Figure 13.3. SIMPLE LINEAR REGRESSION: The Method of Least Squares

Our main concern in this Figure 13.3 is the method of *least squares* – we use it to find the straight line of *best fit* to a set of bivariate data. [A more *evocative* name for the process would be the method of *minimum squared deviations*].

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

Our first model considers the response $\mathbf{Y}$ of respondent population element i to be made up of a *straight-line dependence* on the value of explanatory variate $\mathbf{X}$ for element i plus *noise* [*deviation* from this (straight-line) dependence] – see equation (13.3.1).

$$\mathbf{Y}_i = \mathbf{B}_0 + \mathbf{B}_1\mathbf{X}_i + \mathbf{R}_i \quad \text{-----(13.3.1)}$$

$$i = 1, 2, \ldots\ldots, \mathbf{N} \quad \text{-----(13.3.2)}$$

$${}_{\mathrm{reg}}\overline{\mathbf{Y}} = \mathbf{B}_0 + \mathbf{B}_1\mathbf{X} \quad \text{-----(13.3.3)}$$

$${}_{\mathrm{reg}}\overline{\mathbf{Y}}_i(\mathbf{X}_i) = \mathbf{B}_0 + \mathbf{B}_1\mathbf{X}_i \quad \text{-----(13.3.4)}$$

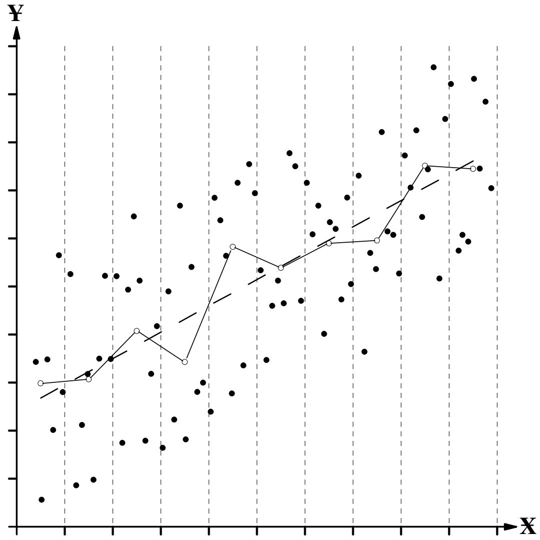
$$\mathbf{R}_i = \mathbf{Y}_i(\mathbf{X}_i) - {}_{\mathrm{reg}}\overline{\mathbf{Y}}_i(\mathbf{X}_i) \quad \text{-----(13.3.5)}$$

$${}_{\mathrm{reg}}\overline{\mathbf{Y}} = \mathbf{A} + \mathbf{B}_1(\mathbf{X} - \overline{\mathbf{X}}) \quad \text{-----(13.3.6)}$$

$$\mathbf{A} = \mathbf{B}_0 + \mathbf{B}_1\overline{\mathbf{X}} \quad \text{-----(13.3.7)}$$

- This 'model' has *nothing* to do with probability – it is just a useful way of thinking about how the value arises for a respondent population element response.
- The noise component in (13.3.1) in referred to as a *residual* and is denoted $\mathbf{R}_i$.
- The population *size* (*i.e.*, its number of elements) is denoted $\mathbf{N}$ – see equation (13.3.2).
- $\mathbf{Y}_i$ in the model (13.3.1) can 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 (13.3.5).

A respondent 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 (13.3.3).



- ${}_{\mathrm{reg}}\overline{\mathbf{Y}}$ 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, and their idealized straight line in long dashes, in the scatter diagram at the right.
  - 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* is usually omitted.
- In the context of these Course Materials, it is understood that the model (13.3.3) would be fitted to the respondent population response values by the method of *least squares* and the fitted model is referred to as the *regression of* $\mathbf{Y}$ *on* $\mathbf{X}$.
  - Strictly, (13.3.3) is the *respondent population* regression of $\mathbf{Y}$ on $\mathbf{X}$ and (13.3.12) overleaf is the *model* regression of $\mathbf{Y}$ on $\mathbf{X}$.
- The *ordinate* of the point *on* the regression line (*i.e.*, the *value* of ${}_{\mathrm{reg}}\overline{\mathbf{Y}}$) at $\mathbf{X} = \mathbf{X}_i$ is denoted ${}_{\mathrm{reg}}\overline{\mathbf{Y}}_i(\mathbf{X}_i)$ – see equation (13.3.4); if there is no ambiguity, this may be shortened to ${}_{\mathrm{reg}}\overline{\mathbf{Y}}_i$
- The (population) residual $\mathbf{R}_i$ is the difference between the *actual* value of $\mathbf{Y}$ and its *model average* when $\mathbf{X} = \mathbf{X}_i$, as shown in equation (13.3.5),
- Equation (13.3.3) can also be written in the *centred* form (13.3.6); we do so for two reasons:
  - algebraic convenience in least squares derivations, due to the fact that, over the *sample*: $\Sigma(x_j - \overline{x}) \equiv 0$ [equation (13.3.26)];
  - 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 respondent 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 may try 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) which can be measured;
- ⊙ 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).

