Figure 16.1. SIMPLE LINEAR REGRESSION: Background for Regression Modelling

After a regression model has been *fitted* to a data set and its *adequacy* for the purpose(s) of the investigation has been confirmed, we are in a position to *estimate* model parameters and make *predictions* about response variate values. Specifically, this Figure 16.1 gives the background for interval estimating of:

∗ the *slope* of the straight-line model, denoted β_1;
∗ the *intercept* of the straight-line model, denoted β_0;
∗ the *mean* of Y when $\mathrm{x} = \mathrm{x}_j$, denoted $\mu_Y(\mathrm{x} = \mathrm{x}_j)$ or $\mu_Y(\mathrm{x}_j)$;
∗ the *value* of Y when $\mathrm{x} = \mathrm{x}_j$, denoted $y(\mathrm{x} = \mathrm{x}_j)$ or $y(\mathrm{x}_j)$.

Our estimating processes will, as usual, be based on a suitable *probability* model, so we first review key ideas associated with formulating the *Gaussian* probability model we use.

1. The *Study Population* Regression of $\mathbf{Y}$ on $\mathbf{X}$

Our first model considers the response $\mathbf{Y}$ of population unit i to be made up of a *straight-line dependence* on the value of explanatory variate $\mathbf{X}$ for unit i plus *noise* [*deviation* from this (straight-line) dependence] – see equation (16.1.1).

$$\mathbf{Y}_\mathrm{i} = \mathbf{B}_0 + \mathbf{B}_1\mathbf{X}_\mathrm{i} + \mathbf{R}_\mathrm{i} \qquad \text{-----(16.1.1)}$$

$$\mathrm{i} = 1, 2, \ldots\ldots, \mathbf{N} \qquad \text{-----(16.1.2)}$$

$${}_{\mathrm{reg}}\overline{\mathbf{Y}} = \mathbf{B}_0 + \mathbf{B}_1\mathbf{X} \qquad \text{-----(16.1.3)}$$

$${}_{\mathrm{reg}}\overline{\mathbf{Y}}_\mathrm{i}(\mathbf{X}_\mathrm{i}) = \mathbf{B}_0 + \mathbf{B}_1\mathbf{X}_\mathrm{i} \qquad \text{-----(16.1.4)}$$

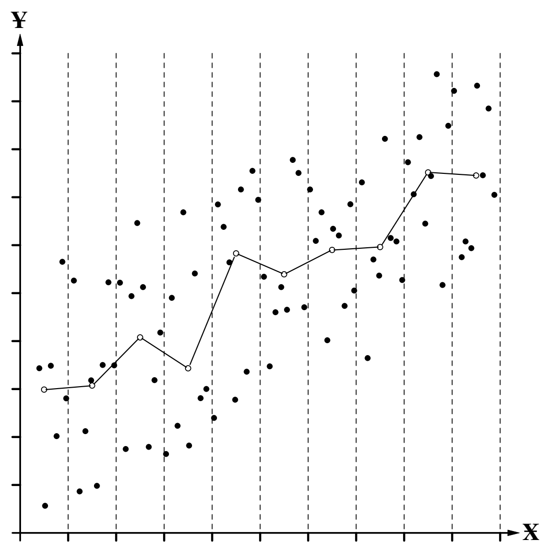
$$\mathbf{R}_\mathrm{i} = \mathbf{Y}_\mathrm{i}(\mathbf{X}_\mathrm{i}) - {}_{\mathrm{reg}}\overline{\mathbf{Y}}_\mathrm{i}(\mathbf{X}_\mathrm{i}) \qquad \text{-----(16.1.5)}$$

$${}_{\mathrm{reg}}\overline{\mathbf{Y}} = \mathbf{A} + \mathbf{B}_1(\mathbf{X} - \overline{\mathbf{X}}) \qquad \text{-----(16.1.6)}$$

$$\mathbf{A} = \mathbf{B}_0 + \mathbf{B}_1\overline{\mathbf{X}} \qquad \text{-----(16.1.7)}$$

- 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 unit response.
- The noise component in (16.1.1) in referred to as a *residual* and is denoted $\mathbf{R}_\mathrm{i}$.
- The population *size* (*i.e.*, its number of units) is denoted $\mathbf{N}$ – see equation (16.1.2).
- $\mathbf{Y}_\mathrm{i}$ in the model (16.1.1) *could* also be written as $\mathbf{Y}_\mathrm{i}(\mathbf{X}_\mathrm{i})$ as an explicit reminder of its (straight-line) dependence on $\mathbf{X}_\mathrm{i}$.

A study population *attribute* is ${}_{\mathrm{reg}}\overline{\mathbf{Y}}$, the *average* of $\mathbf{Y}$ under the straight-line dependence on $\mathbf{X}$ – this line is equation (16.1.3).



