# Lecture 2: Simple Linear Regression

R.G. Pierse

## 1 The Classical Linear Regression Model

$$Y_i = a + bX_i + u_i \quad , \quad i = 1, \cdots, n \tag{1.1}$$

### 1.1 Assumptions of the Model

$$E(u_i) = 0 \quad , \quad i = 1, \cdots, n \tag{A1}$$

$$E(u_i^2) = \sigma^2 \quad , \quad i = 1, \cdots, n \tag{A2}$$

$$E(u_i u_j) = 0 \quad , \quad i, j = 1, \cdots, n \quad j \neq i \tag{A3}$$

X values are *fixed in repeated sampling* (A4)

$Y_i$ is a random variable with the following properties:

$$E(Y_i) = a + bX_i \quad , \quad i = 1, \cdots, n$$

$$\begin{aligned} Var(Y_i) &= E(Y_i - E(Y_i))^2 \\ &= E(u_i^2) = \sigma^2 \quad , \quad i = 1, \cdots, n \end{aligned}$$

## 2 The Ordinary Least Squares Estimator

$$Y_i = \widehat{a} + \widehat{b}X_i + e_i \quad , \quad i = 1, \cdots, n \tag{2.1}$$

The OLS estimator is the estimator that minimises the sum of squared residuals $s = \sum_{i=1}^n e_i^2$.

$$\min_{\widehat{a},\widehat{b}} \; s = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n \left( Y_i - \widehat{a} - \widehat{b}X_i \right)^2$$

$$s = \sum_{i=1}^{n} Y_i^2 - 2\widehat{a}\sum_{i=1}^{n} Y_i - 2\widehat{b}\sum_{i=1}^{n} X_i Y_i + \sum_{i=1}^{n} \widehat{a}^2 + 2\widehat{a}\widehat{b}\sum_{i=1}^{n} X_i + \widehat{b}^2 \sum_{i=1}^{n} X_i^2$$

Differentiating this expression with respect to $\widehat{a}$ and $\widehat{b}$ gives the two first order conditions:

$$\frac{\partial s}{\partial \widehat{a}} = -2\sum_{i=1}^{n} Y_i + 2n\widehat{a} + 2\widehat{b}\sum_{i=1}^{n} X_i = 0 \tag{2.2}$$

and

$$\frac{\partial s}{\partial \widehat{b}} = -2\sum_{i=1}^{n} X_i Y_i + 2\widehat{a}\sum_{i=1}^{n} X_i + 2\widehat{b}\sum_{i=1}^{n} X_i^2 = 0 \tag{2.3}$$

Rearranging (2.2) gives

$$\widehat{a} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \widehat{b}\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{Y} - \widehat{b}\,\overline{X} \tag{2.4}$$

where $\overline{Y} = \frac{1}{n}\sum_i^n Y_i$ and $\overline{X} = \frac{1}{n}\sum_i^n X_i$ are the sample means of $Y$ and $X$ respectively.

Substituting this expression into (2.3) and cancelling the factor of 2 gives

$$\sum_{i=1}^{n} X_i Y_i - n\overline{Y}\ \overline{X} - \widehat{b}n\overline{X}^2 - \widehat{b}\sum_{i=1}^{n} X_i^2 = 0$$

which can be rearranged as:

$$\widehat{b} = \frac{\sum_{i=1}^n X_i Y_i - n\overline{Y}\ \overline{X}}{\sum_{i=1}^n X_i^2 - n\overline{X}^2}$$

or

$$\widehat{b} = \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}. \tag{2.5}$$

