Econometrics Lecture Notes

Brennan Thompson*

11th September 2006

1 Review of probability and statistics

1.1 Random variables

In the simplest of terms, a **random variable** (or **RV**) is a variable whose value is determined by the **outcome** of a random experiment.[1] In what follows, we will denote an RV by an uppercase letter (such as X or Y), and an outcome by a lowercase letter (such as x or y).

As a simple example, consider the random experiment of rolling a single die. Let X be the RV determined by the outcome of this experiment. In this experiment, there are six possible outcomes: If the die lands with 1, 2, 3, 4, 5, or 6, facing up, we assign that value to the RV X.

This example illustrates what is known as a **discrete random variable**: an RV which can only take on a finite (or countably infinite) number of values. Here, the RV X can take only take on the values only the value of 1, 2, 3, 4, 5, or 6.

A discrete RV differs from a **continuous random variable**: an RV which can take on any value in some real interval (and therefore an uncountably infinite number of values). An example of continuous RV is a person's height. For an adult, this could take on any value between (say) 4 and 8 feet, depending on the precision of measurement.

1.2 Probability functions

Associated with every RV is a **probability function**, which in the case of a discrete RV, we will call a **probability mass function**, and in the case of a continuous RV, we will call a **probability density function** (or **PDF**).

Consider first the discrete case. The probability mass function $f(x)$ tells us the probability that the discrete RV X will take on the value x. More formally,

$$f(x) = \text{Prob}(X = x).$$

*Department of Economics, Ryerson University, 350 Victoria Street, Toronto, Ontario, M5B 2K3, Canada. Email: brennan@ryerson.ca

[1] From this definiton, it should be obvious that an RV differs from a **constant**, which is a variable whose value is fixed.

It should be noted that a probability mass function for a discrete RV requires that

$$0 \leq f(x) \leq 1,$$

and[2]

$$\sum_x f(x) = 1.$$

In our die rolling example, it should be clear that, since each outcome (1, 2, 3, 4, 5, or 6) is equally likely, the probability mass function for X, $f(x)$, is equal to $\frac{1}{6}$ for each x, i.e.

$$f(x) = \frac{1}{6}, \text{ for } x = 1, \ldots, 6.$$

Such a probability mass function is said to represent a **uniform distribution**, since each outcome has an equal (or uniform) probability of occurring.

For a continuous RV, the probability associated with any particular value is zero, so we can only assign positive probabilities to ranges of values. For example, if we are interested in the probability that the continuous RV X will take on a value in the range between a and b, the PDF $f(x)$ is implicitly defined as

$$\text{Prob}(a \leq X \leq b) = \int_a^b f(x)dx.$$

As in the discrete case, a PDF for a continuous RV has certain (roughly analogous) requirements that must be satisfied:

$$f(x) \geq 0,$$

and[3]

$$\int_x f(x)dx = 1.$$

1.3 Distribution functions

Closely related to the concept of a probability function is a **distribution function** (or **cumulative distribution function** or **CDF**), which tells us the probability that an RV will take on any value *less than or equal to* some specific value. That is, for an RV X (whether discrete or continuous), the CDF $F(x)$ is defined as

$$F(x) = \text{Prob}(X \leq x).$$

It should be fairly easy to see that for a discrete RV X, this implies that

$$F(x) = \sum_{X \leq x} f(x).$$

[2] The notation $\sum_x$ is used to denote the sum over the entire range of values that the discrete RV X can take on.

[3] The notation $\int_x$ is used to denote the integral over the entire range of values that the continuous RV X can take on.

In our example of rolling a single die, it should be easy to see that the CDF of X is

$$F(x) = \frac{x}{6}, \text{ for } x = 1, \ldots, 6.$$

For a continuous RV X, the definition of a CDF implies that

$$F(x) = \int_{-\infty}^{x} f(t)dt,$$

and

$$f(x) = \frac{dF(x)}{dx}.$$

Finally, it should be noted that for any RV X (whether discrete or continuous), the CDF $F(x)$ must satisfy

$$0 \leq F(x) \leq 1,$$

$$F(+\infty) = 1,$$

$$F(-\infty) = 0,$$

and

$$F(x_1) \geq F(x_2) \text{ for } x_1 > x_2.$$

1.4 Joint probability functions

A probability function defined over more than one RV is known as a **joint probability function**. For instance, in the case of two discrete RVs, X and Y, the **joint probability mass function** $f(x, y)$ tells us the probability that the RV X will take on the value x *and* the RV Y will take on the value y (simultaneously). That is,

$$f(x, y) = \text{Prob}(X = x, Y = y).$$

Similar to the requirements of a probability mass function (for a single discrete RV), we must have

$$0 \leq f(x, y) \leq 1,$$

and

$$\sum_x \sum_y f(x, y) = 1.$$

For example, consider two dice, one red and one blue. Denote by X the RV whose value is determined by the random experiment of rolling the red die, and by Y the RV whose value is determined by the random experiment of rolling the blue die. It should be emphasized that these are two *separate* random experiments (i.e. do not confuse this with the *single* random experiment of tossing a pair of dice). However, the possible outcomes of these two random experiments can be combined in the following manner: let (x, y) denote the combined outcome that X takes on the value x, and Y takes on the value y.

For example, the combined outcome $(1, 5)$ represents rolling a 1 with the red die, and a 5 with the blue die. In total, there are 36 such combined outcomes, each with equal probability, $\frac{1}{36}$. Accordingly, we can write the joint probability mass function for X and Y as

$$f(x, y) = \frac{1}{36}, \text{ for } x = 1, \ldots, 6 \text{ and } y = 1, \ldots, 36.$$

In the continuous case, we may be interested in the probability that the RV X takes on some value in the range between a and b and the RV Y will take on some value between c and d (simultaneously). The **joint probability density function** (or **joint PDF**) $f(x, y)$ is defined so that

$$\text{Prob}(a \leq x \leq b, c \leq y \leq d) = \int_a^b \int_c^d f(x, y) dx dy.$$

The requirements here are similar to that of a PDF (for a single continuous RV):

$$f(x, y) \geq 0,$$

and

$$\int_x \int_y f(x, y) dx dy = 1.$$

1.5 Conditional probability functions

Closely related to the concept of a joint probability function is a **conditional probability function**. For two discrete RVs, X and Y, the **conditional probability mass function**, $f(y|x)$, tell us the probability that Y takes on the value y, given that X takes on the value x, i.e.

$$f(y|x) = \Pr(Y = y|X = x),$$

and is equal to

$$\begin{aligned} f(y|x) &= \frac{\Pr(Y = y, X = x)}{\Pr(\mathrm{X = x})} \\ &= \frac{f(y, x)}{f(x)}. \end{aligned}$$

In the continuous case, a **conditional probability density function** (or **conditional PDF**) can be used to tell us the probability that a continuous RV takes on a value in some range, given that another continuous RV takes on some specific value. Of course, as noted above, the probability that a continuous RV takes on a specific outcome is zero, so we need to make some approximation to this.

Suppose that we want to know the probability that the continuous RV Y takes on some value in the range between a and b, given that the continuous RV X takes on some value extremely close to x. The probability that the continuous

RV X takes on some value extremely close to x, say in the range between x and $x+h$, where h is some extremely small number, is

$$\begin{aligned}\Pr(x \leq X \leq x+h) &= \int_x^{x+h} f(v)dv. \\ &\approx hf(x).\end{aligned}$$

Now, the probability that Y takes on some value in the range between a and b and X takes on some value in the range between x and $x+h$ is

$$\begin{aligned}\Pr(a \leq Y \leq b, x \leq X \leq x+h) &= \int_a^b \int_x^{x+h} f(y,v)dydv \\ &\approx \int_a^b hf(y,x)dy.\end{aligned}$$

So, similar to the discrete case, the probability that Y takes on some value in the range between a and b given X takes on some value in the range between x and $x+h$ is

$$\begin{aligned}\Pr(a \leq Y \leq b | x \leq X \leq x+h) &= \frac{\Pr(a \leq Y \leq b, x \leq X \leq x+h)}{\Pr(x \leq X \leq x+h)} \\ &\approx \frac{\int_a^b hf(y,x)dy}{hf(x)} \\ &= \frac{\int_a^b f(y,x)dy}{f(x)}.\end{aligned}$$

From this approximation, we implicitly define that conditional PDF $f(y|x)$ as

$$\Pr(a \leq Y \leq b | X \approx x) = \int_a^b f(y|x)dy.$$

1.6 Expected value

Roughly speaking, the **expected value** or **mean** of an RV, is the average value we would expect it to take on if we were to repeat the random experiment which determines its value many times over. It is what is known as a measure of *central tendency*.[4] We generally use the notation $E(X)$ when we speak of the expected value of the the RV X, and the notation μ_x when we speak of its mean. However, since these terms are interchangeable, so is the notation.

The mathematical expectation of a discrete RV X is defined as

$$E(X) = \sum_x xf(x),$$

[4] Other measures of central tendency are the **median** and **mode**. The median of the RV X (whether discrete or continuous) is the value m such that $\text{Prob}(X \leq m) \geq 0.5$ and $\text{Prob}(X \geq m) \leq 0.5$, while the mode of X is the value of x at which $f(x)$ takes it maximum.

while for a continuous RV X we have

$$E(X) = \int xf(x)dx.$$

It should be clear that these definitions are quite similar, in that the different possible values of the RV are weighted by the probability attached to them. In the discrete case, these weighted values are summed, while in the continuous case, integration is used.

It should go without saying that *the expected value of a constant is the value of the constant itself.* That is, for any constant c, we have

$$E(c) = c.$$

On the other hand, it is not so obvious that *the expected value of an RV is a constant.* Using the preceding rule, we therefore have

$$E[E(x)] = E(x),$$

which might be a little less confusing if we use the notation

$$E(\mu_X) = \mu_X.$$

In our example of rolling a single die, we can calculate the expected value of X as

$$E(X) = \frac{1}{6}(1) + \frac{1}{6}(2) + \frac{1}{6}(3) + \frac{1}{6}(4) + \frac{1}{6}(5) + \frac{1}{6}(6) = 3.5.$$

In a more general way, we can also calculate the expected value of a function of a RV. Letting $g(X)$ be some function of the discrete RV X, the expected value of this function is

$$E[g(X)] = \sum_x g(x)f(x).$$

Similarly, letting $g(X)$ be some function of a continuous RV X, the expected value of this function is

$$E[g(X)] = \int g(x)f(x)dx.$$

Note that *a function of an RV is an RV itself.* To see this, it might help to imagine a new RV, call it Z, which is a function of our original RV, X (i.e. $Z = g(X)$).

As an example, let's again consider the RV X, whose value is determined from the outcome of tossing a single die. Suppose we have $g(X) = 4X$ (this is what is known as a **linear function**). Again, it might help to imagine a new RV called Z, where, here, $Z = g(X) = 4X$. The expected value of Z can be calculated as

$$E(Z) = E(4X) = \frac{1}{6}(4) + \frac{1}{6}(8) + \frac{1}{6}(12) + \frac{1}{6}(16) + \frac{1}{6}(20) + \frac{1}{6}(24) = 14.$$

Notice that this is just four times the expected value of X (which, as we saw above, is 3.5).

This example illustrates an important rule: *The expected value of an RV multiplied by a constant, is the constant multiplied by the expected value of that RV.* That is, if X is an RV and c is a constant, we have

$$E(cX) = cE(X).$$

The proof of this follows directly from the definition of the mathematical expectation of a discrete RV:[5]

$$\begin{aligned} E(cX) &= \sum_x cxf(x) \\ &= c\sum_x xf(x) \\ &= cE(X). \end{aligned}$$

Continuing with our random experiment of tossing a single die, consider the function $g(X) = X^2$ (this is what is known as a **nonlinear function**). Once again, it might help to imagine a new RV called Z, where $Z = g(X) = X^2$. The expected value of Z can be calculated as

$$E(Z) = E(X^2) = \frac{1}{6}(1) + \frac{1}{6}(4) + \frac{1}{6}(9) + \frac{1}{6}(16) + \frac{1}{6}(25) + \frac{1}{6}(36) = \frac{91}{6} \approx 15.17.$$

It is important to note that $E(X^2)$ (the expected value of the square of the RV X) is *not* equal to $[E(X)]^2$ (the square of the expected value of the RV Y), which in this case would be $3.5^2 = 12.25$.

Quite often, we are interested in functions of more than one RV. Consider a function, $g(X, Y)$, of two discrete RVs, X and Y. The expected value of this function is

$$E[g(X, Y)] = \sum_x \sum_y g(x, y)f(x, y),$$

where $f(x, y)$ is the joint probability mass function of X and Y.

Similarly, for two continuous RVs, X and Y, the expected value of the function $g(X, Y)$ is

$$E[g(X, Y)] = \int_x \int_y g(x, y)f(x, y)dxdy,$$

where $f(x, y)$ is the joint PDF of X and Y.

Consider now our example of rolling two separate die, one red and one blue. Now, consider the function $g(X, Y) = X + Y$. As above, it might help to once

[5] This rule also holds for continuous RVs, and a similar proof could be made.

again imagine a new RV called Z, where $Z = g(X,Y) = X + Y$.[6] The expected value of Z can be calculated as

$$\begin{aligned} E(Z) &= E(X+Y) \\ &= \frac{1}{36}(2) + \frac{2}{36}(3) + \frac{3}{36}(4) + \frac{4}{36}(5) + \frac{5}{36}(6) + \frac{6}{36}(7) \\ &\quad + \frac{5}{36}(8) + \frac{4}{36}(9) + \frac{3}{36}(10)\frac{2}{36}(11) + \frac{1}{36}(12) \\ &= 7. \end{aligned}$$

Note that this is just equal to the expected value of X (3.5) plus the expected value of Y (3.5).

This example illustrates another important rule: *The expected value of the sum of two (or more) RVs is the sum of their expected values.* In general terms, for any two RVs, X and Y, we have[7]

$$E(X+Y) = E(X) + E(Y).$$

After seeing this, it might be tempting to that that the same thing holds for the product of several RVs, (i.e. that $E(XY) = E(X)E(Y)$).While this is true in a certain special case (known as **statistical independence**), it is not true in general.

1.7 Conditional expectation

Often, we are interested in knowing what the expected value of an RV is, given that another RV takes on a certain value. Consider two discrete RVs, X and Y. The **conditional expectation** of Y given that X takes on the value x is

$$E(Y|X = x) = \sum_y f(y|x).$$

For two continuous RVs, X and Y, the conditional expectation of Y given that X takes on some value extremely close to x (see above) is

$$E(Y|X \approx x) = \int_y f(y|x)dy.$$

1.8 Variance

Roughly speaking, the **variance** of an RV, is the average squared distance from its expected value that we would expect it to take on if we were to repeat the

[6] Note that, while Z can be seen as the RV whose value is determined by the outcome of the *single* random experiment of rolling a pair of dice, we are defining it here as the RV whose value is determined by a function of two *separate* RVs, determined by the outcomes of two *separate* random experiments.

[7] As above, this rule also holds for continuous RVs. We omit the proof for both cases here.

random experiment which determines its value many times over. More formally, the variance of the RV X (whether discrete or continuous), usually denoted σ_X^2 (or sometimes $\text{Var}(X)$), is defined as[8]

$$\sigma_X^2 = E[(X - \mu_X)^2].$$

From this definition, it should be clear that *the variance of a constant is zero*. The proof of this is quite trivial: for some constant c we have

$$\begin{aligned} \sigma_c^2 &= E[(c - \mu_c)^2] \\ &= E[(c - c)^2] \\ &= 0. \end{aligned}$$

The variance of an RV is known as a measure of *dispersion*. Another measure of dispersion for an RV is its **standard deviation**, which is just the positive square root root of its variance. The standard deviation of the RV X is usually denoted σ_X.

Notice that the variance of an RV is just the expectation of a function of that RV (i.e. the variance of the RV X is the expected value of the function $(X - \mu_X)^2$, as defined above).[9] Keeping this in mind, we can write the variance of the discrete RV X as

$$\sigma_X^2 = \sum_x (x - \mu_X)^2 f(x),$$

while the variance of the continuous RV X can be written as

$$\sigma_X^2 = \int (x - \mu_X)^2 f(x) dx.$$

In our example of tossing a single die, the variance of X (remembering that $\mu_X = 3.5$), is calculated as

$$\begin{aligned} \sigma_X^2 &= \frac{1}{6}(1 - 3.5)^2 + \frac{1}{6}(2 - 3.5)^2 + \frac{1}{6}(3 - 3.5)^2 \\ &\quad + \frac{1}{6}(4 - 3.5)^2 + \frac{1}{6}(5 - 3.5)^2 + \frac{1}{6}(6 - 3.5)^2 \\ &= 17.5, \end{aligned}$$

which implies

$$\sigma_X = \sqrt{17.5} \approx 4.18.$$

Before moving on, we can also consider the variance of a function of an RV. Consider the linear function $g(X) = cX$ where X is an RV and c is a constant.[10]

[8] At times, it may be convenient to define the variance the RV X as $\sigma_X^2 = E(X^2) - \mu^2$. It is a useful exercise to show that these two definitions are equivalent.

