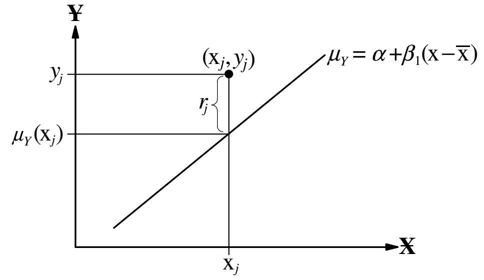
Figure 13.5. SIMPLE LINEAR REGRESSION: Assessing the Fit of a Regression Model

As soon a regression model has been fitted to a data set by the method of least squares, our concern should be to assess *how well* the model fits the data; this matter is taken up in this Figure 13.5. We discuss three methods of assessment:

- *visual assessment* of the scatter diagram with the (estimated) regression line superimposed on it;
- how close r^2, called the *coefficent of determination*, is to 1;
- the value of the *F*-ratio, which we obtain from the analysis of variance (ANOVA) table for the data.

1. Residuals and Estimated Residuals



Suppose we have a *sample* of bivariate values (x_j, y_j), $j = 1, 2, \ldots\ldots n$, from which we wish to *estimate* the regression of $\mathbf{Y}$ on $\mathbf{X}$;

we write the regression line as: $\mu_Y = \beta_0 + \beta_1 x$, -----(13.5.1)

or, in *centred* form, as: $\mu_Y = \alpha + \beta_1(x - \bar{x})$, -----(13.5.2)

where: $\beta_0 = \alpha - \beta_1\bar{x}$. -----(13.5.3)

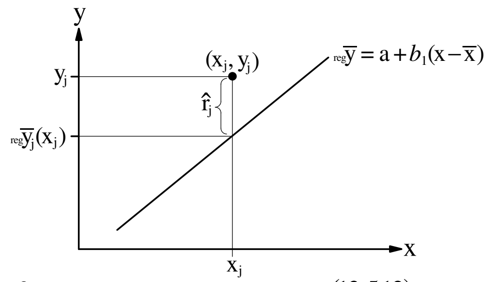
Consider the regression of $\mathbf{Y}$ on $\mathbf{X}$ and a typical point with coordinate values (x_j, y_j), as shown in the diagram at the right;
the point *on* the line at the same value of $\mathbf{X}$ is $[x_j, \mu_Y(x_j)]$;
the vertical distance of the point from the line is the *residual* $r_j = y_j - \mu_Y(x_j)$; -----(13.5.4)
our *estimated* regression of $\mathbf{Y}$ on $\mathbf{X}$ (or the least squares line) is the one which *minimizes the sum of the squares of the* ***sample*** *of* n *residuals.*

As in Section 4 of Figure 13.3, we write: $g(\alpha, \beta_1) = \sum_{j=1}^{n} r_j^2 = \sum_{j=1}^{n}[y_j - \mu_Y(x_j)]^2 = \sum_{j=1}^{n}[y_j - \alpha - \beta_1(x_j - \bar{x})]^2$; -----(13.5.5)

to find [based on *data* (x_j, y_j), $j = 1, 2, \ldots\ldots n$] 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;

i.e.: $\frac{\partial g}{\partial \alpha} = -2\sum_{j=1}^{n}[y_j - \alpha - \beta_1(x_j - \bar{x})]$ so that: $\sum_{j=1}^{n}[y_j - a - b_1(x_j - \bar{x})] = 0$, -----(13.5.6)

and: $\frac{\partial g}{\partial \beta_1} = -2\sum_{j=1}^{n}[y_j - \alpha - \beta_1(x_j - \bar{x})](x_j - \bar{x})$, $\sum_{j=1}^{n}[y_j - a - b_1(x_j - \bar{x})](x_j - \bar{x}) = 0$. -----(13.5.7)



The *estimated* regression line is then: ${}_{\text{reg}}\bar{y} = b_0 + b_1 x$, -----(13.5.8)

or, in *centred* form: ${}_{\text{reg}}\bar{y} = a + b_1(x - \bar{x})$, -----(13.5.9)

where: $b_0 = a - b_1\bar{x}$; ----(13.5.10)

this line, and a typical *data* point (x_j, y_j), are shown in the diagram at the right;
the point *on* the line at the same value of x is $[x_j, \hat{\mu}_Y(x_j)] \equiv [x_j, {}_{\text{reg}}\bar{y}_j(x_j)] \equiv (x_j, {}_{\text{reg}}\bar{y}_j)$;
the vertical distance of the point from
the line is the *estimated residual* $\hat{r}_j = y_j - {}_{\text{reg}}\bar{y}_j = a - b_1(x_j - \bar{x})$. ----(13.5.11)

We now recognize that setting the two partial derivatives to zero in (13.5.6) and (13.5.7) imposes on the estimated residuals the two *constraints* (13.5.12) and (13.5.13) given at the right; the second constraint implies the x_js are *known*.

$\sum_{j=1}^{n}\hat{r}_j = 0$, -----(13.5.12)

$\sum_{j=1}^{n}\hat{r}_j(x_j - \bar{x}) = \sum_{j=1}^{n}\hat{r}_j x_j = 0$. -----(13.5.13)

NOTES: 1. The second form of the constraint (13.5.13) follows from its first form because of the first constraint (13.5.12) and the fact that $\bar{x}$ is not subscripted and so can be taken outside the sigma sign; equation (13.5.13) assumes that the x_js are *not* all equal (in which case b_1 is undefined).