(*continued overleaf*)

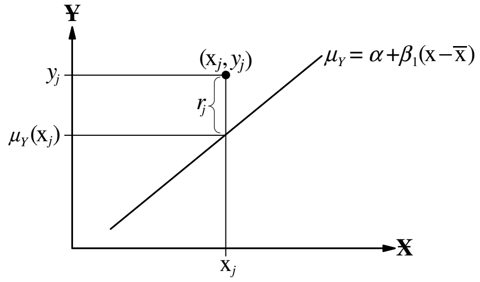
2. The *Model* Regression of $\mathbb{Y}$ on $\mathbb{X}$

In most investigating, there is access to only a *sample* of respondent population elements and an Answer to a Question must be based on a sample *estimate* of the value of the respondent population attribute of interest; assessing how close such an estimate is likely to be to the *true* value is usually based on a *model* for repeated selecting and measuring.

The model for Y_j is given in equation (13.3.8).

$$Y_j = \beta_0 + \beta_1 x_j + R_j \quad \text{-----(13.3.8)}$$
$$j = 1, 2, \ldots, n \quad \text{-----(13.3.9)}$$
$$Y_j = \alpha + \beta_1(x_j - \overline{x}) + R_j \quad \text{----(13.3.10)}$$
$$\alpha = \beta_0 + \beta_1\overline{x} \quad \text{----(13.3.11)}$$
$$\mu_Y = \beta_0 + \beta_1 x \quad \text{----(13.3.12)}$$
$$\mu_Y = \alpha + \beta_1(x - \overline{x}) \quad \text{----(13.3.13)}$$
$$\mu_Y(x_j) = \beta_0 + \beta_1 x_j \quad \text{----(13.3.14)}$$
$$r_j = y_j - \mu_Y(x_j) \quad \text{----(13.3.15)}$$
$$y_j = \beta_0 + \beta_1 x_j + r_j \quad \text{----(13.3.16)}$$

- Y_j is a *random variable* whose distribution represents the possible values of the measured response variate $\mathbb{Y}$ for the jth element in the sample of n elements selected equiprobably from the respondent 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.
  - The model parameters β_0 and β_1 represent the respondent population attributes $\mathbb{B}_0$ and $\mathbb{B}_1$.
  - The model (13.3.8) takes no account of the *finite* size ($\mathbb{N}$ elements) of the respondent population – that is, it assumes $\mathbb{N} \gg n$.
  - The model (13.3.8) can also be written in its centred form (13.3.10).
- The equation of the straight line represented by the structural component of the model (13.3.8) is given in (13.3.12).
  - Equation (13.3.12) is, strictly, the *model* regression of $\mathbb{Y}$ on $\mathbb{X}$, but the adjective *model* is usually omitted.
  - Equation (13.3.12) can also be written in *centred* form (13.3.13).
- The *ordinate* of the point *on* the regression line at $\mathbb{X} = x_j$ is denoted $\mu_Y(x_j)$ – see equation (13.3.14); it denotes the *mean* of Y_j when $\mathbb{X} = x_j$ and it represents the respondent population attribute ${}_{\text{reg}}\overline{\mathbb{Y}}_i(\mathbb{X}_i) \equiv {}_{\text{reg}}\overline{\mathbb{Y}}_i$.
- 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 $\mathbb{Y}$ on $\mathbb{X}$; the diagram at the right shows the regression of $\mathbb{Y}$ on $\mathbb{X}$ written in centred form (13.3.13) and a typical point with coordinate values (x_j, y_j).
  - The point *on* the line at the same value of $\mathbb{X}$ is $[x_j, \mu_Y(x_j)]$ – see equation (13.3.14);
  - as shown in equation (13.3.15), the vertical distance of the point from the line is the *(realized) residual* r_j;
    + r_j occurs in equation (13.3.16) as the value of the random variable R_j of equation (13.3.8).
  - The *estimated* regression of $\mathbb{Y}$ on $\mathbb{X}$ (or the estimated least squares line) is the one which *minimizes the sum of the* sample *of* n *squared (realized) residuals – i.e.*, of the n r_j^2.