- ${}_{\mathrm{reg}}\overline{\mathbf{Y}}$ is a curious population attribute – it is an *idealization* (or 'linearization') of the trend in the average of $\mathbf{Y}$ for ranges of $\mathbf{X}$ values, illustrated by the lines joining the circles in the scatter diagram at the right – see also the last paragraph (–) on page 16.6.
- If we refer to ${}_{\mathrm{reg}}\overline{\mathbf{Y}}$ as an 'attribute', then $\mathbf{B}_0$ and $\mathbf{B}_1$ should be *sub*-attributes, but the prefix *sub* would usually be omitted.
- In this Figure 16.1 and Chapter 16 of the Course Notes, it is understood that the model (16.1.3) is fitted to the study population response variate values by the method of *least squares* and this model is referred to as the *regression of* $\mathbf{Y}$ *on* $\mathbf{X}$.
  – Strictly, (16.1.3) is the *study population* regression of $\mathbf{Y}$ on $\mathbf{X}$ and (16.1.15) overleaf is the *model* regression of $\mathbf{Y}$ on $\mathbf{X}$.
- The *ordinate* of the point *on* the regression line at $\mathbf{X} = \mathbf{X}_\mathrm{i}$ is denoted ${}_{\mathrm{reg}}\overline{\mathbf{Y}}_\mathrm{i}(\mathbf{X}_\mathrm{i})$ – see equation (16.1.4); if there is no ambiguity, this may be shortened to ${}_{\mathrm{reg}}\overline{\mathbf{Y}}_\mathrm{i}$.
- As shown in equation (16.1.5), the (population) residual $\mathbf{R}_\mathrm{i}$ is the difference between the *actual* value of $\mathbf{Y}$ and its *model average* when $\mathbf{X} = \mathbf{X}_\mathrm{i}$.
- Equation (16.1.3) can also be written in the *centred* form (16.1.6); we do so for two reasons:
  – algebraic convenience in least squares derivations, due to the fact that, over the *sample*: $\Sigma(\mathrm{x_j} - \overline{\mathrm{x}}) \equiv 0$;
  – there are data sets (usually involving $\mathbf{X}$ values far from the origin, such as years AD) where the $\mathbf{Y}$-intercept $\mathbf{B}_0$ is of little interest and, instead, $\mathbf{A}$ of the centred form is the relevant study population attribute.

Two *reasons* why straight lines (or more general equations) are fitted to bi-(or multi-)variate data are:

∗ to see if there *is* a relationship between $\mathbf{Y}$ and $\mathbf{X}$, what *form* it has and, perhaps, *why* it has this form;
∗ to use the equation as a convenient *summary* of a bi-(or multi-)variate data set, perhaps so we can calculate (or 'predict') the average of $\mathbf{Y}$ for some specified $\mathbf{X}$ (or the *change* in the average of $\mathbf{Y}$ corresponding to a change in $\mathbf{X}$).

For instance:
⊙ a chemical plant manager may wish to model product yield in terms of a process variable(s) that can be set to optimize the yield;
⊙ a personnel manager or university registrar tries to predict the likely success of job or university applicants in terms of an available measure(s) of previous performance;
⊙ a biologist or psychologist may try to model the behaviour of an organism or animal in terms of a characteristic(s) [explanatory variate(s)] which can be measured;

(*continued overleaf*)

- ⊙ a political scientist may want to relate the outcome of an election to a characteristic(s) of the candidates and the election campaign;
- ⊙ a stockbroker would like to be able to predict the behaviour of the stock market in terms of an available performance indicator(s).

2. The *Model* Regression of ¥ on **X**

In most investigating, there is access to only a *sample* of study population units 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 (16.1.8).

$$Y_j = \beta_0 + \beta_1 x_j + R_j \qquad \text{-----(16.1.8)}$$
$$j = 1, 2, \ldots, n; \ \text{EPS} \qquad \text{-----(16.1.9)}$$
$$Y_j = \alpha + \beta_1(x_j - \bar{x}) + R_j \qquad \text{----(16.1.10)}$$
$$\alpha = \beta_0 + \beta_1 \bar{x} \qquad \text{----(16.1.11)}$$
$$R_j \sim G(0, \sigma) \qquad \text{----(16.1.12)}$$
$$R_j \text{ independent for all } j \qquad \text{----(16.1.13)}$$
$$Y_j \sim G(\beta_0 + \beta_1 x_j, \sigma) \qquad \text{----(16.1.14)}$$
$$\mu_Y = \beta_0 + \beta_1 x \qquad \text{----(16.1.15)}$$
$$\mu_Y = \alpha + \beta_1(x - \bar{x}) \qquad \text{----(16.1.16)}$$
$$\mu_Y(x_j) = \beta_0 + \beta_1 x_j \qquad \text{----(16.1.17)}$$
$$r_j = y_j - \mu_Y(x_j) \qquad \text{----(16.1.18)}$$
$$y_j = \beta_0 + \beta_1 x_j + r_j \qquad \text{----(16.1.19)}$$

- Y_j is a *random variable* whose distribution represents the possible values of the measured response variate ¥ for the jth unit in the sample of n units 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 under EPS.
  - The model (16.1.8) takes no account of the *finite* size (**N** units) of the study population – that is, it assumes **N** $\gg$ n.
  - The model (16.1.8) can also be written in its centred form (16.1.10).
- β_0 is a model parameter which represents the ¥-*intercept* of the straight-line model for the relationship between **X** and the average of ¥ measured in the study population; *i.e.*, the ordinate of this straight line when **X** $= 0$;

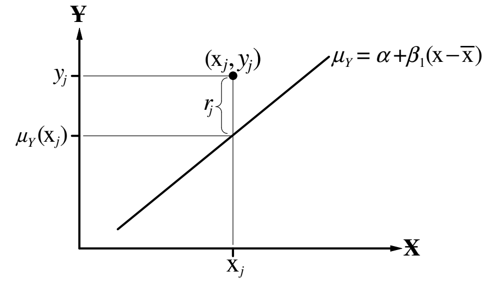
  β_1 is a model parameter which represents the *slope* of the straight-line model for the relationship between **X** and the average of ¥ measured in the study population; *i.e.*, the change in the average of ¥ for unit change in **X** over the units of the study population;