2. The foregoing discussion deals with the usual situation in practice of a *sample* of bivariate data (x_j, y_j), $j = 1, 2, \ldots\ldots n$; if the data were instead a *census* $(\mathbf{X}_i, \mathbf{Y}_i)$, $i = 1, 2, \ldots\ldots, \mathbf{N}$, of a respondent population of $\mathbf{N}$ elements, the upper limit of the sums would be $\mathbf{N}$ instead of n and we would obtain the *values* of β_0 and β_1 (representing the respondent population attributes $\mathbf{B}_0$ and $\mathbf{B}_1$), *not* their estimates, b_0 and b_1. Thus, like the *estimated* residuals (13.5.11), the *residuals* r_j in the model [equation (13.5.4) above] and $\mathbf{R}_i$ in the respondent population [equations (13.3.1) and (13.3.5) on page 13.17] also satisfy two constraints [corresponding to equations (13.5.12) and (13.5.13)].

- In the rare instances where census data for the respondent population *are* available, no *probabilistic assumptions* (as in Figure 13.7) are needed because no *estimating* of respondent population attributes is involved – census data provide values of β_0 and β_1 (and $\mathbf{B}_0$ and $\mathbf{B}_1$) free from sample (but *not* measurement) error.

The importance of the two constraints (13.5.12) and (13.5.13) to the discussion of this Figure 13.5 is that, in constructing the ANOVA table [first given at the bottom of the fourth side (page 13.32) of the Figure], when we find the *average* of the squared estimated residuals, we divide by $n - 2$ – because of the two constraints, only $n - 2$ of the squares of the n estimated residuals are *functionally independent*, which is the relevant consideration in setting the value of the divisor for finding their average.

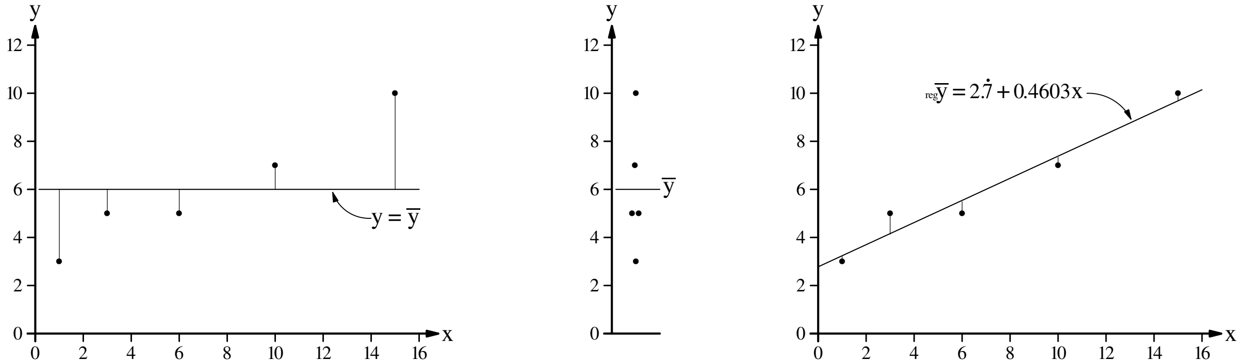
 (*continued overleaf*)

- This is the *same* idea as dividing by $n-1$ (or $\mathbf{N}-1$) to obtain the *average* squared deviation from the average for the standard deviation of a set of *uni*variate data – the *one* constraint in calculating the s.d. is that the observations have average $\overline{y}$ (or $\overline{\mathbf{Y}}$) which means that only $n-1$ of the y_js (or $\mathbf{N}-1$ of the $\mathbf{Y}_i$s) are functionally independent.

The quantites that go into the two calculations just described are compared pictorially below for the same data set (given at the right) of five observations used in Figures 13.1 and 13.3.

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

- The five vertical lines in the *left*-hand diagram represent the *deviations* of the y_js from $\overline{y}$ – they are the basis of the calculation of the *standard deviation* of y;
- the *middle* diagram is a 'collapsed' version of the left-hand diagram, in which the x values have been ignored;
- the five vertical lines in the *right*-hand diagram represent the *estimated residuals* $\hat{r}_j$, the vertical distances of the points from the *estimated* regression of $\mathbf{Y}$ on $\mathbf{X}$ – they are the basis of the calculation of the *estimated residual mean square*, for example in the ANOVA Table 13.5.3 at the bottom of the fourth side (page 13.32) of this Figure 13.5 – see also Note 3 below.
  - As we will see in Figure 13.7, the *square root* of the estimated residual mean square estimates the standard deviation of the points about the estimated regression of $\mathbf{Y}$ on $\mathbf{X}$.



It is unfortunate that, in the usual statistical terminology, the basically similar quantities represented by the vertical lines in the left- and right-hand diagrams above are given *different* names – *deviations* and *estimated residuals.* Likewise, Notes 3, 4 and 5 below (and Note 13 on page 13.34) discuss unevocative historical choices for terminology that is now used so widely it is not feasible to rectify these unhelpful choices.

NOTES: 3. The terminology in regression associated with the squared estimated residuals is summarized in the first two lines of Table 13.5.1 at the right, which also includes in its last two lines the *same* matters for the y values alone to show the *new* ideas in relation to familiar ones.

