CHAPTER 2. THE CLASSICAL LINEAR REGRESSION MODEL (CLRM)

In Chapter 1, we showed how we estimate an LRM by the method of least squares. As noted in Chapter 1, estimation and hypothesis testing are the twin branches of statistical inference. Based on the OLS, we obtained the sample regression, such as the one shown in Equation (1.40). This is of course a sample regression function (SRF) because it is based on a specific sample drawn randomly from the purported population. What can we say about the true population regression function (PRF) from the SRF? In practice, we do not observe the PRF and have to "guess" it from the SRF. To obtain the best possible guess, we need a framework, which is provided by the **classical linear regression model (CLRM)**. The CLRM is based on several assumptions, which are discussed below.

2.1 Assumptions of the CLRM

We now discuss these assumptions. In Chapters 5 and 6, we will examine these assumptions more critically. However, keep in mind that in any scientific inquiry we start with a set of simplified assumptions and gradually proceed to more complex situations.

Assumption 1: The regression model is *linear in the parameters* as in Equation (1.1); it may or may not be linear in the variables, the *Y*s and *X*s.

Assumption 2: The regressors are assumed fixed, or nonstochastic, in the sense that their values are fixed in repeated sampling. However, if the regressors are stochastic, we assume that each regressor is independent of the error term or at least uncorrelated with it. We will discuss this assumption in more detail in Chapter 6.

Assumption 3: Given the values of the X variables, the expected, or mean, value of the error term u_i is 0.

$$E(u_i \mid \boldsymbol{X}) = 0 \tag{2.1}$$

In matrix notation, we have

$$E(\boldsymbol{u} \mid \boldsymbol{X}) = \mathbf{0} \tag{2.1a}$$

where **0** is the **null vector**.

More explicitly,

$$E\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} E(u_1) \\ E(u_2) \\ E(u_3) \\ \vdots \\ E(u_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Because of this *critical* assumption, and given the values of the regressors, we can write Equation (1.5) as

$$\begin{aligned} E(\boldsymbol{y} \mid \boldsymbol{X}) &= \boldsymbol{BX} + E(\boldsymbol{u} \mid \boldsymbol{X}) \\ &= \boldsymbol{BX} \end{aligned} \tag{2.2}$$

This is the PRF. In regression analysis, our primary objective is to estimate this function. The PRF thus gives the mean value of the regressand corresponding to the given values of the regressors, noting that conditional on these values the mean value of the error term is 0.

Assumption 4: The variance of each u_i, given the values of $\boldsymbol{X}$, is constant or **homoscedastic** (i.e., of equal variance). That is,

$$\text{var}(u_i \mid \boldsymbol{X}) = \sigma^2 \tag{2.3}$$

In matrix notation,

$$\text{var}(\boldsymbol{u} \mid \boldsymbol{X}) = E(\boldsymbol{uu'}) = \sigma^2 \boldsymbol{I} \tag{2.3a}$$

where $\boldsymbol{I}$ is an $n \times n$ identity matrix (see also Assumption 5).

If $\text{var}(u_i \mid \boldsymbol{X}) = \sigma_i^2$, the error variance is said to be heteroscedastic, or of unequal variance. We will discuss this case in Chapter 5.

Figure 2.1 is a picture of both homoscedasticity and heteroscedasticity.

Assumption 5: There is no correlation between error terms belonging to two different observations. That is,

$$\text{cov}(u_i, u_j \mid \boldsymbol{X}) = 0, \quad i \neq j \tag{2.4}$$

where cov stands for covariance, and i and j are two different error terms. Of course, if $i=j$, we get the variance of u_i given in Equation (2.3).

Figure 2.2 shows a likely pattern of autocorrelation.

Figure 2.1 Homoscedasticity and Heteroscedasticity

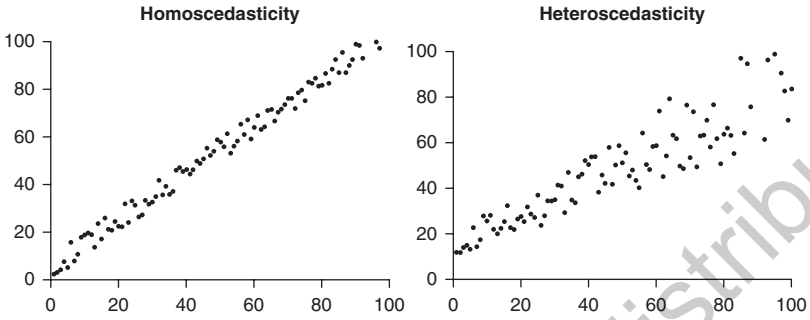
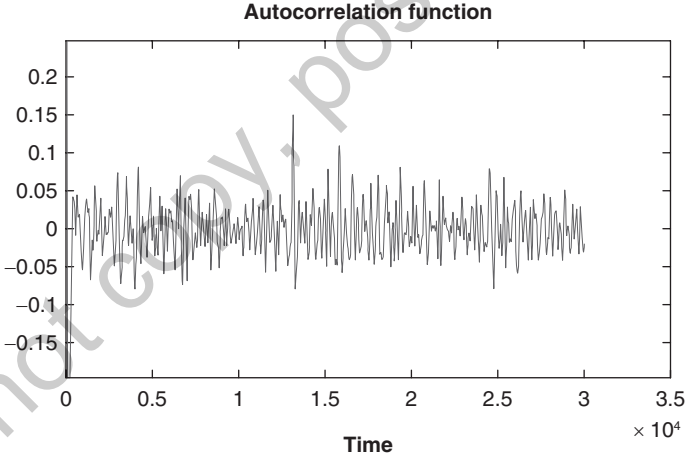