[9] While this *function* is also an RV, its expected value (i.e. its variance) is actually a constant (just like its expected value, as we pointed out above).

[10] Here, we will focus on linear functions, but the analysis could be extended to more general function. We will consider the variance of a function of more than one RV in the next section.

The variance of this function is

$$\begin{aligned}\text{Var}(cX) &= E[(cX - \mu_{cX})^2] \\ &= E[(cX - c\mu_X)^2] \\ &= E[c^2(X - \mu_X)^2] \\ &= c^2 E[(X - \mu_X)^2] \\ &= c^2\sigma_X^2.\end{aligned}$$

1.9 Covariance

Closely related to the concept of variance is **covariance**, which is a measure of *association* between two RVs. The covariance between the two RVs X and Y (whether discrete or continuous), usually denoted $\sigma_{X,Y}$ (or sometimes $\text{Cov}(X, Y)$), is defined as[11]

$$\sigma_{X,Y} = E[(X - \mu_X)(Y - \mu_Y)].$$

This definition should make it clear that the covariance of an RV with itself is just equal to the variance of that RV. In other words, the covariance of the RV X with itself is

$$\begin{aligned}\sigma_{X,X} &= E[(X - \mu_X)(X - \mu_X)] \\ &= E[(X - \mu_X)^2] \\ &= \sigma_X^2.\end{aligned}$$

Just as we viewed variance as the expectation of a function of a (single) RV, we can view covariance as the expectation of a function of two RVs. For two discrete RVs X and Y, this means[12]

$$\sigma_{X,Y} = \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x, y),$$

while for two continuous RVs X and Y, this means

$$E[g(X, Y)] = \int_x \int_y (x - \mu_X)(y - \mu_Y) f(x, y) dx dy.$$

As mentioned above, two RVs X and Y are said to be statistically independent if

$$E(XY) = E(X)E(Y).$$

[11] We can also define the covariance of the RVs X and Y as $\sigma_{X,Y} = E(XY) - \mu_X\mu_Y$. As was the case with the definitions of variance, it is a useful exercise to show that these two are equivalent.

[12] Here, the expectation is taken of the function $g(X, Y) = (X - \mu_X)(Yy - \mu_Y)$.

Therefore, using the alternate definition of covariance, it can be see that if two RVs X and Y are statistically independent, their covariance is zero:

$$\begin{aligned}\sigma_{X,Y} &= E(XY) - \mu_X \mu_y \\ &= E(X)E(Y) - \mu_X \mu_y \\ &= 0\end{aligned}$$

Using the definition of covariance, we can now consider the variance of a function of more than one RV. Consider the linear function $g(X,Y) = X + Y$, where X and Y are two RVs.[13] The variance of this function is

$$\begin{aligned}Var(X+Y) &= E[(X+Y-\mu_{X+Y})^2] \\ &= E[(X+Y-\mu_X-\mu_Y)^2] \\ &= E(X^2+Y^2+2XY-2X\mu_X-2X\mu_Y \\ &\quad -2Y\mu_X-2Y\mu_Y+2\mu_X\mu_Y+\mu_X^2+\mu_Y^2) \\ &= E(X^2-2X\mu_X+\mu_X^2)+E(Y^2-2Y\mu_Y+\mu_Y^2) \\ &\quad +2E(XY-X\mu_X-Y\mu_Y+\mu_X\mu_Y) \\ &= \sigma_X^2+\sigma_Y^2+2\sigma_{X,Y}.\end{aligned}$$

Of course, if X and Y are statistically independent,

$$\text{Var}(X+Y) = \sigma_X^2 + \sigma_Y^2$$

since $\sigma_{X,Y}$ is equal to zero.

A similar analysis (omitted here) could be used to show that the variance of the function $g(X,Y) = X - Y$ is equal to

$$\text{Var}(X-Y) = \sigma_X^2 + \sigma_Y^2 - 2\sigma_{X,Y},$$

unless X and Y are statistically independent, in which case

$$\text{Var}(X-Y) = \sigma_X^2 + \sigma_Y^2.$$

A related measure of association is **correlation**. The **coefficient of correlation** between the two RVs X and Y, denoted $\rho_{X,Y}$, is defined as

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}.$$

Note that if X and Y are statistically independent, the coefficient of correlation is zero (since $\sigma_{X,Y} = 0$), in which case we would say they are said to be uncorrelated.

[13] We again focus on only linear functions, but more general functions could be considered.

1.10 Estimators

In order to calculate the measures considered above (expected value, variance, and covariance), it is necessary to know something about the probability function associated with the RV (or, in the case of covariance, the joint probability function associated with the RVs) of interest. In the die rolling example, the RV X was known *a priori* to be uniformly distributed. Given this, we were able to calculate the expected value and variance of X quite easily.

However, the probability function for an RV determined by anything but the most trivial random experiment is usually unknown. In such cases, it is common to *assume* that the RV follows some theoretical distribution (such as the uniform distribution). For example, if we assume that the continuous RV X follows a **normal distribution**, we will use the PDF

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(\frac{-(x-\mu_X)^2}{2\sigma^2}\right), \text{ for } -\infty < x < +\infty,$$

which requires us to know the parameters μ_X and σ_X^2.[14] The problem is that we can't derive these parameters unless we first know the PDF (which requires knowing the parameters). In some cases we may choose to assume values for them, but in others, we need to *estimate* them using a sample of observations.[15]

A **random sample** of n independently drawn observations on the RV X is a set of RVs, $x_1, \ldots, x_n$, each of which has the same distribution as X. For this reason, we say that the RVs $x_1, \ldots, x_n$ are **independent, identically distributed** (or **IID**).

Let's start by considering an **estimator** of the mean of an RV, called the **sample mean**. The usual estimator of the mean of the RV X, denoted $\bar{X}$, is

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

where $x_1, \ldots, x_n$ is a random sample of observations on the RV X.

Next, consider an estimator of the the variance of an RV, called the **sample variance**. For the RV X, the usual estimator of variance, denoted s_X^2, is

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2.$$

It should be clear right away that both the sample mean and sample variance of an RV are functions of RVs (the random sample), and are therefore RVs themselves. Simply put, *estimators are RVs.*

[14] If the RV X follows a normal distribution with mean μ_x and variance σ_X^2, it is common to write $X \sim N(\mu_X, \sigma_X^2)$.

[15] In fact, if we are not comfortable in assuming some theoretical distribution, it is even possible that we can estimate the probability function itself using a sample of observations. This approach is considered *nonparametric*, since it does not involve estimating the parameters of some theoretical distribution (which is a *parametric* approach).

As RVs, estimators have their own probability function, distribution function, expected value, variance, etc. For example, the expected value of the sample mean of the RV X is

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\ &= \frac{1}{n}E\left(\sum_{i=1}^{n} x_i\right) \\ &= \frac{1}{n}[E(x_1) + \ldots + E(x_n)] \\ &= \frac{1}{n}(\mu_X + \ldots + \mu_X) \\ &= \mu_X. \end{aligned}$$

This demonstrates that the sample mean is what is a called an **unbiased** estimator: an estimator of a parameter whose expected value is the true value of that parameter.

The variance of the sample mean of the RV X is

$$\begin{aligned} \mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\ &= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} x_i\right) \\ &= \frac{1}{n^2}\left[\mathrm{Var}(x_1) + \ldots + \mathrm{Var}(x_n)\right] \\ &= \frac{1}{n^2}\left[\sigma_X^2 + \ldots + \sigma_X^2\right] \\ &= \frac{\sigma_X^2}{n}. \end{aligned}$$

Turning to the sample variance, it turns out that this estimator is also unbiased. The proof is a little more complicated, but it will help to begin by rewriting the sample variance of the RV X as

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}[(x_i - \mu_X) - (\bar{X} - \mu_X)]^2$$

(note that this is equivalent to our original expression since the μ_X terms cancel). The expected value of s_X^2 is thus

$$E(s_X^2) = E\left(\frac{1}{n-1}\sum_{i=1}^{n}[(x - \mu_X) - (\bar{X} - \mu_X)]^2\right)$$

$$
\begin{aligned}
&= \frac{1}{n-1} E\left(\sum_{i=1}^{n} \left[(x_i - \mu_X)^2 - 2(\bar{X} - \mu_X)(x_i - \mu_X) + (\bar{X} - \mu_X)^2 \right] \right) \\
&= \frac{1}{n-1} E\left(\sum_{i=1}^{n} (x_i - \mu_X)^2 - 2(\bar{X} - \mu_X) \sum_{i=1}^{n} (x_i - \mu_X) + n(\bar{X} - \mu_X)^2 \right) \\
&= \frac{1}{n-1} E\left(\sum_{i=1}^{n} (x_i - \mu_X)^2 - 2n(\bar{X} - \mu_X)^2 + n(\bar{X} - \mu_X)^2 \right) \\
&= \frac{1}{n-1} E\left(\sum_{i=1}^{n} (x_i - \mu_X)^2 \right) - \frac{n}{n-1} E[(\bar{X} - \mu_X)^2] \\
&= \frac{1}{n-1} \left[E(x_1 - \mu_X)^2 + \ldots + E(x_n - \mu_X)^2 \right] - \frac{n}{n-1} E\{[\bar{X} - E(\bar{X})]^2\} \\
&= \frac{1}{n-1} (\sigma_X^2 + \ldots + \sigma_X^2) - \frac{n}{n-1} \mathrm{Var}(\bar{X}) \\
&= \frac{n\sigma_X^2}{n-1} - \frac{n}{n-1} \frac{\sigma_X^2}{n} \\
&= \sigma_X^2.
\end{aligned}
$$

We won't do so here, but it can shown that the variance of s_X^2, if X is a normally distributed RV, is

$$\mathrm{Var}(s_X^2) = \frac{2\sigma_X^4}{n-1}.$$

It is useful to compare the estimator s_X^2 with an alternative estimator of variance. Another commonly used estimator of variance of the RV X, denoted $\hat{\sigma}_X^2$, is defined as

$$\hat{\sigma}_X^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{X})^2 = \frac{n-1}{n} s_X^2.$$

The expected value of $\hat{\sigma}_X^2$ is

$$
\begin{aligned}
E(\hat{\sigma}_X^2) &= \frac{n-1}{n} E(s_X^2) \\
&= \frac{n-1}{n} \sigma_X^2,
\end{aligned}
$$

which implies that the estimator $\hat{\sigma}_X^2$ is **biased** (i.e. it is not unbiased). The **bias** of an estimator is equal to the expected value of the estimator minus the parameter being estimated. For $\hat{\sigma}_X^2$, this is

$$
\begin{aligned}
\mathrm{Bias}(\hat{\sigma}_X^2) &= E(\hat{\sigma}_X^2 - \sigma_X^2) \\
&= E(\hat{\sigma}_X^2) - E(\sigma_X^2) \\
&= \frac{n-1}{n} \sigma_X^2 - \sigma_X^2 \\
&= \frac{-1}{n} \sigma_X^2
\end{aligned}
$$

If X is normally distributed, we can use the variance of s_X^2 to find the variance of $\hat{\sigma}_X^2$:

$$\begin{aligned}
\text{Var}(\hat{\sigma}_X^2) &= \text{Var}\left(\frac{n-1}{n}s_X^2\right) \\
&= \left(\frac{n-1}{n}\right)^2 \text{Var}(s_X^2) \\
&= \left(\frac{n-1}{n}\right)^2 \frac{2\sigma_X^4}{n-1} \\
&= \frac{2(n-1)\sigma_X^4}{n^2}.
\end{aligned}$$

Note that this is less than the variance of s_X^2 , i.e.

$$\text{Var}(\hat{\sigma}_X^2) < \text{Var}(s_X^2).$$

Since it has a lower variance, we would say that $\hat{\sigma}_X^2$ is a more **efficient** estimator of variance than s_X^2. Note that efficiency is a relative measure (i.e. we can only speak of the efficiency of an estimator in comparison to the efficiency of other estimators). For this reason, we often refer to the **relative efficiency** of two estimators, which is just equal to the ratio of their variances. For example, the relative efficiency of s_X^2 and $\hat{\sigma}_X^2$ is

$$\frac{\text{Var}(s_X^2)}{\text{Var}(\hat{\sigma}_X^2)} = \left(\frac{n}{n-1}\right)^2.$$

However, we may say that an estimator is "efficient", if it is more efficient than any other estimator.

Although s_X^2 has a lower bias than $\hat{\sigma}_X^2$, it is not as efficient. This raises the important question of which estimator is "better". Typically (as in this case), we see a trade-off between bias and efficiency. A commonly used measure that takes both of these factors into consideration is the **mean squared error** (or **MSE**) of an estimator, which is simply equal to the variance of the estimator plus the square of the bias of the estimator.[16] For s_X^2, we have

$$\begin{aligned}
\text{MSE}(s_X^2) &= \text{Var}(s_X^2) + [\text{Bias}(s_X^2)]^2 \\
&= \frac{2\sigma_X^4}{n-1} + [0]^2 \\
&= \frac{2\sigma_X^4}{n-1},
\end{aligned}$$

[16] Actually, the MSE of an estimator is defined as its squared expected deviation from its true value. For example, the MSE of s_X^2 is defined as $E[(s_X^2 - \sigma_X^2)^2]$. However, since this is equivalent to the estimator's variance plus its squared bias, we usually write it that way. It would be a useful exercise to show this equivalence.

while for $\hat{\sigma}_X^2$, we have

$$\begin{aligned}
\text{MSE}(\hat{\sigma}_X^2) &= \text{Var}(\hat{\sigma}_X^2) + [\hat{\sigma}_X^2]^2 \\
&= \frac{2(n-1)\sigma_X^4}{n^2} + \left[\frac{-1}{n}\sigma_X^2\right]^2 \\
&= \frac{2(n-1)\sigma_X^4}{n^2} + \frac{\sigma_X^4}{n^2} \\
&= \frac{(2n-1)\sigma_X^4}{n^2}.
\end{aligned}$$

To compare the two estimators, we can use the difference in their MSEs:

$$\begin{aligned}
\text{MSE}(s_X^2) - \text{MSE}(\hat{\sigma}_X^2) &= \frac{2\sigma_X^4}{n-1} - \frac{(2n-1)\sigma_X^4}{n^2} \\
&= \frac{2}{n^2}\sigma_X^4 \\
&> 0,
\end{aligned}$$

which implies

$$\text{MSE}(s_X^2) > \text{MSE}(\hat{\sigma}_X^2).$$

So, on the basis of MSE, $\hat{\sigma}_X^2$ would be the preferred estimator. Of course, MSE is not the only basis on which to compare two (or more) estimators.

Before moving on, let's consider an estimator of the covariance of two RVs, called the **sample covariance**. From n observations on each of two RVs, X and Y, the usual estimator of covariance, denoted $s_{X,Y}$, is defined as

$$s_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y}).$$

We can use the sample covariance of X and Y to calculate the **sample coefficient of correlation**, denoted $r_{X,Y}$, as

$$r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y},$$

where s_X, the **sample standard deviation** of X, is just the square root of the sample variance of X. s_Y is defined similarly.

1.11 Asymptotic properties of estimators

The **asymptotic properties** of an estimator are based on random samples of infinite size (i.e. where n tends towards infinity). These differ from the **small sample properties** that we considered above (where n was finite). Although working with random samples of infinite size may seem unrealistic (and it is), asymptotic analysis provides us with an approximation of how estimators perform in *large* random samples.[17]

[17] For this reason, asymptotic properties are sometimes called **large sample properties**.

In order to examine the asymptotic properties of estimators, it is necessary to understand the concept of **convergence in probability**. We begin by introducing some new notation. Let X_n be an RV (such as an estimator), which depends on n. X_n is said to converge in probability to a constant c if

$$\lim_{n\to\infty} \text{Prob}(|X_n - c| > \epsilon) = 0, \text{ for } \epsilon > 0.$$

If X_n converges in probability to c (as above), we will usually write

$$X_n \xrightarrow{p} c.$$

Alternatively, we may say that c is the **probability limit** (or **plim**) of X_n and write

$$\text{plim}(X_n) = c.$$

Using the concept of convergence in probability, we can now consider some useful asymptotic properties of estimators.