3. The *Estimated* Regression of $\mathbb{Y}$ on $\mathbb{X}$

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

$$y_j = b_0 + b_1 x_j + \hat{r}_j \quad \text{-----(13.3.17)}$$
$$y_j = a + b_1(x_j - \overline{x}) + \hat{r}_j \quad \text{-----(13.3.18)}$$
$$a = b_0 + b_1\overline{x} \quad \text{-----(13.3.19)}$$
$${}_{\text{reg}}\overline{y} = b_0 + b_1 x \quad \text{-----(13.3.20)}$$
$${}_{\text{reg}}\overline{y} = a + b_1(x - \overline{x}) \quad \text{-----(13.3.21)}$$
$$\hat{\mu}_Y(x_j) = b_0 + b_1 x_j \quad \text{-----(13.3.22)}$$
$$\hat{\mu}_Y(x_j) = a + b_1(x_j - \overline{x}) \quad \text{-----(13.3.23)}$$

- In this Figure 13.3, we denote the sample of bivariate data by (x_j, y_j), $j = 1, 2, \ldots$ n; we use these data 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 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 the investigators' responsibility to ensure there is a basis for assuming reasonable equivalence of real-world and model quantities.
- Corresponding to equation (13.3.16), we have equation (13.3.17); 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 (13.3.18).
- The equation of the straight line we call the *estimated* regression of $\mathbb{Y}$ on $\mathbb{X}$ is written as (13.3.20);
  - its centred form is equation (13.3.21).
- 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 (13.3.22);
  - its centred form is equation (13.3.23);
  - the notation reminds us that $\hat{\mu}_Y(x_j)$ is the *estimate* of the model quantity $\mu_Y(x_j)$ of equation (13.3.14);
  - $\hat{\mu}_Y(x_j)$ can also be denoted ${}_{\text{reg}}\overline{y}_j(x_j)$ (or even ${}_{\text{reg}}\overline{y}_j$), depending on context – for example, see Figure 13.5.
  - 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: $\hat{y} = b_0 + b_1 x$ (or even as: $y = b_0 + b_1 x$). A reason to *avoid* this notation is its lack of em-

Figure 13.3. SIMPLE LINEAR REGRESSION: The Method of Least Squares (continued 1)

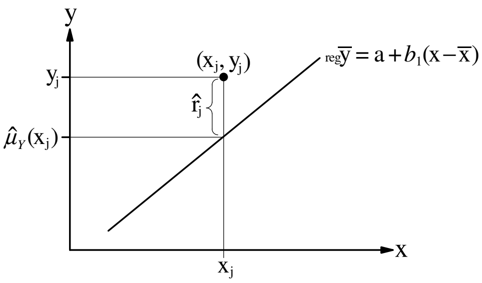
– phasis on the equations of regression models as statements about the *average* of $\mathbf{Y}$.

NOTE: 1. To assist with learning the notation we use to maintain distinctions among the responent population, the model model and the sample, a summary of key equations from the preceding three sections is given in Table 13.3.1 at the right above.

Table 13.3.1	**RESPONDENT POPULATION**	**MODEL** for repeated selecting and measuring	**DATA** from the sample (The real world)
Response:	$\mathbf{Y}_i = \mathbf{B}_0+\mathbf{B}_1\mathbf{X}_i+\mathbf{R}_i$	$Y_j = \beta_0+\beta_1 x_j+R_j$ $y_j = \beta_0+\beta_1 x_j+r_j$	$y_j = b_0+b_1x_j+\hat{r}_j$
Regression line:	${}_{\text{reg}}\overline{\mathbf{Y}} = \mathbf{B}_0+\mathbf{B}_1\mathbf{X}$	$\mu_Y = \beta_0+\beta_1 x$	${}_{\text{reg}}\overline{y} = b_0+b_1x$
Ordinate of point on line:	${}_{\text{reg}}\overline{\mathbf{Y}}_i(\mathbf{X}_i)$ or ${}_{\text{reg}}\overline{\mathbf{Y}}_i$	$\mu_Y(x_j)$	$\hat{\mu}_Y(x_j)$ or ${}_{\text{reg}}\overline{y}_j(x_j)$ or ${}_{\text{reg}}\overline{y}_j$
In centred form:	$\mathbf{B}_0 = \mathbf{A}-\mathbf{B}_1\overline{\mathbf{X}}$	$\beta_0 = \alpha-\beta_1\overline{x}$	$b_0 = a-b_1\overline{x}$

4. Finding the Estimated Regression of $\mathbf{Y}$ on $\mathbf{X}$

Suppose we have a *sample* of bivariate observations (x_j, y_j), $j = 1, 2, \ldots\ldots n$, from which we wish to *estimate* the regression of $\mathbf{Y}$ on $\mathbf{X}$. The diagram at the right shows this line and a typical data point (x_j, y_j);
the point *on* the line at the same value of x is $[x_j, \hat{\mu}_Y(x_j)]$;
the vertical distance of the point from the line is the *estimated residual* $\hat{r}_j = y_j - \hat{\mu}_Y(x_j)$;
the least squares line (or the estimated regression of $\mathbf{Y}$ on $\mathbf{X}$) is the one which *minimizes the sum of the* n *squared residuals* r_j^2.



We write: $g(\alpha, \beta_1) = \sum_{j=1}^{n} r_j^2 = \sum_{j=1}^{n}[y_j - \mu_Y(x_j)]^2 = \sum_{j=1}^{n}[y_j - \alpha - \beta_1(x_j-\overline{x})]^2$;

to find the least squares estimates, a and b_1, of α and β_1, differentiate $g(\alpha, \beta_1)$ with respect to α and with respect to β_1, set the resulting partial derivative expressions to zero, and solve for a and b_1.

