q \in C_n) \to 1-\alpha \tag{4.1}$$

Because we know that $\sqrt{n}\frac{\bar{X}_n-\mu(P)}{S_n} \to N(0,1)$,

$$\sqrt{n}\frac{\bar{X}_n - q}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} \to N(0,1)$$

which implies

$$C_n = \left[\bar{X}_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right]$$

And we know that $C_n$ satisfies (4.1). So we hope that for $n \gg 0$, (4.1) is approximately true. However, in a sense, hope is false. In fact, $\forall n$, $\exists q \in (0,1)$ such that $P(q \in C_n) \simeq 0$.

Let $\epsilon > 0$. Set $q = (1-\alpha)^{\frac{1}{n}}$. With probability $q^n = 1-\alpha$, $X_1 = X_2 = ... = X_n = 1$. Therefore the confidence region $C_n$ for $\alpha$ in this example is $C_n = \{1\}$, which does not even depend on $n$. This implies $q \notin C_n$. So, we have

$$1-\alpha \leq P(q \notin C_n)(=1) \Rightarrow P(q \in C_n)(=0) < \alpha$$

In this example, we have shown that $\forall q \in (0,1)$, $P(q \in C_n) \to 1-\alpha$ does not imply $\inf_q P(q \in C_n) \to 1-\alpha$. As this example illustrates, finite sample property may be quite different from the asymptotic properties.

**Example 4.3.** (Chi-squared Distribution and Testing) Let $X_1, ..., X_n (\in \mathbb{R}^k)$ be iid sequence of random vectors with distribution $P$. Suppose $\Sigma(P) < \infty$. Then, by the central limit theorem,

$$\sqrt{n}(\bar{X}_n - \mu(P)) \to_d z \sim (0, \Sigma(P))$$

In addition, if $\Sigma(P)$ is nonsingular, then $z'[\Sigma(P)]^{-1}z \sim \chi^2_k$. So, by the continuous mapping theorem,

$$n(\bar{X}_n - \mu(P))'[\Sigma(P)]^{-1}(\bar{X}_n - \mu(P)) \to_d \chi^2_k$$

[10] On the notation. Many textbooks use $z_\alpha$ as the probability to denote the point having probability $\alpha$ to the right of it for a standard normal pdf. However, this lecture denotes $z_\alpha$ as the $\alpha$'th quantile of the standard normal, which is opposite from usual.

We want to replace $\Sigma(P)$ with a consistent estimator

$$\hat{\Sigma}_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)(X_i - \bar{X}_n)' \to_p \Sigma(P)$$

Again, by the continuous mapping theorem, $\hat{\Sigma}_n^{-1} \to_p [\Sigma(P)]^{-1}$ because inverse operation of a matrix is a continuous mapping as long as its determinant is nonzero.

Therefore, we conclude that

$$n(\bar{X}_n - \mu(P))'\hat{\Sigma}_n^{-1}(\bar{X}_n - \mu(P)) \to_d \chi_k^2$$

Now, consider testing $H_0 : \mu(P) = 0$ versus $H_1 : \mu(P) \neq 0$ in a way that satisfies $P(T_n > c_n) \to \alpha$ when $H_0$ is true. (i.e. significance level $\alpha$) A choice of $T_n = n\bar{X}_n'\hat{\Sigma}_n\bar{X}_n$.

Part II

# Linear Regression

## 5 Conditional Expectations

In this section, we define the conditional expectations function. It is quite difficult to define rigorously. We take a way to define the conditional expectation via a minimization problem.

**Definition.** (Conditional Expectation) Let $Y \in \mathbb{R}$ and $X \in \mathbb{R}^k$ be random vectors. Suppose $E[Y^2] < \infty$. Let

$$\mathbb{M} = \{m(X) : m : \mathbb{R}^k \to \mathbb{R} \text{ and } E[(m(X))^2] < \infty\}$$

Consider

$$\inf_{m(X)\in\mathbb{M}} E\left[(Y - m(X))^2\right] \tag{5.1}$$

Then, the conditional expectation $E[Y|X]$ is defined as any solution for (5.1).

*Remark.* It is possible to show that a solution for (5.1) exists, i.e. $\exists m^*(X) \in \mathbb{M}$ such that

$$E\left[(Y - m^*(X))^2\right] = \inf_{m(X)\in\mathbb{M}} E\left[(Y - m(X))^2\right]$$

*Remark.* In a sense, $E[Y|X]$ is the *best predictor* of $Y$ given $X$.

**Theorem 5.1.** *(Equivalent Formulation of Conditional Expectation) $m^*(X) \in \mathbb{M}$ solves the minimization problem (5.1) if and only if the following orthogonality conditions hold:*

$$\forall m(X) \in \mathbb{M},\ E\left[(Y - m^*(X))m(X)\right] = 0$$

*Proof.* We have

$$\begin{aligned} E\left[(Y - m(X))^2\right] &= E\left[(Y - m^*(X) + m^*(X) - m(X))^2\right] \\ &= E\left[(Y - m^*(X))^2\right] + 2E\left[(Y - m^*(X))(m^*(X) - m(X))\right] + E\left[(m^*(X) - m(X))^2\right] \quad (5.2) \end{aligned}$$

($\Leftarrow$) Let $m(X) \in \mathbb{M}$. Suppose $m^*(X)$ satisfies $E\left[(Y - m^*(X))m(X)\right] = 0$. Then

$$E\left[(Y - m^*(X))(m^*(X) - m(X))\right] = 0$$

(since $m^*(X) - m(X)$ is orthogonal to $Y - m^*(X)$ as well.) Therefore, we obtain

$$E\left[(Y - m(X))^2\right] \geq E\left[(Y - m^*(X))^2\right]$$

($\Rightarrow$) Suppose $m^*(X)$ solves the minimization problem.
(Idea: compare $E[(Y-m^*(X))^2]$ with $E[(Y-m(X))^2]$, and modify to make the orthogonality condition appears.) Let $m(X) \in \mathbb{M}$. Set $m^*(X) + \alpha m(X)$.

