Figure 13.7. SIMPLE LINEAR REGRESSION: Estimating Model Parameters and Responses

After a regression model has been *fitted* to a data set (Figure 13.3) and its *adequacy* for the purpose(s) of the investigation has been confirmed (Figure 13.5), we are in a position to *estimate* model parameters and make *predictions* about response variate values. Specifically, in this Figure, we deal with *interval* estimating of:

- the *slope* of the straight-line model, denoted β_1;
- the **Y**-*intercept* of the straight-line model, denoted β_0;
- the *mean* of Y when $x = x_j$, denoted $\mu_Y(x_j)$;
- the *value* of Y when $x = x_j$, denoted $y(x_j)$.

1. The *Study Population* Regression of Y on X

Our first model considers the response **Y** of study population element i to be made up of a *straight-line dependence* on the value of explanatory variate **X** for element i plus *noise* [*deviation* from this (straight-line) dependence] – see equation (13.7.1).

- This 'model' has *nothing* to do with probability – it is just a useful way of thinking about how the value arises for a study population element response.
- The noise component in (13.7.1) in referred to as a *residual* and is denoted $\mathbf{R}_i$.
- The population *size* (*i.e.*, its number of elements) is denoted **N** – see equation (13.7.2).
- $\mathbf{Y}_i$ in the model (13.7.1) *could* also be written as $\mathbf{Y}_i(\mathbf{X}_i)$ as an explicit reminder of its (straight-line) dependence on $\mathbf{X}_i$ – see equation (13.7.5).

$$\mathbf{Y}_i = \mathbf{B}_0 + \mathbf{B}_1\mathbf{X}_i + \mathbf{R}_i \qquad \text{-----(13.7.1)}$$

$$i = 1, 2, \ldots\ldots, \mathbf{N} \qquad \text{-----(13.7.2)}$$

$$_{reg}\overline{\mathbf{Y}} = \mathbf{B}_0 + \mathbf{B}_1\mathbf{X} \qquad \text{-----(13.7.3)}$$

$$_{reg}\overline{\mathbf{Y}}_i(\mathbf{X}_i) = \mathbf{B}_0 + \mathbf{B}_1\mathbf{X}_i \qquad \text{-----(13.7.4)}$$

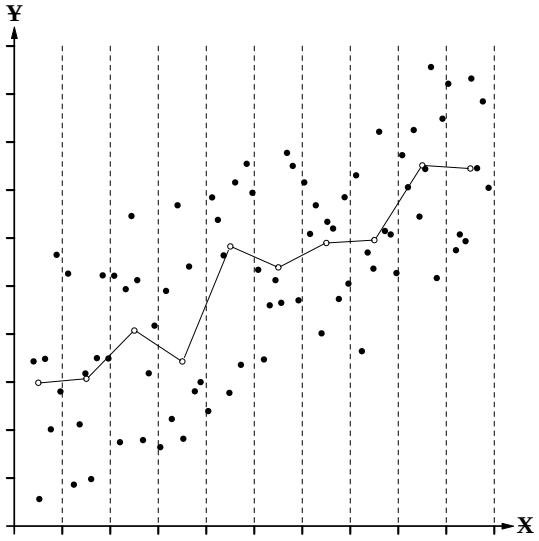
$$\mathbf{R}_i = \mathbf{Y}_i(\mathbf{X}_i) - {}_{reg}\overline{\mathbf{Y}}_i(\mathbf{X}_i) \qquad \text{-----(13.7.5)}$$

$$_{reg}\overline{\mathbf{Y}} = \mathbf{A} + \mathbf{B}_1(\mathbf{X} - \overline{\mathbf{X}}) \qquad \text{-----(13.7.6)}$$

$$\mathbf{A} = \mathbf{B}_0 + \mathbf{B}_1\overline{\mathbf{X}} \qquad \text{-----(13.7.7)}$$

A study population *attribute* is $_{reg}\overline{\mathbf{Y}}$, the *average* of **Y** under the straight-line dependence on **X** – this line is equation (13.7.3).

- $_{reg}\overline{\mathbf{Y}}$ is a curious population attribute – it is an *idealization* (or 'linearization') of the trend in the average of **Y** for ranges of **X** values, illustrated by the lines joining the circles in the scatter diagram at the right.
- If we refer to $_{reg}\overline{\mathbf{Y}}$ as an 'attribute', then $\mathbf{B}_0$ and $\mathbf{B}_1$ should be *sub*attributes, but the prefix *sub* is usually omitted.
- In the context of these Course Materials, it is understood that the model (13.7.3) would be fitted to the study population response variate values by the method of *least squares* and this model is referred to as the *regression of* **Y** *on* **X**.
  - Strictly, (13.7.3) is the *study population* regression of **Y** on **X** and (13.7.15) overleaf on page 13.40 is the *model* regression of **Y** on **X**.
- The *ordinate* of the point *on* the regression line at $\mathbf{X} = \mathbf{X}_i$ is denoted $_{reg}\overline{\mathbf{Y}}_i(\mathbf{X}_i)$ [or $_{reg}\mathbf{Y}_i(\mathbf{X}_i)$] – see equation (13.7.4); if there is no ambiguity, this may be shortened to $_{reg}\overline{\mathbf{Y}}_i$ [or $_{reg}\mathbf{Y}_i$].
- As shown in equation (13.7.5), the (population) residual $\mathbf{R}_i$ is the difference between the *actual* value of **Y** and its *model average* when $\mathbf{X} = \mathbf{X}_i$.