Alternatively, using lower case characters to denote deviations from sample means, we can write

$$\widehat{b} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}. \tag{2.6}$$

# 3 Properties of the OLS Estimator

Summing (1.1) and dividing by $n$ gives

$$\overline{Y} = a + b\overline{X} + \overline{u}$$

and, subtracting from (1.1),

$$Y_i - \overline{Y} = b(X_i - \overline{X}) + (u_i - \overline{u})$$

or

$$y_i = bx_i + u_i - \overline{u}. \tag{3.1}$$

## 3.1 The OLS Estimators $\widehat{a}$ and $\widehat{b}$ are Unbiased

$$E(\widehat{b}) = b \quad , \quad E(\widehat{a}) = a$$

**Proposition 3.1.** $E(\widehat{b}) = b$

*Proof.* Substituting (3.1) into (2.6)

$$\begin{aligned} \widehat{b} &= b + \frac{\sum_{i=1}^n x_i(u_i - \overline{u})}{\sum_{i=1}^n x_i^2} && (3.2) \\ &= b + \frac{\sum_{i=1}^n x_i u_i}{\sum_{i=1}^n x_i^2} && (3.3) \end{aligned}$$

where the second equality follows because $\sum_{i=1}^n x_i = 0$ .
Taking expectations,

$$E(\widehat{b}) = b + \frac{\sum_{i=1}^n x_i E(u_i)}{\sum_{i=1}^n x_i^2}$$

but, from assumption (A1), the second term is zero so that

$$E(\widehat{b}) = b \tag{3.4}$$

and the estimator $\widehat{b}$ is *unbiased*. □

**Proposition 3.2.** $E(\widehat{a}) = a$

*Proof.* Substituting (1.1) into (2.4)

$$\begin{aligned} \widehat{a} &= \frac{1}{n}\sum_{i=1}^n (a + bX_i + u_i) - \widehat{b}\frac{1}{n}\sum_{i=1}^n X_i \\ &= a + (b - \widehat{b})\frac{1}{n}\sum_{i=1}^n X_i + \frac{1}{n}\sum_{i=1}^n u_i && (3.5) \end{aligned}$$

and, taking expectations,

$$E(\widehat{a}) = a + (b - E(\widehat{b}))\frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\sum_{i=1}^{n} E(u_i)$$

and the last two terms are zero, by (3.4) and (A1), so that

$$E(\widehat{a}) = a$$

and the estimator $\widehat{a}$ is *unbiased*. □

## 3.2 The Variance of OLS Estimators

From (3.3)

$$\widehat{b} - E(\widehat{b}) = \frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}$$

so that

$$Var(\widehat{b}) \equiv E(\widehat{b} - E(\widehat{b}))^2 = E\left(\frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}\right)^2.$$

Expanding the numerator gives

$$Var(\widehat{b}) = \frac{\sum_{i=1}^{n} x_i^2 E(u_i^2) + \sum_{j\neq i}\sum_{i=1}^{n} x_i x_j E(u_i u_j)}{\left(\sum_{i=1}^{n} x_i^2\right)^2}$$

and from (A2) and (A3) it follows that

$$Var(\widehat{b}) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{\left(\sum_{i=1}^{n} x_i^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2} \tag{3.6}$$

Similarly, for $\widehat{a}$, from (3.5)

$$\widehat{a} - E(\widehat{a}) = (b - \widehat{b})\frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\sum_{i=1}^{n} u_i$$

so that

$$Var(\widehat{a}) = E\left((b - \widehat{b})\frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\sum_{i=1}^{n} u_i\right)^2.$$

It can be proved (although here the result will only be stated) that

$$Var(\widehat{a}) = \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} x_i^2}\right) = \frac{\sigma^2 \sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2} \tag{3.7}$$

## 3.3 The OLS estimator of $\sigma^2$

The error variance $\sigma^2$ that appears in formulae (3.7) and (3.6) is itself unknown and so in practice we need to estimate it. We now show that the estimator

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} \tag{3.8}$$

is an unbiased estimator of $\sigma^2$.

$$\begin{aligned}\sum_{i=1}^n e_i^2 &= \sum_{i=1}^n (Y_i - \widehat{a} - \widehat{b}X_i)^2 \\ &= \sum_{i=1}^n (y_i - \widehat{b}x_i)^2\end{aligned}$$

from (2.4). Substituting from (3.1)

$$\begin{aligned}\sum_{i=1}^n e_i^2 &= \sum_{i=1}^n (bx_i + u_i - \overline{u} - \widehat{b}x_i)^2 \\ &= (\widehat{b}-b)^2 \sum_{i=1}^n x_i^2 + \sum_{i=1}^n (u_i - \overline{u})^2 - 2(\widehat{b}-b)\sum_{i=1}^n x_i(u_i - \overline{u})\,. \end{aligned} \tag{3.9}$$