Then: $\dfrac{\partial g}{\partial \alpha} = -2\sum_{j=1}^{n}[y_j - \alpha - \beta_1(x_j-\overline{x})]$, which is zero when: $\sum_{j=1}^{n}[y_j - a - b_1(x_j-\overline{x})] = 0$,

i.e., $\sum_{j=1}^{n} y_j - na - b_1\overset{0}{\cancel{\sum_{j=1}^{n}(x_j-\overline{x})}} = 0$, *i.e.*, $\sum_{j=1}^{n} y_j - na = 0$ or: $a = \frac{1}{n}\sum_{j=1}^{n} y_j = \overline{y}$. -----(13.3.24)

Also: $\dfrac{\partial g}{\partial \beta_1} = -2\sum_{j=1}^{n}[y_j - \alpha - \beta_1(x_j-\overline{x})](x_j-\overline{x})$, which is zero when: $\sum_{j=1}^{n}[y_j - a - b_1(x_j-\overline{x})](x_j-\overline{x}) = 0$;

i.e., $\sum_{j=1}^{n}(y_j-\overline{y})(x_j-\overline{x}) - b_1\sum_{j=1}^{n}(x_j-\overline{x})^2 = 0$, or: $b_1 = \dfrac{\sum_{j=1}^{n}(y_j-\overline{y})(x_j-\overline{x})}{\sum_{j=1}^{n}(x_j-\overline{x})^2} \equiv \dfrac{\sum_{j=1}^{n}y_j(x_j-\overline{x})}{\sum_{j=1}^{n}(x_j-\overline{x})^2} \equiv \dfrac{\sum_{j=1}^{n}x_j(y_j-\overline{y})}{\sum_{j=1}^{n}(x_j-\overline{x})^2}$; -----(13.3.25)

the three expressions for b_1 are equivalent because: $\sum_{j=1}^{n}\overline{y}(x_j-\overline{x}) = \overline{y}\sum_{j=1}^{n}(x_j-\overline{x}) \equiv 0$; similarly $\sum_{j=1}^{n}\overline{x}(y_j-\overline{y}) \equiv 0$. -----(13.3.26)

Also: $b_0 = a - b_1\overline{x}$ and the estimated regression of $\mathbf{Y}$ on $\mathbf{X}$ is: ${}_{\text{reg}}\overline{y} = b_0 + b_1 x_j = a + b_1(x_j-\overline{x})$. -----(13.3.27)

To systematize the calculations of the values of b_0 and b_1, we define three '*sums of squares*':

$SS_{xy} = \sum_{j=1}^{n}(x_j-\overline{x})(y_j-\overline{y}) \equiv \sum_{j=1}^{n}(y_j-\overline{y})(x_j-\overline{x}) = \sum_{j=1}^{n}x_jy_j - (\sum_{j=1}^{n}x_j)(\sum_{j=1}^{n}y_j)/n = \sum_{j=1}^{n}x_jy_j - n\overline{x}\,\overline{y}$, -----(13.3.28)

$SS_x = \sum_{j=1}^{n}(x_j-\overline{x})^2 = \sum_{j=1}^{n}x_j^2 - (\sum_{j=1}^{n}x_j)^2/n = \sum_{j=1}^{n}x_j^2 - n\overline{x}^2$; -----(13.3.29)

$SS_y = \sum_{j=1}^{n}(y_j-\overline{y})^2 = \sum_{j=1}^{n}y_j^2 - (\sum_{j=1}^{n}y_j)^2/n = \sum_{j=1}^{n}y_j^2 - n\overline{y}^2$; -----(13.3.30)

we then write: $a = \overline{y}$, $b_1 = \dfrac{SS_{xy}}{SS_x}$, $b_0 = a - b_1\overline{x} = \overline{y} - b_1\overline{x}$. -----(13.3.31)

NOTES: 2. Throughout this Figure 13.3, (lower case) *roman* letters have been used to denote sample x and y values, because our primary concern is with the method of least squares; statistical considerations (such as distributional assumptions and the method of selecting the sample), needed to assess the *precision* of sample estimates of respondent population attributes, are discussed in Figure 13.7.

3. In the derivations given above of the results (13.3.24) and (13.3.25), note that the *variables* are α and β_1 and the *constants* are x and y (the data values) – this is the *opposite* of the notation we are accustomed to from calculus courses, for instance.

4. In the diagram at the right above, the point (x_j, y_j) is shown *above* the line and so its estimated residual is *positive*; if a point is *below* the line, its estimated residual is *negative*.

