# Contents

# Chapter 2

# Conditional Expectation and Projection

## 2.2. The Distribution of Wages

Suppose that we are interested in wage rates in the US. Since wage rates vary across workers we cannot describe wage rates by a single number. Instead, we can describe wages using a probability distribution. Formally, we view the wage of an individual worker as a random variable $w$ with the *probability distribution*

$$F(u) = \mathbb{P}\left[w \leq u\right].$$

*When we say that a person's wage is random, we mean that we do not know their wage before it is measured, and we treat observed wage rates as realizations from the distribution $F$.*

When a distribution $F$ is differentiable we define the *probability density function*

$$f(u) = \frac{\mathrm{d}\, F(u)}{\mathrm{d}\, u}.$$

The density contains the same information as the distribution function, but the density is typically easier to visually interpret.

Important measures of central tendency are the median and mean. The *median* $m$ of a continuous function $F$ is the unique solution to

$$F(m) = \frac{1}{2}.$$

The median is a robust measure of central tendency, but it is tricky to use for many calculations as it is not a linear operator.

The *mean* or *expectation* of a random variable $Y$ with discrete support is

$$\mu = \mathbb{E}\left[Y\right] = \sum_{j=1}^{\infty} \tau_j \, \mathbb{P}\left[Y = \tau_j\right].$$

For a continuous random variable with density $f(y)$ the expectation is

$$\mu = \mathbb{E}[Y] = \int_{-\infty}^{\infty} yf(y)dy.$$

The expectation is a convenient measure of central tendency because it is a linear operator and arises naturally in many economic models. A disadvantage of the expectation is that it is not robust especially in the presence of substantial skewness or thick tails, which are both features of the wage distribution.

## 2.4. Logs and Percentages

In this section, we want to motivate and clarify the use of the logarithm in regression analysis by making two observations.

Take two positive numbers $a$ and $b$. The percentage difference between $a$ and $b$ is

$$p = 100\left(\frac{a-b}{b}\right).$$

Rewriting

$$\frac{a}{b} = 1 + \frac{p}{100}.$$

Taking natural logarithms,

$$\log a - \log b = \log\left(1 + \frac{p}{100}\right). \tag{2.2}$$

A useful approximation for small $x$ is

$$\log(1+x) \simeq x. \tag{2.3}$$

Applying (2.3) to (2.2) and multiplying by 100 we find

$$p \simeq 100(\log a - \log b).$$

*This shows that 100 multiplied by the difference in logarithms is approximately the percentage difference.*

Now consider the difference in the expectation of log transformed random variables. Take two random variables $X_1, X_2 > 0$. It will be useful to define their *geometric means*

$$\theta_1 = \exp(\mathbb{E}[\log X_1]), \quad \theta_2 = \exp(\mathbb{E}[\log X_2]).$$

Similarly, we are interested in the percentage difference between $\theta_1$ and $\theta_2$, i.e.,

$$p = 100\left(\frac{\theta_2 - \theta_1}{\theta_1}\right)$$

The difference in the expectation of the log transforms (multiplied by 100) is

$$100\left(\mathbb{E}\left[\log X_2\right] - \mathbb{E}\left[\log X_1\right]\right) = 100\left(\log \theta_2 - \log \theta_1\right) \simeq p,$$

the percentage difference between $\theta_1$ and $\theta_2$. *In words, the difference between the average of the log transformed variables is (approximately) the percentage difference in the geometric means.*

## 2.5. Conditional Expectation Function

Conditional expectations can be written with the generic notation

$$\mathbb{E}\left[Y \mid X_1 = x_1, \ldots, X_k = x_k\right] = m(x_1, \ldots, x_k).$$

We call this the *conditional expectation function (CEF)*. The CES if a function of $(x_1, x_2, \ldots, x_k)$ as it varies with the variables.

For greater compactness we typically write the conditioning variables as a vector in $\mathbb{R}^k$:

$$\overrightarrow{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}. \tag{2.4}$$

Given this notation, the CEF can be compactly written as

$$\mathbb{E}\left[Y \mid \overrightarrow{X} = \overrightarrow{x}\right] = m(\overrightarrow{x}).$$

The CEF $m(\overrightarrow{x}) = \mathbb{E}\left[Y \mid \overrightarrow{X} = \overrightarrow{x}\right]$ is a function of $\overrightarrow{x} \in \mathbb{R}^k$. It says: “When $\overrightarrow{X}$ takes the value $\overrightarrow{x}$ then the average value of $Y$ is $m(\overrightarrow{x})$.” Sometimes it is useful to view the CEF as a function of the random variable $\overrightarrow{X}$. In this case we evaluate the function $m(\overrightarrow{x})$ at $\overrightarrow{X}$, and write $m(\overrightarrow{X})$ or $\mathbb{E}\left[Y \mid \overrightarrow{X}\right]$. This is random as it is a function of the random variable $\overrightarrow{X}$.

## 2.7. Law of Iterated Expectations

**Theorem 2.1. Simple Law of Iterated Expectations**

If $\mathbb{E}\left|Y\right| < \infty$, then for any random vector $\overrightarrow{X}$,

$$\mathbb{E}\left[\mathbb{E}\left[Y \mid \overrightarrow{X}\right]\right] = \mathbb{E}\left(Y\right).$$

This states that the expectation of the conditional expectation is the unconditional expectation. In other words, the average of the conditional averages is the unconditional average. For discrete $X$,

$$\mathbb{E}\left[\mathbb{E}\left[Y \mid X\right]\right] = \sum_{j=1}^{\infty} \mathbb{E}\left[Y \mid X = x_j\right] \mathbb{P}\left[X = x_j\right].$$

For continuous $X$,

$$\mathbb{E}\left[\mathbb{E}\left[Y \mid X\right]\right] = \int_{\mathbb{R}} \mathbb{E}\left[Y \mid X = x\right] f_X(x) \mathrm{d}x.$$

**Theorem 2.2. Law of Iterated Expectations**

If $\mathbb{E}|Y| < \infty$, then for any random vectors $X_1$ and $X_2$,

$$\mathbb{E}\left[\mathbb{E}\left[Y \mid X_1, X_2\right] \mid X_1\right] = \mathbb{E}\left(Y \mid X_1\right).$$

Notice the way the law is applied. The inner expectation conditions on $X_1$ and $X_2$, while the outer expectation conditions only on $X_1$. The iterated expectation yields the simple answer $\mathbb{E}\left[Y \mid X_1\right]$, the expectation conditional on $X_1$ alone. Sometimes we phrase this as: "*The smaller information set wins!*"

**Theorem 2.3. Conditioning Theorem**

If $\mathbb{E}|Y| < \infty$, then

$$\mathbb{E}\left[g(X)Y \mid X\right] = g(X)\,\mathbb{E}\left[Y \mid X\right]. \tag{2.7}$$

If in addition $E|g(X)| < \infty$,

$$\mathbb{E}\left[g(X)Y\right] = \mathbb{E}\left[g(X)\,\mathbb{E}\left[Y \mid X\right]\right]. \tag{2.8}$$

## 2.8. CEF Error

The CEF error is defined as the difference between $Y$ and the CEF evaluated at $X$:

$$e = Y - m(\overrightarrow{X}).$$

By construction, this yields the formula

$$Y = m(\overrightarrow{X}) + e \tag{2.9}$$

In (2.9) it is useful to understand that the error $e$ is derived from the joint distribution of $(Y, \overrightarrow{X})$, and so its properties are derived from this construction.

*A key property of the CEF error is that it has a conditional mean of zero.* To see this, by the linearity of expectations, the definition $m(\overrightarrow{X}) = \mathbb{E}[Y \mid \overrightarrow{X}]$ and the conditioning theorem

$$\begin{aligned}\mathbb{E}[e \mid \overrightarrow{X}] &= \mathbb{E}[Y - m(\overrightarrow{X}) \mid \overrightarrow{X}] \\ &= \mathbb{E}[Y \mid \overrightarrow{X}] - \mathbb{E}[m(\overrightarrow{X}) \mid \overrightarrow{X}] \\ &= m(\overrightarrow{X}) - m(\overrightarrow{X}) = 0.\end{aligned}$$

This fact can be combined with the law of iterated expectations to show that *the unconditional mean is also zero*.

$$\mathbb{E}[e] = \mathbb{E}[\mathbb{E}[e \mid \overrightarrow{X}]] = \mathbb{E}[0] = 0.$$

We state this and some other results formally.

**Theorem 2.4. Properties of the CEF Error**

If $\mathbb{E}[Y] < \infty$, then

(1) $\mathbb{E}[e \mid \overrightarrow{X}] = 0$.

(2) $\mathbb{E}[e] = 0$.

(3) If $\mathbb{E}\left[|Y|^r\right] < \infty$ for $r \geq 1$, then $\mathbb{E}\left[|e|^r\right] < \infty$.

(4) For any function $h(\overrightarrow{x})$ such that $\mathbb{E}[h(\overrightarrow{X})e] < \infty$ then $\mathbb{E}[h(\overrightarrow{X})e] = 0$.

The equations

$$\begin{aligned}Y &= m(\overrightarrow{X}) + e \\ \mathbb{E}[e \mid \overrightarrow{X}] &= 0\end{aligned}$$

together imply that $m(\overrightarrow{X})$ is the CEF of $Y$ given $\overrightarrow{X}$. *It is important to understand that this is not a restriction. These equations hold true by definition.*

The condition $\mathbb{E}[e \mid \overrightarrow{X}] = 0$ is implied by the definition of $e$ as the difference between $Y$ and the CEF $m(\overrightarrow{X})$. The equation $\mathbb{E}[e \mid \overrightarrow{X}] = 0$ is sometimes called a *conditional mean restriction*, since the conditional mean of the error $e$ is restricted to equal zero. The property is also sometimes called *mean independence*, for the conditional mean of $e$ is 0 and thus independent of $X$. *However, it does not imply that the distribution of $e$ is independent of $X$.*

## 2.9. Intercept-Only Model

A special case of the regression model is when there are no regressors $X$. In this case $m(\overrightarrow{X}) = \mathbb{E}[Y] = \mu$, the unconditional mean of $Y$. We can still write an equation for $Y$ in the regression format:

$$\begin{aligned}Y &= \mu + e \\ \mathbb{E}[e] &= 0,\end{aligned}$$

where $\mu = \mathbb{E}[Y]$. This is useful for it unifies the notation

## 2.10. Regression Variance

An important measure of the dispersion about the CEF function is the unconditional variance of the CEF error $e$. We write this as

$$\sigma^2 = \operatorname{var}[e] = \mathbb{E}\left[(e - \mathbb{E}[e])^2\right] = \mathbb{E}\left[e^2\right].$$

Theorem 2.4, item (3) implies the following simple but useful result.

**Theorem 2.5.**

If $\mathbb{E}\left[Y^2\right] < \infty$, then $\sigma^2 < \infty$.

We can call $\sigma^2$ the *regression variance* or the *variance of the regression error*. The magnitude of $\sigma^2$ measures the amount of variation in $Y$ which is not "explained" or accounted for in the conditional expectation $\mathbb{E}[Y \mid \overrightarrow{X}]$.

The regression variance depends on the regressors. Consider two regressions

$$\begin{aligned} Y &= \mathbb{E}[Y \mid X_1] + e_1 \\ Y &= \mathbb{E}[Y \mid X_1, X_2] + e_2. \end{aligned}$$

We write the two errors distinctly as $e_1$ and $e_2$ as they are different – changing the conditioning information changes the conditional expectation and therefore the regression error as well.

In our discussion of iterated expectations we have seen that by increasing the conditioning set the conditional expectation reveals greater detail about the distribution of $Y$. What is the implication for the regression error? It turns out that there is a simple relationship. The simple relationship we now derive shows that the variance of this unexplained portion decreases when we condition on more variables. *This relationship is monotonic in the sense that increasing the amount of information always decreases the variance of the unexplained portion.*

**Theorem 2.6.**

If $\mathbb{E}\left[Y^2\right] < \infty$, then

$$\operatorname{var}[Y] \geq \operatorname{var}[Y - \mathbb{E}[Y \mid X_1]] \geq \operatorname{var}[Y - \mathbb{E}[Y \mid X_1, X_2]].$$

Theorem 2.6 says that the variance of the difference between $Y$ and its conditional expectation (weakly) decreases whenever an additional variable is added to the conditioning information.

## 2.11. Best Predictor

Suppose that given a random vector $\overrightarrow{X}$ we want to predict or forecast $Y$. We can write any predictor as a function $g(\overrightarrow{X})$ of $\overrightarrow{X}$. The (ex-post) prediction error is the realized difference $Y - g(\overrightarrow{X})$. A non-stochastic measure of the magnitude of the prediction error is the expectation of its square

$$\mathbb{E}\left[(Y - g(\overrightarrow{X}))^2\right]. \tag{2.10}$$

We can define the best predictor as the function $g(\overrightarrow{X})$ which minimizes (2.10). What function is the best predictor? It turns out that the answer is the CEF $m(\overrightarrow{X})$. This holds regardless of the joint distribution of $(Y, \overrightarrow{X})$.

**Theorem 2.7. Conditional Expectation as Best Predictor**

If $\mathbb{E}\left[Y^2\right] < \infty$, then for any predictor $g(\overrightarrow{X})$,

$$\mathbb{E}\left[(Y - g(\overrightarrow{X}))^2\right] \geq \mathbb{E}\left[(Y - m(\overrightarrow{X}))^2\right],$$

where $m(\overrightarrow{X}) = \mathbb{E}\left[Y \mid \overrightarrow{X}\right]$.

To see this, note that the mean squared error of a predictor $g(\overrightarrow{X})$ is

$$\begin{aligned}
\mathbb{E}\left[[Y - g(\overrightarrow{X})]^2\right] &= \mathbb{E}\left[(e + m(\overrightarrow{X}) - g(\overrightarrow{X}))^2\right] \\
&= \mathbb{E}\left[e^2\right] + 2\,\mathbb{E}\left[e\left[m(\overrightarrow{X}) - g(\overrightarrow{X})\right]\right] + \mathbb{E}\left[[m(\overrightarrow{X}) - g(\overrightarrow{X})]^2\right] \\
&= \mathbb{E}\left[e^2\right] + \mathbb{E}\left[[m(\overrightarrow{X}) - g(\overrightarrow{X})]^2\right] \\
&\geq \mathbb{E}\left[e^2\right] = \mathbb{E}\left[(Y - m(\overrightarrow{X}))^2\right].
\end{aligned}$$

The first equality makes the substitution $Y = m(\overrightarrow{X}) + e$ and the third equality uses Theorem 2.4, item (4). The RHS after the third equality is minimized by setting $g(\overrightarrow{X}) = m(\overrightarrow{X})$, yielding the inequality in the fourth line.

## 2.12. Conditional Variance

While the conditional mean is a good measure of the location of a conditional distribution it does not provide information about the spread of the distribution. A common measure of the dispersion is the *conditional variance*. We first give the general definition of the conditional variance of a random variable $W$.

**Definition 2.1.**

If $\mathbb{E}\left[W^2\right] < \infty$, the *conditional variance* of $W$ given $X = x$ is

$$\sigma^2(\vec{x}) = \operatorname{var}[W \mid \vec{X} = \vec{x}] = \mathbb{E}\left[(W - \mathbb{E}[W \mid \vec{X} = \vec{x}])^2 \mid \vec{X} = \vec{x}\right].$$

The conditional variance treated as a random variable is $\operatorname{var}[W \mid \vec{X}] = \sigma^2(\vec{X})$.

The conditional variance is distinct from the unconditional variance $\operatorname{var}[W]$. The difference is that the conditional variance is a function of the conditioning variables. Notice that the conditional variance is the conditional second moment, centered around the conditional first moment.

Given this definition we define the *conditional variance of the regression error*.

**Definition 2.2.**

If $\mathbb{E}\left[e^2\right] < \infty$, the *conditional variance of the regression error* $e$ given $\vec{X} = \vec{x}$ is

$$\sigma^2(\vec{x}) = \operatorname{var}[e \mid \vec{X} = \vec{x}] = \mathbb{E}\left[e^2 \mid \vec{X} = \vec{x}\right].$$

The conditional variance of $e$ treated as a random variable is $\operatorname{var}[e \mid \vec{X}] = \sigma^2(\vec{X})$.

Again, the conditional variance $\sigma^2(\vec{x})$ is distinct from the unconditional variance $\sigma^2$. *The conditional variance is a function of the regressors, the unconditional variance is not.*

Notice as well that $\sigma^2(\vec{x}) = \operatorname{var}[Y \mid \vec{X} = \vec{x}]$ so it is equivalently the conditional variance of the dependent variable. Generally, $\sigma^2(\vec{x})$ is a non-trivial function of $\vec{x}$ and can take any form subject to the restriction that it is non-negative. One way to think about $\sigma^2(\vec{x})$ is that it is the conditional mean of $e^2$ given $\vec{X}$. Notice as well that

$$\sigma^2(\vec{x}) \coloneqq \mathbb{E}\left[e^2 \mid \vec{X} = \vec{x}\right] = \mathbb{E}\left[Y^2 \mid \vec{X} = \vec{x}\right],$$

so the *conditional variance of the regression error is equivalently the conditional variance of the dependent variable*.

The variance of $Y$ is in a different unit of measurement than $Y$. To convert the variance to the same unit of measure we define the *conditional standard deviation* as its square root $\sigma(\vec{x}) = \sqrt{\sigma^2(\vec{x})}$.

The unconditional variance is related to the conditional variance by the following identity.

**Theorem 2.8.**

If $\mathbb{E}\left[Y^2\right] < \infty$ then

$$\operatorname{var}[Y] = \mathbb{E}[\operatorname{var}[Y \mid \vec{W}]] + \operatorname{var}[\mathbb{E}[Y \mid \vec{W}]].$$

Theorem 2.8 decomposes the unconditional variance into what are sometimes called the "within group variance" and the "across group variance".

The regression error has a conditional mean of zero, so its unconditional error variance equals the expected conditional variance, or equivalently can be found by the law of iterated expectations

$$\sigma^2 = \mathbb{E}\left[e^2\right] = \mathbb{E}\left[\mathbb{E}\left[e^2 \mid \overrightarrow{X}\right]\right] = \mathbb{E}\left[\sigma^2(\overrightarrow{X})\right].$$

That is, *the unconditional error variance is the average conditional variance*.

Given the conditional variance we can define a *rescaled error*

$$u = \frac{e}{\sigma(\overrightarrow{X})}. \tag{2.11}$$

We can calculate that since $\sigma(\overrightarrow{X})$ is a function of $\overrightarrow{X}$

$$\mathbb{E}\left[u \mid \overrightarrow{X}\right] = \mathbb{E}\left[\frac{e}{\sigma(\overrightarrow{X})} \mid \overrightarrow{X}\right] = \frac{1}{\sigma(\overrightarrow{X})}\mathbb{E}\left[e \mid \overrightarrow{X}\right] = 0$$

and

$$\operatorname{var}\left[u \mid \overrightarrow{X}\right] = \mathbb{E}\left[u^2 \mid \overrightarrow{X}\right] = \frac{1}{\sigma^2(\overrightarrow{X})}\mathbb{E}\left[e^2 \mid \overrightarrow{X}\right] = 1.$$

*Thus $u$ has a conditional mean of zero and a conditional variance of 1.*

Notice that (2.11) can be rewritten as

$$e = \sigma(\overrightarrow{X})u,$$

and substituting this for $e$ in the CEF equation (2.9), we find that

$$Y = m(\overrightarrow{X}) + \sigma(\overrightarrow{X})u.$$

This is an alternative (mean-variance) representation of the CEF equation.

Many econometric studies focus on the conditional mean $m(\overrightarrow{x})$ and either ignore the conditional variance $\sigma^2(\overrightarrow{x})$, treat it as a constant $\sigma^2(\overrightarrow{x}) = \sigma^2$, or treat it as a nuisance parameter (a parameter not of primary interest). This is appropriate when the primary variation in the conditional distribution is in the mean but can be short-sighted in other cases.

## 2.13. Homoskedasticity and Heteroskedasticity

An important special case obtains when the conditional variance $\sigma^2(\overrightarrow{x})$ is a constant and independent of $\overrightarrow{x}$. This is called *homoskedasticity*.

**Definition 2.3.**

The error is *homoskedastic* if $\sigma^2(\overrightarrow{x}) = \sigma^2$ does not depend on $\overrightarrow{x}$.

In the general case where $\sigma^2(\overrightarrow{x})$ depends on $\overrightarrow{x}$, we say that the error $e$ is heteroskedastic.

**Definition 2.4.**

The error is *heteroskedastic* if $\sigma^2(\overrightarrow{x})$ depends on $\overrightarrow{x}$.

*It is helpful to understand that the concepts homoskedasticity and heteroskedasticity concern the conditional variance, not the unconditional variance.* By definition, the unconditional variance $\sigma^2$ is a constant and independent of the regressors $\overrightarrow{X}$. So when we talk about the variance as a function of the regressors we are talking about the conditional variance $\sigma^2(\overrightarrow{x})$.

Older textbooks also tend to describe homoskedasticity as a component of a correct regression specification and describe heteroskedasticity as an exception or deviance. This description has influenced many generations of economists but it is unfortunately backwards. *The correct view is that heteroskedasticity is generic and "standard", while homoskedasticity is unusual and exceptional.* The default in empirical work should be to assume that the errors are heteroskedastic, not the converse.

## 2.14. Regression Derivative

One way to interpret the CEF $m(\overrightarrow{x}) = \mathbb{E}[Y \mid \overrightarrow{X} = \overrightarrow{x}]$ is in terms of how marginal changes in the regressors $\overrightarrow{x}$ imply changes in the conditional mean of the response variable $Y$. We can unify the continuous and discrete cases with the notation

$$\nabla_1 m(\overrightarrow{x}) = \begin{cases} \frac{\partial}{\partial x_1} m\left(x_1, \ldots, x_k\right), & \text{if } X_1 \text{ is continuous} \\ m\left(1, x_2, \ldots, x_k\right) - m\left(0, x_2, \ldots, x_k\right), & \text{if } X_1 \text{ is binary.} \end{cases}$$

Collecting the $k$ effects into one $k \times 1$ vector, we define the *regression derivative* with respect to $\overrightarrow{X}$:

$$\overrightarrow{\nabla m(\overrightarrow{x})} = \begin{bmatrix} \nabla_1 m(\overrightarrow{x}) \\ \nabla_2 m(\overrightarrow{x}) \\ \vdots \\ \nabla_k m(\overrightarrow{x}) \end{bmatrix}.$$

*First, the effect of each variable is calculated holding the other variables constant.* This is the *ceteris paribus* concept commonly used in economics. But in the case of a regression derivative, the conditional mean does not literally hold all else constant. It only holds constant the variables included in the conditional mean. This means that the regression derivative depends on which regressors are included. *Second, the regression derivative is the change in the conditional expectation of Y, not the change in the actual value of Y for an individual.*

## 2.15. Linear CEF

An important special case is when the CEF $m(\overrightarrow{x}) = \mathbb{E}[Y \mid \overrightarrow{X} = \overrightarrow{x}]$ is linear in $\overrightarrow{x}$. In this case we can write the mean equation as

$$m(\overrightarrow{x}) = x_1\beta_1 + x_2\beta_2 + \ldots + x_k\beta_k + \beta_{k+1}$$

Notationally it is convenient to write this as a simple function of the vector $\overrightarrow{x}$. An easy way to do so is to augment the regressor vector $\overrightarrow{X}$ by listing the number “1” as an element. We call this the “constant” and the corresponding coefficient is called the “intercept”. Equivalently, specify that the final element of the vector $\overrightarrow{x}$ is $x_k = 1$. $\overrightarrow{X}$ has been redefined as the $k \times 1$ vector

$$\overrightarrow{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_{k-1} \\ 1 \end{pmatrix} \tag{2.12}$$

With this redefinition, the CEF is

$$m(\overrightarrow{x}) = x_1\beta_1 + x_2\beta_2 + \ldots + x_{k-1}\beta_{k-1} + \beta_k = \overrightarrow{x}'\overrightarrow{\beta} \tag{2.13}$$

where

$$\overrightarrow{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \tag{2.14}$$

is a $k \times 1$ coefficient vector. This is the *linear CEF model*. It is also often called the *linear regression model*, or the regression of $Y$ on $\overrightarrow{X}$.

In the linear CEF model the regression derivative is simply the coefficient vector. That is $\overrightarrow{\nabla m(\overrightarrow{x})} = \overrightarrow{\beta}$. This is one of the appealing features of the linear CEF model. The coefficients have simple and natural interpretations as the marginal effects of changing one variable, holding the others constant.

Therefore, a linear CEF model is defined as follows.

$$\begin{aligned} Y &= \overrightarrow{X}'\overrightarrow{\beta} + e \\ \mathbb{E}[e \mid \overrightarrow{X}] &= 0 \end{aligned}$$

If in addition the error is homoskedastic we call this the homoskedastic linear CEF model.

$$\begin{aligned} Y &= \overrightarrow{X}'\overrightarrow{\beta} + e \\ \mathbb{E}[e \mid \overrightarrow{X}] &= 0 \\ \mathbb{E}\left[e^2 \mid \overrightarrow{X}\right] &= \sigma^2 \end{aligned}$$

## 2.17. Linear CEF with Dummy Variables

When all regressors take a finite set of values, it turns out the CEF can be written as a linear function of regressors.

This simplest example is a *binary* variable which takes only two distinct values. For example, in traditional data sets the variable gender takes only the values man and woman (or male and female). Binary variables are extremely common in econometric applications and are alternatively called *dummy variables* or *indicator variables*.

In general, if there are $p$ dummy variables $X_1, \dots, X_p$, then the CEF takes at most $2^p$ distinct values and can be written as a linear function of the $2^p$ regressors including $X_1, X_2, \dots, X_p$ and all cross-products. A linear regression model which includes all $2^p$ binary interactions is called a *saturated dummay variable regression model*. It is a complete model of the conditional mean.

## 2.18. Best Linear Predictor

While the condition mean $m(\overrightarrow{X}) = \mathbb{E}[Y \mid \overrightarrow{X}]$ is the best predictor of $Y$ among all functions of $\overrightarrow{X}$, its functional form is typically known. In particular, the linear CEF model is empirically unlikely to be accurate unless $\overrightarrow{X}$ is discrete and low-dimensional so all interactions are included. Consequently, in most cases it is more realistic to view the linear specification (2.13) as an approximation. In this section we derive a specific approximation with a simple interpretation.

Theorem 2.7 showed that the conditional mean $m(\overrightarrow{X})$ is the best predictor in the sense that it has the lowest mean squared error among all predictors. *By extension, we can define an approximation to the CEF by the linear function with the lowest mean squared error among all linear predictors.*

For this derivation we require the following regularity condition.

**Assumption 2.1.**

(1) $\mathbb{E}\left[Y^2\right] < \infty$.

(2) $\mathbb{E}\left[\|\overrightarrow{X}\|^2\right] < \infty$.

(3) $\boldsymbol{Q}_{XX} = \mathbb{E}\left[\overrightarrow{X}\overrightarrow{X}'\right]$ is positive definite.

We use $\|x\| = \sqrt{\overrightarrow{x}'\overrightarrow{x}}$ to denote the Euclidean length of the vector $\overrightarrow{x}$.

The first two parts of Assumption 2.1 imply that the variables $Y$ and $\overrightarrow{X}$ have finite means, variances, and covariances. The third part of the assumption is more technical, and its role will become apparent shortly. It is equivalent to imposing that the columns of the matrix $\boldsymbol{Q}_{XX}$ are linearly independent, or that the matrix is invertible.

A linear predictor for $Y$ is a function $\vec{X}'\vec{\beta}$ for some $\vec{\beta} \in \mathbb{R}^k$. The mean squared prediction error is

$$S(\vec{\beta}) = \mathbb{E}\left[(Y - \vec{X}'\vec{\beta})^2\right]. \tag{2.17}$$

The best linear predictor of $Y$ given $X$, written $\mathcal{P}[Y \mid \vec{X}]$, is found by selecting $\vec{\beta}$ which minimizes the $S(\vec{\beta})$.

**Definition 2.5.**

The *best linear predictor* of $Y$ given $X$ is

$$\mathcal{P}[Y \mid \vec{X}] = \vec{X}'\vec{\beta}$$

where $\vec{\beta}$ which minimizes the mean squared prediction error

$$S(\vec{\beta}) = \mathbb{E}\left[(Y - \vec{X}'\vec{\beta})^2\right].$$

The minimizer

$$\vec{\beta} = \operatorname*{argmin}_{\vec{b} \in \mathbb{R}^k} S(\vec{b}) \tag{2.18}$$

is called the *linear projection coefficient*.

We now calculate an explicit expression for its value. The mean squared prediction error (2.17) can be written out as a quadratic function of

$$S(\vec{\beta}) = \mathbb{E}\left[Y^2\right] - 2\vec{\beta}'\,\mathbb{E}[\vec{X}Y] + \vec{\beta}'\,\mathbb{E}[\vec{X}\vec{X}']\,\vec{\beta}. \tag{2.19}$$

The quadratic structure of $S(\vec{\beta})$ means that we can solve explicitly for the minimizer. The first-order condition for minimization is

$$\vec{0} = \frac{\partial}{\partial \vec{\beta}} S(\vec{\beta}) = -2\,\mathbb{E}[\vec{X}Y] + 2\,\mathbb{E}[\vec{X}\vec{X}']\,\vec{\beta}. \tag{2.20}$$

Rewriting (2.20) as

$$\vec{Q}_{XY} = \boldsymbol{Q}_{XX}\vec{\beta} \tag{2.21}$$

where $\vec{Q}_{XY} = \mathbb{E}[\vec{X}Y]$ is $k \times 1$ and $\boldsymbol{Q}_{XX} = \mathbb{E}[\vec{X}\vec{X}']$ is $k \times k$. The solution is

$$\vec{\beta} = (\boldsymbol{Q}_{XX})^{-1}\vec{Q}_{XY} = (\mathbb{E}[\vec{X}\vec{X}'])^{-1}\,\mathbb{E}[\vec{X}Y]. \tag{2.22}$$

The *projection error* is

$$\varepsilon = Y - \vec{X}'\vec{\beta}. \tag{2.23}$$

This equals the error (2.9) from the regression equation when (and only when) the conditional mean is linear in $\vec{X}$, otherwise they are distinct.

Rewriting, we obtain a decomposition of $Y$ into linear predictor and error

$$Y = \overrightarrow{X}'\overrightarrow{\beta} + \varepsilon. \tag{2.24}$$

An important property of the projection error is

$$\mathbb{E}\left[\overrightarrow{X}\varepsilon\right] = \overrightarrow{0}. \tag{2.25}$$

To see this,

$$\begin{aligned}\mathbb{E}\left[\overrightarrow{X}\varepsilon\right] &= \mathbb{E}\left[\overrightarrow{X}\left(Y - \overrightarrow{X}'\overrightarrow{\beta}\right)\right] \\ &= \mathbb{E}\left[\overrightarrow{X}Y\right] - \mathbb{E}\left[\overrightarrow{X}\overrightarrow{X}'\right]\left(\mathbb{E}\left[\overrightarrow{X}\overrightarrow{X}'\right]\right)^{-1}\mathbb{E}\left[\overrightarrow{X}Y\right] \\ &= \overrightarrow{0}.\end{aligned} \tag{2.26}$$

Equation (2.25) is a set of $k$ equations, one for each regressor. In other words, (2.25) is equivalent to

$$\mathbb{E}\left[X_j\varepsilon\right] = 0, \text{ for } j = 1, \ldots, k. \tag{2.27}$$

As in (2.12), the regressor vector $\overrightarrow{X}$ typically contains a constant, e.g., $X_k = 1$. In this case, the $k$th equation is

$$\mathbb{E}[\varepsilon] = 0. \tag{2.28}$$

*Thus the projection error has a mean of zero when the regressor vector contains a constant.*

It is also useful to observe that since $\operatorname{cov}\left(X_j, \varepsilon\right) = \mathbb{E}\left[X_j\varepsilon\right] - \mathbb{E}\left[X_j\right]\mathbb{E}[\varepsilon]$, then (2.27)-(2.28) together imply that the variables $X_j$ and $\varepsilon$ are uncorrelated.

We summarize some of the most important properties.

**Theorem 2.9. Properties of Linear Projection Model**

Under Assumption 2.1,

(1) The moments $\mathbb{E}\left[\overrightarrow{X}\overrightarrow{X}'\right]$ and $\mathbb{E}\left[\overrightarrow{X}Y\right]$ exist with finite elements.

(2) The linear projection coefficient (2.18) exists, is unique, and equals

$$\overrightarrow{\beta} = \left(\mathbb{E}\left[\overrightarrow{X}\overrightarrow{X}'\right]\right)^{-1}\mathbb{E}\left[\overrightarrow{X}Y\right].$$

(3) The best linear predictor of $y$ given $\overrightarrow{x}$ is

$$\mathcal{P}\left[Y \mid \overrightarrow{X}\right] = \overrightarrow{X}'\left(\mathbb{E}\left[\overrightarrow{X}\overrightarrow{X}'\right]\right)^{-1}\mathbb{E}\left[\overrightarrow{X}Y\right].$$

(4) The projection error $\varepsilon = Y - \overrightarrow{X}'\overrightarrow{\beta}$ exists. It satisfies $\mathbb{E}\left[\varepsilon^2\right] < \infty$ and $\mathbb{E}\left[\overrightarrow{X}\varepsilon\right] = 0$.

(5) If $\overrightarrow{X}$ contains a constant, then $\mathbb{E}[\varepsilon] = 0$.

(6) If $\mathbb{E}\left[|Y|^r\right] < \infty$ and $\mathbb{E}\left[\|X\|^r\right] < \infty$ for $r \geq 2$ then $\mathbb{E}\left[|\varepsilon|^r\right] < \infty$.

It is useful to reflect on the generality of Theorem 2.9. The only restriction is Assumption 2.1. Thus for any random variables $(Y, \overrightarrow{X})$ with finite variances we can define a linear equation (2.24) with the properties listed in Theorem 2.9. Stronger assumptions (such as the linear CEF model) are not necessary. In this sense the linear model (2.24) exists quite generally. However, it is important not to misinterpret the generality of this statement. *The linear equation (2.24) is defined as the best linear predictor. It is not necessarily a conditional mean, nor a parameter of a structural or causal economic model.*

To summarize, a linear projection model is

$$\begin{aligned} Y &= \overrightarrow{X}'\overrightarrow{\beta} + e, \\ \mathbb{E}\left[\overrightarrow{X}e\right] &= \overrightarrow{0}, \\ \overrightarrow{\beta} &= \left(\mathbb{E}\left[\overrightarrow{X}\overrightarrow{X}'\right]\right)^{-1}\mathbb{E}\left[\overrightarrow{X}Y\right]. \end{aligned}$$

## 2.20. Linear Predictor Error Variance

As in the CEF model, we define the error variance as $\varsigma^2 = \mathbb{E}\left[\varepsilon^2\right]$. Setting $Q_{YY} = \mathbb{E}\left[Y^2\right]$ and $\overrightarrow{Q}_{YX} = \mathbb{E}\left[Y\overrightarrow{X}'\right]$, we can write $\varsigma^2$ as

$$\begin{aligned} \varsigma^2 &= \mathbb{E}\left[(Y - \overrightarrow{X}'\overrightarrow{\beta})^2\right] \\ &= \mathbb{E}\left[Y^2\right] - 2\,\mathbb{E}\left[Y\overrightarrow{X}'\right]\overrightarrow{\beta} + \overrightarrow{\beta}'\,\mathbb{E}\left[\overrightarrow{X}\overrightarrow{X}'\right]\overrightarrow{\beta} \\ &= Q_{YY} - 2\overrightarrow{Q}_{YX}\mathbf{Q}_{XX}^{-1}\overrightarrow{Q}_{XY} + \overrightarrow{Q}_{YX}\mathbf{Q}_{XX}^{-1}\mathbf{Q}_{XX}\mathbf{Q}_{XX}^{-1}\overrightarrow{Q}_{XY} \\ &= Q_{YY} - \overrightarrow{Q}_{YX}\mathbf{Q}_{XX}^{-1}\overrightarrow{Q}_{XY} \\ &=: Q_{YY\cdot X}. \end{aligned} \tag{2.36}$$

One useful feature of this formula is that it shows that $Q_{YY\cdot X} = Q_{YY} - \overrightarrow{Q}_{YX}\mathbf{Q}_{XX}^{-1}\overrightarrow{Q}_{XY}$ equals the variance of the error from the linear projection of $Y$ on $\overrightarrow{X}$.

## 2.21. Regression Coefficients

Sometimes it is useful to separate the constant from the other regressors and write the linear projection equation in the format

$$Y = \overrightarrow{X}'\overrightarrow{\beta} + \alpha + \varepsilon \tag{2.37}$$

where $\alpha$ is the intercept and $\overrightarrow{X}$ does not contain a constant.

Taking expectations of this equation, we find

$$\mathbb{E}\left[Y\right] = \mathbb{E}\left[\overrightarrow{X}'\overrightarrow{\beta}\right] + \mathbb{E}\left[\alpha\right] + \mathbb{E}\left[\varepsilon\right]$$

or $\mu_Y = \overrightarrow{\mu_X}\overrightarrow{\beta} + \alpha$, where $\mu_Y = \mathbb{E}[Y]$ and $\overrightarrow{\mu_X} = \mathbb{E}[\overrightarrow{X}]$, since $\mathbb{E}[\varepsilon] = 0$ from (2.28).Substituting $\alpha$ from (2.37), we find

$$Y - \mu_Y = (\overrightarrow{X} - \overrightarrow{\mu_X})'\,\overrightarrow{\beta} + \varepsilon, \tag{2.38}$$

a linear equation between the centered variables $Y - \mu_Y$ and $\overrightarrow{X} - \overrightarrow{\mu_X}$. They are centered at their means so are mean-zero random variables. Because $\overrightarrow{X} - \overrightarrow{\mu_X}$ is uncorrelated with $\varepsilon$, (2.38) is also a linear projection. Thus by the formula for the linear projection model,

$$\begin{aligned}\overrightarrow{\beta} &= \left((\overrightarrow{X} - \overrightarrow{\mu_X})\,(\overrightarrow{X} - \overrightarrow{\mu_X})'\right)^{-1}\mathbb{E}\left[(\overrightarrow{X} - \overrightarrow{\mu_X})\,(Y - \mu_Y)\right]\\ &= (\operatorname{var}[\overrightarrow{X}])^{-1}\operatorname{cov}(\overrightarrow{X}, Y)\end{aligned}$$

a function only of the covariances of $\overrightarrow{X}$ and $Y$.

**Theorem 2.10.**

In the linear projection model $Y = \overrightarrow{X}'\overrightarrow{\beta} + \alpha + \varepsilon$,

$$\alpha = \mu_Y - \overrightarrow{\mu_X}'\overrightarrow{\beta} \tag{2.39}$$

and

$$\overrightarrow{\beta} = (\operatorname{var}[\overrightarrow{X}])^{-1}\operatorname{cov}(\overrightarrow{X}, Y). \tag{2.40}$$

## 2.22. Regression Sub-Vectors

*The partitioned matrix operations should be summarized in the appendix!*

Let the regressors be partitioned as

$$\overrightarrow{X} = \begin{pmatrix} \overrightarrow{X}_1 \\ \overrightarrow{X}_2 \end{pmatrix}. \tag{2.41}$$

We can write the projection of $Y$ on $\overrightarrow{X}$ as

$$\begin{aligned}Y &= \overrightarrow{X}'\overrightarrow{\beta} + e\\ &= \overrightarrow{X}'_1\overrightarrow{\beta}_1 + \overrightarrow{X}'_2\overrightarrow{\beta}_2 + e\\ \mathbb{E}&[\overrightarrow{X}e] = \overrightarrow{0}.\end{aligned} \tag{2.42}$$

In this section we derive formula for the sub-vectors $\overrightarrow{\beta}_1$ and $\overrightarrow{\beta}_2$.

Partion $\boldsymbol{Q}_{XX}$ conformably with $\overrightarrow{X}$

$$\boldsymbol{Q}_{XX} = \begin{bmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{bmatrix} := \begin{bmatrix} \mathbb{E}\left[\overrightarrow{X}_1\overrightarrow{X}'_1\right] & \mathbb{E}\left[\overrightarrow{X}_1\overrightarrow{X}'_2\right] \\ \mathbb{E}\left[\overrightarrow{X}_2\overrightarrow{X}'_1\right] & \mathbb{E}\left[\overrightarrow{X}_2\overrightarrow{X}'_2\right] \end{bmatrix},$$

and similarly

$$\overrightarrow{Q}_{XY} = \begin{bmatrix} \overrightarrow{Q}_{1Y} \\ \overrightarrow{Q}_{2Y} \end{bmatrix} := \begin{bmatrix} \mathbb{E}\,[\overrightarrow{X}_1 Y] \\ \mathbb{E}\,[\overrightarrow{X}_2 Y] \end{bmatrix}.$$

By the partitioned matrix inversion formula,

$$\boldsymbol{Q}_{XX}^{-1} = \begin{bmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{bmatrix}^{-1} := \begin{bmatrix} \boldsymbol{Q}^{11} & \boldsymbol{Q}^{12} \\ \boldsymbol{Q}^{21} & \boldsymbol{Q}^{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Q}_{11\cdot 2}^{-1} & -\boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1} \\ -\boldsymbol{Q}_{22\cdot 1}^{-1}\mathbf{Q}_{21}\boldsymbol{Q}_{11}^{-1} & \boldsymbol{Q}_{22\cdot 1}^{-1} \end{bmatrix} \tag{2.43}$$

where

$$\boldsymbol{Q}_{11\cdot 2} := \boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}, \quad \boldsymbol{Q}_{22\cdot 1} := \boldsymbol{Q}_{22} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}.$$

Thus,

$$\begin{aligned} \beta &= \begin{pmatrix} \overrightarrow{\beta}_1 \\ \overrightarrow{\beta}_2 \end{pmatrix} \\ &= \begin{bmatrix} \boldsymbol{Q}_{11\cdot 2}^{-1} & -\boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1} \\ -\boldsymbol{Q}_{22\cdot 1}^{-1}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1} & \boldsymbol{Q}_{22\cdot 1}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{Q}_{1Y} \\ \boldsymbol{Q}_{2Y} \end{bmatrix} \\ &= \begin{pmatrix} \boldsymbol{Q}_{11\cdot 2}^{-1}\left(\boldsymbol{Q}_{1Y} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{2Y}\right) \\ \boldsymbol{Q}_{22\cdot 1}^{-1}\left(\boldsymbol{Q}_{2Y} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{1Y}\right) \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{1Y\cdot 2} \\ \boldsymbol{Q}_{22\cdot 1}^{-1}\boldsymbol{Q}_{2Y\cdot 1} \end{pmatrix}. \end{aligned}$$

## 2.23. Coefficient Decomposition

In the previous section we derived formulae for the coefficient sub-vectors $\overrightarrow{\beta}_1$ and $\overrightarrow{\beta}_2$. We now use these formulae to give a useful interpretation of the coefficients in terms of an iterated projection.