Table 13.5.1: SYMBOLS	**NAME**	**ABBRE-VIATION**
$\sum_{j=1}^{n} \hat{r}_j^2 = \sum_{j=1}^{n} (y_j - {}_{\text{reg}}\overline{y}_j)^2$	Estimated residual sum of squares	SSE
$\sum_{j=1}^{n} \hat{r}_j^2/(n-2)$	Estimated residual mean square	MSE
$SS_y = \sum_{j=1}^{n} (y_j - \overline{y})^2$	Total sum of squares	SST
$s_y^2 = SS_y/(n-1)$	Square of the standard deviation of the y_js	$[s.d.(y)]^2$

4. In the context of analysis of variance, the divisors (like $n-2$ and $n-1$) used to convert *sums* of squares to *mean* squares are referred to as *degrees of freedom* (abbreviated *df*).
   - Because these Course Materials distinguish *averages* for *data* from *means* for *random variables* (*e.g.*, recall Note 1 on page 5.4 of Figure 5.1 in STAT 220), we would prefer *average square* to *mean square* but, unfortunately, the latter terminology is used almost universally elsewhere and so it has also been used here.

5. In the text and elsewhere (including regression software packages), our *residuals* may be called *errors* and our *estimated* residuals may be called *residuals*; our terminology and the text's are compared in the Table 13.5.2 above. The Course Materials use *error* to mean the *difference of an estimate of a population attribute from its true value* – we must not confuse *either* of these technical meanings with the common English meaning of *mistake*.

Table 13.5.2: Course Materials		**Text 3rd Edition**		
Residual	r_j	Deviation (or error)	ε_i	p. 665
Estimated residual	$\hat{r}_j$	Residual	e_i	p. 666
Estimated residual sum of squares	SSE	Sum of squares for error	SSE	p. 682
Estimated residual mean square	MSE	Mean square for error	MSE	p. 683

   - 'Error' also occurs in *standard error* (meaning the *estimated* standard deviation of a parameter estimate), which is commonly used in, for instance, the output of regression software packages (see text pages 668, 675, 677 and 685).

Against the background of the foregoing discussion, we now take up the main concern of this Figure 13.5 – assessing fit.

Figure 13.5. SIMPLE LINEAR REGRESSION: Assessing the Fit of a ..... Model (continued 1)

2. Fit of the Model 1: Visual Assessment

Our *primary* method of assessing the fit of a regression model to the data is *visual* inspection of the scatter diagram with the estimated regression of $\mathbf{Y}$ on $\mathbf{X}$ superimposed on it; we look for the following:

- the points should lie (reasonably) *close* to the fitted line – we may be able to tolerate a *small proportion* (typically one or two points) with larger estimated residuals, but it is then important to follow up on such observations to try to find out *why* they are more deviant than the other observations;
  - *if* the reason(s) for one or two deviant observations can be found, it *may* be useful to recalculate the estimated regression of $\mathbf{Y}$ on $\mathbf{X}$ with the more deviant observation(s) omitted (see also Data set 3 on pages 13.35 and 13.36);
- there should be no obvious *pattern* in the way the observations fall on the two sides of the fitted line;
- there should not be any systematic *change* in the magnitude of the estimated residuals with increasing values of $\mathbf{X}$ – this is consistent with the assumption of *constant standard deviation about the regression line* which we come to in Figure 13.7.

For the second and third of these criteria, it is important to recognize that human judgement of patterns in, and variation across, a scatter diagram are generally somewhat *subjective* and, particularly with diagrams containing fewer than 20 or so points, we must exercise due caution in claiming too definite an interpretation of visual appearance.

NOTE: 6. Regression software will assist in visual assessment of fit by producing *estimated residual plots*, in which the estimated residuals are plotted against their $\mathbf{X}$ values. Such a plot is actually just the scatter diagram and superimposed estimated regression line *rotated* until the latter becomes the horizontal axis of the diagram; estimated residual plots are therefore assessed by the *same* three criteria given above.

- Elsewhere, you are likely to find estimated residual plots called just *residual plots.*

3. Partitioning SS_y and the Model Sum of Squares, SSM

We approach this partitioning *in*directly by algrebraic expansion of the *estimated residual* sum of squares into the *difference* of two parts, which is then rearranged into a *sum*:

$$\begin{aligned}
\text{SSE} \equiv \sum_{j=1}^{n}\hat{r}_j^2 = \sum_{j=1}^{n}(y_j - {}_{\text{reg}}\bar{y}_j)^2 &= \sum_{j=1}^{n}[y_j - \bar{y} - b_1(x_j - \bar{x})]^2 \qquad \text{[using (13.5.11) and setting } a = \bar{y}\text{]}\\
&= \sum_{j=1}^{n}(y_j-\bar{y})^2 - 2b_1\sum_{j=1}^{n}(x_j-\bar{x})(y_j-\bar{y}) + b_1^2\sum_{j=1}^{n}(x_j-\bar{x})^2\\
&= \sum_{j=1}^{n}(y_j-\bar{y})^2 - b_1\sum_{j=1}^{n}(x_j-\bar{x})(y_j-\bar{y}) \qquad \left[\text{because: } b_1 = \frac{\sum_{j=1}^{n}(y_j-\bar{y})(x_j-\bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2}\right], \qquad \text{-----(13.5.14)}
\end{aligned}$$

i.e., $\text{SSE} = \text{SS}_y - b_1\text{SS}_{xy}$;

hence: $\text{SS}_y = b_1\text{SS}_{xy} + \text{SSE}$ or: $\text{SST} = \text{SSM} + \text{SSE}$; -----(13.5.15)