(*continued overleaf*)

NOTES: (cont.) 5. Of the three *algebraically* equivalent expressions (13.3.25) for b_1 (overleaf on page 13.19), the *statistically* useful one is the second – it showns that, in a linear regression context where the xs are considered *given* values, b_1 is a *linear combination* of the ys. This provides a basis for using the properties of random variables (from probability theory) to develop statistical theory for assessing the behaviour of B_1 or $\tilde{\beta}_1$ (the random variable corresponding the b_1 or $\hat{\beta}_0$) as an estimator of β_1 – for example, see Section 5 on pages 13.2 and 13.43 in Figure 13.7.

6. The three 'sums of squares' expressions (13.3.28) to (13.3.30) overleaf are useful for hand calculation and learning the ideas but they involve subtracting two numbers that may be large and nearly equal and so their difference may have unacceptable numerical inaccuracy. For automated calculation, particularly for large data sets, a more accurate process is to calculate the average, subtract it from each observation and then add the squares of these differences.

REFERENCE: Chan, T.F., Golub, G.H. and R.J. LeVeque: Algorithms for Computing the Sample Variance: Analysis and Recommendations. *The American Statistician* **37** (#3), 242-247 (1983). Stable URL: http://www.jstor.org/stable/2683386.

Exercises: 1. Establish the result: $\sum_{j=1}^{n}(x_j - \bar{x}) \equiv 0$, shown overleaf in (13.3.26) and used in both (13.3.24) and (13.3.25).

2. Starting from $g(\beta_0, \beta_1) = \sum_{j=1}^{n}[y_j - \beta_0 - \beta_1 x_j)]^2$ derive the expressions for b_0 and b_1 given overleaf on page 13.19 in equations (13.3.27) and (13.3.25) – note the *less* convenient algebra compared with starting from $g(\alpha, \beta_1)$.

3. Find appropriate *second* partial derivatives of $g(\alpha, \beta_1)$ and use them to show that the expressions (13.3.24) and (13.3.25) correspond to a *minimum* (not a *maximum*).

Example 13.3.1: For a bivariate sample data set, such as the one with five observations given at the right, it is convenient to proceed in the following steps to find the estimated regression of **Y** on **X**.

x	1	3	6	10	15
y	3	5	5	7	10

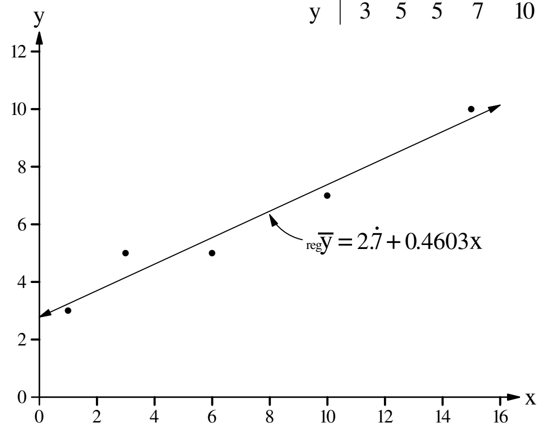
$\Sigma x_j = 35,\quad \Sigma x_j^2 = 371;\quad \Sigma x_j y_j = 268;$
$\Sigma y_j = 30,\quad \Sigma y_j^2 = 208;\quad (n = 5)$
$\bar{x} = 7,\quad \bar{y} = 6,$
$SS_{xy} = 268 - 35\times30/5 = 58,$
$SS_x = 371 - 35^2/5 = 126,$
$SS_y = 208 - 30^2/5 = 28;$

hence, the estimates of β_1 (the slope) and β_0 (the intercept) of the regression of **Y** on **X** are:

$b_1 = \frac{58}{126} = 0.460\,317\,460,\quad b_0 = 6 - b_1\times7 = 2.\dot{7};$

so the equation of the straight-line model is:

${}_{\text{reg}}\bar{y} = 2.\dot{7} + 0.4603x = 6 + 0.4603(x - 7).$



A scatter diagram of the data, with the estimated regression of **Y** on **X** superimposed on it, is given above at the right.

NOTES: 7. The four steps in the calculations in Example 13.3.1 are:

- find the five numerical data summaries Σx_j, Σx_j^2, Σy_j, Σy_j^2 and $\Sigma x_j y_j$;
- find the two averages $\bar{x}$ and $\bar{y}$;
- find the three 'sums of squares' SS_{xy}, SS_x and SS_y;
- find b_1 and b_0.

8. We can use the estimated regression of **Y** on **X** to 'predict' the estimated value of the average of **Y** for a given value of **X**; for example, when **X** = 8, we estimate ${}_{\text{reg}}\bar{\mathbf{Y}}(8)$ as: $\hat{\mu}_Y(8) = 2.\dot{7} + 0.4603\times8 \simeq 6.46.$

- If the straight-line model is a reasonable fit to the data (a matter taken up formally in Figure 13.5), 'predictions' with*in* the interval of the observed **X** values should be reasonably accurate; *out*side this interval, estimated values of the average of **Y** should be treated with caution, because there is no longer empirical evidence that the straight-line relationship holds between the average of **Y** and **X** – the *greater* the degree of extrapolation, the *greater* should be the caution.