Take equation (2.42) for the case $\dim(X_1) = 1$ so that $\beta_1 \in \mathbb{R}$.

$$Y = X_1\beta_1 + \overrightarrow{X}_2'\overrightarrow{\beta}_2 + e. \tag{2.44}$$

Now consider the projection of $X_1$ on $\overrightarrow{X_2}$:

$$\begin{aligned} X_1 &= \overrightarrow{X}_2'\overrightarrow{\gamma}_2 + u_1 \\ \mathbb{E}\,[\overrightarrow{X}_2 u_1] &= 0. \end{aligned}$$

From (2.22) and (2.36), $\overrightarrow{\gamma}_2 = \boldsymbol{Q}_{22}^{-1}\overrightarrow{Q}_{21}$ and $\mathbb{E}\left[u_1^2\right] = Q_{11\cdot 2} = Q_{11} - \overrightarrow{Q}_{12}\boldsymbol{Q}_{22}^{-1}\overrightarrow{Q}_{21}$. We can also calculate that

$$\mathbb{E}\,[u_1 Y] = \mathbb{E}\left[\left(X_1 - \overrightarrow{\gamma}_2'\overrightarrow{X}_2\right)Y\right] = \mathbb{E}\,[X_1 Y] - \overrightarrow{\gamma}_2'\,\mathbb{E}\,[\overrightarrow{X}_2 Y] = Q_{1Y} - \overrightarrow{Q}_{12}\boldsymbol{Q}_{22}^{-1}\overrightarrow{Q}_{2Y} = Q_{1Y\cdot 2}.$$

We have found that

$$\beta_1 = Q_{11\cdot 2}^{-1} Q_{1Y\cdot 2} = \frac{\mathbb{E}\left[u_1 Y\right]}{\mathbb{E}\left[u_1^2\right]},$$

the coefficient from the simple regression of $Y$ on $u_1$.

*What this means is that in the multivariate projection equation (2.44), the coefficient $\beta_1$ equals the projection coefficient from a regression of $Y$ on $u_1$, the error from a projection of $X_1$ on the other regressors $\overrightarrow{X}_2$.* The error $u_1$ can be thought of as the component of $X_1$ which is not linearly explained by the other regressors. Thus the coefficient $\beta_1$ equals the linear effect of $X_1$ on $Y$ after stripping out the effects of the other variables.

There was nothing special in the choice of the variable $X_1$. This derivation applies symmetrically to all coefficients in a linear projection. Each coefficient equals the simple regression of $Y$ on the error from a projection of that regressor on all the other regressors. Each coefficient equals the linear effect of that variable on $Y$ after linearly controlling for all the other regressors.

## 2.24. Omitted Variable Bias

Again, let the regressors be partitioned as in (2.41). Consider the projection of $Y$ on $\overrightarrow{X}_1$ only. Perhaps this is done because the variables $\overrightarrow{X}_2$ are not observed. This is the equation

$$\begin{aligned} Y &= \overrightarrow{X}_1' \overrightarrow{\gamma}_1 + u. \\ \mathbb{E}\left[\overrightarrow{X}_1 u\right] &= \overrightarrow{0}. \end{aligned} \tag{2.45}$$

Notice that we have written the coefficient as $\overrightarrow{\gamma}_1$ rather than $\overrightarrow{\beta}_1$ and the error as $u$ rather than $e$. This is because (2.45) is different than (2.42). Sometimes (2.42) is labelled as *long regression* and (2.45) as *short regression*.

Typically, $\overrightarrow{\beta}_1 \neq \overrightarrow{\gamma}_1$, except in special cases. To see this, we calculate

$$\begin{aligned} \overrightarrow{\gamma}_1 &= \left(\mathbb{E}\left[\overrightarrow{X}_1 \overrightarrow{X}_1'\right]\right)^{-1} \mathbb{E}\left[\overrightarrow{X}_1 Y\right] \\ &= \left(\mathbb{E}\left[\overrightarrow{X}_1 \overrightarrow{X}_1'\right]\right)^{-1} \mathbb{E}\left[\overrightarrow{X}_1 \left(\overrightarrow{X}_1' \overrightarrow{\beta}_1 + \overrightarrow{X}_2' \overrightarrow{\beta}_2 + e\right)\right] \\ &= \overrightarrow{\beta}_1 + \left(\mathbb{E}\left[\overrightarrow{X}_1 \overrightarrow{X}_1'\right]\right)^{-1} \mathbb{E}\left[\overrightarrow{X}_1 \overrightarrow{X}_2'\right] \overrightarrow{\beta}_2 \\ &= \overrightarrow{\beta}_1 + \mathbf{\Gamma}_{12} \overrightarrow{\beta}_2, \end{aligned}$$

where $\mathbf{\Gamma}_{12} \coloneqq \boldsymbol{Q}_{11}^{-1} \boldsymbol{Q}_{12}$ is the coefficient matrix from a projection of $\overrightarrow{X}_2$ on $\overrightarrow{X}_1$.

Observe that $\overrightarrow{\gamma}_1 = \overrightarrow{\beta}_1 + \mathbf{\Gamma}_{12} \overrightarrow{\beta}_2 \neq \overrightarrow{\beta}_1$ unless $\mathbf{\Gamma}_{12} = \mathbf{0}$ or $\overrightarrow{\beta}_2 = \overrightarrow{0}$. Thus the short and long regressions have different coefficients. They are the same only under one of two conditions. First, if the projection of $\overrightarrow{X}_2$ on $\overrightarrow{X}_1$ yields a set of zero coefficients (they are uncorrelated), or second, if the coefficient on $\overrightarrow{X}_2$ in (2.42) is zero. The difference $\mathbf{\Gamma}_{12} \overrightarrow{\beta}_2$ between $\overrightarrow{\gamma}_1$ and $\overrightarrow{\beta}_1$ is known as the *omitted variable bias*. It is the consequence of omission of a relevant correlated variable.

## 2.25. Best Linear Approximation

There are alternative ways we could construct a linear approximation $\overrightarrow{X}'\overrightarrow{\beta}$ to the conditional mean $m(\overrightarrow{X})$. In this section we show that one alternative approach turns out to yield the same answer as the best linear predictor.

We start by defining the mean-square approximation error of $\overrightarrow{X}'\overrightarrow{\beta}$ to $m(\overrightarrow{X})$ as the expected squared difference between $\overrightarrow{X}'\overrightarrow{\beta}$ and the conditional mean $m(\overrightarrow{X})$

$$d(\overrightarrow{\beta}) = \mathbb{E}\left[(m(\overrightarrow{X}) - \overrightarrow{X}'\overrightarrow{\beta})^2\right].$$

The function $d(\overrightarrow{\beta})$ is a measure of the deviation of $\overrightarrow{X}'\overrightarrow{\beta}$ from $m(\overrightarrow{X})$. We can also view the mean-square difference $d(\overrightarrow{\beta})$ as a density-weighted average function $(m(\overrightarrow{X}) - \overrightarrow{X}'\overrightarrow{\beta})^2$ since

$$d(\overrightarrow{\beta}) = \int_{\mathbb{R}^k} (m(\overrightarrow{X}) - \overrightarrow{X}'\overrightarrow{\beta})^2 f_X(\overrightarrow{x})\mathrm{d}\overrightarrow{x},$$

where $f_X(\overrightarrow{x})$ is the marginal density of $\overrightarrow{X}$.

We can then define the best linear approximation to $m(\overrightarrow{X})$ as the function $\overrightarrow{X}'\overrightarrow{\beta}$ obtained by selecting $\overrightarrow{\beta}$ to minimize $d(\overrightarrow{\beta})$:

$$\overrightarrow{\beta} = \underset{\overrightarrow{b}\in\mathbb{R}^k}{\operatorname{argmin}}\, d(\overrightarrow{b}). \tag{2.46}$$

Similar to the best linear predictor we are measuring accuracy by expected squared error. The difference is that the best linear predictor (2.18) selects $\overrightarrow{\beta}$ to minimize the expected squared prediction error, while the best linear approximation (2.46) selects $\overrightarrow{\beta}$ to minimize the expected squared approximation error.

Despite the different definitions, it turns out that the best linear predictor and the best linear approximation are identical. By the same steps as in (2.18) plus an application of conditional expectations we can find that

$$\overrightarrow{\beta} = (\mathbb{E}[\overrightarrow{X}\overrightarrow{X}'])^{-1}\mathbb{E}[\overrightarrow{X}m(\overrightarrow{X})]. \tag{2.47}$$

This is exactly

$$\overrightarrow{\beta} = (\mathbb{E}[\overrightarrow{X}\overrightarrow{X}'])^{-1}\mathbb{E}[\overrightarrow{X}Y], \tag{2.48}$$

Thus (2.46) equals (2.18). We conclude that the definition (2.46) can be viewed as an alternative motivation for the linear projection coefficient.

## 2.26. Regression to the Mean

To understand the origin of the name "regression," consider the simple linear regression

$$Y = X\beta + \alpha + e, \tag{2.49}$$

where $Y$ equals the height of the child and $X$ equals the height of the parent. Assume that $Y$ and $X$ have the same mean so that $\mu_Y = \mu_X = \mu$. When the height distribution is stable across generations so that $\operatorname{var}[X] = \operatorname{var}[Y]$, then the slope is the simple correlation between $X$ and $Y$. Using (2.39), $\alpha = (1-\beta)\,\mu$ so we can write the linear projection (2.49) as

$$\mathcal{P}(Y \mid X) = (1-\beta)\,\mu + X\beta.$$

If we exclude the degenerate case when $Y = X$, $\beta$ is strictly less than 1. This shows that the projected height of the child is a weighted average of the population average height $\mu$ and the parent's height $X$ with the weight equal to $\beta$. Using (2.40),

$$\beta = \frac{\operatorname{cov}(X,Y)}{\operatorname{var}[X]} = \operatorname{corr}(X,Y).$$

This means that on average a child's height is more mediocre (closer to the population average) than the parent's.

A common error – known as the *regression fallacy* – is to infer from $\beta < 1$ that the population is converging meaning that its variance is declining towards zero. This is a fallacy because we derived the implication $\beta < 1$ under the assumption of constant means and variances. So certainly $\beta < 1$ does not imply that the variance $Y$ is less than than the variance of $X$.

Another way of seeing this is to examine the conditions for convergence in the context of equation (2.49). Since $X$ and $e$ are uncorrelated, it follows that

$$\operatorname{var}[Y] = \beta^2 \operatorname{var}[X] + \operatorname{var}[e].$$

Then $\operatorname{var}[X] < \operatorname{var}[Y]$ if and only if

$$\beta^2 > 1 - \frac{\operatorname{var}[e]}{\operatorname{var}[X]},$$

which is not implied by the simple condition $|\beta| < 1$.

## 2.29. Random Coefficient Model

A model which is notationally similar to but conceptually distinct from the linear CEF model is the linear random coefficient model. It takes the form $Y = \overrightarrow{X}'\overrightarrow{\eta}$ where the individual-specific coefficient $\overrightarrow{\eta}$ is random and independent of $\overrightarrow{X}$.

For example, if $X$ is years of schooling and $Y$ is log wages, then $\eta$ is the individual-specific returns to schooling. If a person obtains an extra year of schooling, $\eta$ is the actual change in their wage. The random coefficient model allows the returns to schooling to vary in the population. Some individuals might have a high return to education (a high $\eta$) and others a low return, possibly 0, or even negative.

In the linear CEF model the regressor coefficient equals the regression derivative – the change in the conditional mean due to a change in the regressors,

$$\overrightarrow{\beta} = \overrightarrow{\nabla m\left(\overrightarrow{X}\right)}.$$

*This is not the effect on a given individual, it is the effect on the population average. In contrast, in the random coefficient model the random vector*

$$\overrightarrow{\eta} = \overrightarrow{\nabla\left(\overrightarrow{X}'\overrightarrow{\eta}\right)}$$

*is the true causal effect – the change in the response variable $Y$ itself due to a change in the regressors.*

It is interesting, however, to discover that the linear random coefficient model implies a linear CEF. To see this, let $\overrightarrow{\beta} = \mathbb{E}\left[\overrightarrow{\eta}\right]$ and $\boldsymbol{\Sigma} = \operatorname{var}\left[\overrightarrow{\eta}\right]$ denote the mean and covariance matrix of $\overrightarrow{\eta}$ and then decompose the random coefficient as $\overrightarrow{\eta} = \overrightarrow{\beta} + \overrightarrow{u}$ where $\overrightarrow{u}$ is distributed independently of $\overrightarrow{X}$ with mean zero and covariance matrix $\boldsymbol{\Sigma}$. Then we can write

$$\mathbb{E}\left[Y \mid \overrightarrow{X}\right] = \overrightarrow{X}'\,\mathbb{E}\left[\overrightarrow{\eta} \mid \overrightarrow{X}\right] = \overrightarrow{X}'\overrightarrow{\eta} = \overrightarrow{X}'\overrightarrow{\beta},$$

so the CEF is linear in $\overrightarrow{X}$, and the coefficient $\overrightarrow{\beta}$ equals the mean of the random coefficient $\overrightarrow{\eta}$.

We can thus write the equation as a linear CEF $Y = \overrightarrow{X}'\overrightarrow{\beta} + e$ where $e = \overrightarrow{X}'\overrightarrow{u}$ and $\overrightarrow{u} = \overrightarrow{\eta} - \overrightarrow{\beta}$. The error is conditionally mean zero: $\mathbb{E}\left[e \mid \overrightarrow{X}\right] = 0$. Furthermore,

$$\operatorname{var}\left[e \mid \overrightarrow{X}\right] = \overrightarrow{X}'\boldsymbol{\Sigma}\overrightarrow{X},$$

so the error is conditionally heteroskedastic with its variance a quadratic function of $\overrightarrow{X}$.

**Theorem 2.11.**

In the linear random coefficient model $Y = \overrightarrow{X}'\overrightarrow{\eta}$ with $\overrightarrow{\eta}$ independent of $\overrightarrow{X}$, $\mathbb{E}\left[\|\overrightarrow{X}\|^2\right] < \infty$, and $\mathbb{E}\left[\|\overrightarrow{\eta}\|^2\right] < \infty$, then

$$\mathbb{E}\left[Y \mid \overrightarrow{X}\right] = \overrightarrow{X}'\overrightarrow{\beta}$$
$$\operatorname{var}\left[Y \mid \overrightarrow{X}\right] = \overrightarrow{X}'\boldsymbol{\Sigma}\overrightarrow{X}.$$

where

$$\overrightarrow{\beta} \coloneqq \mathbb{E}\left[\overrightarrow{\eta}\right] \quad \text{and} \quad \boldsymbol{\Sigma} \coloneqq \operatorname{var}\left[\overrightarrow{\eta}\right].$$

## 2.30. Causal Effects

Let $Y$ be a scalar outcome (for example, wages) and $D$ be a binary treatment (for example, college attendance). The specification of treatment as binary is not essential but simplifies the notation. A flexible model describing the impact of the treatment on the outcome is

$$Y = h(D, \overrightarrow{U}) \tag{2.52}$$

where $\overrightarrow{U}$ is an $\ell \times 1$ unobserved random factor and $h$ is a functional relationship. It is also common to use the simplified notation

$$Y(0) = h(0, \overrightarrow{U})$$

and

$$Y(1) = h(1, \overrightarrow{U})$$

for the potential outcomes associated with non-treatment and treatment, respectively. The notation implicitly holds $\overrightarrow{U}$ fixed. The potential outcomes are specific to each individual as they depend on $\overrightarrow{U}$.

Rubin described the effect as causal when we vary $D$ while holding $\overrightarrow{U}$ constant. In our example this means changing an individual's education while holding constant their other attributes.

**Definition 2.6.**

In the model (2.52) the *causal effect* of $D$ on $Y$ is

$$C(\overrightarrow{U}) = Y(1) - Y(0) = h(1, \overrightarrow{U}) - h(0, \overrightarrow{U}), \tag{2.53}$$

the change in $Y$ due to treatment while holing $\overrightarrow{U}$ constant.

Perhaps it would be more appropriate to label (2.53) as a structural effect (the effect within the structural model).

The causal effect of treatment $C(\overrightarrow{U})$ defined in (2.53) is heterogeneous and random as the potential outcomes $Y(0)$ and $Y(1)$ vary across individuals. We do not observe both $Y(0)$ and $Y(1)$ for a given individual but rather only the realized value

$$Y = \begin{cases} Y(0) \text{ if } D = 0 \\ Y(1) \text{ if } D = 1 \end{cases}$$

**Definition 2.7.**

In the model (2.52) the *average causal effect* of $D$ on $Y$ is

$$\text{ACE} = \mathbb{E}\left[C(\overrightarrow{U})\right] = \int_{\mathbb{R}^\ell} C(\overrightarrow{u}) f(\overrightarrow{u}) d\overrightarrow{u},$$

where $f(\overrightarrow{u})$ is the density of $\overrightarrow{U}$.

To estimate the ACE a reasonable starting place is to compare average $Y$ for treated and untreated individuals. This is the same as the coefficient in a regression of the outcome $Y$ on the treatment $D$. Does this equal the ACE?

The answer depends on the relationship between treatment $D$ and the unobserved component $\overrightarrow{U}$. If $D$ is randomly assigned as in an experiment then $D$ and $\overrightarrow{U}$ are independent and the regression coefficient equals the ACE. However, if $D$ and $\overrightarrow{U}$ are dependent then the regression coefficient and ACE are different. To see this, observe that the difference between the average outcomes of the treated and untreated populations are

$$\mathbb{E}\left[Y \mid D=1\right] - \mathbb{E}\left[Y \mid D=0\right] = \int_{\mathbb{R}^\ell} h(1,\overrightarrow{u})f(\overrightarrow{u} \mid D=1)d\overrightarrow{u} - \int_{\mathbb{R}^\ell} h(0,\overrightarrow{u})f(\overrightarrow{u} \mid D=0)d\overrightarrow{u}$$

where $f(\overrightarrow{u} \mid D)$ is the conditional density of $\overrightarrow{U}$ given $D$. If $\overrightarrow{U}$ is independent of $D$ then $f(\overrightarrow{u} \mid D) = f(\overrightarrow{u})$ and the above expression equals to ACE. However, if $D$ and $\overrightarrow{U}$ are dependent this equality fails.

Suppose that the observables include a set of covariates $\overrightarrow{X}$ in addition to the outcome $Y$ and treatment $D$. We extend the potential outcomes model (2.52) to include $\overrightarrow{X}$:

$$Y = h(D,\overrightarrow{X},\overrightarrow{U}). \tag{2.54}$$

We also extend the definition of a causal effect to allow conditioning on $\overrightarrow{X}$.

**Definition 2.8.**

In the model (2.54) the *causal effect* of $D$ on $Y$ is

$$C(\overrightarrow{X},\overrightarrow{U}) = h(1,\overrightarrow{X},\overrightarrow{U}) - h(0,\overrightarrow{X},\overrightarrow{U}),$$

the change in $Y$ due to treatment holding $\overrightarrow{X}$ and $\overrightarrow{U}$ constant.

The *conditional average causal effect* of $D$ on $Y$ conditional on $\overrightarrow{X} = \overrightarrow{x}$ is

$$\operatorname{ACE}(\overrightarrow{x}) = \mathbb{E}\left[C(\overrightarrow{X},\overrightarrow{U}) \mid \overrightarrow{X} = \overrightarrow{x}\right] = \int_{\mathbb{R}^\ell} C(\overrightarrow{x},\overrightarrow{u})f(\overrightarrow{u} \mid \overrightarrow{x})d\overrightarrow{u},$$

where $f(\overrightarrow{u} \mid \overrightarrow{x})$ is the conditional density of $\overrightarrow{U}$ given $\overrightarrow{X}$.

The *unconditional average causal effect* of $D$ on $Y$ is

$$\operatorname{ACE} = \mathbb{E}\left[C(\overrightarrow{X},\overrightarrow{U})\right] = \int \operatorname{ACE}(\overrightarrow{x})f(\overrightarrow{x})d\overrightarrow{x}$$

where $f(\overrightarrow{x})$ is the density of $\overrightarrow{X}$.

The conditional average causal effect $operatorname*ACE(\overrightarrow{x})$ is the ACE for the sub-population with characteristics $\overrightarrow{X} = \overrightarrow{x}$.

**Definition 2.9. Conditional Independence Assumption (CIA)**

Conditional on $\overrightarrow{X}$ the random variables $D$ and $\overrightarrow{U}$ are statistically independent.

The CIA implies that the conditional density of $\overrightarrow{U}$ given $(D,\overrightarrow{X})$ only depends on $\overrightarrow{X}$, thus

$$f(\overrightarrow{u} \mid D,\overrightarrow{X}) = f(\overrightarrow{u} \mid \overrightarrow{X}).$$

This implies that the regression of $Y$ on $(D, \vec{X})$ equals

$$\begin{aligned} m(d, \vec{x}) &= \mathbb{E}[Y \mid D = d, \vec{X} = \vec{x}] \\ &= \mathbb{E}[h(D, \vec{X}, \vec{U}) \mid D = d, \vec{X} = \vec{x}] \\ &= \int h(d, \vec{x}, \vec{u}) f(\vec{u} \mid \vec{x}) d\vec{u} \end{aligned}$$

Under the CIA the treatment effect measured by the regression is

$$\begin{aligned} \nabla m(d, \vec{x}) &= m(1, \vec{x}) - m(0, \vec{x}) \\ &= \int h(1, \vec{x}, \vec{u}) f(\vec{u} \mid \vec{x}) d\vec{u} - \int h(0, \vec{x}, \vec{u}) f(\vec{u} \mid \vec{x}) d\vec{u} \\ &= \int C(\vec{x}, \vec{u}) f(\vec{u} \mid \vec{x}) d\vec{u} \\ &= \mathrm{ACE}(\vec{x}). \end{aligned} \tag{2.55}$$

This is the conditional ACE. Thus under the CIA the regression coefficient equals the ACE.

We deduce that the regression of $Y$ on $(D, \vec{X})$ reveals the causal impact of treatment when the CIA holds. This means that regression analysis can be interpreted causally when we can make the case that the regressors $\vec{X}$ are sufficient to control for factors which are correlated with treatment.

**Theorem 2.12.**

In the structural model (2.54), the Conditional Independence Assumption implies $\nabla m(d, \vec{x}) = \mathrm{ACE}(\vec{x})$, that the regression derivative with respect to treatment equals the conditional ACE.

This is a fascinating result. It shows that whenever the unobservable is independent of the treatment variable after conditioning on appropriate regressors, the regression derivative equals the conditional causal effect. This means the CEF has causal economic meaning, giving strong justification to estimation of the CEF.

# Chapter 3

# The Algebra of Least Squares

## 3.2. Samples

In Section 2.18 we derived and discussed the best linear predictor of $Y$ given $\overrightarrow{X}$ for a pair of random variables $(Y, \overrightarrow{X}) \in \mathbb{R} \times \mathbb{R}^k$ can called this the linear projection model. We are now interested in estimating the parameters of the linear projection model, in particular the projection coefficient

$$\overrightarrow{\beta} = \left(\mathbb{E}\left[\overrightarrow{X}\overrightarrow{X}'\right]\right)^{-1} \mathbb{E}\left[\overrightarrow{X}Y\right]. \tag{3.1}$$

Notationally we wish to distinguish observations (realizations) from the underlying random variables. The random variables are $(Y, \overrightarrow{X})$. THe observations are $(Y_i, \overrightarrow{X}_i)$. From the vantage of the researcher the latter are numbers. From the vantage of statistical theory we view them as realizations of random variables. For individual observations we append a subscript $i$ which runs from 1 to $n$, thus the $i$th observation is $(Y_i, \overrightarrow{X}_i)$.

The individual observations could be draws from a common (homogeneous) distribution or could be draws from heterogeneous distributions. The simplest approach is to assume homogeneity – that the observations are realizations from an identical underlying population $F$.

**Assumption 3.1.**

The random variables $\{(Y_1, \overrightarrow{X_1}), \ldots, (Y_i, \overrightarrow{X_i}), \ldots, (Y_n, \overrightarrow{X_n})\}$ are identically distributed; they are draws from a common distribution $F$.

The assumption does not need to be viewed as literally true. Rather it is a useful modeling device so that parameters such as $\overrightarrow{\beta}$ are well defined. This assumption should be interpreted as how we view an observation a priori, before we actually observe it. In econometric theory we refer to the underlying common distribution $F$ as the *population*. Some authors prefer the label the *data-generating-process (DGP)*. You can think of it as a theoretical concept or an infinitely-large potential population. In contrast, we refer to the observations available to us $\{(Y_i, \overrightarrow{X_i}) : i = 1, \ldots, n\}$ as the *dataset* or *sample*.

The linear projection model, which applies to the random variables $(Y, \overrightarrow{X})$, is

$$Y = \overrightarrow{X}'\overrightarrow{\beta} + e \tag{3.2}$$

where the linear projection coefficient $\overrightarrow{\beta}$ is defined as

$$\overrightarrow{\beta} = \operatorname*{argmin}_{\overrightarrow{b} \in \mathbb{R}^k} S(\overrightarrow{b}), \tag{3.3}$$

the minimizer of the expected squared error

$$S(\overrightarrow{b}) = \mathbb{E}\left[(Y - \overrightarrow{X}'\overrightarrow{\beta})^2\right]. \tag{3.4}$$

The coefficient has the explicit solution (3.1).

## 3.3. Moment Estimators

We want to estimate the coefficient $\overrightarrow{\beta}$ defined in (3.1) from the sample of observations. Notice that $\overrightarrow{\beta}$ is written as a function of certain population expectations. In this context an appropriate estimator is the same function of the sample moments.

To start, suppose that we are interested in the population mean $\mu$ of a random variable $Y$ with distribution function $F$,

$$\mu = \mathbb{E}\left[Y\right] = \int_{-\infty}^{\infty} y dF(y). \tag{3.5}$$

The expectation $\mu$ is a function of the distribution $F$. To estimate $\mu$ given $n$ random variables $Y_i$ from $F$, a natural estimator is the sample mean

$$\widehat{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Now suppose that we are interested in a set of population expectations of possibly nonlinear functions of a random vector $\overrightarrow{Y}$, say $\overrightarrow{\mu} = \mathbb{E}\left[h(\overrightarrow{Y})\right]$. In this case the natural estimator is the vector of sample means

$$\widehat{\overrightarrow{\mu}} = \frac{1}{n}\sum_{i=1}^{n} h(\overrightarrow{Y}).$$

We call $\widehat{\overrightarrow{\mu}}$ the *moment estimator* for $\overrightarrow{\mu}$. For example, if $h(y) = \left(y, y^2\right)'$, then $\widehat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\widehat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$.

Now suppose that we are interested in a nonlinear function of a set of moments. For example, consider the variance of $Y$

$$\sigma^2 = \operatorname{var}[Y] = \mathbb{E}\left[Y^2\right] - (\mathbb{E}[Y])^2 .$$

In general, many parameters of interest can be written as a function of moments of $Y$. Notationally, $\beta = g(\mu)$ and $\mu = \mathbb{E}[h(\vec{Y})]$. Here, $Y$ are the random variables, $h(Y)$ are functions (transformations) of the random variables, and $\mu$ is the expectation of these functions. $\beta$ is the parameter of interest, and is the (nonlinear) function $g(\cdot)$ of these expectations.

In this context a natural estimator of $\beta$ is obtained by replacing $\mu$ with $\widehat{\mu}$. Thus $\widehat{\beta} = g(\widehat{\mu})$. The estimator $\widehat{\beta}$ is often called a *plug-in estimator*. We also call $\widehat{\beta}$ a moment, or moment-based, estimator of $\beta$ since it is a natural extension of the moment estimator $\widehat{\mu}$.

Take the example of the variance $\sigma^2 = \operatorname{var}[Y]$. Its moment estimator is

$$\widehat{\sigma}^2 = \widehat{\mu}_2 - \widehat{\mu}_1^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2 .$$

*This is not the only possible estimator for $\sigma^2$ (there is also the well-known bias-corrected estimator) but $\widehat{\sigma}^2$ is a straightforward and simple choice.*

## 3.4. Least Squares Estimator

The linear projection coefficient $\vec{\beta}$ is defined in (3.3) as the minimizer of the expected squared error $S(\vec{\beta})$ defined in (3.4). For a given $\vec{\beta}$, the expected squared error is the expectation of the squared error $(Y - \vec{X}'\vec{\beta})^2$. The moment estimator of $S(\vec{\beta})$ is the sample average:

$$\widehat{S}(\vec{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \vec{X}_i'\vec{\beta})^2 = \frac{1}{n}\operatorname{SSE}(\vec{\beta}) \tag{3.6}$$

where

$$\operatorname{SSE}(\vec{\beta}) = \sum_{i=1}^{n}(Y_i - \vec{X}_i'\vec{\beta})^2$$

is called the *sum of squared errors* function.

**Definition 3.1.**

The *least squares estimator* is $\widehat{\vec{\beta}} = \operatorname*{argmin}_{\vec{\beta}\in\mathbb{R}^k} \widehat{S}(\vec{\beta})$, where

$$\widehat{S}(\vec{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \vec{X}_i'\vec{\beta})^2 .$$

As $\widehat{S}(\overrightarrow{\beta})$ is a scale multiple of $\mathrm{SSE}\left(\overrightarrow{\beta}\right)$ we may equivalently define $\widehat{\overrightarrow{\beta}}$ as the minimizer of $\mathrm{SSE}\left(\overrightarrow{\beta}\right)$. Hence $\widehat{\overrightarrow{\beta}}$ is commonly called the *least squares (LS) estimator* of $\overrightarrow{\beta}$. The estimator is also commonly referred to as the *ordinary least squares (OLS) estimator*.

## 3.5. Solving for Least Squares with One Regressor

For simplicity, we start by considering the case $k = 1$ so that there is a scalar regressor $X$ and a scalar coefficient $\beta$.

The sum of squared errors is the function

$$\mathrm{SSE}\left(\beta\right) = \sum_{i=1}^{n} \left(Y_i - X_i \beta\right)^2 = \left(\sum_{i=1}^{n} Y_i^2\right) - 2\beta \left(\sum_{i=1}^{n} X_i Y_i\right) + \beta^2 \left(\sum_{i=1}^{n} X_i^2\right).$$

The OLS estimator $\widehat{\beta}$ minimizes this function. The minimizer of $\mathrm{SSE}\left(\beta\right)$ is

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}. \tag{3.7}$$

The intercept-only model is the special case $X_i = 1$. In this case, we find

$$\widehat{\beta} = \frac{1}{n} \sum_{i=1}^{n} Y_i = \overline{Y}, \tag{3.8}$$

the sample mean of $Y_i$.

## 3.6. Solving for Least Squares with Multiple Regressors

We now consider the case with $k > 1$ so that the coefficient $\overrightarrow{\beta} \in \mathbb{R}^k$ is a vector.

The sum of squared errors can be written as

$$\mathrm{SSE}\left(\overrightarrow{\beta}\right) = \sum_{i=1}^{n} \left(Y_i - \overrightarrow{X_i}' \overrightarrow{\beta}\right)^2 = \left(\sum_{i=1}^{n} Y_i^2\right) - 2\overrightarrow{\beta}' \left(\sum_{i=1}^{n} \overrightarrow{X_i} Y_i\right) + \overrightarrow{\beta}' \left(\sum_{i=1}^{n} \overrightarrow{X_i}\overrightarrow{X_i}'\right) \overrightarrow{\beta}.$$

As in the single regressor case this is a quadratic function in $\overrightarrow{\beta}$.

The first-order conditions are

$$0 = \frac{\partial}{\partial \overrightarrow{\beta}} \mathrm{SSE}\left(\overrightarrow{\beta}\right) = -2 \left(\sum_{i=1}^{n} \overrightarrow{X_i} Y_i\right) + 2 \left(\sum_{i=1}^{n} \overrightarrow{X_i}\overrightarrow{X_i}'\right) \overrightarrow{\beta}. \tag{3.9}$$

Dividing (3.9) by 2, we obtain

$$\left(\sum_{i=1}^{n} \overrightarrow{X_i}\overrightarrow{X_i}'\right)\overrightarrow{\beta} = \left(\sum_{i=1}^{n} \overrightarrow{X_i}Y_i\right). \tag{3.10}$$

Then, we can find an explicit formula for the least squares estimator

$$\widehat{\overrightarrow{\beta}} = \left(\sum_{i=1}^{n} \overrightarrow{X_i}\overrightarrow{X_i}'\right)^{-1}\left(\sum_{i=1}^{n} \overrightarrow{X_i}Y_i\right). \tag{3.11}$$

To be complete we should verify the second-order conditions. We calculate that

$$\frac{\partial^2}{\partial\overrightarrow{\beta}\partial\overrightarrow{\beta}'}\mathrm{SSE}\left(\overrightarrow{\beta}\right) = 2\left(\sum_{i=1}^{n} \overrightarrow{X_i}\overrightarrow{X_i}'\right) > 0$$

which is a positive definite matrix.

Alternatively, equation (3.1) writes the projection coefficient $\overrightarrow{\beta}$ as an explicit function of the population moments $\boldsymbol{Q}_{XY}$ and $\boldsymbol{Q}_{XX}$. Their moment estimators are the sample moments

$$\widehat{\boldsymbol{Q}}_{XY} = \frac{1}{n}\sum_{i=1}^{n} \overrightarrow{X_i}Y_i$$

$$\widehat{\boldsymbol{Q}}_{XX} = \frac{1}{n}\sum_{i=1}^{n} \overrightarrow{X_i}\overrightarrow{X_i}'.$$

The moment estimator of $\overrightarrow{\beta}$ replaces the population moments in (3.1) with the sample moments:

$$\widehat{\overrightarrow{\beta}} = \widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\boldsymbol{Q}}_{XY} = \left(\sum_{i=1}^{n} \overrightarrow{X_i}\overrightarrow{X_i}'\right)^{-1}\left(\sum_{i=1}^{n} \overrightarrow{X_i}Y_i\right), \tag{3.12}$$

which is identical with (3.11).

Technically, the estimator $\widehat{\overrightarrow{\beta}}$ is unique and equals (3.11) only if the inverted matrix is actually invertible, which holds if (and only if) this matrix is positive definite. This excludes the case that $X_i$ contains redundant regressors.

**Theorem 3.1.**

If $\sum_{i=1}^{n} \overrightarrow{X_i}\overrightarrow{X_i}' > \mathbf{0}$, the least squares estimator is unique and equals

$$\widehat{\overrightarrow{\beta}} = \left(\sum_{i=1}^{n} \overrightarrow{X_i}\overrightarrow{X_i}'\right)^{-1}\left(\sum_{i=1}^{n} \overrightarrow{X_i}Y_i\right).$$

## 3.8. Least Squares Residuals

As a by-product of estimation we define the *fitted value* $\widehat{Y}_i = \overrightarrow{X_i}'\overrightarrow{\widehat{\beta}}$ and the *residual*

$$\widehat{e}_i = Y_i - \widehat{Y}_i = Y_i - \overrightarrow{X_i}'\overrightarrow{\widehat{\beta}}. \tag{3.14}$$

Note that $Y_i = \widehat{Y}_i + \widehat{e}_i$ and

$$Y_i = \overrightarrow{X_i}'\overrightarrow{\widehat{\beta}} + \widehat{e}_i. \tag{3.15}$$

We make a distinction between the *error* $e_i$ and the *residual* $\widehat{e}_i$. *The error $e_i$ is unobservable while the residual $\widehat{e}_i$ is an estimator.* These two variables are frequently mislabeled which can cause confusion.

Equation (3.9) implies that

$$\sum_{i=1}^{n} \overrightarrow{X_i}\widehat{e}_i = \overrightarrow{0}. \tag{3.16}$$

To see this by a direct calculation, using (3.14) and (3.11),

$$\begin{aligned}
\sum_{i=1}^{n} \overrightarrow{X_i}\widehat{e}_i &= \sum_{i=1}^{n} \overrightarrow{X_i}\left(Y_i - \overrightarrow{X_i}'\overrightarrow{\widehat{\beta}}\right) \\
&= \sum_{i=1}^{n} \overrightarrow{X_i}Y_i - \sum_{i=1}^{n} \overrightarrow{X_i}\overrightarrow{X_i}'\overrightarrow{\widehat{\beta}} \\
&= \sum_{i=1}^{n} \overrightarrow{X_i}Y_i - \sum_{i=1}^{n} \overrightarrow{X_i}\overrightarrow{X_i}'\left[\left(\sum_{i=1}^{n} \overrightarrow{X_i}\overrightarrow{X_i}'\right)^{-1}\left(\sum_{i=1}^{n} \overrightarrow{X_i}Y_i\right)\right] \\
&= \sum_{i=1}^{n} \overrightarrow{X_i}Y_i - \sum_{i=1}^{n} \overrightarrow{X_i}Y_i = \overrightarrow{0}.
\end{aligned}$$

When $\overrightarrow{X_i}$ contains a constant an implication of (3.16) is

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{e}_i = 0. \tag{3.17}$$

*Thus the residuals have a sample mean of* 0 *and the sample correlation between the regressors and the residual is* 0.

## 3.9. Demeaned Regressors

Sometimes it is useful to separate the constant from the other regressors and write the linear projection equation in the format

$$Y_i = \overrightarrow{X_i}' \overrightarrow{\beta} + \alpha + e_i$$

where $\alpha$ is the intercept and $\overrightarrow{X_i}$ does not contain a constant.

In this case (3.16) can be written as the equation system

$$\sum_{i=1}^{n} \left( Y_i - \overrightarrow{X}_i' \widehat{\overrightarrow{\beta}} - \widehat{\alpha} \right) = 0$$
$$\sum_{i=1}^{n} \overrightarrow{X}_i \left( Y_i - \overrightarrow{X}_i' \widehat{\overrightarrow{\beta}} - \widehat{\alpha} \right) = \overrightarrow{0}.$$

The first equation implies

$$\widehat{\alpha} = \overline{Y} - \overline{\overrightarrow{X}}' \widehat{\overrightarrow{\beta}}.$$

Subtracting from the second we obtain

$$\sum_{i=1}^{n} \overrightarrow{X_i} \left[ \overrightarrow{\beta} \left( Y_i - \overline{Y} \right) - \left( \overrightarrow{X}_i - \overline{\overrightarrow{X}} \right)' \right] = \overrightarrow{0}.$$

We can solve for $\widehat{\overrightarrow{\beta}}$ as

$$\widehat{\overrightarrow{\beta}} = \left[ \sum_{i=1}^{n} \left( \overrightarrow{X_i} - \overline{\overrightarrow{X}} \right) \left( \overrightarrow{X_i} - \overline{\overrightarrow{X}} \right)' \right]^{-1} \left[ \sum_{i=1}^{n} \left( \overrightarrow{X_i} - \overline{\overrightarrow{X_i}} \right) \left( Y_i - \overline{Y} \right) \right]. \tag{3.18}$$

Thus the OLS estimator for the slope coefficients is OLS with demeaned data and no intercept.

The representation (3.18) is known as the demeaned formula for the least squares estimator.

## 3.10. Model in Matrix Notation

We can stack these $n$ equations together as

$$\overrightarrow{Y} = \boldsymbol{X} \overrightarrow{\beta} + \overrightarrow{e}. \tag{3.19}$$

Sample sums can be written in matrix notation. For example

$$\sum_{i=1}^{n} \vec{X}_i \vec{X}_i{}' = \boldsymbol{X}'\boldsymbol{X}$$

$$\sum_{i=1}^{n} \vec{X}_i Y_i = \boldsymbol{X}'\vec{Y}.$$

Therefore the least squares estimator can be written as

$$\widehat{\vec{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\vec{Y}.$$

**Theorem 3.2. Important Matrix Expressions**

$$\widehat{\vec{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\vec{Y}$$

$$\widehat{\vec{e}} = \vec{Y} - \boldsymbol{X}\widehat{\vec{\beta}}$$

$$\boldsymbol{X}'\widehat{\vec{e}} = \vec{0}.$$

## 3.11. Projection Matrix

Define the matrix

$$\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'. \tag{3.20}$$

Observe that

$$\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{X} = \boldsymbol{X}.$$

This is a property of a *projection matrix*. More generally, for any matrix $\boldsymbol{Z}$ which can be written as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\Gamma}$ for some matrix $\boldsymbol{\Gamma}$ (we say that $\boldsymbol{Z}$ lies in the range space of $\boldsymbol{X}$), then

$$\boldsymbol{P}\boldsymbol{Z} = \boldsymbol{P}\boldsymbol{X}\boldsymbol{\Gamma} = \boldsymbol{Z}.$$

The projection matrix $\boldsymbol{P}$ has the algebraic property that it is idempotent: $\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$.

*The matrix $\boldsymbol{P}$ creates the fitted values in a least squares regression*:

$$\boldsymbol{P}\vec{Y} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}\vec{Y} = \boldsymbol{X}\widehat{\vec{\beta}} = \widehat{\vec{Y}}.$$

Because of this property $\boldsymbol{P}$ is also known as the *hat matrix*.

A special example of a projection matrix occurs when $\vec{X} = \vec{1}_n$ is an $n$-vector of ones. Then

$$\boldsymbol{P} = \vec{1}_n \left(\vec{1}_n' \vec{1}_n\right)^{-1} \vec{1}_n' = \frac{1}{n}\vec{1}_n \vec{1}_n'.$$

Note that in this case,

$$\boldsymbol{P}\overrightarrow{Y} = \overrightarrow{1}_n\overline{Y}$$

creates an $n$-vector whose elements are the sample mean $\overline{Y}$.

**Theorem 3.3. Properties of the Projection Matrix**

The projection matrix

$$\boldsymbol{P} := \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$$

for any $n \times k$ $\boldsymbol{X}$ with $n \geq k$ has the following algebraic properties.

(1) $\boldsymbol{P}$ is symmetric ($\boldsymbol{P}' = \boldsymbol{P}$).

(2) $\boldsymbol{P}$ is idempotent ($\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$).

(3) $\operatorname{tr} \boldsymbol{P} = k$.

(4) The eigenvalues of $\boldsymbol{P}$ are 1 and 0. There are $k$ eigenvalues equalling 1 and $n-k$ equalling to 0.

(5) $\operatorname{rank}(\boldsymbol{P}) = k$.

*Matrix decomposition results related to idempotent symmetric matrices should be in the appendix.*

## 3.12. Annihilator Matrix

Define

$$\boldsymbol{M} := \boldsymbol{I}_n - \boldsymbol{P} = \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. Note that

$$\boldsymbol{M}\boldsymbol{X} = \boldsymbol{0}. \tag{3.21}$$

Thus $\boldsymbol{M}$ and $\boldsymbol{X}$ are orthogonal. We call $\boldsymbol{M}$ the annihilator matrix due to the property that for any matrix $\boldsymbol{Z}$ in the range space of $\boldsymbol{X}$ (i.e., there exists a matrix $\boldsymbol{\Gamma}$ such that $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\Gamma}$) then

$$\boldsymbol{M}\boldsymbol{Z} = \boldsymbol{0}.$$

The annihilator matrix $\boldsymbol{M}$ has similar properties with $\boldsymbol{P}$, including that $\boldsymbol{M}$ is symmetric ($\boldsymbol{M}' = \boldsymbol{M}$) and idempotent ($\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$). It is thus a projection matrix. Also,

$$\operatorname{tr} \boldsymbol{M} = n - k. \tag{3.22}$$

One implication is that the rank of $\boldsymbol{M}$ is $n - k$.