Because $m^*(X)$ achieves the minimization of $E\left[(Y - m^*(X))^2\right]$, $\forall\alpha$ we have

$$\begin{aligned} E\left[(Y - m^*(X))^2\right] - E\left[(Y - m^*(X) - \alpha m(X))^2\right] &\leq 0 \\ E\left[(Y - m^*(X))^2\right] - E\left[(Y - m^*(X))^2\right] + 2\alpha E[(Y - m^*(X))m(X)] - \alpha^2 E\left[(Y - m(X))^2\right] &\leq 0 \\ 2\alpha E[(Y - m^*(X))m(X)] - \alpha^2 E\left[(Y - m(X))^2\right] &\leq 0 \quad (5.3) \end{aligned}$$

Since (5.3) has to hold for any $\alpha$, we have $E[(Y - m^*(X))m(X)] = 0$, which is desired. □

**Corollary.** *If $\tilde{m}(X), m^*(X) \in \mathbb{M}$ both solve the minimization problem (5.1), then $P(m^*(X) = \tilde{m}(X)) = 1$.*

*Proof.* Suppose $\tilde{m}(X), m^*(X)$ both solve the problem. From (5.2) of the proof of above theorem, $E\left[(m^*(X) - \tilde{m}(X))^2\right] = 0$. So, $m^*(X) =_{L^2} \tilde{m}(X)$, therefore $P(m^*(X) = \tilde{m}(X)) = 1$. □

In fact, we can define the conditional expectations without the assumption $E[Y^2] < \infty$. We only need that $E[|Y|] < \infty$. We present an alternative definition here. Many useful properties follow immediately from the definition.

**Definition.** (Conditional Expectation) Let $Y \in \mathbb{R}$ and $X \in \mathbb{R}^k$ be random vectors. Suppose $E[|Y|] < \infty$. Let

$$\mathbb{M} = \{m(X) : m : \mathbb{R}^k \to \mathbb{R} \text{ and } E[|m(X)|] < \infty\}$$

Suppose, for any set $B$, $m^*(X)$ satisfies

$$E\left[(Y - m^*(X))I(X \in B)\right] = 0 \tag{5.4}$$

The conditional expectation $E[Y|X]$ is defined as any solution $m^*(X)$ which satisfies (5.4).

From the previous definition, the properties of conditional expectation follows as below.

**Proposition 5.1.** *(Properties of Conditional Expectation)*
*(i) If $Y = f(X)$, then $E[Y|X] = f(X)$*
*(ii) $E[Y + Z|X] = E[Y|X] + E[Z|X]$*
*(iii) $E[f(X)Y|X] = f(X)E[Y|X]$*
*(iv) If $P(Y \geq 0) = 1$, then $P(E[Y|X] \geq 0) = 1$.*
*(v) (Law of Iterated Expectations) $E[Y] = E[E[Y|X]]$*[11]
*(vi) If $X$ is independent of $Y$, then $E[Y|X] = E[Y]$*

*Proof.* (i) We need to verify the orthogonality condition for this choice of $m^*$.

$$E\left[(Y - f(X))I(X \in B)\right] = E\left[(Y - Y)I(X \in B)\right] = 0$$

Therefore, $E[Y|X] = f(X)$.
(ii) ($\Rightarrow$) Suppose $m^*(X)$ satisfies

$$E\left[(Y + Z - m^*(X))I(X \in B)\right] = 0$$

For each $X$, there exists $\tilde{m}(X)$ and $\hat{m}(X)$ such that $\tilde{m} + \hat{m} = m^*$ and that

I am not
about
part.

$$E\left[(Y - \hat{m}(X))I(X \in B)\right] + E\left[(Z - \tilde{m}(X))I(X \in B)\right] = 0$$

By defining $\hat{m}(X) = E[Y|X]$ and $\tilde{m}(X) = E[Z|X]$, we have the desired result.
($\Leftarrow$) Suppose

$$E\left[(Y - E[Y|X])I(X \in B)\right] = 0 \quad \textit{and} \quad E\left[(Z - E[Z|X])I(X \in B)\right] = 0$$

Summing these two equations up, we have

$$E\left[(Y + Z - E[Y|X] - E[Z|X])I(X \in B)\right] = 0$$

By defining $E[Y|X] + E[Z|X] = E[Y + Z|X]$, we have the desired result.
(iii) The proof is easy when $f(X) = I(X \in \tilde{B})$.

$$E\left[(f(X)Y - f(X)E[Y|X])I(X \in B)\right] = E\left[(Y - E[Y|X])I(X \in B \cap \tilde{B})\right] = 0$$

But I do
know if it
general pro

(iv) Exercise
(v) Let $B \in \mathbb{R}^k$. Let $X \in B$. Then, $I(X \in B) = 1$, so $E\left[(Y - E[Y|X])\right] = 0$. Therefore we have

$$E[Y] = E[E[Y|X]]$$

(vi) Suppose $X$ is independent of $Y$. Let $m^*(X)$ satisfies $E\left[(Y - m^*(X))I(X \in B)\right]$. Then,

$$0 = E\left[(Y - m^*(X))I(X \in B)\right] \quad = \quad E\left[(Y - m^*(X))\right]E[I(X \in B)]$$

Because the right-hand side has to be zero for any $B$, $E[Y] = E[m^*(X)] = E[E[Y|X]]$. Using (v), we know that $E[E[Y|X]] = E[Y]$. □

[11] More generally, if $X = (X_1, X_2)$, $E\left[E[Y|X_1, X_2]|X_1\right] = E[Y|X_1]$

*Remark.* If $E[Y|X]$ is constant, we say that $Y$ is mean independent of $X$. Then, we have

$$E[E[Y|X]] = E[Y|X] = E[Y]$$

The first equality holds because $E[Y|X]$ is constant, the second by law of iterated expectations. Therefore, $Y$ is mean independent of $X$ implies $E[XY] = E[X]E[Y]$. Which implies that $Y$ is uncorrelated with $X$.

NOTE. Independent → Mean independent → Uncorrelated. However, the converse is not generally true.

# 6 Linear Regression

## 6.1 Interpretations of Linear Regression

Let $(Y, X, u)$ be a random vector, where $Y \in \mathbb{R}$, $u \in \mathbb{R}$, $X \in \mathbb{R}^{k+1}$. Let $X = (X_0, X_1, ..., X_k)'$. Assume $X_0 = 1$. Suppose further that $\beta \in \mathbb{R}^{k+1}$ so that $Y = X'\beta + u$. We are now to present three different interpretations of linear regression.

**Interpretation 1. Linear Conditional Expectation**

Assume that $E[Y|X] = X'\beta$. Define $u := Y - E[Y|X] = Y - X'\beta$. So, $Y = X'\beta + u$ and $E[u|X] = 0$ by construction. So, by law of iterated expectations,

$$E[u] = E[E[u|X]] = 0$$

Moreover, $u$ is mean-independent of $X$. This implies that $u$ and $X$ are uncorrelated, so using the law of iterated expectations, we obtain $E[uX] = 0$.