Figure 2.2 Autocorrelation



Assumptions 4 and 5 can be expressed as

$$E(\boldsymbol{uu}') = \sigma^2 \boldsymbol{I} \quad \text{if } i = j$$
$$= \mathbf{0} \quad \text{if } i \neq j$$

where **0** is the **null matrix** and ***I*** is the identity matrix. We discuss this assumption further in Chapter 5. More compactly, we can express Assumptions 4 and 5 as

$$E(\boldsymbol{uu}') = \sigma^2 \boldsymbol{I} = \begin{pmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{pmatrix}$$

Assumption 6: There is *no perfect linear relationship* among the ***X*** variables. This is the assumption of no **multicollinearity**. Strictly speaking, *multicollinearity* refers to the existence of more than one exact linear relationship, and *collinearity* refers to the existence of a single exact linear relationship. But this distinction is rarely maintained in practice, and multicollinearity refers to both cases. Imagine what would happen in the wage regression given in Equation (1.5), if we were to include work experience both in years and in months!

In matrix notation, this assumption means that the ***X*** matrix is of *full column rank*. In other words, *the columns of the* **X** *matrix are linearly independent*. This requires that the number of observations, *n*, is greater than the number of parameters estimated (i.e., the *k* regression coefficients). We discuss this assumption further in Chapter 7.

Assumption 7: The regression model used in the analysis is **correctly specified**, that is, there is no (model) **specific error** or **bias**. In practice, this is a tall assumption, but in Chapter 7, we discuss fully the import of this assumption.

Assumption 8: Although not a part of the original CLRM, for statistical inference (hypothesis testing), we assume that the error term u_i follows the normal distribution with 0 mean and (constant) variance σ^2. Symbolically,

$$u_i \sim N(0, \sigma^2) \tag{2.5}$$

Or in matrix notation,

$$\boldsymbol{u} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \tag{2.5a}$$

The assumption of the normality of the error term is crucial if the sample size is rather small; it is not essential if we have a very large sample. However, we will revisit this assumption in Chapter 7. With this assumption, CLRM is known as the **classical normal linear regression model (CNLRM)**.

Since we are assuming that the $\boldsymbol{X}$ matrix is nonstochastic but $\boldsymbol{u}$ is stochastic, the regressand Y is also stochastic. In addition, since $\boldsymbol{u}$ is normally distributed with 0 mean and constant variance, Y inherits the properties of $\boldsymbol{u}$. More specifically,

$$\boldsymbol{y} \sim N(\boldsymbol{BX}, \sigma^2 \boldsymbol{I}) \tag{2.6}$$

That is, the regressand is distributed normally with mean $\boldsymbol{BX}$ and the (constant) variance σ^2.

Under Assumption 8, we can use the method of **maximum likelihood (ML)** as an alternative to OLS. We will discuss ML more thoroughly in Chapter 3 because of its general applicability in many areas of statistics.

With one or more of the preceding assumptions, in this chapter, we discuss the following topics:

1. The sampling distribution of the OLS estimators, $\boldsymbol{b}$
2. An estimator of the unknown variance, σ^2
3. The relationship between the residual $\boldsymbol{e}$ and the error $\boldsymbol{u}$
4. Small-sample properties of the OLS estimators
5. Large-sample properties of the OLS estimators

2.2 The Sampling or Probability Distributions of the OLS Estimators

Remember that the population parameters in $\boldsymbol{B}$, although unknown, are constants. However, this is not true of the estimated $\boldsymbol{b}$ coefficients, for their values depend on the sample data at hand. In other words, the $\boldsymbol{b}$ coefficients are random. As such, we would like to find their sampling or probability distributions to establish properties of the (OLS) estimators.

Recall that

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \quad \text{using Equation } (1.16)$$

Therefore,

$$\begin{aligned} \boldsymbol{b} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\left[(\boldsymbol{XB} + \boldsymbol{u})\right], \quad \text{using Equation } (1.5) \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{XB} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u} \\ &= \boldsymbol{B} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u} \end{aligned} \tag{2.7}$$

By the definition of covariance, we obtain

$$\begin{aligned}\text{cov}(\boldsymbol{b}) &= E(\boldsymbol{b}-\boldsymbol{B})(\boldsymbol{b}-\boldsymbol{B})' = E\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}\right]\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}\right]', \\ &\quad \text{using Equation } (2.7) \\ &= E\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}\boldsymbol{u}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{u}\boldsymbol{u}')\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \end{aligned} \tag{2.8}$$

In deriving this expression, we have used properties of the transpose of an inverse matrix, and the assumption that $\boldsymbol{X}$ is fixed and that the variance of u_i is constant and the us are uncorrelated. *Notice that we can move the expectations operator through* **X** *because it is assumed fixed.* The variances of the individual elements of $\boldsymbol{b}$ are on the main diagonal (running from upper left to lower right), and the off-diagonal elements give the covariances between pairs of coefficients in $\boldsymbol{b}$.