An estimator of a parameter is said to be **consistent** if it converges in probability to the true value of the parameter. A *sufficient* condition for an estimator to be consistent is that its MSE tends to zero as n approaches infinity. This condition is known as **convergence in quadratic mean** and always implies convergence in probability. For example, consider the sample mean, $\bar{X}$, of the RV X. The MSE of $\bar{X}$ is

$$\text{MSE}(\bar{X}) = \frac{\sigma_X^2}{n},$$

since $\bar{X}$ is unbiased. So

$$\begin{aligned} \lim_{n\to\infty}[\text{MSE}(\bar{X})] &= \lim_{n\to\infty}\left(\frac{\sigma_X^2}{n}\right) \\ &= 0, \end{aligned}$$

which implies that,

$$\bar{X} \xrightarrow{p} \mu_X.$$

So, we can conclude that $\bar{X}$ is a consistent estimator of the mean of X, μ_X.

Next, consider the estimators of the variance of the RV X, s_X^2 and $\hat{\sigma}_X^2$. We have

$$\begin{aligned} \lim_{n\to\infty}[\text{MSE}(s_X^2)] &= \lim_{n\to\infty}\left(\frac{2\sigma_X^4}{n-1}\right) \\ &= 0, \end{aligned}$$

and

$$\begin{aligned} \lim_{n\to\infty}[\text{MSE}(\hat{\sigma}_X^2)] &= \lim_{n\to\infty}\left[\frac{(2n-1)\sigma_X^4}{n^2}\right] \\ &= 0, \end{aligned}$$

implying that

$$s_X^2 \xrightarrow{p} \sigma_X^2,$$

and

$$\hat{\sigma}_X^2 \xrightarrow{p} \sigma_X^2.$$

So, both s_X^2 and $\hat{\sigma}_X^2$ are consistent estimators of the variance of X, σ_X^2.

Note that this is true even though $\hat{\sigma}_X^2$ is a biased estimator of the variance of X. If the bias of an estimator goes to zero as n approaches infinity, it is said to be **asymptotically unbiased**. Recall that the bias of the estimator $\hat{\sigma}_X^2$ was

$$\text{Bias}(\hat{\sigma}_X^2) = \frac{-1}{n}\sigma_X^2.$$

As n approaches infinity, this becomes zero, and therefore, $\hat{\sigma}_X^2$ is asymptotically unbiased.

Similarly, an estimator is said to be **asymptotically efficient**, if, as n approaches infinity, it is more efficient than any other estimator.

Before concluding, we should make some mention about the *distribution* of estimators. Specifically, we are often interested in the **asymptotic distribution** of estimators (i.e. the distribution of estimators which are based on random samples of infinite size). This requires understanding of the concept of **convergence in distribution**.

The RV X_n, with distribution function $F_n(x)$, is said to converge in distribution to the RV X, with distribution function $F(x)$, if

$$\lim_{n\to\infty} |F_n(x) = F(x)| = 0$$

at all continuity points of $F(x)$. This is usually written as

$$X_n \xrightarrow{d} X.$$

In considering the asymptotic distribution of an estimator, a **central limit theorem** (or **CLT**) often comes in handy. For example, the asymptotic distribution of the sample mean can be found by using the Lindeberg-Levy CLT. This CLT states that if $x_1, \ldots, x_n$ is a random sample of n observations on the RV X, where X has mean μ_X and variance σ_X^2, then

$$n^{-1/2}\sum_{i=1}^{n}(x_i - \mu_X) \xrightarrow{d} N(0, \sigma_X^2).$$

Alternatively, this can be written as

$$\sqrt{n}(\bar{X} - \mu_X) \xrightarrow{d} N(0, \sigma_X^2),$$

since

$$\begin{aligned} n^{-1/2}\sum_{i=1}^{n}(x_i - \mu_X) &= n^{-1/2}\left(\sum_{i=1}^{n} x_i - n\mu_X\right) \\ &= n^{-1/2}(n\bar{X} - n\mu_X) \\ &= \sqrt{n}(\bar{X} - \mu_X) \end{aligned}$$

where $\bar{X}$ is as defined earlier. In this case, $\bar{X}$ is said to be **asymptotically normal**. What is remarkable about this theorem is that it holds true even if the RV X does not follow a normal distribution (e.g. it may be uniformly distributed).

2 Regression Models

2.1 Introduction

Natural and social scientists use abstract models to describe the behaviour of the systems that they study. Typically, these models are formulated in mathematical terms, using one or more equations. For example, in an introductory macroeconomics course, the aggregate expenditures model is used to describe the behaviour of an economy. In this model, the equation

$$C = cD,$$

is used to describe how the level of consumption, C, depends on the level of disposable income, D. Here, C is a linear function of D, whose slope is given by the parameter c, called the marginal propensity to consume (which is some arbitrary constant). Such a relationship is said to be *deterministic*: given the level of D, we can determine the exact level of C.

Deterministic relationships, however, rarely (if ever) hold in observed data. This is for two reasons: First, there will almost always be some error in measuring the variables involved; and second, there will very often be some other, *non-systematic* factors that affect the relationship but which cannot be easily identified. Of course, there may be *some* relationship between the variables of interest, but it will not be *exact* (i.e. completely deterministic). Such a relationship is said to be *statistical*.

To illustrates the difference, consider a model which posits that the variable Y is determined by some function of the variable X, i.e.

$$Y = m(X). \tag{1}$$

Now, suppose that we have n observations on each of the variables Y and X, which we denote $y_1, \ldots, y_n$ and $x_1, \ldots, x_n$, respectively. For any given pair of observations, (y_i, x_i), we would like to allow the possibility that the relationship in (1) does not hold exactly. To do this, we add in what is called a **disturbance term**, denoted u_i, which gives us the **regression model**

$$y_i = m(x_i) + u_i, \qquad i = 1, \ldots, n. \tag{2}$$

By adding this disturbance term, we are able to capture any measurement errors or other non-systematic factors not identified in (1). On average, we would hope that each disturbance term is zero (otherwise, the model would contain some systematic error). However, for each pair of observations, (y_i, x_i), $i = 1, \ldots, n$ the actual size of u_i will differ.

Since the disturbance term is, by nature, unpredictable, it is treated as an RV. Like any other RV, it has its own distribution. The most important characteristic of this distribution is that its expected value is zero, i.e. for each u_i,

$$E(u_i) = 0.$$

Typically, we assume that each u_i, $i = 1, \ldots, n$, is IID. This means, among other things, that each has the same variance, i.e. for each u_i,

$$\text{Var}(u_i) = \sigma_u^2.$$

In this case, the disturbance terms are said to be **homoskedastic** (which literally means "same variance"). Given the IID assumption, it is common to write

$$u_i \sim \text{IID}(0, \sigma_u^2).$$

Note that, since y_i is a function of the RV u_i (see Equation (2)), it also must be an RV itself. Of course, y_i is also dependent on x_i. For this reason, y_i is usually referred to as the **dependent variable**, and x_i is usually referred to as the **independent variable** (or **explanatory variable**).

We haven't yet indicated whether x_i is an RV. This is because sometimes it is, and sometimes it isn't. If we our observations are generated by a *controlled* experiment, then $x_1, \ldots, x_n$ are held constant (and are obviously not RVs). As an example, consider an experiment where we give patients different doses of a blood pressure medication, and measure the reduction in their blood pressures over some period of time. Here, $x_1, \ldots, x_n$ are the doses of the medication given to patients $1, \ldots, n$, and $y_1, \ldots, y_n$ are the reduction in the blood pressures of patients $1, \ldots, n$. If we repeat this experiment several times over (with, say, a new group of patients each time), holding $x_1, \ldots, x_n$ constant (i.e. keeping the doses unchanged), we would still expect that $y_1, \ldots, y_n$ would be different each time (due to the disturbances $u_1, \ldots, u_n$). In this case, $y_1, \ldots, y_n$ are RVs (since $u_1, \ldots, u_n$ are RVs), but $x_1, \ldots, x_n$ are not.

On the other hand, if our observations are generated by a *field* experiment, then $x_1, \ldots, x_n$ are RVs. For example, consider an experiment where we observe different workers' education and income levels (perhaps we want to see if higher levels of education result in higher levels of income). Here, $x_1, \ldots, x_n$ are the education levels of workers $1, \ldots, n$, and $y_1, \ldots, y_n$ are the income levels of workers $1, \ldots, n$. If we repeat this experiment several times over (with a new group of workers each time), $x_1, \ldots, x_n$ will be different each time, and therefore can be considered RVs (since we don't know their values until we complete the experiment).

Up to this point, we have been dealing with what is called a **bivariate regression model**, since it involves just two variables: y_i and x_i. Of course, most regression models used in the natural and social sciences involve more than just two variables. A model of this type, referred to as a **multivariate regression model**, has one dependent variable and more than one independent

variables. A multivariate regression model with k different independent variables takes the form

$$y_i = m(x_{1,i}, \ldots x_{k,i}) + u_i, \qquad i = 1, \ldots .n,$$

where $x_{1,1}, \ldots, x_{1,n}$ are observations on the variable X_1, $x_{2,1}, \ldots, x_{2,n}$ are observations on the variable X_2, and so on.

Aler natively, it is possible to have a **univariate regression model**, where only one variable is involved. Such models are usually only relevant when the variable involved is in the form of a time-series. An example of such a model is what is known as an **autoregressive process**, where the dependent variable is regressed on lagged values of itself. Letting $y_1, \ldots, y_n$ be observations on the variable Y over time, a ***p*th-order autoregressive process** (or **AR(*p*) process**) takes the form

$$y_t = m(y_{t-1}, \ldots y_{t-p}) + u_t, \qquad t = 1, \ldots .n.$$

For ease of exposition, the rest of this lecture will focus on bivariate regression models, but all of the results are perfectly applicable to multivariate and univariate regression models.

2.2 Estimation of regression models

Usually, the exact form of the function $m(\cdot)$ is unknown. Therefore, the task of the statistician is to come up with some estimate of this function, which we will denote $\hat{m}(\cdot)$. Of course, since any estimate of $m(\cdot)$ will be based on the the RVs $y_1, \ldots, y_n$ (and $x_1, \ldots, x_n$, which also may be RVs), $\hat{m}(\cdot)$, like any other estimator, will also be an RV. Exactly how $m(\cdot)$ is estimated, and how $\hat{m}(\cdot)$ is distributed (which will depend on the method of estimation), will occupy us for the remainder of the course.

Assuming we have an estimate of $m(\cdot)$ in hand, it can be used in the following way. Given a specific observation x_i, the estimate $\hat{m}(x_i)$ will give us an estimate of y_i, which we will denote $\hat{y}_i$, i.e.

$$\hat{y}_i = \hat{m}(x_i).$$

Clearly, the estimate $\hat{y}_i$ will not always be equal to the observed value y_i. So, we add in what is called an **error term** (or **residual**), denoted $\hat{u}_i$, and write

$$y_i = \hat{m}(x_i) + \hat{u}_i.$$

Rearranging, we have

$$\begin{aligned} \hat{u}_i &= y_i - \hat{m}(x_i) \\ &= y_i - \hat{y}_i. \end{aligned}$$

That is, the error term is equal to the difference between the observed and estimated values of y_i. Clearly, any estimator of $m(\cdot)$ should have the property that it minimizes such errors. This will be a guiding principle in developing the estimators we are interested in.

2.3 Specification of regression models

While the exact form of $m(\cdot)$ is usually unknown, it is common for the statistician (or even the theorist) to assume some function form. If $m(\cdot)$ is given a specific form, (2) is referred to as a **parametric regression model** (since the function must involve some parameters). On the other hand, if $m(\cdot)$ is left unspecified, the regression model in (2) is referred to as a **nonparametric regression model**. For now, we will focus on parametric regression models, but we will consider nonparametric regression models later in the course.

The most common form assumed for regression models involves specifying $m(\cdot)$ as an **affine function**[18], i.e.

$$m(x_i) = \alpha + \beta x_i,$$

which gives use the **linear regression model**

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

Here α and β are parameters to be estimated. Denoting these estimates $\hat{\alpha}$ and $\hat{\beta}$, we have

$$\hat{m}(x_i) = \hat{\alpha} + \hat{\beta} x_i.$$

Arriving at these estimates will be the focus of the next topic.

Of course, not all parametric regression models are linear. A **nonlinear regression model** is one in which $m(\cdot)$ is nonlinear. For example, if

$$m(x_i) = x_i^{\gamma}$$

we have the nonlinear regression model

$$y_i = x_i^{\gamma} + u_i.$$

As above, γ is a parameter to be estimated. Denoting this estimate $\hat{\gamma}$, we have

$$\hat{m}(x_i) = x_i^{\hat{\gamma}}.$$

Estimating nonlinear regression models is beyond the scope of this course.

3 Ordinary least squares

3.1 Estimating bivariate linear regression models

The most common method for estimating linear regression models (both bivariate and multivariate) is the method of **ordinary least squares** (or **OLS**). This method is based on minimizing the **sum of squared residuals** (or **SSR**) from the estimated regression model.

[18] An affine function, $g(x)$, takes the form $g(x) = a + bx$, where a and b are any real numbers.

If we have the bivariate linear regression model

$$y_i = \alpha + \beta x_i + u_i, \qquad i = 1, \ldots . n \tag{3}$$

then the ith residual from the estimate of this model is

$$\hat{u}_i = y_i - \hat{\alpha} - \hat{\beta} x_i.$$

Summing the square of each of these terms (from $i = 1 \ldots, n$), we have

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2.$$

From the left hand side of this equation, it is evident that the sum of squared residuals is a function of both $\hat{\alpha}$ and $\hat{\beta}$. Therefore, we normally write

$$\text{SSR}(\hat{\alpha}, \hat{\beta}) = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2.$$

Since the method of OLS is based on minimizing this function, we are faced with a very simple optimization problem: Minimize the function $\text{SSR}(\hat{\alpha}, \hat{\beta})$ by choosing $\hat{\alpha}$ and $\hat{\beta}$, i.e.

$$\min_{\hat{\alpha}, \hat{\beta}} \text{SSR}(\hat{\alpha}, \hat{\beta}) = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

The first-order, necessary conditions for this optimization problem are

$$\frac{\partial \text{SSR}(\hat{\alpha}, \hat{\beta})}{\partial \hat{\alpha}} = -2 \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0, \tag{4}$$

and

$$\frac{\partial \text{SSR}(\hat{\alpha}, \hat{\beta})}{\partial \hat{\beta}} = -2 \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)(x_i) = 0. \tag{5}$$

Equation (4) implies

$$\sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0.$$

Carrying through the summation operator, we have

$$\sum_{i=1}^{n} y_i - n\hat{\alpha} - \sum_{i=1}^{n} \hat{\beta} x_i = 0.$$

Finally, solving for $\hat{\alpha}$ gives

$$\hat{\alpha} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \hat{\beta}\sum_{i=1}^{n} x_i\right),$$

which can be rewritten as

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \tag{6}$$

where, $\bar{y}$ is the sample mean of the observations $y_i, \ldots, y_n$, i.e.

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

and $\bar{x}$ is the sample mean of the observations $x_i, \ldots, x_n$, i.e.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Moving on to Equation (5), we have

$$\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)(x_i) = 0.$$

Multiplying through by x_i, we have

$$\sum_{i=1}^{n}(y_i x_i - \hat{\alpha}x_i - \hat{\beta}x_i^2) = 0.$$

Next, carrying through the summation operator, we have

$$\sum_{i=1}^{n} y_i x_i - \hat{\alpha}\sum_{i=1}^{n} x_i - \hat{\beta}\sum_{i=1}^{n} x_i^2 = 0.$$

Now, substituting for $\hat{\alpha}$ from Equation (6) gives

$$\sum_{i=1}^{n} y_i x_i - (\bar{y} - \hat{\beta}\bar{x})\sum_{i=1}^{n} x_i - \hat{\beta}\sum_{i=1}^{n} x_i^2 = 0.$$

Expanding the term in brackets, we have

$$\sum_{i=1}^{n} y_i x_i - \bar{y}\sum_{i=1}^{n} x_i + \hat{\beta}\bar{x}\sum_{i=1}^{n} x_i - \hat{\beta}\sum_{i=1}^{n} x_i^2 = 0.$$

Finally, solving for $\hat{\beta}$, we have

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i - \bar{y}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i}. \tag{7}$$

It is often desirable to write $\hat{\beta}$ as

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{8}$$

To show that Equations (7) and (8) are equivalent, we show that the numerator and denominator of each are equivalent. First, expand the numerator in Equation (8) and rearrange:

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) &= \sum_{i=1}^{n}(y_i x_i - y_i\bar{x} - \bar{y}x_i + \bar{y}\bar{x}) \\
&= \sum_{i=1}^{n} y_i x_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + n\bar{y}\bar{x} \\
&= \sum_{i=1}^{n} y_i x_i - \bar{x}n\bar{y} - \bar{y}\sum_{i=1}^{n} x_i + ny\bar{x}) \\
&= \sum_{i=1}^{n} y_i x_i - \bar{y}\sum_{i=1}^{n} x_i,
\end{aligned}$$

which is exactly the numerator in Equation (7). Second, expand the denominator in Equation (8) and rearrange:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i,
\end{aligned}$$

which is exactly the denominator in Equation (7).