- Equation (13.7.3) can also be written in the *centred* form (13.7.6); we do so for two reasons:
  - algebraic convenience in least squares derivations, due to the fact that, over the *sample*: $\Sigma(x_j - \overline{x}) \equiv 0$ – see Figure 13.3;
  - there are data sets (usually involving **X** values far from the origin, such as calendar years) where the **Y**-intercept $\mathbf{B}_0$ is of little interest and, instead, **A** of the centred form is the relevant study population attribute.

2. The *Model* Regression of Y on X

In most investigating, there is access to only a *sample* of study population elements and Answer(s) to Question(s) must be based on sample *estimates* of the values of study population attributes; assessing how close these estimates are likely to be to the true values is usually based on a *model* for repetition of selecting and measuring.

The model for Y_j is given in equation (13.7.8).

- Y_j is a *random variable* whose distribution represents the possible values of the measured response variate **Y** for the jth element in the sample of n elements selected equiprobably from the study population, if the selecting and measuring processes were to be repeated over and over.
  - Strictly, only a *known* selecting probability is needed but we confine our attention to the simpler case of *equal* selecting probabilities.

$$Y_j = \beta_0 + \beta_1 x_j + R_j, \qquad \text{-----(13.7.8)}$$

$$j = 1, 2, \ldots\ldots, n, \quad \text{EPS} \qquad \text{-----(13.7.9)}$$

$$Y_j = \alpha + \beta_1(x_j - \overline{x}) + R_j \qquad \text{----(13.7.10)}$$

$$\alpha = \beta_0 + \beta_1\overline{x} \qquad \text{----(13.7.11)}$$

$$R_j \sim N(0, \sigma) \qquad \text{----(13.7.12)}$$

(*continued overleaf*)

– The model (13.7.8) takes no account of the *finite* size (**N** elements) of the study population – that is, it assumes **N** ≫ n.

– The model (13.7.8) can also be written in its centred form (13.7.10).

- β_0 is a model parameter which represents the **Y**-*intercept* of the straight-line model for the relationship between **X** and the average of **Y** in the study population; *i.e.*, the ordinate of this straight line when **X** = 0;

  β_1 is a model parameter which represents the *slope* of the straight-line model for the relationship between **X** and the average of **Y** in the study population; *i.e.*, the change in the average of **Y** for unit change in **X** over the elements of the study population;

  α is a model parameter which represents the *average* of **Y** for the elements of the study population whose value of **X** is $\bar{x}$, the *sample* average; it can be convenient to think of α as an 'intercept' – the ordinate of the point on the straight-line model for the relationship between **X** and the average of **Y** when **X** = $\bar{x}$.

  – The model parameters β_0 and β_1 represent the study population attributes $\mathbf{B}_0$ and $\mathbf{B}_1$.

$$R_j \sim N(0, \sigma), \quad \text{-----(13.7.12)}$$

$$R_j \text{ independent for all } j \quad \text{-----(13.7.13)}$$

$$Y_j \sim N(\beta_0 + \beta_1 x_j, \sigma) \quad \text{-----(13.7.14)}$$

$$\mu_Y = \beta_0 + \beta_1 x \quad \text{-----(13.7.15)}$$

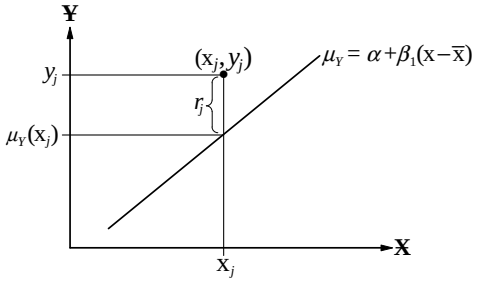
$$\mu_Y = \alpha + \beta_1 (x - \bar{x}) \quad \text{-----(13.7.16)}$$