5. Correlation

Recall from Figure 9.3 in Part 9 of STAT 220 that the *correlation coefficient* of a set of bivariate data, (x_j, y_j), j = 1, 2, ... n, is given by the expression (13.3.32) at the right; in the notation of *this* Figure, this expression can be written as (13.3.33).

$$r = \frac{\sum_{j=1}^{n}x_j y_j - (\sum_{j=1}^{n}x_j)(\sum_{j=1}^{n}y_j)/n}{\sqrt{[\sum_{j=1}^{n}x_j^2 - (\sum_{j=1}^{n}x_j)^2/n][\sum_{j=1}^{n}y_j^2 - (\sum_{j=1}^{n}y_j)^2/n]}} \quad \text{-----(13.3.32)}$$

$$r = \frac{SS_{xy}}{\sqrt{SS_x \times SS_y}} \quad \text{-----(13.3.33)}$$

As an illustration, for the data set of five observations in Example 13.3.1, we have: $r = \frac{58}{\sqrt{126\times28}} = 0.976\,481.$

(*continued*)

Figure 13.3. SIMPLE LINEAR REGRESSION: The Method of Least Squares (continued 2)

NOTES: 9. A value close to 1 for the correlation coefficent in this instance reminds us that, for small data sets, there can be a high correlation even when there is a noticable departure from perfect linearity of the points – look back also at Figures 9.4, 9.5 and 9.6 of the STAT 220 Course Materials.

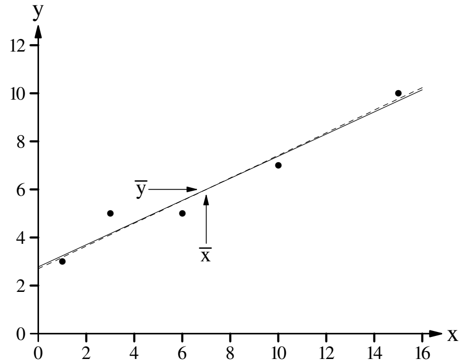
10. In the context of straight lines fitted to bivariate data by the method of least squares, it seems to be more usual to refer to the 'correlation' rather than the 'correlation coefficient'.

11. We meet r again, although with a somewhat different method of calculation, in on page 13.32 in Figure 13.5.

12. Using expressions (13.3.31) overleaf on page 3.19 for b_1 and (13.3.33) above for r, we obtain equation (13.3.34) at the right; although this expression is not usually a useful way of *calculating* b_1, it shows that the slope of the estimated regression of $\mathbf{Y}$ on $\mathbf{X}$ is the correlation times the ratio of the standard deviations of y and x – the line whose slope is this ratio of standard deviations is sometimes called *the s.d. line*.

$$b_1 = \frac{SS_{xy}}{SS_x} = \frac{SS_{xy}}{\sqrt{SS_x \times SS_y}}\sqrt{\frac{SS_y/(n-1)}{SS_x/(n-1)}} = r \times \frac{s_y}{s_x}; \qquad \text{-----(13.3.34)}$$

- For the data set of five observations in Example 13.3.1, the scatter diagram, with the estimated regression of $\mathbf{Y}$ on $\mathbf{X}$ *and* the s.d. line (dashed) superimposed on it, is shown at the right – the two lines intersect at $(\overline{x}, \overline{y})$, called the *point of averages*.



6. Finding the Estimated Regression of X on Y

Another line of 'best' fit to a set of bivariate sample data is obtained by minimizing the sum of the squared *horizontal* distances of the points from the line, to give *the estimated regression of* $\mathbf{X}$ *on* $\mathbf{Y}$. The theory associated with this case is the *same* as that for the estimated regression of $\mathbf{Y}$ on $\mathbf{X}$ *except* that x and y are interchanged in each expression; we therefore obtain the results given at the right, which correspond to equations (13.3.12), (13.3.13), (13.3.11) and (13.3.31).

$$\mu_x = \gamma_0 + \gamma_1 y \qquad \text{-----(13.3.35)}$$

$$\mu_x = \delta + \gamma_1(y - \overline{y}) \qquad \text{-----(13.3.36)}$$

$$\delta = \gamma_0 + \gamma_1 \overline{y} \qquad \text{-----(13.3.37)}$$

$$\left.\begin{array}{l} d = \overline{x}, \qquad g_1 = \dfrac{SS_{xy}}{SS_y} \\ g_0 = d - g_1\overline{y} = \overline{x} - g_1\overline{y} \end{array}\right\} \qquad \text{-----(13.3.38)}$$

The expressions (13.3.32) and (13.3.33) for r on the facing page 13.20 are *symmetrical* in x and y and so do *not* change for the regression of $\mathbf{X}$ on $\mathbf{Y}$ – looked at another way, the correlation of y and x is the *same* as the correlation of x and y.