At other times, it will be convenient to to rewrite $\hat{\beta}$ as

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})x_i}. \tag{9}$$

To show that this is equivalent to Equation (8) (and therefore Equation (7)), we again show that the numerator and denominator of each are equivalent. To

do so, it is helpful to realize that

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x}) &= \sum_{i=1}^{n} x_i - n\bar{x} \\ &= nx - n\bar{x} \\ &= 0,\end{aligned}$$

and similarly,

$$\sum_{i=1}^{n}(y_i - \bar{y}) = 0.$$

Now, the numerator in Equation (8) can be expanded to give

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^{n}[(x_i - \bar{x})y_i - (x_i - \bar{x})\bar{y}] \\ &= \sum_{i=1}^{n}(x_i - \bar{x})y_i - \bar{y}\sum_{i=1}^{n}(x_i - \bar{x}) \\ &= \sum_{i=1}^{n}(x_i - \bar{x})y_i - 0 \\ &= \sum_{i=1}^{n}(x_i - \bar{x})y_i,\end{aligned}$$

which is the numerator in Equation (9). Similarly, the denominator in Equation (8) can be expanded to give

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}[(x_i - \bar{x})x_i - (x_i - \bar{x})\bar{x}] \\ &= \sum_{i=1}^{n}(x_i - \bar{x})x_i - \bar{x}\sum_{i=1}^{n}(x_i - \bar{x}) \\ &= \sum_{i=1}^{n}(x_i - \bar{x})x_i - 0 \\ &= \sum_{i=1}^{n}(x_i - \bar{x})x_i,\end{aligned}$$

which is the denominator in Equation (9).

So, we have three different (but equivalent) expressions for $\hat{\beta}$: Equations (7), (8), and (9). We most often use Equation (8), but, as we will see in the next section, Equation (9) will often come in handy.

3.2 Properties of OLS estimators of bivariate linear regression models

The properties of OLS estimators depend on the properties of the error terms in the linear regression model we are estimating. For now, we will assume that, for each u_i,

$$u_i \sim \text{IID}(0, \sigma_u^2).$$

However, as we progress, we will relax this assumption and see how the properties of OLS estimators change.

In addition, we will also assume that x_i, $i = 1, \ldots, n$ are not RVs. This assumption does not have nearly as large an impact on the properties of OLS estimators as the IID assumption, but will make the following analysis a little easier.

The first question we want to ask is whether or not OLS estimators are unbiased. Let's start with $\hat{\beta}$. To do so, we substitute for y_i in Equation (9) from Equation (3):

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(\alpha + \beta x_i + u_i)}{\sum_{i=1}^n (x_i - \bar{x}) x_i}.$$

Multiplying through, we have

$$\begin{aligned}
\hat{\beta} &= \frac{\alpha \sum_{i=1}^n (x_i - \bar{x}) + \beta \sum_{i=1}^n (x_i - \bar{x}) x_i + \sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x}) x_i} \\
&= \frac{0 + \beta \sum_{i=1}^n (x_i - \bar{x}) x_i + \sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x}) x_i} \\
&= \beta + \frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x}) x_i}. \qquad (10)
\end{aligned}$$

Next, we find the expected value of $\hat{\beta}$:

$$\begin{aligned}
E(\hat{\beta}) &= E\left[\beta + \frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x}) x_i}\right] \\
&= \beta + \frac{\sum_{i=1}^n (x_i - \bar{x}) E(u_i)}{\sum_{i=1}^n (x_i - \bar{x}) x_i} \\
&= \beta,
\end{aligned}$$

since $E(u_i) = 0$. So, $\hat{\beta}$ is unbiased.

Let's now move on to $\hat{\alpha}$. From Equation (6), we have

$$\hat{\alpha} = \frac{1}{n} \sum_{i=1}^n y_i - \hat{\beta} \bar{x}.$$

Substituting for y_i from Equation (3), we have

$$\begin{aligned}
\hat{\alpha} &= \frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i + u_i) - \hat{\beta}\bar{x} \\
&= \frac{1}{n}\left[n\alpha + \beta\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} u_i)\right] - \hat{\beta}\bar{x} \\
&= \alpha + \beta\bar{x} + \frac{1}{n}\sum_{i=1}^{n} -\hat{\beta}\bar{x} \\
&= \alpha + \bar{x}(\beta - \hat{\beta}) + \frac{1}{n}\sum_{i=1}^{n} u_i. \qquad (11)
\end{aligned}$$

Taking the expected value of $\hat{\alpha}$, we have

$$\begin{aligned}
E(\hat{\alpha}) &= E\left[\alpha + \bar{x}(\beta - \hat{\beta}) + \frac{1}{n}\sum_{i=1}^{n} u_i\right] \\
&= \alpha + \bar{x}[\beta - E(\hat{\beta})] + \frac{1}{n}\sum_{i=1}^{n} E(u_i) \\
&= \alpha + \bar{x}(\beta - \beta) \\
&= \alpha
\end{aligned}$$

So, $\hat{\alpha}$ is also unbiased.

Another important property of OLS estimators is their variance. Since $\hat{\beta}$ is unbiased, we have

$$\mathrm{Var}(\hat{\beta}) = E[(\hat{\beta} - \beta)^2].$$

From Equation (10), this means

$$\mathrm{Var}(\hat{\beta}) = E\left(\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})x_i}\right]^2\right)$$

To make handling the term in brackets a little easier, let's introduce some new notation. Let

$$k_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

so that

$$\mathrm{Var}(\hat{\beta}) = E\left[\left(\sum_{i=1}^{n} k_i u_i\right)^2\right].$$

Expanding this, we have

$$\mathrm{Var}(\hat{\beta}) = E\left[\sum_{i=1}^{n} k_i^2 u_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_i k_j u_i u_j\right]$$

$$= \sum_{i=1}^{n} k_i^2 E(u_i^2) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_i k_j E(u_i u_j).$$

Two things should be apparent here: First, notice that

$$E(u_i^2) = \sigma_u^2$$

since

$$\begin{aligned} \sigma_u^2 &= \text{Var}(u_i) \\ &= E\left([u_i - E(u_i)]^2\right) \\ &= E(u_i^2) \end{aligned}$$

Second, for all $i \neq j$, we have

$$\begin{aligned} E(u_i u_j) &= E(u_i)E(u_j) \\ &= 0, \end{aligned}$$

since we have assumed that each u_i is IID (and therefore statistically independent from one another). So,

$$\begin{aligned} \text{Var}(\hat{\beta}) &= \sigma_u^2 \sum_{i=1}^{n} k_i^2 \\ &= \sigma_u^2 \sum_{i=1}^{n} \left[\frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^2 \\ &= \frac{\sigma_u^2}{[\sum(x_i - \bar{x})^2]^2} \sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{\sigma_u^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \end{aligned}$$

We now move on to $\hat{\alpha}$. Since $\hat{\alpha}$ is unbiased, we have

$$\text{Var}(\hat{\alpha}) = E[(\hat{\alpha} - \alpha)^2].$$

From Equation (11), this means

$$\begin{aligned} \text{Var}(\hat{\alpha}) &= E\left(\left[\bar{x}(\beta - \hat{\beta}) + \frac{1}{n}\sum_{i=1}^{n} u_i\right]^2\right) \\ &= E\left[\bar{x}^2(\beta - \hat{\beta})^2 + 2\bar{x}(\beta - \hat{\beta}) + \frac{1}{n^2}\left(\sum_{i=1}^{n} u_i\right)^2\right] \\ &= \bar{x}^2 E[(\beta - \hat{\beta})^2] + 2\bar{x}[\beta - E(\hat{\beta})] + \frac{1}{n^2}E\left(\sum_{i=1}^{n} u_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} u_i u_j\right) \end{aligned}$$

$$
\begin{aligned}
&= \bar{x}^2 \text{Var}(\hat{\beta}) + 2\bar{x}[\beta - \beta] + \frac{1}{n^2}\sum_{i=1}^{n} E(u_i^2) + \frac{2}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E(u_i u_j) \\
&= \frac{\sigma_u^2 \bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \frac{1}{n^2} n\sigma_u^2 \\
&= \sigma_u^2 \left(\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \frac{1}{n} \right).
\end{aligned}
$$

Given the variance of OLS estimators, we might want to ask how efficient they are. The **Gauss-Markov theorem** states that, in the class of linear, unbiased estimators, they are in fact the most efficient. This property is often stated as **BLUE** (Best Linear Unbiased Estimator). By "best", we mean most efficient. Here, we will prove this only for $\hat{\beta}$, but the proof for $\hat{\alpha}$ is very similar.

Using the k_i notation we introduced above, Equation (9) implies

$$\hat{\beta} = \sum_{i=1}^{n} k_i y_i.$$

This should make it clear that $\hat{\beta}$ is a linear estimator – the above function is linear in y_i. Now, let's consider some linear, unbiased estimator of β, call it $\tilde{\beta}$:

$$\tilde{\beta} = \sum_{i=1}^{n} g_i y_i,$$

where

$$g_i = k_i + w_i,$$

and w_i is some non-zero term which depends on $x_1, \ldots, x_n$. This implies that

$$
\begin{aligned}
\tilde{\beta} &= \sum_{i=1}^{n} (k_i + w_i) y_i \\
&= \sum_{i=1}^{n} k_i y_i + \sum_{i=1}^{n} w_i y_i \\
&= \hat{\beta} + \sum_{i=1}^{n} w_i y_i. \qquad (12)
\end{aligned}
$$

Taking expectations, we have

$$
\begin{aligned}
E(\tilde{\beta}) &= E\left(\hat{\beta} + \sum_{i=1}^{n} w_i y_i \right) \\
&= \beta + E\left(\sum_{i=1}^{n} w_i y_i \right).
\end{aligned}
$$

Substituting for y_i from Equation (3), we have

$$\begin{aligned} E(\tilde{\beta}) &= \beta + E\left[\sum_{i=1}^{n} w_i(\alpha + \beta x_i + u_i)\right] \\ &= \beta + \alpha\sum_{i=1}^{n} w_i + \beta\sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} E(u_i) \\ &= \beta + \alpha\sum_{i=1}^{n} w_i + \beta\sum_{i=1}^{n} w_i x_i, \end{aligned}$$

since $E(u_i) = 0$. In order for $\tilde{\beta}$ to be unbiased, we must have

$$\sum_{i=1}^{n} w_i = 0,$$

and

$$\sum_{i=1}^{n} w_i x_i = 0,$$

since, in general, α and β are non-zero. This allows us to rewrite the second term on the left-hand side of Equation (12) as

$$\begin{aligned} \sum_{i=1}^{n} w_i y_i &= \sum_{i=1}^{n} w_i(\alpha + \beta x_i + u_i) \\ &= \alpha\sum_{i=1}^{n} w_i + \beta\sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} w_i u_i \\ &= \sum_{i=1}^{n} w_i u_i. \end{aligned}$$

Also, note that Equation (10) implies

$$\hat{\beta} = \beta + \sum_{i=1}^{n} k_i u_i.$$

Substituting these last two results into Equation (12), we have

$$\begin{aligned} \tilde{\beta} &= \beta + \sum_{i=1}^{n} k_i u_i + \sum_{i=1}^{n} w_i u_i \\ &= \beta + \sum_{i=1}^{n} (k_i + w_i) u_i. \end{aligned}$$

Finally, the variance of $\tilde{\beta}$ is

$$\mathrm{Var}(\tilde{\beta}) = E\left(\left[\tilde{\beta} - E(\tilde{\beta})\right]^2\right)$$

$$
\begin{aligned}
&= E\left(\left[\sum_{i=1}^{n}(k_i + w_i)u_i\right]^2\right) \\
&= E\left[\sum_{i=1}^{n}(k_i + w_i)^2 u_i^2\right] \\
&= \sum_{i=1}^{n}(k_i + w_i)^2 E(u_i^2) \\
&= \sigma_u^2 \sum_{i=1}^{n}(k_i + w_i)^2 \\
&= \sigma_u^2 \sum_{i=1}^{n}(k_i^2 + 2k_i w_i + w_i^2) \\
&= \sigma_u^2 \sum_{i=1}^{n} k_i^2 + 2\sigma_u^2 \sum_{i=1}^{n} k_i w_i + \sigma_u^2 \sum_{i=1}^{n} w_i^2 \\
&= \mathrm{Var}(\hat{\beta}) + 2\sigma_u^2 \sum_{i=1}^{n} k_i w_i + \sigma_u^2 \sum_{i=1}^{n} w_i^2.
\end{aligned}
$$

The second term on the right-hand-side of this equation is

$$
\begin{aligned}
\sum_{i=1}^{n} k_i w_i &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})w_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n} x_i w_i - \bar{x}\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= 0,
\end{aligned}
$$

since $\sum_{i=1}^{n} x_i w_i = 0$ and $\sum_{i=1}^{n} w_i = 0$. Therefore, we have

$$
\mathrm{Var}(\tilde{\beta}) = \mathrm{Var}(\hat{\beta}) + \sigma_u^2 \sum_{i=1}^{n} w_i^2,
$$

which implies

$$
\mathrm{Var}(\tilde{\beta}) > \mathrm{Var}(\hat{\beta}),
$$

since w_i is non-zero. This proves the Gauss-Markov theorem for $\hat{\beta}$. We won't go through it here, but the proof for $\hat{\alpha}$ is quite similar.

3.3 Estimating multivariate linear regression models

Now consider the multivariate linear regression model[19]

$$
y_i = \beta_1 + \beta_2 x_{2,i} + \ldots + \beta_k x_{k,i} + u_i, \qquad i = 1, \ldots n
$$

[19] Note that the bivariate linear regression model considered above is just a special case of the multivariate linear regression model considered here. For this reason, what follows is applicable to both bivariate and multivariate linear regression models.

To make the analysis of such regression models much easier, we typically express this in matrix form:[20]

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} \tag{13}$$

where $\mathbf{y}$ and $\mathbf{u}$ are n-vectors, $\mathbf{X}$ is an $n \times k$ matrix, and β is a k-vector, defined as follows:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{2,1} & \dots & x_{k,1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{2,n} & \dots & x_{k,n} \end{pmatrix},$$

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix},$$

and

$$\mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}.$$

Letting $\hat{\mathbf{u}}$ denote the residuals from the estimate of this model, we have

$$\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\mathbf{b}$$

where $\mathbf{b}$ is an estimator of β. The sum squared residuals is thus

$$\begin{aligned} \text{SSR}(\mathbf{b}) &= \hat{\mathbf{u}}'\hat{\mathbf{u}} \\ &= (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) \\ &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}. \end{aligned}$$

Here, we make use of the fact that $\mathbf{y}'\mathbf{X}\mathbf{b}$ is a scalar (this can be confirmed by checking its dimensions), and is therefore equal to its transpose, i.e.

$$\begin{aligned} \mathbf{y}'\mathbf{X}\mathbf{b} &= (\mathbf{y}'\mathbf{X}\mathbf{b})' \\ &= \mathbf{b}'\mathbf{X}'\mathbf{y}. \end{aligned}$$

As we saw in the bivariate case, the method of OLS is based on minimizing this function. Our optimization problem is thus

$$\min_{\mathbf{b}} \text{SSR}(\mathbf{b}) = \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}.$$

[20] In these notes, we make the assumption that the reader is reasonably well-versed in matrix algebra.

The first-order necessary condition for this optimization problem is

$$\frac{\partial \text{SSR}(\mathbf{b})}{\partial \mathbf{b}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} = 0,$$

which implies

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b}.$$

Finally, solving for $\mathbf{b}$, we have

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{14}$$

3.4 Properties of OLS estimators of multivariate linear regression models[21]

In order to analyze the properties of $\mathbf{b}$, we again need to make some assumptions about the error terms. As before, we assume that each u_i has mean zero. This means that the n-vector $\mathbf{u}$ has the following expected value:

$$\begin{aligned} E(\mathbf{u}) &= E\begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \\ &= \mathbf{0}, \end{aligned}$$

where $\mathbf{0}$ is a k-vector of zeros.

The variance of $\mathbf{u}$ is

$$\begin{aligned} \text{Var}(\mathbf{u}) &= E\left([\mathbf{u} - E(\mathbf{u})][\mathbf{u} - E(\mathbf{u})]'\right) \\ &= E(\mathbf{u}\mathbf{u}') \\ &= E\left[\begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}\begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix}\right] \\ &= E\begin{pmatrix} u_1^2 & \cdots & u_1 u_n \\ \vdots & & \vdots \\ u_n u_1 & \cdots & u_n^2 \end{pmatrix}, \end{aligned}$$

which is called the **error covariance matrix**. For now, we will assume that each u_i is IID with variance σ_u^2, which means that

$$E(u_i^2) = \sigma_u^2,$$

[21] As noted above, what follows is also applicable to bivariate linear regression models (since the bivariate linear regression model is just a special case of the multivariate linear regression model).

and, for all $i \neq j$,

$$E(u_i u_j) = 0.$$

Therefore,

$$\begin{aligned} \text{Var}(\mathbf{u}) &= \begin{pmatrix} \sigma_u^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_u^2 \end{pmatrix} \\ &= \sigma_u^2 \mathbf{I}_n, \end{aligned}$$

where $\mathbf{I}_n$ is an $n \times n$ identity matrix.