  α is a model parameter which represents the *average* of ¥ for the units 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 ¥ when **X** $= \bar{x}$.
  - The model parameters β_0 and β_1 represent the study population attributes $\mathbf{B}_0$ and $\mathbf{B}_1$.
- We assume a *Gaussian* probability model (16.1.12) 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 ¥ for the jth unit in the sample of n units 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 Gaussian model for the distribution of the residual, is a model parameter which represents the (data) *standard deviation* of the measured response variate ¥ of the units of the study population with value x_j for explanatory variate **X**; this standard deviation (and, hence, σ) quantifies the *variation* of the response variate of the units 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 (16.1.8) and (16.1.12) is that Y_j also has a Gaussian distribution with standard deviation σ but with mean $\beta_0 + \beta_1 x_j$ – see equation (16.1.14).
- The sample size is n – equation (16.1.9) – and the R_js (and, hence, the Y_js) are *assumed* to be *probabilistically independent* for all values of j – equation (16.1.13).
- The equation of the straight line represented by the structural component of the model (16.1.8) is given in (16.1.15).
  - Equation (16.1.15) can be called the *model* regression of ¥ on **X**, but the adjective *model* is usually omitted.
  - Equation (16.1.15) can also be written in *centred* form (16.1.16).
- The *ordinate* of the point *on* the regression line at **X** $= x_j$ is denoted $\mu_Y(x_j)$ – see equation (16.1.17); it denotes the *mean* of Y_j when **X** $= x_j$ and it represents the study population attribute $_{\text{reg}}\bar{\mathbf{Y}}_i(\mathbf{X}_i) \equiv {}_{\text{reg}}\bar{\mathbf{Y}}_i$.
- 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 ¥ on **X**; the diagram at the top of the facing page 16.3 shows the regression of ¥ on **X** written in centred form (16.1.16) 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 (16.1.17);
  - as shown in equation (16.1.18), the vertical distance of the point from the line is the *(realized) residual* r_j;

(*continued*)

Figure 16.1. SIMPLE LINEAR REGRESSION: Background for Regression M... (continued 1)

+ r_j occurs in equation (16.1.19) as the *value* of the random variable R_j of equation (16.1.8).
– 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.

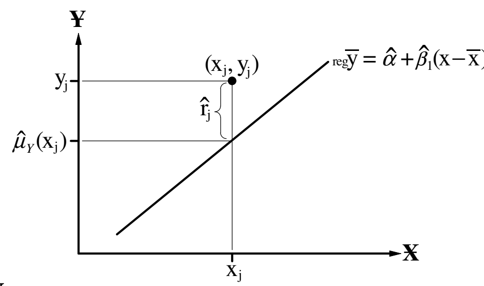


3. The *Estimated* Regression of Y on X

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

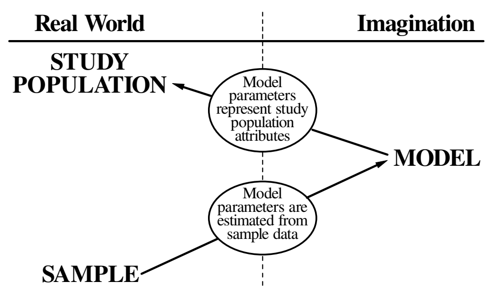
- We denote by (x_j, y_j), $j = 1, 2, \ldots n$, the sample of bivariate data available to evaluate $\hat{\beta}_0$ and $\hat{\beta}_1$, the *estimates* of the model parameters β_0 and β_1.
  – The use of roman letters for j and y for the *data*, compared with italics for *model* values, reminds us the data is the *real world*, the model is an *idealization*; it is investigators' responsibility to ensure that the Pan provides a basis for assuming reasonable equivalence of real-world and model quantities.
- Corresponding to equation (16.1.19), we have equation (16.1.20); the change 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 (16.1.21).
- The equation of the straight line called the *estimated* regression of **Y** on **X** is written as (16.1.23) and is shown (in centred form) in the diagram at the right;
  – its centred form is equation (16.1.24).
- The ordinate of the point on the *estimated* regression line at $x = x_j$ is denoted $\hat{\mu}_Y(x_j)$ and is given by equation (16.1.25);
  – its centred form is equation (16.1.26);
  – the notation reminds us that $\hat{\mu}_Y(x_j)$ is the *estimate* of the model quantity $\mu_Y(x_j)$ of equation (16.1.17);
  – $\hat{\mu}_Y(x_j)$ can also be denoted ${}_{\text{reg}}\bar{y}_j(x_j)$ (or even ${}_{\text{reg}}\bar{y}_j$), 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 $x = x_j$; you may then see the equation of the line written as: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ (or even as: $y = \hat{\beta}_0 + \hat{\beta}_1 x$). The *dis*advantage of this useage is its lack of emphasis on a regression equation as a statement about the *average* of **Y**.

$$y_j = \hat{\beta}_0 + \hat{\beta}_1 x_j + \hat{r}_j \qquad \text{-----(16.1.20)}$$
$$y_j = \hat{\alpha} + \hat{\beta}_1(x_j - \bar{x}) + \hat{r}_j \qquad \text{-----(16.1.21)}$$
$$\hat{\alpha} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \qquad \text{-----(16.1.22)}$$
$${}_{\text{reg}}\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 x \qquad \text{-----(16.1.23)}$$
$${}_{\text{reg}}\bar{y} = \hat{\alpha} + \hat{\beta}_1(x - \bar{x}) \qquad \text{-----(16.1.24)}$$
$$\hat{\mu}_Y(x_j) = \hat{\beta}_0 + \hat{\beta}_1 x_j \qquad \text{-----(16.1.25)}$$
$$\hat{\mu}_Y(x_j) = \hat{\alpha} + \hat{\beta}_1(x_j - \bar{x}) \qquad \text{-----(16.1.26)}$$



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 (16.1.1), (16.1.3), (16.1.8) + (16.1.12) ≡ (16.1.14), only the *last* involves *probability.* However, all *three* expressions, together with (16.1.9) and (16.1.13), comprise the model *we* use as the basis of *estimating.*

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. The *assumptions* underlying the estimating processes we use 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) $Y_j \sim G(\beta_0 + \beta_1 x_j, \sigma)$ [equation (16.1.14) ≡ (16.1.8) + (16.1.12) on the facing page 16.2].
   (5) The parameter σ, the *standard deviation* of the Gaussian 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.