*While $\boldsymbol{P}$ creates fitted values, $\boldsymbol{M}$ creates least squares residuals*:

$$\boldsymbol{M}\boldsymbol{Y} = \widehat{\overrightarrow{e}}. \tag{3.23}$$

As discussed in the previous section, a special example of a projection matrix occurs when $\overrightarrow{X} = \overrightarrow{1}_n$ is an $n$-vector of ones, so that $\boldsymbol{P} = \overrightarrow{1}_n \left(\overrightarrow{1}'_n \overrightarrow{1}_n\right)^{-1} \overrightarrow{1}'_n$.The associated annihilator matrix is

$$\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{P} = \boldsymbol{I}_n - \overrightarrow{1}_n \left(\overrightarrow{1}'_n \overrightarrow{1}_n\right)^{-1} \overrightarrow{1}'_n.$$

While $\boldsymbol{P}$ creates a vector of sample means, $\boldsymbol{M}$ creates demeaned values:

$$\boldsymbol{M}\overrightarrow{Y} = \overrightarrow{Y} - \overrightarrow{1}_n \overline{Y}.$$

For simplicity we will often write the right-hand side as $\overrightarrow{Y} - \overrightarrow{\overline{Y}}$. The $i$th element is $Y_i - \overline{Y}$, the demeaned value of $Y_i$.

We can also use (3.23) to write an alternative expression for the residual vector.

$$\widehat{\overrightarrow{e}} = \boldsymbol{M}\overrightarrow{Y} = \boldsymbol{M}\left(\boldsymbol{X}\overrightarrow{\beta} + \overrightarrow{e}\right) = \boldsymbol{M}\overrightarrow{e}, \tag{3.24}$$

which is free of dependence on the regression coefficient $\beta$.

## 3.13. Estimation of Error Variance

The error variance $\sigma^2 = \mathbb{E}\left[e^2\right]$ is a moment, so a natural estimator is a moment estimator. If $e_i$ were observed we would estimate $\sigma^2$ by

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2. \tag{3.25}$$

However, this is infeasible as $e_i$ is not observed. In this case it is common to take a two-step approach to estimation. The residuals $\widehat{e}_i$ are calculate in the first step, and then we substitute $\widehat{e}_i$ for $e_i$ in expression (3.25) to obtain the feasible estimator

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{e}_i^2. \tag{3.26}$$

In matrix notation, we can write (3.25) and (3.26) as $\widetilde{\sigma}^2 = n^{-1}\overrightarrow{e}'\overrightarrow{e}$ and

$$\widehat{\sigma}^2 = \frac{1}{n}\widehat{\overrightarrow{e}}'\widehat{\overrightarrow{e}}. \tag{3.27}$$

Recall the expressions $\widehat{\overrightarrow{e}} = \boldsymbol{M}\overrightarrow{Y} = \boldsymbol{M}\overrightarrow{e}$ from (3.23) and (3.24). Applied to (3.27) we find

$$\widehat{\sigma}^2 = \frac{1}{n}\widehat{\overrightarrow{e}}'\widehat{\overrightarrow{e}} = \frac{1}{n}\overrightarrow{e}'\boldsymbol{M}\overrightarrow{e}. \tag{3.28}$$

An interesting implication is that

$$\widetilde{\sigma}^2 - \widehat{\sigma}^2 = \frac{1}{n}\left(\overrightarrow{e}'\overrightarrow{e} - \overrightarrow{e}'\boldsymbol{M}\overrightarrow{e}\right) = \frac{1}{n}\overrightarrow{e}'\boldsymbol{P}\overrightarrow{e} \geq 0.$$

The final inequality holds because $P$ is positive semi-definite and $\overrightarrow{e}'\boldsymbol{P}\overrightarrow{e}$ is a quadratic form. *This shows that the feasible estimator $\widehat{\sigma}^2$ is numerically smaller than the idealized estimator (3.25).*

## 3.14. Analysis of Variance

Another way of writing (3.23) is

$$\overrightarrow{Y} = \boldsymbol{P}\overrightarrow{Y} + \boldsymbol{M}\overrightarrow{Y} = \widehat{\overrightarrow{Y}} + \widehat{\overrightarrow{e}}. \tag{3.29}$$

This decomposition is *orthogonal*, that is

$$\widehat{\overrightarrow{Y}}'\widehat{\overrightarrow{e}} = (\boldsymbol{P}\overrightarrow{Y})'(\boldsymbol{M}\overrightarrow{Y}) = \overrightarrow{Y}'\boldsymbol{PM}\overrightarrow{Y} = 0. \tag{3.30}$$

It follows that

$$\overrightarrow{Y}'\overrightarrow{Y} = \widehat{\overrightarrow{Y}}'\widehat{\overrightarrow{Y}} + 2\widehat{\overrightarrow{Y}}'\widehat{\overrightarrow{e}} + \widehat{\overrightarrow{e}}'\widehat{\overrightarrow{e}} = \widehat{\overrightarrow{Y}}'\widehat{\overrightarrow{Y}} + \widehat{\overrightarrow{e}}'\widehat{\overrightarrow{e}}$$

or

$$\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} \widehat{Y}_i^2 + \sum_{i=1}^{n} \widehat{e}_i^2.$$

Subtracting $\overline{Y}$ from both sides of (3.29) we obtain

$$\overrightarrow{Y} - \overrightarrow{1}_n\overline{Y} = \widehat{\overrightarrow{Y}} - \overrightarrow{1}_n\overline{Y} + \widehat{\overrightarrow{e}}.$$

This decomposition is also orthogonal when $\boldsymbol{X}$ contains a constant, as

$$\left(\widehat{\overrightarrow{Y}} - \overrightarrow{1}_n\overline{Y}\right)'\widehat{\overrightarrow{e}} = \widehat{\overrightarrow{Y}}'\widehat{\overrightarrow{e}} - \overline{Y}\overrightarrow{1}_n'\widehat{\overrightarrow{e}} = 0$$

under (3.17). It follows that

$$\left(\overrightarrow{Y} - \overrightarrow{1}_n\overline{Y}\right)'\left(\overrightarrow{Y} - \overrightarrow{1}_n\overline{Y}\right) = \left(\widehat{\overrightarrow{Y}} - \overrightarrow{1}_n\overline{Y}\right)'\left(\widehat{\overrightarrow{Y}} - \overrightarrow{1}_n\overline{Y}\right) + \widehat{\overrightarrow{e}}'\widehat{\overrightarrow{e}}$$

or

$$\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=\sum_{i=1}^{n}\left(\widehat{Y}_i-\overline{Y}\right)^2+\sum_{i=1}^{n}\widehat{e}_i^2.$$

This is commonly called the *analysis-of-variance* formula for least squares regression.

A commonly reported statistic is the *coefficient of determination* or *R-squared*:

$$R^2=\frac{\sum_{i=1}^{n}\left(\widehat{Y}_i-\overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}=1-\frac{\sum_{i=1}^{n}\widehat{e}_i^2}{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}.$$

*It is often described as "the fraction of the sample variance of $Y$ which is explained by the least squares fit".* $R^2$ is a crude measure of regression fit. We have better measures of fit, but these require a statistical (not just algebraic) analysis and we will return to these issues later. One deficiency with $R^2$ is that it increases when regressors are added to a regression so the "fit" can be always increased by increasing the number of regressors.

## 3.16. Regression Components

Partition $\boldsymbol{X}=[\boldsymbol{X}_1 \quad \boldsymbol{X}_2]$ and $\overrightarrow{\beta}=(\overrightarrow{\beta_1},\overrightarrow{\beta_2})$. The regression model can be written as

$$\overrightarrow{Y}=\boldsymbol{X}_1\overrightarrow{\beta_1}+\boldsymbol{X}_2\overrightarrow{\beta_2}+\overrightarrow{e}. \tag{3.31}$$

The OLS estimator of $\overrightarrow{\beta}$ is obtained by regression of $\overrightarrow{Y}$ on $\boldsymbol{X}=[\boldsymbol{X}_1 \quad \boldsymbol{X}_2]$ and can be written as

$$\overrightarrow{Y}=\boldsymbol{X}\widehat{\overrightarrow{\beta}}+\widehat{\overrightarrow{e}}=\boldsymbol{X}_1\widehat{\overrightarrow{\beta}}_1+\boldsymbol{X}_2\widehat{\overrightarrow{\beta}}_2+\widehat{\overrightarrow{e}}. \tag{3.32}$$

We are interested in algebraic expressions for $\widehat{\overrightarrow{\beta}}_1$ and $\widehat{\overrightarrow{\beta}}_2$.

Let's first focus on $\widehat{\overrightarrow{\beta}}_1$. The least squares estimator by definition is found by the joint minimization

$$\left(\widehat{\overrightarrow{\beta}}_1,\widehat{\overrightarrow{\beta}}_2\right)=\underset{\overrightarrow{\beta}_1,\overrightarrow{\beta}_2}{\operatorname{argmin}}\,\mathrm{SSE}\left(\overrightarrow{\beta}_1,\overrightarrow{\beta}_2\right) \tag{3.33}$$

where

$$\mathrm{SSE}\,(\overrightarrow{\beta}_1,\overrightarrow{\beta}_2)=(\overrightarrow{Y}-\boldsymbol{X}_1\overrightarrow{\beta}_1-\boldsymbol{X}_2\overrightarrow{\beta}_2)'\,(\overrightarrow{Y}-\boldsymbol{X}_1\overrightarrow{\beta}_1-\boldsymbol{X}_2\overrightarrow{\beta}_2)\,.$$

The solution (3.33) can be written as

$$\widehat{\overrightarrow{\beta}}_1=\underset{\overrightarrow{\beta}_1}{\operatorname{argmin}}\left(\underset{\overrightarrow{\beta}_2}{\operatorname{argmin}}\,\mathrm{SSE}\left(\overrightarrow{\beta}_1,\overrightarrow{\beta}_2\right)\right). \tag{3.34}$$

The inner expression minimizes over $\overrightarrow{\beta}_2$ while holding $\overrightarrow{\beta}_1$ fixed. it is the lowest possible sum of squared errors given $\overrightarrow{\beta}_1$. The outer minimization finds the coefficient $\overrightarrow{\beta}_1$ which minimizes the "lowest possible sum of squared errors given $\overrightarrow{\beta}_1$."

This means that $\widehat{\overrightarrow{\beta}}_1$ as defined in (3.33) and (3.34) are algebraically equivalent.

Examine the inner minimization problem in (3.34). This is simply the least squares regression of $\overrightarrow{Y} - \boldsymbol{X}_1\overrightarrow{\beta}_1$ on $\boldsymbol{X}_2$. This has solution

$$\operatorname*{argmin}_{\overrightarrow{\beta}_2} \mathrm{SSE}\left(\overrightarrow{\beta}_1, \overrightarrow{\beta}_2\right) = \left(\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\left(\overrightarrow{Y} - \boldsymbol{X}_1\overrightarrow{\beta}_1\right)\right)$$

with residuals

$$\boldsymbol{M}_2\left(\overrightarrow{Y} - \boldsymbol{X}_1\overrightarrow{\beta}_1\right),$$

where

$$\boldsymbol{M}_2 = \boldsymbol{I}_n - \boldsymbol{X}_2\left(\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2' \tag{3.35}$$

is the annihilator matrix for $\boldsymbol{X}_2$. This means that the inner minimization problem has minimized value

$$\min_{\overrightarrow{\beta}_2} \mathrm{SSE}\left(\overrightarrow{\beta}_1, \overrightarrow{\beta}_2\right) = \left(\overrightarrow{Y} - \boldsymbol{X}_1\overrightarrow{\beta}_1\right)'\boldsymbol{M}_2\left(\overrightarrow{Y} - \boldsymbol{X}_1\overrightarrow{\beta}_1\right).$$

Substituting this into (3.34) we find

$$\widehat{\overrightarrow{\beta}}_1 = \operatorname*{argmin}_{\overrightarrow{\beta}_1}\left(\overrightarrow{Y} - \boldsymbol{X}_1\overrightarrow{\beta}_1\right)'\boldsymbol{M}_2\left(\overrightarrow{Y} - \boldsymbol{X}_1\overrightarrow{\beta}_1\right) = \left(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}_1'\boldsymbol{M}_2\overrightarrow{Y}\right).$$

By a similar argument we find

$$\widehat{\overrightarrow{\beta}}_2 = \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{M}_1\overrightarrow{Y}\right)$$

where

$$\boldsymbol{M}_1 = \boldsymbol{I}_n - \boldsymbol{X}_1\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1' \tag{3.36}$$

is the annihilator matrix for $\boldsymbol{X}_1$.

**Theorem 3.4.**

The least squares estimator $\left(\widehat{\overrightarrow{\beta}}_1, \widehat{\overrightarrow{\beta}}_2\right)$ for (3.32) has the algebraic solution

$$\widehat{\overrightarrow{\beta}}_1 = \left(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}_1'\boldsymbol{M}_2\overrightarrow{Y}\right) \tag{3.37}$$

$$\widehat{\overrightarrow{\beta}}_2 = \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{M}_1\overrightarrow{Y}\right) \tag{3.38}$$

where $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are defined as follows,

$$\boldsymbol{M}_1 = \boldsymbol{I}_n - \boldsymbol{X}_1 \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1} \boldsymbol{X}_1'$$
$$\boldsymbol{M}_2 = \boldsymbol{I}_n - \boldsymbol{X}_2 \left(\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1} \boldsymbol{X}_2'.$$

## 3.17. Regression Components (Alternative Derivation)

An alternative proof of Theorem 3.4 uses an algebraic argument based on the population calculations from Section 2.22. Since this is a classic derivation we present it here for completeness.

Partition $\widehat{\boldsymbol{Q}}_{XX}$ as

$$\widehat{\boldsymbol{Q}}_{XX} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11} & \widehat{\boldsymbol{Q}}_{12} \\ \widehat{\boldsymbol{Q}}_{21} & \widehat{\boldsymbol{Q}}_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_1 & \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_2 \\ \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_1 & \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_2 \end{bmatrix}$$

and similarly $\widehat{\boldsymbol{Q}}_{XY}$ as

$$\widehat{\boldsymbol{Q}}_{XY} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{1Y} \\ \widehat{\boldsymbol{Q}}_{2Y} \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\boldsymbol{X}_1'\overrightarrow{Y} \\ \frac{1}{n}\boldsymbol{X}_2'\overrightarrow{Y} \end{bmatrix}.$$

By the partitioned matrix inversion formula,

$$\widehat{\boldsymbol{Q}}_{XX}^{-1} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11} & \widehat{\boldsymbol{Q}}_{12} \\ \widehat{\boldsymbol{Q}}_{21} & \widehat{\boldsymbol{Q}}_{22} \end{bmatrix}^{-1} \overset{\text{def}}{=} \begin{bmatrix} \widehat{\boldsymbol{Q}}^{11} & \widehat{\boldsymbol{Q}}^{12} \\ \widehat{\boldsymbol{Q}}^{21} & \widehat{\boldsymbol{Q}}^{22} \end{bmatrix} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1} & -\widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1} \\ -\widehat{\boldsymbol{Q}}_{22\cdot 1}^{-1}\widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1} & \widehat{\boldsymbol{Q}}_{22\cdot 1}^{-1} \end{bmatrix}, \tag{3.39}$$

where $\widehat{\boldsymbol{Q}}_{11\cdot 2} = \widehat{\boldsymbol{Q}}_{11} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{21}$ and $\widehat{\boldsymbol{Q}}_{22\cdot 1} = \widehat{\boldsymbol{Q}}_{22} - \widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{12}$. Thus

$$\begin{aligned} \overrightarrow{\widehat{\beta}} &= \begin{pmatrix} \overrightarrow{\widehat{\beta}}_1 \\ \overrightarrow{\widehat{\beta}}_2 \end{pmatrix} \\ &= \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1} & -\widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1} \\ -\widehat{\boldsymbol{Q}}_{22\cdot 1}^{-1}\widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1} & \widehat{\boldsymbol{Q}}_{22\cdot 1}^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{Q}}_{1Y} \\ \widehat{\boldsymbol{Q}}_{2Y} \end{bmatrix} \\ &= \begin{pmatrix} \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{1Y\cdot 2} \\ \widehat{\boldsymbol{Q}}_{22\cdot 1}^{-1}\widehat{\boldsymbol{Q}}_{2Y\cdot 1} \end{pmatrix}. \end{aligned}$$

Now

$$\begin{aligned}\widehat{\boldsymbol{Q}}_{11\cdot 2} &= \widehat{\boldsymbol{Q}}_{11} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{21}\\ &= \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_1 - \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_2\left(\frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_1\\ &= \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1\end{aligned}$$

and

$$\begin{aligned}\widehat{\boldsymbol{Q}}_{1Y\cdot 2} &= \widehat{\boldsymbol{Q}}_{1Y} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{2Y}\\ &= \frac{1}{n}\boldsymbol{X}_1'\overrightarrow{Y} - \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_2\left(\frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\frac{1}{n}\boldsymbol{X}_2'\overrightarrow{Y}\\ &= \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{M}_2\overrightarrow{Y}.\end{aligned}$$

Equation (3.38) follows.

## 3.18. Residual Regression

Take (3.38). Since $\boldsymbol{M}_1$ is idempotent, $\boldsymbol{M}_1 = \boldsymbol{M}_1\boldsymbol{M}_1$ and thus

$$\begin{aligned}\widehat{\overrightarrow{\beta}}_2 &= \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{M}_1\overrightarrow{Y}\right)\\ &= \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{M}_1\overrightarrow{Y}\right)\\ &= \left(\widetilde{\boldsymbol{X}}_2'\widetilde{\boldsymbol{X}}_2\right)^{-1}\left(\widetilde{\boldsymbol{X}}_2'\widetilde{\overrightarrow{e}}_1\right)\end{aligned}$$

where $\widetilde{\boldsymbol{X}}_2 = \boldsymbol{M}_1\boldsymbol{X}_2$ and $\widetilde{\overrightarrow{e}}_1 = \boldsymbol{M}_1\overrightarrow{Y}$.

*Thus the coefficient estimator $\widehat{\overrightarrow{\beta}}_2$ is algebraically equal to the least squares regression of $\widetilde{\overrightarrow{e}}$ on $\widetilde{\boldsymbol{X}}_2$.* Notice that these two are $\overrightarrow{Y}$ and $\boldsymbol{X}_2$, respectively, premultiplied by $\boldsymbol{M}_1$. But we know that pre-multiplication by $\boldsymbol{M}_1$ creates least squares residuals. *Therefore, $\widetilde{\overrightarrow{e}}_1$ is simply the least squares residual form a regression of $\overrightarrow{Y}$ on $\boldsymbol{X}_1$, and the columns of $\widehat{\boldsymbol{X}}_2$ are the least squares residuals from the regresions of the columns of $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$.*

**Theorem 3.5. Frisch-Waugh-Lovell (FWL)**

In the model (3.31), the OLS estimator of $\overrightarrow{\beta}_2$ and the OLS residuals $\widehat{\overrightarrow{e}}$ may be computed by either the OLS regression (3.32) or via the following algorithm:

(1) Regress $\overrightarrow{Y}$ on $\boldsymbol{X}_1$, obtain residuals $\widehat{\overrightarrow{e}}_1$;

(2) Regress $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$, obtain residuals $\widetilde{\boldsymbol{X}}_2$;

(3) Regress $\widehat{\overrightarrow{e}}_1$ on $\widetilde{\boldsymbol{X}}_2$, obtain OLS estimates $\widehat{\overrightarrow{\beta}}_2$ and residuals $\widehat{\overrightarrow{e}}$.

In some contexts (such as panel data models, to be introduced in Chapter 17), the FWL theorem can be used to greatly speed computation.

A common application of the FWL theorem is the demeaning formula for regression obtained in (3.18). Partition $\boldsymbol{X} = [\overrightarrow{X}_1 \boldsymbol{X}_2]$ where $\overrightarrow{X}_1 = \overrightarrow{1}_n$ is a vector of ones and $\boldsymbol{X}_2$ is a matrix of observed regressors. In this case,

$$\boldsymbol{M}_1 = \boldsymbol{I}_n - \overrightarrow{1}_n \left(\overrightarrow{1}_n' \overrightarrow{1}_n\right)^{-1} \overrightarrow{1}_n'.$$

Observe that

$$\widetilde{\boldsymbol{X}}_2 = \boldsymbol{M}_1 \boldsymbol{X}_2 = \boldsymbol{X}_2 - \overline{\boldsymbol{X}}_2$$

and

$$\boldsymbol{M}_2 \overrightarrow{Y} = \overrightarrow{Y} - \overline{\overrightarrow{Y}}$$

are the "demeaned" variables.

## 3.19. Leverage Values

*Derivations of the leverage values should be reviewed in the appendix.*

The *leverage values* for the regressor matrix $\boldsymbol{X}$ are the diagonal elements of the projection matrix $\boldsymbol{P} = \boldsymbol{X} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'$. There are $n$ leverage values, and are typically written as $h_{ii}$ for $i = 1, \ldots, n$. Since

$$\boldsymbol{P} = \begin{pmatrix} \overrightarrow{X}_1' \\ \overrightarrow{X}_2' \\ \vdots \\ \overrightarrow{X}_n' \end{pmatrix} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \begin{pmatrix} \overrightarrow{X}_1 & \overrightarrow{X}_2 & \cdots & \overrightarrow{X}_n \end{pmatrix}$$

they are

$$h_{ii} = \overrightarrow{X}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \overrightarrow{X}_i. \tag{3.40}$$

*The leverage value $h_{ii}$ is a normalized length of the observed regressor vector $\overrightarrow{X}_i$.* They appear frequently in the algebraic and statistical analysis of least squares regression, including leave-one-out regression, influential observations, robust covariance matrix estimation, and cross-validation.

**Theorem 3.6. Properties of the Leverage Values**

(1) $0 \le h_{ii} \le 1$.

(2) $h_{ii} \ge \frac{1}{n}$ if $\overrightarrow{X}$ includes an intercept.

(3) $\sum_{i=1}^{n} h_{ii} = k$.

*The leverage value $h_{ii}$ measures how unusual the ith observation $X_i$ is relative to the other observations in the sample. A large $h_{ii}$ occurs when $X_i$ is quite different from the other sample values.* A measure of overall unusualness is the maximum leverage value

$$\overline{h} = \max_{1 \le i \le n} h_{ii}. \tag{3.41}$$

It is common to say that a regression design is *balanced* when the leverage values are all roughly equal to one another. We know that complete balance occurs when $h_{ii} = \overline{h} = k/n$. An example of complete balance is when the regressors are all orthogonal dummy variables, each of which have equal occurrance of 0's and 1's.

A regression design is *unbalanced* if some leverage values are highly unequal from the others. The most extreme case is $\overline{h} = 1$. An example where this occurs is when there is a dummy regressor which takes the value 1 for only one observation in the sample.

*Some inference procedures (such as robust covariance matrix estimation and cross-validation) are sensitive to high leverage values. We will return to these issues later.*

*Proof.* We now prove Theorem 3.6. For part (1), let $\overrightarrow{s_i}$ be an $n \times 1$ unit vector with a 1 in the $i$th place and zeros elsewhere so that $h_{ii} = \overrightarrow{s_i}'\boldsymbol{P}\overrightarrow{s_i}$. Then we know

$$h_{ii} = \overrightarrow{s_i}'\boldsymbol{P}\overrightarrow{s_i} \le \overrightarrow{s_i}'\overrightarrow{s_i}\lambda_{\max}(\boldsymbol{P}) = 1$$

as claimed.

For part (2), partition $\overrightarrow{X}_i = \left(1, \overrightarrow{Z}_i'\right)'$. Without loss of generality we can replace $\overrightarrow{Z}_i$ with the demeaned values $\overrightarrow{Z}_i^* = \overrightarrow{Z}_i - \overline{\overrightarrow{Z}}_i$. Then since $\overrightarrow{Z}_i^*$ and the intercept are orthogonal

$$\begin{aligned} h_{ii} &= \left(1, \overrightarrow{Z}_i^{*\prime}\right) \begin{bmatrix} n & 0 \\ 0 & \overrightarrow{Z}^{*\prime}\overrightarrow{Z}^{*} \end{bmatrix}^{-1} \begin{pmatrix} 1 \\ \overrightarrow{Z}_i^{*} \end{pmatrix} \\ &= \frac{1}{n} + \overrightarrow{Z}_i^{*\prime}\left(\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^{*}\right)^{-1}\overrightarrow{Z}_i^{*} \ge \frac{1}{n}. \end{aligned}$$

For part (3), $\sum_{i=1}^{n} h_{ii} = \operatorname{tr}\boldsymbol{P} = k$. □

## 3.20. Leave-One-Out Regression

There are a number of statistical procedures – residual analysis, jackknife variance estimation, cross-validation, two-step estimation, hold-out sample evaluation – which make use of estimators constructed on sub-samples. Of particular importance is the case where we exclude a single observation and then repeat this for all observations. This is called *leave-one-out (LOO)* regression.

Specifically, the leave-one-out estimator of the regression coefficient $\overrightarrow{\beta}$ is the least squares estimator constructed using the full sample excluding a single observation $i$. This can be written as

$$\begin{aligned}\widehat{\overrightarrow{\beta}}_{(-i)} &= \left(\sum_{j\neq i}\overrightarrow{X}_j\overrightarrow{X}_j'\right)^{-1}\left(\sum_{j\neq i}\overrightarrow{X}_jY_j\right)\\ &= \left(\boldsymbol{X}'\boldsymbol{X}-\overrightarrow{X}_i\overrightarrow{X}_i'\right)^{-1}\left(\boldsymbol{X}'\overrightarrow{Y}-\overrightarrow{X}_iY_i\right)\\ &= \left(\boldsymbol{X}_{(-i)}'\boldsymbol{X}_{(-i)}\right)^{-1}\left(\boldsymbol{X}_{(-i)}'\overrightarrow{Y}_{(-i)}\right).\end{aligned} \tag{3.42}$$

Here, $\boldsymbol{X}_{(-i)}$ and $\overrightarrow{Y}_{(-i)}$ are the data matrices omitting the $i$th row. There is a leave-one-out estimator for each observation, $i = 1, \ldots, n$, so we have $n$ such estimators.

The leave-one-out predicted value for $Y_i$ is $\widetilde{Y}_i = \overrightarrow{X}_i'\widehat{\overrightarrow{\beta}}_{(-i)}$. This is the predicted value obtained by estimating $\overrightarrow{\beta}$ on the sample without observation $i$ and then using the covariate vector $\overrightarrow{X}_i$ to predict $Y_i$. Notice that $\widetilde{Y}_i$ is an authentic prediction as $Y_i$ is not used to construct $\widetilde{Y}_i$. This is in contrast to the fitted values $\widehat{Y}_i$ which are functions of $Y_i$.

The *leave-one-out residual*, *prediction error*, or *prediction residual* is $\widetilde{e}_i = Y_i - \widetilde{Y}_i$. The prediction errors may be used as estimators of the errors instead of the residuals. The prediction errors are better estimators than the residuals since the former are based on authentic predictions.

**Theorem 3.7.**

The leave-one-out estimator and prediction error equal

$$\widehat{\overrightarrow{\beta}}_{(-i)} = \widehat{\overrightarrow{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\overrightarrow{X}_i\widetilde{e}_i \tag{3.43}$$

and

$$\widetilde{e}_i = (1-h_{ii})^{-1}\widehat{e}_i \tag{3.44}$$

where $h_{ii}$ are the leverage values as defined in (3.40).

Equation (3.43) shows that the leave-one-out coefficients can be calculated by a simple linear operation and do not need to be calculated using $n$ separate regressions. Another interesting feature of equation (3.44) is that the prediction errors $\widetilde{e}_i$ are a simple scaling of the least squares residuals $\widehat{e}_i$ with the scaling dependent on the leverage values $h_{ii}$. If $h_{ii}$ is small then $\widetilde{e}_i \simeq \widehat{e}_i$. However, if $h_{ii}$ is large then $\widetilde{e}_i$ can be quite different from $\widehat{e}_i$. Thus the difference between the residuals and predicted values depends on the leverage values, that is, how unusual is $X_i$.

To write (3.44) in vector notation, define

$$\boldsymbol{M}^{*} = \left(\boldsymbol{I}_n - \operatorname{diag}\left(h_{11}, \ldots, h_{nn}\right)\right)^{-1} = \operatorname{diag}\left(\left(1-h_{11}\right)^{-1}, \ldots, \left(1-h_{nn}\right)^{-1}\right).$$

Then (3.44) is equivalent to

$$\overrightarrow{\widetilde{e}} = \boldsymbol{M}^{*}\overrightarrow{\widehat{e}}. \tag{3.45}$$

One use of the prediction errors is to estimate the out-of-sample mean squared error:

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\widetilde{e}_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left(1-h_{ii}\right)^{-2}\widehat{e}_i^2. \tag{3.46}$$

This is known as the *sample mean squared prediction error*. Its square root is the *prediction standard error*.

*Proof.* We complete the section with a proof of Theorem 3.7. The leave-one-out estimator (3.42) can be written as

$$\overrightarrow{\widehat{\beta}}_{(-i)} = \left(\boldsymbol{X}'\boldsymbol{X} - \overrightarrow{X}_i\overrightarrow{X}_i'\right)^{-1}\left(\boldsymbol{X}'\overrightarrow{Y} - \overrightarrow{X}_iY_i\right). \tag{3.47}$$

Multiple (3.47) by $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{X} - \overrightarrow{X}_i\overrightarrow{X}_i'\right)$. We obtain

$$\overrightarrow{\widehat{\beta}}_{(-i)} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\overrightarrow{X}_i\overrightarrow{X}_i'\overrightarrow{\widehat{\beta}}_{(-i)} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\overrightarrow{Y} - \overrightarrow{X}_iY_i\right) = \overrightarrow{\widehat{\beta}} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\overrightarrow{X}_iY_i.$$

Rewriting

$$\overrightarrow{\widehat{\beta}}_{(-i)} = \overrightarrow{\widehat{\beta}} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\overrightarrow{X}_i\left(Y_i - \overrightarrow{X}_i'\overrightarrow{\widehat{\beta}}_{(-i)}\right) = \overrightarrow{\widehat{\beta}} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\overrightarrow{X}_i\widetilde{e}_i$$

which is (3.43). Premultiplying this expression by $\overrightarrow{X}_i'$ and using the definition (3.40) we obtain

$$\overrightarrow{X}_i'\overrightarrow{\widehat{\beta}}_{(-i)} = \overrightarrow{X}_i\overrightarrow{\widehat{\beta}} - \overrightarrow{X}_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\overrightarrow{X}_i\widetilde{e}_i = \overrightarrow{X}_i\overrightarrow{\widehat{\beta}} - h_{ii}\widetilde{e}_i. \tag{3.48}$$

Using the definitions for $\widehat{e}_i$ and $\widetilde{e}_i$ we obtain $\widetilde{e}_i = \widehat{e}_i + h_{ii}\widetilde{e}_i$. Rewriting we obtain (3.44). □

## 3.21. Influential Observations

Another use of the leave-one-out estimator is to investigate the impact of *influential observations*, sometimes called *outliers*. *We say that observation i is influential if its omission from the sample induces a substantial change in a parameter estimate of interest.*

From (3.43) we know that

$$\overrightarrow{\widehat{\beta}} - \overrightarrow{\widehat{\beta}}_{(-i)} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\overrightarrow{X}_i\widetilde{e}_i. \tag{3.49}$$

By direct calculation of this quantity for each observation $i$, we can directly discover if a specific observation $i$ is influential for a coefficient estimate of interest.

For a general assessment, we can focus on the predicted values. The difference between the full-sample and leave-one-out predicted values is

$$\widehat{Y}_i - \widetilde{Y}_i = \overrightarrow{X}_i' \widehat{\overrightarrow{\beta}} - \overrightarrow{X}_i' \widehat{\overrightarrow{\beta}}_{(-i)} = \overrightarrow{X}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \overrightarrow{X}_i \widetilde{e}_i = h_{ii}\widetilde{e}_i,$$

which is a simple function of the leverage values $h_{ii}$ and prediction errors $\widetilde{e}_i$. Observation $i$ is influential for the predicted value if $\left|h_{ii}\widetilde{e}_i\right|$ is large, which requires that both $h_{ii}$ and $\left|\widetilde{e}_i\right|$ are large.

One way to think about this is that a large leverage value $h_{ii}$ gives the potential for observation $i$ to be influential. A large $h_{ii}$ means that observation $i$ is unusual in the sense that the regressor $\overrightarrow{X}_i$ is far from its sample mean. We call an observation with large $h_{ii}$ a *leverage point*. A leverage point is not necessarily influential as the latter also requires that the prediction error $\widetilde{e}_i$ is large.

If an observation is determined to be influential what should be done? As a common cause of influential observations is data error, the influential observations should be examined for evidence that the observation was mis-recorded. Perhaps the observation falls outside of permitted ranges, or some observables are inconsistent. If it is determined that an observation is incorrectly recorded, then the observation is typically deleted from the sample. This process is often called “cleaning the data”. The decisions made in this process involve a fair amount of individual judgment.

## 3.22. Collinearity Errors

How can we numerically check if a matrix $\boldsymbol{A}$ is singular? A standard diagnostic is the reciprocal condition number

$$C = \frac{\lambda_{\min}(\boldsymbol{A})}{\lambda_{\max}(\boldsymbol{A})}.$$

If $C = 0$, then $\boldsymbol{A}$ is singular. If $C = 1$, then $\boldsymbol{A}$ is perfectly balanced. If $C$ is extremely small we say that $\boldsymbol{A}$ is *ill-conditioned*.

# Chapter 4

# Least Squares Regression

## 4.1. Random Sampling

Assumption 3.1 specified that the observations have identical distributions. To derive the finite-sample properties of the estimators we will need to additionally specify the dependence structure across the observations.

The simplest context is when the observations are mutually independent in which case we say that they are *independent and identically distributed* or *i.i.d.* It is also common to describe i.i.d. observations as a *random sample*.

**Assumption 4.1.**

The random variables $\{(Y_1, \overrightarrow{X}_1), \ldots, (Y_i, \overrightarrow{X}_i), \ldots, (Y_n, \overrightarrow{X}_n)\}$ are independent and identically distributed.

For most of this chapter we will use Assumption 4.1 to derive properties of the OLS estimator. It means that if you take any two individuals $i \neq j$ in a sample, the values $(Y_i, \overrightarrow{X}_i)$ are independent of the values $\left(Y_j, \overrightarrow{X}_j\right)$ yet have the same distribution. Independence means that the decisions and choices of individual $i$ do not affect the decisions of individual $j$ and conversely.

This assumption may be violated if individuals in the sample are connected in some way, for example, if they are neighbors, members of the same village, classmates at a school, or even firms within a specific industry. In this case it seems plausible that decisions may be inter-connected and thus mutually dependent rather than independent. Allowing for such interactions complicates inference and requires specialized treatment. A currently popular approach which allows for mutual dependence is known as *clustered dependence* which assumes that observations are grouped into “clusters” (for example, schools). We will discuss clustering in more detail in Section 4.23.

## 4.2. Sample Mean

We start with the simplest setting of the intercept-only model.

$$\begin{aligned} Y &= \mu + e \\ \mathbb{E}[e] &= 0, \end{aligned}$$

which is equivalent to the regression model with $k = 1$ and $X_i = 1$. In the intercept model $\mu = \mathbb{E}[Y]$ is the expectation of $Y_i$. The least squares estimator $\widehat{\mu} = \overline{Y}$ equals the sample mean as shown in equation (3.8).

It is easy to calculate the expectation of the estimator $\overline{Y}$ under Assumption 4.1.

$$\mathbb{E}\left[\overline{Y}\right] = \mu.$$

This shows that the expected value of the least squares estimator (the sample mean) equals the projection coefficient (the population expectation). An estimator with the property that its expectation equals the parameter it is estimating is called *unbiased*.

**Definition 4.1.**

An estimator $\widehat{\theta}$ for $\theta$ is unbiased if $\mathbb{E}\left[\widehat{\theta}\right] = \theta$.

We next calculate the variance of the estimator $\overline{Y}$ under Assumption 4.1. Making the substitution $Y_i = \mu + e_i$, we find

$$\overline{Y} - \mu = \frac{1}{n}\sum_{i=1}^{n} e_i.$$

Then,

$$\begin{aligned} \operatorname{var}\left[\overline{Y}\right] &= \mathbb{E}\left[\left(\overline{Y} - \mu\right)^2\right] \\ &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} e_i\right)\left(\frac{1}{n}\sum_{j=1}^{n} e_j\right)\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[e_i e_j\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 \\ &= \frac{1}{n}\sigma^2. \end{aligned}$$

The fourth equality is because $\mathbb{E}\left[e_i e_j\right] = \sigma^2$ for $i = j$ yet $\mathbb{E}\left[e_i e_j\right] = 0$ for $i \neq j$ due to independence.

## 4.3. Linear Regression Model

**Assumption 4.2. Linear Regression Model**

The variables $(Y, \overrightarrow{X})$ satisfy the linear regression equation

$$Y = \overrightarrow{X}'\overrightarrow{\beta} + e \tag{4.1}$$

$$\mathbb{E}[e \mid \overrightarrow{X}] = 0. \tag{4.2}$$

The variables have finite second moments

$$\mathbb{E}\left[Y^2\right] < \infty,$$

$$\mathbb{E}\left[\|\overrightarrow{X}\|^2\right] < \infty,$$

and an invertible design matrix

$$\boldsymbol{Q}_{XX} = \mathbb{E}[\overrightarrow{X}\overrightarrow{X}'] > 0.$$

We will consider both the general case of heteroskedastic regression where the conditional variance $\mathbb{E}\left[e^2 \mid \overrightarrow{X}\right] = \sigma^2(\overrightarrow{X})$ is unrestricted, and the specialized case of homoskedastic regression where the conditional variance is constant. In the latter case we add the following assumption.

**Assumption 4.3. Homoskedastic Linear Regression Model**

In addition to Assumption 4.2,

$$\mathbb{E}\left[e^2 \mid \overrightarrow{X}\right] = \sigma^2(\overrightarrow{X}) = \sigma^2 \tag{4.3}$$

is independent of $\overrightarrow{X}$.

## 4.4. Expectation of Least Squares Estimator

Observe first that under (4.1) and (4.2),

$$\mathbb{E}[Y_i \mid \overrightarrow{X}_1, \ldots, \overrightarrow{X}_n] = \mathbb{E}[Y_i \mid \overrightarrow{X}_i] = \overrightarrow{X}_i'\overrightarrow{\beta}. \tag{4.4}$$

The first equality states that the conditional expectation of $Y_i$ given $\{\overrightarrow{X}_1, \ldots, \overrightarrow{X}_n\}$ only depends on $\overrightarrow{X}_i$ since the observations are independent across $i$. The second equality is the assumption of a linear conditional expectation.

Using definition (3.11), the conditioning theorem (Theorem 2.3), the linearity of expectations, (4.4), and properties of the matrix inverse,

$$\begin{aligned}
\mathbb{E}\left[\widehat{\beta} \mid \overrightarrow{X}_1, \dots, \overrightarrow{X}_n\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{n} \overrightarrow{X}_i \overrightarrow{X}_i'\right)^{-1} \left(\sum_{i=1}^{n} \overrightarrow{X}_i Y_i\right) \mid \overrightarrow{X}_1, \dots, \overrightarrow{X}_n\right] \\
&= \left(\sum_{i=1}^{n} \overrightarrow{X}_i \overrightarrow{X}_i'\right)^{-1} \mathbb{E}\left[\left(\sum_{i=1}^{n} \overrightarrow{X}_i Y_i\right) \mid \overrightarrow{X}_1, \dots, \overrightarrow{X}_n\right] \\
&= \left(\sum_{i=1}^{n} \overrightarrow{X}_i \overrightarrow{X}_i'\right)^{-1} \sum_{i=1}^{n} \mathbb{E}\left[\overrightarrow{X}_i Y_i \mid \overrightarrow{X}_1, \dots, \overrightarrow{X}_n\right] \\
&= \left(\sum_{i=1}^{n} \overrightarrow{X}_i \overrightarrow{X}_i'\right)^{-1} \sum_{i=1}^{n} \overrightarrow{X}_i \mathbb{E}\left[Y_i \mid \overrightarrow{X}_i\right] \\
&= \left(\sum_{i=1}^{n} \overrightarrow{X}_i \overrightarrow{X}_i'\right)^{-1} \sum_{i=1}^{n} \overrightarrow{X}_i \overrightarrow{X}_i' \overrightarrow{\beta} \\
&= \overrightarrow{\beta}.
\end{aligned}$$

Now let's show the same result using matrix notation. (4.4) implies

$$\mathbb{E}\left[\overrightarrow{Y} \mid \boldsymbol{X}\right] = \boldsymbol{X}\overrightarrow{\beta}. \tag{4.5}$$

Similarly,

$$\mathbb{E}\left[\overrightarrow{e} \mid \boldsymbol{X}\right] = \overrightarrow{0}.$$

Using $\widehat{\overrightarrow{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\overrightarrow{Y})$, the conditioning theorem, the linearity of expectations, (4.5), and the properties of the matrix inverse,

Then we can get

$$\begin{aligned}
\mathbb{E}\left[\widehat{\overrightarrow{\beta}} \mid \boldsymbol{X}\right] &= \mathbb{E}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\overrightarrow{Y} \mid \boldsymbol{X}\right] \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathbb{E}\left[\overrightarrow{Y} \mid \boldsymbol{X}\right] \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\overrightarrow{\beta} \\
&= \overrightarrow{\beta}.
\end{aligned}$$

Another way to calculate the same result is as follows. Insert $\overrightarrow{Y} = \boldsymbol{X}\overrightarrow{\beta} + e$ into the formula for $\widehat{\overrightarrow{\beta}}$ to obtain

$$\begin{aligned}
\widehat{\overrightarrow{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\boldsymbol{X}'\left(\boldsymbol{X}\overrightarrow{\beta} + \overrightarrow{e}\right)\right) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})\overrightarrow{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\overrightarrow{e} \\
&= \overrightarrow{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\overrightarrow{e}.
\end{aligned} \tag{4.6}$$

This is a useful linear decomposition of the estimator $\widehat{\vec{\beta}}$ into the true parameter $\vec{\beta}$ and the stochastic component $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\vec{e}$. Once again, we can calculate that

$$\begin{aligned}\mathbb{E}\left[\widehat{\vec{\beta}}-\vec{\beta}\mid \boldsymbol{X}\right] &= \mathbb{E}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\vec{e}\mid \boldsymbol{X}\right]\\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathbb{E}\left[\vec{e}\mid \boldsymbol{X}\right]=\vec{0}.\end{aligned}$$

**Theorem 4.1. Expectation of Least Squares Estimator**

In the linear regression model (Assumption 4.2) with i.i.d. sampling (Assumption 4.1),

$$\mathbb{E}\left[\widehat{\vec{\beta}}\mid \boldsymbol{X}\right]=\vec{\beta}. \tag{4.7}$$

Equation (4.7) says that the estimator $\widehat{\vec{\beta}}$ is unbiased for $\vec{\beta}$, conditional on $\boldsymbol{X}$. This means that the conditional distribution of $\widehat{\vec{\beta}}$ is centered at $\vec{\beta}$. By "conditional on $\boldsymbol{X}$" this means that the distribution is unbiased (centered at $\vec{\beta}$) for any realization of the regressor matrix $\boldsymbol{X}$. *It is worth mentioning that Theorem 4.1, and all finite sample results in this chapter, make the implicit assumption that $\boldsymbol{X}'\boldsymbol{X}$ is full rank with probability one.*

## 4.5. Variance of Least Squares Estimator

For any $r\times 1$ random vector $\vec{Z}$ define the $r\times r$ covariance matrix,

$$\operatorname{var}[\vec{Z}]=\mathbb{E}\left[(\vec{Z}-\mathbb{E}[\vec{Z}])(\vec{Z}-\mathbb{E}[\vec{Z}])'\right]=\mathbb{E}[\vec{Z}\vec{Z}']-(\mathbb{E}[\vec{Z}])(\mathbb{E}[\vec{Z}])'$$

and for any pair $(\vec{Z},X)$ define the conditional covariance matrix

$$\operatorname{var}[\vec{Z}\mid X]=\mathbb{E}\left[(\vec{Z}-\mathbb{E}[\vec{Z}\mid X])(\vec{Z}-\mathbb{E}[\vec{Z}\mid X])'\mid X\right].$$

We define

$$\boldsymbol{V}_{\widehat{\vec{\beta}}}\coloneqq\operatorname{var}\left[\widehat{\vec{\beta}}\mid \boldsymbol{X}\right]$$

as the conditional covariance matrix of the regression coefficient estimators.