In this interpretation, $\beta$ just summarizes $E[Y|X]$, and there is no causal interpretations.

The dir
sion, is
matrix or
a vector?
then,
come $Y$ c
be scalar w
$\beta \in \mathbb{R}^{k+1}$
Why by
struction?

**Interpretation 2. Best Linear Approximation to $E[Y|X]$, or Best Linear Predictor $Y$ given $X$**

In general, $E[Y|X]$ may not be linear. However, we can try to find the closest linear function to $E[Y|X]$, i.e. choose $b$ to solve

$$\min_{b \in \mathbb{R}^{k+1}} E\left[(Y - X'b)^2\right] \tag{6.1}$$

or equivalently one could solve

$$\min_{b \in \mathbb{R}^{k+1}} E\left[(E[Y|X] - X'b)^2\right] \tag{6.2}$$

Why the problems (6.1) and (6.2) are equivalent?

$$E\left[(E[Y|X] - X'b)^2\right] = E\left[(E[Y|X] - Y + Y - X'b)^2\right]$$

And let us define $V := E[Y|X] - Y$. The above expression becomes:

$$E[V^2 + 2V(Y - X'b) + (Y - X'b)^2] \quad = \quad E[V^2] + 2E[VY] - 2E[VX']b + E[(Y - X'b)^2]$$

Notice that $E[VX'] = E[(E[Y|X] - Y)X'] = E[E[Y|X]X'] - E[YX'] = E[E[YX'|X]] - E[YX'] = 0$. Moreover, $E[V^2]$ and $E[VY]$ are constants. So, the maximization problems are equivalent.

Consider the problem (6.2).

$$D_b E[(Y - X'b)^2] = E[-2X(Y - X'b)]$$

Imposing the first-order conditions and replacing $b$ with $\beta$, we have

$$E[X(Y - X'\beta)] = 0 \tag{6.3}$$

and we define $u := Y - X'\beta$. That is, $u$ is constructed from $X$ and $Y$. From the first-order condition for $X = 1$, $E[u] = 0$. From (6.3), it follows immediately $E[Xu] = 0$.

In this interpretation, we do not have any causal interpretation between $X$ and $Y$ yet.

**Interpretation 3. Causal Model**

Assume $Y = g(X, u)$, where $X$ is the observed determinant of $Y$ and $u$ is unobserved determinant of $Y$. This is a model for how $Y$ is determined. The effect of $X_j$ on $Y$ holding $X_{-j}$ and $u$ constant is captured by $g(X, u)$. Here, $g(X, u)$ is set by theoretical models.

Note that in this setup, we do not know anything about $E[Xu]$ or $E[u|X]$. However, we can normalize $\beta_0$ such that $E[u] = 0$.

If $g$ is differentiable, then

$$\frac{\partial Y}{\partial X_j} = \frac{\partial g(X, u)}{\partial X_j} \tag{6.4}$$

In addition, if we make an assumption $g(X, u) = X'\beta$ is independent of $u$, then (6.4) equals $\beta_j$. Moreover, under this strong assumption, $E[u|X] = E[Xu] = 0$.

## 6.2 Population Regression

In this section, we want to find the expression for $\beta$. We repeat the settings presented in the above section.

Let $(Y, X, u)$ be a random vector, where $Y \in \mathbb{R}$, $u \in \mathbb{R}$, $X \in \mathbb{R}^{k+1}$. Let $X = (X_0, X_1, ..., X_k)'$. Assume $X_0 = 1$. Suppose further that $\beta \in \mathbb{R}^{n+1}$ so that $Y = X'\beta + u$. Additionally, we add further assumptions on $X$ and $u$.

Assumption 1. $E[Xu] = 0$

Assumption 2. $E[XX'] < \infty$

Assumption 3. (No Perfect Collinearity of $X$) $\forall c \in \mathbb{R}^{k+1}$, $P(X'c = 0) = 0$.

**Proposition 6.1.** *$E[XX']$ is full rank if and only if the Assumption 3 holds.*

*Proof.* ($\Rightarrow$) Suppose $\exists c \in \mathbb{R}^{n+1}$ nonzero, such that $P(X'c = 0) = 1$. Then, $E[XX']c = E[X(X'c)'] = 0$. So, $E[XX']$ is not full rank, and therefore not invertible.

($\Leftarrow$) Suppose $E[XX']$ is not invertible. Then, $\exists c \in \mathbb{R}^{n+1}$ nonzero, such that $E[XX']c = 0$. This implies

$$\begin{aligned} 0 &= c'E[XX']c \\ &= E[(X'c)^2] \end{aligned}$$

So, $P(X'c = 0) = 1$. □

Recall (6.3). $E[X(Y - X'\beta)] = 0 \Rightarrow E[XY] = E[XX']\beta$. Because $E[XX']$ is assumed to be invertible, we obtain

$$\beta = E[XX']^{-1}E[XY] \tag{6.5}$$

*Remark.* If $E[XX']$ is not invertible, then there may exist multiple solutions for (6.3).
*Remark.* It is possible to show that for any solution $\beta$ and $\tilde{\beta}$ satisfy (6.3), $P(X'\beta = X'\tilde{\beta}) = 1$.

## 6.3 Partitioned Regression

### Basic Properties

Let $X_1, X_2$ be a partition of $X$ such that $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$. Then,

$$Y = X'\beta = X_1'\beta_1 + X_2'\beta_2$$

and (6.5) changes as:

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} E[X_1X_1'] & E[X_1X_2] \\ E[X_2X_1'] & E[X_2X_2'] \end{pmatrix}^{-1} \begin{pmatrix} E[X_1Y] \\ E[X_2Y] \end{pmatrix}$$

This is the partitioned regression formula. From the formula, we obtain the expressions for $\beta_1$ and $\beta_2$.

**Definition.** (Best Linear Predictor) $BLP(A|B)$ is the best linear interpretation of $A$ given $B$.

*Remark.* $BLP(A|B)$ is defined componentwise if $A$ is a vector.