Since

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \quad (1.16)$$

and the $\boldsymbol{X}$ matrix is fixed, $\boldsymbol{b}$ is a linear function of $\boldsymbol{y}$. Using Assumption 8, we know that $\boldsymbol{y}$ is normally distributed. It is a property of the normal distribution that any linear function of a normally distributed variable is also normally distributed. Therefore, $\boldsymbol{b}$ is ipso facto normally distributed as follows:

$$\boldsymbol{b} \sim N(\boldsymbol{B}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}) \tag{2.9}$$

That is, $\boldsymbol{b}$ is normally distributed with $\boldsymbol{B}$ as its mean (see Equation 2.20) and the variance established in Equation (2.8). In other words, under the normality assumption, the **sampling distribution** of the OLS estimator is normal, as shown in Equation (2.9). This finding will aid us in testing hypotheses about any element of $\boldsymbol{B}$ or any linear combination thereof. It may be noted that a sampling distribution is a probability distribution of an estimator or of any test statistic. In other words, it describes the variation in the values of a statistic over all possible samples, here the variation in $\boldsymbol{b}$ over all possible samples.[1]

[1]Suppose we draw several independent samples and for each sample we compute a (test) statistic, such as the mean, and draw a frequency distribution of all these statistics. Roughly speaking, this frequency distribution is the sampling distribution of that statistic. In our case, under the assumed conditions, the probability or sampling distribution of any component of $\boldsymbol{b}$ is normal as shown in Equation (2.9).

For any single element of $\boldsymbol{b}$, b_k, we can express Equation (2.9) as

$$b_k \sim N(B_k, \sigma^2 x^{kk}) \tag{2.9a}$$

where x^{kk} is the kth diagonal element of $(\boldsymbol{X'X})^{-1}$. The square root of $\sigma^2 x^{kk}$ will give the standard error of b_k (see Figure 2.3).

However, before we can engage in hypothesis testing, we need to estimate the unknown σ^2. Remember that σ^2 refers to the variance of the error term $\boldsymbol{u}$. Since we do not observe $\boldsymbol{u}$ directly, we have to rely on the estimated residuals, $\boldsymbol{e}$, to learn about the true variance. Toward that end, we need to establish the relationship between $\boldsymbol{u}$ and $\boldsymbol{e}$. Recall that

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} \tag{2.10}$$

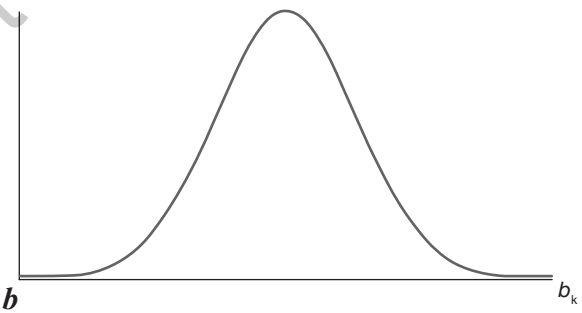
Substituting for $\hat{y}$ from (1.23), we obtain

$$\begin{aligned}
\boldsymbol{e} &= \boldsymbol{y} - \boldsymbol{Xb} \\
&= \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'y}, \text{ substituting for } \boldsymbol{b} \text{ from Eq. (1.16)} \\
&= \boldsymbol{My} \\
&= \boldsymbol{M}(\boldsymbol{XB} + \boldsymbol{u})
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{M} &= [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}] \\
&= \boldsymbol{Mu}, \text{ because } \boldsymbol{MXB} = \boldsymbol{0}.^{2}
\end{aligned} \tag{2.11}$$

Figure 2.3 The Distribution of b_k, a Component of the Vector $\boldsymbol{b}$



[2] $\boldsymbol{MXB} = [\boldsymbol{XB} - \boldsymbol{X}(\boldsymbol{XX'})^{-1}\boldsymbol{X'XB}] = \boldsymbol{XB} - \boldsymbol{IXB} = \boldsymbol{0}$, where $\boldsymbol{I}$ is the identity matrix.

As noted in Chapter 1, $\boldsymbol{M}$ is a very important matrix in the analysis of LRMs. It is an **idempotent matrix**, a square matrix with the property that $\boldsymbol{M}=\boldsymbol{M}^2$. For further properties of the idempotent matrices, see Appendix A on linear algebra.

Since $\boldsymbol{M}$ is constant because it is a function of (fixed) $\boldsymbol{X}$, we can write

$$\begin{aligned} E(\boldsymbol{e}) &= E(\boldsymbol{Mu}) \\ &= \boldsymbol{M}E(\boldsymbol{u}) \\ &= \boldsymbol{0} \end{aligned} \tag{2.12}$$

because $E(\boldsymbol{u})=\boldsymbol{0}$, by Assumption 1. We have thus shown that the expectation of each element of $\boldsymbol{e}$ is 0.

Now,

$$\begin{aligned} \text{cov}(\boldsymbol{e}) = E(\boldsymbol{ee'}) &= E(\boldsymbol{Muu'M'}) \\ &= \boldsymbol{M}E(\boldsymbol{uu'})\boldsymbol{M'} \\ &= \sigma^2\boldsymbol{IM} = \sigma^2\boldsymbol{M} \end{aligned} \tag{2.13}$$

recalling the properties of $\boldsymbol{M}$. This equation gives the covariance matrix of $\boldsymbol{e}$.

Since $\boldsymbol{e}$ is a linear function of $\boldsymbol{u}$ and since $\boldsymbol{u}$ is normally distributed by Assumption 8, we have

$$\boldsymbol{e} \sim N(0,\ \sigma^2\boldsymbol{M}) \tag{2.14}$$

Therefore, like the mean of $\boldsymbol{u}$, the mean of $\boldsymbol{e}$ is 0, but unlike $\boldsymbol{u}$, the residuals are heteroscedastic as well as autocorrelated.[3]

What Equations (2.13) and (2.14) show is that the residuals $e_1, e_2, \ldots, e_n$ have zero mean values, generally have different variances, and have nonzero covariances. Remember that in the (population) CLRM errors, $u_1, u_2, \ldots, u_n$ have zero expectations, have constant variance, and are not autocorrelated (by assumption). In other words, the properties that hold for $\boldsymbol{u}$ generally do not hold for $\boldsymbol{e}$, except for zero expectations.