The conditional covariance matrix of the $n\times 1$ regression error $\vec{e}$ is the $n\times n$ matrix

$$\operatorname{var}[\vec{e}\mid \boldsymbol{X}]=\mathbb{E}[\vec{e}\,\vec{e}'\mid \boldsymbol{X}]\coloneqq \boldsymbol{D}.$$

The $i$th diagonal element of $\boldsymbol{D}$ is

$$\mathbb{E}\left[e_i^2\mid \boldsymbol{X}\right]=\mathbb{E}[e_i\mid X_i]=\sigma_i^2$$

while the $ij$th off-diagonal element of $\boldsymbol{D}$ is

$$\mathbb{E}\left[e_i e_j \mid \boldsymbol{X}\right] = \mathbb{E}\left[e_i \mid X_i\right] \mathbb{E}\left[e_j \mid X_j\right] = 0,$$

where the first equality uses independence of the observations (Assumption 4.1) and the second is equation (4.2). Thus $\boldsymbol{D}$ is a diagonal matrix with $i$th diagonal element $\sigma_i^2$:

$$\boldsymbol{D} = \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_n^2\right) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix} \tag{4.8}$$

In the special case of the linear homoskedastic regression model (Assumption 4.3), $\boldsymbol{D} = \boldsymbol{I}_n \sigma^2$. In general, however, $\boldsymbol{D}$ need not necessarily take this simplified form.

For any $n \times r$ matrix $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{X})$,

$$\operatorname{var}\left[\boldsymbol{A}'\overrightarrow{Y} \mid \boldsymbol{X}\right] = \operatorname{var}\left[\boldsymbol{A}'\overrightarrow{e} \mid \boldsymbol{X}\right] = \boldsymbol{A}'\boldsymbol{D}\boldsymbol{A}. \tag{4.9}$$

In particular, we can write $\widehat{\overrightarrow{\beta}} = \boldsymbol{A}'\boldsymbol{Y}$, where $\boldsymbol{A} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ and thus

$$\boldsymbol{V}_{\widehat{\overrightarrow{\beta}}} = \operatorname{var}\left[\widehat{\overrightarrow{\beta}} \mid \boldsymbol{X}\right] = \boldsymbol{A}'\boldsymbol{D}\boldsymbol{A} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

It is useful to note that

$$\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X} = \sum_{i=1}^{n} \overrightarrow{X}_i \overrightarrow{X}_i' \sigma_i^2,$$

a weighted version of $\boldsymbol{X}'\boldsymbol{X}$.

In the special case of the linear homoskedastic regression model, $\boldsymbol{D} = \boldsymbol{I}_n\sigma^2$, so $\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X} = \boldsymbol{X}'\boldsymbol{X}\sigma^2$, and the covariance matrix simplifies to $\boldsymbol{V}_{\widehat{\overrightarrow{\beta}}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2$.

**Theorem 4.2. Variance of Least Squares Estimator**

In the linear regression model (Assumption 4.2) with i.i.d. sampling (Assumption 4.1),

$$\boldsymbol{V}_{\widehat{\overrightarrow{\beta}}} = \operatorname{var}\left[\widehat{\overrightarrow{\beta}} \mid \boldsymbol{X}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \tag{4.10}$$

where $\boldsymbol{D}$ is defined in (4.8), as

$$\boldsymbol{D} = \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_n^2\right) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix},$$

where $\sigma_i^2 = \mathbb{E}\left[e_i^2 \mid \overrightarrow{X}_i\right]$.

If in addition the error is homoskedastic (Assumption 4.3) then (4.10) simplifies to

$$\boldsymbol{V}_{\widehat{\overrightarrow{\beta}}} = \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

## 4.6. Unconditional Moments

The previous sections derived the form of the conditional mean and variance of the least squares estimator where we conditioned on the regressor matrix $\boldsymbol{X}$. What about the unconditional mean and variance?

Indeed, it is not obvious if $\widehat{\overrightarrow{\beta}}$ has a finite mean or variance. Take the case of a single dummy variable regressor $D_i$ with no intercept. Assume $\mathbb{P}\left[D_i\right] = p < 1$. Then

$$\widehat{\beta} = \frac{\sum_{i=1}^n D_i Y_i}{\sum_{i=1}^n D_i}$$

is well defined if $\sum_{i=1}^n D_i > 0$. However, $\mathbb{P}\left[\sum_{i=1}^n D_i = 0\right] = (1-p)^n > 0$. This means that with positive (but small) probability $\widehat{\beta}$ does not exist. Consequently, $\widehat{\beta}$ has no finite moments! We ignore this complication in practice but it does pose a conundrum for theory. This existence problem arises whenever there are discrete regressors.

This dilemma is avoided when the regressors have continuous distributions. A clean statement was obtained by Kinal (1980) under the assumption of normal regressors and errors.

**Theorem 4.3. Kinal (1980)**

In the linear regression model with i.i.d. sampling, if in addition $(X, e)$ have a joint normal distribution then for any $r$, $\mathbb{E}\left[\left\|\widehat{\overrightarrow{\beta}}\right\|^r\right] < \infty$ if and only if $r < n - k + 1$.

This shows that when the errors and regressors are normally distributed that the least squares estimator possesses all moments up to $n-k$ which includes all moments of practical interest. The normality assumption is not critical for this result. What is key is the assumption that the regressors are continuously distributed.

Therefore, if $n > k$,

$$\mathbb{E}\left[\widehat{\overrightarrow{\beta}}\right] = \mathbb{E}\left[\mathbb{E}\left[\widehat{\overrightarrow{\beta}} \mid \boldsymbol{X}\right]\right] = \overrightarrow{\beta}.$$

Here $\widehat{\overrightarrow{\beta}}$ is unconditionally unbiased as asserted.

Furthermore, if $n - k > 1$, then $\mathbb{E}\left[\left\|\widehat{\beta}\right\|^2\right] < \infty$ and $\widehat{\beta}$ has a finite unconditional variance. We can calculate explicitly that

$$\operatorname{var}\left[\widehat{\overrightarrow{\beta}}\right] = \mathbb{E}\left[\operatorname{var}\left[\widehat{\overrightarrow{\beta}} \mid \boldsymbol{X}\right]\right] + \operatorname{var}\left[\mathbb{E}\left[\widehat{\overrightarrow{\beta}} \mid \boldsymbol{X}\right]\right] = \mathbb{E}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}\right],$$

the second equality since $\mathbb{E}\left[\widehat{\overrightarrow{\beta}} \mid \boldsymbol{X}\right] = \overrightarrow{\beta}$ has zero variance.

In the homoskedastic case this simplifies to

$$\operatorname{var}\left[\widehat{\overrightarrow{\beta}}\right] = \sigma^2\,\mathbb{E}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right].$$

*In both cases the expectation cannot pass through the matrix inverse since this is a nonlinear function. Thus there is not a simple expression for the unconditional variance, other than stating that is it the mean of the conditional variance.*

## 4.7. Gauss-Markov Theorem

Consider the class of estimators of $\overrightarrow{\beta}$ which are linear functions of the vector $\overrightarrow{Y}$ and thus can be written as $\widetilde{\overrightarrow{\beta}} = \boldsymbol{A}'\overrightarrow{Y}$, where $\boldsymbol{A}$ is an $n\times k$ function of $\boldsymbol{X}$. As noted before, the least squares estimator is the special case obtained by setting $\boldsymbol{A} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. What is the best choice of $\boldsymbol{A}$? *The Gauss-Markov Theorem says that the least squares estimator is the best choice among linear unbiased estimators when the errors are homoskedastic, in the sense that the least squares estimator has the smallest variance among all unbiased linear estimators.*

To see this, since $\mathbb{E}[\overrightarrow{Y} \mid \boldsymbol{X}] = \boldsymbol{X}\overrightarrow{\beta}$ then for any linear estimator $\widetilde{\overrightarrow{\beta}} = \boldsymbol{A}'\overrightarrow{Y}$ we have

$$\mathbb{E}\left[\widetilde{\overrightarrow{\beta}} \mid \boldsymbol{X}\right] = \boldsymbol{A}'\,\mathbb{E}[\overrightarrow{Y} \mid \boldsymbol{X}] = \boldsymbol{A}'\boldsymbol{X}\overrightarrow{\beta},$$

so $\widetilde{\overrightarrow{\beta}}$ is unbiased if and only if $\boldsymbol{A}'\boldsymbol{X} = \boldsymbol{I}_k$. Further more, we saw in (4.9) that

$$\operatorname{var}\left[\widetilde{\overrightarrow{\beta}} \mid \boldsymbol{X}\right] = \operatorname{var}[\boldsymbol{A}'\overrightarrow{Y} \mid \boldsymbol{X}] = \boldsymbol{A}'\boldsymbol{D}\boldsymbol{A} = \sigma^2(\boldsymbol{A}'\boldsymbol{A}),$$

where the last equality using the homoskedasticity assumption $\boldsymbol{D} = \mathbb{I}_n\sigma^2$. The "best" unbiased linear estimator is obtained by finding the matrix $\boldsymbol{A}_0$ satisfying

$$\boldsymbol{A}_0'\boldsymbol{X} = \boldsymbol{I}_k,$$

and $\boldsymbol{A}'\boldsymbol{A}$ is minimized in the positive definite sense, which means that for any other matrix $\boldsymbol{A}$ satisfying $\boldsymbol{A}'\boldsymbol{X} = \boldsymbol{I}_k$ then

$$\boldsymbol{A}'\boldsymbol{A} - \boldsymbol{A}_0'\boldsymbol{A}_0 \text{ is positive semi-definite.}$$

**Theorem 4.4. Gauss-Markov**

In the homoskedastic linear regression model (Assumption 4.3) with iid sampling (Assumption 4.1), if $\widetilde{\overrightarrow{\beta}}$ is a linear unbiased estimator of $\overrightarrow{\beta}$ then

$$\operatorname{var}\left[\widetilde{\overrightarrow{\beta}} \mid \boldsymbol{X}\right] \geq \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

The Gauss-Markov theorem provides a lower bound on the covariance matrix of unbiased linear estimators under the assumption of homoskedasticity. It says that no unbiased linear estimator can have a variance matrix smaller (in the positive definite sense) than $\sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$. Since the variance of the OLS estimator is exactly equal to this bound this means that the OLS estimator is efficient in the class of linear unbiased estimators. This gives rise to the description of OLS as BLUE, standing for *best linear unbiased estimator*. This is an efficiency justification for the least squares estimator. *The justification is limited because the class of models is restricted to homoskedastic regressions and the class of potential estimators is restricted to linear unbiased estimators.* This latter restriction is particularly unsatisfactory as there is no sensible motivation for focusing on linear estimators.

*Proof.* We complete this section with a proof of the Gauss-Markov theorem.

Let $\boldsymbol{A}$ be any $n \times k$ function of $\boldsymbol{X}$ such that $\boldsymbol{A}'\boldsymbol{X} = \boldsymbol{I}_k$. This ensures that the estimator $\boldsymbol{A}'\overrightarrow{Y}$ is unbiased for $\overrightarrow{\beta}$. Also, this estimator has variance $\boldsymbol{A}'\boldsymbol{A}\sigma^2$. Since the least squares estimator is unbiased and has variance $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2$, it is sufficient to show that the difference in the two variance matrices is positive semi-definite, or

$$\boldsymbol{C} = \boldsymbol{A}'\boldsymbol{A} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \geq 0. \tag{4.11}$$

Let $\boldsymbol{D} = \boldsymbol{A}' - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ or $\boldsymbol{A}' = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' + \boldsymbol{D}$. Notice that $\boldsymbol{D}\boldsymbol{X} = \boldsymbol{A}'\boldsymbol{X} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_k - \boldsymbol{I}_k = \boldsymbol{0}$.

Then

$$\begin{aligned}
\boldsymbol{C} &= \boldsymbol{A}'\boldsymbol{A} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' + \boldsymbol{D}\right)\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' + \boldsymbol{D}\right)' - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{D}' + \boldsymbol{D}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} + \boldsymbol{D}\boldsymbol{D}' - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \boldsymbol{D}\boldsymbol{D}' \geq \boldsymbol{0}.
\end{aligned}$$

The final inequality states that the matrix $\boldsymbol{D}\boldsymbol{D}'$ is positive semi-definite which is a property of quadratic forms. □

## 4.8. Modern Gauss-Markov Theorem

**Theorem 4.5. Modern Gauss-Markov**

In the linear regression model with i.i.d. sampling, if $\mathbb{E}\left[\widetilde{\overrightarrow{\beta}} \mid \boldsymbol{X}\right] = \overrightarrow{\beta}$ and Assumption 4.3 holds then $\operatorname{var}\left[\widetilde{\overrightarrow{\beta}} \mid \boldsymbol{X}\right] \geq \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$.

## 4.9. Generalized Least Squares

Take the liner regression model in matrix format

$$\overrightarrow{Y} = \boldsymbol{X}\overrightarrow{\beta} + \overrightarrow{e}. \tag{4.12}$$

Consider a generalized situation where the observation errors are possibly correlated and/or heteroskedastic. Specifically, suppose that

$$\mathbb{E}\left[\overrightarrow{e} \mid \boldsymbol{X}\right] = 0 \tag{4.13}$$

$$\operatorname{var}\left[\overrightarrow{e} \mid \boldsymbol{X}\right] = \boldsymbol{\Omega} \tag{4.14}$$

for some $n \times n$ covariance matrix $\boldsymbol{\Omega}$, possibly a function of $\boldsymbol{X}$. This includes the iid sampling framework where $\boldsymbol{\Omega} = \boldsymbol{D}$ as defined in (4.8) but allows for non-diagonal covariance matrices as well. As a covariance matrix, $\boldsymbol{\Omega}$ is necessarily symmetric and positive semi-definite.

Under these assumptions, by arguments similar to the previous section, we can calculate the mean and variance of the OLS estimator:

$$\mathbb{E}\left[\widehat{\overrightarrow{\beta}} \mid \boldsymbol{X}\right] = \overrightarrow{\beta} \tag{4.15}$$

$$\operatorname{var}\left[\widehat{\overrightarrow{\beta}} \mid \boldsymbol{X}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \tag{4.16}$$

We have an analog of the Gauss-Markov Theorem.

**Theorem 4.6. Generalized Gauss-Markov**

In the linear regression model (Assumption 4.2) and $\boldsymbol{\Omega} > 0$, if $\widetilde{\overrightarrow{\beta}}$ is a linear unbiased estimator of $\beta$ then

$$\operatorname{var}\left[\widetilde{\overrightarrow{\beta}} \mid \boldsymbol{X}\right] \geq \left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}.$$

The theorem provides a lower bound on the covariance matrix of unbiased linear estimators. *The bound is different from the variance matrix of the OLS estimator as stated in (4.16) except when $\mathbf{\Omega} = \boldsymbol{I}_n\sigma^2$. The fact that the variance bound is different (and lower) than the least squares variance suggests that we can improve on the OLS estimator.* In the i.i.d. sampling case the variance lower bound is $\left(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1}$ since $\mathbf{\Omega} = \boldsymbol{D}$.

This is indeed the case when $\mathbf{\Omega}$ is known up to scale. That is, suppose that $\mathbf{\Omega} = c^2\mathbf{\Sigma}$ where $c^2 > 0$ is real and $\mathbf{\Sigma}$ is $n \times n$ and known. Take the linear model (4.12) and pre-multiply by $\mathbf{\Sigma}^{-1/2}$. This produces the equation

$$\overrightarrow{\widetilde{Y}} = \widetilde{\boldsymbol{X}}\overrightarrow{\beta} + \overrightarrow{\widetilde{e}}$$

where

$$\overrightarrow{\widetilde{Y}} = \Sigma^{-1/2}\overrightarrow{Y}, \widetilde{\boldsymbol{X}} = \Sigma^{-1/2}\boldsymbol{X}, \overrightarrow{\widetilde{e}} = \Sigma^{-1/2}\overrightarrow{e}.$$

Consider OLS estimation of $\overrightarrow{\beta}$ in this equation.

$$\begin{aligned} \widetilde{\beta}_{\text{gls}} &= \left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\right)^{-1}\widetilde{\boldsymbol{X}}'\overrightarrow{\widetilde{Y}} \\ &= \left(\left(\Sigma^{-1/2}\boldsymbol{X}\right)'\left(\Sigma^{-1/2}\boldsymbol{X}\right)\right)^{-1}\left(\Sigma^{-1/2}\boldsymbol{X}\right)'\left(\Sigma^{-1/2}\overrightarrow{Y}\right) \qquad (4.17) \\ &= \left(\boldsymbol{X}'\Sigma^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\Sigma^{-1}\overrightarrow{Y} \end{aligned}$$

This is called the *Generalized Least Squares (GLS) estimator* of $\overrightarrow{\beta}$.

It is easy to calculate that

$$\mathbb{E}\left[\overrightarrow{\widetilde{\beta}}_{\text{gls}} \mid \boldsymbol{X}\right] = \overrightarrow{\beta}, \tag{4.18}$$

$$\operatorname{var}\left[\overrightarrow{\widetilde{\beta}}_{\text{gls}} \mid \boldsymbol{X}\right] = \left(\boldsymbol{X}'\mathbf{\Omega}^{-1}\boldsymbol{X}\right)^{-1}. \tag{4.19}$$

This shows that the GLS estimator is unbiased and has a covariance matrix which equals the lower bound from Theorem 4.6. This shows that the lower bound is sharp when $\mathbf{\Sigma}$ is known. GLS is thus efficient in the class of linear unbiased estimators.

In the linear regression model with independent observations and known conditional variances, so that $\mathbf{\Omega} = \mathbf{\Sigma} = \boldsymbol{D} = \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_n^2\right)$, the GLS estimator takes the form

$$\begin{aligned} \overrightarrow{\widetilde{\beta}}_{\text{gls}} &= \left(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{D}^{-1}\overrightarrow{Y} \\ &= \left(\sum_{i=1}^{n}\sigma_i^{-2}X_iX_i'\right)^{-1}\left(\sum_{i=1}^{n}\sigma_i^{-2}X_iY_i\right). \end{aligned}$$

The assumption $\mathbf{\Omega} > 0$ in this case reduces to $\sigma_i^2 > 0$ for $i = 1, \ldots, n$.

In practice, the covariance matrix $\mathbf{\Omega}$ is unknown so the GLS estimator as presented here is not feasible. However, the form of the GLS estimator motivates feasible versions, effectively by replacing $\boldsymbol{\Omega}$ with an estimator. We do not pursue this here as it is not common in current applied econometric practice.

## 4.10. Modern Generalized Gauss Markov Theorem

**Theorem 4.7. Modern Generalized Gauss-Markov**

In the linear regression model with i.i.d. sampling, if $\mathbb{E}\left[\widetilde{\vec{\beta}} \mid \boldsymbol{X}\right] = \beta$ then $\operatorname{var}\left[\widetilde{\vec{\beta}} \mid \boldsymbol{X}\right] \geq \left(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1}$.

## 4.11. Residuals

What are some properties of the residuals $\widehat{e}_i = Y_i - \vec{X}_i'\widehat{\vec{\beta}}$ and prediction errors $\widetilde{e}_i = Y_i - \vec{X}_i'\widehat{\vec{\beta}}_{(-i)}$ in the context of the linear regression model?

Recall from (3.24) that we can write the residuals in vector notation as $\widehat{\vec{e}} = \boldsymbol{M}\vec{e}$ where $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)\boldsymbol{X}'$ is the orthogonal projection matrix. Using the properties of conditional expectation

$$\mathbb{E}\left[\widehat{\vec{e}} \mid \boldsymbol{X}\right] = \mathbb{E}\left[\boldsymbol{M}\vec{e} \mid \boldsymbol{X}\right] = \boldsymbol{M}\,\mathbb{E}\left[\vec{e} \mid \boldsymbol{X}\right] = 0$$

and

$$\operatorname{var}\left[\widehat{\vec{e}} \mid \boldsymbol{X}\right] = \operatorname{var}\left[\boldsymbol{M}\vec{e} \mid \boldsymbol{X}\right] = \boldsymbol{M}\operatorname{var}\left[\vec{e} \mid \boldsymbol{X}\right]\boldsymbol{M} = \boldsymbol{M}\boldsymbol{D}\boldsymbol{M} \tag{4.20}$$

where $\boldsymbol{D}$ is defined in (4.8) under i.i.d. sampling.

We can simplify this expression under the assumption of conditional homoskedasticity

$$\mathbb{E}\left[e^2 \mid \vec{X}\right] = \sigma^2.$$

In this case, (4.20) simplifies to

$$\operatorname{var}\left[\widehat{\vec{e}} \mid \boldsymbol{X}\right] = \boldsymbol{M}\sigma^2. \tag{4.21}$$

In particular, for a single observation $i$ we can find the variance of $\widehat{e}_i$ by taking the $i$th diagonal element of (4.21). Since the $i$th element of $\boldsymbol{M}$ is $1 - h_{ii}$, we obtain,

$$\operatorname{var}\left[\widehat{e}_i \mid \boldsymbol{X}\right] = \mathbb{E}\left[\widehat{e}_i^2 \mid \boldsymbol{X}\right] = (1 - h_{ii})\sigma^2. \tag{4.22}$$

Similarly, the prediction errors $\widetilde{e}_i = (1 - h_{ii})^{-1}\widehat{e}_i$ can be written in vector notation as $\widetilde{\vec{e}} = \boldsymbol{M}^*\widehat{\vec{e}}$ where $\boldsymbol{M}^*$ is a diagonal matrix with $i$th element $(1 - h_{ii})^{-1}$. Thus, $\widetilde{\vec{e}} = \boldsymbol{M}^*\boldsymbol{M}\vec{e}$. We can calculate that

$$\mathbb{E}\left[\widetilde{\vec{e}} \mid \boldsymbol{X}\right] = \boldsymbol{M}^{*}\boldsymbol{M}\,\mathbb{E}\left[\vec{e} \mid \boldsymbol{X}\right] = \vec{0}$$

$$\operatorname{var}\left[\widetilde{\vec{e}} \mid \boldsymbol{X}\right] = \boldsymbol{M}^{*}\boldsymbol{M}\operatorname{var}\left[\vec{e} \mid \boldsymbol{X}\right]\boldsymbol{M}\boldsymbol{M}^{*} = \boldsymbol{M}^{*}\boldsymbol{M}\boldsymbol{D}\boldsymbol{M}\boldsymbol{M}^{*}$$

which simplifies under homoskedasticity to

$$\operatorname{var}\left[\widetilde{\vec{e}} \mid \boldsymbol{X}\right] = \boldsymbol{M}^{*}\boldsymbol{M}\boldsymbol{M}\boldsymbol{M}^{*}\sigma^{2} = \boldsymbol{M}^{*}\boldsymbol{M}\boldsymbol{M}^{*}\sigma^{2}.$$

The variance of $i$th prediction error is then

$$\operatorname{var}\left[\widetilde{e}_i \mid \boldsymbol{X}\right] = (1-h_{ii})^{-1}\sigma^{2}.$$

A residual with constant conditional variance can be obtained by rescaling. The *standardized residuals* are

$$\bar{e}_i = (1-h_{ii})^{-1/2}\widehat{e}_i, \tag{4.23}$$

and in vector notation

$$\overline{\vec{e}} = \left(\bar{e}_1,\ldots,\bar{e}_n\right)' = \boldsymbol{M}^{*1/2}\boldsymbol{M}\vec{e}. \tag{4.24}$$

From the above calculations, under homoskedasticity,

$$\operatorname{var}\left[\overline{\vec{e}} \mid \boldsymbol{X}\right] = \boldsymbol{M}^{*1/2}\boldsymbol{M}\boldsymbol{M}^{*1/2}\sigma^{2}$$

and

$$\operatorname{var}\left[\bar{e}_i \mid \boldsymbol{X}\right] = \sigma^{2}$$

and thus *these standardized residuals have the same bias and variance as the original errors when the latter are homoskedastic*.

## 4.12. Estimation of Error Variance

The error variance $\sigma^2 = \mathbb{E}\left[e^2\right]$ can be a parameter of interest even in a heteroskedastic regression or a projection model. $\sigma^2$ measures the variation in the “unexplained” part of the regression. Its method of moments estimator (MME) is the sample average of the squared residuals:

$$\widehat{\sigma}^{2} = \frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^{2}.$$

In the linear regression model we can calculate the mean of $\widehat{\sigma}^2$:

$$\widehat{\sigma}^{2} = \frac{1}{n}\vec{e}\,'\boldsymbol{M}\vec{e} = \frac{1}{n}\operatorname{tr}\left(\vec{e}\,'\boldsymbol{M}\vec{e}\right) = \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\vec{e}\,\vec{e}\,'\right).$$

Note that the first equality $\widehat{\sigma}^2 = \frac{1}{n}\widehat{\vec{e}}'\widehat{\vec{e}} = \frac{1}{n}\vec{e}'\boldsymbol{M}\vec{e}$ is due to $\widehat{\vec{e}} = \boldsymbol{M}\vec{e}$. Then

$$\begin{aligned}
\mathbb{E}\left[\widehat{\sigma}^2 \mid \boldsymbol{X}\right] &= \frac{1}{n}\operatorname{tr}\left(\mathbb{E}\left[\boldsymbol{M}\vec{e}\vec{e}' \mid \boldsymbol{X}\right]\right) \\
&= \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\mathbb{E}\left[\vec{e}\vec{e}' \mid \boldsymbol{X}\right]\right) \\
&= \frac{1}{n}\operatorname{tr}(\boldsymbol{M}\boldsymbol{D}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(1 - h_{ii}\right)\sigma_i^2.
\end{aligned} \tag{4.25}$$

The final equality holds since the trace is the sum of the diagonal elements of $\boldsymbol{MD}$, and since $\boldsymbol{D}$ is diagonal the diagonal elements of $\boldsymbol{MD}$ are the product of the diagonal elements of $\boldsymbol{M}$ and $\boldsymbol{D}$, which are $1 - h_{ii}$, and $\sigma_i^2$, respectively.

Adding the assumption of conditional homoskedasticity $\mathbb{E}\left[e^2 \mid \boldsymbol{X}\right] = \sigma^2$ so that $\boldsymbol{D} = \boldsymbol{I}_n\sigma^2$, then (4.25) simplified to

$$\mathbb{E}\left[\widehat{\sigma}^2 \mid \boldsymbol{X}\right] = \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\sigma^2\right) = \sigma^2\left(\frac{n-k}{n}\right).$$

Note that the above derivation uses the property of leverage values $\sum_i h_{ii} = k$. *This calculation shows that $\widehat{\sigma}^2$ is biased towards zero under the homoskedastic case.* The order of the bias depends on $k/n$, ratio of the number of estimated coefficients to the sample size.

Another way to see this is to use (4.22). Note that

$$\mathbb{E}\left[\widehat{\sigma}^2 \mid \boldsymbol{X}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\widehat{e}_i^2 \mid \boldsymbol{X}\right] = \frac{1}{n}\sum_{i=1}^{n}(1 - h_{ii})\sigma^2 = \left(\frac{n-k}{n}\right)\sigma^2.$$

Since the bias takes a scale from a classic method to obtain an unbiased estimator is by rescaling. Define

$$s^2 = \frac{1}{n-k}\sum_{i=1}^{n}\widehat{e}_i^2. \tag{4.26}$$

By the above calculation $\mathbb{E}\left[s^2 \mid \boldsymbol{X}\right] = \sigma^2$ and $\mathbb{E}\left[s^2\right] = \sigma^2$. Hence the estimator $s^2$ is unbiased for $\sigma^2$. Consequently, $s^2$ is known as the *bias-corrected estimator for $\sigma^2$* and in empirical practice $s^2$ is the most widely used estimator for $\sigma^2$.

Interestingly, this is not the only method to construct an unbiased estimator for $\sigma^2$. An estimator constructed with the standardized residuals $\bar{e}_i$ from (4.23) is

$$\bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\bar{e}_i^2 = \frac{1}{n}\sum_{i=1}^{n}(1 - h_{ii})^{-1}\widehat{e}_i^2. \tag{4.27}$$

You can show that $\bar{\sigma}^2$ is unbiased for $\sigma^2$ (in the homoskedastic linear regression model).

## 4.13. Mean-Square Forecast Error

One use of an estimated regression is to predict out-of-sample. Consider an out-of-sample realization $(Y_{n+1}, \vec{X}_{n+1})$ where $\vec{X}_{n+1}$ is observed but not $Y_{n+1}$. Then the standard point estimator is $\widetilde{Y}_{n+1} = \vec{X}'_{n+1}\widetilde{\vec{\beta}}$. The forecast error is the difference between the actual value $Y_{n+1}$ and the point forecast $\widetilde{Y}_{n+1}$. This is the forecast error $\widetilde{e}_{n+1} = Y_{n+1} - \widetilde{Y}_{n+1}$. The *mean-squared forecast error (MSFE)* is the expected value $\mathrm{MSFE}_n = \mathbb{E}\left[\widetilde{e}^2_{n+1}\right]$. In the linear regression model, $\widetilde{e}_{n+1} = e_{n+1} - \vec{X}'_{n+1}\left(\widehat{\vec{\beta}} - \vec{\beta}\right)$ so

$$\mathrm{MSFE}_n = \mathbb{E}\left[e^2_{n+1}\right] - 2\,\mathbb{E}\left[e_{n+1}\vec{X}'_{n+1}\left(\widehat{\vec{\beta}} - \vec{\beta}\right)\right] + \mathbb{E}\left[\vec{X}'_{n+1}\left(\widehat{\vec{\beta}} - \vec{\beta}\right)\left(\widehat{\vec{\beta}} - \vec{\beta}\right)'\vec{X}_{n+1}\right]. \tag{4.28}$$

The first term in (4.28) is $\sigma^2$. The second term in (4.28) is zero since $e_{n+1}\vec{X}'_{n+1}$ is independent of $\widehat{\vec{\beta}} - \vec{\beta}$ and both are mean zero. Using the properties of the trace operator the third term in (4.28) is

$$\begin{aligned}
&\mathrm{tr}\left(\mathbb{E}\left[\vec{X}_{n+1}\vec{X}'_{n+1}\right]\mathbb{E}\left[\left(\widehat{\vec{\beta}} - \vec{\beta}\right)\left(\widehat{\vec{\beta}} - \vec{\beta}\right)'\right]\right)\\
&= \mathrm{tr}\left(\mathbb{E}\left[\vec{X}_{n+1}\vec{X}'_{n+1}\right]\mathbb{E}\left[\mathbb{E}\left(\left(\widehat{\vec{\beta}} - \vec{\beta}\right)\left(\widehat{\vec{\beta}} - \vec{\beta}\right)'\right) \mid \boldsymbol{X}\right]\right)\\
&= \mathrm{tr}\left(\mathbb{E}\left[\vec{X}_{n+1}\vec{X}'_{n+1}\right]\mathbb{E}\left[\boldsymbol{V}_{\widehat{\vec{\beta}}}\right]\right) \qquad (4.29)\\
&= \mathbb{E}\left[\mathrm{tr}\left(\left(\vec{X}_{n+1}\vec{X}'_{n+1}\right)\boldsymbol{V}_{\widehat{\vec{\beta}}}\right)\right]\\
&= \mathbb{E}\left[\vec{X}'_{n+1}\boldsymbol{V}_{\widehat{\vec{\beta}}}\vec{X}_{n+1}\right]
\end{aligned}$$

where we use the fact that $\vec{X}_{n+1}$ is independent of $\widehat{\vec{\beta}}$, the definition $\boldsymbol{V}_{\widehat{\vec{\beta}}} = \mathbb{E}\left[\left(\widehat{\vec{\beta}} - \vec{\beta}\right)\left(\widehat{\vec{\beta}} - \vec{\beta}\right)' \mid \boldsymbol{X}\right]$, and the fact that $\vec{X}_{n+1}$ is independent of $\boldsymbol{V}_{\widehat{\vec{\beta}}}$. Thus,

$$\mathrm{MSFE}_n = \sigma^2 + \mathbb{E}\left[\vec{X}'_{n+1}\boldsymbol{V}_{\widehat{\vec{\beta}}}\vec{X}_{n+1}\right].$$

Under conditional homoskedasticity this simplifies to

$$\mathrm{MSFE}_n = \sigma^2\left(1 + \mathbb{E}\left[\vec{X}'_{n+1}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\vec{X}_{n+1}\right]\right).$$

A simple estimator for the MSFE is obtained by averaging the squared prediction errors (3.46)

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\widetilde{e}^2_i,$$

where $\widetilde{e}_i = Y_i - \vec{X}_i' \widehat{\vec{\beta}}_{(-i)} = \widehat{e}_i (1 - h_{ii})^{-1}$. Indeed, we can calculate that

$$\begin{aligned}
\mathbb{E}\left[\widetilde{\sigma}^2\right] &= \mathbb{E}\left[\widetilde{e}_i^2\right] \\
&= \mathbb{E}\left[\left(e_i - \vec{X}_i'\left(\widehat{\vec{\beta}}_{(-i)} - \vec{\beta}\right)\right)^2\right] \\
&= \sigma^2 + \mathbb{E}\left[\vec{X}_i'\left(\widehat{\vec{\beta}}_{(-i)} - \vec{\beta}\right)\left(\widehat{\vec{\beta}}_{(-i)} - \vec{\beta}\right)'\vec{X}_i\right].
\end{aligned}$$

By a similar calculation as in (4.29), we find

$$\mathbb{E}\left[\widetilde{\sigma}^2\right] = \sigma^2 + \mathbb{E}\left[\vec{X}_i' \boldsymbol{V}_{\widehat{\vec{\beta}}_{(-i)}} \vec{X}_i\right] = \mathrm{MSFE}_{n-1}\,.$$

This is the MSFE based on a sample of size $n-1$ rather than size $n$. The difference arises because the in-sample prediction errors $\widetilde{e}_i$ for $i \le n$ are calculated using an effective sample size of $n-1$, while the out-of sample prediction error $\widetilde{e}_{n+1}$ is calculated from a sample with the full $n$ observations. Unless $n$ is very small, we should expect $\mathrm{MSFE}_{n-1}$ to be close to $\mathrm{MSFE}_n$. Thus, $\widetilde{\sigma}^2$ is a reasonable estimator for $\mathrm{MSFE}_n$.

**Theorem 4.8. MSFE**

In the linear regression model (Assumption 4.2) and iid sampling (Assumption 4.1),

$$\mathrm{MSFE}_n = \sigma^2 + \mathbb{E}\left[\vec{X}_{n+1}' \boldsymbol{V}_{\widehat{\vec{\beta}}} \vec{X}_{n+1}\right],$$

where $\boldsymbol{V}_{\widehat{\vec{\beta}}} = \operatorname{var}\left[\widehat{\vec{\beta}} \mid \boldsymbol{X}\right]$. Furthermore, $\widetilde{\sigma}^2$ defined in (3.46) is an unbiased estimator of $\mathrm{MSFE}_{n-1}$, since $\mathbb{E}\left[\widetilde{\sigma}^2\right] = \mathrm{MSFE}_{n-1}$.

## 4.14. Covariance Matrix Estimation Under Homoskedasticity

For inference we need an estimator of the covariance matrix $\boldsymbol{V}_{\widehat{\vec{\beta}}}$ of the least squares estimator. In this section we consider the homoskedastic regression model (Assumption 4.3).

Under homoskedasticity the covariance matrix takes the simple form

$$\boldsymbol{V}^0_{\widehat{\vec{\beta}}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2$$

which is known up to the scale $\sigma^2$. In Section 4.12 we discussed three estimators of $\sigma^2$. The most common used choice is $s^2$ leading to the classic covariance matrix estimator

$$\widehat{\boldsymbol{V}}^0_{\widehat{\vec{\beta}}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} s^2. \tag{4.30}$$

Since $s^2$ is conditionally unbiased for $\sigma^2$ it is simple to calculate that $\widehat{\boldsymbol{V}}^0_{\hat{\vec{\beta}}}$ is conditionally unbiased for $\boldsymbol{V}^0_{\hat{\vec{\beta}}}$ under the assumption of homoskedasticity:

$$\mathbb{E}\left(\widehat{\boldsymbol{V}}^0_{\hat{\vec{\beta}}} \mid \boldsymbol{X}\right) = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \mathbb{E}\left(s^2 \mid \boldsymbol{X}\right) = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \sigma^2 = \boldsymbol{V}_{\hat{\vec{\beta}}}.$$

*This was the dominant covariance matrix estimator in applied econometrics for many years and is still the default method in most regression packages.* For example, Stata uses the covariance matrix estimator (4.30) by default in linear regression unless an alternative is specified.

*If the estimator (4.30) is used but the regression error is heteroskedastic it is possible for $\widehat{\boldsymbol{V}}^0_{\hat{\vec{\beta}}}$ to be quite biased for the correct covariance matrix $\boldsymbol{V}_{\hat{\vec{\beta}}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$.*

For example, suppose $k = 1$ and $\sigma_i^2 = X_i^2$ with $\mathbb{E}[X] = 0$. The ratio of the true variance of the least squares estimator to the expectation of the variance estimator is

$$\frac{\boldsymbol{V}_{\hat{\vec{\beta}}}}{\mathbb{E}\left[\boldsymbol{V}_{\hat{\vec{\beta}}^0} \mid \boldsymbol{X}\right]} = \frac{\sum_{i=1}^n X_i^4}{\sigma^2 \sum_{i=1}^n X_i^2} \simeq \frac{\mathbb{E}\left[X^4\right]}{\left(\mathbb{E}\left[X^2\right]^2\right)} \coloneqq \kappa.$$

Notice that we use the fact that $\sigma_i^2 = X_i^2$ implies $\sigma^2 = \mathbb{E}\left[\sigma_i^2\right] = \mathbb{E}\left[X^2\right]$. The constant $\kappa$ is the standardized fourth moment (or kurtosis) of the regressor $X$ and can be any number greater than one. For example, if $X \sim N\left(0, \sigma^2\right)$, then $\kappa = 3$, so the true variance $\boldsymbol{V}_{\hat{\vec{\beta}}}$ is three times larger than the expected homoskedastic estimator $\boldsymbol{V}_{\hat{\vec{\beta}}^0}$.

*I didn't quite understand the calculation of the above example.*

## 4.15. Covariance Matrix Estimation Under Heteroskedasticity

Recall that the general form for the covariance matrix is

$$\boldsymbol{V}_{\hat{\vec{\beta}}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1},$$

with $\boldsymbol{D}$ defined in (4.8). This depends on the unknown matrix $\boldsymbol{D}$ which we can write as

$$\boldsymbol{D} = \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_n^2\right) = \mathbb{E}\left[\vec{e}\,\vec{e}\,' \mid \boldsymbol{X}\right] = \mathbb{E}\left[\widetilde{\boldsymbol{D}} \mid \boldsymbol{X}\right]$$

where $\widetilde{\boldsymbol{D}} = \operatorname{diag}\left(e_1^2, \ldots, e_n^2\right)$. Thus $\widetilde{\boldsymbol{D}}$ is a conditionally unbiased estimator for $\boldsymbol{D}$. If the squared errors $e_i^2$ were observable, we could construct an unbiased estimator for $\boldsymbol{V}_{\widehat{\vec{\beta}}}$ as

$$\begin{aligned}\boldsymbol{V}_{\widehat{\vec{\beta}}}^{\text{ideal}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\widetilde{\boldsymbol{D}}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\vec{X}_i\vec{X}_i'e_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.\end{aligned}$$

Indeed,

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{V}_{\widehat{\vec{\beta}}}^{\text{ideal}}\right] &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}X_iX_i'\,\mathbb{E}\left[e_i^2 \mid \boldsymbol{X}\right]\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\vec{X}_i\vec{X}_i'\sigma_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} = \boldsymbol{V}_{\widehat{\vec{\beta}}},\end{aligned}$$

verifying that $\boldsymbol{V}_{\widehat{\vec{\beta}}}^{\text{ideal}}$ is unbiased for $\boldsymbol{V}_{\widehat{\vec{\beta}}}$.

Since the errors $e_i^2$ were unobservable, $\boldsymbol{V}_{\widehat{\vec{\beta}}}^{\text{ideal}}$ is not a feasible estimator. However, we can replace $e_i^2$ with the squared residuals $\widehat{e}_i^2$. Making this substitution we obtain the estimator

$$\boldsymbol{V}_{\widehat{\vec{\beta}}}^{\text{HC0}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\vec{X}_i\vec{X}_i'\widehat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{4.31}$$

The label "HC" refers to "heteroskedasticity-consistent". The label "HC0" refers to this being the *baseline heteroskedasticity-consistent covariance matrix estimator*.

We know, however, that $\widehat{e}_i^2$ is biased towards zero (recall equation (4.22)). To estimate the variance $\sigma^2$ scales the moment estimator $\widehat{\sigma}^2$ by $n/(n-k)$. Making the same adjustment we obtain the estimator

$$\boldsymbol{V}_{\widehat{\vec{\beta}}}^{\text{HC1}} = \left(\frac{n}{n-k}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\vec{X}_i\vec{X}_i'\widehat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{4.32}$$

While the scaling by $n/(n-k)$ is ad hoc, HC1 is often recommended over the unscaled HC0 estimator.