$$\mu_Y(x_j) = \beta_0 + \beta_1 x_j \quad \text{-----(13.7.17)}$$

$$r_j = y_j - \mu_Y(x_j) \quad \text{-----(13.7.18)}$$

$$y_j = \beta_0 + \beta_1 x_j + r_j \quad \text{-----(13.7.19)}$$

- We assume a *normal* probability model (13.7.12) for the distribution of the residual (here, the random variable R_j), whose distribution represents the possible *differences*, from the (straight-line) structural component of the model, of the measured value of the response variate **Y** for the *j*th element in the sample of n elements selected equiprobably from the study population, if the selecting and measuring processes were to be repeated over and over.
  – σ, the (probabilistic) *standard deviation* of the normal model for the distribution of the residual, is a model parameter which represents the (data) *standard deviation* of the measured response variate of the elements of the study population with value x_j for explanatory variate **X**; this (data) standard deviation (and, hence, σ) quantifies the *variation* of the measured response variate of the elements of the study population with value x_j for explanatory variate **X** – as this variation increases, so does σ.
  – The *absence* of subscript *j* from σ corresponds to an *assumption* that the amount of variation in Y_j does *not* change with **X**, sometimes stated as σ is assumed to be *constant* across the range of **X** values.
  – A consequence of equations (13.7.8) and and (13.7.12) is that Y_j also has a normal distribution with standard deviation σ but with mean $\beta_0 + \beta_1 x_j$ – see equation (13.7.14).
- The sample size is n [equation (13.7.9)] and the R_js (and, hence, the Y_js) are *assumed* to be *probabilistically independent* for all values of *j* [equation (13.7.13)].
- The equation of the straight line represented by the structural component of the model (13.7.8) is given in (13.7.15).
  – Equation (13.7.15) can be called the *model* regression of **Y** on **X**, but the adjective *model* is usually omitted.
  – Equation (13.7.15) can also be written in *centred* form (13.7.16).
- The *ordinate* of the point *on* the regression line at **X** = x_j is denoted $\mu_Y(x_j)$ – see equation (13.7.17); it denotes the *mean* of Y_j when **X** = x_j and it represents the study population attribute ${}_{\text{reg}}\overline{\mathbf{Y}}_i(\mathbf{X}_i) \equiv {}_{\text{reg}}\overline{\mathbf{Y}}_i$.
- Suppose we have a *sample* of bivariate values (x_j, y_j), $j = 1, 2, \ldots .. n$, from which we wish to *estimate* the regression of **Y** on **X**; the diagram at the right shows the regression of **Y** on **X** written in centred form (13.7.16) and a typical point with coordinate values (x_j, y_j).



  – The point *on* the line at the same value of **X** is $[x_j, \mu_Y(x_j)]$ – see equation (13.7.17);
  – as shown in equation (13.7.18), the vertical distance of the point from the line is the *(realized) residual* r_j;
    + r_j occurs in equation (13.7.19) as the *value* of the random variable R_j of equations (13.7.8) and (13.7.10).
  – The *estimated* regression of **Y** on **X** (or the estimated least squares line) is the line which *minimizes the sum of the* sample *of* n *squared residuals* r_j^2.

3. The *Estimated* Regression of Y on X

In an investigation, a *particular* sample is selected which provides the data.

- We denote by (x_j, y_j), $j = 1, 2, \ldots .. n$, the sample of bivariate data available to evaluate b_0 and b_1, the *estimates* of the model parameters β_0 and β_1.
  – Elsewhere, the estimates b_0 and b_1 may be denoted $\hat{\beta}_0$ and $\hat{\beta}_1$.
  – The use of roman letters for y and j for the *data*, compared with italics for *model* values, reminds us the data is the *real world*, the model is an *idealization*; it is investigators' responsibility to ensure there is a reasonable basis for using real-world quantities as model quantities.
- Corresponding to equation (13.7.19), we have equation (13.7.20); the change

$$y_j = b_0 + b_1 x_j + \hat{r}_j \quad \text{-----(13.7.20)}$$

$$y_j = a + b_1 (x_j - \bar{x}) + \hat{r}_j \quad \text{-----(13.7.21)}$$

$$a = b_0 + b_1 \bar{x} \quad \text{-----(13.7.22)}$$

$${}_{\text{reg}}\bar{y} = b_0 + b_1 x \quad \text{-----(13.7.23)}$$

$${}_{\text{reg}}\bar{y} = a + b_1 (x - \bar{x}) \quad \text{-----(13.7.24)}$$

$$\hat{\mu}_Y(x_j) = b_0 + b_1 x_j \quad \text{-----(13.7.25)}$$

$$\hat{\mu}_Y(x_j) = a + b_1 (x_j - \bar{x}) \quad \text{-----(13.7.26)}$$

(*continued*)

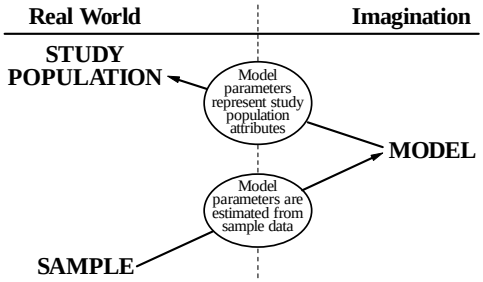
Figure 13.7. SIMPLE LINEAR REGRESSION: Estimating Model Parameters ..... (continued 1)

from model parameters to their estimates means that the last term on the right-hand side is now the *estimated* residual $\hat{r}_j$;
- the centred form is equation (13.7.21).