**Proposition 6.2.** *Let $\tilde{Y} = Y - BLP(Y|X_1)$. Let $\tilde{X}_1 = X_1 - BLP(X_1|X_2)$. Consider the regression $\tilde{Y} = \tilde{X}_1'\tilde{\beta}_1 + \tilde{u}$ when $E[\tilde{X}_1\tilde{u}] = 0$. Then, $\tilde{\beta}_1 = \beta_1$.*

*Proof.* We have:

$$\begin{aligned} \tilde{\beta}_1 &= E[\tilde{X}_1\tilde{X}_1']^{-1}E[\tilde{X}_1\tilde{Y}] \qquad (6.6) \\ &= E[\tilde{X}_1\tilde{X}_1']^{-1}\left\{E[\tilde{X}_1Y'] - E[\tilde{X}_1BLP(Y|X_2)]\right\} \\ &= E[\tilde{X}_1\tilde{X}_1']^{-1}\left\{E[\tilde{X}_1X_1']\beta_1 - E[\tilde{X}_1X_2']\beta_2 + E[\tilde{X}_1u]\right\} \because E[\tilde{X}_1BLP(Y|X_2)] = 0 \\ &= E[\tilde{X}_1\tilde{X}_1']^{-1}\left\{E[\tilde{X}_1\tilde{X}_1']\beta_1 + E[\tilde{X}_1BLP(X_1|X_2)']\beta_1\right\} \because E[\tilde{X}_1X_2']\beta_2 = E[\tilde{X}_1u] = 0,\ \tilde{X}_1 + BLP(X_1|X_2) = X_1 \\ &= \beta_1 \because E[\tilde{X}_1BLP(X_1|X_2)']\beta_1 = 0 \end{aligned}$$

□

*Remark.* If $X_2$ is a constant, $\tilde{Y} = Y - E[Y]$ and $\tilde{X} = X_1 - E[X_1]$. By substituting these back into (6.6), we have $\beta_1 = Var[X_1]^{-1}Cov[X_1, Y]$.

**Omitted Variable Bias**

Consider the true regression $Y = X_1'\beta_1 + X_2'\beta_2 + u$ with the same usual assumptions. Consider $Y = X_1'\beta_1^* + u^*$, where $E[X_1u^*] = 0$.

$$\begin{aligned}
\beta_1^* &= E[X_1X_1']^{-1}E[X_1Y] \\
&= \beta_1 + E[X_1X_1']^{-1}E[X_1X_2']\beta_2 + E[X_1X_1']^{-1}E[X_1u] \\
&= \beta_1 + E[X_1X_1']^{-1}E[X_1X_2']\beta_2 \\
&\neq \beta_1
\end{aligned}$$

The equality may hold only if $E[X_1X_2'] = 0$ or $\beta_2 = 0$, i.e. (1) there is no correlation between $X_1$ and $X_2$ or (2) $\beta_2$ does not explain any change of $Y$.

If $Y = \beta_0 + \beta_1X_1 + \beta_2X_2$ where $X_1, X_2 \in \mathbb{R}$, and if $Y = X_1'\beta_1^* + u^*$, then

$$\beta_1^* = \beta_1 + \frac{Cov(X_1, X_2)}{Var(X_1)}\beta_2$$

## 6.4 The Ordinary Least Squares Estimator and its Properties

**OLS estimatior**

Let $(Y, X, u)$[12] satisfies the following assumptions:

Assumption 1. $E[Xu] = 0$

Assumption 2. $E[XX'] < \infty$

Assumption 3. (No Perfect Collinearity of $X$) $\forall c \in \mathbb{R}^{k+1}$, $P(X'c = 0) = 0$.

Let $(Y_1, X_1)$, $(Y_2, X_2)$, ..., $(Y_n, X_n)$ be an iid sample from distribution $P$. By the analogy of the population regression, we could suggest an estimator:

$$\hat{\beta}_n = \left(\frac{1}{n}\sum_{i=1}^{n} X_iX_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_iY_i\right)$$

[13] and $\hat{\beta}_n$ is called an OLS estimator. The estimator solves

$$\min_{b\in\mathbb{R}^{k+1}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'b)^2$$

So, by the first-order conditions, $\hat{\beta}_n$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n} X_i'(Y_i - X_i'\hat{\beta}_n) = 0$$

The $i$-th residual $\hat{u}_i$ is defined as

$$\hat{u}_i := Y_i - \hat{Y}_i = Y_i - X_i\hat{\beta}_n$$

We immediately know that $\frac{1}{n}\sum_{i=1}^{n} X_i'\hat{u}_i = 0$ by the first-order conditions.

[12] $Y \in \mathbb{R}$, $X \in \mathbb{R}^{k+1}$, $u \in \mathbb{R}$

[13] $\forall i$, $X_iX_i'$ is a $k+1$ by $k+1$ matrix since $X_i$ is a $k+1$ vector. $\left(\frac{1}{n}\sum_{i=1}^{n} X_iY_i\right)$ is a $k+1$ vector. So, $\hat{\beta}_n$ is a $k+1$ vector.

### Projection Interpretation of OLS

Let $\mathbf{y}_n = (Y_1, Y_2, ..., Y_n)'$. Let $\mathbf{X}_{n\times(k+1)} = (X_1, X_2, ..., X_n)'$. Let $\mathbf{u}_n = (u_1, u_2, ..., u_n)$ and $\hat{\mathbf{u}}_n = (\hat{u}_1, \hat{u}_2, ..., \hat{u}_n)$. Let $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}_n$. $\hat{\beta}_n$ solves

$$\min_{b\in\mathbb{R}^{k+1}} |\mathbf{y} - \mathbf{X}b|^2$$

Then we obtain the familiar expression

$$\hat{\beta}_n = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

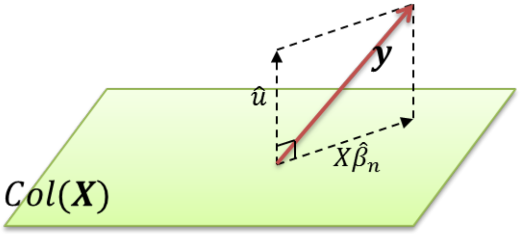
The condition $\frac{1}{n}\sum_{i=1}^{n} X_i'\hat{u}_i = 0$ could be expressed as $\mathbf{X}'\hat{\mathbf{u}} = 0$.