[3]Actually, the distribution of $\boldsymbol{e}$ is degenerate as its variance–covariance matrix is singular. On this, see Vogelvang, B. (2005). *Econometrics: Theory and applications with Eviews* (chapter 4). Harlow, England: Pearson-Addison Wesley.

Although we have assumed that the variance of $\boldsymbol{u}$ (not of $\boldsymbol{e}$) is constant, equal to σ^2, we are yet to estimate it from the sample data. Toward that end, we proceed as follows.

Even though we do not observe $\boldsymbol{u}$, we observe $\boldsymbol{e}$ (after the regression is estimated). Naturally, we will have to estimate the unknown variance from the estimated $\boldsymbol{e}$. From Equation (2.11), we know that

$$\boldsymbol{e} = \boldsymbol{M}\boldsymbol{u} \qquad (2.11)$$

Therefore,

$$\begin{aligned} E(\boldsymbol{e}'\boldsymbol{e}) &= E(\boldsymbol{u}'\boldsymbol{M}'\boldsymbol{M}\boldsymbol{u}) \\ &= E(\boldsymbol{u}'\boldsymbol{M}\boldsymbol{u}) \end{aligned} \qquad (2.15)$$

because of the properties of $\boldsymbol{M}$. Now,

$$\begin{aligned} E(\boldsymbol{e}'\boldsymbol{e}) &= E(\boldsymbol{u}'\boldsymbol{M}\boldsymbol{u}) \\ &= E\,[\mathrm{tr}(\boldsymbol{u}'\boldsymbol{M}\boldsymbol{u})], \text{ since } \boldsymbol{u}'\boldsymbol{M}\boldsymbol{u} \text{ is a scalar} \\ &= E\,[\mathrm{tr}(\boldsymbol{M}\boldsymbol{u}\boldsymbol{u}')] \text{ changing the order of multiplication inside the trace} \\ &= \mathrm{tr}[\boldsymbol{M}(E\boldsymbol{u}\boldsymbol{u}')], \text{ since the trace and expectations operators are both linear} \\ &= \mathrm{tr}[\boldsymbol{M}(\sigma^2\boldsymbol{I})] \\ &= \sigma^2\,\mathrm{tr}(\boldsymbol{M}) \\ &= \sigma^2\left[\mathrm{tr}(\boldsymbol{I}) - \mathrm{tr}(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\right], \text{ using the definition of } \boldsymbol{M} \\ &= \sigma^2\left[\mathrm{tr}(\boldsymbol{I}_n) - \mathrm{tr}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X})\right] \\ &= \sigma^2\left[\mathrm{tr}(\boldsymbol{I}_n) - \mathrm{tr}(\boldsymbol{I}_k)\right], \text{ since} (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_k \\ &= \sigma^2\,(n-k), \text{ since } \mathrm{tr}(\boldsymbol{I}) = n \text{ and } \mathrm{tr}(\boldsymbol{I}_k) = k \end{aligned} \qquad (2.16)$$

The notation $\mathrm{tr}(\boldsymbol{M})$ means the trace of the matrix $\boldsymbol{M}$, which is simply the sum of the entries of the main diagonal of $\boldsymbol{M}$. In deriving the steps in Equation (2.16), we have made use of several properties of the *trace* of a matrix, such as the fact that trace is a linear operator and if $\boldsymbol{AB}$ and $\boldsymbol{BA}$ are both square matrices, then $\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA})$.

As a result, we can now write

$$E\left(\frac{\boldsymbol{e}'\boldsymbol{e}}{n-k}\right)=\sigma^2 \tag{2.17}$$

If we now define

$$S^2=\frac{\boldsymbol{e}'\boldsymbol{e}}{n-k}=\frac{\Sigma e_i^2}{n-k} \tag{2.18}$$

then

$$E(S^2)=\sigma^2 \tag{2.19}$$

In words, S^2 is an unbiased estimator of the true error variance σ^2. S, the square root of S^2, is called the **standard error (*se*) of the estimate** or the **standard error of the regression**. In practice, therefore, we use S^2 in place of σ^2.

2.3 Properties of OLS Estimators: The Gauss–Markov Theorem[4]

The OLS estimators possess some ideal or optimum properties, which are contained in the well-known **Gauss–Markov theorem**:[5] Given the assumptions of the classical regression model, in the class of unbiased linear estimators, the least-squares estimators have minimum variance; that is, they are **best linear unbiased estimators**, **BLUE** for short. In other words, no other linear, unbiased estimator of $\boldsymbol{B}$ can have a smaller variance than the OLS estimator given in Equation (2.8).

To establish this theorem, first note that $\boldsymbol{b}$, the OLS estimator of $\boldsymbol{B}$, is a linear function of the regressand $\boldsymbol{y}$, as we have established in Chapter 1 (see Equation 1.16).[6] To prove that $\boldsymbol{b}$ is unbiased, we proceed as follows:

$$\begin{aligned}\boldsymbol{b}&=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \quad (1.16)\\ &=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\left[\boldsymbol{X}\boldsymbol{B}+\boldsymbol{u}\right], \text{ substituting for } \boldsymbol{y}\\ &=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B}+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}\\ &=\boldsymbol{B}+(\boldsymbol{X}'\boldsymbol{X})^{1}\boldsymbol{X}'\boldsymbol{u}\end{aligned}$$

[4]In Appendix C, we discuss both small-sample and large-sample properties of OLS and ML estimators.

[5]Although known as the *Gauss–Markov theorem*, the least-squares approach of Gauss antedates (1821) the minimum-variance approach of Markov (1900).