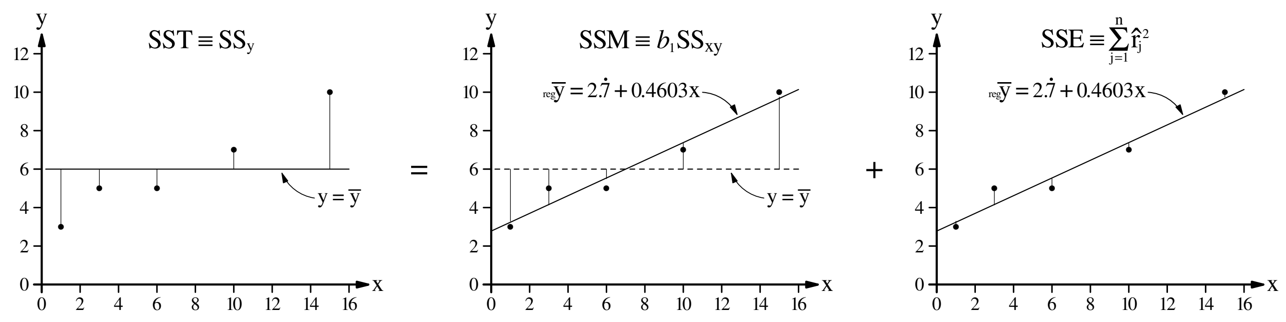
i.e., in the rearranged form (13.5.15) of the expression (13.5.14), the *total* sum of squares, SS_y, has been *partitioned* into (the *sum* of) two parts:

- $b_1 SS_{xy}$, which is called the *model sum of squares* and is abbreviated SSM (*not* previously discussed in this Figure 13.5),
- SSE, the *estimated residual sum of squares* (already discussed).

We can show this partitioning of SS_y graphically for the the same data set (given at the right) of five observations used on the facing page 13.30 and in Figures 13.1 and 13.3:

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

- the sum of the squares of the five vertical lines on the left-hand diagram is $\text{SST} \equiv \text{SS}_y = \sum_{j=1}^{n}(y_j-\bar{y})^2$ – it quantifies (on a squared scale) the variation of the y_js about their average, $\bar{y}$;
- the sum of the squares of the five vertical lines on the middle diagram is $\text{SSM} \equiv b_1\text{SS}_{xy} = \sum_{j=1}^{n}({}_{\text{reg}}\bar{y}_j - \bar{y})^2$ – it quantifies (on a squared scale) the variation of the ${}_{\text{reg}}\bar{y}_j$s about *their* average, $\bar{y}$;
- the sum of the squares of the five vertical lines on the right-hand diagram is $\text{SSE} \equiv \sum_{j=1}^{n}\hat{r}_j^2 = \sum_{j=1}^{n}(y_j - {}_{\text{reg}}\bar{y}_j)^2$ – it quantifies (on a squared scale) the variation of the y_js about the ${}_{\text{reg}}\bar{y}_j$s [*i.e.*, about the estimated regression line].



(*continued overleaf*)

The partioning (13.5.15) of the *total* sum of squares, SS_y, into two parts is the basis of the methods of assessing the fit of a regression model discussed in the next two sections – the *coefficient of determination* and *analysis of variance*.

Exercises: 1. Establish the result: $b_1 SS_{xy} = b_1^2 SS_x$, which is involved in simplifying to *two* terms on the right-hand side of equation (13.5.14) from *three* terms in the previous line.

2. Show that: $SSM \equiv b_1 SS_{xy} = \sum_{j=1}^{n}({}_{reg}\bar{y}_j - \bar{y})^2$ by writing $\sum_{j=1}^{n}(y_j - \bar{y})^2$ as: $\sum_{j=1}^{n}[(y_j - {}_{reg}\bar{y}_j) + ({}_{reg}\bar{y}_j - \bar{y})]^2$ and multiplying out the square; you will also need to use the results (13.5.11), (13.5.12) and (13.5.13) to show that the cross-product term is *zero* in this instance (*un*like the situation in the expansion of SSE overleaf on page 13.31).

3. Show that *each* of the three sums: $\sum_{j=1}^{n}(y_j - \bar{y})$, $\sum_{j=1}^{n}({}_{reg}\bar{y}_j - \bar{y})$ and $\sum_{j=1}^{n}(y_j - {}_{reg}\bar{y}_j)$ is identically *zero.*

4. Fit of the Model 2: The Coefficient of Determination, r^2

The partition of the total sum of squares, SST = SSM + SSE, and its graphical illustration at the bottom of page 13.31 overleaf, lead to the two *extreme* cases for the fit of a regression model to a set of bivariate data;

- the most *successful* regression is when the points lie *exactly on* the line – in this situation, SSE is *zero* and so SSM = SST;
- the most *un*successful regression is when there is *no relationship* between $\mathbf{Y}$ and $\mathbf{X}$ and the ‘regression’ line is the *horizontal* line $y = \bar{y}$ – in this situation, SSE = SST and so SSM is *zero.*

Thus, the model sum of squares, SSM, takes values between SST in a completely *successful* regression and 0 in a completely *un*successful one; most real situations lie somewhere between these extremes.

These ideas can be *quantified* by calculating the quotient SSM/SST, called the *coefficent of determination* and denoted r^2:

$$r^2 = \frac{SSM}{SST} = \frac{b_1 SS_{xy}}{SS_y} = \frac{SS_{xy}^2}{SS_x \times SS_y} \qquad \text{[because: } b_1 = \frac{SS_{xy}}{SS_x}\text{]} \qquad (0 \le r^2 \le 1); \qquad \text{-----(13.5.16)}$$

the two extreme cases correspond to:

- $r^2 = 1$ (completely successful),
- $r^2 = 0$ (completely *un*successful).