$\mathbf{X}\hat{\beta}_n$ is a vector in column space of $\mathbf{X}$, closest to $\mathbf{y}$. $\mathbf{X}\hat{\beta}_n = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, and we define

$$\mathbf{P} := \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

. Note that $\mathbf{P}$ is symmetric and idempotent, i.e. $\mathbf{P}^2 = \mathbf{P}$. Likewise, we define $\mathbf{M} := \mathbf{I} - \mathbf{P}$, which is a projection matrix which projects a vector orthogonal to the column space of $\mathbf{X}$. $\mathbf{M}$ is called as the residual matrix as well, since

$$\mathbf{M}\mathbf{y} = \mathbf{y} - \mathbf{P}\mathbf{y} = \mathbf{y} - \mathbf{X}\hat{\beta}_n = \hat{\mathbf{u}}$$



### Estimating Subvectors

Again, assume that $Y = X_1'\beta_1 + X_2'\beta_2 + u$. Define $\mathbf{X}_1, \mathbf{X}_2, \mathbf{P}_1, \mathbf{P}_2, \mathbf{M}_1, \mathbf{M}_2$ analogous with above so that we have

$$\mathbf{y} = \mathbf{X}_1\hat{\beta}_{n,1} + \mathbf{X}_2\hat{\beta}_{n,2} + \mathbf{u}$$

Then,

$$\begin{aligned}\mathbf{M}_2\mathbf{y} &= \mathbf{M}_2\mathbf{X}_1\hat{\beta}_{n,1} + \mathbf{M}_2\mathbf{X}_2\hat{\beta}_{n,2} + \mathbf{M}_2\hat{\mathbf{u}} \\ &= \mathbf{M}_2\mathbf{X}_1\hat{\beta}_{n,1} + \hat{\mathbf{u}}\end{aligned}$$

Analogously we obtain

$$(\mathbf{M}_2\mathbf{X}_1)'(\mathbf{M}_2\mathbf{y}) = (\mathbf{M}_2\mathbf{X}_1)'\mathbf{M}_2\mathbf{X}_1\hat{\beta}_{n,1}$$

so that

$$\hat{\beta}_{n,1} = \left[(\mathbf{M}_2\mathbf{X}_1)'(\mathbf{M}_2\mathbf{X}_1)\right]^{-1}(\mathbf{M}_2\mathbf{X}_1)'(\mathbf{M}_2\mathbf{y})$$

Here, $(\mathbf{M}_2\mathbf{X}_1)$ is the residual of regressing $\mathbf{X}_1$ on $\mathbf{X}_2$.

### Bias

ASSUMPTION 1'. $E[u|X] = 0$

ASSUMPTION 2. $E[XX'] < \infty$

ASSUMPTION 3. (No Perfect Collinearity of $X$) $\forall c \in \mathbb{R}^{k+1}$, $P(X'c = 0) = 0$.

ASSUMPTION 4. $(Y, X) \sim P$ and $(Y_1, X_1)$, ... ,$(Y_n, X_n)$ iid $\sim P$.

Under assumptions 1'~4,

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

Take conditional expectations, we have

$$\begin{aligned} E[\hat{\beta}_n|X_1, ..., X_n] &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{u}|X_1, ..., X_n] \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[u_i|X_i] \text{ By independence of } X_i \\ &= \beta \end{aligned}$$

[14] Therefore, $E[\hat{\beta}_n] = E\left[E\left[\hat{\beta}_n|X_1, ..., X_n\right]\right] = \beta$.

### Gauss-Markov Theorem

The Gauss-Markov theorem is one of the most important finite sample properties of OLS estimator.

Assumption 1'. $E[u|X] = 0$

Assumption 2. $E[XX'] < \infty$

Assumption 3. (No Perfect Collinearity of $X$) $\forall c \in \mathbb{R}^{k+1}$, $P(X'c = 0) = 0$.

Assumption 4. $(Y, X) \sim P$ and $(Y_1, X_1)$, ... ,$(Y_n, X_n)$ iid $\sim P$.

Assumption 5. (Homoskedasticity) $Var[u|X] = \sigma^2$.[15]

Under the assumption 1'~5, the following theorem holds.

**Theorem 6.1.** *(Gauss-Markov) Let $Y = X\beta + u$. Let $\mathbf{A}'\mathbf{y}$ be a linear estimator for $\beta$, where $\mathbf{A}$ is a function of $(X_1, ..., X_n)$ such that $E[\mathbf{A}'\mathbf{y}|X_1, ..., X_n] = \beta$. Then, the OLS estimator $\mathbf{A}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ achieves the smallest conditional variance $Var[\mathbf{A}'\mathbf{y}|X_1, ..., X_n]$, i.e. $\forall\Sigma$, $(\Sigma - Var[\mathbf{A}'\mathbf{y}|X_1, ..., X_n])$ is positive semidefinite.*

*Proof.* From $\mathbf{y} = \mathbf{X}'\beta + \mathbf{u}$, premultiply $\mathbf{A}$ to obtain

$$\mathbf{A}'\mathbf{y} = \mathbf{A}'\mathbf{X}\beta + \mathbf{A}'\mathbf{u}$$

Take conditional expectations to obtain

$$\begin{aligned} E[\mathbf{A}'\mathbf{y}|X_1, ..., X_n] &= \mathbf{A}'\mathbf{X}\beta + \mathbf{A}'E[\mathbf{u}|X_1, ..., X_n] \\ &= \mathbf{A}'\mathbf{X}\beta \\ &= \beta \text{ if } \mathbf{A}'\mathbf{X} = \mathbf{I} \end{aligned}$$

To see whether the OLS estimator achieves the minimum variance, we have

$$\begin{aligned} Var[\mathbf{A}'\mathbf{y}|X_1, ..., X_n] &= \mathbf{A}'Var[\mathbf{y}|X_1, ..., X_n]\mathbf{A} \\ &= \mathbf{A}'Var[\mathbf{u}|X_1, ..., X_n]\mathbf{A} \\ &= \sigma^2\mathbf{A}'\mathbf{A} \end{aligned}$$

And notice that when $\mathbf{A}'$ =$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $\mathbf{A}'\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}$. Therefore, it is sufficient to show that $\forall\mathbf{A}$ such that $\mathbf{A}'\mathbf{X} = \mathbf{I}$,[16] $(\mathbf{A}'\mathbf{A}) - (\mathbf{X}'\mathbf{X})^{-1}$ is positive semidefinite. Take $\mathbf{c} = \mathbf{A} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$. Notice that $\mathbf{X}'\mathbf{c} = \mathbf{X}'\mathbf{A} - (\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} = 0$.

Now, we have

$$\begin{aligned} \mathbf{A}'\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1} &= (\mathbf{c} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1})'(\mathbf{c} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) - (\mathbf{X}'\mathbf{X})^{-1} \\ &= \mathbf{c}'\mathbf{c} + \mathbf{c}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{c} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1} \\ &= \mathbf{c}'\mathbf{c} \end{aligned}$$

which is positive semidefinite. □

[14]

$$E[\mathbf{u}|X_1, ..., X_n] = E\left[(u_i - E[u_i|X_i])\right] I(X_i \in B_i, X_{-i} \in B^*)$$

[15] Note. $Var[u|X] = \sigma^2$ does not generally imply $E[u|X] = 0$.

[16] For unbiasedness.