Alternatively, we could use the standardized residuals $\overline{e}_i$ or the prediction errors $\widetilde{e}_i$ , yielding the HC2 and HC3 estimators

$$\begin{aligned}\boldsymbol{V}_{\widehat{\vec{\beta}}}^{\text{HC2}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\vec{X}_i\vec{X}_i'\overline{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\left(1-h_{ii}\right)^{-1}\vec{X}_i\vec{X}_i'\widehat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\end{aligned} \tag{4.33}$$

and

$$\begin{aligned} \boldsymbol{V}_{\widehat{\vec{\beta}}}^{\mathrm{HC3}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\sum_{i=1}^{n} \vec{X}_i \vec{X}_i' \widetilde{e}_i^2\right) \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\sum_{i=1}^{n} \left(1-h_{ii}\right)^{-2} \vec{X}_i \vec{X}_i' \widehat{e}_i^2\right) \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \end{aligned} \tag{4.34}$$

The four estimators HC0, HC1, HC2 and HC3 are collectively called *robust, heteroskedasticity-consistent, or heteroskedasticity-robust* covariance matrix estimators. *The degree-of-freedom adjustment in HC1 is the default robust covariance matrix estimator implemented in Stata.*

Since $\left(1-h_{ii}\right)^{-2} > \left(1-h_{ii}\right)^{-1} > 1$, it is straightforward to show that

$$\boldsymbol{V}_{\widehat{\vec{\beta}}}^{\mathrm{HC0}} < \boldsymbol{V}_{\widehat{\vec{\beta}}}^{\mathrm{HC2}} < \boldsymbol{V}_{\widehat{\vec{\beta}}}^{\mathrm{HC3}}. \tag{4.35}$$

The inequality $\boldsymbol{A} < \boldsymbol{B}$ when applied to matrices means that the matrix $\boldsymbol{B} - \boldsymbol{A}$ is positive definite.

In general, the bias of the covariance matrix estimators is complicated but simplify under the assumption of homoskedasticity (Assumption 4.3). For example, using (4.22),

$$\begin{aligned} \mathbb{E}\left[\widehat{\boldsymbol{V}}_{\widehat{\vec{\beta}}}^{\mathrm{HC0}} \mid \boldsymbol{X}\right] &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\sum_{i=1}^{n} \vec{X}_i \vec{X}_i' \mathbb{E}\left[\widehat{e}_i^2 \mid \boldsymbol{X}\right]\right) \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\sum_{i=1}^{n} \vec{X}_i \vec{X}_i' \left(1-h_{ii}\right)\sigma^2\right) \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\sum_{i=1}^{n} \vec{X}_i \vec{X}_i' h_{ii}\right) \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 \\ &< \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 = \boldsymbol{V}_{\widehat{\vec{\beta}}}. \end{aligned}$$

This calculation shows that $\boldsymbol{V}_{\widehat{\vec{\beta}}}^{\mathrm{HC0}}$ is biased towards zero under homoskedasticity.

By a similar calculation (again under homoskedasticity) we can calculate that the HC2 estimator is unbiased

$$\mathbb{E}\left[\boldsymbol{V}_{\widehat{\vec{\beta}}}^{\mathrm{HC2}} \mid \boldsymbol{X}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2. \tag{4.36}$$

*It might seem rather odd to compare the bias of heteroskedasticity-robust estimators under the assumption of homoskedasticity but it does give us a baseline for comparison.*

Another interesting calculation shows that *in general (that is, without assuming homoskedasticity) the HC3 estimator is biased away from zero*. Indeed, using the definition of the prediction errors (3.44)

$$\widehat{e}_i = Y_i - \vec{X}_i' \widehat{\vec{\beta}}_{-i} = e_i - \vec{X}_i' \left(\widehat{\vec{\beta}}_{-i} - \vec{\beta}\right)$$

so

$$\widehat{e}_i^2 = e_i^2 - 2\overrightarrow{X}_i'\left(\widehat{\overrightarrow{\beta}}_{-i} - \overrightarrow{\beta}\right)e_i + \left(\overrightarrow{X}_i'\left(\widehat{\overrightarrow{\beta}}_{-i} - \overrightarrow{\beta}\right)\right)^2.$$

Note that $e_i$ and $\widehat{\overrightarrow{\beta}}_{-i}$ are functions of non-overlapping observations and are thus independent. Hence $\mathbb{E}\left[\left(\widehat{\overrightarrow{\beta}}_{-i} - \overrightarrow{\beta}\right)e_i \mid \boldsymbol{X}\right] = 0$ and

$$\mathbb{E}\left[\widehat{e}_i^2 \mid \boldsymbol{X}\right] = \sigma_i^2 + \mathbb{E}\left[\left(\overrightarrow{X}_i'\left(\widehat{\overrightarrow{\beta}}_{-i} - \overrightarrow{\beta}\right)\right)^2 \mid \boldsymbol{X}\right] \geq \sigma_i^2.$$

If follows that

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{V}_{\widehat{\overrightarrow{\beta}}}^{\mathrm{HC3}} \mid \boldsymbol{X}\right] &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\left(1-h_{ii}\right)^{-2}\overrightarrow{X}_i\overrightarrow{X}_i'\,\mathbb{E}\left[\widehat{e}_i^2\right] \mid \boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &\geq \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\overrightarrow{X}_i\overrightarrow{X}_i'\sigma_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} = \boldsymbol{V}_{\widehat{\overrightarrow{\beta}}}.\end{aligned}$$

This means that the HC3 estimator is conservative in the sense that it is weakly larger (in expectation) than the correct variance for any realization of $\boldsymbol{X}$.

## 4.16. Standard Errors

**Definition 4.2.**

A *standard error* $s\left(\widehat{\beta}\right)$ for a real-valued estimator $\widehat{\beta}$ is an estimator of the standard deviation of the distribution of $\widehat{\beta}$.

When $\overrightarrow{\beta}$ is a vector with estimator $\widehat{\overrightarrow{\beta}}$ and covariance matrix estimator $\widehat{\boldsymbol{V}}_{\widehat{\overrightarrow{\beta}}}$, standard errors for individual elements are the square roots of the diagonal elements of $\widehat{\boldsymbol{V}}_{\widehat{\overrightarrow{\beta}}}$. That is,

$$s\left(\widehat{\beta}_j\right) = \sqrt{\widehat{\boldsymbol{V}}_{\widehat{\overrightarrow{\beta}}_j}} = \sqrt{\left[\widehat{\boldsymbol{V}}_{\widehat{\overrightarrow{\beta}}}\right]_{jj}}.$$

When the classical covariance matrix estimator (4.30) is used the standard error takes the simple form

$$s\left(\widehat{\beta}_j\right) = s\sqrt{\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]_{jj}}. \tag{4.37}$$

## 4.17. Estimation with Sparse Dummy Variables

The heteroskedasticity-robust covariance matrix estimators can be quite imprecise in some contexts. One is in the presence of *sparse dummy variables* – when a dummy variable only takes the value 1 or 0 for very few observations. In these contexts one component of the covariance matrix is estimated on just those few observations and will be imprecise. This is effectively hidden from the user.

To see the problem, let $D$ be a dummy variable and consider the dummy variable regression

$$Y = \beta_1 D + \beta_2 + e. \tag{4.40}$$

The number of observations for which $D_i = 1$ is $n_1 = \sum_{i=1}^n D_i$. The number of observations for which $D_i = 0$ is $n_2 = n - n_1$. We say the design is *sparse* if $n_1$ or $n_2$ is small.

To simplify our analysis, we take the extreme case $n_1 = 1$. The ideas extend to the case of $n_1 > 1$ but small, though with less dramatic effects.

In the regression model (4.40), we can calculate that the true covariance matrix of the least squares estimator for the coefficients under the simplifying assumption of conditional homoskedasticity is

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} = \sigma^2 \begin{pmatrix} 1 & 1 \\ 1 & n \end{pmatrix}^{-1} = \frac{\sigma^2}{n-1} \begin{pmatrix} n & -1 \\ -1 & 1 \end{pmatrix}$$

In particular, the variance of the estimator for the coefficient on the dummy variable is

$$V_{\widehat{\beta}_1} = \frac{\sigma^2 \cdot n}{n-1}.$$

Essentially, the coefficient $\widehat{\beta}_1$ is estimated from a single observation so its variance is roughly unaffected by sample size. *An important message is that certain coefficient estimators in the presence of sparse dummy variables will be imprecise, regardless of the sample size. A large sample alone is not sufficient to ensure precise estimation.*

Now lets examine the standard HC1 covariance matrix estimator (4.32). The regression has perfect fit for the observation for which $D_i = 1$ so the corresponding residual is $\widehat{e}_i = 0$. It follows that $D_i \widehat{e}_i = 0$ for all $i$ (either $D_i = 0$ or $\widehat{e}_i = 0$). Hence

$$\sum_{i=1}^n X_i X_i' \widehat{e}_i^2 = \begin{pmatrix} 0 & 0 \\ 0 & \sum_{i=1}^n \widehat{e}_i^2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & (n-2)s^2 \end{pmatrix}.$$

Together we find that

$$\begin{aligned} \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{HC1}} &= \left(\frac{n}{n-2}\right) \frac{1}{(n-1)^2} \begin{pmatrix} n & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & (n-2)s^2 \end{pmatrix} \begin{pmatrix} n & -1 \\ -1 & 1 \end{pmatrix} \\ &= s^2 \frac{n}{(n-1)^2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \end{aligned}$$

In particular, the estimator for $V_{\widehat{\beta}_1}$ is

$$\widehat{V}^{\mathrm{HC1}}_{\widehat{\beta}_1} = s^2 \frac{n}{(n-1)^2}.$$

It has expectation

$$\mathbb{E}\left[\widehat{V}^{\mathrm{HC1}}_{\widehat{\beta}_1}\right] = \frac{\sigma^2 \cdot n}{(n-1)^2} = \frac{V_{\widehat{\beta}_1}}{n-1} \ll V_{\widehat{\beta}_1}.$$

The variance $\widehat{V}^{\mathrm{HC1}}_{\widehat{\beta}_1}$ is extremely biased for the true variance $V_{\widehat{\beta}_1}$. The reported variance – and standard error – is misleadingly small. The variance estimate erroneously mis-states the precision of $\widehat{\beta}_1$.

The fact that $\widehat{V}^{\mathrm{HC1}}_{\widehat{\beta}_1}$ is biased is unlikely to be noticed by an applied researcher. Nothing in the reported output will alert a researcher to the problem.

One insight is to examine the the leverage values. The single observation observation with $D_i = 1$ has

$$h_{ii} = \frac{1}{n-1} \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} n & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 1$$

This is an extreme leverage value.

A possible solution is to replace the biased covariance matrix estimator $\widehat{V}^{\mathrm{HC1}}_{\widehat{\beta}_1}$ with $\widehat{V}^{\mathrm{HC2}}_{\widehat{\beta}_1}$ or $\widehat{V}^{\mathrm{HC3}}_{\widehat{\beta}_1}$. Neither approach can be done in the sparse case $n_1 = 1$, since they cannot be calculated.

It is unclear if there is a best practice to avoid this situation. Once possibility is to calculate the maximum leverage value. If it is very large calculate the standard errors using several methods to see if variation occurs.

## 4.20. Measures of Fit

As we described in the previous chapter a commonly reported measure of regression fit is the regression $R^2$ defined as

$$R^2 = 1 - \frac{\sum_{i=1}^n \widehat{e}_i^2}{\sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2} = 1 - \frac{\widehat{\sigma}^2}{\widehat{\sigma}_Y^2},$$

where $\widehat{\sigma}_Y^2 = n^{-1} \sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2$. $R^2$ is an estimator of the population parameter,

$$\rho^2 = \frac{\operatorname{var}[\overrightarrow{X}' \overrightarrow{\beta}]}{\operatorname{var}[Y]} = 1 - \frac{\sigma^2}{\sigma_Y^2}.$$

However, $\widehat{\sigma}^2$ and $\widehat{\sigma}_Y^2$ are biased. Theil (1961) proposed replacing these by the unbiased versions $s^2$ and

$$\widetilde{\sigma}_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$$

yielding what is known as *R-bar-squared* or *adjusted R-squared*

$$\overline{R}^2 = 1 - \frac{s^2}{\widetilde{\sigma}_Y^2} = 1 - \frac{(n-k)^{-1}\sum_{i=1}^{n}\widehat{e}_i^2}{(n-1)^{-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}.$$

While $\overline{R}^2$ is an improvement on $R^2$, a much better improvement is

$$\widetilde{R}^2 = 1 - \frac{\sum_{i=1}^{n}\widetilde{e}_i^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = 1 - \frac{\widetilde{\sigma}^2}{\widetilde{\sigma}_Y^2}$$

where $\widetilde{e}_i$ are the prediction errors (3.44) and $\widetilde{\sigma}^2$ is the MSPE from (3.46). As described in Section 4.13, $\widetilde{\sigma}^2$ is a good estimator of the out-of-sample mean-squared forecast error so *$\widetilde{R}^2$ is a good estimator of the percentage of the forecast variance which is explained by the regression forecast*.

One problem with $R^2$ which is partially corrected by $\overline{R}^2$ and fully corrected by $\widetilde{R}^2$ is that $R^2$ necessarily increases when regressors are added to a regression model. This occurs because $R^2$ is a negative function of the sum of squared residuals which cannot increase when a regressor is added. *In contrast, $\overline{R}^2$ and $\widetilde{R}^2$ are non-monotonic in the number of regressors. $\widetilde{R}^2$ can even be negative, which occurs when an estimated model predicts worse than a constant-only model.*

In the statistical literature the MSPE $\widetilde{\sigma}^2$ is known as the *leave-one-out cross validation* criterion and is popular for model comparison and selection, especially in high-dimensional and nonparametric contexts. It is equivalent to use $\widetilde{R}^2$ or $\widetilde{\sigma}^2$ to compare and select models. Models with high $\widetilde{R}^2$ (or low $\widetilde{\sigma}^2$) are better models in terms of expected out of sample squared error. In contrast, $R^2$ cannot be used for model selection as it necessarily increases when regressors are added to a regression model. $\overline{R}^2$ is also an inappropriate choice for model selection as it tends to select models with too many parameters though a justification of this assertion requires a study of the theory of model selection. Unfortunately, $\overline{R}^2$ is routinely used by some economists, possibly as a hold-over form previous generations.

## 4.22. Multicollinearity

If $\boldsymbol{X}'\boldsymbol{X}$ is singular then $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $\widehat{\overrightarrow{\beta}}$ are not defined. This situation is called *strict multicollinearity* as the columns of $\boldsymbol{X}$ are linearly dependent, i.e., there is some $\overrightarrow{\alpha} \neq \overrightarrow{0}$ such that

$\boldsymbol{X}\vec{\alpha} = 0$. Most commonly this arises when sets of regressors are included which are identically related.

A related common situation is *near multicollinearity* which is often called "multicollinearity" for brevity. This is the situation when the regressors are highly correlated. An implication of near multicollinearity is that individual coefficient estimates will be imprecise. This is not necessarily a problem for econometric analysis if the reported standard errors are accurate. However, robust standard errors can be sensitive to large leverage values which can occur under near multicollinearity. This leads to the undesirable situation where the coefficient estimates are imprecise yet the standard errors are misleadingly small.

## 4.23. Clustered Sampling

In clustering contexts it is convenient to double index the observations as $(Y_{ig}, \vec{X}_{ig})$ where $g = 1, \ldots, G$ indexes the cluster and $i = 1, \ldots, n_g$ indexes the individual within the $g$th cluster. The number of observations per cluster $n_g$ may vary across clusters. The number of clusters is $G$. The total number of observations is $n = \sum_{i=g}^{G} n_g$.

Let $\vec{Y}_g = \left(Y_{1g}, \ldots, Y_{n_g g}\right)'$ and $\boldsymbol{X}_g = \left(\vec{X}_{1g}, \ldots, \vec{X}_{n_g g}\right)'$ denote the $n_g \times 1$ vector of dependent variables and $n_g \times k$ matrix of regressors for the $g$th cluster. A linear regression model can be written by individual as

$$Y_{ig} = \vec{X}_{ig}' \vec{\beta} + e_{ig}$$

and using cluster notation as

$$\vec{Y}_g = \boldsymbol{X}_g \vec{\beta} + \vec{e}_g \tag{4.43}$$

where $\vec{e}_g = \left(e_{1g}, \ldots, e_{n_g g}\right)'$ is a $n_g \times 1$ error vector. We can also stack the observations into full sample matrices and write the model as

$$\vec{Y} = \boldsymbol{X}\vec{\beta} + \vec{e}.$$

The OLS estimator can be written as

$$\begin{aligned}
\widehat{\vec{\beta}} &= \left(\sum_{g=1}^{G} \sum_{i=1}^{n_g} \vec{X}_{ig} \vec{X}_{ig}'\right)^{-1} \left(\sum_{g=1}^{G} \sum_{i=1}^{n_g} \vec{X}_{ig} Y_{ig}\right) \\
&= \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{X}_g\right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \vec{Y}_g\right) \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}'\vec{Y}\right).
\end{aligned} \tag{4.44}$$

The residuals are $\widehat{e}_{ig} = Y_{ig} - \vec{X}_{ig}' \widehat{\vec{\beta}}$ in individual level notation and $\widehat{\vec{e}}_g = \vec{Y}_g - \boldsymbol{X}_g \widehat{\vec{\beta}}$ in cluster level notation.

**Assumption 4.4.**

The clusters $(\overrightarrow{Y}_g, \boldsymbol{X}_g)$ are mutually independent across clusters $g$.

The model is a linear regression under the assumption

$$\mathbb{E}\left[\overrightarrow{e}_g \mid \boldsymbol{X}_g\right] = \overrightarrow{0}. \tag{4.45}$$

This is the same as assuming that the individual errors are conditionally mean zero

$$\mathbb{E}\left[e_{ig} \mid \boldsymbol{X}_g\right] = 0,$$

or that the conditional mean of $\overrightarrow{Y}_g$ given $\boldsymbol{X}_g$ is linear. *As in the independent case equation (4.45) means that the linear regression model is correctly specified.* In the clustered regression model this requires that all interaction effects within clusters have been accounted for in the specification of the individual regressors $\overrightarrow{X}_{ig}$.

Given (4.45) we can calculate the mean of the OLS estimator. Substituting (4.43) into (4.44) we find

$$\widehat{\overrightarrow{\beta}} - \overrightarrow{\beta} = \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{X}_g\right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \overrightarrow{e}_g\right).$$

The mean of $\widehat{\overrightarrow{\beta}} - \overrightarrow{\beta}$ conditioning on all the regressors is

$$\begin{aligned}
\mathbb{E}[\widehat{\overrightarrow{\beta}} - \overrightarrow{\beta} \mid \boldsymbol{X}] &= \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{X}_g\right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \mathbb{E}\left[\boldsymbol{e}_g \mid \boldsymbol{X}\right]\right) \\
&= \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{X}_g\right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \mathbb{E}\left[\boldsymbol{e}_g \mid \boldsymbol{X}_g\right]\right) \\
&= 0.
\end{aligned}$$

The fist equality holds by linearity, the second by Assumption 4.4, and the third by (4.45).

This shows that OLS is unbiased under clustering if the conditional mean is linear, as summarized in Theorem 4.9.

**Theorem 4.9.**

In the clustered linear regression model (Assumption 4.4 and (4.45)),

$$\mathbb{E}\left[\widehat{\overrightarrow{\beta}} \mid \boldsymbol{X}\right] = \overrightarrow{\beta}.$$

Now consider the covariance matrix of $\widehat{\overrightarrow{\beta}}$. Let $\boldsymbol{\Sigma}_g = \mathbb{E}\left[\overrightarrow{e}_g \overrightarrow{e}'_g \mid \boldsymbol{X}_g\right]$ denote the $n_g \times n_g$ conditional covariance matrix of the errors within the $g$th cluster. Since the observations are independent across clusters,

$$\begin{aligned} \operatorname{var}\left[\left(\sum_{g=1}^{G} \boldsymbol{X}_g' \overrightarrow{e}_g\right) \mid \boldsymbol{X}\right] &= \sum_{g=1}^{G} \operatorname{var}\left[\boldsymbol{X}_g' \overrightarrow{e}_g \mid \boldsymbol{X}\right] \\ &= \sum_{g=1}^{G} \boldsymbol{X}_g' \operatorname{var}\left[\overrightarrow{e}_g \mid \boldsymbol{X}\right] \boldsymbol{X}_g \\ &= \sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{\Sigma}_g \boldsymbol{X}_g \\ &=: \boldsymbol{\Omega}_n. \end{aligned} \tag{4.46}$$

It follows that

$$\boldsymbol{V}_{\widehat{\overrightarrow{\beta}}} = \operatorname{var}\left[\widehat{\overrightarrow{\beta}} \mid \boldsymbol{X}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{\Omega}_n \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{4.47}$$

This differs from the formula in the independent case due to the correlation between observations within clusters. The magnitude of the difference depends on the degree of correlation between observations within clusters and the number of observations within clusters. To see this, suppose that all clusters have the same number of observations $n_g = N$, $\mathbb{E}\left[e_{ig}^2 \mid \boldsymbol{X}_g\right] = \sigma^2, \mathbb{E}\left[e_{ig}e_{\ell g} \mid \boldsymbol{X}_g\right] = \sigma^2\rho$ for $i \neq \ell$, and the regressors $\overrightarrow{X}_{ig}$ do not vary within a cluster. In this case, the exact variance of the OLS estimator equals

$$\boldsymbol{V}_{\widehat{\overrightarrow{\beta}}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \sigma^2 \left(1 + \rho\left(N-1\right)\right). \tag{4.48}$$

Arellano proposed a cluster-robust covariance matrix estimator which is an extension of the White estimator. Recall that the insight of the White covariance estimator is that the squared error $e_i^2$ is unbiased for $\mathbb{E}\left[e_i^2 \mid \overrightarrow{X}_i\right] = \sigma_i^2$. Similarly with cluster dependence the matrix $\overrightarrow{e}_g \overrightarrow{e}_g'$ is unbiased for $\mathbb{E}\left[\overrightarrow{e}_g \overrightarrow{e}_g' \mid \boldsymbol{X}_g\right] = \boldsymbol{\Sigma}_g$. This means that an unbiased estimator for (4.46) is $\widetilde{\boldsymbol{\Omega}}_n = \sum_{g=1}^{G} \boldsymbol{X}_g' \overrightarrow{e}_g \overrightarrow{e}_g' \boldsymbol{X}_g$. This is not feasible, but we can replace the unknown errors by the OLS residuals to obtain Arellano's estimator

$$\begin{aligned} \widehat{\boldsymbol{\Omega}}_n &= \sum_{g=1}^{G} \boldsymbol{X}_g' \widehat{\overrightarrow{e}}_g \widehat{\overrightarrow{e}}_g' \boldsymbol{X}_g \\ &= \sum_{g=1}^{G} \sum_{i=1}^{n_g} \sum_{\ell=1}^{n_g} \overrightarrow{X}_{ig} \overrightarrow{X}_{\ell g}' \widehat{e}_{ig} \widehat{e}_{\ell g} \\ &= \sum_{g=1}^{G} \left(\sum_{i=1}^{n_g} \overrightarrow{X}_{ig} \widehat{e}_{ig}\right) \left(\sum_{\ell=1}^{n_g} \overrightarrow{X}_{\ell g} \widehat{e}_{\ell g}\right)'. \end{aligned} \tag{4.49}$$

Given the expressions (4.46)-(4.47), a natural cluster covariance matrix estimator takes the form

$$\boldsymbol{V}_{\widehat{\overrightarrow{\beta}}} = a_n \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \widehat{\boldsymbol{\Omega}}_n \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \tag{4.50}$$

where $a_n$ is a possible finite-sample adjustment. The Stata `cluster` command uses

$$a_n = \left(\frac{n-1}{n-k}\right)\left(\frac{G}{G-1}\right). \tag{4.51}$$

The factor $\frac{G}{G-1}$ was derived by Chris Hansen (2007) in the context of equal-sized clusters to improve performance when the number of clusters $G$ is small. The factor $\frac{n-1}{n-k}$ is an ad hoc generalization which nests the adjustment used in (4.32) since $G = n$ implies the simplification an $a_n = n/(n-k)$.

Alternative cluster-robust covariance matrix estimators can be constructed using cluster-level prediction errors such as $\widetilde{\overrightarrow{e}}_g = \overrightarrow{Y}_g - \boldsymbol{X}_g \widehat{\overrightarrow{\beta}}_{(-g)}$ where $\widehat{\overrightarrow{\beta}}_{(-g)}$ is the least squares estimator omitting cluster $g$. As in Section 3.20, we can show that

$$\widetilde{\overrightarrow{e}}_g = \left(\boldsymbol{I}_{n_g} - \boldsymbol{X}_g \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'_g\right)^{-1} \widehat{\overrightarrow{e}}_g \tag{4.52}$$

and

$$\widehat{\overrightarrow{\beta}}_{(-g)} = \widehat{\overrightarrow{\beta}} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'_g \widetilde{\overrightarrow{e}}_g. \tag{4.53}$$

We then have the robust covariance matrix estimator

$$\widehat{\boldsymbol{V}}^{\mathrm{CR3}}_{\widehat{\overrightarrow{\beta}}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}'_g \widetilde{\overrightarrow{e}}_g \widetilde{\overrightarrow{e}}'_g \boldsymbol{X}_g\right) \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{4.54}$$

The label "CR" refers to "cluster-robust" and "CR3" refers to the analogous formula for the HC3 estimator.

Similarly to the heteroskedastic-robust case you can show that CR3 is a conservative estimator for $\boldsymbol{V}_{\widehat{\overrightarrow{\beta}}}$ in the sense that the conditional expectation for $\widehat{\boldsymbol{V}}^{\mathrm{CR3}}_{\widehat{\overrightarrow{\beta}}}$ exceeds $\boldsymbol{V}_{\widehat{\overrightarrow{\beta}}}$. This covariance matrix estimator may be more cumbersome to implement, however, as the cluster-level prediction errors (4.52) cannot be calculated in a simple linear operation and appear to require a loop (across clusters) to calculate.

## 4.24. Inference with Clustered Samples

*In many respects cluster-robust inference should be viewed similarly to heteroskedasticity-robust inference where a "cluster" in the cluster-robust case is interpreted similarly to an "observation" in the heteroskedasticity-robust case.*

*In particular, the effective sample size should be viewed as the number of clusters not the "sample size" n.* This is because the cluster-robust covariance matrix estimator effectively treats each cluster as a single observation and estimates the covariance matrix based on the variation across cluster means.

Furthermore, most cluster-robust theory (for example, the work of Chris Hansen (2007)) assumes that the clusters are homogeneous including the assumption that the cluster sizes are all identical. This turns out to be a very important simplication. When this is violated – *when, for example, cluster sizes are highly heterogeneous – the regression should be viewed as roughly equivalent to the heteroskedastic case with an extremely high degree of heteroskedasticity*. Cluster sums have variances which are proportional to the cluster sizes so if the latter is heterogeneous so will be the variances of the cluster sums. This also has a large effect on finite sample inference. When clusters are heterogeneous then cluster-robust inference is similar to heteroskedasticity-robust inference with highly heteroskedastic observations.

Put together, if the number of clusters $G$ is small and the number of observations per cluster is highly varied then we should interpret inferential statements with a great degree of caution. Unfortunately, small $G$ with heterogeneous cluster size is commonplace. Many empirical studies on U.S. data cluster at the "state" level meaning that there are 50 or 51 clusters (the District of Columbia is typically treated as a state). The number of observations vary considerably across states since the populations are highly unequal. Thus when you read empirical papers with individual-level data but clustered at the "state" level you should be cautious and recognize that this is equivalent to inference with a small number of extremely heterogeneous observations.

A further complication occurs when we are interested in treatment as in the tracking example given in the previous section. In many cases the interest is in the effect of a treatment applied at the cluster level (e.g., schools). In many cases, the number of treated clusters is small relative to the total number of clusters; in an extreme case there is just a single treated cluster. Based on the reasoning given above these applications should be interpreted as equivalent to heteroskedasticity-robust inference with a sparse dummy variable as discussed in Section 4.17. As discussed there, standard error estimates can be erroneously small. In the extreme of a single treated cluster (in the example, if only a single school was tracked) then the estimated coefficient on tracking will be very imprecisely estimated yet will have a misleadingly small cluster standard error. In general, reported standard errors will greatly understate the imprecision of parameter estimates.

## 4.25. At What Level to Cluster?

*First, suppose cluster dependence is ignored or imposed at too fine a level (e.g. clustering by households instead of villages). Then variance estimators will be biased as they will omit covariance terms. As correlation is typically positive, this suggests that standard errors will be too small giving rise to spurious indications of significance and precision.*

Second, suppose cluster dependence is imposed at too aggregate a measure (e.g. clustering by states rather than villages). This does not cause bias. But the variance estimators will contain many extra components so the precision of the covariance matrix estimator will be poor. This means that reported standard errors will be imprecise – more random – than if clustering had been less aggregate.

These considers show that there is a trade-off between bias and variance in the estimation of the covariance matrix by cluster-robust methods. It is not at all clear – based on current theory – what to do.

One challenge is that in empirical practice many people have observed: "Clustering is important. Standard errors change a lot whether or not we cluster. Therefore we should only report clustered standard errors." The flaw in this reasoning is that we do not know why in a specific empirical example the standard errors change under clustering. One possibility is that clustering reduces bias and thus is more accurate. The other possibility is that clustering adds sampling noise and is thus less accurate. In reality it is likely that both factors are present.

*Recent advancements and survey on clustering standard errors should be reviewed separately.*

# Chapter 5

# Normal Regression

## 5.2. The Normal Distribution

We say that a random variable $Z$ has the *standard normal distribution*, or *Gaussian*, written $Z \sim N(0,1)$, if it has the density

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \quad -\infty < x < \infty.$$

The standard normal density is typically written with the symbol $\phi(x)$ and the corresponding distribution by $\Phi(x)$.

**Theorem 5.1.**

If $Z \sim N(0,1)$, then

(1) All integer moments of $Z$ are finite.

(2) All odd moments of $Z$ equal 0.

(3) For any positive integer $m$

$$\mathbb{E}\left[Z^{2m}\right] = (2m-1)!! = (2m-1) \times (2m-3) \times \ldots \times 1.$$

(4) For any $r > 0$,

$$\mathbb{E}\left[|Z|^r\right] = \frac{2^{r/2}}{\sqrt{\pi}} \Gamma\left(\frac{r+1}{2}\right)$$

where $\Gamma(t) = \int_0^\infty u^{t-1} e^{-u} du$ is the gamma function.

If $Z \sim N(0,1)$ and $X = \mu + \sigma Z$ for $\mu \in \mathbb{R}$ and $\sigma \geq 0$ then $X$ has the *univariate normal distribution*, written $X \sim N\left(\mu, \sigma^2\right)$. $X$ has the density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty.$$

The mean and variance of $X$ are $\mu$ and $\sigma^2$, respectively.

The normal distribution and its relatives (the chi-square, student $t$, $F$, non-central chi-square, and non-central $F$) are frequently used for inference to calculate critical values and $p$-values. This involves evaluating the normal cdf $\Phi(x)$ and its inverse. Since the cdf $\Phi(x)$ is not available in closed form statistical textbooks have traditionally provided tables for this purpose.

## 5.3. Multivariate Normal Distribution

We say that the $k$-vector $\overrightarrow{Z}$ has a *multivariate standard normal distribution*, written $\overrightarrow{Z} \sim N(\overrightarrow{0}, \boldsymbol{I}_k)$, if it has the joint density,

$$f(\overrightarrow{x}) = \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{\overrightarrow{x}'\overrightarrow{x}}{2}\right), \quad x \in \mathbb{R}^k.$$

The mean and covariance matrix of $\overrightarrow{Z}$ are $\overrightarrow{0}$ and $\boldsymbol{I}_k$, respectively. The multivariate joint density factors into the product of univariate normal densities, so the elements of $\overrightarrow{Z}$ are mutually independent standard normals.

If $\overrightarrow{Z} \sim N(\overrightarrow{0}, \boldsymbol{I}_k)$ and $\overrightarrow{X} = \overrightarrow{\mu} + \boldsymbol{B}\overrightarrow{Z}$ then the $k$-vector $\overrightarrow{X}$ has a *multivariate normal distribution*, written $\overrightarrow{X} \sim N(\overrightarrow{\mu}, \Sigma)$, where $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}'$. If $\boldsymbol{\Sigma} > \boldsymbol{0}$, then by change-of-variables $\overrightarrow{X}$ has the joint density function

$$f(\overrightarrow{x}) = \frac{1}{(2\pi)^{k/2}\det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{(\overrightarrow{x} - \overrightarrow{\mu})'\Sigma^{-1}(\overrightarrow{x} - \overrightarrow{\mu})}{2}\right), \quad \overrightarrow{x} \in \mathbb{R}^k.$$

An important property of normal random vectors is that affine functions are multivariate normal.

**Theorem 5.2.**

If $\overrightarrow{X} \sim N(\overrightarrow{\mu}, \Sigma)$ and $Y = \overrightarrow{a} + \boldsymbol{B}\overrightarrow{X}$, then $\overrightarrow{Y} \sim N(\overrightarrow{a} + \boldsymbol{B}\overrightarrow{\mu}, \boldsymbol{B\Sigma B}')$.

One simple implication of Theorem 5.2 is that if $\overrightarrow{X}$ is multivariate normal then each component of $\overrightarrow{X}$ is univariate normal.

Another useful property of the multivariate normal distribution is that uncorrelatedness is the same as independence. *That is, if a vector is multivariate normal, subsets of variables are independent if and only if they are uncorrelated.*

**Theorem 5.3. Properties of the Multivariate Normal Distribution**

(1) The mean and covariance matrix of $\overrightarrow{X} \sim N(\overrightarrow{\mu}, \boldsymbol{\Sigma})$ are $\mathbb{E}[\overrightarrow{X}] = \overrightarrow{\mu}$ and $\operatorname{var}[\overrightarrow{X}] = \boldsymbol{\Sigma}$.

(2) If $(\overrightarrow{X}, \overrightarrow{Y})$ are multivariate normal, $\overrightarrow{X}$ and $\overrightarrow{Y}$ are uncorrelated if and only if they are independent.

(3) If $\overrightarrow{X} \sim N(\overrightarrow{\mu}, \Sigma)$ and $Y = \overrightarrow{a} + \boldsymbol{B}\overrightarrow{X}$, then $\overrightarrow{Y} \sim N(\overrightarrow{a} + \boldsymbol{B}\overrightarrow{\mu}, \boldsymbol{B}\Sigma\boldsymbol{B}')$.

(4) If $\overrightarrow{X} \sim N(\overrightarrow{0}, \boldsymbol{I}_k)$ then $\overrightarrow{X}'\overrightarrow{X} \sim \chi^2_k$, chi-square with $k$ degrees of freedom.

(5) If $\overrightarrow{X} \sim N(\overrightarrow{0}, \Sigma)$ with $\boldsymbol{\Sigma} > \boldsymbol{0}$ then $\overrightarrow{X}'\boldsymbol{\Sigma}^{-1}\overrightarrow{X} \sim \chi^2_k$ where $k = \dim(X)$.

(6) If $\overrightarrow{X} \sim N(\overrightarrow{\mu}, \Sigma)$ with $\boldsymbol{\Sigma} > \boldsymbol{0}, r \times r$, then $\overrightarrow{X}'\boldsymbol{\Sigma}^{-1}\overrightarrow{X} \sim \chi^2_k(\lambda)$ where $\lambda = \overrightarrow{\mu}'\boldsymbol{\Sigma}^{-1}\overrightarrow{\mu}$.

(7) If $Z \sim N(0,1)$ and $Q \sim \chi^2_k$ are independent then $\frac{Z}{\sqrt{Q/k}} \sim t_k$, student $t$ with $k$ degrees of freedom.

(8) If $(\overrightarrow{Y}, \overrightarrow{X})$ are multivariate normal,

$$\begin{pmatrix} \overrightarrow{Y} \\ \overrightarrow{X} \end{pmatrix} \sim \mathrm{N}\left( \begin{pmatrix} \overrightarrow{\mu_Y} \\ \overrightarrow{\mu_X} \end{pmatrix}, \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix} \right)$$

where $\boldsymbol{\Sigma_{YY}} > \boldsymbol{0}$ and $\boldsymbol{\Sigma_{XX}} > \boldsymbol{0}$ then the conditional distributions are

$$\overrightarrow{Y} \mid \overrightarrow{X} \sim \mathrm{N}\left( \overrightarrow{\mu_Y} + \Sigma_{YX}\Sigma_{XX}^{-1}(\overrightarrow{X} - \overrightarrow{\mu_X}), \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} \right)$$
$$\overrightarrow{X} \mid \overrightarrow{Y} \sim \mathrm{N}\left( \overrightarrow{\mu_X} + \Sigma_{XY}\Sigma_{YY}^{-1}(\overrightarrow{Y} - \overrightarrow{\mu_Y}), \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} \right).$$

*Multivariate normal distributions and corresponding derived distributions should be reviewed in the appendix.*

## 5.4. Joint Normality and Linear Regression

Suppose the variables $(Y, \overrightarrow{X})$ are jointly normally distributed. Consider the best linear predictor of $Y$ given $\overrightarrow{X}$

$$Y = \overrightarrow{X}'\overrightarrow{\beta} + \alpha + e.$$

By the properties of the best linear predictor, $\mathbb{E}[\overrightarrow{X}e] = \overrightarrow{0}$ and $\mathbb{E}[e] = 0$, so $\overrightarrow{X}$ and $e$ are uncorrelated. Since $(e, \overrightarrow{X})$ is an affine transformation of the normal vector $(Y, \overrightarrow{X})$ it follows that $(e, \overrightarrow{X})$ is jointly normal (Theorem 5.2). Since $(e, \overrightarrow{X})$ is jointly normal and uncorrelated they are independent (Theorem 5.3). Independence implies that

$$\mathbb{E}[e \mid \overrightarrow{X}] = \mathbb{E}[e] = 0$$

and

$$\mathbb{E}\left[e^2 \mid \overrightarrow{X}\right] = \mathbb{E}\left[e^2\right] = \sigma^2$$

which are properties of a homoskedastic linear CEF.

We have shown that when $(Y, \overrightarrow{X})$ are jointly normally distributed they satisfy a normal linear CEF

$$Y = \overrightarrow{X}'\overrightarrow{\beta} + \alpha + e$$

where

$$e \sim N\left(0, \sigma^2\right)$$

is independent of $\overrightarrow{X}$.

This is a classical motivation for the linear regression model.

## 5.5. Normal Regression Model

The *normal regression model* is the linear regression model with an independent normal error

$$Y = \overrightarrow{X}'\overrightarrow{\beta} + e \quad e \sim N\left(0, \sigma^2\right). \tag{5.1}$$

As we learned in Section 5.4 the normal regression model holds when $(Y, \overrightarrow{X})$ are jointly normally distributed. *Normal regression, however, does not require joint normality. All that is required is that the conditional distribution of $Y$ given $\overrightarrow{X}$ is normal (the marginal distribution of $\overrightarrow{X}$ is unrestricted).* In this sense the normal regression model is broader than joint normality. Notice that for notational convenience we have written (5.1) so that $\overrightarrow{X}$ contains the intercept.

Normal regression is a parametric model where likelihood methods can be used for estimation, testing, and distribution theory. The *likelihood* is the name for the joint probability density of the data, evaluated at the observed sample, and viewed as a function of the parameters. The maximum likelihood estimator is the value which maximizes this likelihood function. Let us now derive the likelihood of the normal regression model.

First, observe that model (5.1) is equivalent to the statement that the conditional density of $Y$ given $\overrightarrow{X}$ takes the form

$$f(y \mid \overrightarrow{x}) = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\left(y - \overrightarrow{x}'\overrightarrow{\beta}\right)^2\right).$$

Under *the assumption that the observations are mutually independent* this implies that the conditional density of $(Y_1, \ldots, Y_n)$ given $(\overrightarrow{X}_1, \ldots, \overrightarrow{X}_n)$ is

$$\begin{aligned} f\left(y_1, \ldots, y_n \mid \overrightarrow{x}_1, \ldots, \overrightarrow{x}_n\right) &= \prod_{i=1}^{n} f\left(y_i \mid \overrightarrow{x}_i\right) \\ &= \prod_{i=1}^{n} \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \overrightarrow{x}_i'\overrightarrow{\beta}\right)^2\right) \\ &= \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \overrightarrow{x}_i'\overrightarrow{\beta}\right)^2\right) \\ &=: L_n\left(\overrightarrow{\beta}, \sigma^2\right). \end{aligned}$$

This is called the *likelihood function* when evaluated at the sample data.

For convenience it is typical to work with the natural logarithm

$$\log L_n\left(\overrightarrow{\beta}, \sigma^2\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \overrightarrow{X}_i'\overrightarrow{\beta}\right)^2 =: \ell_n\left(\overrightarrow{\beta}, \sigma^2\right) \tag{5.2}$$

which is called the *log-likelihood function*.

The *maximum likelihood estimator* (MLE) $\left(\widehat{\overrightarrow{\beta}}_{\text{mle}}, \widehat{\sigma}^2_{\text{mle}}\right)$ is the value which maximizes the log-likelihood. We can write the maximization problem as

$$\left(\widehat{\overrightarrow{\beta}}_{\text{mle}}, \widehat{\sigma}^2_{\text{mle}}\right) = \underset{\overrightarrow{\beta} \in \mathbb{R}^k, \sigma^2 > 0}{\operatorname{argmax}}\ \ell_n\left(\overrightarrow{\beta}, \sigma^2\right). \tag{5.3}$$

In most applications of maximum likelihood the MLE must be found by numerical methods. However in the case of the normal regression model we can find an explicit expression for $\widehat{\overrightarrow{\beta}}_{\text{mle}}$ and $\widehat{\sigma}^2_{\text{mle}}$.

$$0 = \frac{\partial}{\partial\overrightarrow{\beta}}\ell_n\left(\overrightarrow{\beta}, \sigma^2\right)\bigg|_{\overrightarrow{\beta} = \widehat{\overrightarrow{\beta}}_{\text{mle}}, \sigma^2 = \widehat{\sigma}^2_{\text{mle}}} = \frac{1}{\widehat{\sigma}^2_{\text{mle}}}\sum_{i=1}^{n}\overrightarrow{X}_i\left(Y_i - \overrightarrow{X}_i'\widehat{\overrightarrow{\beta}}_{\text{mle}}\right) \tag{5.4}$$

$$0 = \frac{\partial}{\partial\sigma^2}\ell_n\left(\overrightarrow{\beta}, \sigma^2\right)\bigg|_{\overrightarrow{\beta} = \widehat{\overrightarrow{\beta}}_{\text{mle}}, \sigma^2 = \widehat{\sigma}^2_{\text{mle}}} = -\frac{n}{2\widehat{\sigma}^2_{\text{mle}}} + \frac{1}{2\widehat{\sigma}^4_{\text{mle}}}\sum_{i=1}^{n}\left(Y_i - \overrightarrow{X}_i'\widehat{\overrightarrow{\beta}}_{\text{mle}}\right)^2. \tag{5.5}$$

It follows that the MLE satisfies

$$\widehat{\overrightarrow{\beta}}_{\text{mle}} = \left(\sum_{i=1}^{n}\overrightarrow{X}_i\overrightarrow{X}_i'\right)^{-1}\left(\sum_{i=1}^{n}\overrightarrow{X}_iY_i\right) = \widehat{\overrightarrow{\beta}}_{\text{ols}}.$$

That is, the MLE for $\overrightarrow{\beta}$ is algebraically identical to the OLS estimator.