(*continued overleaf*)

NOTES: (cont.) 3. Our use of *Roman* letters $(\mathrm{x_j}, \mathrm{y_j})$ j = 1, 2, ... n, for the sample data denotes restricting attention to just the data *values*. The derivations of distributions of estimators *broaden* this initial view to include the *method of selecting* the units which comprise the sample and which yield the data. We therefore distinguish $\mathrm{y_j}$ [a *data* value of y] and y_j [a value of the *random variable* Y_j in the model (16.1.8) [and (16.1.10)]. It is the task of investigation *design* (proper selecting, proper measuring, etc.) to provide a basis for being able to assume reasonable equivalence of these two ys.

- The *italic* subscript *j* in equations (16.1.8), (16.1.9), (16.1.10), (16.1.12), (16.1.13), (16.1.14), (16.1.17), (16.1.18) and (16.1.19) remindes us that *probability* (*e.g.*, *equi*probable) selecting is one *modelling* assumption and so undelies our *estimating* processes.
- As in other developments of statistical theory in these Course Materials, 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 the field of Survey Sampling, 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 that it is then difficult mathematically to assess formally the uncertainty due to *measurement* error.

4. To assist with learning the notation we use to maintain distinctions among the study population, the model and the sample, a summary of key equations from the preceding three sections is given in Table 16.1.1 at the right above.

Table 16.1.1:	**STUDY POPULATION**	**MODEL** for repeated selecting and measuring	**DATA** from the sample
Response:		$Y_j = \beta_0 + \beta_1 \mathrm{x}_j + R_j$	
	$\mathbf{Y}_\mathrm{i} = \mathbf{B}_0 + \mathbf{B}_1\mathbf{X}_\mathrm{i} + \mathbf{R}_\mathrm{i}$	$y_j = \beta_0 + \beta_1 \mathrm{x}_j + r_j$	$\mathrm{y_j} = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{x_j} + \hat{\mathrm{r}}_\mathrm{j}$
Regression line:	${}_{\mathrm{reg}}\overline{\mathbf{Y}} = \mathbf{B}_0 + \mathbf{B}_1\mathbf{X}$	$\mu_Y = \beta_0 + \beta_1 \mathrm{x}$	${}_{\mathrm{reg}}\overline{\mathrm{y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{x}$
Ordinate of point on line:	${}_{\mathrm{reg}}\overline{\mathbf{Y}}_\mathrm{i}(\mathbf{X}_\mathrm{i})$	$\mu_Y(\mathrm{x}_j)$	$\hat{\mu}_Y(\mathrm{x_j})$
In centred form:	$\mathbf{B}_0 = \mathbf{A} - \mathbf{B}_1\overline{\mathbf{X}}$	$\beta_0 = \alpha - \beta_1\overline{\mathrm{x}}$	$\hat{\beta}_0 = \hat{\alpha} - \hat{\beta}_1\overline{\mathrm{x}}$

5. The centred form (16.1.10) of the model (16.1.8) has one *dis*advantage relating to the interpretation of α. In (16.1.6), $\mathbf{A}$ is the average ${}_{\mathrm{reg}}\overline{\mathbf{Y}}_\mathrm{i}$ when $\mathbf{X} = \overline{\mathbf{X}}$, but α represents the average ${}_{\mathrm{reg}}\overline{\mathbf{Y}}_\mathrm{i}$ when $\mathbf{X} = \overline{\mathrm{x}}$. Thus, α does *not* represent $\mathbf{A}$ and its interpretation involves the $\mathbf{X}$-values in the *particular* sample.

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 16.3; 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 $\mathbf{X}$ and the average of $\mathbf{Y}$ 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 $\mathbf{X}$ 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 *Gaussian* model (16.1.12) for the distribution of the residuals may not be reasonable.

Assessment: Construct a *histogram* of the estimated residuals and look for the characteristic Gaussian shape; **OR**
construct a *Gaussian quantile plot* of the estimated residuals and look for a roughly straight line on the plot – departures from a straight line may indicate the Gaussian 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 Gaussian model, does not change with increasing $\mathbf{X}$ values – there is no subscript *j* on σ in equation (16.1.12).
One form of this model error is that the variation in $\mathbf{Y}$ *in*creases as $\mathbf{X}$ increases;
less common is the variation in $\mathbf{Y}$ *de*creasing as $\mathbf{X}$ increases or increasing and then decreasing as $\mathbf{X}$ increases.

Assessment: Plot the estimated residuals against their $\mathbf{X}$ 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.

(*continued*)

Figureo 16.1. SIMPLE LINEAR REGRESSION: Background for Regression M... (continued 2)

(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 probabilistic 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: 6. 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).

7. 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 modelling assumption is not met.

5. Summary of Estimating Results

Table 16.1.2 at the right summarizes the notation and formulae for five estimating processes in simple linear regression; these results are used throughout the Illustrations in Figures 16.2, 16.3 and 16.4. Interval estimates are equations (16.1.27) to (16.1.31).