When using r^2 to assess fit, the closer its value is to 1 (or the closer $100r^2$ is to 100%), the more ‘successful’ the regression.

For the data set of 5 observations used overleaf on page 13.31, $SS_{xy} = 58$, $SS_x = 126$ and $SS_y = 28$ so that $r^2 = 0.953\,515$; we would then comment that, because over 95% of the variation in the y_js about their average has been accounted for by the estimated regression of $\mathbf{Y}$ on $\mathbf{X}$, the value of r^2 indicates a *successful* regression.

NOTES: 7. The interpretation of r^2 = SSM/SST as the *fraction of the variation in the* y_js *about their average accounted for by the regression* is apparent from the diagrams overleaf at the bottom of page 13.31 – SS_y is the variation of the y_js about $\bar{y}$ and SSM is the variation of the ${}_{reg}\bar{y}_j$s about $\bar{y}$.

- Elsewhere, you may find the phrase *accounted for* in this interpretation of r^2 replaced by *explained by*; the latter is generally *not* used in these Course Materials because of it may be taken as implying a *causal* relationship between $\mathbf{Y}$ and $\mathbf{X}$, when there may be little or no evidence that (or interest in whether) this is the actual state of affairs.

8. We now see the reason for the notation r^2 for the coefficent of determination – equation (13.5.16) shows it is the *square* of the *correlation coefficient* r, where $-1 \le r \le 1$.

- When taking the square root of the coefficient of determination to find the value of the correlation, its *sign* is that of SS_{xy} (and of the *slope* of the estimated regression line).
  - When r is *negative*, it is easy to forget, in the process of taking the square root of r^2, to include the minus sign in the value of r.

9. Elsewhere, you may see the *model* sum of squares called the *regression* sum of squares and abbreviated SSR.

- It can be a source of confusion in regression acronyms that, in English, ‘residual’ and ‘regression’ both start with R and ‘model’ and ‘mean’ both begin with M.

5. Fit of the Model 3: Analysis of Variance, ANOVA

The ANOVA table in simple linear regression is mainly a systematic tabular presentation of information already given in this Figure 13.5; its structure is as follows:

Table 13.5.3:

SOURCE	SUM of SQUARES	*Df*	MEAN SQUARE	*F*-RATIO
Model	$SSM = b_1 SS_{xy}$	1	$MSM = SSM \div 1$	$\frac{MSM}{MSE} = \frac{MSM}{\hat{\sigma}^2}$
Estimated residual	$SSE = SS_y - SSM$	$n-2$	$MSE = SSE \div (n-2)$	
Total	SS_y	$n-1$	$[s_y^2 = SS_y \div (n-1)]$	-----

(*continued*)

Figure 13.5. SIMPLE LINEAR REGRESSION: Assessing the Fit of a ..... Model (continued 2)

Noteworthy features of the table are:

Column 1: The conventional heading is 'SOURCE' and the three rows are 'Model', 'Estimated residual' and 'Total'.

Column 2: The three sums of squares are calculated in the order: SS_y, SSM (as $b_1 SS_{xy}$) and SSE (found by *difference*).

- The calculation of SSE may involve subtracting two *large* and nearly *equal* values; it is important to do the arithmetic in a way that does *not* introduce significant *rounding* inaccuracy into the value of SSE, which is needed throughout subsequent calculations for interval estimating of model parameters and responses, as discussed in Figure 13.7.

Column 3: The *degrees of freedom* (abbreviated *Df*) are the divisors used to convert *sums* of squares to *mean* squares; they add down the table, reflecting the fact that the model has only *one* degree of freedom in simple linear regression.

- *One* degree of freedom for SSM means that, if *one* of its n terms is given, all the other $n-1$ are determined or, expressed another way, the n values satisfy $n-1$ constraints – the reason is the straight line that determines the ${}_{\text{reg}}\overline{y}_j$s is defined by the point of averages $(\overline{x}, \overline{y})$ and any *one* other point $(x_j, {}_{\text{reg}}\overline{y}_j)$.

Column 4: The mean (really average) squares are calculated by dividing the sums of squares by their degrees of freedom; with only *one* degree of freedom for SSM in simple linear regression, the *value* of SSM and MSM is the *same*;

- s_y^2 is usually *not* part of ANOVA tables given elsewhere; it is included in these Course Materials to remind us that mean squares, despite their name, are essentially *squared standard deviations.*

Column 5: The *new* idea in the ANOVA table is the *F*-ratio, which can be used to assess the strength of the evidence in the sample data against the hypothesis H: $\beta_1 = 0$ – the *larger* the value of the ratio, the *stronger* the evidence of a non-zero slope and the more 'successful' the regression.

The background to this *F*-test (which is analogous to the z-, t- and χ^2-tests discussed in Part 12) is as follows:

- the bivariate data (x_j, y_j), $j = 1, 2, ... n$, are assumed to come from a *sample* obtained from the respondent population by *equiprobable* (also called simple random) selecting;
  - to estimate the *slope* β_1 efficiently, this equiprobable selecting should be done *conditional* on a suitable set of **X** values, because we want to *spread out* the **X** values over their possible range when estimating β_1;
  - to estimate the respondent population *correlation coefficent* **R**, the sample needs to be obtained in the more *usual* way where the units (and, hence, their **Y** *and* **X** values) are selected equiprobably from the *whole* respondent population (*i.e.*, with*out* conditioning on **X**);
- conditional on each value of **X** in the sample, the random variable Y representing **Y** is assumed to have a normal distribution with mean $\beta_0 + \beta_1 x$ and standard deviation σ (which is assumed to be *constant*, that is, *not* to depend on **X** – see Figure 13.7);
- the distributions of $Y_j | x_j$ and $Y_k | x_k$ must be *probabilistically independent* for $j \neq k$.