- The equation of the straight line called the *estimated* regression of **Y** on **X** is written as (13.7.23);
  - its centred form is equation (13.7.24).
- The ordinate of the point on the *estimated* regression line at $\mathbf{X} = x_j$ is denoted $\hat{\mu}_Y(x_j)$ and is given by equation (13.7.25);
  - its centred form is equation (13.7.26);
  - the notation reminds us that $\hat{\mu}_Y(x_j)$ is the *estimate* of the model quantity $\mu_Y(x_j)$ of equation (13.7.17);
  - $\hat{\mu}_Y(x_j)$ can also be denoted ${}_{reg}\overline{y}_j(x_j)$ (or even ${}_{reg}\overline{y}_j$), depending on context.
  - Elsewhere, you may see the notation $\hat{y}_j$ [or $\hat{y}_j(x_j)$] for the value of **Y** *on* the fitted line when $\mathbf{X} = x_j$; you may then see the equation of the line written as: $\hat{y} = b_0 + b_1 x$ (or even as: $y = b_0 + b_1 x$). The *dis*advantage of this usage is its lack of emphasis on a regression equation as a statement about the *average* of **Y**.

NOTES: 1. *All* mathematical models are idealizations and all are products of the intellect and the imagination, but we can distinguish two senses of the word *model* in the foregoing discussion – of the three expressions (13.7.1), (13.7.3), (13.7.8) + (13.7.12) ≡ (13.7.14), only the *last* involves *probability*. However, all *three* expressions, together with (13.7.9) and (13.7.13), comprise the model we use as the basis of *estimating* in this Figure.

It may be helpful to think of the model as a *link* between the *sample* and the *study population*; a pictorial representation of this idea is shown at the right.



2. *Assumptions* underlying the estimating processes which follow in this Figure can be summarized as:
   (1) We have chosen the *correct form* for the structural component of the model.
   (2) A statistically acceptable process is used to select the *sample*, and the **Y** and **X** values are measured with a *calibrated* measuring system of suitable *precision*.
   (3) The estimating processes are *conditional on* the observed **X** values, which have *no* measurement error.
   (4) $Y_j \sim N(\beta_0 + \beta_1 x_j, \sigma)$ [equation (13.7.14) ≡ (13.7.8) + (13.7.12) on the facing page 13.40].
   (5) The parameter σ, the (probabilistic) *standard deviation* of the normal model, is the *same* for all values of **X**.
   (6) The random variables Y_j are *probabilistically independent* for all *j*.

   Assumption (3) removes from the values of **X** two sources of error: sampling and measuring; as a consequence, **X** can be treated as *constant*, *not* a random variable, in derivations of distributions of estimators. This *greatly* simplifies the mathematics involved.
   - The assumption of no measurement error in the **X** values translates, in practice, into a reqirement that measurement error in the **X**s should be much *smaller* than that in the **Y**s.

3. Our use of *Roman* letters (x_j, y_j) $j = 1, 2, \ldots .. n$, for the sample data denotes restricting attention to just the data *values*. The derivations which follow in this Figure *broaden* this initial view to include the *method of selecting* the units which comprise the sample and which yield the data. We therefore distinguish y_j [a *data* value of y] and y_j [a value of the *random variable* Y_j in the model (13.7.8) [and (13.7.10)]. It is the task of investigation *design* (proper selecting, proper measuring, etc.) to provide a reasonable basis for being able to use the y_js as y_js.
   - The *italic* subscript *j* in equations (13.7.8), (13.7.9), (13.7.10), (13.7.12), (13.7.13), (13.7.14), (13.7.17), (13.7.18) and (13.7.19) is a reminder that *probability* (*e.g.*, *equi*probable) selecting is one basis for the *model* and, hence, for our *estimating* processes.
   - As in other developments of statistical theory in these Course Materials (*except* in Part 8 of STAT 220 on Sample Surveys), we model the uncertainty due to *sample* and *measurement* error; our *model-based* approach is founded on *conceptual repetition* of the processes of selecting the sample and measuring response variate values. In Part 8, the so-called *design-based* approach to interval estimating of study population attributes has the advantage of including the role of the *finite size* of the study population, but the *dis*advantage it is then too difficult mathematically for us to assess formally the uncertainty due to *measurement* error.

4. The centred form (13.7.10) of the model (13.7.8) has one *dis*advantage relating to the interpretation of α. In (13.7.6), **A** is the average ${}_{reg}\overline{\mathbf{Y}}_i$ when $\mathbf{X} = \overline{\mathbf{X}}$, but α represents the average ${}_{reg}\overline{\mathbf{Y}}_i$ when $\mathbf{X} = \overline{x}$. Thus, α does *not* represent **A** and its interpretation involves the **X**-values in the *particular* sample.