Table 16.1.2

Study population attribute	...Model... Idea	Symbol	Estimator	Estimate	Interval Estimate	Reference Equation number	Reference Course Notes
Ƀ$_1$	slope	β_1	$\tilde{\beta}_1$	$\hat{\beta}_1$	$\hat{\beta}_1 \pm {}_{\alpha}t^*_{n-2}\,\hat{\sigma}\sqrt{\frac{1}{SS_x}}$	(16.1.27)	p. 16.7
Ƀ$_0$	mean of Y when X̶ $= 0$	β_0	$\tilde{\beta}_0$	$\hat{\beta}_0$	$\hat{\beta}_0 \pm {}_{\alpha}t^*_{n-2}\,\hat{\sigma}\sqrt{\frac{\bar{x}^2}{SS_x}+\frac{1}{n}}$	(16.1.28)	----
$_{reg}$Ɏ$_i(\bar{x})$ [≃ $\bar{}$Ɏ?]	mean of Y when X̶ $= \bar{x}$	α	$\tilde{\alpha}$	$\hat{\alpha}$	$\hat{\alpha} \pm {}_{\alpha}t^*_{n-2}\,\hat{\sigma}\sqrt{\frac{1}{n}}$	(16.1.29)	[see Note 5, page 16.4]
$_{reg}\bar{}$Ɏ$_i$(X̶$_i$)	mean of Y when X̶ $= x_j$	$\mu_Y(x_j)$	$\tilde{\mu}_Y(x_j)$	$\hat{\mu}_Y(x_j)$	$\hat{\mu}_Y(x_j) \pm {}_{\alpha}t^*_{n-2}\,\hat{\sigma}\sqrt{\frac{(x_j-\bar{x})^2}{SS_x}+\frac{1}{n}}$	(16.1.30)	p. 16.5
$_{reg}$Ɏ$_i$(X̶$_i$)	value of Y when X̶ $= x_j$	$y(x_j)$	$Y(x_j)$	$\hat{\mu}_Y(x_j)$	$\hat{\mu}_Y(x_j) \pm {}_{\alpha}t^*_{n-2}\,\hat{\sigma}\sqrt{\frac{(x_j-\bar{x})^2}{SS_x}+\frac{1}{n}+1}$	(16.1.31)	p. 16.8

NOTES: 8. It is of interest to interpret equation (16.1.31) 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 (16.1.10).

- ${}_{\alpha}t^*_{n-2}$ – 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').

9. In the expression under the square root in equation (16.1.31), 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 (16.1.30) gets *narrower* as n increases (for a given set of data).

6. Appendix: Background for Regression

We think first of *one* population of N̶ units, each with a value for a response variate Ɏ and an explanatory variate X̶; such a data set (X̶$_i$, Ɏ$_i$), $i = 1, 2, \ldots\ldots,$ N̶, would usually be shown graphically as a scatter diagram like that at the centre right of page 16.1. The Question we want to answer is: *What is the* form *of the relationship between* Ɏ *and* X̶ *in the study population*?

One answer to this question, for the N̶ (X̶, Ɏ) values, is Ɏ $= f($X̶$)$, where $f($X̶$)$ is a polynomial of degree N̶ passing through *all* the N̶ points (which are assumed to have *distinct* X̶ values); such a model for the form of the relationship contains N̶ + 1 parameters. Depending on the positions of the points (*i.e.*, depending on the units' X̶ and Ɏ values), a polynomial of *lower* degree may also pass through all the points but, in general, as the degree *de*creases, more and more of the points will lie *off* the curve representing the polynomial. *Our* concern *here* is with using a *straight line* (a *first*-degree polynomial with *two* parameters) to model the linear component of the *trend* in the points; this straight line *may* pass through one or more of the points or it may pass through *none* of them.

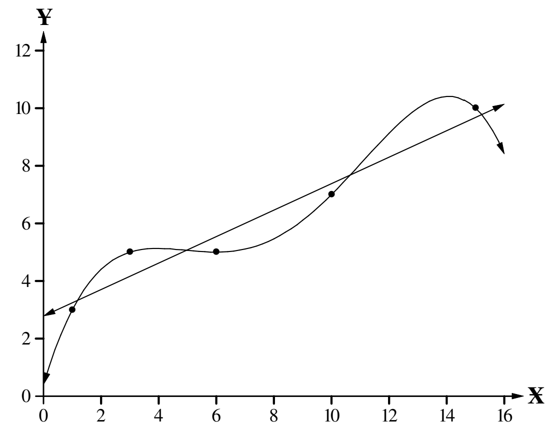
(*continued overleaf*)

There is only *one* polynomial of degree $\mathbf{N}$, specified by a unique set of values of its $\mathbf{N}+1$ parameters, which passes through all $\mathbf{N}$ points, but there is *more* than one line which could be thought of as representing the straight-line component of their *trend*. The criterion we use to select *one* of these lines is called *least squares*: the *line* of *best fit* to the $\mathbf{N}$ points is the one which *minimizes the sum of the squares of the vertical distances of the points from the line* (*i.e*., the one which minimizes the sum of the squared *residuals*). Such a line is often referred to, for historical reasons summarized on page 16.8, as *the regression of* $\mathbf{Y}$ *on* $\mathbf{X}$; its two parameters $\mathbf{B}_0$ (the $\mathbf{Y}$-intercept) and $\mathbf{B}_1$ (the slope) are *attributes* of the population.