Consider the last term in equation (3.9). Note that $\sum_{i=1}^n x_i(u_i - \overline{u}) = \sum_{i=1}^n x_i u_i$ since $\overline{u}\sum_{i=1}^n x_i = 0$ and, from equation (3.3) $\sum_{i=1}^n x_i u_i = \sum_{i=1}^n x_i^2(\widehat{b}-b)$, so that, substituting and taking expectations,

$$E\left(\sum_{i=1}^n e_i^2\right) = E(\widehat{b}-b)^2 \sum_{i=1}^n x_i^2 + \sum_{i=1}^n E\,(u_i - \overline{u})^2 - 2E(\widehat{b}-b)^2 \sum_{i=1}^n x_i^2\,. \tag{3.10}$$

From the expression for $Var(\widehat{b})$ (3.6) the first and third terms in (3.10) are $\sigma^2$ and $-2\sigma^2$ respectively. Expanding the second term

$$\begin{aligned}\sum_{i=1}^n E\,(u_i - \overline{u})^2 &= \sum_{i=1}^n E(u_i)^2 - 2E\sum_{i=1}^n (u_i\overline{u}) + nE(\overline{u}^2) \\ &= \sum_{i=1}^n E(u_i)^2 - 2nE\,(\overline{u}\overline{u}) + nE(\overline{u}^2) \\ &= \sum_{i=1}^n E(u_i)^2 - nE(\overline{u}^2)\end{aligned} \tag{3.11}$$

but

$$E(\overline{u}^2) = E\left(\frac{1}{n}\sum_{i=1}^{n} u_i\right)^2 = \frac{1}{n^2}E(\sum_{i=1}^{n} u_i^2 + \sum_{j\neq i}\sum_{i=1}^{n} u_i u_j) = \frac{1}{n}\sigma^2$$

so that, substituting into (3.11),

$$\sum_{i=1}^{n} E\left(u_i - \overline{u}\right)^2 = \sum_{i=1}^{n} E(u_i)^2 - nE(\overline{u}^2) = n\sigma^2 - \sigma^2 = (n-1)\sigma^2\,.$$

Finally, substituting these expressions back into (3.10), we have

$$E\left(\sum_{i=1}^{n} e_i^2\right) = \sigma^2 + (n-1)\sigma^2 - 2\sigma^2 = (n-2)\sigma^2$$

so that

$$E(\widehat{\sigma}^2) = \frac{E\left(\sum_{i=1}^{n} e_i^2\right)}{(n-2)} = \sigma^2$$

which shows that $\widehat{\sigma}^2$ is an unbiased estimator of the error variance $\sigma^2$.

# 4 The Gauss-Markov Theorem

**Definition 4.1.** *A linear estimator is one that can be written in the form*

$$\widetilde{b} = \sum_{i}^{n} w_i Y_i$$

*where $w_i$ are fixed weights.*

Note that the OLS estimator $\widehat{b}$ is a linear estimator since

$$\widehat{b} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} = \sum_{i=1}^{n} k_i Y_i$$

where the weights $k_i$ are given by

$$k_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}\,.$$

**Theorem 4.1.** *The OLS estimator $\widehat{b}$ is the Best Linear Unbiased Estimator (BLUE) of the classical regresssion model. By **best** we mean the estimator in the class that achieves **minimum variance**.*