Given these assumptions, we typically write

$$\mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma_u^2 \mathbf{I}_n).$$

Also, to make things easier, we will continue to assume that $\mathbf{X}$ is not a RV. Later in the course, we will see what the consequences are if we relax this assumption.

Let's start the analysis by asking whether $\mathbf{b}$ is unbiased or not. Substituting Equation (13) into Equation (14), we have

$$\begin{aligned} \mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\mathbf{b} + \mathbf{u}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}. \end{aligned} \tag{15}$$

Now, taking expectations we have

$$\begin{aligned} E(\mathbf{b}) &= E\left[\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\right] \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u}) \\ &= \beta, \end{aligned}$$

since $E(\mathbf{u}) = \mathbf{0}$. So $\mathbf{b}$ is unbiased.

Let's now move on to the variance of $\mathbf{b}$. Given that $\mathbf{b}$ is unbiased, we have

$$\begin{aligned} \text{Var}(\mathbf{b}) &= E\left([\mathbf{b} - E(\mathbf{b})][\mathbf{b} - E(\mathbf{b})]'\right) \\ &= E\left[(\mathbf{b} - \beta)(\mathbf{b} - \beta)'\right]. \end{aligned}$$

Substituting from Equation (15), we have

$$\begin{aligned} \text{Var}(\mathbf{b}) &= E\left([(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}][(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}]'\right) \\ &= E\left[(\mathbf{X}'\mathbf{X})\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u}\mathbf{u}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \tag{16}$$

Having derived the variance of $\mathbf{b}$, we now want to show that the Gauss-Markov theorem holds here. Consider some other linear, unbiased estimator of β, call it $\mathbf{B}$:

$$\mathbf{B} = \mathbf{G}\mathbf{y},$$

where

$$\mathbf{G} = \mathbf{W} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}',$$

and $\mathbf{W}$ is some non-zero matrix which depends on $\mathbf{X}$. The expected value of $\mathbf{B}$ is thus

$$\begin{aligned} E(\mathbf{B}) &= E(\mathbf{G}\mathbf{y}) \\ &= \mathbf{G}E(\mathbf{y}). \end{aligned}$$

Note that the expected value of $\mathbf{y}$ is

$$\begin{aligned} E(\mathbf{y}) &= E(\mathbf{X}\beta + \mathbf{u}) \\ &= \mathbf{X}\beta + E(\mathbf{u}) \\ &= \mathbf{X}\beta, \end{aligned}$$

so

$$E(\mathbf{B}) = \mathbf{G}\mathbf{X}\beta.$$

This means that, in order for $\mathbf{B}$ to be unbiased, we must have

$$\mathbf{G}\mathbf{X} = \mathbf{I}_k,$$

where $\mathbf{I}_k$ is an $k \times k$ identity matrix (check the dimensions of $\mathbf{G}\mathbf{X}$ to confirm this). Given this condition, notice that

$$\begin{aligned} \mathbf{B} &= \mathbf{G}\mathbf{y} \\ &= \mathbf{G}(\mathbf{X}\beta + \mathbf{u}) \\ &= \mathbf{G}\mathbf{X}\beta + \mathbf{G}\mathbf{u} \\ &= \beta + \mathbf{G}\mathbf{u}, \end{aligned}$$

and that

$$\begin{aligned} \mathbf{W}\mathbf{X} &= [\mathbf{G} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{X} \\ &= \mathbf{G}\mathbf{X} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} \\ &= \mathbf{I}_k - \mathbf{I}_k \\ &= \mathbf{0}. \end{aligned}$$

The variance of $\mathbf{B}$ is therefore

$$\begin{aligned} \text{Var}(\mathbf{B}) &= E\left([\mathbf{B} - E(\mathbf{B})][\mathbf{b} - E(\mathbf{B})]'\right) \\ &= E[(\beta + \mathbf{G}\mathbf{u} - \beta)(\beta + \mathbf{G}\mathbf{u} - \beta)'] \end{aligned}$$

$$
\begin{aligned}
&= E[(\mathbf{Gu})(\mathbf{Gu})'] \\
&= E(\mathbf{Guu'G'}) \\
&= \mathbf{G}E(\mathbf{uu'})\mathbf{G'} \\
&= \sigma_u^2\mathbf{GG'} \\
&= \sigma_u^2[\mathbf{W} + (\mathbf{X'X})^{-1}\mathbf{X'}][\mathbf{W} + (\mathbf{X'X})^{-1}\mathbf{X'}]' \\
&= \sigma_u^2[\mathbf{WW'} + \mathbf{WX}(\mathbf{X'X})^{-1} + (\mathbf{X'X})^{-1}\mathbf{X'W'} + \\
&\quad (\mathbf{X'X})^{-1}\mathbf{X'X}(\mathbf{X'X})^{-1}] \\
&= \sigma_u^2\mathbf{WW'} + \sigma_u^2(\mathbf{W'W})^{-1}.
\end{aligned}
$$

Finally, since $\mathbf{W}$ is a non-zero matrix,

$$\text{Var}(\mathbf{B}) > \text{Var}(\mathbf{b}),$$

which proves the Gauss-Markov theorem.

Before moving on, let's briefly consider some of the asymptotic properties of OLS estimators.[22] Arguably the most important of these is consistency. Recall that a sufficient condition for estimator to be consistent is that its MSE approaches zero as n approaches infinity. Since $\mathbf{b}$ is unbiased, its MSE is equal to its variance, i.e.

$$\text{MSE}(\mathbf{b}) = \sigma_u^2(\mathbf{X'X})^{-1}.$$

Now, consider the matrix $\mathbf{X'X}$. Since each element of this matrix is a sum of n numbers, as n approaches infinity, we would expect that each such element would also approach infinity. However, it would be safe to assume that $\frac{1}{n}$ times this times each of these elements would approach some constant. We can write this assumption as

$$\lim_{n\to\infty}\left(\frac{1}{n}\mathbf{X'X}\right) = \mathbf{Q}, \tag{17}$$

where $\mathbf{Q}$ is a finite, non-singular matrix.

We can use this assumption to show that $\mathbf{b}$ is consistent by rewriting its MSE as

$$\text{MSE}(\mathbf{b}) = \frac{1}{n}\sigma_u^2(\frac{1}{n}\mathbf{X'X})^{-1}.$$

Notice that this is equivalent to our previous expression for the MSE of $\mathbf{b}$ since the $\frac{1}{n}$ terms cancel each other out. Taking the limit as n approaches infinity, we have

$$
\begin{aligned}
\lim_{n\to\infty}[\text{MSE}(\mathbf{b})] &= \lim_{n\to\infty}\left[\frac{1}{n}\sigma_u^2\left(\frac{1}{n}\mathbf{X'X}\right)^{-1}\right] \\
&= \lim_{n\to\infty}\left(\frac{1}{n}\sigma_u^2\right)\left[\lim_{n\to\infty}\left(\frac{1}{n}\mathbf{X'X}\right)\right]^{-1} \\
&= (0)\mathbf{Q}^{-1} \\
&= \mathbf{0},
\end{aligned}
$$

[22] We didn't do this when we focused on the bivariate linear regression model, but, once again, all of what follows is applicable to that special case.

which implies

$$\mathbf{b} \xrightarrow{p} \beta.$$

So, $\mathbf{b}$ is consistent.

Finally, let's consider the asymptotic distribution of $\mathbf{b}$.[23] To do so, begin by rewriting Equation (15) as

$$\mathbf{b} = \beta + \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'\mathbf{u}\right).$$

Notice that, as above, the $\frac{1}{n}$ terms cancel each other out, leaving us with Equation (15). Rearranging, we have

$$\sqrt{n}(\mathbf{b} - \beta) = \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} n^{-1/2}\mathbf{X}'\mathbf{u}. \tag{18}$$

Now, let $\mathbf{X}_i'\mathbf{u}_i$ be the ith row of the matrix $\mathbf{X}'\mathbf{u}$. Using a CLT, we have the result that

$$n^{-1/2}[\mathbf{X}_i'\mathbf{u}_i - E(\mathbf{X}_i'\mathbf{u}_i)] \xrightarrow{d} N\left(\mathbf{0}, \lim_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n}\text{Var}(\mathbf{X}_i'\mathbf{u}_i)\right]\right).$$

The expected value of $\mathbf{X}_i'\mathbf{u}_i$ is

$$\begin{aligned} E(\mathbf{X}_i'\mathbf{u}_i) &= \mathbf{X}_i'E(\mathbf{u}_i) \\ &= 0, \end{aligned}$$

so its variance is

$$\begin{aligned} \text{Var}(\mathbf{X}_i'\mathbf{u}_i) &= E[(\mathbf{X}_i'\mathbf{u}_i)(\mathbf{X}_i'\mathbf{u}_i)'] \\ &= E(\mathbf{X}_i'\mathbf{u}_i\mathbf{u}_i'\mathbf{X}_i) \\ &= \mathbf{X}_i'E(\mathbf{u}_i\mathbf{u}_i')\mathbf{X}_i \\ &= \sigma_u^2\mathbf{X}_i'\mathbf{X}_i. \end{aligned}$$

Using Assumption (17), we thus have

$$\begin{aligned} \lim_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n}\text{Var}(\mathbf{X}_i'\mathbf{u}_i)\right] &= \lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_u^2\mathbf{X}_i'\mathbf{X}_i\right) \\ &= \sigma_u^2\lim_{n\to\infty}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right) \\ &= \sigma_u^2\mathbf{Q}. \end{aligned}$$

[23] In small samples, we can't generally say anything about the distribution of $\mathbf{b}$ unless we know the exact distribution of $\mathbf{X}$ and $\mathbf{u}$. For example, if $\mathbf{X}$ is not an RV, and $\mathbf{u}$ is normally distributed, then $\mathbf{b}$ is also normally distributed since it is a linear function of $\mathbf{u}$ (see Equation (15)).

Therefore, we can write

$$n^{-1/2}(\mathbf{X}_i'\mathbf{u}_i) \xrightarrow{d} N(\mathbf{0}, \sigma_u^2\mathbf{Q}).$$

Combining this with Assumption (17), we have

$$\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} n^{-1/2}(\mathbf{X}_i'\mathbf{u}_i) \xrightarrow{d} N(\mathbf{0}, \sigma_u^2\mathbf{Q}^{-1}\mathbf{Q}\mathbf{Q}^{-1}),$$

which, by Equation 18, means

$$\sqrt{n}(\mathbf{b} - \beta) \xrightarrow{d} N(0, \sigma_u^2\mathbf{Q}^{-1}).$$

That is, $\mathbf{b}$ is asymptotically normal.

3.5 Estimating the variance of the error terms of linear regression models

Even if we are willing to make the assumption that each of the error terms is IID with mean zero and variance σ_u^2, we do not usually know the value of σ_u^2. So, we typically would like to get some estimate of it. We start here by proposing the following estimator σ_u^2, which we denote s_u^2:

$$s_u^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k}.$$

While it may not be immediately clear where this comes from, we can show that it is an unbiased estimator of σ_u^2.

Let's start by substituting Equation (14) into the definition of $\hat{\mathbf{u}}$:

$$\begin{aligned}
\hat{\mathbf{u}} &= \mathbf{y} - \mathbf{X}\mathbf{b} \\
&= \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} \\
&= \mathbf{M}\mathbf{y},
\end{aligned}$$

where

$$\mathbf{M} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}').$$

It is easy to show that the matrix $\mathbf{M}$ is both symmetric and idempotent, i.e. that[24]

$$\mathbf{M}' = \mathbf{M},$$

and

$$\mathbf{M}\mathbf{M} = \mathbf{M}.$$

[24] This proof of this will be left as an assignment question.

Substituting Equation (13) into the above expression for $\hat{\mathbf{u}}$, we have

$$\begin{aligned}\hat{\mathbf{u}} &= \mathbf{M}\mathbf{y} \\ &= \mathbf{M}(\mathbf{X}\beta + \mathbf{u}) \\ &= \mathbf{M}\mathbf{X}\beta + \mathbf{M}\mathbf{u} \\ &= \mathbf{M}\mathbf{u},\end{aligned}$$

since

$$\begin{aligned}\mathbf{M}\mathbf{X} &= (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X} \\ &= \mathbf{X} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} \\ &= \mathbf{X} - \mathbf{X} \\ &= \mathbf{0}.\end{aligned}$$

Taking expectations of s_u^2, we have

$$\begin{aligned}E(s_u^2) &= \frac{E(\hat{\mathbf{u}}'\hat{\mathbf{u}})}{n-k} \\ &= \frac{E[(\mathbf{M}\mathbf{u})'(\mathbf{M}\mathbf{u})]}{n-k} \\ &= \frac{E(\mathbf{u}'\mathbf{M}'\mathbf{M}\mathbf{u})}{n-k} \\ &= \frac{E(\mathbf{u}'\mathbf{M}\mathbf{u})}{n-k}.\end{aligned}$$

Notice that $\mathbf{u}'\mathbf{M}\mathbf{u}$ is a scalar (check the dimensions to confirm this), and is therefore equal to its trace, i.e.

$$\mathbf{u}'\mathbf{M}\mathbf{u} = \mathrm{Tr}(\mathbf{u}'\mathbf{M}\mathbf{u}).$$

So, we can write

$$\begin{aligned}E(s_u^2) &= \frac{E[\mathrm{Tr}(\mathbf{u}'\mathbf{M}\mathbf{u})]}{n-k} \\ &= \frac{E[\mathrm{Tr}(\mathbf{M}\mathbf{u}\mathbf{u}')]}{n-k} \\ &= \frac{\mathrm{Tr}[\mathbf{M}E(\mathbf{u}\mathbf{u}')]}{n-k} \\ &= \frac{\sigma_u^2\mathrm{Tr}(\mathbf{M})}{n-k} \\ &= \sigma_u^2.\end{aligned}$$

since

$$\mathrm{Tr}(\mathbf{M}) = \mathrm{Tr}[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']$$

$$
\begin{aligned}
&= \text{Tr}(\mathbf{I}_n) - \text{Tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \\
&= n - \text{Tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] \\
&= n - \text{Tr}(\mathbf{I}_k) \\
&= n - k.
\end{aligned}
$$

So, we conclude that s_u^2 is an unbiased estimator of σ_u^2.

Note that since the variance of $\mathbf{b}$, depends on σ_u^2 (see Equation (31)), we can use s_u^2 to get an estimate of the variance of $\mathbf{b}$. This estimate, which we denote by $\hat{\text{Var}}(\mathbf{b})$, is found by replacing σ_u^2 by s_u^2 in Equation (31), i.e.

$$\hat{\text{Var}}(\mathbf{b}) = s_u^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Since s_u^2 is an unbiased estimator of σ_u^2, $\hat{\text{Var}}(\mathbf{b})$ is an unbiased estimate of $\text{Var}(\mathbf{b})$. The square root of $\hat{\text{Var}}(\mathbf{b})$ is often called the **standard error** of $\mathbf{b}$.

4 Hypothesis testing

4.1 Testing linear restrictions

Having established the properties of the OLS estimator, we now move on to showing how this estimator can be used to test various hypotheses. For illustrative purposes, let's start by considering the following linear regression model:

$$y_i = \beta_1 + \beta_2 x_{2,i} + \beta_3 x_{3,i} + u_i, \qquad i = 1, \ldots, n. \tag{19}$$

Some examples of null hypotheses we might wish to test are:

1. $H_0 : \beta_2 = \beta_{2,0}$, where $\beta_{2,0}$ is some specified value (often zero); and
2. $H_0 : \beta_2 = \beta_3 = 0$.

Both of these examples fit into the general linear framework

$$H_0 : \mathbf{R}\beta = \mathbf{r},$$

where, here, $\beta = (\ \beta_1 \quad \beta_2 \quad \beta_3\ )'$, and for

1. $\mathbf{R} = (\ 0 \quad 1 \quad 0\ )$, and $\mathbf{r} = \beta_{2,0}$; and
2. $\mathbf{R} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, and $\mathbf{r} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

Written this way, it can be seen that each of these null hypotheses imposes some linear restriction(s) to the original regression model in (19), which we will call the **unrestricted model**. Each of these null hypotheses can therefore be rewritten in terms of a **restricted model**:

1. $H_0 : y_i = \beta_1 + \beta_{2,0} x_{2,i} + \beta_3 x_{3,i} + u_i$; and

2. $H_0 : y_i = \beta_1 + u_i$.