An illustration of the matters described in the previous paragraph is shown at the right. There are five $(\mathbf{X}, \mathbf{Y})$ values:

$\mathbf{X}$	1	3	6	10	15
$\mathbf{Y}$	3	5	5	7	10,

and the fourth-degree polynomial which passes through the five points is: $$7560\mathbf{Y} = 3060 + 25236\mathbf{X} - 6205\mathbf{X}^2 + 608\mathbf{X}^3 - 19\mathbf{X}^4$$
[or: $\mathbf{Y} \simeq 0.4048 + 3.3381\mathbf{X} - 0.8208\mathbf{X}^2 + 0.08042\mathbf{X}^3 - 0.002513\mathbf{X}^4$];
also, using the results in Section 16.2 of the Course Notes, we find the regression of $\mathbf{Y}$ on $\mathbf{X}$ is:
$$_{\text{reg}}\overline{\mathbf{Y}} = 2.7778 + 0.4603\mathbf{X}.$$
The diagram at the right shows a scatter diagram of the five ($\mathbf{X}$, $\mathbf{Y}$) values with the polynomial and the straight line of best fit superimposed on it.

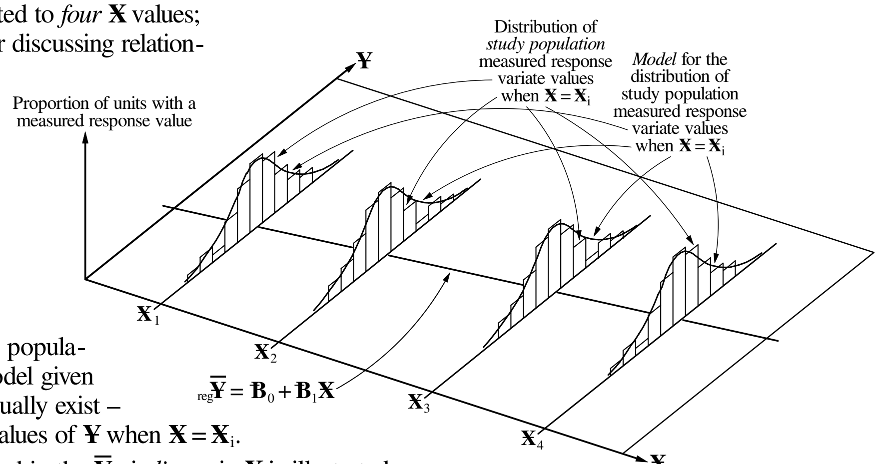


The discussion so far is in terms of a *population* of $\mathbf{N}$ $(\mathbf{X}, \mathbf{Y})$ values but the more usual data set in practice is a *sample* of n (x, y) values, where the sample size n is generally much *smaller* than the population size $\mathbf{N}$. As discussed in Sections 2 and 3 on pages 16.2 and 16.3, the modification to the foregoing discussion is that, for *sample* data, the method of least squares yields only *estimates*, $\hat{\beta}_0$ and $\hat{\beta}_1$, of the *model parameters* β_0 and β_1 representing the population attributes $\mathbf{B}_0$ and $\mathbf{B}_1$; *we* also refer to the line as the *estimated* regression of $\mathbf{Y}$ on $\mathbf{X}$. As always, the *method* of selecting the sample is *central* to assessing the likely size of the error associated with inference(s) from the sample about the relationship in the study population; one statistically acceptable method is, of course, *equiprobable* selecting (EPS).

- Statistical inference (interval estimating, significance testing) from a sample about the straight-line component of the relationship in the study population usually requires *distributional* assumptions about the population; for simple linear regression, the common *Gaussian* model is the five expressions (16.1.8,9) and (16.1.12,13,14) rewritten from page 16.2 at the right, where the Y_j (and the R_j) are *probabilistically independent* random variables.

$$Y_j \sim G(\beta_0 + \beta_1 x_j,\ \sigma), \qquad \text{-----(16.1.8)}$$
$$j = 1, 2, \ldots.., \text{n}. \qquad \text{-----(16.1.9)}$$
$$Y_j = \beta_0 + \beta_1 x_j + R_j, \quad R_j \sim G(0, \sigma), \qquad \text{--(16.1.14,12)}$$
$$j = 1, 2, \ldots.., \text{n}. \quad \text{independent, EPS.} \qquad \text{-----(16.1.13)}$$

This model somewhat *changes* our original picture of *one* population of $\mathbf{N}$ $(\mathbf{X}, \mathbf{Y})$ values – there is now *conceptually* a 'population' of $\mathbf{Y}$ values with a Gaussian distribution at *any* $\mathbf{X}$ value $\mathbf{X}_i$. This idea is shown pictorially in the diagram at the right below which, for convenience, is limited to *four* $\mathbf{X}$ values; the diagram shows how the model for discussing relationships *appears* as an extension from two to *many* populations, one for each value of $\mathbf{X}$.



  - This diagram shows *only* the populations and omits the samples; if they *were* to be included, there would typically be only *four* observations – one from each 'population' at the four $\mathbf{X}$ values.
  - There may be situations where the population histograms, implied by the model given at the right above for Y_j, do not actually exist – there may be only *one* (or a *few*) values of $\mathbf{Y}$ when $\mathbf{X} = \mathbf{X}_i$.
  - The model assumption that the trend in the $\overline{\mathbf{Y}}_i$s is *linear* in $\mathbf{X}$ is illustrated in the diagram at the centre right of page 16.1; as indicated by the dashed vertical lines, the points of the scatter diagram have been *grouped* into ten intervals on the basis of their $\mathbf{X}$ values and the *average* of $\mathbf{Y}$ (indicated by a circle ◦) calculated for each group; these averages have then been connected by lines. The regression of $\mathbf{Y}$ on $\mathbf{X}$ can be thought of as an *idealization* of this roughly linear trend in these *averages.*
    - A notational difficulty is that $\overline{\mathbf{Y}}_i$ is the natural symbol for the *actual* average of $\mathbf{Y}$ when $\mathbf{X} = \mathbf{X}_i$, but we also need a symbol for the *hypothetical* average of $\mathbf{Y}$ under the idealization of the straight-line regression model. In these Course