For the same data set (given at the right) of five observations used previously in this Figure 13.5 and in Figures 13.1 and 13.3, the ANOVA table is:

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

Table 13.5.4:

SOURCE	SUM of SQUARES	*Df*	MEAN SQUARE	*F*-RATIO
Model	$b_1 SS_{xy}$ = 26.698 413	1	MSM = 26.698 413	$\frac{\text{MSM}}{\text{MSE}} = \frac{\text{MSM}}{\hat{\sigma}^2} \simeq 61.54$
Estimated residual	SSE = 1.301 587	3	MSE = 0.433 862	
Total	SS_y = 28	4	(s_y^2 = 7)	-----

We would then comment that, because the value of the *F*-ratio is *high* (appreciably greater than 10), there is *highly statistically significant* evidence against H: $\beta_1 = 0$, indicating a 'successful' regression.

NOTES: 10. To assess more quantitatively the *strength* of the evidence against H: $\beta_1 = 0$ provided by the value of the *F*-ratio, we use Table 13.13; the relevant entries are for the $F_{1,3}$ distribution, found in the first column on page 1 of 6 (page 13.61). The (numerator and denominator) degrees of freedom for the F distribution are those for the model and the estimated residuals – here, 1 and 3, more generally 1 and $n-2$ – and are found in the third column of the ANOVA table.

We see that our value of 61.54 lies between the 99*th* and 99.9*th* percentiles of 34.116 and 167.03, meaning our value provides highly statistically significant evidence in a *two*-sided test against H: $\beta_1 = 0$.

Other information about this test of H: $\beta_1 = 0$ is:

- The F distribution is a model for the behaviour of the *ratio of the squares of two standard deviation estimates* under assumptions of equiprobable selecting of samples from two respondent populations whose responses can be modelled by probabilistically independent normal distributions; we met this situation previously in Case 2 of Figure 12.19 when discussing the difference of the means of two normal models. In simple linear regression, the squares of the two 'standard deviation estimates' are MSM and MSE.

(*continued overleaf*)

NOTES: (cont.) 10. ● – Because MSM in simple linear regression is associated with only *one* degree of freedom, the *F* distribution is equivalent to the square of a *t* distribution with $n-2$ degrees of freedom (the number associated with the estimated residuals) using a *two*-sided test; thus, for a given number of degrees of freedom, five of a group of six entries in the first column of Figure 13.13 are (apart from rounding inaccuracy) the squares of the entries in the columns headed 0.95, 0.975, 0.995, 0.9995 and 0.99995 in the *t* distribution table in Figure 12.16.

○ From Figure 12.16, we would find the *P*-value for the test above is between 0.01 and 0.002 – our *F*-ratio of 61.54 is between $5.84091^2 = 34.116$ and $10.2145^2 = 104.34$, the table entries for the columns headed 0.995 and 0.999.

11. In this Figure, we have used r^2 as our *second* of three assessments of fit – visual, r^2 and ANOVA. If a problem requires the ANOVA table for its solution, it is then usually more convenient to use the quotient of SSM and SS_y from this table to find r^2 (and r) than it is to use the third expression for r^2 in equation (13.5.16); r^2 then becomes the *third* criterion of fit rather than the second.

● This discussion reminds us that the *closer* the value of r^2 is to 1, the *greater* the fraction of SS_y that has gone into SSM at the expense of SSE – this is something we look for as we construct an ANOVA table.

12. In places other than these Course Materials, an ANOVA table may have 'Regression' in place of 'Model' and 'Error' instead of 'Estimated residual', omit s_y^2 and also omit the reminder that $\hat{\sigma} = \sqrt{\text{MSE}}$ in the *F*-ratio column; the table would then look like:

Table 13.5.5:

SOURCE	SUM of SQUARES	*Df*	MEAN SQUARE	*F*-RATIO
Regression	$SSR = b_1 SS_{xy}$	1	$MSR = SSR \div 1$	$\frac{MSR}{MSE}$
Error	$SSE = SS_y - SSR$	$n-2$	$MSE = SSE \div (n-2)$	
Total	SS_y	$n-1$	-----	----- .

13. The form of ANOVA table used in these Course Materials is a compromise between practice elsewhere and our terminology; ideally, the writer would prefer the following:

Table 13.5.6:

SOURCE	SUM of SQUARES	*Df*	AVERAGE SQUARE	*F*-RATIO
Model	$SSM = b_1 SS_{xy}$	1	$ASM = SSM \div 1$	$\frac{ASM}{ASE} = \frac{ASM}{\hat{\sigma}^2}$
Estimated residual	$SSE = SS_y - SSM$	$n-2$	$ASE = SSE \div (n-2)$	
Total	SS_y	$n-1$	$[s_y^2 = SS_y \div (n-1)]$	----- .

The mean squares are *average* squares and the acronym PASSDI (partition of sums of squared differences) is more evocative of what is actually going on in 'analysis of variance'.

14. Using our now-familiar data set of five observations (given at the right), we can calculate various quantities which illustrate earlier ideas numerically (except for possible rounding inaccuracy) as shown in Table 13.5.7 at the right below. We see that:

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

Table 13.5.7:

x_j	y_j	$_{reg}\overline{y}_j$	$\hat{r}_j$	$\hat{r}_j^2$	$\hat{r}_j x_j$
1	3	3.238 095	−0.238 095	0.056 689	−0.238 095
3	5	4.158 730	0.841 270	0.707 735	2.523 810
6	5	5.539 683	−0.539 683	0.291 257	−3.238 095
10	7	7.380 952	−0.380 952	0.145 125	−3.809 524
15	10	9.682 540	0.317 460	0.100 781	4.761 904
Sum	30	30.	0.	1.301 587	0.