In general, for the linear regression model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u},$$

where $\mathbf{X}$ is an $n \times k$ matrix, and β is a k-vector, $\mathbf{R}$ will be a $q \times k$ matrix, and $\mathbf{r}$ will be a q-vector, where q is the number of restrictions. Note that the first example above imposes a single restriction (i.e. $q = 1$), while the second imposes two restrictions (i.e. $q = 2$).

In order to test any such linear restriction(s), we use the OLS estimator of β,

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Recall that, if we assume that

$$\mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma_u^2\mathbf{I}_n),$$

then $\text{E}(\mathbf{b}) = \beta$ and $\text{Var}(\mathbf{b}) = \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}$. Therefore, we have

$$\begin{aligned} \text{E}(\mathbf{Rb}) &= \mathbf{R}\text{E}(\mathbf{b}) \\ &= \mathbf{R}\beta, \end{aligned}$$

and

$$\begin{aligned} \text{Var}(\mathbf{Rb}) &= \text{E}([\mathbf{Rb} - \text{E}(\mathbf{Rb})][\mathbf{Rb} - \text{E}(\mathbf{Rb})]') \\ &= \text{E}[(\mathbf{Rb} - \mathbf{R}\beta)(\mathbf{Rb} - \mathbf{R}\beta)'] \\ &= \text{E}[(\mathbf{Rb} - \mathbf{R}\beta)(\mathbf{b}'\mathbf{R}' - \beta'\mathbf{R}')] \\ &= \text{E}[\mathbf{R}(\mathbf{b} - \beta)(\mathbf{b}' - \beta')\mathbf{R}'] \\ &= \mathbf{R}\text{E}[(\mathbf{b} - \beta)(\mathbf{b} - \beta)']\mathbf{R}' \\ &= \mathbf{R}\text{Var}(\mathbf{b})\mathbf{R}' \\ &= \sigma_u^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'. \end{aligned}$$

If we make the much stronger assumption that

$$\mathbf{u} \sim \text{N}(\mathbf{0}, \sigma_u^2\mathbf{I}_n),$$

then

$$\mathbf{b} \sim \text{N}[\beta, \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}],$$

which implies that

$$\mathbf{Rb} \sim \text{N}[\mathbf{R}\beta, \sigma_u^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'],$$

or, alternatively, that

$$\mathbf{Rb} - \mathbf{R}\beta \sim \text{N}[\mathbf{0}, \sigma_u^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'].$$

Note that, under the null hypothesis, $\mathbf{R}\beta = \mathbf{r}$, so we can rewrite the above as

$$\mathbf{Rb} - \mathbf{r} \sim \mathrm{N}[\mathbf{0}, \sigma_u^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'].$$

Finally, we can rewrite this in quadratic form (see Appendix A) to get the test statistic

$$(\mathbf{Rb} - \mathbf{r})'[\sigma_u^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{r}) \sim \chi^2_{(q)}. \tag{20}$$

Unfortunately, the parameter σ_u^2 is usually unknown, so we have to estimate it using

$$s_u^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k}.$$

Replacing σ_u^2 in (20) by s_u^2 gives

$$\frac{(\mathbf{Rb} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{r})}{\hat{\mathbf{u}}'\hat{\mathbf{u}}/(n-k)}.$$

Dividing both the numerator and denominator of the above by σ_u^2, we have

$$\frac{(\mathbf{Rb} - \mathbf{r})'[\sigma_u^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{r})}{(\hat{\mathbf{u}}'\hat{\mathbf{u}}/\sigma_u^2)/(n-k)}. \tag{21}$$

Due to the randomness of $\hat{\mathbf{u}}$, this quantity is no longer distributed as $\chi^2_{(q)}$. However, as we saw earlier, the numerator is. Furthermore, as shown in Appendix A

$$\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{\sigma_u^2} \sim \chi^2_{(n-k)},$$

and is independent of the numerator in (21). Therefore, if we divide the numerator in (21) by its degrees of freedom (q), then we have a quantity which is distributed as $F_{(q,n-k)}$. That is,

$$\frac{(\mathbf{Rb} - \mathbf{r})'[\sigma_u^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{r})/q}{(\hat{\mathbf{u}}'\hat{\mathbf{u}}/\sigma_u^2)/(n-k)} \sim F_{(q,n-k)}.$$

Eliminating the σ_u^2 terms and substituting in for the definition of s_u^2, we have the test statistic

$$(\mathbf{Rb} - \mathbf{r})'[s_u^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{r})/q \sim F_{(q,n-k)}. \tag{22}$$

4.2 Testing a single restriction

Let's now see how we can construct specific test statistics for the first example considered above. First, note that

$$\mathbf{Rb} - \mathbf{r} = \mathbf{b}_2 - \beta_{2,0},$$

which is a scalar. Second, note that $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ "picks out" the 2nd diagonal element element the 3×3 matrix $(\mathbf{X}'\mathbf{X})^{-1}$. Therefore, since

$$\hat{\mathrm{Var}}(\mathbf{b}) = s_u^2(\mathbf{X}'\mathbf{X})^{-1},$$

we have

$$s_u^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}' = \hat{\text{Var}}(\mathbf{b}_2),$$

which is also a scalar (check the dimensions).

In general, if there are k parameters, and we want to test the null hypothesis that the jth one is equal to some specified value, i.e.

$$H_0 : \beta_j = \beta_{j,0},$$

we will have

$$\mathbf{Rb} - \mathbf{r} = \mathbf{b}_j - \beta_{j,0},$$

and

$$s_u^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}' = \hat{\text{Var}}(\mathbf{b}_j).$$

Substituting these into (22), we have the test statistic

$$\frac{(\mathbf{b}_j - \beta_{j,0})^2}{\hat{\text{Var}}(\mathbf{b}_j)} \sim F_{(1,n-k)}.$$

Alternatively, taking the square root of this statistic, we have what is known as a ***t*-statistic**:

$$\frac{\mathbf{b}_j - \beta_{j,0}}{\text{Se}(\mathbf{b}_j)} \sim t_{(n-k)}$$

4.3 Testing several restrictions

When there is more than one restriction to test, constructing an appropriate test statistic is not so simple. The main problem is that it will usually involve having to estimate both the unrestricted and restricted models. To estimate the unrestricted model, OLS can be used. However, to estimate the restricted model, we need to use a different method of estimation.

The idea is to estimate the model

$$y = X\beta + u,$$

subject to the restriction

$$\mathbf{R}\beta = \mathbf{r}.$$

To do so, we use the **restricted least squares** estimator of β, which we will denote $\mathbf{b}_*$. Letting $\hat{\mathbf{u}}_*$denote the residuals from the estimate of this model, we have

$$\hat{\mathbf{u}}_* = \mathbf{y} - \mathbf{X}\mathbf{b}_*. \tag{23}$$

The sum squared residuals is thus

$$\begin{aligned} \text{SSR}(\mathbf{b}_*) &= \hat{\mathbf{u}}_*'\hat{\mathbf{u}}_* \\ &= (\mathbf{y} - \mathbf{X}\mathbf{b}_*)'(\mathbf{y} - \mathbf{X}\mathbf{b}_*) \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{b}_*'\mathbf{X}'\mathbf{y} + \mathbf{b}_*'\mathbf{X}'\mathbf{X}\mathbf{b}_*. \end{aligned}$$

As with OLS, the restricted least squares estimator is based on minimizing this function. Our (constrained) optimization problem is thus

$$\min_{\mathbf{b}_*} \text{SSR}(\mathbf{b}_*) = \mathbf{y}'\mathbf{y} - 2\mathbf{b}_*'\mathbf{X}'\mathbf{y} + \mathbf{b}_*'\mathbf{X}'\mathbf{X}\mathbf{b}_*,$$

subject to

$$\mathbf{R}\mathbf{b}_* = \mathbf{r}.$$

The Lagrangian for this problem is

$$L(\mathbf{b}_*, \lambda) = \mathbf{y}'\mathbf{y} - 2\mathbf{b}_*'\mathbf{X}'\mathbf{y} + \mathbf{b}_*'\mathbf{X}'\mathbf{X}\mathbf{b}_* - 2\lambda(\mathbf{R}\mathbf{b}_* - \mathbf{r}),$$

and the first-order necessary conditions are

$$\frac{\partial L(\mathbf{b}_*, \lambda)}{\partial \mathbf{b}_*} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}_* - 2\mathbf{R}'\lambda = 0, \tag{24}$$

and

$$\frac{\partial L(\mathbf{b}_*, \lambda)}{\partial \lambda} = -2(\mathbf{R}\mathbf{b}_* - \mathbf{r}) = 0. \tag{25}$$

Equation (24) implies

$$\mathbf{X}'\mathbf{X}\mathbf{b}_* = \mathbf{X}'\mathbf{y} + \mathbf{R}'\lambda.$$

Solving for $\mathbf{b}_*$, we have

$$\begin{aligned} \mathbf{b}_* &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}'\lambda \\ &= \mathbf{b} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}'\lambda, \end{aligned} \tag{26}$$

where $\mathbf{b}$ is the usual OLS estimator. Premultiplying by $\mathbf{R}$, we have

$$\mathbf{R}\mathbf{b}_* = \mathbf{R}\mathbf{b} + \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}'\lambda,$$

which implies

$$\mathbf{R}\mathbf{b}_* - \mathbf{R}\mathbf{b} = \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}'\lambda,$$

and therefore,

$$[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b}_* - \mathbf{R}\mathbf{b}) = \lambda, \tag{27}$$

Equation (25) implies

$$\mathbf{R}\mathbf{b}_* = \mathbf{r}.$$

Substituting this into Equation (27), we have

$$[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{R}\mathbf{b}) = \lambda.$$

Finally, substituting this into Equation (26), we have

$$\mathbf{b}_* = \mathbf{b} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{R}\mathbf{b}). \tag{28}$$

Let's now focus on the residuals from restricted model. Adding and subtracting $\mathbf{Xb}$ to Equation (23), we have

$$\begin{aligned}\hat{\mathbf{u}}_* &= \mathbf{y} - \mathbf{Xb} - \mathbf{Xb}_* + \mathbf{Xb} \\ &= \hat{\mathbf{u}} - \mathbf{X}(\mathbf{b}_* - \mathbf{b}).\end{aligned}$$

The sum of squared residuals from the restricted model is thus[25]

$$\begin{aligned}\hat{\mathbf{u}}_*'\hat{\mathbf{u}}_* &= [\hat{\mathbf{u}} - \mathbf{X}(\mathbf{b}_* - \mathbf{b})]'[\hat{\mathbf{u}} - \mathbf{X}(\mathbf{b}_* - \mathbf{b})] \\ &= [\hat{\mathbf{u}}' - (\mathbf{b}_* - \mathbf{b})'\mathbf{X}'][\hat{\mathbf{u}} - \mathbf{X}(\mathbf{b}_* - \mathbf{b})] \\ &= \hat{\mathbf{u}}'\hat{\mathbf{u}} - \hat{\mathbf{u}}'\mathbf{X}(\mathbf{b}_* - \mathbf{b}) - (\mathbf{b}_* - \mathbf{b})'\mathbf{X}'\hat{\mathbf{u}} + (\mathbf{b}_* - \mathbf{b})'\mathbf{X}'\mathbf{X}(\mathbf{b}_* - \mathbf{b}) \\ &= \hat{\mathbf{u}}'\hat{\mathbf{u}} + (\mathbf{b}_* - \mathbf{b})'\mathbf{X}'\mathbf{X}(\mathbf{b}_* - \mathbf{b}).\end{aligned}$$

Substituting for $\mathbf{b}_* - \mathbf{b}$ from Equation (28), we have

$$\begin{aligned}\hat{\mathbf{u}}_*'\hat{\mathbf{u}}_* &= \hat{\mathbf{u}}'\hat{\mathbf{u}} + \{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{Rb})\}' \\ &\quad \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{Rb}) \\ &= \hat{\mathbf{u}}'\hat{\mathbf{u}} + (\mathbf{r} - \mathbf{Rb})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}']^{-1}\mathbf{RX}(\mathbf{X}'\mathbf{X})^{-1} \\ &\quad \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{Rb}) \\ &= \hat{\mathbf{u}}'\hat{\mathbf{u}} + (\mathbf{r} - \mathbf{Rb})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{Rb}),\end{aligned}$$

Denoting $\hat{\mathbf{u}}_*'\hat{\mathbf{u}}_*$ (the sum of squared residuals from the restricted model) by $\mathrm{SSR_R}$, and $\hat{\mathbf{u}}'\hat{\mathbf{u}}$ (the sum of squared residuals from the unrestricted model) by $\mathrm{SSR_U}$, we have

$$\mathrm{SSR_R} - \mathrm{SSR_U} = (\mathbf{r} - \mathbf{Rb})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{Rb}).$$

Notice that the term on the right hand side is in the numerator of (22). This leads to what is known as the ***F*-statistic**:

$$\frac{(\mathrm{SSR_R} - \mathrm{SSR_U})/q}{\mathrm{SSR_U}/(n-k)} \sim F_{(q,n-k)}. \tag{29}$$

4.4 Bootstrapping

The test statistics considered above were built on the assumption that $\mathbf{u}$ is normally distributed. However, quite often, we may be unwilling to make this assumption. In such cases, we therefore do not know the distribution of our test statistics.

This problem leads to a procedure known as the **bootstrap**.[26] The basic idea is to use the observed data to try to estimate the distribution of the relevant test statistic. In general, this procedure requires the following steps:

[25] Here, we use make use of the fact that $\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$. Proving this would be a useful exercise.

[26] Here, we present a very brief introduction to the bootstrap. A more complete introduction can be found in Davidson and MacKinnon, (2004, Section 4.6).

1. Using the observed data, estimate both the unrestricted and restricted models. Save the fitted values and residuals from the restricted model. Call these $\hat{\mathbf{y}}_{\mathrm{R}}$ and $\hat{\mathbf{u}}_{\mathrm{R}}$, respectively.

2. Using the estimates from the previous step, calculate the test statistic of interest (e.g. the t-statistic or the F-statistic). Cal this $\hat{T}$.

3. Randomly sample, with replacement, n of the residuals from the restricted model (this is known as **resampling**). Call these $\hat{\mathbf{u}}^*$. Generate the **bootstrap sample** as:
$$\mathbf{y}^* = \hat{\mathbf{y}}_{\mathrm{R}} + \hat{\mathbf{u}}^*$$

4. Using the bootstrap sample, re-estimate both the unrestricted and restricted models.

5. Using the estimates from the previous step, calculate the test statistic of interest (this is known as the **bootstrap test statistic**, since it is based on the bootstrap sample). Call this $\hat{T}_b^*$.

Finally, repeat Steps 3-5 B (some large number[27]) times. What this leaves you with is the original test statistic, $\hat{T}$, and B different bootstrap test statistics, $\hat{T}_b^*$, $b = 1, \ldots, B$.

It turns out that these B bootstrap test statistics provide a fairly good estimate of the distribution of the test statistic of interest. For inference purposes, we can calculate the **simulated *P*-value** of the original test statistic , $\hat{T}$, as

$$\hat{p}^*(\hat{T}) = \frac{1}{B}\sum_{b=1}^{B} I(\hat{T}_b^* > \hat{T}),$$

where $I(\cdot)$ is an indicator function, equal to one if $\hat{T}_b^* > \hat{T}$, and zero otherwise.

5 Generalized least squares

5.1 Introduction

In estimating the linear regression model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \tag{30}$$

by OLS, we made the (very strong) assumption that

$$\mathbf{u} \sim \mathrm{IID}(\mathbf{0}, \sigma_u^2 \mathbf{I}_n).$$

[27] For a chosen level of significance, α, for the test, B should be chosen so that $\alpha(B+1)$ is an integer (see Davidson and MacKinnon, (2004, Section 4.6)). For $\alpha = 0.01$, appropriate values of B would be 99, 199, 299, and so on. Since computing costs are now so low, B is often chosen to be 999 or even 9999.

We would now like to relax this assumption, and consider what happens when the error terms are not IID. While we will continue to assume that

$$\mathrm{E}(\mathbf{u}) = \mathbf{0},$$

we will now let

$$\begin{aligned} \mathrm{Var}(\mathbf{u}) &= \mathrm{E}\begin{pmatrix} u_1^2 & \cdots & u_1 u_n \\ \vdots & & \vdots \\ u_n u_1 & \cdots & u_n^2 \end{pmatrix} \\ &= \mathbf{\Omega}, \end{aligned}$$

where $\mathbf{\Omega}$, the error covariance matrix, is some $n \times n$ positive definite matrix. That is, we allow the possibility that there is some $\mathrm{E}(u_i u_j) \neq 0$ for $i \neq j$ (and that therefore u_i and u_j are not independent), and that there is some $\mathrm{E}(u_i^2) \neq \mathrm{E}(u_j^2)$ for $i \neq j$ (and that therefore u_i and u_j are not identically distributed). In other words, we allow the possibility that each u_i is not IID.