Figure 16.1. SIMPLE LINEAR REGRESSION: Background for Regression M... (continued 3)

Materials, we use $_{\text{reg}}\overline{\mathbf{Y}}_{\text{i}}(\mathbf{X}_{\text{i}})$ where:

- ⊙ the **bar** reminds us we are dealing with an *average* response;
- ⊙ the subscript **reg** reminds us this average is a *model* quantity;
- ⊙ the $\mathbf{X}_{\text{i}}$ reminds us the model average is a (straight-line) function of the value of the explanatory variate $\mathbf{X}$;
  [When there is no ambiguity, ($\mathbf{X}_{\text{i}}$) can be omitted and the population model average ordinate shortened to $_{\text{reg}}\overline{\mathbf{Y}}_{\text{i}}$].

When *probability* modelling is introduced to provide a basis for *inferences* based on *sample* data, the symbol for this average ordinate is $\mu_Y(\text{x}_j)$, which follows naturally from $_{\text{reg}}\overline{\mathbf{Y}}_{\text{i}}(\mathbf{X}_{\text{i}})$; also, we write the *equation* of the straight line fitted by the method of least squares as $_{\text{reg}}\overline{\mathbf{Y}} = \mathbf{B}_0 + \mathbf{B}_1\mathbf{X}$ or, for our illustrative population data set of 5 observations on page 16.6, as $_{\text{reg}}\overline{\mathbf{Y}} = 2.7778 + 0.4603\mathbf{X}$. We need to remember from the start that a regression model deals with an *average* response, although it may take time to get used to this change to our long-standing *mathematical* description of a straight line as $\text{y} = \text{b} + \text{mx}$.

It is the concern with *averges* that distinguishes most sharply regression modelling from the idea of a polynomial *passing through all the observations*; we are nor longer concerned with this latter idea in Chapter 16. The two approaches to fitting equations to points have different purposes and outcomes.

– Note the steps leading to the *Gaussian* probability model for Y_j: the (discrete) population *histogram* for $\mathbf{Y}$ at each value of $\mathbf{X}$ has a shape that is approximated by a (continuous) Gaussian p.d.f. and, under *equiprobable* selecting, this Gaussianicity for the *populations* carries through into the models for the random variables Y_j for each value x_j of $\mathbf{X}$ in the *sample*.

NOTES: 10. Five diagrams on the facing page 16.6 and pages 16.1 and 16.3 show the case of a *direct* relationship between $\mathbf{Y}$ and $\mathbf{X}$ – both variables tend to increase or decrease together (*i.e.*, $\mathbf{B}_1 > 0$); of course, *inverse* relationships, where one variable tends to *in*crease as the other *de*creases (*i.e.*, $\mathbf{B}_1 < 0$) are also important in practice.

11. A straight line fitted to *bi*variate data (or, equivalently, the values of its two parameters) can be regarded as a useful data *summary*, in the same way an *average* is a useful summary of *uni*variate data.

12. One purpose of fitting a straight line or other function to a set of bivariate data may be to *extrapolate* beyond the interval of the observed $\mathbf{X}$ values to 'predict' a value for (the average of) $\mathbf{Y}$. The diagram at the upper right of the facing page 16.6 reminds us of the importance of choosing a 'sensible' model and of the dangers of extrapolating;
    - the points in this diagram show an increasing roughly linear trend, but we have no *empirical* information about whether this form of relationship between $\mathbf{Y}$ and $\mathbf{X}$ persists *out*side the range of the data (*i.e.*, $\mathbf{X}$ values outside the interval from about 0 to 16);
    - the 'predicted' $\mathbf{Y}$ value for $\mathbf{X}$ *out*side the interval 0-16 would differ *substantially* between the *straight line* of best fit (which passes through *none* of the points but follows their *trend*) and the fourth degree polynomial (which passes through *all* the points but departs sharply from their trend ouside the interval of observation) – in the context of this Figure 16.1, the polynomial is introduced *only* for illustrative purposes and *our* real concern is with the straight line.

13. If the Question involves whether a change in $\mathbf{X}$ *causes* a change in $\mathbf{Y}$, the degree of limitation on the Answer depends on whether the data come from an *experimental* Plan [in which the *investigator(s)* determine the $\mathbf{X}$ values and then measure $\mathbf{Y}$] or from an *observational* Plan [in which the $\mathbf{Y}$ values arise from the values of $\mathbf{X}$ that occur in the units that happen to be investigated].