*Proof.* Taking expectations

$$\begin{aligned} E(\widetilde{b}) &= \sum_i^n w_i E(Y_i) \\ &= a\sum_i^n w_i + b\sum_i^n w_i X_i \end{aligned}$$

so the conditions for unbiasedness of $\widetilde{b}$ are that

$$\sum_i^n w_i = 0 \quad \text{and} \quad \sum_i^n w_i X_i = 1 \tag{4.1}$$

The variance of the estimator $\widetilde{b}$ is given by

$$Var(\widetilde{b}) = \sum_i^n w_i^2 Var(Y_i) = \sigma^2 \sum_i^n w_i^2$$

We now use a trick and add and subtract the OLS weights $k_i$ from this expression to give

$$\begin{aligned} Var(\widetilde{b}) &= \sigma^2 \sum_i^n (w_i - k_i + k_i)^2 \\ &= \sigma^2 \sum_i^n (w_i - k_i)^2 + \sigma^2 \sum_i^n k_i^2 + 2\sigma^2 \sum_i^n (w_i - k_i)k_i \end{aligned} \tag{4.2}$$

but the third term in (4.2) is zero from the unbiasedness conditions (4.1) since

$$\sum_i^n (w_i - k_i)k_i = \frac{\sum_{i=1}^n w_i X_i - \overline{X}\sum_{i=1}^n w_i}{\sum_{i=1}^n x_i^2} - \frac{1}{\sum_{i=1}^n x_i^2} = 0$$

Thus

$$\begin{aligned} Var(\widetilde{b}) &= \sigma^2 \sum_i^n (w_i - k_i)^2 + \sigma^2 \sum_i^n k_i^2 \\ &= \sigma^2 \sum_i^n (w_i - k_i)^2 + \frac{\sigma^2}{\sum_i^n x_i^2} \\ &= \sigma^2 \sum_i^n (w_i - k_i)^2 + Var(\widehat{b}) \end{aligned} \tag{4.3}$$

and the first term in (4.3) is greater than or equal to zero, achieving its lower bound for the OLS estimator where

$$w_i = k_i \quad \text{or equivalently} \quad \widetilde{b} = \widehat{b}\,.$$

□

# 5 Hypothesis Testing

In order to make statistical inferences on the parameter estimates $\widehat{a}$ , $\widehat{b}$ and $\widehat{\sigma}^2$we must make a further assumption:

$$u_i \sim iid\, N(0, \sigma^2) \quad , \quad i = 1, \cdots, n \tag{A5}$$

that the errors $u_i$ are distributed as independent normal variables. It then follows that, since $y_i$, $\widehat{a}$ and $\widehat{b}$ are all linear combinations of $u_i$, they are also distributed normally with

$$y_i \sim N(a + bX_i, \sigma^2) \quad , \quad i = 1, \cdots, n$$

$$\widehat{a} \sim N\left(a, \frac{\sigma^2 \sum_{i=1}^n X_i^2}{n \sum_{i=1}^n x_i^2}\right)$$

and

$$\widehat{b} \sim N\left(b, \frac{\sigma^2}{\sum_{i=1}^n x_i^2}\right) .$$

To make this practical we need to replace the unknown parameter $\sigma^2$ with the estimator $\widehat{\sigma}^2$ defined in (3.8). Then, defining

$$\widehat{\sigma}_{\widehat{a}} = \sqrt{\frac{\widehat{\sigma}^2 \sum_{i=1}^n X_i^2}{n \sum_{i=1}^n x_i^2}} \tag{5.1}$$

and

$$\widehat{\sigma}_{\widehat{b}} = \sqrt{\frac{\widehat{\sigma}^2}{\sum_{i=1}^n x_i^2}} \tag{5.2}$$

it can be shown that

$$\frac{\widehat{a} - a}{\widehat{\sigma}_{\widehat{a}}} \sim t_{n-2} \tag{5.3}$$

and

$$\frac{\widehat{b} - b}{\widehat{\sigma}_{\widehat{b}}} \sim t_{n-2} \tag{5.4}$$

where $t_{n-2}$ is the Student t distribution with $n-2$ degrees of freedom. $\widehat{\sigma}_{\widehat{a}}$ and $\widehat{\sigma}_{\widehat{b}}$ are known as the *standard errors* of the estimators $\widehat{a}$ and $\widehat{b}$ respectively.

It can also be shown that

$$(n-2)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2} \tag{5.5}$$

where $\chi^2_{n-2}$ is the Chi-squared distribution with $n-2$ degrees of freedom. $\widehat{\sigma}$ is known as the *equation standard error*.