(*continued overleaf*)

4. Assessing Model Error

Proper use of regression models requires checking the modelling assumptions numbered (1), (4), (5) and (6) in Note 2 overleaf on page 13.41; these checks usually involve looking at appropriate scatter diagrams involving the estimated residuals, and such diagrams are typically produced by software used to calculate regression models. A summary of the four types of diagrams, and what to look for in them, is as follows:

(1) **Structural component:** The structural component of the model may have the wrong mathematical form of relationship. For instance, we may have used a *straight-line* relationship between **X** and the average of **Y** when the true relationship is *non*linear – the true relationship might be better modelled as a parabola or as a cubic polynomial, for example.

Assessment: Plot the estimated residuals against their **X** values;
if the structural component has the *correct* form, there should be no pattern in the scatter diagram;
a *patttern* indicates a wrong form for the structural component and it is often possible to infer an appropriate form from the type of pattern.

(4) **Distribution of the residuals:** The *normal* model (13.7.12) for the distribution of the residuals may not be reasonable.

Assessment: Construct a *histogram* of the estimated residuals and look for the characteristic normal shape; **OR**
construct a *normal quantile plot* of the estimated residuals and look for a roughly straight line on the plot – departures from a straight line may indicate the normal model is not reasonable.
[Constructing a histogram is more straight forward but an adequate histogram requires substantially more data (say, 50 or more observations) than a quantile plot.]

(5) **Constant standard deviation:** The estimating process assumes that σ, the (probabilistic) standard deviation of the normal model, does not change with increasing **X** values – there is no subscript j on σ in equation (13.7.12).
One form of this model error is that the variation in **Y** *in*creases as **X** increases;
less common is the variation in **Y** *de*creasing as **X** increases or increasing and then decreasing as **X** increases.

Assessment: Plot the estimated residuals against their **X** values; if the assumption of constant σ is reasonable, there should be no obvious change in the width of the band of points across the diagram;
the three instances mentioned above of this model error would be seen as:

- an *in*creasing width of the band of points across the diagram;
- a *de*creasing width of the band of points across the diagram;
- a band of points that first *in*creases and then *de*creases in width across the diagram.

(6) **Independence:** The model assumes the residuals R_j (and, hence, the Y_js) are *probabilistically independent*.

Assessment: Plot the estimated residuals in the *time order* in which the data were collected;
the independence assumption is accepted if there is no pattern in the scatter diagram;
evidence of *lack* of independence is when adjacent points tend to form clusters, indicating the estimated residuals of adjacent points are more similar in value than expected under independence.

NOTES: 5. We see that assessing model error involves graphical presentations of the estimated residuals; the statistical software that produces these presentations is likely to refer to 'residuals' rather than '*estimated* residuals'.

– As with any interpretation of a *pattern* (or its absence) in a scatter diagram, we must be careful not to *over*-interpret a diagram, particularly if the number of points is small (say, 20 or fewer).

6. The *same* scatter diagram is used in checks (1) and (5) above; what *differs* is the how the pattern (*i.e.*, the type of departure from the desired 'random-scatter') is interpreted in terms of which assumption is not met.

5. Estimating β_1, the Slope Parameter in the Model

The expressions [equation (13.3.19)] for the *estimate* b_1 of the parameter β_1 are given at the right as equation (13.7.27) except the j subscripts are now *italic*; the corresponding *estimator* is the random variable B_1 (of which b_1 is a *value*) in equation (13.7.28) – the difference between the two equations is that the second contains the *random variable* Y_j instead of its *value* y_j. We now find successively the *mean*, the *standard deviation* and the the *distribution* of B_1, starting from the *second* form of equation (13.7.28). We remember that, under assumption (3) overleaf, only Y_j in the numerator of this expression is a random variable; we also recall $\sum_{j=1}^{n}(x_j-\bar{x}) \equiv 0$ and the notation: $SS_x = \sum_{j=1}^{n}(x_j-\bar{x})^2$. A final point is that σ is *estimated* by the standard deviation of the points about the estimated regression of **Y** on **X**; *i.e.*,

$$b_1 = \frac{\sum_{j=1}^{n}(y_j-\bar{y})(x_j-\bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2} \equiv \frac{\sum_{j=1}^{n}y_j(x_j-\bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2} \qquad \text{-----(13.7.27)}$$

$$B_1 = \frac{\sum_{j=1}^{n}(Y_j-\bar{Y})(x_j-\bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2} \equiv \frac{\sum_{j=1}^{n}Y_j(x_j-\bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2} \qquad \text{-----(13.7.28)}$$

$$\hat{\sigma} = \sqrt{\sum_{j=1}^{n}\hat{r}_j^2/(n-2)} = \sqrt{\text{MSE}}. \qquad \text{-----(13.7.29)}$$