[6]See the discussion in Darnell, A. C. (1994). *A dictionary of econometrics* (p. 155). Cheltenham, England: Edward Elgar.

Now

$$\begin{aligned} E(\boldsymbol{b}) &= \boldsymbol{B} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{u}) \\ &= \boldsymbol{B} \end{aligned} \tag{2.20}$$

In words, the expected value of $\boldsymbol{b}$ is equal to $\boldsymbol{B}$, thus proving that $\boldsymbol{b}$ is unbiased. (Recall the definition of unbiased estimator.) Note that $E(\boldsymbol{u}|\boldsymbol{X})=\boldsymbol{0}$ by assumption.

To prove that in the class of unbiased linear estimators the least-squares estimators have the least variance (i.e., they are efficient), we proceed as follows:

Let $\boldsymbol{b}^*$ be another linear estimator of $\boldsymbol{B}$ such that

$$\boldsymbol{b}^* = [\boldsymbol{A} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{y} \tag{2.21}$$

where $\boldsymbol{A}$ is some nonstochastic $k \times n$ matrix, similar to $\boldsymbol{X}$. Simplifying, we obtain

$$\begin{aligned} \boldsymbol{b}^* &= \boldsymbol{A}\boldsymbol{y} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= \boldsymbol{A}\boldsymbol{y} + \boldsymbol{b} \end{aligned} \tag{2.22}$$

where $\boldsymbol{b}$ is the least-squares estimator given in Equation (1.16).

Now

$$\begin{aligned} E(\boldsymbol{b}^*) &= \left[\boldsymbol{A} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]E(\boldsymbol{y}) \\ &= \left[\boldsymbol{A} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right](\boldsymbol{X}\boldsymbol{B}) \\ &= (\boldsymbol{A}\boldsymbol{X} + \boldsymbol{I})\boldsymbol{B} \end{aligned} \tag{2.23}$$

Now $\boldsymbol{E}(\boldsymbol{b}^*)=\boldsymbol{B}$ if and only if $\boldsymbol{A}\boldsymbol{X}=\boldsymbol{0}$. In other words, for the linear estimator $\boldsymbol{b}^*$ to be unbiased, $\boldsymbol{A}\boldsymbol{X}$ must be $\boldsymbol{0}$.

Thus,

$$\begin{aligned} \boldsymbol{b}^* &= \left[\boldsymbol{A} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right][\boldsymbol{X}\boldsymbol{B} + \boldsymbol{u}], \quad \text{substituting for } (\boldsymbol{y}) \\ &= \boldsymbol{B} + \left[\boldsymbol{A} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]\boldsymbol{u}, \quad \text{because } \boldsymbol{A}\boldsymbol{X} = \boldsymbol{0} \end{aligned}$$

Given that $\boldsymbol{u}$ has zero mean and constant variance ($= \sigma^2\boldsymbol{I}$), we can now find the variance of $\boldsymbol{b}^*$ as follows:

$$\begin{aligned} \text{cov}(\boldsymbol{b}^*) &= E\left[\boldsymbol{A} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]\boldsymbol{u}\boldsymbol{u}'\left[\boldsymbol{A} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]' \\ &= \left[\boldsymbol{A} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]E(\boldsymbol{u}\boldsymbol{u}')\left[\boldsymbol{A} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]' \\ &= \sigma^2\left[\boldsymbol{A}\boldsymbol{A}' + (\boldsymbol{X}'\boldsymbol{X})^{-1}\right] \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{A}\boldsymbol{A}'\sigma^2 \\ &= \text{var}(\boldsymbol{b}) + \boldsymbol{A}\boldsymbol{A}'\sigma^2 \end{aligned} \tag{2.24}$$

Since $\boldsymbol{AA'}$ is a positive semidefinite matrix, Equation (2.24) shows that the covariance matrix of $\boldsymbol{b}^*$ is equal to the covariance matrix of $\boldsymbol{b}$ plus a positive semidefinite matrix. That is, $\text{cov}(\boldsymbol{b}^*) > \text{cov}(\boldsymbol{b})$, unless $\boldsymbol{A} = \boldsymbol{0}$. This shows that in the class of unbiased linear estimators, the least-square estimator $\boldsymbol{b}$ has the least variance, that is, it is efficient compared with any other linear unbiased estimator of $\boldsymbol{B}$.

It is important to note that in establishing the Gauss–Markov theorem we do not have to assume that the error term $\boldsymbol{u}$ follows a particular probability distribution, such as the normal. To establish the theorem, we only need Assumptions 1 to 5.

It is also important to note that if one or more assumptions underlying the Gauss–Markov theorem are not satisfied, the OLS estimators will not be BLUE. Also, bear in mind that the Gauss–Markov theorem holds only for linear estimators, that is, linear functions of the observation vector $\boldsymbol{y}$. There are situations where nonlinear (in-the-parameter) estimators are more efficient than the linear estimators. In this book, we do not deal with nonlinear estimators, for that requires a separate book.[7]

To sum up, we have shown that under the Gauss–Markov assumptions, $\boldsymbol{b}$, the least-square estimator of $\boldsymbol{B}$, is BLUE, that is, in the class of unbiased linear estimators, $\boldsymbol{b}$ has the least variance. We also showed how to estimate $\boldsymbol{B}$ and the variance of the estimated $\boldsymbol{B}$.