*Remark.* In the sense that the OLS estimator achieves the minimum variance, OLS is called the best linear unbiased estimator (BLUE) under given assumption 1'~5.

### Large Sample Properties of OLS Estimator

**Limiting Distribution of $\hat{\beta}_n$** Let us assume that $E[XX'] < \infty$ and $Var[Xu] < \infty$.

Recall that $\beta = E[XX']^{-1}E[XY]$ and

$$\hat{\beta}_n = \beta + \left(\frac{1}{n}\sum_{i=1}^{n} X_iX_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_iu_i\right)$$

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left(\frac{1}{n}\sum_{i=1}^{n} X_iX_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_iu_i\right)$$

Since $\left(\frac{1}{n}\sum_{i=1}^{n} X_iX_i'\right) \to_p E[XX']$ and $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_iu_i\right) \to_d N(0, Var[Xu])$. So, by continuous mapping theorem,

$$\sqrt{n}(\hat{\beta}_n - \beta) \to_d N(0, E[XX']^{-1}Var[Xu]E[XX']^{-1})$$

### Consistent Estimator of $\Omega$ Under Homoskedasticity

Assumption 1'. $E[u|X] = 0$

Assumption 2. $E[XX'] < \infty$

Assumption 3. (No Perfect Collinearity of $X$) $\forall c \in \mathbb{R}^{k+1}$, $P(X'c = 0) = 0$.

Assumption 4. $(Y, X) \sim P$ and $(Y_1, X_1)$, ... ,$(Y_n, X_n)$ iid $\sim P$.

Assumption 5. (Homoskedasticity) $Var[u|X] = \sigma^2$.

Define $\Omega := E[XX']^{-1}Var[Xu]E[XX']^{-1}$. Assume that $E[u|X] = 0$ and $Var[u|X] = \sigma^2$. Then, $\Omega$ simplifies as:

$$\begin{aligned} Var[Xu] &= E[XX'u^2] - E[Xu]E[Xu] \\ &= E[E[XX'u^2|X]] \\ &= E[XX'E[u^2|X]] \\ &= E[XX']\sigma^2 \end{aligned}$$

Therefore, under this strong assumption, $\Omega = E[XX']^{-1}\sigma^2$. It is not difficult to consistently estimate $E[XX']^{-1}$, since $\left(\frac{1}{n}\sum_{i=1}^{n} X_iX_i'\right)$ is a consistent estimator of $E[XX']^{-1}$. The problem arises here to estimate $\sigma^2$ consistently.

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2$$

We have

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

Take sample average on $\hat{u}_i^2$ to get

$$\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 = \frac{1}{n}\sum_{i=1}^{n} u_i^2 - 2\frac{1}{n}\sum_{i=1}^{n} u_iX_i'(\hat{\beta}_n - \beta) + \frac{1}{n}\sum_{i=1}^{n}(X_i'(\hat{\beta}_n - \beta))^2$$

$\frac{1}{n}\sum_{i=1}^{n} u_i^2 \to_p \sigma^2$ (by WLLN) and $(\hat{\beta}_n - \beta) \to_p 0$ by consistency of $\hat{\beta}_n$. The remaining part is:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(X_i'(\hat{\beta}_n - \beta))^2 &= \frac{1}{n}\sum_{i=1}^{n} |X_i'(\hat{\beta}_n - \beta)|^2 \\ &\leq \frac{1}{n}\sum_{i=1}^{n} |X_i|^2|\hat{\beta}_n - \beta|^2 \text{ Cauchy-Schwarz} \\ &\to_p 0 \end{aligned}$$

Since $(\hat{\beta}_n - \beta) \to_p 0$ and $\frac{1}{n}\sum_{i=1}^n |X_i|^2 = O_p(1)$.[17]

Therefore, under homoskedasticity assumption,

$$\hat{\Omega} = \left(\frac{1}{n}\sum_i X_i X_i'\right)^{-1} \hat{\sigma}_n^2 \to_p \Omega$$

with $\hat{\sigma}_n^2 = \frac{1}{n}\sum_i \hat{u}_i^2$

**Consistent Estimator of $\Omega$ Under Heteroskedasticity**

Assumption 1. $E[uX] = 0$

Assumption 2. $E[XX'] < \infty$

Assumption 3. (No Perfect Collinearity of $X$) $\forall c \in \mathbb{R}^{k+1}$, $P(X'c = 0) = 0$.

Assumption 4. $(Y, X) \sim P$ and $(Y_1, X_1)$, ... ,$(Y_n, X_n) \sim$ iid $P$.

Under the assumptions 1~4, we know that

$$\sqrt{n}(\hat{\beta}_n - \beta) \to_d N(0, \Omega)$$

Under heteroskedasticity, we define $\Omega := E[XX']^{-1}Var[Xu]E[XX']^{-1}$.

**Proposition 6.3.** *(White Estimator)*[18] *Let $\hat{\Omega}$ be an estimator of $\Omega$ such that:*

$$\hat{\Omega} = \left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n X_i X_i' \hat{u}_i^2\right) \left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1}$$

*Then, $\hat{\Omega} \to_p \Omega$.*

*Proof.* We know that $\left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1} \to_p E[XX']^{-1}$. The main difficulty to prove the fact that $\hat{\Omega} \to_p \Omega$ is to prove $\left(\frac{1}{n}\sum_{i=1}^n X_i X_i' \hat{u}_i^2\right) \to_p Var[uX] = E[uXX'u'] - E[uX]^2 = E[XX'u^2]$ under heteroskedasticity. We have:

$$\frac{1}{n}\sum_{i=1}^n X_i X_i' \hat{u}_i^2 = \frac{1}{n}\sum_{i=1}^n X_i X_i' u_i^2 + \frac{1}{n}\sum_{i=1}^n X_i X_i'(\hat{u}_i^2 - u_i^2)$$

And we know that $\frac{1}{n}\sum_{i=1}^n X_i X_i' u_i^2 \to_p E[XX'u^2]$ by the WLLN. So, it remains to show that

$$\frac{1}{n}\sum_{i=1}^n X_i X_i'(\hat{u}_i^2 - u_i^2) \to_p 0 \tag{6.7}$$

We will do this elementwise. Consider $(s,t)$'th element of (6.7).

$$\begin{aligned} \left|\frac{1}{n}\sum_{i=1}^n X_{i,s}X_{i,t}(\hat{u}_i^2 - u_i^2)\right| &\leq \frac{1}{n}\sum_{i=1}^n |X_{i,s}X_{i,t}|\,|\hat{u}_i^2 - u_i^2| \text{ By triangle inequality} \\ &\leq \left(\frac{1}{n}\sum_{i=1}^n |X_{i,s}X_{i,t}|\right)\left(\max_i |\hat{u}_i^2 - u_i^2|\right) \end{aligned}$$