Corresponding to equation (13.3.34), we have: $g_1 = r \times \frac{s_x}{s_y}$, so that: $\frac{1}{g_1} = \frac{1}{r} \times \frac{s_y}{s_x}$; -----(13.3.39)

i.e., when we express the equation of the estimated regression of $\mathbf{X}$ on $\mathbf{Y}$ in the form $y = f(x)$ instead of $x = g(y)$, its slope is *greater* than that of the s.d. line by a factor of 1/r; this means that a scatter diagram containing the three lines would show the regression of $\mathbf{Y}$ on $\mathbf{X}$ with a slope *less* than that of the s.d. line by a factor of r and the regression of $\mathbf{X}$ on $\mathbf{Y}$ with a slope *greater* than that of the s.d. line by a factor of 1/r.

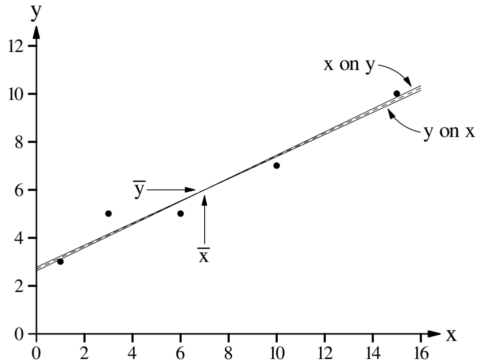
As an illustration, for the data set of five observations in Example 13.3.1, we have:

$$g_1 = \tfrac{58}{28} = 2.07\,\dot{1}\dot{4}\dot{2}\,\dot{8}\dot{5}\dot{7}, \quad g_0 = 7 - g_1 \times 6 = -5.\dot{4}\dot{2}\dot{8}\,\dot{5}\dot{7}\dot{1},$$

so that the equation is: $_{\text{reg}}\overline{x} = -5.4286 + 2.0714y$
$= 7 + 2.0714(y - 6),$

equivalent to: $_{\text{reg}}\overline{y} = 2.6207 + 0.4828x.$

A scatter diagram of the data, with the estimated regression of $\mathbf{Y}$ on $\mathbf{X}$ (labelled 'y on x'), the s.d. line (dashed) and the estimated regression of $\mathbf{X}$ on $\mathbf{Y}$ (labelled 'x on y') superimposed on it, is given at the right; the three lines intersect at $(\overline{x}, \overline{y})$, the point of averages.



NOTES: 13. Although the two estimated regression lines are very similar in this instance (because r is close to 1), they are *not* the same line; this reminds us that, in general:

- the usual designation of the regression of $\mathbf{Y}$ on $\mathbf{X}$ (based on minimizing *squares* of *vertical* distances) as the line of *best* fit to a set of bivariate data is contingent on the matter of interest being the behaviour of the *average* of $\mathbf{Y}$ *conditional* on the value of $\mathbf{X}$;
- in real situations, it is important to decide correctly which variable should be identified as the *response* variate and which as the *explanatory* variate.

(*continued overleaf*)

NOTES: **(cont.)** 14. The s.d. line has the following properties of interest.

- It is the line of 'best' fit obtained by minimizing the squares of the *perpendicular* distances of the points from the line – the s.d. line thus involves **X** and **Y** *equivalently*, rather than (as in the two regressions) considering one variable as a *response* to the other.
- If the bivariate normal distribution is an appropriate description of the *joint* distribution of **Y** and **X**, the cloud of points on the scatter diagram will lie mainly in an elliptical region, of which the s.d. line is the *major axis*;
  - For such a roughly elliptical cloud of points on a scatter diagram, the *s.d. line* has more intuitive appeal *visually* as the line of 'best' fit than does the regression of **Y** on **X** (or of **X** on **Y**).
  - If a scatter diagram is drawn with one standard deviation of **X** and of **Y** occupying the *same* distance on its vertical and horizontal axes:
    ○ the s.d. line has slope 1, ○ the regression of **Y** on **X** has slope r, ○ the regression of **X** on **Y** has slope 1/r.
  - When there is no linear relationship between **X** and **Y** (*i.e.*, when r = 0), the points of a bivariate normal scatter diagram lie in a roughly *circular* (not elliptical) region and the regression of **Y** on **X** s a *horizontal* line.
  - When there is a *perfect* positive or negative correlation (*i.e.*, when the magnitude of r is 1), the regression of **Y** on **X**, the regression of **X** on **Y**, and the s.d. line are *coincident*.

15. When investigating **X**-**Y** relationships in the context of this Figure 13.3, '*strength of relationship*' involves *two* matters:

- how *well* the model *fits* the data – we usually regard an **X**-**Y** relationship as 'stronger' when the points generally lie *closer* to the line (when the *correlation* is closer to 1 in magnitude) [but recall Note 7 near the bottom of page 13.20];
- the magnitude of (the estimate of) β_1 – we usually regard a *steeper* slope as indicating a 'stronger' **X**-**Y** relationship.

How correlation and slope are related is discussed in Section 8 in Figure 9.3 of the STAT 220 Course Materials.