Solving the second FOC (5.5) for $\widehat{\sigma}^2_{\text{mle}}$ we find

$$\widehat{\sigma}^2_{\text{mle}} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overrightarrow{X}_i'\widehat{\overrightarrow{\beta}}_{\text{mle}}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overrightarrow{X}_i'\widehat{\overrightarrow{\beta}}_{\text{ols}}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2 = \widehat{\sigma}^2_{\text{ols}}.$$

Thus theMLE for $\sigma^2$ is identical to the OLS/moment estimator from (3.26).

Plugging the estimators into (5.2) we obtain the maximized log-likelihood

$$\ell_n\left(\widehat{\overrightarrow{\beta}}_{\text{mle}}, \widehat{\sigma}^2_{\text{mle}}\right) = -\frac{n}{2}\log\left(2\pi\widehat{\sigma}^2_{\text{mle}}\right) - \frac{n}{2}. \tag{5.6}$$

The log-likelihood is typically reported as a measure of fit.

It may seem surprising that the MLE $\widehat{\overrightarrow{\beta}}_{\text{mle}}$ is numerically equal to the OLS estimator despite emerging from quite different motivations. It is not completely accidental. The least squares estimator minimizes a particular sample loss function – the sum of squared error criterion – and *most loss functions are equivalent to the likelihood of a specific parametric distribution, in this case the normal regression model*. In this sense it is not surprising that the least squares estimator can be motivated as either the minimizer of a sample loss function or as the maximizer of a likelihood function.

## 5.6. Distribution of OLS Coefficient Vector

In the normal linear regression model we can derive exact sampling distributions for the OLS/MLE estimator, residuals, and variance estimator. In this section we derive the distribution of the OLS coefficient estimator.

The normality assumption $e \mid \overrightarrow{X} \sim N\left(0, \sigma^2\right)$ combined with independence of the observations has the multivariate implication

$$\overrightarrow{e} \mid \boldsymbol{X} \sim N\left(\overrightarrow{0}, \boldsymbol{I}_n\sigma^2\right).$$

That is, the error vector $\overrightarrow{e}$ is independent of $\boldsymbol{X}$ and is normally distributed.

Recall that the OLS estimator satisfies

$$\widehat{\overrightarrow{\beta}} - \overrightarrow{\beta} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\overrightarrow{e}$$

which is a linear function of $\overrightarrow{e}$. Since linear functions of normals are also normal (Theorem 5.2) this implies that conditional on $\boldsymbol{X}$,

$$\begin{aligned}\widehat{\overrightarrow{\beta}} - \overrightarrow{\beta} \mid \boldsymbol{X} &\sim \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\mathrm{N}\left(\overrightarrow{0}, \boldsymbol{I}_n\sigma^2\right) \\ &\sim \mathrm{N}\left(\overrightarrow{0}, \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right) \\ &= \mathrm{N}\left(\overrightarrow{0}, \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right).\end{aligned}$$

**Theorem 5.4.**

In the normal regression model,

$$\widehat{\overrightarrow{\beta}} \mid \boldsymbol{X} \sim N\left(\overrightarrow{\beta}, \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right).$$

Any affine function of the OLS estimator is also normally distributed including individual components. Letting $\beta_j$ and $\widehat{\beta}_j$ denote the $j$th element of $\overrightarrow{\beta}$ and $\widehat{\overrightarrow{\beta}}$, we have

$$\widehat{\beta}_j \mid \boldsymbol{X} \sim N\left(\beta_j, \sigma^2\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]_{jj}\right). \tag{5.7}$$

Theorem 5.4 is a statement about the conditional distribution. What about the unconditional distribution? In Theorem 4.3 we presented Kinal's theorem about the existence of moments for the joint normal regression model. We re-state the result here.

**Theorem 5.5. Kinal (1980)**

If $(Y, \overrightarrow{X})$ are jointly normal, then for any $r$, $\mathbb{E}\left[\left\|\widehat{\overrightarrow{b}}\right\|^r\right] < \infty$ if and only if $r < n - k + 1$.

## 5.7. Distribution of OLS Residual Vector

Consider the OLS residual vector. Recall from (3.24) that $\widehat{\overrightarrow{e}} = \boldsymbol{M}\overrightarrow{e}$ where $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$. This shows that $\widehat{\overrightarrow{e}}$ is linear in $\overrightarrow{e}$. So conditional on $\boldsymbol{X}$

$$\widehat{\overrightarrow{e}} \mid \boldsymbol{X} = \boldsymbol{M}\overrightarrow{e} \mid \boldsymbol{X} \sim N\left(\overrightarrow{0}, \sigma^2\boldsymbol{M}\boldsymbol{M}\right) = N\left(\overrightarrow{0}, \sigma^2\boldsymbol{M}\right)$$

the final equality since $\boldsymbol{M}$ is idempotent.

Furthermore, it is useful to find the joint distribution of $\widehat{\overrightarrow{\beta}}$ and $\widehat{\overrightarrow{e}}$. This is easiest done by writing the two as a stacked linear function of the error $\overrightarrow{e}$. Indeed,

$$\begin{pmatrix} \widehat{\overrightarrow{\beta}} - \overrightarrow{\beta} \\ \widehat{\overrightarrow{e}} \end{pmatrix} = \begin{pmatrix} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\overrightarrow{e} \\ \boldsymbol{M}\overrightarrow{e} \end{pmatrix} = \begin{pmatrix} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' \\ \boldsymbol{M} \end{pmatrix}\overrightarrow{e}$$

which is a linear function of $\overrightarrow{e}$. The vector thus has a joint normal distribution with covariance matrix

$$\begin{pmatrix} \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma^2 \boldsymbol{M}. \end{pmatrix}$$

The off-diagonal block is zero because $\boldsymbol{X}'\boldsymbol{M} = \mathbf{0}$. Since this is zero it follows that $\widehat{\overrightarrow{\beta}}$ and $\widehat{\overrightarrow{e}}$ are statistically independent.

**Theorem 5.6.**

In the normal regression model, $\widehat{\overrightarrow{e}} \mid \boldsymbol{X} \sim N\left(\overrightarrow{0}, \sigma^2 \boldsymbol{M}\right)$ and is independent of $\widehat{\overrightarrow{\beta}}$.

*The fact that $\widehat{\overrightarrow{\beta}}$ and $\widehat{\overrightarrow{e}}$ are independent implies that $\widehat{\overrightarrow{\beta}}$ is independent of any function of the residual vector including individual residuals $\widehat{e}_i$ and the variance estimators $s^2$ and $\widehat{\sigma}^2$.*

## 5.8. Distribution of Variance Estimator

Next, consider the variance estimator $s^2$ defined as

$$s^2 = \frac{1}{n-k} \sum_{i=1}^{n} \widehat{e}_i^2.$$

Using (3.28) it satisfies $(n-k)\, s^2 = \widehat{\overrightarrow{e}}'\widehat{\overrightarrow{e}} = \overrightarrow{e}'\boldsymbol{M}\overrightarrow{e}$.

The spectral decomposition of $\boldsymbol{M}$ is $\boldsymbol{M} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ where $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}$ and $\boldsymbol{\Lambda}$ is diagonal with the eigenvalues of $\boldsymbol{M}$ on the diagonal. Since $\boldsymbol{M}$ is idempotent with rank $n-k$ it has $n-k$ eigenvalues equalling 1 and $k$ eigenvalues equalling 0, so

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{I}_{n-k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_k \end{bmatrix}.$$

Let $\overrightarrow{u} = \boldsymbol{H}'\overrightarrow{e} \sim N\left(\overrightarrow{0}, \boldsymbol{I}_n \sigma^2\right)$ and partition $\overrightarrow{u} = \left(\overrightarrow{u}_1', \overrightarrow{u}_2'\right)'$. Then

$$\begin{aligned}
(n-k)s^2 &= \overrightarrow{e}'\boldsymbol{M}\overrightarrow{e} \\
&= \overrightarrow{e}'\boldsymbol{H} \begin{bmatrix} \boldsymbol{I}_{n-k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \boldsymbol{H}'\overrightarrow{e} \\
&= \overrightarrow{u}' \begin{bmatrix} \boldsymbol{I}_{n-k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \overrightarrow{u} \\
&= \overrightarrow{u}_1'\overrightarrow{u}_1 \\
&\sim \sigma^2 \chi^2_{n-k}.
\end{aligned}$$

We see that in the normal regression model the exact distribution of $s^2$ is scaled chi-squared.

Since $\widehat{\overrightarrow{e}}$ is independent of $\widehat{\overrightarrow{\beta}}$ it follows that $s^2$ is independent of $\overrightarrow{\beta}$ as well.

**Theorem 5.7.**

In the normal regression model,

$$\frac{(n-k)\,s^2}{\sigma^2} \sim \chi^2_{n-k},$$

and is independent of $\widehat{\overrightarrow{\beta}}$.

## 5.9. *t*-Statistic

An alternative way of writing (5.7) is

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\sigma^2\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]_{jj}}} \sim N(0,1).$$

This is sometimes called a *standardized* statistic as the distribution is the standard normal.

Now take the standardized statistic and replace the unknown variance $\sigma^2$ with its estimator $s^2$. We call this a *t-ratio* or *t-statistic*

$$T = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{s^2\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]_{jj}}} = \frac{\widehat{\beta}_j - \beta_j}{s\left(\widehat{\beta}_j\right)},$$

where $s\left(\widehat{\beta}_j\right)$ is the classical (homoskedastic) standard error for $\widehat{\beta}_j$ from (4.37). We will sometimes write the $t$-statistic as $T\left(\beta_j\right)$ to explicitly indicate its dependence on the parameter value $\beta_j$, and sometimes will simplify notation and write the $t$-statistic as $T$ when the dependence is clear from the context.

With algebraic re-scaling we can write the t-statistic as the ratio of the standardized statistic and the square root of the scaled variance estimator. Since the distributions of these two components are normal and chi-square, respectively, and independent, we deduce that the $t$-statistic has the distribution

$$\begin{aligned} T &= \frac{\widehat{\beta}_j - \beta_j}{\sqrt{\sigma^2\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]_{jj}}} \Big/ \sqrt{\frac{(n-k)s^2}{\sigma^2}\Big/(n-k)} \\ &\sim \frac{\mathrm{N}(0,1)}{\sqrt{\chi^2_{n-k}/(n-k)}} \\ &\sim t_{n-k}, \end{aligned}$$

a student $t$ distribution with $n - k$ degrees of freedom.

*This derivation shows that the t-ratio has a sampling distribution which depends only on the quantity $n - k$. The distribution does not depend on any other features of the data.* In this context, we say that the distribution of the $t$-ratio is *pivotal*, meaning that it does not depend on unknowns.

*The trick behind this result is scaling the centered coefficient by its standard error, and recognizing that each depends on the unknown $\sigma$ only through scale.* Thus the ratio of the two does not depend on $\sigma$. This trick (scaling to eliminate dependence on unknowns) is known as studentization.

**Theorem 5.8.**

In the normal regression model, $T \sim t_{n-k}$.

*An important caveat about Theorem 5.8 is that it only applies to the t-statistic constructed with the homoskedastic (old-fashioned) standard error. It does not apply to a t-statistic constructed with any of the robust standard errors.* In fact, the robust $t$-statistics can have finite sample distributions which deviate considerably from $t_{n-k}$ even when the regression errors are independent $N\left(0, \sigma^2\right)$. *Thus the distributional result in Theorem 5.8 and the use of the t distribution in finite samples is only exact when applied to classical t-statistics under the normality assumption.*

## 5.10. Confidence Intervals for Regression Coefficients

The OLS estimator $\widehat{\overrightarrow{\beta}}$ is a *point estimator* for a coefficient $\overrightarrow{\beta}$. A broader concept is a *set* or *interval estimator* which takes the form $\widehat{C} = \left[\widehat{L}, \widehat{U}\right]$. The goal of an interval estimator $\widehat{C}$ is to contain the true value, e.g., $\overrightarrow{\beta} \in \widehat{C}$ with high probability.

The interval estimator $\widehat{C}$ is a function of the data and hence is random.

An interval estimator $\widehat{C}$ is called a $1-\alpha$ *confidence interval* when $\mathbb{P}\left[\overrightarrow{\beta} \in \widehat{C}\right] = 1-\alpha$ for a selected value of $\alpha$. The value $1 - \alpha$ is called the *coverage probability*.

The probability calculation $\mathbb{P}\left[\overrightarrow{\beta} \in \widehat{C}\right]$ is easily mis-interpreted as treating $\overrightarrow{\beta}$ as random and $\widehat{C}$ as fixed. (The probability that $\overrightarrow{\beta}$ is in $\widehat{C}$.) This is not the appropriate interpretation. *Instead, the correct interpretation is that the probability $\mathbb{P}\left[\overrightarrow{\beta} \in \widehat{C}\right]$ treats the point $\overrightarrow{\beta}$ as fixed and the set $\widehat{C}$ as random. It is the probability that the random set $\widehat{C}$ covers (or contains) the fixed true coefficient $\overrightarrow{\beta}$.*

There is not a unique method to construct confidence intervals. For example, one simple (yet silly) interval is

$$\widehat{C} = \begin{cases} \mathbb{R} & \text{with probability } 1 - \alpha \\ \left\{\widehat{\overrightarrow{\beta}}\right\} & \text{with probability } \alpha. \end{cases}$$

If $\widehat{\overrightarrow{\beta}}$ has a continuous distribution, then by construction $\mathbb{P}\left[\overrightarrow{\beta} \in \widehat{C}\right] = 1 - \alpha$, so this confidence interval has perfect coverage. However, $\widehat{C}$ is uninformative and is therefore not useful.

A good choice for a confidence interval for the regression coefficient $\overrightarrow{\beta}$ is obtained by adding and subtracting from the estimator $\widehat{\overrightarrow{\beta}}$ a fixed multiple of its standard error:

$$\widehat{C} = \left[\widehat{\overrightarrow{\beta}} - c \times s\left(\widehat{\overrightarrow{\beta}}\right), \widehat{\overrightarrow{\beta}} + c \times s\left(\widehat{\overrightarrow{\beta}}\right)\right] \tag{5.8}$$

where $c > 0$ is a pre-specified constant. This confidence interval is symmetric about the point estimator $\widehat{\overrightarrow{\beta}}$ and its length is proportional to the standard error $s\left(\widehat{\overrightarrow{\beta}}\right)$.

Equivalently, $\widehat{C}$ is the set of parameter values for $\overrightarrow{\beta}$ such that the $t$-statistic $T(\beta)$ is smaller (in absolute value) than $c$, that is

$$\widehat{C} = \{\overrightarrow{\beta} : |T(\overrightarrow{\beta})| \le c\} = \left\{\overrightarrow{\beta} : -c \le \frac{\widehat{\overrightarrow{\beta}} - \overrightarrow{\beta}}{s\left(\widehat{\overrightarrow{\beta}}\right)} \le c\right\}.$$

The coverage probability of this confidence interval is

$$\begin{aligned} \mathbb{P}[\overrightarrow{\beta} \in \widehat{C}] &= \mathbb{P}\left[|T(\overrightarrow{\beta})| \le c\right] \\ &= \mathbb{P}\left[-c \le T(\overrightarrow{\beta}) \le c\right]. \end{aligned} \tag{5.9}$$

Since the $t$-statistic $T(\overrightarrow{\beta})$ has the $t_{n-k}$ distribution (5.9) equals $F(c) - F(-c)$, where $F(u)$ is the student $t$ distribution function with $n-k$ degrees of freedom. Since $F(-c) - 1 - F(c)$ we can write (5.9) as

$$\mathbb{P}\left[\overrightarrow{\beta} \in \widehat{C}\right] = 2F(c) - 1.$$

This is the coverage probability of the interval $\widehat{C}$, and only depends on the constant $c$.

As we mentioned before, a confidence interval has the coverage probability $1-\alpha$. This requires selecting the constant $c$ so that $F(c) = 1-\alpha/2$. This holds if $c$ equals the $1-\alpha/2$ quantile of the $t_{n-k}$ distribution. As there is no closed form expression for these quantiles we compute their values numerically. *By default, Stata reports 95% confidence intervals $\widehat{C}$ for each estimated regression coefficient using the same formula.*

**Theorem 5.9.**

In the normal regression model, the confidence interval constructed in (5.8), i.e.,

$$\widehat{C} = \left[\widehat{\overrightarrow{\beta}} - c \times s\left(\widehat{\overrightarrow{\beta}}\right), \widehat{\overrightarrow{\beta}} + c \times s\left(\widehat{\overrightarrow{\beta}}\right)\right]$$

with $c = F^{-1}(1-\alpha/2)$ has coverage probability $\mathbb{P}\left[\overrightarrow{\beta} \in \widehat{C}\right] = 1-\alpha$.

When the degree of freedom is large the distinction between the student $t$ and the normal distribution is negligible. In particular, for $n-k \geq 61$, we have $c \leq 2.00$ for a 95% interval. Using this value we obtain the most commonly used confidence interval in applied econometric practice:

$$\widehat{C} = \left[\widehat{\overrightarrow{\beta}} - 2s\left(\widehat{\overrightarrow{\beta}}\right), \widehat{\overrightarrow{\beta}} + 2s\left(\widehat{\overrightarrow{\beta}}\right)\right] \tag{5.10}$$

This is a useful rule-of-thumb.

**Theorem 5.10.**

In the normal regression model, if $n-k \geq 61$, then (5.10) has coverage probability $\mathbb{P}\left[\overrightarrow{\beta} \in \widehat{C}\right] \geq 0.95$.

## 5.11. Confidence Intervals for Error Variance

We can also construct a confidence interval for the regression error variance $\sigma^2$ using the sampling distribution of $s^2$ from Theorem 5.7. This states that in the normal regression model

$$\frac{(n-k)\,s^2}{\sigma^2} \sim \chi^2_{n-k}. \tag{5.11}$$

Let $F(u)$ denote the $\chi^2_{n-k}$ distribution function and for some $\alpha$ set $c_1 = F^{-1}(\alpha/2)$ and $c_2 = F^{-1}(1-\alpha/2)$. Equation (5.11) implies that

$$\mathbb{P}\left[c_1 \leq \frac{(n-k)\,s^2}{\sigma^2} \leq c_2\right] = 1-\alpha.$$

Rewriting the inequalities we find

$$\mathbb{P}\left[\frac{(n-k)\,s^2}{c_1} \leq \sigma^2 \leq \frac{(n-k)\,s^2}{c_2}\right] = 1-\alpha.$$

This shows that an exact $1-\alpha$ confidence interval for $\sigma^2$ is

$$\widehat{C} = \left[\frac{(n-k)\,s^2}{c_1}, \frac{(n-k)\,s^2}{c_2}\right]. \tag{5.12}$$

**Theorem 5.11.**

In the normal regression model, (5.12) has coverage probability $\mathbb{P}\left[\sigma^2 \in \widehat{C}\right] = 1-\alpha$.

## 5.12. *t* Test

A typical goal in an econometric exercise is to assess whether or not a coefficient $\overrightarrow{\beta}$ equals a specific value $\overrightarrow{\beta}_0$. Often the specific value to be tested is $\overrightarrow{\beta}_0 = \overrightarrow{0}$, but this is not essential. This is called *hypothesis testing*. In this section and the following we give a short introduction specific to the normal regression model.

For simplicity write the coefficient to be tested as $\beta$. The null hypothesis is

$$\mathbb{H}_0 : \beta = \beta_0. \tag{5.13}$$

This states that the hypothesis is that the true value of $\beta$ equals to the hypothesized value $\beta_0$.

The alternative hypothesis is the complement of $\mathbb{H}_0$, and is written as

$$\mathbb{H}_1 : \beta \neq \beta_0.$$

This states that the true value of $\beta_0$ does not equal to the hypothesized value.

The standard statistic to test $\mathbb{H}_0$ against $\mathbb{H}_1$ is the absolute value of the $t$-statistic

$$|T| = \left| \frac{\widehat{\beta} - \beta}{s\left(\widehat{\beta}\right)} \right|. \tag{5.14}$$

If $\mathbb{H}_0$ is true then we expect $|T|$ to be small, but if $\mathbb{H}_1$ is true then we expect $|T|$ to be large. Hence the standard rule is to reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ for large values of the $t$-statistic $|T|$ and otherwise fail to reject $\mathbb{H}_0$. Thus, the hypothesis test takes the form

$$\text{Reject } \mathbb{H}_0 \text{ if } |T| > c.$$

The constant $c$ which appears in the statement of the test is called the *critical value*. Its value is selected to control the probability of false rejections. When the null hypothesis is true $T$ has an exact $t_{n-k}$ distribution in the normal regression model. Thus for a given value of $c$ the probability of false rejection is

$$\begin{aligned} \mathbb{P}\left[\text{ Reject } \mathbb{H}_0 \mid \mathbb{H}_0\right] &= \mathbb{P}\left[|T| > c \mid \mathbb{H}_0\right] \\ &= \mathbb{P}\left[T > c \mid \mathbb{H}_0\right] + \mathbb{P}\left[T < -c \mid \mathbb{H}_0\right] \\ &= 1 - F(c) + F(-c) \\ &= 2(1 - F(c)), \end{aligned}$$

where $F(u)$ is the $t_{n-k}$ distribution function. *This is the probability of false rejection and is decreasing in the critical value c.* We select the value $c$ so that this probability equals a pre-selected value called the *significance level* which is typically written as $\alpha$. It is conventional to set $\alpha = 0.05$, though this is not a hard rule. We then select $c$ so that $F(c) = 1 - \alpha/2$, which means that $c$ is the $1 - \alpha/2$ quantile of the $t_{n-k}$ distribution, the same as used for confidence intervals. With this choice the decision rule "Reject $\mathbb{H}_0$ if $|T| > c$" has a significance level (false rejection probability) of $\alpha$.

**Theorem 5.12.**

In the normal regression model if the null hypothesis (5.13) is true, then for $|T|$ defined in (5.14) $T \sim t_{n-k}$. If $c$ is set so that $\mathbb{P}\left[|t_{n-k}| \geq c\right] = \alpha$ then the test "Reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $|T| > c$" has significance level $\alpha$.

To report the result of a hypothesis test we need to pre-determine the significance level $\alpha$ in order to calculate the critical value c. This can be inconvenient and arbitrary. In general, when a test takes the form "Reject $\mathbb{H}_0$ if $S > c$" and $S$ has null distribution $G(u)$ then the $p$-value of the test is $p = 1 - G(c)$. A simplification is to report what is known as the *p-value* of the test. It is sufficient to report the $p$-value $p$ and *we can interpret the value of p as indexing the test's strength of rejection of the null hypothesis*. Thus a $p$-value of 0.07 might be interpreted as "nearly significant", 0.05 as "borderline significant", and 0.001 as "highly significant". In the context of the normal regression model the $p$-value of a $t$-statistic $|T|$ is $p = 2\left(1 - F_{n-k}(|T|)\right)$ where $F_{n-k}$ is the $t_{n-k}$ CDF.

*A p-value reports the strength of evidence against $\mathbb{H}_0$ but is not itself a probability.* A common mistake understanding is that the $p$-value is the "probability that the null hypothesis is true." This is incorrect interpretation. It is a static, is random, and is a measure of the evidence against $\mathbb{H}_0$. Nothing more.

## 5.13. Likelihood Ratio Test

In the previous section we described the $t$-test as the standard method to test a hypothesis on a single coefficient in a regression. In many contexts, however, we want to simultaneously assess a set of coefficients. In the normal regression model, this can be done by an $F$ test which can be derived from the likelihood ratio test.

Partition the regressors as $\overrightarrow{X} = \left(\overrightarrow{X}'_1, \overrightarrow{X}'_2\right)'$ and similarly partition the coefficient vector as $\overrightarrow{\beta} = \left(\overrightarrow{\beta}'_1, \overrightarrow{\beta}'_2\right)'$. The regression model can be written as

$$Y = \overrightarrow{X}_1 \overrightarrow{\beta}_1 + \overrightarrow{X}_2 \overrightarrow{\beta}_2 + e. \tag{5.15}$$

Let $k = \dim\left(\overrightarrow{X}\right), k_1 = \dim\left(\overrightarrow{X}_1\right), q = \dim\left(\overrightarrow{X}_2\right)$, so that $k = k_1 + q$. Partition the variables so that the hypothesis is that the second set of coefficients are zero, or

$$\mathbb{H}_0 : \overrightarrow{\beta}_2 = \overrightarrow{0}. \tag{5.16}$$

If $\mathbb{H}_0$ is true then the regressors $\overrightarrow{X}_2$ can be omitted from the regression. In this case we can write (5.15) as

$$Y = \overrightarrow{X}_1 \overrightarrow{\beta}_1 + e. \tag{5.17}$$

We call (5.17) the *null model*. The alternative hypothesis is that at least one element of $\overrightarrow{\beta}_2$ is non-zero and is written as $\mathbb{H}_1 : \overrightarrow{\beta}_2 \neq \overrightarrow{0}$.

When models are estimated by maximum likelihood a well-accepted testing procedure is to reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ for large values of the *Likelihood Ratio* – the ratio of the maximized likelihood function under $\mathbb{H}_1$ and $\mathbb{H}_0$, respectively. We now construct this statistic in the normal regression model. Recall from (5.6) that the maximized log-likelihood equals

$$\ell_n\left(\widehat{\overrightarrow{\beta}}_{\text{mle}}, \widehat{\sigma}^2_{\text{mle}}\right) = -\frac{n}{2}\log\left(2\pi\widehat{\sigma}^2_{\text{mle}}\right) - \frac{n}{2}.$$

We similarly calculate the maximized log-likelihood for the constrained model (5.17). By the same steps for derivation of the unconstrained MLE we find that the MLE for (5.17) is OLS of $Y$ on $\overrightarrow{X}_1$. We can write the estimator as

$$\widetilde{\overrightarrow{\beta}}_1 = \left(\overrightarrow{X}_1'\overrightarrow{X}_1\right)^{-1}\overrightarrow{X}_1'Y,$$

with residual $\widetilde{e}_i = Y_i - \overrightarrow{X}_1'\widetilde{\overrightarrow{\beta}}_1$ and the error variance estimator $\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \widetilde{e}_i^2$. You can calculate similar to (5.6) that the maximized constrained log-likelihood is

$$\ell_n\left(\widetilde{\overrightarrow{\beta}}_1, \widetilde{\sigma}^2\right) = -\frac{n}{2}\log\left(2\pi\widetilde{\sigma}^2\right) - \frac{n}{2}.$$

A classic testing procedure is to reject $\mathbb{H}_0$ for large values of the ratio of the maximized likelihoods. Equivalently the test rejects $\mathbb{H}_0$ for large values of the twice the difference in the log-likelihood functions. (Multiplying the likelihood difference by two turns out to be a useful scaling.) This equals

$$\begin{aligned} LR &= 2\left(\ell_n\left(\widehat{\overrightarrow{\beta}}_1, \widehat{\sigma}^2\right) - \ell_n\left(\widetilde{\overrightarrow{\beta}}, \widetilde{\sigma}^2\right)\right) \\ &= n\log\left(\frac{\widetilde{\sigma}^2}{\widehat{\sigma}^2}\right). \end{aligned} \tag{5.18}$$

The likelihood ratio test rejects $\mathbb{H}_0$ for large values of $LR$, or equivalently for large values of[1]

$$F = \frac{\left(\widetilde{\sigma}^2 - \widehat{\sigma}^2\right)/q}{\widehat{\sigma}^2/(n-k)}. \tag{5.19}$$

This is known as the $F$ statistic for the test of hypothesis $\mathbb{H}_0$ against $\mathbb{H}_1$.

To develop an appropriate critical value we need the null distribution of $F$. Recall from (3.28) that $n\widehat{\sigma}^2 = \overrightarrow{e}'\boldsymbol{M}\overrightarrow{e}$ where $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{P}$ with $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$. Similarly, under $\mathbb{H}_0$, $n\widetilde{\sigma}^2 = \overrightarrow{e}'\boldsymbol{M}_1\overrightarrow{e}$ where $\boldsymbol{M}_1 = \boldsymbol{I}_n - \boldsymbol{P}_1$ with $\boldsymbol{P}_1 = \boldsymbol{X}_1\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'$. You can calculate that $\boldsymbol{M}_1 - \boldsymbol{M} = \boldsymbol{P} - \boldsymbol{P}_1$ is idempotent with rank $q$. Furthermore, $\left(\boldsymbol{M}_1 - \boldsymbol{M}\right)\boldsymbol{M} = \boldsymbol{0}$. It follows that

$$\overrightarrow{e}'\left(\boldsymbol{M}_1 - \boldsymbol{M}\right)\overrightarrow{e} \sim \chi^2_q$$

[1] There is a one-to-one mapping between $LR$ and $F$:

$$F = \left[\exp\left(\frac{LR}{n}\right) - 1\right] \times \frac{n-k}{q}.$$

and is independent of $\vec{e}\,'\boldsymbol{M}\vec{e}$. Hence

$$F = \frac{\vec{e}\,'\left(\boldsymbol{M}_1 - \boldsymbol{M}\right)\vec{e}/q}{\vec{e}\,'\boldsymbol{M}\vec{e}/(n-k)} \sim \frac{\chi^2_q/q}{\chi^2_{n-k}/(n-k)} \sim F_{q,n-k},$$

an exact $F$ distribution with degrees of freedom $q$ and $n-k$, respectively. Thus under $\mathbb{H}_0$, the $F$ statistic has an exact $F_{q,n-k}$ distribution.

The critical values are selected from the upper tail of the $F$ distribution. For a given significance level $\alpha$ (typically $\alpha = 0.05$) we select the critical value $c$ so that

$$\mathbb{P}\left[F_{q,n-k} \geq c\right] = \alpha.$$

The test rejects $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $F > c$ and does not reject $\mathbb{H}_0$ otherwise. The p-value of the test is $p = 1 - G_{q,n-k}(F))$ where $G_{q,n-k}(\cdot)$ is the $F_{q,n-k}$ distribution function.

**Theorem 5.13.**

In the normal regression model if the null hypothesis (5.16) is true then for $F$ defined in (5.19), i.e.,

$$F = \frac{\left(\tilde{\sigma}^2 - \hat{\sigma}^2\right)/q}{\hat{\sigma}^2/(n-k)}.$$

$F \sim F_{q,n-k}$. If $c$ is set so that $\mathbb{P}\left[F_{q,n-k} \geq c\right] = \alpha$ then the test "Reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $F > c$" has significance level $\alpha$.

*The derivation of the F distribution?*

## 5.14. Information Bound for Normal Regression

The likelihood scores for the normal regression model are

$$\frac{\partial}{\partial \vec{\beta}}\ell_n\left(\vec{\beta},\sigma^2\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\vec{X}_i\left(Y_i - \vec{X}_i'\vec{\beta}\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\vec{X}_i\vec{e}_i$$

and

$$\frac{\partial}{\partial \sigma^2}\ell_n\left(\vec{\beta},\sigma^2\right) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(Y_i - \vec{X}_i'\vec{\beta}\right)^2 = \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(e_i^2 - \sigma^2\right)$$

It follows that the information matrix is

$$\mathscr{I} = \operatorname{var}\left[\begin{array}{c|c} \frac{\partial}{\partial \vec{\beta}}\ell\left(\vec{\beta},\sigma^2\right) & \boldsymbol{X} \\ \frac{\partial}{\partial \sigma^2}\ell\left(\vec{\beta},\sigma^2\right) & \boldsymbol{X} \end{array}\right] = \begin{pmatrix} \frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

The Cramér-Rao Lower Bound is

$$\mathscr{I}^{-1} = \begin{pmatrix} \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

This shows that the lower bound for estimation of $\overrightarrow{\beta}$ is $\sigma_2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ and the lower bound for $\sigma_2$ is $2\sigma^4/n$.

Since in the homoskedastic linear regression model the OLS estimator is unbiased and has variance $\sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$, *it follows that the OLS coefficient estimator $\widehat{\overrightarrow{\beta}}$ is Cramér-Rao efficient in the normal regression model*. *Cramér-Rao efficiency means that no unbiased estimator has a lower covariance matrix.* This expands on the Gauss-Markov theorem which stated that no linear unbiased estimator has a lower variance matrix in the homoskedastic regression model. Notice that that the results are complementary. *Gauss-Markov efficiency concerns a more narrow class of estimators (linear) but allows a broader model class (linear homoskedastic rather than normal regression). The Cramér-Rao efficiency result is more powerful in that it does not restrict the class of estimators (beyond unbiasedness) but is more restrictive in the class of models allowed (normal regression).*

The unbiased variance estimator $s^2$ of $\sigma^2$ has variance $2\sigma^4/(n-k)$, which is larger than the Cramér-Rao lower bound $2\sigma^4/n$. Thus in contrast to the coefficient estimator, the variance estimator is not Cramér-Rao efficient.

# Chapter 7

# Asymptotic Theory for Least Squares

## 7.1. Introduction

It turns out that the asymptotic theory of least squares estimation applies equally to the projection model and the linear CEF model. Therefore the results in this chapter will be stated for the broader projection model described in Section 2.18. Recall that the model is $Y = \overrightarrow{X}'\overrightarrow{\beta} + e$ with the linear projection coefficient $\overrightarrow{\beta} = \left(\mathbb{E}\left[\overrightarrow{X}\overrightarrow{X}'\right]\right)^{-1}\mathbb{E}\left[\overrightarrow{X}Y\right]$.

Throughout the chapter, the following assumptions are made.

**Assumption 7.1. Random Sampling and Finite Second Moments**

(1) The variables $\left(Y_i, \overrightarrow{X}_i\right), i = 1, \ldots, n$ are iid.

(2) $\mathbb{E}\left[Y^2\right] < \infty$.

(3) $\mathbb{E}\left[\|\overrightarrow{X}\|^2\right] < \infty$.

(4) $\boldsymbol{Q}_{XX} = \mathbb{E}\left[\overrightarrow{X}\overrightarrow{X}'\right]$ is positive definite.

The distributional results will require a strengthening of these assumptions to finite fourth moments. We discuss the specific conditions in Section 7.3.

## 7.2. Consistency of Least Squares Estimator

In this section we use the weak law of large numbers (WLLN) and continuous mapping theorem (CMT) to show that the least squares estimator $\widehat{\overrightarrow{\beta}}$ is consistent for the projection coefficient $\overrightarrow{\beta}$.

This derivation is based on three key components. First, the OLS estimator can be written as a continuous function of a set of sample moments. Second, the WLLN shows that sample moments converge in probability to population moments. And third, the CMT states that continuous functions preserve convergence in probability. We now explain each step in brief and then in greater detail.

First, observe that the OLS estimator

$$\widehat{\vec{\beta}} = \left(\frac{1}{n}\sum_{i=1}^{n}\vec{X}_i\vec{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\vec{X}_iY_i\right) = \widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\vec{Q}}_{XY}$$

is a function of the sample moments

$$\widehat{\boldsymbol{Q}}_{XX} = \frac{1}{n}\sum_{i=1}^{n}\vec{X}_i\vec{X}_i'$$

and

$$\widehat{\vec{Q}}_{XY} = \frac{1}{n}\sum_{i=1}^{n}\vec{X}_iY_i.$$

Second, by an application of the WLLN these sample moments converge in probability to their population expectations. *Specifically, the fact that $(Y_i, \vec{X}_i)$ are mutually i.i.d. implies that any function of $(Y_i, \vec{X}_i)$ is i.i.d., including $\vec{X}_iY_i$ and $\vec{X}_i\vec{X}_i'$. These variables also have finite expectations under Assumption 7.1.* Under these conditions, the WLLN implies that as $n \to \infty$,

$$\widehat{\boldsymbol{Q}}_{XX} = \frac{1}{n}\sum_{i=1}^{n}\vec{X}_i\vec{X}_i' \xrightarrow{p} \mathbb{E}\left[\vec{X}_1\vec{X}_1'\right] = \boldsymbol{Q}_{XX} \tag{7.1}$$

and

$$\widehat{\vec{Q}}_{XY} = \frac{1}{n}\sum_{i=1}^{n}\vec{X}_iY_i' \xrightarrow{p} \mathbb{E}\left[\vec{X}_1Y_1'\right] = \vec{Q}_{XY}.$$

Third, the CMT allows us to combine these equations to show that $\widehat{\vec{\beta}}$ converges in probability to $\vec{\beta}$. Specifically, as $n \to \infty$,

$$\widehat{\vec{\beta}} = \widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\vec{Q}}_{XY} \xrightarrow{p} \boldsymbol{Q}_{XX}^{-1}\vec{Q}_{XY} = \vec{\beta}. \tag{7.2}$$

To fully understand the application of the CMT we walk through it in detail. We can write

$$\widehat{\vec{\beta}} = g\left(\widehat{\boldsymbol{Q}}_{XX}, \widehat{\vec{Q}}_{XY}\right),$$

where $g(\boldsymbol{A}, \vec{b}) = \boldsymbol{A}^{-1}\vec{b}$ is a function of $\boldsymbol{A}$ and $\vec{b}$. The function $g(\boldsymbol{A}, \vec{b})$ is a continuous function of $\boldsymbol{A}$ and $\vec{b}$ at all values of the arguments such that $\boldsymbol{A}^{-1}$ exists. Assumption 7.1 specifies that $\boldsymbol{Q}_{XX}$

is positive definite, which means that $\boldsymbol{Q}_{XX}^{-1}$ exists. Thus $g(\boldsymbol{A}, \overrightarrow{b})$ is continuous at $\boldsymbol{A} = \boldsymbol{Q}_{XX}$. This justifies the application of the CMT in (7.2).

For a slightly different demonstration of (7.2) recall that (4.6) implies that

$$\widehat{\overrightarrow{\beta}} - \overrightarrow{\beta} = \widehat{\boldsymbol{Q}}_{XX}^{-1} \widehat{\overrightarrow{Q}}_{Xe} \tag{7.3}$$

where

$$\widehat{\overrightarrow{Q}}_{Xe} = \frac{1}{n} \sum_{i=1}^{n} \overrightarrow{X}_i e_i.$$

The WLLN and (2.25) imply

$$\widehat{\overrightarrow{Q}}_{Xe} \xrightarrow{p} \mathbb{E}\left[\overrightarrow{X} e\right] = \overrightarrow{0}.$$

Therefore,

$$\widehat{\overrightarrow{\beta}} - \overrightarrow{\beta} = \widehat{\boldsymbol{Q}}_{XX}^{-1} \widehat{\overrightarrow{Q}}_{Xe} \xrightarrow{p} \widehat{\boldsymbol{Q}}_{XX}^{-1} \overrightarrow{0} = \overrightarrow{0},$$

which is the same as $\widehat{\overrightarrow{\beta}} \xrightarrow{p} \overrightarrow{\beta}$.

**Theorem 7.1. Consistency of Least squares**

Under Assumption 7.1,

$$\widehat{\boldsymbol{Q}}_{XX} \xrightarrow{p} \boldsymbol{Q}_{XX},$$

$$\widehat{\overrightarrow{Q}}_{XY} \xrightarrow{p} \overrightarrow{Q}_{XY},$$

$$\widehat{\boldsymbol{Q}}_{XX}^{-1} \xrightarrow{p} \boldsymbol{Q}_{XX}^{-1},$$

$$\widehat{\overrightarrow{Q}}_{Xe} \xrightarrow{p} \overrightarrow{0},$$

and

$$\widehat{\overrightarrow{\beta}} \xrightarrow{p} \overrightarrow{\beta}$$

as $n \longrightarrow \infty$.

In the stochastic order notation, Theorem 7.1 can be equivalently written as

$$\widehat{\overrightarrow{\beta}} = \beta + o_p(1). \tag{7.4}$$

## 7.3. Asymptotic Normality

We started this chapter discussin the need for an approximation to the distribution of the OLS estimator $\widehat{\overrightarrow{\beta}}$. In Section 7.2 we showed that $\widehat{\overrightarrow{\beta}}$ converges in probability to $\overrightarrow{\beta}$. Consistency is a good first step, but in itself does not describe the distribution of the estimator. In this section we derive an approximation typically called the *asymptotic distribution*.

*The derivation starts by writing the estimator as a function of sample moments. One of the moments must be written as a sum of zero-mean random vectors and normalized so that the central limit theorem can be applied.* The steps are as follows.

Take equation (7.3) and multiply it by $\sqrt{n}$. This yields the expression

$$\sqrt{n}\left(\widehat{\vec{\beta}}-\vec{\beta}\right)=\left(\frac{1}{n}\sum_{i=1}^{n}\vec{X}_i\vec{X}_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\vec{X}_ie_i\right). \tag{7.5}$$

This shows that the normalized and centered estimator $\sqrt{n}\left(\widehat{\vec{\beta}}-\vec{\beta}\right)$ is a function of the sample average $n^{-1}\sum_{i=1}^{n}\vec{X}_i\vec{X}_i'$ and the normalized sample average $n^{-1/2}\sum_{i=1}^{n}\vec{X}_ie_i$.

The random pairs $(Y_i,\vec{X}_i)$ are i.i.d., meaning that they are independent across $i$ and identically distributed. Any function of $(Y_i,\vec{X}_i)$ is also i.i.d.. This includes $e_i=Y_i-\vec{X}_i'\vec{\beta}$ and the product $\vec{X}_ie_i$. The latter is mean zero ($\mathbb{E}\left[\vec{X}e\right]=0$) and has $k\times k$ covariance matrix

$$\mathbf{\Omega}=\mathbb{E}\left[(\vec{X}e)(\vec{X}e)'\right]=\mathbb{E}\left[\vec{X}\vec{X}'e^2\right].$$

We show below that $\mathbf{\Omega}$ has finite elements under a strengthening of Assumption 7.1. Since $\vec{X}_ie_i$ is i.i.d., mean zero, and finite variance, the central limit theorem implies

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\vec{X}_ie_i\Rightarrow N(0,\mathbf{\Omega}).$$

We state the required conditions here.

**Assumption 7.2. Random Sampling and Finite Fourth Moments**

(1) The variables $(Y_i,\vec{X}_i), i=1,\ldots,n$ are iid.

(2) $\mathbb{E}\left[Y^4\right]<\infty$.

(3) $\mathbb{E}\left[\|\vec{X}\|^4\right]<\infty$.

(4) $\boldsymbol{Q}_{XX}=\mathbb{E}\left[\vec{X}\vec{X}'\right]$ is positive definite.