Notice that the OLS estimator is still unbiased since

$$\begin{aligned} E(\mathbf{b}) &= \mathrm{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= \mathrm{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\mathbf{b} + \mathbf{u})] \\ &= \mathrm{E}\left[\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\right] \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' E(\mathbf{u}) \\ &= \beta. \end{aligned}$$

However, the variance of the OLS estimator is now different:

$$\begin{aligned} \mathrm{Var}(\mathbf{b}) &= E\left([(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}][(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}]'\right) \\ &= E\left[(\mathbf{X}'\mathbf{X})\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' E(\mathbf{u}\mathbf{u}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{\Omega}(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \tag{31}$$

It turns out (as we will see shortly), that, in this more general setting, OLS is no longer the "best" linear unbiased estimator (BLUE). That is, there is some other linear unbiased estimator which has a variance smaller than (32).

In deriving this "better" estimator, the basic strategy is to transform the linear regression model in (30) so that the error terms become IID. To do so, start by letting

$$\mathbf{\Omega}^{-1} = \mathbf{\Psi}\mathbf{\Psi}'.$$

where $\mathbf{\Psi}$ is some $n \times n$ matrix, which is usually triangular.

Next, multiply both sides of Equation (30) by $\mathbf{\Psi}'$:

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

It is convenient to rewrite this as

$$\mathbf{y}_* = \mathbf{X}_*\beta + \mathbf{u}_* \tag{32}$$

where

$$\mathbf{y}_* = \mathbf{\Psi}'\mathbf{y},$$
$$\mathbf{X}_* = \mathbf{\Psi}'\mathbf{X},$$

and

$$\mathbf{u}_* = \mathbf{\Psi}'\mathbf{u}.$$

Note that

$$\begin{aligned} E(\mathbf{u}_*) &= E(\mathbf{\Psi}'\mathbf{u}) \\ &= \mathbf{\Psi}'E(\mathbf{u}) \\ &= \mathbf{0} \end{aligned}$$

and that

$$\begin{aligned} \text{Var}(\mathbf{u}_*) &= E\{[\mathbf{u}_* - E(\mathbf{u}_*)][\mathbf{u}_* - E(\mathbf{u}_*)']\} \\ &= E(\mathbf{u}_*\mathbf{u}_*') \\ &= E[(\mathbf{\Psi}'\mathbf{u})(\mathbf{\Psi}'\mathbf{u})'] \\ &= E(\mathbf{\Psi}'\mathbf{u}\mathbf{u}'\mathbf{\Psi}) \\ &= \mathbf{\Psi}'E(\mathbf{u}\mathbf{u}')\mathbf{\Psi} \\ &= \mathbf{\Psi}'\text{Var}(\mathbf{u})\mathbf{\Psi} \\ &= \mathbf{\Psi}'\Omega\mathbf{\Psi} \\ &= \mathbf{\Psi}'(\mathbf{\Psi}\mathbf{\Psi}')^{-1}\mathbf{\Psi} \\ &= \mathbf{\Psi}'(\mathbf{\Psi}')^{-1}\mathbf{\Psi}^{-1}\mathbf{\Psi} \\ &= \mathbf{I}_n. \end{aligned}$$

That is,

$$\mathbf{u}_* \sim \text{IID}(\mathbf{0}, \mathbf{I}_n).$$

Applying OLS to Equation (32), we get the **generalized least squares** (or **GLS**) estimator

$$\begin{aligned} \mathbf{b}_{\text{GLS}} &= (\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{y}_* \\ &= [(\mathbf{\Psi}'\mathbf{X})'\mathbf{\Psi}'\mathbf{X}]^{-1}(\mathbf{\Psi}'\mathbf{X})'\mathbf{\Psi}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{\Psi}\mathbf{\Psi}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Psi}\mathbf{\Psi}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y}. \end{aligned} \tag{33}$$

We now want to show that $\mathbf{b}_{\text{GLS}}$ is unbiased. Substituting Equation (30) into Equation (33), we have

$$\begin{aligned}
\mathbf{b}_{\text{GLS}} &= (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}(\mathbf{X}\mathbf{b}+\mathbf{u}) \\
&= (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\mathbf{b} + (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{u} \\
&= \beta + (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{u}. \qquad (34)
\end{aligned}$$

Taking expectations we have

$$\begin{aligned}
E(\mathbf{b}_{\text{GLS}}) &= \text{E}\left[\beta + (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{u}\right] \\
&= \beta + (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}E(\mathbf{u}) \\
&= \beta,
\end{aligned}$$

since $E(\mathbf{u}) = \mathbf{0}$. So $\mathbf{b}_{\text{GLS}}$ is unbiased.

Let's now move on to the variance of $\mathbf{b}_{\text{GLS}}$:

$$\begin{aligned}
\text{Var}(\mathbf{b}_{\text{GLS}}) &= \text{E}\left([\mathbf{b}_{\text{GLS}} - E(\mathbf{b}_{\text{GLS}})][\mathbf{b}_{\text{GLS}} - E(\mathbf{b}_{\text{GLS}})]'\right) \\
&= \text{E}\left[(\mathbf{b}_{\text{GLS}} - \beta)(\mathbf{b}_{\text{GLS}} - \beta)'\right].
\end{aligned}$$

Substituting from Equation (34), we have

$$\begin{aligned}
\text{Var}(\mathbf{b}_{\text{GLS}}) &= \text{E}\{[(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{u}][(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{u}]'\} \\
&= \text{E}[(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{u}\mathbf{u}'\mathbf{\Omega}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}] \\
&= (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}E(\mathbf{u}\mathbf{u}')\mathbf{\Omega}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\text{Var}(\mathbf{u}\mathbf{u}')\mathbf{\Omega}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{\Omega}\mathbf{\Omega}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}. \qquad (35)
\end{aligned}$$

We now want to consider a generalization of the Gauss-Markov theorem, which states that, if $\text{Var}(\mathbf{u}) = \mathbf{\Omega}$, $\mathbf{b}_{\text{GLS}}$ is BLUE. The proof of this will be very similar to the proof of the Gauss-Markov theorem for $\mathbf{b}$ (except that, there, we assumed that $\text{Var}(\mathbf{u}) = \sigma_u^2\mathbf{I}_n$). We again consider some other linear, unbiased estimator of β, which we will again call $\mathbf{B}$:

$$\mathbf{B} = \mathbf{G}\mathbf{y},$$

where, here,

$$\mathbf{G} = \mathbf{W} + (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1},$$

and $\mathbf{W}$ is some non-zero matrix which depends on $\mathbf{X}$. The expected value of $\mathbf{B}$ is thus

$$\begin{aligned}
\text{E}(\mathbf{B}) &= \text{E}(\mathbf{G}\mathbf{y}) \\
&= \mathbf{G}\text{E}(\mathbf{y}) \\
&= \mathbf{G}\text{E}(\mathbf{X}\beta + \mathbf{u}) \\
&= \mathbf{G}\mathbf{X}\beta + \mathbf{G}\text{E}(\mathbf{u}) \\
&= \mathbf{G}\mathbf{X}\beta
\end{aligned}$$

This means that, in order for $\mathbf{B}$ to be unbiased, we must have

$$\mathbf{GX} = \mathbf{I}_k,$$

where $\mathbf{I}_k$ is an $k \times k$ identity matrix. Given this condition, notice that

$$\begin{aligned} \mathbf{B} &= \mathbf{Gy} \\ &= \mathbf{G}(\mathbf{X}\beta + \mathbf{u}) \\ &= \mathbf{GX}\beta + \mathbf{Gu} \\ &= \beta + \mathbf{Gu}, \end{aligned}$$

and that

$$\begin{aligned} \mathbf{WX} &= [\mathbf{G} - (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}]\mathbf{X} \\ &= \mathbf{GX} - (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X} \\ &= \mathbf{I}_k - \mathbf{I}_k \\ &= \mathbf{0}. \end{aligned}$$

The variance of $\mathbf{B}$ is therefore

$$\begin{aligned} \mathrm{Var}(\mathbf{B}) &= \mathrm{E}\left([\mathbf{B} - \mathrm{E}(\mathbf{B})][\mathbf{b} - \mathrm{E}(\mathbf{B})]'\right) \\ &= \mathrm{E}[(\beta + \mathbf{Gu} - \beta)(\beta + \mathbf{Gu} - \beta)'] \\ &= \mathrm{E}[(\mathbf{Gu})(\mathbf{Gu})'] \\ &= \mathrm{E}(\mathbf{Guu}'\mathbf{G}') \\ &= \mathbf{G}\mathrm{E}(\mathbf{uu}')\mathbf{G}' \\ &= \mathbf{G\Omega G}' \\ &= [\mathbf{W} + (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}]\mathbf{\Omega}[\mathbf{W} + (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}]' \\ &= [\mathbf{W\Omega} + (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{\Omega}][\mathbf{W}' + \mathbf{\Omega}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}] \\ &= \mathbf{W\Omega W}' + \mathbf{W\Omega\Omega}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}' \\ &\quad +(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1} \\ &= \mathbf{W\Omega W}' + (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}' \end{aligned}$$

Finally, since $\mathbf{W}$ is a non-zero matrix and $\mathbf{\Omega}$ is a positive definite matrix,

$$\mathrm{Var}(\mathbf{B}) > \mathrm{Var}(\mathbf{b}_{\mathrm{GLS}}),$$

which proves our generalized Gauss-Markov theorem. It should be noted that, if $\mathrm{Var}(\mathbf{u}) \neq \sigma_u^2\mathbf{I}_n$, then $\mathbf{B}$ includes $\mathbf{b}$ as a special case[28], and we have

$$\mathrm{Var}(\mathbf{b}) > \mathrm{Var}(\mathbf{b}_{\mathrm{GLS}}).$$

[28] On the other hand, $\mathbf{b}_{\mathrm{GLS}}$ includes $\mathbf{b}$ as a special case when $\mathrm{Var}(\mathbf{u}) = \sigma_u^2\mathbf{I}_n$. Proving this would be a useful exercise.

5.2 Feasible Generalized Least Squares

While the GLS estimator has some highly desirable properties, the main drawback with this method is that $\mathbf{\Omega}$ is almost never actually known in practice. That is, GLS is not usually feasible. However, we can use some estimate of $\mathbf{\Omega}$, which we denote $\hat{\mathbf{\Omega}}$. Replacing $\mathbf{\Omega}$ by $\hat{\mathbf{\Omega}}$ in Equation (33), we have the **feasible generalized least squares** (or **FGLS**) estimator

$$\mathbf{b}_{\text{FGLS}} = (\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{y}.$$

How we actually go about estimating $\mathbf{\Omega}$ depends on the particular problem. Below, we analyze two of the most common problems encountered in practice.

5.3 Heteroskedasticity

The problem of **heteroskedasticity** occurs when $\mathbf{\Omega}$ is a diagonal matrix (meaning each u_i is independent), but each diagonal element may not be identical (meaning each u_i is not identically distributed). That is, letting $\text{Var}(u_i) = \sigma_{u_i}^2$, we have

$$\mathbf{\Omega} = \begin{pmatrix} \sigma_{u_1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{u_2}^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_{u_1}^2 \end{pmatrix}.$$

Since $\mathbf{\Omega}$ is diagonal, we have

$$\mathbf{\Omega}^{-1} = \begin{pmatrix} 1/\sigma_{u_1}^2 & 0 & \cdots & 0 \\ 0 & 1/\sigma_{u_2}^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1/\sigma_{u_1}^2 \end{pmatrix}.$$

So, if we let

$$\mathbf{\Psi} = \mathbf{\Psi}' = \begin{pmatrix} 1/\sigma_{u_1} & 0 & \cdots & 0 \\ 0 & 1/\sigma_{u_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1/\sigma_{u_n} \end{pmatrix},$$

then $\mathbf{\Omega}^{-1} = \mathbf{\Psi}\mathbf{\Psi}'$, as desired.

Therefore, transforming our variables as above, we have

$$\begin{aligned} \mathbf{y}_* &= \mathbf{\Psi}'\mathbf{y} \\ &= \begin{pmatrix} 1/\sigma_{u_1} & 0 & \cdots & 0 \\ 0 & 1/\sigma_{u_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1/\sigma_{u_n} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \end{aligned}$$

$$= \begin{pmatrix} y_1/\sigma_{u_1} \\ y_2/\sigma_{u_2} \\ \vdots \\ y_n/\sigma_{u_n} \end{pmatrix},$$

$$\begin{aligned} \mathbf{X}_* &= \mathbf{\Psi}'\mathbf{X} \\ &= \begin{pmatrix} 1/\sigma_{u_1} & 0 & \cdots & 0 \\ 0 & 1/\sigma_{u_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1/\sigma_{u_n} \end{pmatrix} \begin{pmatrix} x_{1,1} & \cdots & x_{k,1} \\ x_{1,2} & \cdots & x_{k,2} \\ \vdots & & \vdots \\ x_{1,n} & \cdots & x_{k,n} \end{pmatrix}, \\ &= \begin{pmatrix} x_{1,1}/\sigma_{u_1} & \cdots & x_{k,1}/\sigma_{u_1} \\ x_{1,2}/\sigma_{u_2} & \cdots & x_{k,2}/\sigma_{u_2} \\ \vdots & & \vdots \\ x_{1,n}/\sigma_{u_n} & \cdots & x_{k,n}/\sigma_{u_n} \end{pmatrix}, \end{aligned}$$

and

$$\begin{aligned} \mathbf{u}_* &= \mathbf{\Psi}'\mathbf{u} \\ &= \begin{pmatrix} 1/\sigma_{u_1} & 0 & \cdots & 0 \\ 0 & 1/\sigma_{u_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1/\sigma_{u_n} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \\ &= \begin{pmatrix} u_1/\sigma_{u_1} \\ u_2/\sigma_{u_2} \\ \vdots \\ u_n/\sigma_{u_n} \end{pmatrix}. \end{aligned}$$

So, in this case, the transformed model,

$$\mathbf{y}_* = \mathbf{X}_*\beta + \mathbf{u}_*,$$

can be written as

$$\begin{pmatrix} y_1/\sigma_{u_1} \\ y_2/\sigma_{u_2} \\ \vdots \\ y_n/\sigma_{u_n} \end{pmatrix} = \begin{pmatrix} x_{1,1}/\sigma_{u_1} & \cdots & x_{k,1}/\sigma_{u_1} \\ x_{1,2}/\sigma_{u_2} & \cdots & x_{k,2}/\sigma_{u_2} \\ \vdots & & \vdots \\ x_{1,n}/\sigma_{u_n} & \cdots & x_{k,n}/\sigma_{u_n} \end{pmatrix} \beta + \begin{pmatrix} u_1/\sigma_{u_1} \\ u_2/\sigma_{u_2} \\ \vdots \\ u_n/\sigma_{u_n} \end{pmatrix}.$$

Alternatively, for the ith observation, we have

$$\frac{y_i}{\sigma_{u_i}} = \beta_1 \frac{x_{1,i}}{\sigma_{u_i}} + \ldots + \beta_k \frac{x_{k,i}}{\sigma_{u_i}} + \frac{u_i}{\sigma_{u_i}}, \qquad i = 1, \ldots, n.$$

Estimating this model by OLS is a special case of GLS known as **weighted least squares** (or **WLS**). Of course, this requires knowing the value of each

σ_{u_i} (i.e. knowing $\mathbf{\Omega}$), which, as mentioned above, is almost never possible. Accordingly, we need some method to estimate $\mathbf{\Omega}$ so that we can use FGLS (or, in this case, what might be referred to as feasible WLS).