● the $_{reg}\overline{y}_j$s have the *same* sum as the y_js as a consequence of having the same *average* $\overline{y}$;

● the estimated residuals $\hat{r}_j$ sum to zero, illustrating the constraint (13.5.12);

● the sum of the *squares* of the estimated residuals, $\sum_{j=1}^{n}\hat{r}_j^2$, is SSE – one feature of the ANOVA table is it provides a computationally convenient route to SSE in place of this onerous direct calculation.

● the $\hat{r}_j x_j$ sum to zero, illustrating the constraint (13.5.13) on page 13.29.

15. From (13.5.16) and (13.5.15): $1 - r^2 = \frac{SS_y - SSM}{SS_y} = \frac{SSE}{SS_y}$ so that: $SSE = (1 - r^2)SS_y$; -----(13.5.17)

this result shows two matters of interest:

● dividing by the relevant degrees of freedom ($n-2$ and $n-1$) on the two sides of this expression, we see that MSE (which is used to estimate the model parameter σ) is *approximately* $(1 - r^2)$ times the square of the standard deviation of the y_js; the larger the value of n, the closer the approximation; (Why?)

● while *sum of squares* obey *exact* algebraic relationships, the corresponding *mean squares* – which provide us with useful *estimates* – do not.

– This is reminiscent of Case 2 of the test of significant for the difference of the means of two normal models in Figure 12.19 – when we calculate the *pooled* estimate of σ, the numerator of the quantity under the square root is actually a *sum of squares.* In that situation, the two sample standard deviations are estimating the model parameter representing the *common* standard deviation of the two respondent populations; it is somewhat more complicated in (13.5.17) for simple linear regression because MSE and s_y^2 estimate *different* quantities.

Figure 13.5. SIMPLE LINEAR REGRESSION: Assessing the Fit of a ..... Model (continued 3)

6. Anscombe's Four Regression Data Sets

Of interest in the context of this Figure 13.5 is an article in *The American Statistician* **27**(#1) 17-21 (1973) by F. J. Anscombe entitled *Graphs in Statistical Analysis.* The introduction of this article about the usefulness of graphs reminds us of the theme of Part 3 of the STAT 220 Course Materials and is as follows:

> Most textbooks on statistical methods, and most statistical computer programs, pay too little attention to graphs. Few of us escape being indoctrinated with these notions:
>
> - numerical calculations are exact, but graphs are rough;
> - for any particular kind of statistical data, there is just one set of calculations constituting a correct statistical analysis;
> - performing intricate calculations is virtuous, whereas actually looking at the data is cheating.
>
> A computer should make *both* calculations *and* graphs. Both sorts of output should be studied; each will contribute to understanding.
>
> Graphs can have various purposes, such as:
>
> * to help us perceive and appreciate some broad features of the data;
> * to let us look behind those broad features and see what else is there.
>
> Most kinds of statistical calculation rest on assumptions about the behaviour of the data. Those assumptions may be false, and then the calculations may be misleading. We ought always to try to check whether the assumptions are reasonably correct; if they are wrong, we ought to be able to perceive in what ways they are wrong. Graphs are very valuable for these purposes.
>
> Good statistical analysis is not a purely routine matter, and generally calls for more than one pass through the computer. The analysis should be sensitive both to peculiar features in the given numbers and also to whatever background information is available about the variables. The latter is particularly helpful in suggesting alternative ways of setting up the analysis.

Of relevance to simple linear regression are four fictitious data sets given by Anscombe and reproduced (with minor refinements) at the right; the x values are the *same* for the first three data sets. Numerical summaries relevant to simple linear regression are essentially the *same* for all four data sets and are:

$\Sigma x_j = 99$, $\Sigma x_j^2 = 1{,}001$; $\Sigma x_j y_j = 797.5$;

$\Sigma y_j = 82.5$, $\Sigma y_j^2 \simeq 660.0013$; $(n = 11)$

$\bar{x} = 9$, $\bar{y} = 7.5$;

$SS_{xy} = 55$, $SS_x = 110$, $SS_y \simeq 41.2513$;

Data set 1:											
x	4	5	6	7	8	9	10	11	12	13	14
y	4.25	5.68	7.24	4.82	6.98	8.80	8.03	8.32	10.84	7.58	9.96

Data set 2:											
x	4	5	6	7	8	9	10	11	12	13	14
y	3.10	4.74	6.13	7.26	8.14	8.76	9.14	9.26	9.13	8.74	8.10

Data set 3:											
x	4	5	6	7	8	9	10	11	12	13	14
y	5.39	5.69	6.08	6.44	6.80	7.12	7.44	7.83	8.14	12.74	8.83

Data set 4:											
x	8	8	8	8	8	8	8	8	8	8	19
y	6.57	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

hence, the *estimates* of β_1 (the slope) and β_0 (the $\mathbf{Y}$-intercept) of the regression of $\mathbf{Y}$ on $\mathbf{X}$ are: $b_1 = 0.5$, $b_0 = 3$,

so the equation of the straight-line model is: ${}_{\text{reg}}\bar{y} = 3 + 0.5x = 7.5 + 0.5(x - 9)$; also: $SSE \simeq 13.7513$, $r^2 \simeq 0.66665$.