7. The Regression Effect and the Regression Fallacy

The data set which first gave rise to Galton's use of the term 'regression' had three characteristics:

(1) one variable (children's height – our **Y**) was clearly a *response* to the other 'explanatory' variable (parents' height – our **X**); it it therefore seemed 'natural' to examine the behaviour of **Y** – for example, its average – *conditional* on a succession of single (or, in prac- tice, narrow intervals of) **X** values;

(2) the average expected for **Y**, when **X** was in a narrow interval, was 'known' – in Galton's case, children were expected to deviate from their overall average by the *same* amount (*i.e.*, the *same* number of standard deviations) as their parents;

(3) a scatter diagram of the population **Y** and **X** values showed a roughly elliptical cloud of points so that a bivariate normal distribution was a reasonable model for the *joint* distribution of **Y** and **X** in the population; the *marginal* normal models for the histograms of **Y** and **X** individually could be considered to have the *same* mean and the *same* standard deviation – *i.e.*, the histograms of parents' heights and children's heights were essentially identical.

What Galton found, apparently to his surprise, was that, on average, children deviate from average *less* than their parents – parents taller than average tend to have children taller than average but not by as much (in terms of standard deviation) as themselves and, similarly, parents shorter than average tend to have children who are shorter than average but again by less than their parents (in terms of standard deviation). This finding was Galton's 'regression to mediocrity' (a somewhat unfortunate choice of words in light of the current negative connotation of 'mediocrity'), and it is the basis of what is now called the *regression effect.* Unfortunately, there is more to a proper understanding of this phenomenon than its simple beginning might suggest; a key distinction is between what the regression effect entails for groups of elements and for individuals.

We can pursue this discussion using the father-son and mother-daughter height data published by Pearson in 1903, even though the individual data apparently no longer exist. Relevent summaries for the 1,078 father-son pairs are shown on the facing page 13.23.

from our perspective, Galton's finding is a reflection of the fact that, when there is variation in **Y** for given **X**, the regression of **Y** on **X** [which is based on regarding **Y** as *conditional* on **X** and, hence, minimizing the sum of the squared *vertical* distances of the points from the line] has a slope only r times the slope of the s.d. line. This is now known as the *regression effect*, and inferring in a 'test-retest' situation that its occur- rence demonstrates the effect of an (important) explanatory variate is sometimes called the *regression fallacy.*

Taken in isolation, 'regression to mediocrity' implies that the distribution of sons' heights (response variate **Y**) has a *smaller* standard deviation than the distribution of fathers' heights (explanatory variate **X**), but this is *not* the case:

∗ the two standard deviations have essentiually the *same* value (about 2.5 inches); AND:

∗ the *horizontal* strip of the scatter diagram containing sons' heights between (say) 71½ and 72½ inches has much the *same* distribution as the *vertical* strip containing fathers in this height range; however, the majority of the former are *not* the offspring of the latter.

Thus, Galton's 'regression to mediocrity' is only a partial (and possibly misleading) Answer from examining the scatter diagram to the Question about the father-son height relationship; it reminds us that the way a Question is posed [as in the 'natural' approach of (1) above] may compromise the Answer.

continued)

Figure 13.3. SIMPLE LINEAR REGRESSION: The Method of Least Squares (continued 3)

NOTES: 16. Attaching the name *regression effect* to the fact that the regression of **Y** on **X** has a only r times the slope of the s.d. line can obscure two points:

- when there is *variation* in the values of **Y** for a given value of **X**, the less deviant average of **Y** from its average than of **X** from its average is an *inherent* property under the straight-line model and can be demonstated algebraically;
- the name should *not* be taken as implying that 'regression' produces a *change* in the variation in **Y** – for example, it is not *de*creased – and this can again be demonstrated algebraically.

17. Although the *marginal* normal models for the histograms of **Y** and **X** in Galton's data could be considered to have the *same* mean and the *same* standard deviation, this is *not* required for the regression effect to occur; the average height of children may *change* over generations – for instance, it may *in*crease under the influence of improved nutrition.

One rationalization of the regression effect in data sets with characteristics like Galton's is the following. For populations that maintain themselves by reproduction, it is presumably necessary to keep the *variation* in a trait like body size relatively *stable* over many generations; this could be accomplished by either:

- having offspring be *exact* copies of their parents in the trait – all points would then lie on the *s.d. line* and there would be *no* mechanism for variation to act from generation to generation; **OR**
- allowing for *variation* in the trait among offspring of parents who have the *same* value of the trait, and maintaining stability in the distribution of the trait from generation to generation by the regression effect, with its compensatory movements towards and away from the average among offspring.

There are presumably evolutionary advantages in the *usual* state of affairs in nature being the *second* possibility, with its role for variation across generations – *un*like the first possibility, there is *change* from generation to generation as to whose offspring occupy the categories defined by different magnitudes of the trait, despite the stability of the variation in **Y**.