Typically, this is done as follows. Suppose we think that the variance of $\mathbf{u}$ can be explained by $\mathbf{Z}$, an $n \times r$ matrix which may or may not include some of the columns of $\mathbf{X}$. A fairly general non-linear function describing this relationship is

$$\mathrm{E}(u_i^2) = \alpha_1 z_{1,i}^{\gamma_1} + \ldots + \alpha_r z_{r,i}^{\gamma_r}, \qquad i = 1, \ldots, n.$$

Letting v_i denote the difference between $\mathrm{E}(u_i^2)$ and u_i^2, we have

$$u_i^2 = \alpha_1 z_{1,i}^{\gamma_1} + \ldots + \alpha_r z_{r,i}^{\gamma_r} + v_i, \qquad i = 1, \ldots, n,$$

which is in the form of a non-linear regression model. Unfortunately, we don't have time to cover non-linear regression models in this course, but estimating such a model is certainly possible. However, since we can't actually observe u_i^2, we would have to use the residuals from the OLS estimation of model (30) , $\hat{u}_i^2$, in their place. That is, we could estimate the non-linear regression model

$$\hat{u}_i^2 = \alpha_1 z_{1,i}^{\gamma_1} + \ldots + \alpha_r z_{r,i}^{\gamma_r} + w_i, \qquad i = 1, \ldots, n,$$

where w_i is some error term.[29] Using the parameter estimates from this model, we could construct estimates of $\sigma_{u_i}^2$ as follows:

$$\hat{\sigma}_{u_i}^2 = \hat{\alpha}_1 z_{1,i}^{\hat{\gamma}_1} + \ldots + \hat{\alpha}_r z_{r,i}^{\hat{\gamma}_r} + w_i, \qquad i = 1, \ldots, n.$$

Using the square root of these estimates, $\hat{\sigma}_{u_i}$, our FGLS estimators are found by estimating

$$\frac{y_i}{\hat{\sigma}_{u_i}} = \beta_1 \frac{x_{1,i}}{\hat{\sigma}_{u_i}} + \ldots + \beta_k \frac{x_{k,i}}{\hat{\sigma}_{u_i}} + \frac{u_i}{\hat{\sigma}_{u_i}}, \qquad i = 1, \ldots, n,$$

by OLS.

5.4 Autocorrelation

The problem of **autocorrelation** (or **serial correlation**) occurs when the off-diagonal elements in $\mathbf{\Omega}$ are non-zero (meaning each u_i is no longer independent). Here, we will assume that the errors are homoskedastic (meaning each diagonal element in $\mathbf{\Omega}$ is identical), but it is possible that autocorrelation and heteroskedasticity are simultaneously present.

Autocorrelation is usually encountered in time-series data, where the error terms may be generated by some autoregressive process. For example, suppose the error terms follow a (linear) first-order autoregressive process (or AR(1) process):

$$u_t = \rho u_{t-1} + v_t, \qquad t = 1, \ldots, n, \tag{36}$$

[29] This model could also be used to testing for the presence of heteroskedasticity. Here, the null hypothesis would be $H_0 : \alpha_1 = \ldots = \alpha_r = 0$ (implying homoskedasticity). An F-test could be used to test this restriction.

where

$$|\rho| < 1,$$

and

$$v_t \sim \text{IID}(0, \sigma_v^2).$$

The condition that $|\rho| < 1$ is imposed so that the AR(1) process in (36) is what is known as **covariance stationary**. A covariance stationary process for u_t is one in which $\text{E}(u_t)$, $\text{Var}(u_t)$, and $\text{Cov}(u_t, u_{t-j})$, for any given j, are independent of t.

One way to see this is to imagine that, although we only start observing it at $t = 1$, the series for u_t has been in existence for an infinite time. First, note that (36) implies that

$$u_{t-1} = \rho u_{t-2} + v_{t-1},$$
$$u_{t-2} = \rho u_{t-3} + v_{t-2},$$

and so on. Substituting these into (36), we have

$$\begin{aligned} u_t &= \rho(\rho u_{t-2} + v_{t-1}) + v_t \\ &= \rho^2 u_{t-2} + \rho v_{t-1} + v_t \\ &= \rho^2(\rho u_{t-3} + v_{t-2}) + \rho v_{t-1} + v_t \\ &= \rho^3 u_{t-3} + \rho^2 v_{t-2} + \rho v_{t-1} + v_t, \end{aligned}$$

and so on. Alternatively, this can be rewritten as

$$\begin{aligned} u_t &= v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \rho^2 v_{t-2} + \rho^3 v_{t-3} + \dots \\ &= \sum_{i=0}^{\infty} \rho^i v_{t-i}. \end{aligned}$$

Taking expectations, we have

$$\begin{aligned} \text{E}(u_t) &= \text{E}\left(\sum_{i=0}^{\infty} \rho^i v_{t-i}\right) \\ &= \sum_{i=0}^{\infty} \rho^i \text{E}(v_{t-i}) \\ &= 0, \end{aligned}$$

since $E(v_t) = 0$. So $\text{E}(u_t)$ is independent of t.

Next, since v_t is IID,

$$\begin{aligned} \text{Var}(u_t) &= \text{Var}\left(\sum_{i=0}^{\infty} \rho^i v_{t-i}\right) \\ &= \sum_{i=0}^{\infty} \rho^{2i} \text{Var}(v_{t-i}) \\ &= \sum_{i=0}^{\infty} \rho^{2i} \sigma_v^2, \end{aligned}$$

which is a infinite geometric series (since $|\rho| < 1$). Therefore, we can write

$$\text{Var}(u_i) = \frac{\sigma_v^2}{1-\rho^2}.$$

So, $\text{Var}(u_t)$ is also independent of t. On the other hand, if $|\rho| \geq 1$ this infinite series would not converge (rather, it would "explode"), and therefore depend on t (and make the process **non-stationary**).

Finally, let's consider the covariance between u_t and u_{t-j}, for any given j. First, note that

$$\begin{aligned}
\text{Cov}(u_t, u_{t-1}) &= \text{E}(u_t u_{t-1}) \\
&= \text{E}[(\rho u_{t-1} + v_t)u_{t-1}] \\
&= \text{E}(\rho u_{t-1}^2 + v_t u_{t-1}) \\
&= \rho\text{E}(u_{t-1}^2) \\
&= \rho\text{Var}(u_{t-1}) \\
&= \frac{\rho\sigma_v^2}{1-\rho^2}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\text{Cov}(u_t, u_{t-2}) &= \text{E}(u_t u_{t-2}) \\
&= \text{E}[(\rho u_{t-1} + v_t)u_{t-2}] \\
&= \text{E}(\rho u_{t-1}u_{t-2} + v_t u_{t-2}) \\
&= \rho\text{E}(u_{t-1}u_{t-2}) \\
&= \frac{\rho^2\sigma_v^2}{1-\rho^2}.
\end{aligned}$$

Generalizing these two results, we have

$$\text{Cov}(u_t, u_{t-j}) = \frac{\rho^j\sigma_v^2}{1-\rho^2},$$

which is also independent of t. Since this result depends on $\text{Var}(u_t)$ being constant, it therefore depends on the condition $|\rho| < 1$.

Using the above results for $\text{Var}(u_t)$ and $\text{Cov}(u_t, u_{t-j})$, we have

$$\boldsymbol{\Omega} = \frac{\sigma_v^2}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix}.$$

Inverting this matrix, we have

$$\mathbf{\Omega}^{-1} = \sigma_v^2 \begin{pmatrix} 1 & -\rho & 0 & \cdots & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

(this can be confirmed by showing that $\mathbf{\Omega}\mathbf{\Omega}^1 = \mathbf{I}_n$). So, if we let

$$\mathbf{\Psi}' = \sigma_v \begin{pmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 \\ -\rho & 1 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix},$$

then $\mathbf{\Omega}^{-1} = \mathbf{\Psi}\mathbf{\Psi}'$, as desired.

Therefore, transforming our variables as above, we have

$$\begin{aligned} \mathbf{y}_* &= \mathbf{\Psi}'\mathbf{y} \\ &= \sigma_v \begin{pmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 \\ -\rho & 1 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} \\ &= \sigma_v \begin{pmatrix} y_1\sqrt{1-\rho^2} \\ y_2 - \rho y_1 \\ y_3 - \rho y_2 \\ \vdots \\ y_n - \rho y_{n-1} \end{pmatrix}, \end{aligned}$$

$$\begin{aligned} \mathbf{X}_* &= \mathbf{\Psi}'\mathbf{X} \\ &= \sigma_v \begin{pmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 \\ -\rho & 1 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} x_{1,1} & \cdots & x_{k,1} \\ x_{1,2} & \cdots & x_{k,2} \\ \vdots & & \vdots \\ x_{1,n} & \cdots & x_{k,n} \end{pmatrix}, \\ &= \sigma_v \begin{pmatrix} x_{1,1}\sqrt{1-\rho^2} & \cdots & x_{k,1}\sqrt{1-\rho^2} \\ x_{1,2} - \rho x_{1,1} & \cdots & x_{k,2} - \rho x_{k,1} \\ \vdots & & \vdots \\ x_{1,n} - \rho x_{1,n-1} & \cdots & x_{k,n} - \rho x_{k,n-1} \end{pmatrix}, \end{aligned}$$

and

$$\begin{aligned}
\mathbf{u}_* &= \mathbf{\Psi}'\mathbf{u} \\
&= \sigma_v \begin{pmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 \\ -\rho & 1 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{pmatrix} \\
&= \sigma_v \begin{pmatrix} u_1\sqrt{1-\rho^2} \\ u_2 - \rho u_1 \\ u_3 - \rho u_2 \\ \vdots \\ u_n - \rho u_{n-1} \end{pmatrix},
\end{aligned}$$

So, in this case, the transformed model,

$$\mathbf{y}_* = \mathbf{X}_*\beta + \mathbf{u}_*,$$

can be written as

$$\begin{aligned}
\sigma_v \begin{pmatrix} y_1\sqrt{1-\rho^2} \\ y_2 - \rho y_1 \\ y_3 - \rho y_2 \\ \vdots \\ y_n - \rho y_{n-1} \end{pmatrix} &= \sigma_v \begin{pmatrix} x_{1,1}\sqrt{1-\rho^2} & \cdots & x_{k,1}\sqrt{1-\rho^2} \\ x_{1,2} - \rho x_{1,1} & \cdots & x_{k,2} - \rho x_{k,1} \\ \vdots & & \vdots \\ x_{1,n} - \rho x_{1,n-1} & \cdots & x_{k,n} - \rho x_{k,n-1} \end{pmatrix} \beta \\
&\quad + \sigma_v \begin{pmatrix} u_1\sqrt{1-\rho^2} \\ u_2 - \rho u_1 \\ u_3 - \rho u_2 \\ \vdots \\ u_n - \rho u_{n-1} \end{pmatrix}.
\end{aligned}$$

Alternatively, for the 1st observation, we have

$$\sigma_v y_1\sqrt{1-\rho^2} = \beta_1\sigma_v x_{1,1}\sqrt{1-\rho^2} + \ldots + \beta_k\sigma_v x_{k,1}\sqrt{1-\rho^2} + \sigma_v u_1\sqrt{1-\rho^2},$$

and for t-th observation, $t = 2, \ldots, n$, we have

$$\sigma_v(y_t - \rho y_{t-1}) = \beta_1\sigma_v(x_{1,t} - \rho x_{1,t-1}) + \ldots + \beta_k\sigma_v(x_{k,t} - \rho x_{k,t-1}) + \sigma_v(u_t - \rho u_{t-1}).$$

Of course, since σ_v is a constant, we can divide it out of each side of the above to get

$$y_1\sqrt{1-\rho^2} = \beta_1 x_{1,1}\sqrt{1-\rho^2} + \ldots + \beta_k x_{k,1}\sqrt{1-\rho^2} + u_1\sqrt{1-\rho^2},$$

and

$$y_t - \rho y_{t-1} = \beta_1(x_{1,t} - \rho x_{1,t-1}) + \ldots + \beta_k(x_{k,t} - \rho x_{k,t-1}) + v_t, \qquad t = 2, \ldots, n,$$

respectively. Estimating this model by OLS gives us GLS estimates.

Of course, the parameter ρ is not usually known in practice, so we need to somehow estimate it (and therefore $\mathbf{\Omega}$), and get FGLS estimates. In this case, we need to resort to a procedure known as **iterated FGLS**. The steps involved in this procedure are as follows:

1. Use OLS to estimate (30). Save the residuals, $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\mathbf{b}$.
2. Using the residuals from the previous step, compute the jth estimate of ρ, $\hat{\rho}^{(j)}$, as
$$\hat{\rho}^{(j)} = \frac{\sum_{i=2}^{n} \hat{u}_t \hat{u}_{t-1}}{\sum_{i=2}^{n} \hat{u}_t}.$$
3. Use $\hat{\rho}^{(j)}$ to get the jth FGLS estimate of (30), $\mathbf{b}_{\text{FGLS}}^{(j)}$. Save the updated residuals, $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\mathbf{b}_{\text{FGLS}}^{(j)}$.

Repeat Steps 2-3 until

$$|\hat{\rho}^{(j)} - \hat{\rho}^{(j-1)}| < \delta,$$

and

$$|\mathbf{b}_{\text{FGLS}}^{(j)} - \mathbf{b}_{\text{FGLS}}^{(j-1)}| < \delta,$$

where δ is some small number (say 0.0001).

5.5 Hypothesis Testing

Hypothesis testing can be done using GLS (or FGLS) estimates in very similar manner as it was done using OLS estimates. The only difference is that, in testing the linear restriction imposed by

$$H_0 : \mathbf{R}\beta = \mathbf{r},$$

we use $\mathbf{b}_{\text{GLS}}$ (or $\mathbf{b}_{\text{FGLS}}$) instead of $\mathbf{b}$ as an estimate of β. For now, we will use use $\mathbf{b}_{\text{GLS}}$, but, at the end of this section, we will discuss the effects of using $\mathbf{b}_{\text{FGLS}}$.

Note first, that

$$\begin{aligned} \text{E}(\mathbf{R}\mathbf{b}_{\text{GLS}}) &= \mathbf{R}\text{E}(\mathbf{b}_{\text{GLS}}) \\ &= \mathbf{R}\beta, \end{aligned}$$

and

$$\begin{aligned}
\text{Var}(\mathbf{R}\mathbf{b}_{\text{GLS}}) &= \text{E}([\mathbf{R}\mathbf{b}_{\text{GLS}} - \text{E}(\mathbf{R}\mathbf{b}_{\text{GLS}})][\mathbf{R}\mathbf{b}_{\text{GLS}} - \text{E}(\mathbf{R}\mathbf{b}_{\text{GLS}})]') \\
&= \text{E}[(\mathbf{R}\mathbf{b}_{\text{GLS}} - \mathbf{R}\beta)(\mathbf{R}\mathbf{b}_{\text{GLS}} - \mathbf{R}\beta)'] \\
&= \text{E}[(\mathbf{R}\mathbf{b}_{\text{GLS}} - \mathbf{R}\beta)(\mathbf{b}'_{\text{GLS}}\mathbf{R}' - \beta'\mathbf{R}')] \\
&= \text{E}[\mathbf{R}(\mathbf{b}_{\text{GLS}} - \beta)(\mathbf{b}'_{\text{GLS}} - \beta')\mathbf{R}'] \\
&= \mathbf{R}\text{E}[(\mathbf{b}_{\text{GLS}} - \beta)(\mathbf{b}_{\text{GLS}} - \beta)']\mathbf{R}' \\
&= \mathbf{R}\text{Var}(\mathbf{b}_{\text{GLS}})\mathbf{R}' \\
&= \mathbf{R}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}'.
\end{aligned}$$

Here, if we make the assumption that

$$\mathbf{u} \sim \text{N}(\mathbf{0}, \mathbf{\Omega}),$$

then

$$\mathbf{b}_{\text{GLS}} \sim \text{N}[\beta, (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}],$$

which implies that

$$\mathbf{R}\mathbf{b}_{\text{GLS}} \sim \text{N}[\mathbf{R}\beta, \mathbf{R}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}'],$$

or, alternatively, that

$$\mathbf{R}\mathbf{b}_{\text{GLS}} - \mathbf{R}\beta \sim \text{N}[\mathbf{0}, \mathbf{R}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}'].$$

Note that, under the null hypothesis, $\mathbf{R}\beta = \mathbf{r}$, so we can rewrite the above as

$$\mathbf{R}\mathbf{b}_{\text{GLS}} - \mathbf{r} \sim \text{N}[\mathbf{0}, \mathbf{R}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}'].$$

Finally, we can rewrite this in quadratic form to get the test statistic

$$(\mathbf{R}\mathbf{b}_{\text{GLS}} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b}_{\text{GLS}} - \mathbf{r}) \sim \chi^2_{(q)}.$$

Note that, here, knowledge of $\mathbf{\Omega}$ is needed not only to compute $\mathbf{b}_{\text{GLS}}$, but also to construct the the above test statistic. Of course, since we almost never know $\mathbf{\Omega}$, we need to use our estimate, $\hat{\mathbf{\Omega}}$. Substituting $\hat{\mathbf{\Omega}}$ for $\mathbf{\Omega}$ in the above, we have the test statistic

$$(\mathbf{R}\mathbf{b}_{\text{FGLS}} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b}_{\text{FGLS}} - \mathbf{r}),$$

whose distribution is very hard to ascertain (even if $\mathbf{u}$ is normally distributed). As a result, it is suggested that a bootstrap procedure be used in such tests.

Appendix A

We showed these results in class. These notes will be eventually updated.

References

Davidson, R. and MacKinnon, J. G. (2004). *Econometric theory and methods.* New York, Oxford University Press.