Because $\frac{1}{n}\sum_{i=1}^n |X_{i,s}X_{i,t}| \to_p E\left[|X_{i,s}X_{i,t}|\right] < \infty$, it is sufficient to show that $\left(\max_i |\hat{u}_i^2 - u_i^2|\right) \to_p 0$. Before we proceed, we state and prove the following lemma. □

**Lemma 6.1.** *Let $Z_1, ..., Z_n$ be iid with $E[|Z_i|^r] < \infty$. Then, $n^{-\frac{1}{r}}\max_i |Z_i| \to_p 0$.*

[17] $\frac{1}{n}\sum_{i=1}^n |X_i|^2 \to_p E[X^2] < \infty$ since $E[XX'] < \infty$

[18] This estimator is sometimes called the heteroskedasticity robust standard errors estimator.

*Proof.* Let $\epsilon > 0$.

$$\begin{aligned}
P\left(n^{-\frac{1}{r}} \max_i |Z_i| > \epsilon\right) &= P\left(\max_i |Z_i| > n^{\frac{1}{r}}\epsilon\right) \\
&= P\left(\max_i |Z_i|^r > n\epsilon^r\right) \\
&= P\left(\bigcup_{i=1}^n (|Z_i|^r > \epsilon^r n)\right) \\
&\leq \sum_{i=1}^n P\left(|Z_i|^r > \epsilon^r n\right) \text{ By Bonferroni's inequality} \\
&= \sum_{i=1}^n P\left(|Z_i|^r I(|Z_i|^r > \epsilon^r n) > \epsilon^r n\right) \\
&\leq \frac{\sum_{i=1}^n E[|Z_i|^r I(|Z_i|^r > \epsilon^r n)]}{\epsilon^r n} \text{ By Markov's inequality} \\
&= \frac{1}{n}\sum_{i=1}^n \frac{1}{\epsilon^r} E\left[|Z_i|^r I(|Z_i|^r > \epsilon^r n)\right]
\end{aligned}$$

Because as $n \to \infty$, $I(|Z_i|^r > \epsilon^r n) = 1$ much less frequently. Applying the Dominated Convergence Theorem using the fact that $E[|Z_i|^r] < \infty$, we conclude that the last term converges to zero. □

We return to the proof of the consistency of White estimator.

*Proof.* (Con't) We will apply the lemma to $\max_i |\hat{u}_i^2 - u_i^2|$. Recall that

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

Therefore,

$$\begin{aligned}
|\hat{u}_i^2 - u_i^2| &\leq |2u_i X_i'(\hat{\beta}_n - \beta)| + |X_i|^2|\hat{\beta}_n - \beta|^2 \\
&\leq 2|u_i||X_i'||\hat{\beta}_n - \beta| + |X_i|^2|\hat{\beta}_n - \beta|^2 \qquad (6.8)
\end{aligned}$$

We want to show that $\left(\max_i |\hat{u}_i^2 - u_i^2|\right) \to_p 0$. By (6.8),

$$\max_i |\hat{u}_i^2 - u_i^2| \leq 2\max_i |u_i X_i'||\hat{\beta}_n - \beta| + \max_i |X_i|^2|\hat{\beta}_n - \beta|^2 \qquad (6.9)$$

The first-term of (6.9) is:

$$\begin{aligned}
\max_i |u_i||X_i'||\hat{\beta}_n - \beta| &= |\hat{\beta}_n - \beta| \max_i |u_i X_i'| \\
&= \sqrt{n}|\hat{\beta}_n - \beta| \frac{1}{\sqrt{n}} \max_i |u_i X_i'| \\
&= O_p(1) \frac{1}{\sqrt{n}} \max_i |u_i X_i'|
\end{aligned}$$

By the above lemma, $\frac{1}{\sqrt{n}} \max_i |u_i X_i'| \to_p 0$ if $E[|u_i X_i'|^2] < \infty$. By the assumption $Var[Xu] < \infty$, we conclude that $E[|u_i X_i'|^2] < \infty$, so the first term converges to zero in probability.

The second term could be proved to converge to zero in probability by the same method. This concludes our proof. □

## 6.5 Measures of Fit

We define the total sum of squares, explained sum of squares, and the residual sum of squares as below.

**Definition.** (Total Sum of Squares) The total sum of squares in a regression is:

$$TSS := \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2$$

**Definition.** (Explained Sum of Squares) The explained sum of squares in a regression is:

$$ESS := \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_n)^2$$

**Definition.** (Residual Sum of Squares) The residual sum of squares in a regression is:

$$RSS := \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$R^2$, a measure of fit, is defined as the ratio between the explained sum of squares and the total sum of squares.

**Definition.** (Coefficient of Determination $R^2$)

$$R^2 = \frac{ESS}{TSS}$$

**Proposition 6.4.** *In the OLS estimation, if the regressor $X_i$ includes the constant term,*

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

*Proof.* We want to show that $TSS = RSS + ESS$.

$$\begin{aligned}
TSS &= \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_n + \hat{Y}_n - \bar{Y}_n)^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_n)^2 + \sum_{i=1}^{n} (\hat{Y}_n - \bar{Y}_n)^2 + 2\sum_{i=1}^{n} \hat{u}_i(\hat{Y}_i - \bar{Y}_n) \\
&= RSS + ESS + 2\sum_{i=1}^{n} \hat{u}_i \hat{Y}_i - 2\bar{Y}_n \sum_{i=1}^{n} \hat{u}_i \because \bar{Y}_n \text{does not depend on } i \\
&= RSS + ESS + 2\sum_{i=1}^{n} \hat{u}_i (X_i' \hat{\beta}_n) - 2\bar{Y}_n \sum_{i=1}^{n} \hat{u}_i \\
&= RSS + ESS - 2\bar{Y}_n \sum_{i=1}^{n} \hat{u}_i \because \sum_{i=1}^{n} \hat{u}_i (X_i' \hat{\beta}_n) = \left[ \sum_{i=1}^{n} (Y_i - X_i' \hat{\beta}_n) X_i' \right]' \hat{\beta}_n = 0 \text{ by FOC} \\
&= RSS + ESS \because \left[ \sum_{i=1}^{n} (Y_i - X_i' \hat{\beta}_n) \times 1 \right]' \hat{\beta}_n = 0 \text{ since the first element of } X_i \text{is 1.}
\end{aligned}$$

□

It immediately follows that $0 \leq R^2 \leq 1$. $R^2 = 1$ if and only if $RSS = 0$ so that $\hat{u}_i = 0 \; \forall i$. This is a perfect fit. $R^2 = 0$ if and only if $ESS = 0$ so that $\hat{Y}_i = \bar{Y}_i \; \forall i$. This implies the flat regression line. So the slope parameters are simply zero. Also note that $R^2$ is nondecreasing always with an additional regressor. That is, as the number of regressor increases $ESS$ increases as well, so that $RSS$ falls. Because of this feature, some economists prefer the adjusted $R^2$, which is defined as below.