14. Note the sometimes confusing terminology used in the topic of 'regression' and the phrase *simple linear regression*.
    - *Regression*, in its statistical useage, means a process for finding a mathematical form for the relationship between (the average of) a measured *response* variate and one or more *explanatory* variates; like all mathematical models, regression *models* are only an *approximate* description of what is being modelled.
    - *Simple* refers to the *straight-line* regression model with *one* explanatory variate: $_{\text{reg}}\overline{\mathbf{Y}} = \beta_0 + \beta_1\mathbf{X}$, which is also *the regression of* $\mathbf{Y}$ on $\mathbf{X}$; *simple* is to be contrasted with *multiple*.
    - *Linear* refers to a regression model being linear in the *parameters* – it does *not* mean linear in the explanatory variate(s), so that: $_{\text{reg}}\overline{\mathbf{Y}} = \beta_0 + \beta_1\mathbf{X} + \beta_2\mathbf{X}^2$ *is* linear; linear is to be contrasted with *non-linear*.
    - *Multiple* regression models have *two or more* explanatory variates (on their right-hand side) – for example: $_{\text{reg}}\overline{\mathbf{Y}} = \beta_0 + \beta_1\mathbf{X}_1 + \beta_2\mathbf{X}_2$.
    - *Non-linear* refers to a regression model that is *not* linear in the *parameters* – for instance:
      $_{\text{reg}}\overline{\mathbf{Y}} = \beta_1\mathbf{X}^{\alpha}$ is non-linear (because of α) but can be made linear by a logarithmic transformation;
      $_{\text{reg}}\overline{\mathbf{Y}} = \beta_0 + \beta_1\mathbf{X}^{\alpha}$ is (inherently) non-linear.

15. Although we limit discussion to fitting *straight lines* to data in STAT 231, the method of least squares can also be

(*continued overleaf*)

NOTES: **(cont.)** 15. ● used to fit a second or higher degree polynomial (or a more complicated model) to a set of bi-(or multi-)variate data.

16. Information about two matters of historical interest in 'regression', taken from page 45 of an established 'regression' text (*Applied Regression Analysis*, Norman R. Draper and Harry Smith, Third Edition, John Wiley & Sons, New York, 1998), is as follows:

It appears that Sir Francis Galton (1822-1911), a well-known British anthropologist and meteorologist, was responsible for the introduction of the term *regression*. He originally used the word *reversion* in an unpublished address *Typical laws of heredity in man* to the Royal Institution on February 9, 1877. The later word *regression* appears in his Presidential address to Section H of the British Association at Aberdeen, 1885, printed in *Nature*, pages 507-510, September, 1885 [DC Library call number: PER Q1.N2], and also in a paper *Regression towards mediocrity in hereditary stature*, *Journal of the Anthropological Institute*: **15**, 246-263, 1885. In the latter, Galton reports (page 246) on his initial discovery that the offspring of seeds ..... *did* not *tend to resemble their parent seeds in size, but to be always more mediocre* (i.e., *more average) than they – to be smaller than the parents, if the parents were large; to be larger than the parents if the parents were very small. ..... The experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it.*. Galton then describes how the same features were observed in the records of *heights of 930 adult children and their respective parentages, 205 in number.* In essence, he shows that if variate y represents child's height and variate x represents a weighted average of the mother's and father's heights (see the original paper for details), an estimated regression of **Y** on **X** something like $y = \overline{y} + \frac{2}{3}(x - \overline{x})$ is appropriate, although he does not phrase it like this. Galton's paper makes fascinating reading, as does the account in *The History of Statistics* by S.M. Stigler, pages 294-299, Belknap Press of Harvard University, 1986 [DC Library call number: QA276.15.S75 1986]. Today, Galton's analysis would be called *correlation analysis*, a term for which he is also responsible. The term *regression* soon came to be applied to relationships in situations *other than* the one from which it arose, including situations where the explanatory variate can*not* reasonably be represented by a random variable, and its use has persisted to this day. In most current model-fitting situations, there is no element of 'regression' in the original sense, but the word is so established that it is retained in most statistical communication (including these Course Materials).

There has been a dispute about who first discovered the method of least squares. It appears it was discovered independently by Carl Friedrich Gauss (1777-1855) and Adrien Marie Legendre (1752-1833), that Gauss started using it before 1803 (he claimed in about 1795, but there is no corroboration of this earlier date), and that the first account was published by Legendre in 1805. When Gauss wrote in 1809 that he had used the method earlier than the date of Legendre's publication, controversy concerning the priority began. The facts are carefully sifted and discussed by R.L. Plackett in *Studies in the history of probability and statistics. XXIX. The discovery of the method of least squares. Biometrika*: **59**, 239-251, 1972 [DC Library call number: PER QH301.B5], a paper well worth reading. Also of interest are:

- C. Eisenhart: *The meaning of 'least' in least squares*, *Journal of the Washington Academy of Sciences*: **54**, 24-33, 1964 [DC Library call number: PER Q11.W32], which is reprinted in:
  - *Precision, Measurement and Calibration*, edited by H.H. Ku, U.S. National Bureau of Standards Special Publication 300, 1969, Volume 1;
- *Gauss, Carl Friedrich* in: *International Encyclopedia of the Social Sciences*, Macmillan Co., Free Press Division, New York, 1968, Volume 6, pages 74-81 [CG Library call number: H40.A215 1968], and:
  - *Encyclopedia of Statistical Sciences*, John Wiley & Sons, New York, 1983, Volume 3, pages 305-309 [DC Library call number: Ref QA276.14.E5 1982];
- S.M. Stigler: *Gergonne's 1815 paper on the design and analysis of polynomial regression experiments*, *Historia Mathematica*: **1**, 431-447, 1974 (see page 433) [DC Library call number: PER QA21.H58x];
  - see also Stigler's book *The History of Statistics* [DC Library call no.: QA276.15.S75 1986] referenced above.

Newspaper articles which involve statistical matters discussed in this Figure 16.1 are:

EM9701: *Enrolment and test scores*	EM8911: The *Challenger* disaster
EM9201: *Will women soon outrun men?*	EM9204: *Lead, Lies and Data Tape*
EM9901: *Skin-cancer rate levels off as Canadians cover up*	EM9901: *Cutback in supply of MDs urged*

The first, second and third of these articles are also cited for Chapter 7 of the Course Notes.