2.4 Estimating Linear Functions of the OLS Parameters

We have shown how to estimate $\boldsymbol{B}$, that is, each of its elements. Suppose we want to estimate some linear function of the elements of $\boldsymbol{B}$, that is, that of $B_1, B_2, B_3, \ldots, B_k$. More specifically, suppose we want to estimate $\boldsymbol{t'B}$, where $\boldsymbol{t'}$ is a $1 \times k$ vector of real numbers and $\boldsymbol{B}$ is a $k \times 1$ vector of the parameters in $\boldsymbol{B}$. It can be shown that the BLUE of $\boldsymbol{t'B}$ is $\boldsymbol{t'b}$, where $\boldsymbol{b}$ is the least-square estimator of $\boldsymbol{B}$ (see also Appendix C).

What this means is that whether we estimate all the elements of $\boldsymbol{B}$, or one of its elements, or estimate a linear combination ($\boldsymbol{t'B}$), we can use the OLS regression.

Let $\lambda = \boldsymbol{t'B}$. By choosing $\boldsymbol{t}$ appropriately, we can make λ equal to any element of $\boldsymbol{B}$, or to the sum of the elements of $\boldsymbol{B}$ that might be of interest to researchers. Suppose in Equation (1.2), we want the coefficient of $\boldsymbol{B}_1$ equal to 4, and the coefficient of $\boldsymbol{B}_5$ equal to −1, and the rest of the coefficients to be all zeros. In other words, we want $\lambda = 4\boldsymbol{B}_1 - \boldsymbol{B}_5$. Here, $\boldsymbol{t'} = (4, 0, 0, 0, -1, 0, 0, 0, \ldots)'$.

Using the definition of variance, we can now find the variance of the estimated λ ($= \hat{\lambda}$), which is

[7]For examples of nonlinear estimators and their applications, see Gujarati, D. (2015). *Econometrics by example* (2nd ed.). London, England: Palgrave Macmillan.

$$\text{var}(\hat{\lambda}) = \boldsymbol{t}'(\text{var}(\boldsymbol{b}))\boldsymbol{t} = \sigma^2 \boldsymbol{t}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{t} \tag{2.25}$$

But keep in mind that in general the variance of $\hat{\lambda}$ depends on every element of the covariance matrix of $\boldsymbol{b}$, the estimator of $\boldsymbol{B}$. However, if some elements of the vector $\boldsymbol{t}$ are equal to zero, var($\hat{\lambda}$) does not depend on the corresponding rows and columns of the covariance matrix $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

As an example, consider $\lambda = 4B_1 - B_5$. In this case,

$$\begin{aligned}\text{var}(\hat{\lambda}) &= t_1^2 \text{ var}(b_1) + t_5^2 \text{ var}(b_5) + 2t_1 t_5 \text{ cov}(b_1, b_5) \\ &= 16 \text{ var}(b_1) + \text{var}(b_5) - 8 \text{cov}(b_1, b_5)\end{aligned} \tag{2.26}$$

Notice in this example only the variances of b_1 and b_5 and their covariances are involved, as the values of the other parameters in the k-variable regression (1.2) are assumed to be zero. But if there are more nonzero coefficients, the variances and their pairwise covariances will also be involved in computing the variance of the linear combination $\boldsymbol{t}'\boldsymbol{B}$.

2.5 Large-Sample Properties of OLS Estimators

2.5.1 Consistency of OLS Estimators

We have shown that the OLS estimators of the CLRM are unbiased, which is a small, or finite sample, property. We can also show that the OLS estimators are consistent, that is, they converge to their true values as the sample size increases indefinitely. Convergence is a large-sample property.

Proof: A sufficient condition for an unbiased estimator to be consistent is for its variance to converge to zero as the sample size n increases indefinitely. For the OLS estimator $\boldsymbol{b}$, we have already shown that its variance is

$$\text{cov}(\boldsymbol{b}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{2.8}$$

which we can write as

$$\text{cov}(\boldsymbol{b}) = \frac{\sigma^2}{n}(n^{-1}\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{2.27}$$

To see the behavior of this expression as $n \to \infty$, we have

$$\begin{aligned}\underset{n\to\infty}{p\lim}\, \text{cov}(\boldsymbol{b}) &= \underset{n\to\infty}{p\lim}\left[\frac{\sigma^2}{n}(n^{-1}\boldsymbol{X}'\boldsymbol{X})^{-1}\right] \\ &= \underset{n\to\infty}{p\lim}\frac{\sigma^2}{n}\lim_{n\to\infty}(n^{-1}\boldsymbol{X}'\boldsymbol{X})^{-1}\end{aligned} \tag{2.28}$$

where *p*lim is probability limit (see Appendix C for details).

We have assumed that the elements of the matrix $\boldsymbol{X}$ are bounded, which means the second term in the preceding equation is bounded for all n. Therefore, the second term above can be replaced by a matrix of finite constants. Now, the limit of the first term in Equation (2.28) tends to zero as n increases indefinitely. As a result,

$$\underset{n\to\infty}{p\lim}\operatorname{cov}(\boldsymbol{b})=0 \tag{2.29}$$

which establishes that $\boldsymbol{b}$ is a consistent estimator of $\boldsymbol{B}$. In establishing this result, we have used some of the properties of the *p*lim operator.

2.5.2 Consistency of the OLS Estimator of the Error Variance

We have proved that S^2 is an unbiased estimator of σ^2. Assuming values of u_i are independent and identically distributed (iid), we can prove that S^2 is also a consistent estimator of σ^2. The proof is as follows:

$$\begin{aligned} S^2 &= \frac{(\boldsymbol{y}-\boldsymbol{Xb})'(\boldsymbol{y}-\boldsymbol{Xb})}{n-k} \\ &= \frac{\boldsymbol{u}'(\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{u}}{n-k} \\ &= \left(\frac{n}{n-k}\right)\left(\frac{\boldsymbol{u}'\boldsymbol{u}}{n}-\frac{\boldsymbol{u}'\boldsymbol{X}}{n}\cdot\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1}\cdot\frac{\boldsymbol{X}'\boldsymbol{u}}{n}\right) \end{aligned} \tag{2.30}$$

Note: $\boldsymbol{e}=\boldsymbol{My}=\boldsymbol{Mu}$, where $\boldsymbol{M}=[\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{X}']$. Also note how the entries are manipulated by multiplying or dividing them by the sample size or the adjusted sample size without affecting the basic relationships.