Assumption 7.2 implies that $\mathbf{\Omega}<\infty$. To see this, take the $j\ell$th element of $\mathbf{\Omega}$, $\mathbb{E}\left[\vec{X}_j\vec{X}'_\ell e^2\right]$. *First, Theorem 2.9(6) shows that* $\mathbb{E}\left[e^4\right]<\infty$. By the expectation inequality the $j\ell$th element of $\mathbf{\Omega}$ is bounded by

$$\left|\mathbb{E}\left[\vec{X}_j\vec{X}'_\ell e^2\right]\right|\le\mathbb{E}\left[\left|\vec{X}_j\vec{X}'_\ell e^2\right|\right]=\mathbb{E}\left[\left|\vec{X}_j\right|\left|\vec{X}_\ell\right|e^2\right].$$

By two applications of the *Cauchy-Schwarz inequality* this is smaller than

$$\left(\mathbb{E}\left[\left|\vec{X}_j\right|^2\left|\vec{X}_\ell\right|^2\right]\right)^{1/2}\left(\mathbb{E}\left[e^4\right]\right)^{1/2}\le\left(\mathbb{E}\left[\left|\vec{X}_j\right|^4\right]\right)^{1/4}\left(\mathbb{E}\left[\left|\vec{X}_\ell\right|^4\right]\right)^{1/4}\left(\mathbb{E}\left[e^4\right]\right)^{1/2}<\infty$$

where the finiteness holds under Assumption 7.2. Thus $\Omega < \infty$.

An alternative way to show that the elements of $\boldsymbol{\Omega}$ are finite is by using a matrix norm $\|\cdot\|$. Then by the expectation inequality, the Cauchy-Schwarz inequality, Assumption 7.2, and $\mathbb{E}\left[e^4\right] < \infty$,

$$\|\boldsymbol{\Omega}\| \le \mathbb{E}\left[\left\|\overrightarrow{X}\overrightarrow{X}'e^2\right\|\right] = \mathbb{E}\left[\|\overrightarrow{X}\|^2 e^2\right] \le \left(\mathbb{E}\left[\|\overrightarrow{X}\|^4\right]\right)^{1/2}\left(\mathbb{E}\left[e^4\right]\right)^{1/2} < \infty.$$

This is a more compact argument (often described as more elegant) but such manipulations should not be done without understanding the notation and the applicability of each step of the argument.

**Theorem 7.2.**

Assumption 7.2 implies that

$$\boldsymbol{\Omega} = \mathbb{E}\left[\overrightarrow{X}\overrightarrow{X}'e^2\right] < \infty \tag{7.6}$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\overrightarrow{X}_i e_i \Rightarrow N(\overrightarrow{0}, \boldsymbol{\Omega}) \tag{7.7}$$

as $n \to \infty$.

Putting together (7.1), (7.5) and (7.7),

$$\sqrt{n}\left(\widehat{\overrightarrow{\beta}} - \overrightarrow{\beta}\right) \Rightarrow \boldsymbol{Q}_{XX}^{-1}N(0, \boldsymbol{\Omega}) = N\left(\overrightarrow{0}, \boldsymbol{Q}_{XX}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{XX}^{-1}\right)$$

as $n \to \infty$. The final equality follows from the property that linear combinations of normal vectors are also normal (Theorem 5.2).

**Theorem 7.3. Asymptotic Normality of Least Squares Estimator**

Under Assumption 7.2, as $n \to \infty$

$$\sqrt{n}\left(\widehat{\overrightarrow{\beta}} - \overrightarrow{\beta}\right) \Rightarrow N(0, \boldsymbol{V}_{\overrightarrow{\beta}})$$

where $\boldsymbol{Q}_{XX} = \mathbb{E}\left[\overrightarrow{X}\overrightarrow{X}'\right]$, $\boldsymbol{\Omega} = \mathbb{E}\left[\overrightarrow{X}\overrightarrow{X}'e^2\right]$, and

$$\boldsymbol{V}_{\overrightarrow{\beta}} = \boldsymbol{Q}_{XX}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{XX}^{-1}. \tag{7.8}$$

In the stochastic order notation, Theorem 7.3 implies that $\widehat{\overrightarrow{\beta}} = \overrightarrow{\beta} + O_p\left(n^{-1/2}\right)$ which is strong than (7.4).

The matrix $\boldsymbol{V}_{\overrightarrow{\beta}} = \boldsymbol{Q}_{XX}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{XX}^{-1}$ is the variance of the asymptotic distribution of $\sqrt{n}\left(\widehat{\overrightarrow{\beta}} - \overrightarrow{\beta}\right)$. Consequently, $\boldsymbol{V}_{\overrightarrow{\beta}}$ is often refereed to as the *asymptotic covariance matrix* of $\widehat{\overrightarrow{\beta}}$. The expression

$\boldsymbol{V}_{\overrightarrow{\beta}} = \boldsymbol{Q}_{XX}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{XX}^{-1}$ is called a *sandwich* form as the matrix $\boldsymbol{\Omega}$ is sandwiched between two copies of $\boldsymbol{Q}_{XX}^{-1}$.

It is useful to compare the variance of the asymptotic distribution given in (7.8) and the finite-sample conditional variance in the CEF model as given in (4.10):

$$\boldsymbol{V}_{\widehat{\overrightarrow{\beta}}} = \operatorname{var}\left[\widehat{\overrightarrow{\beta}} \mid \boldsymbol{X}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{7.9}$$

*Notice that $\boldsymbol{V}_{\widehat{\overrightarrow{\beta}}}$ is the exact conditional variance of $\widehat{\overrightarrow{\beta}}$ and $\boldsymbol{V}_{\overrightarrow{\beta}}$ is the asymptotic variance of $\sqrt{n}\left(\widehat{\overrightarrow{\beta}} - \overrightarrow{\beta}\right)$.* Then $\boldsymbol{V}_{\overrightarrow{\beta}}$ should be (roughly) $n$ times as large as $\boldsymbol{V}_{\widehat{\overrightarrow{\beta}}}$. Indeed, multiplying (7.9) by $n$ and distributing we find

$$n\boldsymbol{V}_{\widehat{\overrightarrow{\beta}}} = \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

which looks like an estimator of $\boldsymbol{V}_{\overrightarrow{\beta}}$. Indeed, as $n \longrightarrow \infty$, $n\boldsymbol{V}_{\widehat{\overrightarrow{\beta}}} \xrightarrow{p} \boldsymbol{V}_{\overrightarrow{\beta}}$. *The expression $\boldsymbol{V}_{\widehat{\overrightarrow{\beta}}}$ is useful for practical inference (such as computation of standard errors and tests) since it is the variance of the estimator $\overrightarrow{\beta}$, while $\boldsymbol{V}_{\overrightarrow{\beta}}$ is useful for asymptotic theory as it is well defined in the limit as n goes to infinity.* We will make use of both symbols and it will be advisable to adhere to this convention.

There is a special case when $\boldsymbol{\Omega}$ and $\boldsymbol{V}_{\overrightarrow{\beta}}$ simplify. Suppose that

$$\operatorname{cov}\left(\overrightarrow{X}\overrightarrow{X}', e^2\right) = 0. \tag{7.10}$$

Condition (7.10) holds in the homoskedastic linear regression model but it is somewhat broader. Under (7.10) the asymptotic variance formulae simplifies as

$$\boldsymbol{\Omega} = \mathbb{E}\left[\overrightarrow{X}\overrightarrow{X}'\right]\mathbb{E}\left[e^2\right] = \boldsymbol{Q}_{XX}\sigma^2$$

$$\boldsymbol{V}_{\overrightarrow{\beta}} = \boldsymbol{Q}_{XX}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{XX}^{-1} = \boldsymbol{Q}_{XX}^{-1}\sigma^2 =: \boldsymbol{V}_{\overrightarrow{\beta}}^{0}. \tag{7.11}$$

In (7.11) we define $\boldsymbol{V}_{\overrightarrow{\beta}}^{0} := \boldsymbol{Q}_{XX}^{-1}\sigma^2$ whether (7.10) is true or false. When (7.10) is true then $\boldsymbol{V}_{\overrightarrow{\beta}} = \boldsymbol{V}_{\overrightarrow{\beta}}^{0}$, otherwise $\boldsymbol{V}_{\overrightarrow{\beta}} \neq \boldsymbol{V}_{\overrightarrow{\beta}}^{0}$. We call $\boldsymbol{V}_{\overrightarrow{\beta}}^{0}$ the *homoskedastic asymptotic covariance matrix*.

Theorem 7.3 states that the sampling distribution of the least squares estimator, after rescaling, is approximately normal when the sample size $n$ is sufficiently large. This holds true for all joint distributions of $(Y, \overrightarrow{X})$ which satisfy the conditions of Assumption 7.2. Consequently, asymptotic normality is routinely used to approximate the finite sample distribution of $\sqrt{n}\left(\widehat{\overrightarrow{\beta}} - \overrightarrow{\beta}\right)$.

A difficulty is that for any fixed $n$ the sampling distribution of $\widehat{\overrightarrow{\beta}}$ can be arbitrarily far from the normal distribution. The normal approximation improves as $n$ increases, but how large should $n$ be in order for the approximation to be useful? Unfortunately, there is no simple answer to this reasonable question. *The trouble is that no matter how large is the sample size the normal approximation is arbitrarily poor for some data distribution satisfying the assumptions.*

## 7.4. Joint Distribution

Theorem 7.3 gives the point asymptotic distribution of the coefficient estimators. We can use the result to study the covariance between the coefficient estimators. For simplicity, take the case of two regressors, no intercept, and homoskedastic error. Assume the regressors are mean zero, variance one, with correlation $\rho$. Then using the formula for inversion of a $2 \times 2$ matrix,

$$\boldsymbol{V}^0_{\vec{\beta}} = \sigma^2 \boldsymbol{Q}^{-1}_{XX} = \frac{\sigma^2}{1-\rho^2} \left[ \begin{array}{cc} 1 & -\rho \\ -\rho & 1 \end{array} \right].$$

Thus if $X_1$ and $X_2$ are positively correlated ($\rho > 0$), then $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are negatively correlated (and vice-versa).

This finding that the correlation of the regressors is of opposite sign of the correlation of the coefficient estimates is sensitive to the assumption of homoskedasticity. If the errors are heteroskedastic then this relationship is not guaranteed.

We can extend the above analysis to study the covariance between coefficient sub-vectors. For example, partitioning $\boldsymbol{X}' = \left(\boldsymbol{X}_1', \boldsymbol{X}_2'\right)'$ and $\vec{\beta}' = \left(\vec{\beta}_1', \vec{\beta}_2'\right)'$, we can write the general model as

$$Y = \boldsymbol{X}_1' \vec{\beta}_1 + \boldsymbol{X}_2' \vec{\beta}_2 + e$$

and the coefficient estimates as $\widehat{\vec{\beta}}' = \left(\widehat{\vec{\beta}}_1', \widehat{\vec{\beta}}_2'\right)'$. Make the partitions

$$\boldsymbol{Q}_{XX} = \left[ \begin{array}{cc} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{array} \right], \quad \boldsymbol{\Omega} = \left[ \begin{array}{cc} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{array} \right].$$

We know

$$\boldsymbol{Q}^{-1}_{XX} = \left[ \begin{array}{cc} \boldsymbol{Q}^{-1}_{11\cdot 2} & -\boldsymbol{Q}^{-1}_{11\cdot 2}\boldsymbol{Q}_{12}\boldsymbol{Q}^{-1}_{22} \\ -\boldsymbol{Q}^{-1}_{22\cdot 1}\mathrm{Q}_{21}\boldsymbol{Q}^{-1}_{11} & \boldsymbol{Q}^{-1}_{22\cdot 1} \end{array} \right]$$

where $\boldsymbol{Q}_{11\cdot 2} = \boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}^{-1}_{22}\boldsymbol{Q}_{21}$ and $\boldsymbol{Q}_{22\cdot 1} = \boldsymbol{Q}_{22} - \boldsymbol{Q}_{21}\boldsymbol{Q}^{-1}_{11}\boldsymbol{Q}_{12}$. Thus, when the error is homoskedastic,

$$\operatorname{cov}\left(\vec{\beta}_1, \vec{\beta}_2\right) = -\sigma^2 \boldsymbol{Q}^{-1}_{11\cdot 2}\boldsymbol{Q}_{12}\boldsymbol{Q}^{-1}_{22}$$

which is a matrix generalization of the two-regressor case.

In general you can show that

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \left[ \begin{array}{cc} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{array} \right] \tag{7.13}$$

where

$$\boldsymbol{V}_{11} = \boldsymbol{Q}^{-1}_{11\cdot 2}\left(\Omega_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}^{-1}_{22}\Omega_{21} - \Omega_{12}\boldsymbol{Q}^{-1}_{22}\boldsymbol{Q}_{21} + \boldsymbol{Q}_{12}\boldsymbol{Q}^{-1}_{22}\Omega_{22}\boldsymbol{Q}^{-1}_{22}\boldsymbol{Q}_{21}\right)\boldsymbol{Q}^{-1}_{11\cdot 2} \tag{7.14}$$

$$\boldsymbol{V}_{21} = \boldsymbol{Q}^{-1}_{22\cdot 1}\left(\Omega_{21} - \boldsymbol{Q}_{21}\boldsymbol{Q}^{-1}_{11}\Omega_{11} - \Omega_{22}\boldsymbol{Q}^{-1}_{22}\boldsymbol{Q}_{21} + \boldsymbol{Q}_{21}\boldsymbol{Q}^{-1}_{11}\Omega_{12}\boldsymbol{Q}^{-1}_{22}\boldsymbol{Q}_{21}\right)\boldsymbol{Q}^{-1}_{11\cdot 2} \tag{7.15}$$

$$\boldsymbol{V}_{22} = \boldsymbol{Q}^{-1}_{22\cdot 1}\left(\Omega_{22} - \boldsymbol{Q}_{21}\boldsymbol{Q}^{-1}_{11}\Omega_{12} - \Omega_{21}\boldsymbol{Q}^{-1}_{11}\boldsymbol{Q}_{12} + \boldsymbol{Q}_{21}\boldsymbol{Q}^{-1}_{11}\Omega_{11}\boldsymbol{Q}^{-1}_{11}\boldsymbol{Q}_{12}\right)\boldsymbol{Q}^{-1}_{22\cdot 1}. \tag{7.16}$$

## 7.5. Consistency of Error Variance Estimators

Using the methods of Section 7.2, we can show that the estimators $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \widehat{e}_i^2$ and $s^2 = \frac{1}{n-k}\sum_{i=1}^n \widehat{e}_i^2$ are consistent for $\sigma^2$.

The trick is to write the residual $\widehat{e}_i^2$ as equal to the error $e_i$ plus a deviation

$$\widehat{e}_i = Y_i - \vec{X}_i'\widehat{\vec{\beta}} = e_i - \vec{X}_i'\left(\widehat{\vec{\beta}} - \vec{\beta}\right).$$

Thus the squared residual equals the squared error plus a deviation

$$\widehat{e}_i^2 = e_i^2 - 2e_i\vec{X}_i'\left(\widehat{\vec{\beta}} - \vec{\beta}\right) + \left(\widehat{\vec{\beta}} - \vec{\beta}\right)'\vec{X}_i\vec{X}_i'\left(\widehat{\vec{\beta}} - \vec{\beta}\right). \tag{7.17}$$

So when we take the average of the squared residuals we obtain the average of the squared errors, plus two terms which are (hopefully) asymptotically negligible.

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \widehat{e}_i^2 = \frac{1}{n}\sum_{i=1}^n e_i^2 - 2\left(\frac{1}{n}\sum_{i=1}^n e_i\vec{X}_i'\right)\left(\widehat{\vec{\beta}} - \vec{\beta}\right) + \left(\widehat{\vec{\beta}} - \vec{\beta}\right)'\left(\frac{1}{n}\sum_{i=1}^n \vec{X}_i\vec{X}_i'\right)\left(\widehat{\vec{\beta}} - \vec{\beta}\right). \tag{7.18}$$

Indeed, the WLLN shows that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n e_i^2 &\xrightarrow[p]{} \sigma^2 \\ \frac{1}{n}\sum_{i=1}^n e_iX_i' &\xrightarrow[p]{} \mathbb{E}\left[eX'\right] = 0 \\ \frac{1}{n}\sum_{i=1}^n X_iX_i' &\xrightarrow[p]{} \mathbb{E}\left[XX'\right] = \boldsymbol{Q}_{XX} \end{aligned}$$

Theorem 7.1 shows that $\widehat{\vec{\beta}} \xrightarrow{p} \vec{\beta}$. Hence (7.18) converges in probability to $\sigma^2$ as desired.

Finally, since $\frac{n}{n-k} \longrightarrow 1$ as $n \longrightarrow \infty$ it follows that $s^2 = \left(\frac{n}{n-k}\right)\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$. Thus both estimators are consistent.

**Theorem 7.4.**

Under Assumption 7.1, $\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$ and $s^2 \xrightarrow{p} \sigma^2$ as $n \longrightarrow \infty$.

## 7.6. Homoskedastic Covariance Matrix Estimation

Theorem 7.3 shows that $\sqrt{n}\left(\widehat{\vec{\beta}}-\vec{\beta}\right)$ is asymptotically normal with asymptotic covariance matrix $\boldsymbol{V}_{\vec{\beta}}$. For asymptotic inference (confidence intervals and tests) we need a consistent estimator of $\boldsymbol{V}_{\vec{\beta}}$. Under homoskedasticity $\boldsymbol{V}_{\vec{\beta}}$ simplifies to $\boldsymbol{V}_{\vec{\beta}}^{0}=\boldsymbol{Q}_{XX}^{-1}\sigma^{2}$ and in this section we consider the simplified problem of estimating $\boldsymbol{V}_{\vec{\beta}}^{0}$.

The standard moment estimator of $\boldsymbol{Q}_{XX}$ is $\widehat{\boldsymbol{Q}}_{XX}$ defined in (7.1) and thus an estimator for $\boldsymbol{Q}_{XX}^{-1}$ is $\widehat{\boldsymbol{Q}}_{XX}^{-1}$. The standard estimator of $\sigma^{2}$ is the unbiased estimator $s^{2}$ defined as

$$s^{2}=\frac{1}{n-k}\sum_{i=1}^{n}\widehat{e}_{i}^{2}.$$

Thus a natural plug-in estimator for $\boldsymbol{V}_{\vec{\beta}}^{0}=\boldsymbol{Q}_{XX}^{-1}\sigma^{2}$ is $\widehat{\boldsymbol{V}}_{\vec{\beta}}^{0}=\widehat{\boldsymbol{Q}}_{XX}^{-1}s^{2}$.

Consistency of $\widehat{\boldsymbol{V}}_{\vec{\beta}}^{0}$ for $\boldsymbol{V}_{\vec{\beta}}^{0}$ follows from the consistency of the moment estimators $\boldsymbol{Q}_{XX}$ and $s^{2}$ and an application of the continuous mapping theorem.

**Theorem 7.5.**

Under Assumption 7.1, $\widehat{\boldsymbol{V}}_{\vec{\beta}}^{0}\xrightarrow{p}\boldsymbol{V}_{\vec{\beta}}^{0}$ as $n\rightarrow\infty$, where $\widehat{\boldsymbol{V}}_{\vec{\beta}}^{0}$ is constructed as follows,

$$\widehat{\boldsymbol{V}}_{\vec{\beta}}^{0}=\left(\frac{1}{n}\sum_{i=1}^{n}\vec{X}_{i}\vec{X}_{i}^{\prime}\right)^{-1}\left(\frac{1}{n-k}\sum_{i=1}^{n}\widehat{e}_{i}^{2}\right).$$

*It is instructive to notice that Theorem 7.5 does not require the assumption of homoskedasticity.* That is, $\widehat{\boldsymbol{V}}_{\vec{\beta}}^{0}$ is consistent for $\boldsymbol{V}_{\vec{\beta}}^{0}$ regardless if the regression is homoskedastic or heteroskedastic. *However,* $\boldsymbol{V}_{\vec{\beta}}^{0}=\boldsymbol{V}_{\vec{\beta}}=\operatorname{avar}\left[\widehat{\vec{\beta}}\right]$ *only under homoskedasticity. Thus, in the general case* $\widehat{\boldsymbol{V}}_{\vec{\beta}}^{0}$ *is consistent for a well-defined but non-useful object.*

## 7.7. Heteroskedastic Covariance Matrix Estimation

Theorem 7.3 established that the asymptotic covariance matrix of $\sqrt{n}\left(\widehat{\vec{\beta}}-\vec{\beta}\right)$ is $\boldsymbol{V}_{\vec{\beta}}=\boldsymbol{Q}_{XX}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{XX}^{-1}$. We now consider estimation of this covariance matrix without imposing homoskedasticity. The standard approach is to use a plug-in estimator which replaces the unknowns with sample moments.

As described in the previous section a natural estimator for $\boldsymbol{Q}_{XX}^{-1}$ is $\widehat{\boldsymbol{Q}}_{XX}^{-1}$ where $\widehat{\boldsymbol{Q}}_{XX}$ is defined in (7.1).

The moment estimator for $\boldsymbol{\Omega}$ is

$$\widehat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} \overrightarrow{X}_i \overrightarrow{X}_i' \widehat{e}_i^2,$$

leading to the plug-in covariance matrix estimator

$$\widehat{\boldsymbol{V}}_{\overrightarrow{\beta}}^{\mathrm{HC0}} = \widehat{\boldsymbol{Q}}_{XX}^{-1} \widehat{\boldsymbol{\Omega}} \widehat{\boldsymbol{Q}}_{XX}^{-1}. \tag{7.19}$$

You can check that $\widehat{\boldsymbol{V}}_{\overrightarrow{\beta}}^{\mathrm{HC0}} = n\widehat{\boldsymbol{V}}_{\widehat{\overrightarrow{\beta}}}^{\mathrm{HC0}}$ where $\widehat{\boldsymbol{V}}_{\widehat{\overrightarrow{\beta}}}^{\mathrm{HC0}}$ is the HC0 covariance matrix estimator from (4.31).

As shown in Theorem 7.1, $\widehat{\boldsymbol{Q}}_{XX}^{-1} \xrightarrow{p} \boldsymbol{Q}_{XX}^{-1}$, so we just need to verify the consistency of $\widehat{\boldsymbol{\Omega}}$. The key is to replace the squared residual $\widehat{e}_i^2$ with the squared error $e_i^2$, and then show the difference is asymptotically negligible.

Specifically, observe that

$$\begin{aligned}
\widehat{\boldsymbol{\Omega}} &= \frac{1}{n}\sum_{i=1}^{n} \overrightarrow{X}_i \overrightarrow{X}_i' \widehat{e}_i^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \overrightarrow{X}_i \overrightarrow{X}_i' e_i^2 + \frac{1}{n}\sum_{i=1}^{n} \overrightarrow{X}_i \overrightarrow{X}_i' \left(\widehat{e}_i^2 - e_i^2\right).
\end{aligned}$$

The first term is an average of the i.i.d. random variables $\overrightarrow{X}_i \overrightarrow{X}_i' e_i^2$, and therefore by the WLLN converges in probability to its expectation, namely,

$$\frac{1}{n}\sum_{i=1}^{n} \overrightarrow{X}_i \overrightarrow{X}_i' e_i^2 \xrightarrow{p} \mathbb{E}\left[\overrightarrow{X}\overrightarrow{X}' e^2\right] = \boldsymbol{\Omega}.$$

Technically, this requires that $\boldsymbol{\Omega}$ has finite elements, which was shown in (7.6).

It remains to show that

$$\frac{1}{n}\sum_{i=1}^{n} \overrightarrow{X}_i \overrightarrow{X}_i' \left(\widehat{e}_i^2 - e_i^2\right) \xrightarrow{p} 0. \tag{7.20}$$

There are multiple ways to do this. A reasonable straightforward yet slightly tedious derivation is to start by applying the triangle inequality using a matrix norm,

$$\begin{aligned}
\left\| \frac{1}{n}\sum_{i=1}^{n} \overrightarrow{X}_i \overrightarrow{X}_i' \left(\widehat{e}_i^2 - e_i^2\right) \right\| &\le \frac{1}{n}\sum_{i=1}^{n} \left\| \overrightarrow{X}_i \overrightarrow{X}_i' \left(\widehat{e}_i^2 - e_i^2\right) \right\| \\
&= \frac{1}{n}\sum_{i=1}^{n} \|\overrightarrow{X}_i\|^2 \left|\widehat{e}_i^2 - e_i^2\right|.
\end{aligned} \tag{7.21}$$

Then recalling the expression for the squared residual (7.17), apply the triangle inequality and then the Schwarz inequality twice

$$\begin{aligned} \left|\widehat{e}_i^2 - e_i^2\right| &\le 2\left|e_i \vec{X}_i'(\widehat{\vec{\beta}} - \vec{\beta})\right| + (\widehat{\vec{\beta}} - \vec{\beta})'\vec{X}_i\vec{X}_i'(\widehat{\vec{\beta}} - \vec{\beta}) \\ &= 2\left|e_i\right| \left|\vec{X}_i'(\widehat{\vec{\beta}} - \vec{\beta})\right| + \left|(\widehat{\vec{\beta}} - \vec{\beta})'\vec{X}_i\right|^2 \\ &\le 2\left|e_i\right| \|\vec{X}_i\| \, \|\widehat{\vec{\beta}} - \vec{\beta}\| + \|\vec{X}_i\|^2 \, \|\widehat{\vec{\beta}} - \vec{\beta}\|^2 \end{aligned} \tag{7.22}$$

Combining (7.21) and (7.22), we find

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \vec{X}_i\vec{X}_i' \left(\widehat{e}_i^2 - e_i^2\right) \right\| \le 2\left(\frac{1}{n}\sum_{i=1}^{n} \|\vec{X}_i\|^3 \left|e_i\right|\right) \|\widehat{\vec{\beta}} - \vec{\beta}\| + \left(\frac{1}{n}\sum_{i=1}^{n} \|\vec{X}_i\|^4\right) \|\widehat{\vec{\beta}} - \vec{\beta}\|^2 = o_p(1). \tag{7.23}$$

The expression is $o_p(1)$ because $\|\widehat{\vec{\beta}} - \vec{\beta}\| \xrightarrow{p} 0$ and both averages in parenthesis are averages of random variables with finite expectation under Assumption 7.2 (and are thus $O_p(1)$). Indeed, by Holder's inequality,

$$\mathbb{E}\left[\|X\|^3|e|\right] \le \left(\mathbb{E}\left[\left(\|X\|^3\right)^{4/3}\right]\right)^{3/4} \left(\mathbb{E}\left[e^4\right]\right)^{1/4} = \left(\mathbb{E}\|X\|^4\right)^{3/4} \left(\mathbb{E}\left[e^4\right]\right)^{1/4} < \infty.$$

**Theorem 7.6.**

Under Assumption 7.2, as $n \longrightarrow \infty$, $\widehat{\mathbf{\Omega}} \xrightarrow{p} \mathbf{\Omega}$ and $\widehat{\boldsymbol{V}}_{\vec{\beta}}^{\mathrm{HC0}} \xrightarrow{p} \boldsymbol{V}_{\vec{\beta}}$, where $\widehat{\boldsymbol{V}}_{\vec{\beta}}^{\mathrm{HC0}}$ is constructed as follows

$$\widehat{\boldsymbol{V}}_{\vec{\beta}}^{\mathrm{HC0}} = \left(\frac{1}{n}\sum_{i=1}^{n} \vec{X}_i\vec{X}_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \vec{X}_i\vec{X}_i'\widehat{e}_i^2\right) \left(\frac{1}{n}\sum_{i=1}^{n} \vec{X}_i\vec{X}_i'\right)^{-1}.$$

## 7.8. Summary of Covariance Matrix Notation

The notation we have introduced may be somewhat confusing so it is helpful to write it down in one place.

The exact variance of $\widehat{\vec{\beta}}$ (under the assumptions of the linear regression model) and the asymptotic variance of $\sqrt{n}\left(\widehat{\vec{\beta}} - \vec{\beta}\right)$ (under the more general assumptions of the linear projection model) are

$$\begin{aligned} \boldsymbol{V}_{\widehat{\vec{\beta}}} &= \operatorname{var}[\widehat{\vec{\beta}} \mid \boldsymbol{X}] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right) \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ \boldsymbol{V}_{\vec{\beta}} &= \operatorname{avar}\left[\sqrt{n}\left(\widehat{\vec{\beta}} - \vec{\beta}\right)\right] = \boldsymbol{Q}_{XX}^{-1}\mathbf{\Omega}\boldsymbol{Q}_{XX}^{-1}. \end{aligned}$$

The HC0 estimators of these covariance matrices are

$$\widehat{\boldsymbol{V}}^{\mathrm{HC0}}_{\hat{\vec{\beta}}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{i=1}^{n}\vec{X}_i\vec{X}_i'\hat{e}_i^2\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$\widehat{\boldsymbol{V}}^{\mathrm{HC0}}_{\vec{\beta}} = \widehat{\boldsymbol{Q}}^{-1}_{XX}\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{Q}}^{-1}_{XX}$$

and satisfy the simple relationship $\widehat{\boldsymbol{V}}^{\mathrm{HC0}}_{\vec{\beta}} = n\widehat{\boldsymbol{V}}^{\mathrm{HC0}}_{\hat{\vec{\beta}}}$.

Similarly, under the assumption of homoskedasticity the exact and asymptotic variances simplify to

$$\boldsymbol{V}^{0}_{\hat{\vec{\beta}}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$$

$$\boldsymbol{V}^{0}_{\vec{\beta}} = \boldsymbol{Q}^{-1}_{XX}\sigma^2.$$

Their standard estimators are

$$\widehat{\boldsymbol{V}}^{0}_{\hat{\vec{\beta}}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}s^2$$

$$\widehat{\boldsymbol{V}}^{0}_{\vec{\beta}} = \boldsymbol{Q}^{-1}_{XX}s^2.$$

which also satisfy the relationship $\widehat{\boldsymbol{V}}^{0}_{\vec{\beta}} = n\widehat{\boldsymbol{V}}^{0}_{\hat{\vec{\beta}}}$.

## 7.9. Alternative Covariance Matrix Estimators

In Section 7.7, we introduced $\widehat{\boldsymbol{V}}^{\mathrm{HC0}}_{\vec{\beta}}$ as an estimator of $\boldsymbol{V}_{\vec{\beta}}$. $\widehat{\boldsymbol{V}}^{\mathrm{HC0}}_{\vec{\beta}}$ is a scaled version of $\widehat{\boldsymbol{V}}^{\mathrm{HC0}}_{\hat{\vec{\beta}}}$ from Section 4.15, where we also introduced the alternative HC1, HC2, and HC3 heteroskedasticity-robust covariance matrix estimators. We now discuss the consistency properties of these estimators.

To do so we introduce their scaled versions, e.g.,

$$\widehat{\boldsymbol{V}}^{\mathrm{HC1}}_{\vec{\beta}} = n\widehat{\boldsymbol{V}}^{\mathrm{HC1}}_{\hat{\vec{\beta}}}, \quad \widehat{\boldsymbol{V}}^{\mathrm{HC2}}_{\vec{\beta}} = n\widehat{\boldsymbol{V}}^{\mathrm{HC2}}_{\hat{\vec{\beta}}}, \quad \widehat{\boldsymbol{V}}^{\mathrm{HC3}}_{\vec{\beta}} = n\widehat{\boldsymbol{V}}^{\mathrm{HC3}}_{\hat{\vec{\beta}}}.$$

These are alternative estimators of the asymptotic covariance matrix $\boldsymbol{V}_{\vec{\beta}}$.

Notice that

$$\widehat{\boldsymbol{V}}^{\mathrm{HC1}}_{\vec{\beta}} = n\widehat{\boldsymbol{V}}^{\mathrm{HC1}}_{\hat{\vec{\beta}}} = \frac{n}{n-k}\widehat{\boldsymbol{V}}^{\mathrm{HC0}}_{\vec{\beta}},$$

where $\widehat{\boldsymbol{V}}^{\mathrm{HC0}}_{\vec{\beta}}$ was defined in (7.19) and shown consistent for $\boldsymbol{V}_{\vec{\beta}}$ in Theorem 7.6. If $k$ is fixed as $n \to \infty$, then $\frac{n}{n-k} \to 1$ and thus $\widehat{\boldsymbol{V}}^{\mathrm{HC1}}_{\vec{\beta}}$ is consistent for $\boldsymbol{V}_{\vec{\beta}}$.

The alternative estimators $\widehat{\boldsymbol{V}}^{\mathrm{HC2}}_{\vec{\beta}}$ and $\widehat{\boldsymbol{V}}^{\mathrm{HC3}}_{\vec{\beta}}$ take the form (7.19) but with $\widehat{\boldsymbol{\Omega}}$ replaced by

$$\widetilde{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}(1-h_{ii})^{-2}\vec{X}_i\vec{X}_i'\widehat{e}_i^2$$

$$\overline{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}(1-h_{ii})^{-1}\vec{X}_i\vec{X}_i'\widehat{e}_i^2,$$

respectively. To show that these estimators also consistent for $\boldsymbol{V}_{\vec{\beta}}$ given $\widehat{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$ it is sufficient to show that the differences $\widetilde{\boldsymbol{\Omega}} - \widehat{\boldsymbol{\Omega}}$ and $\overline{\boldsymbol{\Omega}} - \widehat{\boldsymbol{\Omega}}$ converge in probability to zero as $n \longrightarrow \infty$.

The trick is the fact that the leverage values are asymptotically negligible:

$$h_n^* = \max_{1\le i\le n} h_{ii} = o_p(1). \tag{7.24}$$

**Theorem 7.7.**

Under Assumption 7.2, as $n \longrightarrow \infty$,

$$\widehat{\boldsymbol{V}}^{\mathrm{HC1}}_{\vec{\beta}} \xrightarrow{p} \boldsymbol{V}^{\mathrm{HC1}}_{\vec{\beta}}, \quad \widehat{\boldsymbol{V}}^{\mathrm{HC2}}_{\vec{\beta}} \xrightarrow{p} \boldsymbol{V}^{\mathrm{HC2}}_{\vec{\beta}}, \quad \widehat{\boldsymbol{V}}^{\mathrm{HC3}}_{\vec{\beta}} \xrightarrow{p} \boldsymbol{V}^{\mathrm{HC3}}_{\vec{\beta}}$$

Theorem 7.7 shows that the alternative covariance matrix estimators are also consistent for the asymptotic covariance matrix. To simplify notation, for the remainder of the chapter we will use the notation $\widehat{\boldsymbol{V}}_{\vec{\beta}}$ and $\widehat{\boldsymbol{V}}_{\widehat{\vec{\beta}}}$ to refer to any of the heteroskedasticity-consistent covariance matrix estimators HC0, HC1, HC2 and HC3, since they all have the same asymptotic limits.

## 7.10. Functions of Parameters

In most serious applications a researcher is actually interested in a specific transformation of the coefficient vector $\vec{\beta}$, where we can write the parameter of interest $\theta$ as a function of the coefficients, e.g., $\theta = r(\vec{\beta})$ for some function $r : \mathbb{R}^k \longrightarrow \mathbb{R}^q$. The estimate of $\theta$ is

$$\widehat{\theta} = r\left(\widehat{\vec{\beta}}\right).$$

**Theorem 7.8.**

Under Assumption 7.1, if $r(\vec{\beta})$ is continuous at the true value of $\vec{\beta}$ then as $n \longrightarrow \infty$, $\widehat{\theta} \xrightarrow{p} \theta$.

Furthermore, if the transformation is sufficiently smooth, by the Delta Method we can show that $\widehat{\theta}$ is asymptotically normal.

**Assumption 7.3.**

$r(\overrightarrow{\beta}) : \mathbb{R}^k \longrightarrow \mathbb{R}^q$ is continuously differentiable at the true value of $\overrightarrow{\beta}$ and

$$\boldsymbol{R} = \frac{\partial}{\partial \overrightarrow{\beta}} r\left(\overrightarrow{\beta}\right)'$$

has rank $q$.

**Theorem 7.9. Asymptotic Distribution of Functions of Parameters**

Under Assumption 7.2 and 7.3, as $n \longrightarrow \infty$,

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow{d} N(0, \boldsymbol{V}_\theta) \tag{7.25}$$

where $\boldsymbol{V}_\theta = \boldsymbol{R}'\boldsymbol{V}_{\overrightarrow{\beta}}\boldsymbol{R}$.

In many cases the function $r(\overrightarrow{\beta})$ is linear:

$$r(\overrightarrow{\beta}) = \boldsymbol{R}'\overrightarrow{\beta}$$

for some $k \times q$ matrix $\boldsymbol{R}$. In particular, if $\boldsymbol{R}$ is a "selector matrix"

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{I} \\ \boldsymbol{0} \end{pmatrix},$$

then we can partition $\overrightarrow{\beta} = (\overrightarrow{\beta}'_1, \overrightarrow{\beta}'_2)'$ so that $\boldsymbol{R}'\overrightarrow{\beta} = \overrightarrow{\beta}_1$. Then

$$\boldsymbol{V}_\theta = \boldsymbol{V}_{11},$$

the upper-left sub-matrix of $\boldsymbol{V}_{11}$ given in (7.14).

To illustrate the case of a nonlinear transformation take the example $\theta = \beta_j/\beta_l$ for $j \neq l$. Then

$$\boldsymbol{R} = \frac{\partial}{\partial \overrightarrow{\beta}} r(\overrightarrow{\beta}) = \begin{pmatrix} \frac{\partial}{\partial \beta_1}\left(\beta_j/\beta_l\right) \\ \vdots \\ \frac{\partial}{\partial \beta_j}\left(\beta_j/\beta_l\right) \\ \vdots \\ \frac{\partial}{\partial \beta_\ell}\left(\beta_j/\beta_l\right) \\ \vdots \\ \frac{\partial}{\partial \beta_k}\left(\beta_j/\beta_l\right) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 1/\beta_l \\ \vdots \\ -\beta_j/\beta_l^2 \\ \vdots \\ 0 \end{pmatrix} \tag{7.26}$$

so

$$\boldsymbol{V}_\theta = \boldsymbol{V}_{jj}/\beta_l^2 + \boldsymbol{V}_{ll}\beta_j^2/\beta_l^4 - 2\boldsymbol{V}_{jl}\beta_j/\beta_l^3$$

where $\boldsymbol{V}_{ab}$ denotes the $ab$th element of $\boldsymbol{V}_\beta$.

For inference we need an estimator of the asymptotic covariance matrix $\boldsymbol{V}_\theta = \boldsymbol{R}'\boldsymbol{V}_{\vec{\beta}}\boldsymbol{R}$. For this it is typical to use the plug-in estimator

$$\widehat{\boldsymbol{R}} = \frac{\partial}{\partial \vec{\beta}} r\left(\widehat{\vec{\beta}}\right)'. \tag{7.27}$$

The derivation in (7.27) may be calculated analytically or numerically.

The estimator for $\boldsymbol{V}_\theta$ is

$$\widehat{\boldsymbol{V}}_\theta = \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\vec{\beta}}\widehat{\boldsymbol{R}}. \tag{7.28}$$

Alternatively, the homoskedastic covariance matrix estimator could be used leading to a homoskedastic covariance matrix estimator for $\theta$.

$$\widehat{\boldsymbol{V}}_\theta^0 = \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\vec{\beta}}^0\widehat{\boldsymbol{R}} = \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\boldsymbol{R}}s^2. \tag{7.29}$$

As the primary justification for $\widehat{\boldsymbol{V}}_\theta$ is the asymptotic approximation (7.25), $\widehat{\boldsymbol{V}}_\theta$ is often called an *asymptotic covariance matrix estimator*.

**Theorem 7.10.**

Under Assumption 7.2 and 7.3, as $n \longrightarrow \infty$, $\widehat{\boldsymbol{V}}_\theta \xrightarrow{p} \boldsymbol{V}_\theta$.

In practice, we may set

$$\widehat{\boldsymbol{V}}_{\widehat{\theta}} = \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\widehat{\vec{\beta}}}\widehat{\boldsymbol{R}} = n^{-1}\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\vec{\beta}}\widehat{\boldsymbol{R}} \tag{7.30}$$

as an estimator of the variance of $\widehat{\theta}$.

## 7.11. Asymptotic Standard Errors

As described in Section 4.16 a standard error is an estimator of the standard deviation of the distribution of an estimator.

$$s\left(\widehat{\theta}\right) = \sqrt{\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\widehat{\vec{\beta}}}\widehat{\boldsymbol{R}}} = \sqrt{n^{-1}\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\vec{\beta}}\widehat{\boldsymbol{R}}}.$$

When the justification is based on asymptotic theory we call $s\left(\widehat{\theta}\right)$ or $s\left(\widehat{\beta}_j\right)$ an *asymptotic standard error* for $\widehat{\theta}$ or $\widehat{\beta}_j$. When reporting your results it is good practice to report standard errors for each reported estimate and this includes functions and transformations of your parameter estimates. This helps users of the work (including yourself ) assess the estimation precision.

## 7.12. t-statistic

Let $\theta = r(\overrightarrow{\beta}) : \mathbb{R}^k \to \mathbb{R}$ be a a parameter of interest, $\widehat{\theta}$ its estimator, and $s\left(\widehat{\theta}\right)$ its asymptotic standard error. Consider the statistic

$$T(\theta) = \frac{\widehat{\theta} - \theta}{s\left(\widehat{\theta}\right)}. \tag{7.33}$$

By Theorems 7.9 and 7.10, $\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow{d} N(0, \boldsymbol{V}_\theta)$ and $\widehat{\boldsymbol{V}}_\theta \xrightarrow{p} \boldsymbol{V}_\theta$. Thus

$$\begin{aligned} T(\theta) &= \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})} \\ &= \frac{\sqrt{n}(\widehat{\theta} - \theta)}{\sqrt{\widehat{V}_\theta}} \\ &\longrightarrow \frac{\mathrm{N}\left(0, V_\theta\right)}{\sqrt{V_\theta}} \\ &= Z \sim \mathrm{N}(0,1). \end{aligned}$$

**Assumption 7.4.**

$\boldsymbol{V}_\theta = \boldsymbol{R}'\boldsymbol{V}_{\overrightarrow{\beta}}\boldsymbol{R} > 0.$

Thus the asymptotic distribution of the t-ratio $T(\theta)$ is standard normal. Since this distribution does not depend on the parameters we say that $T(\theta)$ is *asymptotically pivotal*. In finite samples $T(\theta)$ is not necessarily pivotal but the property means that the dependence on unknowns diminishes as $n$ increases.

It is also useful to consider the distribution of the *absolute t-ratio* $|T(\theta)|$. Since $T(\theta) \xrightarrow{d} Z$ the continuous mapping theorem yields $|T(\theta)| \xrightarrow{d} |Z|$. Letting $\Phi(u) = \mathbb{P}[Z \leq u]$ denote the standard normal distribution function we calculation the distribution of $|Z|$ as

$$\mathbb{P}[|Z| \leq u] = 2\Phi(u) - 1. \tag{7.34}$$

**Theorem 7.11.**

Under Assumption 7.2, 7.3, and 7.4, $T(\theta) \xrightarrow{d} Z = N(0,1)$ and $|T(\theta)| \xrightarrow{d} |Z|$.