*Remark.* High $R^2$ does not necessarily mean that the model is a really good causal model.

**Definition.** (Adjusted $R^2$) $\bar{R}^2$, the adjusted $R^2$ is defined as:

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{RSS}{TSS}$$

*Remark.* The motivation of this definition comes from the unbiased estimator of $RSS$ and $TSS$. An unbiased estimator for $Var(Y)$ is

$$\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2 := \widehat{TSS}$$

An unbiased estimator for $Var(u)$ is

$$\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2 := \widehat{RSS}$$

Therefore, it is natural to think $R^2$ as an estimator of $1 - \frac{Var(u)}{Var(Y)}$. Replacing $Var(u)$ and $Var(Y)$ with unbiased estimators, we have the definition of $\bar{R}^2$. We have $\bar{R}^2 \leq 1$ and $\bar{R}^2$ may be negative. However, $\bar{R}^2$ does not need to increase when a regressor is added.

## 6.6 Hypothesis Testing in Linear Regression

This section will cover the inference using the large-sample properties of least squares estimators. Consider the linear regression $Y = X'\beta + u$ with the following assumptions.

Assumption 1. $E[uX] = 0$

Assumption 2. $E[XX'] < \infty$

Assumption 3. (No Perfect Collinearity of $X$) $\forall c \in \mathbb{R}^{k+1}$, $P(X'c = 0) = 0$.

Assumption 4. $(Y, X) \sim P$ and $(Y_1, X_1)$, ... ,$(Y_n, X_n)$ iid $\sim P$.

Assumption 5. $Var[uX]$ is invertible.

### Testing a Single Linear Restriction

Let $r \in \mathbb{R}^{k+1}$ be a nonzero vector. We want to test:

$$H_0 : r'\beta = c \qquad vs \qquad H_1 : r'\beta \neq c$$

By the central limit theorem, we have

$$r'\left[\sqrt{n}(\hat{\beta}_n - \beta)\right] \to_d N(0, r'\Omega r)$$

Because we have $r'\hat{\Omega}r \to_p r'\Omega r$, using invertiblity,

$$\frac{\sqrt{n}(r'\hat{\beta}_n - r'\beta)}{\sqrt{r'\hat{\Omega}_n r}} \to_d N(0,1)$$

We propose a test statistic

$$|T_n| = \frac{\sqrt{n}(r'\hat{\beta}_n - c)}{\sqrt{r'\hat{\Omega}_n r}}$$

of which the critical value $z_{1-\frac{\alpha}{2}} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$. Under $H_0$,

$$P(|T_n| > z_{1-\frac{\alpha}{2}}) = P(T_n > z_{1-\frac{\alpha}{2}}) + P(T_n < z_{\frac{\alpha}{2}}) \to 1 - \Phi(z_{1-\frac{\alpha}{2}}) + \Phi(z_{\frac{\alpha}{2}}) = 1 - (1 - \frac{\alpha}{2}) + \frac{\alpha}{2} = \alpha$$

**Exercise 6.1.** Modify the test for $H_0 : r'\beta \leq c$ vs $H_1 : r'\beta > c$.

### Testing Multiple Linear Restrictions

Let $\mathbf{R}$ be a $p \times (k+1)$ matrix. We want to test:

$$H_0 : \mathbf{R}\beta = c \qquad vs \qquad H_1 : \mathbf{R}\beta \neq c$$

We assume that the rows of $\mathbf{R}$ is linearly independent, to rule out redundant restrictions.

$$\sqrt{n}(\mathbf{R}\hat{\beta}_n - \mathbf{R}\beta) = \mathbf{R}\sqrt{n}(\hat{\beta}_n - \beta) \to N(0, \mathbf{R}\Omega\mathbf{R}')$$

Let $T_n$ be the test statistic such that

$$T_n := n(\mathbf{R}\hat{\beta}_n - \mathbf{R}\beta)' \left[\mathbf{R}\hat{\Omega}_n\mathbf{R}'\right]^{-1} (\mathbf{R}\hat{\beta}_n - \mathbf{R}\beta)$$

. Because of the consistency of the White estimator, we obtain

$$\mathbf{R}\hat{\Omega}_n\mathbf{R}' \to_p \mathbf{R}\Omega\mathbf{R}'$$

Hence, using the Slutsky's lemma,

$$n(\mathbf{R}\hat{\beta}_n - \mathbf{R}\beta)' \left[\mathbf{R}\hat{\Omega}_n\mathbf{R}'\right]^{-1} (\mathbf{R}\hat{\beta}_n - \mathbf{R}\beta) \to_d \chi^2_p$$

with the critical value is $c_{p,1-\alpha}$; $1-\alpha$'th quantile of $\chi^2_p$.

We can construct $C_n$, a confidence region for $\beta$ such that $P(\beta \in C_n) \to 1-\alpha$, using the duology between hypothesis testing and constructing the confidence regions, i.e. inverting tests. So,

$$C_n = \left\{\beta \in \mathbb{R}^{k+1} : n(\hat{\beta}_n - \beta)'\hat{\Omega}_n^{-1}(\hat{\beta}_n - \beta) \leq c_{k+1,1-\alpha}\right\}$$

### Testing Nonlinear Hypothesis

Let $f : \mathbb{R}^{k+1} \to \mathbb{R}^p$ be continuously differentiable at $\beta$. Define $D_\beta f(\beta)$ be a $p \times (k+1)$ vector of partials evaluated at $\beta$. Assume that rows of $D_\beta f(\beta)$ are all linearly independent. Then, by the Delta method, we obtain

$$\sqrt{n}(f(\hat{\beta}_n) - f(\beta)) \to_d N(0, D_\beta f(\beta)\Omega D_\beta f(\beta)')$$

because $D_\beta f(\hat{\beta}_n)\hat{\Omega}_n D_\beta f(\hat{\beta}_n)' \to_p D_\beta f(\beta)\Omega D_\beta f(\beta)'$.