Mean: $$E(B_1) = E\left(\frac{\sum_{j=1}^{n}Y_j(x_j-\bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2}\right) = \frac{\sum_{j=1}^{n}(x_j-\bar{x})}{SS_x}E(Y_j) = \frac{\sum_{j=1}^{n}(x_j-\bar{x})}{SS_x}(\beta_0+\beta_1 x_j)$$

(*continued*)

Figure 13.7. SIMPLE LINEAR REGRESSION: Estimating Model Parameters ..... (continued 2)

$$= \frac{\beta_0}{SS_x}\sum_{j=1}^{n}(x_j-\bar{x}) + \frac{\beta_1}{SS_x}\sum_{j=1}^{n}(x_j-\bar{x})x_j = 0 + \frac{\beta_1}{SS_x}\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x}) = \beta_1. \quad \text{-----(13.7.30)}$$

S.d.: $$s.d.(B_1) = s.d.\left(\frac{\sum_{j=1}^{n}Y_j(x_j-\bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2}\right) = \sqrt{\frac{\sum_{j=1}^{n}(x_j-\bar{x})^2}{SS_x^2}[s.d.(Y_j)]^2} = \sqrt{\frac{SS_x}{SS_x^2}}\sigma = \sigma\sqrt{\frac{1}{SS_x}}. \quad \text{-----(13.7.31)}$$

Distn.: Because B_1 is a linear combination of n independent *normal* random variables Y_j, B_1 *also* has a normal distribution; hence:

$$B_1 \sim N(\beta_1, \sigma\sqrt{\tfrac{1}{SS_x}}). \quad \text{-----(13.7.32)}$$

Standardizing in equation (13.7.32) and *estimating* σ by $\hat{\sigma} = \sqrt{MSE}$, we obtain equation (13.7.33); this is the *test statistic* for tests of significance about β_1.

$$\frac{B_1-\beta_1}{\hat{\sigma}/\sqrt{SS_x}} \sim t_{n-2} \quad \text{-----(13.7.33)}$$

A $100(1-\alpha)\%$ confidence interval for β_1 is then given by equation (13.7.34).

$$b_1 \pm {}_{\alpha}t^*_{n-2}\hat{\sigma}\sqrt{\tfrac{1}{SS_x}} \quad \text{-----(13.7.34)}$$

NOTE: 7. Under H: $\beta_1 = 0$, the *value* of the left-hand side of (13.7.33) becomes:

$$\frac{b_1}{\hat{\sigma}/\sqrt{SS_x}} = \frac{\sqrt{b_1^2 SS_x}}{\hat{\sigma}} = \frac{\sqrt{b_1 SS_{xy}}}{\hat{\sigma}} = \frac{\sqrt{MSM}}{\hat{\sigma}}; \quad \text{-----(13.7.35)}$$

thus, the F-ratio in the ANOVA table and (13.7.33) provide *equivalent* tests of H: $\beta_1 = 0$. This is because $F_{1,n-2}$ is t^2_{n-2}, provided *both* left and right tail areas of the t distribution are included; hence, the values in the first column of the body of the version of Table 12.19t printed on the overleaf side (page 10.00) of Figure 10.12 (the standard normal table) are the *squares* of those of the relevant t distribution as given, for instance, in Figure 12.16.

6. Estimating $\mu_Y(x_j)$, the *Mean* of Y in the Model

We denote the *estimator* (a random variable) of $\mu_Y(x_j)$ by $\tilde{\mu}_Y(x_j)$ as given in equation (13.7.36) or, in *centred* form, equation (13.7.37).

$$\tilde{\mu}_Y(x_j) = B_0 + B_1 x_j \quad \text{-----(13.7.36)}$$

$$\tilde{\mu}_Y(x_j) = A + B_1(x_j-\bar{x}) \quad \text{-----(13.7.37)}$$

Because the estimate of α is $a = \bar{y}$, its estimator is $A = \bar{Y}$;
also, because $b_0 = \bar{y} - b_1\bar{x}$, the corresponding estimator is $B_0 = \bar{Y} - B_1\bar{x}$.

We now find successively the *mean*, the *standard deviation* and the *distribution* of $\tilde{\mu}_Y(x_j)$, starting from equation (13.7.36).

Mean:
$$E[\tilde{\mu}_Y(x_j)] = E[B_0 + B_1x_j] = E[\bar{Y} - B_1\bar{x} + B_1x_j] = E[\bar{Y}] - \bar{x}E[B_1] + x_jE[B_1]$$
$$= \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} + \beta_1x_j = \beta_0 + \beta_1x_j = \mu_Y(x_j). \quad \text{-----(13.7.38)}$$

S.d.:
$$s.d.[\tilde{\mu}_Y(x_j)] = s.d.[B_0 + B_1x_j] = s.d.[\bar{Y} - B_1\bar{x} + B_1x_j] = s.d.[\bar{Y} + (x_j-\bar{x})B_1]$$
$$= \sqrt{[s.d.(\bar{Y})]^2 + (x_j-\bar{x})^2[s.d.(B_1)]^2} = \sqrt{\frac{\sigma^2}{n} + (x_j-\bar{x})^2\frac{\sigma^2}{SS_x}} = \sigma\sqrt{\frac{1}{n} + \frac{(x_j-\bar{x})^2}{SS_x}} \equiv \sigma\sqrt{\frac{(x_j-\bar{x})^2}{SS_x} + \frac{1}{n}}. \quad \text{-----(13.7.39)}$$