The asymptotic normality of Theorem 7.11 is used to justify confidence intervals and tests for the parameters.

## 7.13. Confidence Intervals

The estimator $\widehat{\theta}$ is a *point estimator* for $\theta$ meaning that $\widehat{\theta}$ is a single value in $\mathbb{R}^q$. A broader concept is a *set estimator* $\widehat{C}$ which is a collection of values in $\mathbb{R}^q$. When the parameter $\theta$ is real-valued then it is common to focus on sets of the form $\widehat{C} = \left[\widehat{L}, \widehat{U}\right]$, which is called an *interval estimator* for $\theta$.

An interval estimate $\widehat{C}$ is a function of the data and hence is random. The *coverage probability* of the interval $\widehat{C} = \left[\widehat{L}, \widehat{U}\right]$ is $\mathbb{P}\left[\theta \in \widehat{C}\right]$. The randomness comes from $\widehat{C}$ as the parameter $\theta$ is treated as fixed. In Section 5.10 we introduced confidence intervals for the normal regression model which used the finite sample distribution of the $t$-statistic. When we are outside the normal regression model we cannot rely on the exact normal distribution theory but instead use asymptotic approximations. A benefit is that we can construct confidence intervals for general parameters of interest $\theta$ not just regression coefficients.

An interval estimator $\widehat{C}$ is called a *confidence interval* when the goal is to set the coverage probability to equal a pre-specified target such as 90% or 95%. $\widehat{C}$ is called a $1-\alpha$ confidence interval is

$$\inf_{\theta} \mathbb{P}\left[\theta \in \widehat{C}\right] = 1 - \alpha.$$

When $\widehat{\theta}$ is asymptotically normal with standard errors $s\left(\widehat{\theta}\right)$ the conventional confidence interval for $\theta$ takes the form

$$\widehat{C} = \left[\widehat{\theta} - c \times s\left(\widehat{\theta}\right), \widehat{\theta} + c \times s\left(\widehat{\theta}\right)\right], \tag{7.35}$$

where $c$ equals the $1-\alpha$ quantile of the distribution of $|Z|$. Using (7.34) we calculate that $c$ is equivalently the $1-\alpha/2$ quantile of the standard normal distribution. Thus, $c$ solves

$$2\Phi(c) - 1 = 1 - \alpha.$$

The confidence interval (7.35) is symmetric about the point estimator $\widehat{\theta}$ and its length is proportional to the standard error $s\left(\widehat{\theta}\right)$.

Equivalently, (7.35) is the set of parameter values for $\theta$ such that the $t$-statistic $T(\theta)$ is smaller (in absolute value) than $c$, that is,

$$\widehat{C} = \{\theta : |T(\theta)| \leq c\} = \left\{\theta : -c \leq \frac{\widehat{\theta} - \theta}{s\left(\widehat{\theta}\right)} \leq c\right\}.$$

The coverage probability of this confidence interval is

$$\mathbb{P}\left[\theta \in \widehat{C}\right] = \mathbb{P}\left[|T(\theta)| \leq c\right] \longrightarrow \mathbb{P}\left[|Z| \leq c\right] = 1 - \alpha,$$

where the limit is taken as $n \longrightarrow \infty$, and holds since $T(\theta)$ is asymptotically standard normal by Theorem 7.11. We call the limit the *asymptotic coverage probability* and call $\widehat{C}$ an asymptotic $1-\alpha$ confidence interval for $\theta$. Since the $t$-ratio is asymptotically pivotal, the asymptotic coverage probability is independent of the parameter $\theta$.

It is useful to contrast the confidence interval (7.35) with (5.8) for the normal regression model. They are similar but there are differences. *The normal regression interval (5.8) only applies to regression coefficients $\beta$ not to functions $\theta$ of the coefficients. The normal interval (5.8) also is constructed with the homoskedastic standard error, while (7.35) can be constructed with a heteroskedastic-robust standard error. Furthermore, the constants $c$ in (5.8) are calculated using the student $t$ distribution, while $c$ in (7.35) are calculated using the normal distribution.* The difference between the student $t$ and normal values are typically small in practice (since sample sizes are large in typical economic applications). However, since the student $t$ values are larger, it results in slightly larger confidence intervals which is reasonable. A practical rule of thumb is that if the sample sizes are sufficiently small that it makes a difference then neither (5.8) nor (7.35) should be trusted. Despite these differences the coincidence of the intervals means that inference on regression coefficients is generally robust to using either the exact normal sampling assumption or the asymptotic large sample approximation, at least in large samples.

*Stata by default reports 95% confidence intervals for each coefficient where the critical values $c$ are calculated using the $t_{n-k}$ distribution. This is done for all standard error methods even though it is only exact for homoskedastic standard errors and under normality.*

The standard coverage probability for confidence intervals is 95%, leading to the choice $c = 1.96$ for the constant in (7.35). Rounding 1.96 to 2, we obtain the most commonly used confidence interval in applied econometric practice

$$\widehat{C} = \left[\widehat{\theta} - 2 \times s\left(\widehat{\theta}\right), \widehat{\theta} + 2 \times s\left(\widehat{\theta}\right)\right].$$

This is a useful rule-of thumb. This asymptotic 95% confidence interval $\widehat{C}$ is simple to compute and can be roughly calculated from tables of coefficient estimates and standard errors. (Technically, it is an asymptotic 95.4% interval due to the substitution of 2.0 for 1.96 but this distinction is overly precise.)

**Theorem 7.12.**

Under Assumptions 7.2, 7.3, and 7.4, for $\widehat{C}$ defined in (7.35), with

$$c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right),$$

we have

$$\mathbb{P}\left[\theta \in \widehat{C}\right] \longrightarrow 1 - \alpha.$$

For $c = 1.96$, $\mathbb{P}\left[\theta \in \widehat{C}\right] \longrightarrow 0.95$.

Confidence intervals are a simple yet effective tool to assess estimation uncertainty. When reading a set of empirical results look at the estimated coefficient estimates and the standard errors. For a parameter of interest compute the confidence interval $\widehat{C}$ and consider the meaning of the spread of the suggested values. If the range of values in the confidence interval are too wide to learn about $\theta$ then do not jump to a conclusion about $\theta$ based on the point estimate alone.

## 7.14. Regression Intervals

In the linear regression model the confidence expectation of $Y$ given $\overrightarrow{X} = \overrightarrow{x}$ is

$$m(\overrightarrow{x}) = \mathbb{E}\left[Y \mid \overrightarrow{X} = \overrightarrow{x}\right] = \overrightarrow{x}'\overrightarrow{\beta}.$$

In some cases we want to estimate $m(\overrightarrow{x})$ at a particular point $\overrightarrow{x}$. Notice that this is a linear function of $\overrightarrow{\beta}$. Letting $r(\overrightarrow{\beta}) = \overrightarrow{x}'\overrightarrow{\beta}$ and $\theta = r(\overrightarrow{\beta})$, we see that $\widehat{m}(\overrightarrow{x}) = \widehat{\theta} = \overrightarrow{x}'\overrightarrow{\beta}$ and $\boldsymbol{R} = \overrightarrow{\xi}$, so

$$s(\overrightarrow{\theta}) = \sqrt{\overrightarrow{x}'\widehat{\boldsymbol{V}}_{\overrightarrow{\beta}}\overrightarrow{x}}.$$

Thus, an asymptotic 95% confidence interval for $m(\overrightarrow{x})$ is

$$\left[\overrightarrow{x}'\overrightarrow{\beta} - 1.96 \times \sqrt{\overrightarrow{x}'\widehat{\boldsymbol{V}}_{\overrightarrow{\beta}}\overrightarrow{x}},\ \overrightarrow{x}'\overrightarrow{\beta} + 1.96 \times \sqrt{\overrightarrow{x}'\widehat{\boldsymbol{V}}_{\overrightarrow{\beta}}\overrightarrow{x}}\right].$$

It is interesting to observe that if this is viewed as a function of $\overrightarrow{x}$ the width of the confidence interval is dependent on $\overrightarrow{x}$.

## 7.15. Forecast Intervals

Suppose we are given a value of the regressor vector $\overrightarrow{X}_{n+1}$ for an individual outside the sample and we want to forecast $Y_{n+1}$ for this individual. This is equivalent to forecasting $Y_{n+1}$ given $\overrightarrow{X}_{n+1} = \overrightarrow{x}$ which will generally be a function of $\overrightarrow{x}$. A reasonable forecasting rule is the conditional expectation $m(\overrightarrow{x})$ as it is the mean-square minimizing forecast. A point forecast is the estimated conditional expectation $\widehat{m}(\overrightarrow{x}) = \overrightarrow{x}'\overrightarrow{\beta}$. We would also like a measure of uncertainty for the forecast.

The forecast error is $\widehat{e}_{n+1} = Y_{n+1} - \widehat{m}(\overrightarrow{x}) = e_{n+1} - \overrightarrow{x}'\left(\widehat{\overrightarrow{\beta}} - \overrightarrow{\beta}\right)$. As the out-of-sample error $e_{n+1}$

is independent of the in-sample estimator $\widehat{\overrightarrow{\beta}}$, this has conditional variance

$$\begin{aligned} \mathbb{E}\left[\widehat{e}_{n+1}^2 \mid \overrightarrow{X}_{n+1} = \overrightarrow{x}\right] &= \mathbb{E}\left[\left(e_{n+1} - \overrightarrow{x}'\left(\widehat{\overrightarrow{\beta}} - \overrightarrow{\beta}\right)\right)^2 \mid \overrightarrow{X}_{n+1} = \overrightarrow{x}\right] \\ &= \mathbb{E}\left[\widehat{e}_{n+1} \mid \overrightarrow{X}_{n+1} = \overrightarrow{x}\right] + \overrightarrow{x}'\,\mathbb{E}\left[\left(\widehat{\overrightarrow{\beta}} - \overrightarrow{\beta}\right)^2\right]\overrightarrow{x} \\ &= \sigma^2(\overrightarrow{x}) + \overrightarrow{x}'\boldsymbol{V}_{\widehat{\overrightarrow{\beta}}}\overrightarrow{x}. \end{aligned} \tag{7.36}$$

Under homoskedasticity $\mathbb{E}\left[\widehat{e}_{n+1} \mid \overrightarrow{X}_{n+1} = \overrightarrow{x}\right] = \sigma^2$. In this case a simple estimator of (7.36) is $\widehat{\sigma}^2 + \overrightarrow{x}'\boldsymbol{V}_{\widehat{\overrightarrow{\beta}}}\overrightarrow{x}$ so a standard error for the forecast can be estimated as

$$\widehat{s}\left(\widehat{m}(\overrightarrow{x})\right) = \sqrt{\widehat{\sigma}^2 + \overrightarrow{x}'\widehat{\boldsymbol{V}}_{\widehat{\overrightarrow{\beta}}}\overrightarrow{x}}.$$

*Notice that this is different from the standard error for the conditional expectation.*

The conventional 95% forecast interval for $Y_{n+1}$ uses a normal approximation and equals

$$\left[\overrightarrow{x}'\widehat{\overrightarrow{\beta}} - 1.96 \times \widehat{s}\left(\widehat{m}(\overrightarrow{x})\right),\ \overrightarrow{x}'\widehat{\overrightarrow{\beta}} + 1.96 \times \widehat{s}\left(\widehat{m}(\overrightarrow{x})\right)\right].$$

However, it is difficult to fully justify this choice. It would be correct if we have a normal distribution to the ratio

$$\frac{e_{n+1} - \overrightarrow{x}'\left(\widehat{\overrightarrow{\beta}} - \overrightarrow{\beta}\right)}{\widehat{s}\left(\widehat{m}(\overrightarrow{x})\right)}.$$

The difficulty is that the equation error $e_{n+1}$ is generally non-normal and asymptotic theory cannot be applied to a single observation. The only special exception is the case where $e_{n+1}$ has the exact distribution $N\left(0, \sigma^2\right)$ which is generally invalid.

An accurate forecast interval would use the conditional distribution of $e_{n+1}$ given $\overrightarrow{X}_{n+1} = \overrightarrow{x}$, which is more challenging to estimate. Due to this difficulty many applied forecasters use the simple approximate interval despite the lack of a convincing justification.

## 7.16. Wald Statistic

Let $\overrightarrow{\theta} = r(\overrightarrow{\beta}) : \mathbb{R}^k \to \mathbb{R}^q$ be any parameter vector of interest, $\widehat{\overrightarrow{\theta}}$ its estimator, and $\widehat{\boldsymbol{V}}_{\widehat{\overrightarrow{\theta}}}$ its covariance matrix estimator. Consider the quadratic form

$$W(\overrightarrow{\theta}) = \left(\widehat{\overrightarrow{\theta}} - \overrightarrow{\theta}\right)'\widehat{\boldsymbol{V}}_{\widehat{\overrightarrow{\theta}}}^{-1}\left(\widehat{\overrightarrow{\theta}} - \overrightarrow{\theta}\right) = n\left(\widehat{\overrightarrow{\theta}} - \overrightarrow{\theta}\right)'\left(\widehat{\boldsymbol{V}}_{\overrightarrow{\theta}}\right)^{-1}\left(\widehat{\overrightarrow{\theta}} - \overrightarrow{\theta}\right), \tag{7.37}$$

where

$$\widehat{\boldsymbol{V}}_{\vec{\theta}} = n\widehat{\boldsymbol{V}}_{\widehat{\vec{\theta}}}.$$

When $q = 1$, then $W(\theta) = T(\theta)^2$ is the square of the $t$-ratio. When $q > 1$, $W(\vec{\theta})$ is typically called a *Wald statistic*. We are interested in its sampling distribution.

The asymptotic distribution of $W(\vec{\theta})$ is simple to derive given Theorem 7.9 and Theorem 7.10. They show that

$$\sqrt{n}\left(\widehat{\vec{\theta}} - \vec{\theta}\right) \xrightarrow{d} \vec{Z} \sim N\left(0, \boldsymbol{V}_{\vec{\theta}}\right),$$

and

$$\widehat{\boldsymbol{V}}_{\vec{\theta}} \xrightarrow{p} \boldsymbol{V}_{\vec{\theta}}.$$

It follows that

$$W(\vec{\theta}) = \sqrt{n}\left(\widehat{\vec{\theta}} - \vec{\theta}\right)' \widehat{\boldsymbol{V}}_{\vec{\theta}}^{-1} \sqrt{n}\left(\widehat{\vec{\theta}} - \vec{\theta}\right) \xrightarrow{d} \vec{Z}'\boldsymbol{V}_{\vec{\theta}}^{-1}\vec{Z}$$

a quadratic int he normal random vector $\vec{Z}$. As shown in Theorem 5.3(5), the distribution of this quadratic form is $\chi_q^2$, a chi-square random variable with $q$ degrees of freedom.

**Theorem 7.13.**

Under Assumptions 7.2, 7.3, and 7.4, as $n \longrightarrow \infty$,

$$W\left(\vec{\theta}\right) \xrightarrow{d} \chi_q^2.$$

Theorem 7.13 is used to justify multivariate confidence regions and multivariate hypothesis tests.

## 7.17. Homoskedastic Wald Statistic

Under the conditional homoskedasticity assumption $\mathbb{E}\left[e^2 \mid \vec{X}\right] = \sigma^2$ we can construct the Wald statistic using the homoskedastic covariance matrix estimator $\widehat{\boldsymbol{V}}^0_{\vec{\theta}}$ defined in (7.29). This yields a homoskedastic Wald statistic

$$W^0\left(\vec{\theta}\right) = \left(\widehat{\vec{\theta}} - \vec{\theta}\right)' \left(\widehat{\boldsymbol{V}}^0_{\widehat{\vec{\theta}}}\right)^{-1} \left(\widehat{\vec{\theta}} - \vec{\theta}\right) = n\left(\widehat{\vec{\theta}} - \vec{\theta}\right)' \left(\widehat{\boldsymbol{V}}^0_{\vec{\theta}}\right)^{-1} \left(\widehat{\vec{\theta}} - \vec{\theta}\right). \tag{7.38}$$

Using the additional assumption of conditional homoskedasticity it has the same asymptotic distribution as $W\left(\vec{\theta}\right)$.

**Theorem 7.14.**

Under Assumptions 7.2, 7.3, and $\mathbb{E}\left[e^2 \mid \overrightarrow{X}\right] = \sigma^2 > 0$, as $n \to \infty$,

$$W^0\left(\overrightarrow{\theta}\right) \xrightarrow{d} \chi^2_q.$$

A confidence region $\widehat{C}$ is a set estimator for $\overrightarrow{\theta} \in \mathbb{R}^q$ when $q > 1$. A confidence region $\widehat{\boldsymbol{C}}$ is a set in $\mathbb{R}^q$ intended to cover the true parameter value with a pre-selected probability $1-\alpha$. Thus an ideal confidence region has the coverage probability $\mathbb{P}\left[\overrightarrow{\theta} \in \widehat{\boldsymbol{C}}\right] = 1 - \alpha$. In practice it is typically not possible to construct a region with exact coverage but we can calculate its asymptotic coverage.

When the parameter estimator satisfies the conditions of Theorem 7.13, a good choice for a confidence region is the ellipse

$$\widehat{\boldsymbol{C}} = \left\{\overrightarrow{\theta} : W\left(\overrightarrow{\theta}\right) \leq c_{1-\alpha}\right\},$$

with $c_{1-\alpha}$ the $1 - \alpha$ quantile of the $\chi^2_q$ distribution, i.e., $F_q(c_{1-\alpha}) = 1 - \alpha$.

Theorem 7.13 implies

$$\mathbb{P}\left[\overrightarrow{\theta} \in \widehat{\boldsymbol{C}}\right] \to \mathbb{P}\left[\chi^2_q \leq c_{1-\alpha}\right] = 1 - \alpha,$$

which shows that $\widehat{\boldsymbol{C}}$ has asymptotic coverage $1 - \alpha$.

## 7.18. Confidence Regions

# Appendix A

# The Multivariate Normal Distribution and Its Application to Statistical Inference

## A.2. The Multivariate Normal Distribution

Now let the components of $\overrightarrow{X} = (X_1, \ldots, X_n)'$ be independent, standard normally distributed random variables. Then, $\mathbb{E}\left[\overrightarrow{X}\right] = \overrightarrow{0}$ and $\operatorname{var}\left[\overrightarrow{X}\right] = \boldsymbol{I}_n$. Moreover, the joint density $f(\overrightarrow{x}) = f(x_1, \ldots, x_n)$ of $\overrightarrow{X}$ in this case is the product of the standard normal marginal densities:

$$\begin{aligned} f(\overrightarrow{x}) &= \prod_{j=1}^{n} \frac{\exp\left(-x_j^2/2\right)}{\sqrt{2\pi}} \\ &= \frac{\exp\left(-\frac{1}{2}\sum_{j=1}^{n} x_j^2\right)}{\left(\sqrt{2\pi}\right)^n} \\ &= \frac{\exp\left(-\frac{1}{2}\overrightarrow{x}'\overrightarrow{x}\right)}{\left(\sqrt{2\pi}\right)^n} \end{aligned}$$

Next, consider the following linear transformation of $\overrightarrow{X}$:

$$\overrightarrow{Y} = \boldsymbol{A}\overrightarrow{X} + \overrightarrow{\mu},$$

where $\overrightarrow{\mu} = (\mu_1, \ldots, \mu_n)'$ is a vector of constants and $\boldsymbol{A}$ is a nonsingular $n \times n$ matrix with nonrandom elements. Because $\boldsymbol{A}$ is nonsingular and therefore invertible, this transformation is a one-to-one mapping with inverse

$$\boldsymbol{X} = \boldsymbol{A}^{-1}\left(\overrightarrow{Y} - \overrightarrow{\mu}\right).$$

Then the density function $g(\vec{y})$ of $\vec{Y}$ is equal to

$$\begin{aligned}
g(\vec{y}) &= f(\vec{x}) \left|\det\left(\frac{\partial \vec{x}}{\partial \vec{y}}\right)\right| \\
&= f\left(\boldsymbol{A}^{-1}\vec{y} - \boldsymbol{A}^{-1}\vec{\mu}\right) \left|\det\left(\frac{\partial \boldsymbol{A}^{-1}\vec{y} - \boldsymbol{A}^{-1}\vec{\mu}}{\partial \vec{y}}\right)\right| \\
&= f\left(\boldsymbol{A}^{-1}\vec{y} - \boldsymbol{A}^{-1}\vec{\mu}\right) \left|\det\left(\boldsymbol{A}^{-1}\right)\right| \\
&= \frac{f(\boldsymbol{A}^{-1}\vec{y} - \boldsymbol{A}^{-1}\vec{\mu})}{|\det(\boldsymbol{A})|} \\
&= \frac{\exp\left[-\frac{1}{2}(\vec{y} - \vec{\mu})'(\boldsymbol{A}^{-1})'\boldsymbol{A}^{-1}(\vec{y} - \vec{\mu})\right]}{\left(\sqrt{2\pi}\right)^n |\det(\boldsymbol{A})|} \\
&= \frac{\exp\left[-\frac{1}{2}(\vec{y} - \vec{\mu})'(\boldsymbol{A}\boldsymbol{A}')^{-1}(\vec{y} - \vec{\mu})\right]}{\left(\sqrt{2\pi}\right)^n \sqrt{|\det(\boldsymbol{A}\boldsymbol{A}')|}}.
\end{aligned}$$

Observe that $\vec{\mu}$ is the expectation vector of $\vec{Y}$: $\mathbb{E}[\vec{Y}] = \vec{\mu} + \boldsymbol{A}(\mathbb{E}[\vec{X}]) = \vec{\mu}$. But what is $\boldsymbol{A}\boldsymbol{A}'$?

$$\begin{aligned}
\operatorname{var}[\vec{Y}] &= \mathbb{E}[\vec{Y}\vec{Y}'] - (\mathbb{E}[\vec{Y}])(\mathbb{E}[\vec{Y}])' \\
&= \mathbb{E}[(\vec{\mu} + \boldsymbol{A}\vec{X})(\vec{\mu}' + \vec{X}'\boldsymbol{A}')] - \vec{\mu}\vec{\mu}' \\
&= \vec{\mu}\vec{\mu}' + \vec{\mu}\,\mathbb{E}[\vec{X}']\boldsymbol{A}' + \boldsymbol{A}\,\mathbb{E}[\vec{X}]\vec{\mu}' + \boldsymbol{A}\,\mathbb{E}[\vec{X}\vec{X}']\boldsymbol{A}' - \vec{\mu}\vec{\mu}' \\
&= \boldsymbol{A}\boldsymbol{A}',
\end{aligned}$$

because $\mathbb{E}[\vec{X}] = \vec{0}$ and $\mathbb{E}[\vec{X}\vec{X}'] = \boldsymbol{I}_n$. Thus, $\boldsymbol{A}\boldsymbol{A}'$ is the variance matrix of $\vec{Y}$. This argument gives rise to the following definition of the $n$-variate normal distribution:

**Definition A.1.**

Let $\vec{Y}$ be an $n \times 1$ random vector satisfying $\mathbb{E}[\vec{Y}] = \vec{\mu}$ and $\operatorname{var}[\vec{Y}] = \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is nonsingular. Then $\vec{Y}$ is distributed $N_n(\vec{\mu}, \boldsymbol{\Sigma})$ if hte density $g(\vec{y})$ of $\vec{Y}$ is of the form

$$g(\vec{y}) = \frac{\exp\left[-\frac{1}{2}(\vec{y} - \vec{\mu})'\boldsymbol{\Sigma}^{-1}(\vec{y} - \vec{\mu})\right]}{\left(\sqrt{2\pi}\right)^n \sqrt{|\det(\boldsymbol{\Sigma})|}} \tag{A.4}$$

In the same way as before we can show that a nonsingular (hence one-to-one) linear transformation of a normal distribution is normal itself:

**Theorem A.1.**

Let $\vec{Z} = \vec{a} + \boldsymbol{B}\vec{Y}$, where $\vec{Y}$ is distributed $N_n(\vec{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{B}$ is a nonsingular matrix of constants of size $n \times n$. Then $\boldsymbol{Z}$ is distributed $N_n(\vec{a} + \boldsymbol{B}\vec{\mu}, \boldsymbol{B\Sigma B}')$.

I will now relax the assumption in Theorem A.1 that the matrix $\boldsymbol{B}$ is a nonsingular $n \times n$ matrix. This more general version of Theorem A.1 can be proved using the moment-generating function or the characteristic function of the multivariate normal distribution.

**Theorem A.2.**

Let $\overrightarrow{Y}$ be distributed $N_n(\overrightarrow{\mu}, \Sigma)$. Then the moment-generating function of $\overrightarrow{Y}$ is

$$m(\overrightarrow{t}) = \exp\left(\overrightarrow{t}'\overrightarrow{\mu} + \frac{1}{2}\overrightarrow{t}'\Sigma\overrightarrow{t}\right),$$

and the characteristic of $\overrightarrow{Y}$ is

$$\varphi(t) = \exp\left(i \cdot \overrightarrow{t}'\overrightarrow{\mu} - \frac{1}{2}\overrightarrow{t}'\Sigma\overrightarrow{t}\right).$$

**Theorem A.3.**

Theorem A.1 holds for any linear transformation $\overrightarrow{Z} = \overrightarrow{a} + \boldsymbol{B}\overrightarrow{Y}$.

Note that this result holds regardless of whether the matrix $\boldsymbol{B}\Sigma\boldsymbol{B}$ is nonsingular or not. In the latter case the normal distribution involved is called “singular”:

**Definition A.2.**

An $n \times 1$ random vector $\overrightarrow{Y}$ has a singular $N_n(\overrightarrow{\mu}, \Sigma)$ distribution if its characteristic function is of the form

$$\varphi_Y(\overrightarrow{t}) = \exp\left(i \cdot \overrightarrow{t}'\overrightarrow{\mu} - \frac{1}{2}\overrightarrow{t}'\Sigma\overrightarrow{t}\right),$$

with $\boldsymbol{\Sigma}$ a singular, positive semidefinite matrix.

Because of the latter condition the distribution of the random vector $\overrightarrow{Y}$ involved is no longer absolutely continuous, but the form of the characteristic function is the same as in the nonsingular case – and that is all that matters.

For example, let $n = 2$ and

$$\overrightarrow{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & \sigma^2 \end{pmatrix}$$

where $\sigma^2 > 0$ but small. The density of the corresponding $N_2(\overrightarrow{mu}, \boldsymbol{\Sigma})$ distribution of $\overrightarrow{Y} = (Y_1, Y_2)'$ is

$$f\,(y_1, y_2 \mid \sigma) = \frac{\exp\left(-y_1^2/2\right)}{\sqrt{2\pi}} \times \frac{\exp\left(-y_2^2/(2\sigma^2)\right)}{\sigma\sqrt{2\pi}}. \tag{A.5}$$

Then $\lim_{\sigma\downarrow 0} f = 0$ if $y_2 \neq 0$, and $\lim_{\sigma\downarrow 0} f = \infty$ if $y_2 = 0$. *Thus, a singular normal density distribution does not have a density.*

The next theorem shows that uncorrelated multivariate normally distributed random variables are independent. Thus, although for most distributions uncorrelatedness does not imply independence, for the multivariate normal distribution it does.

**Theorem A.4.**

Let $\overrightarrow{X}$ be $n$-variate normally distributed, and let $X_1$ and $X_2$ be subvectors of components of $\overrightarrow{X}$. If $\overrightarrow{X}_1$ and $\overrightarrow{X}_2$ are uncorrelated, that is, $\text{cov}\left[\overrightarrow{X}_1, \overrightarrow{X}_2\right] = \mathbf{0}$, then $\overrightarrow{X}_1$ and $\overrightarrow{X}_2$ are independent.

## A.3. Conditional Distributions of Multivariate Normal Random Variables

Let $Y$ be a scalar random variable and $\overrightarrow{X}$ be a $k$-dimensional random vector. Assume that

$$\begin{pmatrix} Y \\ \overrightarrow{X} \end{pmatrix} \sim N_{k+1}\left[\begin{pmatrix} \mu_Y \\ \overrightarrow{\mu}_X \end{pmatrix}, \begin{pmatrix} \Sigma_{YY} & \overrightarrow{\Sigma}_{YX} \\ \overrightarrow{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{pmatrix}\right],$$

where $\mu_Y = \mathbb{E}\left[Y\right]$ and $\overrightarrow{\mu}_X = \mathbb{E}\left[\overrightarrow{X}\right]$, and

$$\begin{aligned} \Sigma_{YY} &= \text{var}\left[Y\right], \\ \overrightarrow{\Sigma}_{YX} &= \text{cov}\left[Y, \overrightarrow{X}\right] = \mathbb{E}\left[(Y - \mu_Y)\left(\overrightarrow{X} - \overrightarrow{\mu}_X\right)'\right], \\ \overrightarrow{\Sigma}_{XY} &= \text{cov}\left[\overrightarrow{X}, Y\right] = \mathbb{E}\left[\left(\overrightarrow{X} - \overrightarrow{\mu}_X\right)(Y - \mu_Y)\right] = \overrightarrow{\Sigma}'_{YX}, \\ \boldsymbol{\Sigma}_{XX} &= \text{var}\left[\overrightarrow{X}\right]. \end{aligned}$$

To derive the conditional distribution of $Y$, given $\overrightarrow{X}$, let $U = Y - \alpha - \beta'\overrightarrow{X}$, where $\alpha$ is a scalar constant and $\beta$ is a $k \times 1$ vector of constants such that $\mathbb{E}\left[U\right] = 0$ and $U$ and $\overrightarrow{X}$ are independent. It follows from Theorem A.1 that

$$\begin{aligned} \begin{pmatrix} U \\ \overrightarrow{X} \end{pmatrix} &= \begin{pmatrix} -\alpha \\ \overrightarrow{0} \end{pmatrix} + \begin{pmatrix} 1 & -\overrightarrow{\beta}' \\ \overrightarrow{0} & \boldsymbol{I}_k \end{pmatrix} \begin{pmatrix} Y \\ \overrightarrow{X} \end{pmatrix} \\ &\sim N_{k+1}\left[\begin{pmatrix} -\alpha + \mu_Y - \overrightarrow{\beta}'\overrightarrow{\mu}_X \\ \overrightarrow{\mu}_X \end{pmatrix}, \begin{pmatrix} 1 & \overrightarrow{\beta}' \\ \overrightarrow{0} & \boldsymbol{I}_k \end{pmatrix} \begin{pmatrix} \Sigma_{YY} & \overrightarrow{\Sigma}_{YX} \\ \overrightarrow{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{pmatrix} \begin{pmatrix} 1 & \overrightarrow{0}' \\ -\overrightarrow{\beta} & \boldsymbol{I}_k \end{pmatrix}\right]. \end{aligned}$$

The variance matrix involved can be rewritten as

$$\text{var}\left[\begin{pmatrix} U \\ \overrightarrow{X} \end{pmatrix}\right] = \begin{pmatrix} \Sigma_{YY} - \overrightarrow{\Sigma}_{YX}\overrightarrow{\beta} - \overrightarrow{\beta}'\overrightarrow{\Sigma}_{XY} + \overrightarrow{\beta}'\boldsymbol{\Sigma}_{XX}\overrightarrow{\beta} & \overrightarrow{\Sigma}_{YX} - \overrightarrow{\beta}'\boldsymbol{\Sigma}_{XX} \\ \overrightarrow{\Sigma}_{XY} - \boldsymbol{\Sigma}_{XX}\overrightarrow{\beta} & \boldsymbol{\Sigma}_{XX} \end{pmatrix}. \qquad \text{(A.6)}$$

Next, choose $\overrightarrow{\beta}$ such that $U$ and $\overrightarrow{X}$ are uncorrelated and hence independent. In view of (A.6), a necessary and sufficient condition for this is

$$\overrightarrow{\Sigma}_{XY} - \mathbf{\Sigma}_{XX}\overrightarrow{\beta} = \overrightarrow{0};$$

hence,

$$\overrightarrow{\beta} = \mathbf{\Sigma}_{XX}^{-1}\overrightarrow{\Sigma}_{XY}.$$

Moreover, $\mathbb{E}\left[U\right] = 0$ if

$$\alpha = \mu_Y - \overrightarrow{b}'\overrightarrow{\mu}_X.$$

Substituting these values of $\alpha$ and $\beta$ into the expression for the joint distribution of $(U, \overrightarrow{X}')'$, we obtain

$$\begin{pmatrix} U \\ \overrightarrow{X} \end{pmatrix} \sim N_{k+1}\left[\begin{pmatrix} 0 \\ \overrightarrow{\mu}_X \end{pmatrix}, \begin{pmatrix} \Sigma_{YY} - \overrightarrow{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\overrightarrow{\Sigma}_{XY} & \overrightarrow{0}' \\ \overrightarrow{0} & \mathbf{\Sigma}_{XX} \end{pmatrix}\right]. \tag{A.7}$$

Thus, $U$ and $\overrightarrow{X}$ are independent normally distributed, and consequently

$$\mathbb{E}\left[U \mid \overrightarrow{X}\right] = \mathbb{E}\left[U\right] = 0.$$

Because $Y = \alpha + \overrightarrow{\beta}' + U$, we now have

$$\mathbb{E}\left[Y \mid \overrightarrow{X}\right] = \alpha + \overrightarrow{\beta}'\,\mathbb{E}\left[\overrightarrow{X} \mid \overrightarrow{X}\right] + \mathbb{E}\left[U \mid \overrightarrow{X}\right] = \alpha + \overrightarrow{\beta}'\overrightarrow{X}.$$

Moreover, it is easy to verify from (A.7) that the conditional density of $Y$, given $\overrightarrow{X} = \overrightarrow{x}$, is

$$f(y \mid \overrightarrow{x}) = \frac{\exp\left[-\frac{1}{2}\left(y - \alpha - \overrightarrow{\beta}'\overrightarrow{x}\right)^2 / \sigma_u^2\right]}{\sigma_u\sqrt{2\pi}},$$

where

$$\sigma_u^2 = \Sigma_{YY} - \overrightarrow{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\overrightarrow{\Sigma}_{XY}.$$

Furthermore, note that $\sigma_u^2$ is just the conditional variance of $Y$, given $\overrightarrow{X}$,

$$\sigma_u^2 = \operatorname{var}\left[Y \mid \overrightarrow{X}\right] := \mathbb{E}\left[\left(Y - \mathbb{E}\left[Y \mid \overrightarrow{X}\right]\right)^2 \mid \overrightarrow{X}\right].$$

These results are summarized in the following theorem:

**Theorem A.5.**

Let

$$\begin{pmatrix} Y \\ \overrightarrow{X} \end{pmatrix} \sim N_{k+1}\left[\begin{pmatrix} \mu_Y \\ \overrightarrow{\mu}_X \end{pmatrix}, \begin{pmatrix} \Sigma_{YY} & \overrightarrow{\Sigma}_{YX} \\ \overrightarrow{\Sigma}_{XY} & \mathbf{\Sigma}_{XX} \end{pmatrix}\right],$$

where $Y \in \mathbb{R}, X \in \mathbb{R}^k$, and $\mathbf{\Sigma}_{XX}$ is nonsingular. Then, conditionally on $\overrightarrow{X}$, $Y$ is normally distributed with conditional expectation $\mathbb{E}[Y \mid \vec{X}] = \alpha + \vec{\beta}'\vec{X}$, where

$$\vec{\beta} = \boldsymbol{\Sigma}_{XX}^{-1}\vec{\Sigma}_{XY}, \quad \alpha = \mu_Y - \vec{\beta}'\vec{\mu}_X,$$

and conditional variance

$$\text{var}[Y \mid \vec{X}] = \Sigma_{YY}^2 - \vec{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\vec{\Sigma}_{XY}$$

The results in Theorem A.5 is the basis for linear regression analysis. Suppose that $Y$ measures an economic activity that is partly caused or influenced by other economic variables measured by the components of the random vector $\vec{X}$. In applied economics, *the relation between $Y$, called the dependent variable, and the components of $\vec{X}$, called the independent variables or the regressors, is often modeled linearly as $Y = \alpha + \vec{\beta}'\vec{X} + U$, where $\alpha$ is the intercept, $\vec{\beta}$ is the vector of slope parameters, and $U$ is an error term that is usually assumed to be independent of $\vec{X}$ and normally $N(0, \sigma^2)$ distributed. Theorem A.5 shows that if $\vec{X}$ and $Y$ are jointly normally distributed, then such a linear relationship between $Y$ and $\vec{X}$ exits.*

## A.4. Independence of Linear and Quadratic Transformation of Multivariate Normal Random Variables

Let $\vec{X}$ be distributed $N_n(\vec{0}, \boldsymbol{I}_n)$ – that is, $\vec{X}$ is $n$-variate, standard, normally distributed. Consider the linear transformation $\vec{Y} = \boldsymbol{B}\vec{X}$, where $\boldsymbol{B}$ is a $k \times n$ matrix of constants, and $\vec{Z} = \boldsymbol{C}\vec{X}$, where $\boldsymbol{C}$ is an $m \times n$ matrix of constants. It follows from Theorem A.4 that

$$\begin{matrix} \vec{Y} \\ \vec{X} \end{matrix} \sim N_{k+m}\left[\begin{pmatrix} \vec{0} \\ \vec{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{BB}' & \boldsymbol{BC}' \\ \boldsymbol{CB}' & \boldsymbol{CC}' \end{pmatrix}\right].$$

Then $\vec{Y}$ and $\vec{Z}$ are uncorrelated and therefore independent if and only if

$$\boldsymbol{BC}' = \mathbf{0}.$$

More generally, we have

**Theorem A.6.**

Let $\vec{X}$ be distributed $N_n(\vec{0}, \boldsymbol{I}_n)$, and consider the linear transformation $\vec{Y} = \vec{b} + \boldsymbol{B}\vec{X}$, where $\vec{b}$ is a $k \times 1$ vector of contants and $\boldsymbol{B}$ a $k \times n$ matrix of constants, and $\vec{Z} = \vec{c}\boldsymbol{C}\vec{X}$, where $\vec{c}$ is an $m \times 1$ vector of constants and $\boldsymbol{C}$ is an $m \times n$ matrix of constants. Then $\vec{Y}$ and $\vec{Z}$ are independent if and only if

$$\boldsymbol{BC}' = \mathbf{0}.$$

This result can be used to set forth conditions for independence of linear and quadratic transformations of standard normal random vectors:

**Theorem A.7.**

Let $\overrightarrow{X}$ and $\overrightarrow{Y}$ be defined as in Theorem A.6, and let $Z = \overrightarrow{X}'\boldsymbol{C}\overrightarrow{X}$, where $\boldsymbol{C}$ is a symmetric $n \times n$ matrix of constants. Then $\overrightarrow{Y}$ and $Z$ are independent if and only if

$$\boldsymbol{BC} = \boldsymbol{0}.$$

Finally, consider the conditions for independence of two quadratic forms of standard normal random vectors:

**Theorem A.8.**

Let $\overrightarrow{X} \sim N_n(\overrightarrow{0}, \boldsymbol{I}_n)$, $Z_1 = \overrightarrow{X}'\boldsymbol{A}\overrightarrow{X}$, $Z_2 = \overrightarrow{X}'\boldsymbol{B}\overrightarrow{X}$, where $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric $n \times n$ matrices of constants. Then $Z_1$ and $Z_2$ are independent if and only if

$$\boldsymbol{AB} = \boldsymbol{0}.$$

As we will see in Section A.5, quadratic forms of multivariate normal random variables play a key role in statistical testing theory. The two most important results are stated in Theorems A.9 and A.10:

**Theorem A.9.**

Let $\overrightarrow{X} \sim N_n(\overrightarrow{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is nonsingular. Then $\overrightarrow{X}'\boldsymbol{\Sigma}^{-1}\overrightarrow{X}$ is distributed as $\chi^2_n$.

*Proof.* Denote $\overrightarrow{Y} = \boldsymbol{\Sigma}^{-1/2}\overrightarrow{X}$. Then $\overrightarrow{Y}$ is $n$-variate, standard, normally distributed; hence, $Y_1, \ldots, Y_n$ are i.i.d. $N(0,1)$, and thus,

$$\overrightarrow{X}'\boldsymbol{\Sigma}^{-1}\overrightarrow{X} = \overrightarrow{Y}'\overrightarrow{Y} = \sum_{j=1}^{n} Y_j^2 \sim \chi^2_n.$$

□

**Theorem A.10.**

Let $\overrightarrow{X} \sim N_n(\overrightarrow{0}, \boldsymbol{I}_n)$, and let $\boldsymbol{M}$ be a symmetric idempotent $n \times n$ matrix of constants with rank $k$. Then $\overrightarrow{X}'\boldsymbol{M}\overrightarrow{X}$ is distributed as $\chi^2_k$.

*Proof.* Recall that a square matrix $\boldsymbol{M}$ is idempotent if $\boldsymbol{M}^2 = \boldsymbol{M}$. If $\boldsymbol{M}$ is also symmetric, then we can write $\boldsymbol{M} = \boldsymbol{Q\Lambda Q}'$, where $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues of $\boldsymbol{M}$ and $\boldsymbol{Q}$ is the corresponding orthogonal matrix of eigenvectors. Then $\boldsymbol{M}^2 = \boldsymbol{M}$ implies $\boldsymbol{\Lambda}^2 = \boldsymbol{\Lambda}$; hence, the eigenvalues of $\boldsymbol{M}$ are either 1 or 0. If all eigenvalues are 1, then $\boldsymbol{\Lambda} = \boldsymbol{I}$; hence $\boldsymbol{M} = \boldsymbol{I}$. *Thus, the only nonsingular symmetric idempotent matrix is the identity matrix.* Consequently, the concept of a symmetric idempotent matrix is only meaningful if the matrix involved is singular.

The rank of a symmetric idempotent matrix $\boldsymbol{M}$ equals the number of nonzero eigenvalues; hence, $\operatorname{tr}(\boldsymbol{M}) = \operatorname{tr}(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}') = \operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{Q}\boldsymbol{Q}') = \operatorname{tr}(\boldsymbol{\Lambda}) = \operatorname{rank}(\boldsymbol{M})$.

From the above explanation, we can write

$$\boldsymbol{M} = \boldsymbol{Q}\begin{pmatrix} \boldsymbol{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\boldsymbol{Q}',$$

where $\boldsymbol{Q}$ is the orthogonal matrix of eigenvectors. Because $\overrightarrow{Y} = (Y_1, \dots, Y_n)' = \boldsymbol{Q}'\overrightarrow{X} \sim N_n(\overrightarrow{0}, \boldsymbol{I}_n)$, we now have

$$\overrightarrow{X}'\boldsymbol{M}\overrightarrow{X} = \overrightarrow{Y}'\begin{pmatrix} \boldsymbol{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\overrightarrow{Y} = \sum_{j=1}^{k} Y_j^2 \sim \chi_k^2.$$

□

## A.5. Applications to Statistical Inference under Normality

### A.5.1. Estimation