Taking the *p*lim of both sides of Equation (2.30), we obtain

$$\begin{aligned} p\lim S^2 &= p\lim\left(\frac{n}{n-k}\right)\left(p\lim\frac{\boldsymbol{u}'\boldsymbol{u}}{n}-p\lim\frac{\boldsymbol{u}'\boldsymbol{X}}{n}\cdot p\lim\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1}\cdot p\lim\frac{\boldsymbol{X}'\boldsymbol{u}}{n}\right) \\ &= 1(\sigma^2-0\cdot Q^{-1}\cdot 0), \quad \text{where } p\lim\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1}=Q^{-1} \\ &= \sigma^2 \end{aligned} \tag{2.31}$$

which establishes the result. Note that for large n, $(n-k)\approx n$.

In deriving the preceding result, we have used **Khinchine's theorem** (see Appendix B) as well as the properties of the *p*lim.

2.5.3 Independence of the OLS Estimators and the Residual Term, e

What this says is that each element of $\boldsymbol{b}$ is uncorrelated with each element of the least-squares residual vector $\boldsymbol{e}$. The proof is as follows:

Recall that

$$\boldsymbol{b} = \boldsymbol{B} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u} \quad (2.7)$$
$$\boldsymbol{e} = \boldsymbol{M}\boldsymbol{u} \quad (2.11)$$

are both linear functions of the error term $\boldsymbol{u}$.

Now the covariance of $\boldsymbol{b}$ and $\boldsymbol{e}$ is

$$\begin{aligned}\text{cov}(\boldsymbol{b},\boldsymbol{e}) &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\text{var}(\boldsymbol{u})\boldsymbol{M}' \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{M}' \\ &= 0, \quad \text{since } \boldsymbol{M}\boldsymbol{X} = \boldsymbol{0} \leftrightarrow \boldsymbol{X}'\boldsymbol{M}'\end{aligned} \quad (2.32)$$

This shows that $\boldsymbol{b}$ and $\boldsymbol{e}$ are uncorrelated.

It may be noted that if we assume that $\boldsymbol{u}$ is normally distributed, $\boldsymbol{b}$ and $\boldsymbol{e}$ are not only uncorrelated but also independent. In Chapter 3, we will consider the **normal linear regression model**, which explicitly assumes that the error term $\boldsymbol{u}$ is normally distributed and will see the consequence of the normality assumption.

2.5.4 Large-Sample Distribution of b*: Asymptotic Normality of the OLS Estimators*

It can be shown that[8]

$$\boldsymbol{b}\,\text{asy} \sim \boldsymbol{N}(\boldsymbol{B}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}) \quad (2.33)$$

where asy means asymptotically (i.e., $n \to \infty$).

In other words, as the sample size n increases indefinitely, $\boldsymbol{b}$ is approximately normally distributed with mean equal to $\boldsymbol{B}$ and variance equal to

[8]The proof is rather complicated and can be found in Theil, H. (1971). *Principles of econometrics* (pp. 380–381). New York, NY: Wiley; see also Mittlehammer, R. C. (1996). *Mathematical statistics for economics and business* (pp. 443–447). New York, NY: Springer.

$\sigma^2(\boldsymbol{X'X})^{-1}$. Each element of $\boldsymbol{b}$ is individually normally distributed with variance equal to the appropriate element of the variance matrix $\sigma^2(\boldsymbol{X'X})^{-1}$. *This result holds whether* u *is normally distributed or not*.

It may be noted that if the errors u_i are not iid, even then $\boldsymbol{b}$ is asymptotically normally distributed as in (2.33) under certain conditions.[9]

2.5.5 Asymptotic Normality of S^2

If in addition to the classical assumptions, it is assumed that values of u_i are iid and have bounded fourth-order moments about the origin, $\boldsymbol{S}^2$ is asymptotically normally distributed. These results also hold true even if the u_i values are not iid.[10]

2.6 Summary

The CLRM, $\boldsymbol{y}=\boldsymbol{XB}+\boldsymbol{u}$, is the foundation of regression analysis. It is based on several assumptions. The basic assumptions are that (1) the data matrix $\boldsymbol{X}$ is nonstochastic, (2) it is of full column rank, (3) the expected value of the error term is zero, and (4) the covariance matrix of the error term $E(\boldsymbol{uu'})=\sigma^2\boldsymbol{I}$. This means the error variance is constant and equal to σ^2 and that the error terms are mutually uncorrelated.

We used the method of OLS to estimate the parameters of an LRM. One reason for using OLS is that it does not require us to make assumptions about the probability distribution of the error term, and it is comparatively easy to estimate. Parameters of the CLRM estimated by OLS are called OLS estimators. OLS estimators have several desirable statistical properties such as (1) they are unbiased and (2) among all linear unbiased estimators of $\boldsymbol{B}$, they have minimum variances. This is called the Gauss–Markov theorem. These are small-sample properties.

OLS estimators have these asymptotic, or large-sample, properties: (1) The OLS estimators of $\boldsymbol{B}$ as well as the estimator of the error variance are consistent estimators and (2) the OLS estimators asymptotically follow the normal distribution.