Distn.: Because $\tilde{\mu}_Y(x_j)$ is a linear combination of the independent *normal* random variables B_0 and B_1, it *also* has a normal distribution; hence:

$$\tilde{\mu}_Y(x_j) \sim N[\mu_Y(x_j), \sigma\sqrt{\frac{(x_j-\bar{x})^2}{SS_x} + \frac{1}{n}}]. \quad \text{-----(13.7.40)}$$

Standardizing in equation (13.7.40) and *estimating* σ by $\hat{\sigma} = \sqrt{MSE}$, we obtain the test statistic (13.7.41) for tests of significance about $\mu_Y(x_j)$.

$$\frac{\tilde{\mu}_Y(x_j) - \mu_Y(x_j)}{\hat{\sigma}\sqrt{\frac{(x_j-\bar{x})^2}{SS_x} + \frac{1}{n}}} \sim t_{n-2} \quad \text{-----(13.7.41)}$$

A $100(1-\alpha)\%$ confidence interval for $\mu_Y(x_j)$ is then given by equation (13.7.42).

$$\hat{\mu}_Y(x_j) \pm {}_{\alpha}t^*_{n-2}\hat{\sigma}\sqrt{\frac{(x_j-\bar{x})^2}{SS_x} + \frac{1}{n}} \quad \text{-----(13.7.42)}$$

NOTE: 8. For random variables denoted by English letters, we are accustomed to the convention of an *upper* case letter for a random variable, a *lower* case letter for its values. For model *parameters*, which we denote by *Greek* letters, in the case of $\mu_Y(x_j)$, we use an *alternative* convention to distinguish:

$\mu_Y(x_j)$ – the model parameter representing a study population attribute;
[here, the *mean* of Y_j when $\mathbf{X} = x_j$, representing ${}_{reg}\bar{\mathbf{Y}}_i(\mathbf{X}_i) \equiv {}_{reg}\bar{\mathbf{Y}}_i$].

$\hat{\mu}_Y(x_j)$ – the *estimate* of the model parameter, a *number* whose value is derived from an appropriate set of *data* [and which is modelled as a *value* of the random variable $\tilde{\mu}_Y(x_j)$];

$\tilde{\mu}_Y(x_j)$ – the *estimator* of the model parameter;
[a *random variable* whose distribution represents the possible values of the estimate $\hat{\mu}_Y(x_j)$ under repetition of the selecting, measuring and estimating processes];

For the *slope*, the corresponding symbols are β_1, b_1 and B_1.

(*continued overleaf*)

7. Estimating β_0, the Y-Intercept of the Model

The intercept β_0 is the special case $\mu_Y(0)$ of $\mu_Y(x_j)$ when x_j is *zero*; the relevant test statistic and confidence interval are therefore equations (13.7.41) and (13.7.42) with $x_j = 0$, although *we* would use the notation β_0 for $\mu_Y(0)$, b_0 for $\hat{\mu}_Y(0)$ and B_0 for $\tilde{\mu}_Y(0)$.

NOTES: 9. If the data involve only $\mathbf{X}$ values *far* from zero, an inference about β_0 should be treated with caution; in such cases, an inference about α in the *centred* model (13.7.10) may be more relevant, but recall Note 4 on page 13.41.

10. If we are interested in the parameter α in the centred model (13.7.10), we can proceed as we do for β_0 *except* we set $\mathbf{X} = \bar{x}$ and use the notation α for $\mu_Y(\bar{x})$, a for $\hat{\mu}_Y(\bar{x})$ and A for $\tilde{\mu}_Y(\bar{x})$.

Because the least squares estimate of α is $a = \bar{y}$, the *average* of the y_js, we are reminded that the standard deviation of A is of the form $\sigma/\sqrt{n}$, where σ is estimated by $\sqrt{MSE}$ in this context.

8. Estimating $y(x_j)$, the *Value* of Y in the Model

Two points to remember are that:

- the *point* estimate of $y(x_j)$ [an *individual value*] is the *same* as that of $\mu_Y(x_j)$ [a *mean*];
- the *interval* estimate of $y(x_j)$ is *wider* than that of $\mu_Y(x_j)$ – intuitively, individuals vary *more* than averages so, for a given data set, there is more uncertainty in estimating an individual value than a mean; more formally, as well as having to estimate $\mu_Y(x_j)$, there is the *additional* source of uncertainty due to selecting the individual from the study population.

The argument leading to the expression for a *prediction* interval for $y(x_j)$ is like that in Figure 12.22.