Scatter diagrams of the four data sets, with the estimated regression of $\mathbf{Y}$ on $\mathbf{X}$ superimposed on each, are given overleaf on page 13.36; *visual* assessment shows the following:

Data set 1: The scatter of the points about the line seems to meet the three visual criteria for acceptability given at the top of the third side of this Figure 13.5 (page 13.31), indicating a 'successful' regression.

Data set 2: The dependence of $\mathbf{Y}$ on $\mathbf{X}$ appears to be *non*-linear and, in place of a *straight-line* model, a *quadratic* polynomial would be a better choice.

Data set 3: The *outlier* at $\mathbf{X} = 13$ pulls the fitted line upwards so it is *not* a good fit *either* to this observation *or* to the remaining ten; if these were real data, we should try to find out *why* this observation differs markedly from the other ten.

Data set 4: The estimate of the slope of the line depends mainly on the *one* observation with $\mathbf{X} = 19$, so it would be important to:

- check that this observation is *accurate*:
- understand *why* the data set has unusual $\mathbf{X}$ values – all but one are the *same.*

In addition, with observations at only *two* $\mathbf{X}$ values, we do not know if the relationship between $\mathbf{Y}$ and $\mathbf{X}$ is *linear* over the interval $\mathbf{X} = 8$ to $\mathbf{X} = 19$.

Thus, the straight-line model is *un*successful for three of the four data sets. The reader is referred to Anscombe's article for more detailed discussion and other matters of statistical interest.

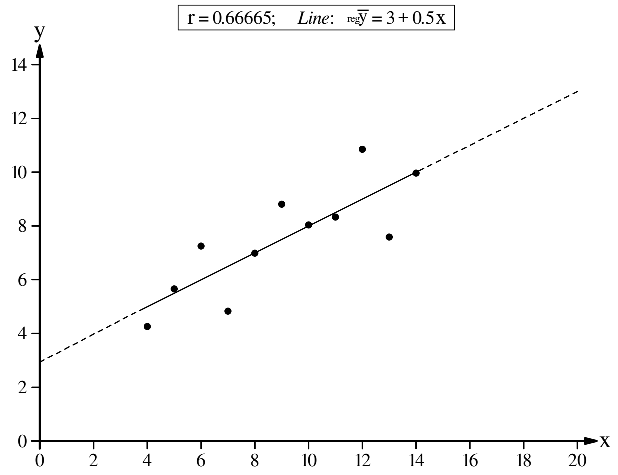
For this Figure 13.5, the lesson from Anscombe's four data sets – with their vastly differing *visual* implications but virtually identical *numerical* summaries – is the importance of *visual inspection* as our *primary* method of assessing the fit of a regression model, and the *supplementary* nature of the (still useful) information provided by the value of r^2 and the ANOVA table. We are also reminded that, essential as numerical summaries are for data analysis, they must be used with due care.

(*continued overleaf*)

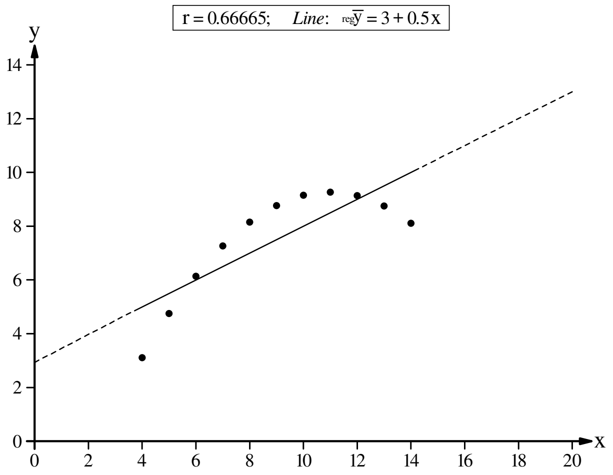
Exercise: 4. For Data set 3, if the outlier at $\mathbf{X} = 13$ is omitted, show that the estimated regression of $\mathbf{Y}$ on $\mathbf{X}$ for the remaining ten points is: ${}_{\text{reg}}\overline{y} = 4.003 + 0.3457x = 6.976 + 0.3457(x - 8.6)$ – this line is shown in *longer* dashes on the lower left diagram below.

- For the recalculated line, confirm that $r^2 = 0.999\,607$ – this is a dramatic *in*crease from the previous value of 0.666 647 and reflects the fact we can *see* that the ten points lie almost *on* the recalculated line.

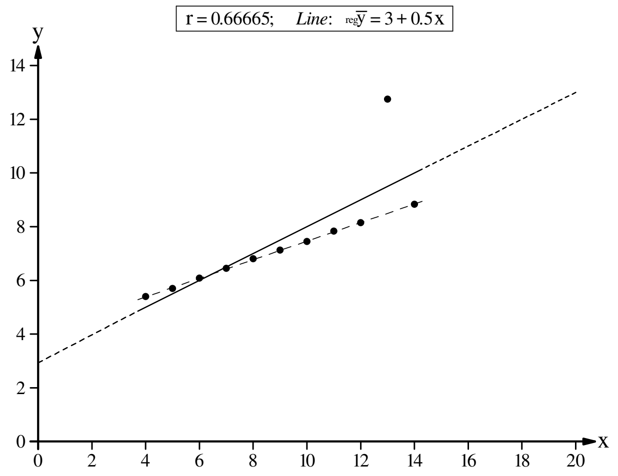
Data set 1



Data set 2



Data set 3



Data set 4