Exercises

2.1 Consider the bivariate regression: $Y_i = B_1 + B_2 X_i + u_i$. Under the classical linear regression assumptions, show that

[9]See Mittlehammer, R. C. (1996). *Mathematical statistics for economics and business* (p. 445). New York, NY: Springer.

[10]See Mittlehammer, R. C. (1996). *Mathematical statistics for economics and business* (pp. 448–449). New York, NY: Springer.

a. $\text{cov}(b_1, b_2) = -\bar{X}\dfrac{\sigma^2}{\Sigma(X_i - \bar{X})^2}$

b. $\text{cov}(\bar{Y}, b_2) = 0$

2.2 Show that for the model in Exercise 2.1,

$$\text{RSS} = \frac{\Sigma x_i^2 \Sigma y_i^2 - (\Sigma x_i y_i)^2}{\Sigma x_i^2}$$

where RSS is the residual sum of squares and

$$x_i = (X_i - \bar{X}); \quad y_i = (Y_i - \bar{Y}); \quad x_i y_i = (X_i - \bar{X})(Y_i - \bar{Y})$$

2.3 Verify the following properties of OLS estimators:

a. The OLS regression line (plane) passes through the sample means of the regressand and the regressors.

b. The mean values of the actual Y and the estimated $Y(=\hat{Y})$ are the same.

c. In the CLRM with intercept, the mean value of the residuals $(\bar{e})$ is zero.

d. As a result of the preceding property, the k-variable sample CLRM can be expressed as

$$y_i = b_2 x_{2i} + b_3 x_{3i} + \cdots + b_k x_{ki} + e_i$$

$$\text{where } y_i = (Y_i - \bar{Y}); x_{ki} = (X_{ki} - \bar{X}_k)$$

2.4 Consider the following bivariate regression model:

$$Y_i^* = B_1^* + B_2^* X_i^* + u_i$$

where

$$Y_i^* = \frac{Y_i - \bar{Y}}{s_Y}; \quad X_i^* = \frac{X_i - \bar{X}}{s_X}$$

where s_Y and s_X are the sample standard deviations of Y and X. Y_i^* and X_i^* are known as **standardized variables**, often known as ***Z*** scores. Since the units of measurement of the Z scores in the numerator and the denominator are the same, they are called "pure" or "unitless" numbers.

a. Show that a standardized variable has a zero mean and unit variance.

b. What are the formulas to estimate B_1^* and B_2^*?

c. What is the relationship between B_1^* and B_1 and between B_2^* and B_2?

2.5 The sample correlation coefficient between variables Y and X, r_{XY}, is defined as

$$r_{XY} = \frac{\Sigma x_i y_i}{\sqrt{\Sigma x_i^2 \Sigma y_i^2}}$$

where

$$x_i = (X_i - \bar{X}); \quad y_i = (Y_i - \bar{Y})$$

If we standardize variables as in Exercise 2.4, does it affect the correlation coefficient between X and Y? Show the necessary calculations.

2.6 Consider variables X_1, X_2, and X_3. Now consider the following correlation coefficients:

$$r_{12} = \text{correlation coefficient between } X_1 \text{ and } X_2$$
$$r_{13} = \text{correlation coefficient between } X_1 \text{ and } X_3$$
$$r_{23} = \text{correlation coefficient between } X_2 \text{ and } X_3$$

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

$r_{12.3}$ is called the **partial correlation coefficient** between X_1 and X_2 holding the influence of the variable X_3. The concept of partial correlation is akin to the concept of a partial regression coefficient.

a. What other partial correlation coefficients can you compute?

b. If we standardize the three variables as in Exercise 2.4, would the correlation coefficients among the standardized variables be different from the unstandardized variables?

c. Would partial correlation coefficients be affected by standardizing the variables? Explain.

2.7 Consider the following LRM:

$$Y_i = B_1 + B_2 X_{2i} + B_3 X_{3i} + B_4 X_{4i} + B_5 X_{5i} + u_i$$

How would you test the following hypotheses?

a. $B_2 = B_3 = B_4 = B_5 = B$, that is, all partial regression coefficients are the same.

b. $B_2 = B_3$ and $B_4 = B_5$

c. $B_2 + B_3 = 2B_4$

2.8 Remember that the hat matrix, $\boldsymbol{H}$, is expressed as

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}$$

Show that the residual vector $\boldsymbol{e}$ can also be expressed as

$$\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{y}$$

2.9 Prove that the matrices $\boldsymbol{H}$ and $(\boldsymbol{I} - \boldsymbol{H})$ are idempotent.

2.10* For the following matrix, compute its eigenvalues:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

(*Optional)

2.11 Consider the following regression model (see Chapter 7, Equation (7.30)):

$$Y_i = B_1 + B_2 X_i + B_3 X_i^2 + u_i$$

Models like this are called polynomial regression models, here a second-degree polynomial.

a. Is this an LRM?

b. Can OLS be used to estimate the parameters of this model?

c. Since X_i^2 is the square of X_i, does this model suffer from perfect collinearity?

2.12 Consider the following model:

$$Y_i = B_1 + B_2 X_{2i} + B_3 X_{3i} + B_4 X_{4i} + u_i$$

You are told that $B_2 = 1$.

a. In this case, is it legitimate to estimate the following regression?

$$(Y_i - X_{2i}) = B_1 + B_3 X_{3i} + B_4 X_{4i} + u_i$$

This model is called a **restricted** linear regression, whereas the preceding model is called an **unrestricted** linear regression (see Chapter 4, Appendix 4A for further details).

b. How would you estimate the restricted regression, taking into account the restriction that $B_2 = 1$?