We have, from equations (13.7.14) and (13.7.17): $$Y(x_j) \sim N[\mu_Y(x_j), \sigma]; \qquad \text{-----(13.7.43)}$$

and equation (13.7.40) is: $$\tilde{\mu}_Y(x_j) \sim N\left[\mu_Y(x_j), \sigma\sqrt{\frac{(x_j-\bar{x})^2}{SS_x} + \frac{1}{n}}\right]; \qquad \text{-----(13.7.44)}$$

subtracting, we obtain: $$Y(x_j) - \tilde{\mu}_Y(x_j) \sim N\left[0, \sigma\sqrt{\frac{(x_j-\bar{x})^2}{SS_x} + \frac{1}{n} + 1}\right], \qquad \text{-----(13.7.45)}$$

so that, when σ is *estimated* by $\hat{\sigma} = \sqrt{MSE}$, a $100(1-\alpha)\%$ *prediction* interval for $y(x_j)$ is given by equation (13.7.46). $$\hat{\mu}_Y(x_j) \pm {}_{\alpha}t^*_{n-2}\hat{\sigma}\sqrt{\frac{(x_j-\bar{x})^2}{SS_x} + \frac{1}{n} + 1} \qquad \text{-----(13.7.46)}$$

NOTES: 11. It is of interest to interpret equation (13.7.46) in terms of how its *components* contribute to quantifying the likely size of *sample* and *measurement* error in estimating $y(x_j)$ – our reference point is the *centred* model (13.7.10).

- ${}_{\alpha}t^*_{n-2}$ – determined by n and by the confidence level (the higher the level, the *greater* the error is likely to be);
- $\hat{\sigma}$ – the component from having to estimate σ, which quantifies variation around the straight-line model due to unknown or uncontrolled explanatory variates;
- SS_x – the component from having to estimate the model parameter β_1;
- $(x_j-\bar{x})^2$ – the component from how far x_j is from $\bar{x}$ (the further it is, the *greater* the error is likely to be);
- $1/n$ – the component from having to estimate the centred model parameter α;
- 1 – the component arising from the *individual* whose response (value) we are estimating (or 'predicting').

12. In the expression in square brackets under the square root in equation (13.7.46), the predominant term is the last '1', so the width of a PI for [an individual] $y(x_j)$ is largely *un*affected by the sample size n; by contrast, the CI for [the mean] $\mu_Y(x_j)$ in equation (13.7.42) overleaf on page 13.43 gets *narrower* as n increases (for a given set of data).

9. SUMMARY

Study population attribute	Model: Idea	Model: Symbol	Model: Estimator	Model: Estimate	Interval Estimate	Reference: Equation	Reference: Text[3]
$\mathbf{B}_1$	slope	β_1	$\tilde{\beta}_1$ or B_1	$\hat{\beta}_1$ or b_1	$b_1 \pm {}_{\alpha}t^*_{n-2}\hat{\sigma}\sqrt{\frac{1}{SS_x}}$	(13.7.34)	p. 672
$\mathbf{B}_0$	mean of Y when $\mathbf{X}=0$	β_0	$\tilde{\beta}_0$ or B_0	$\hat{\beta}_0$ or b_0	$b_0 \pm {}_{\alpha}t^*_{n-2}\hat{\sigma}\sqrt{\frac{\bar{x}^2}{SS_x} + \frac{1}{n}}$	(13.7.42)	p. 672
${}_{reg}\mathbf{Y}_i(\bar{x})$ [$\simeq \bar{\mathbf{Y}}$?]	mean of Y when $\mathbf{X}=\bar{x}$	α	$\tilde{\alpha}$ or A	$\hat{\alpha}$ or a	$a \pm {}_{\alpha}t^*_{n-2}\hat{\sigma}\sqrt{\frac{1}{n}}$	(13.7.42)	[see Note 4, page 13.41]
${}_{reg}\bar{\mathbf{Y}}_i(\mathbf{X}_i)$	mean of Y when $\mathbf{X}=x_j$	$\mu_Y(x_j)$	$\tilde{\mu}_Y(x_j)$ or $M(x_j)$	$\hat{\mu}_Y(x_j)$ or $m(x_j)$	$\hat{\mu}_Y(x_j) \pm {}_{\alpha}t^*_{n-2}\hat{\sigma}\sqrt{\frac{(x_j-\bar{x})^2}{SS_x} + \frac{1}{n}}$	(13.7.42)	p. 674
${}_{reg}\mathbf{Y}_i(\mathbf{X}_i)$	value of Y when $\mathbf{X}=x_j$	$y(x_j)$	$Y(x_j)$	$\hat{\mu}_Y(x_j)$ or $m(x_j)$	$\hat{\mu}_Y(x_j) \pm {}_{\alpha}t^*_{n-2}\hat{\sigma}\sqrt{\frac{(x_j-\bar{x})^2}{SS_x} + \frac{1}{n} + 1}$	(13.7.46)	p. 676