Figure 13.8. SIMPLE LINEAR REGRESSION: Case Study 1

Example 13.8.1: The bivariate data given below, for a sample of 12 apple trees, show the number of apples (in hundreds) on a tree and the percentage of the apples that had been attacked by codling moth larvae ('percent wormy'); it is generally believed that, across a number of apple trees, these two quantities are *inversely* related.

Number of apples [100s] (x)	6	8	11	14	17	18	19	22	23	24	26	40
Percent wormy (y)	58	59	56	50	45	43	39	53	38	42	30	27

(a) Prepare a properly-labelled scatter diagram of these data and show on it the estimated regression of $\mathbf{Y}$ on $\mathbf{X}$.

(b) Give the ANOVA table, coefficient of determination, correlation coefficient and estimate of σ for these data.

(c) Assess how well the regression model fits the data; give your reasons completely but concisely.

(d) Find a 95% confidence interval for the *slope* of the respondent population regression of $\mathbf{Y}$ on $\mathbf{X}$.

(e) Find a 95% confidence interval for the *intercept* of the respondent population regression of $\mathbf{Y}$ on $\mathbf{X}$.

(f) Find a 95% *confidence* interval for the *mean* $\mu_Y(\mathrm{x}_j = 30)$ representing ${}_{\mathrm{reg}}\overline{\mathbf{Y}}_{\mathrm{i}}(\mathbf{X}_{\mathrm{i}})$, the respondent population regression *average* percentage of wormy apples for trees with a crop size specified by explanatory variate $\mathbf{X}_{\mathrm{i}} = 3{,}000$ apples.

(g) Find a 95% *prediction* interval for the *random variable* $Y(\mathrm{x}_j = 30)$ representing ${}_{\mathrm{reg}}\mathbf{Y}_{\mathrm{i}}(\mathbf{X}_{\mathrm{i}})$, the respondent population regression percentage of wormy apples of an equiprobably-selected *individual* tree with a crop size specified by explanatory variate $\mathbf{X}_{\mathrm{i}} = 3{,}000$ apples.

Solution: (a) From the $\mathrm{n} = 12$ observations $(\mathrm{x_j}, \mathrm{y_j})$ given in the statement of the question, we find the following:

$\Sigma \mathrm{x_j} = 228$,
$\Sigma \mathrm{x_j^2} = 5{,}256$;

$\Sigma \mathrm{y_j} = 540$,
$\Sigma \mathrm{y_j^2} = 25{,}522$;

$\Sigma \mathrm{x_j y_j} = 9{,}324$;

$\overline{\mathrm{x}} = 19$,
$\overline{\mathrm{y}} = 45$;

$\mathrm{SS_{xy}} = -936$,
$\mathrm{SS_x} = 924$,
$\mathrm{SS_y} = 1{,}222$;

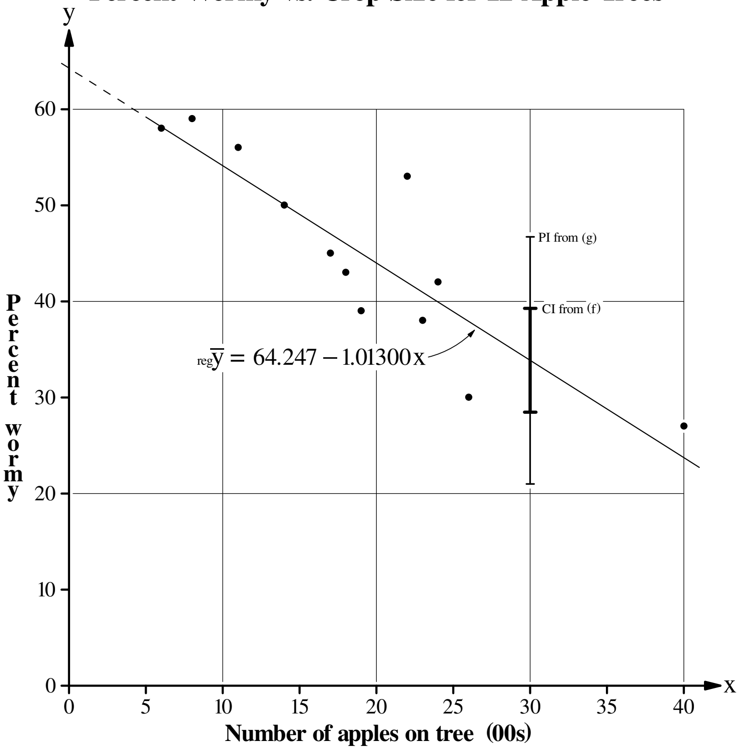
hence, the *estimates* of β_1 (the slope) and β_0 (the intercept) of the regression of $\mathbf{Y}$ on $\mathbf{X}$ are:

$b_1 = -1.012\ 987\ 01$,
$b_0 = 64.246\ 753\ 25$,

so the equation of the straight-line model is:

$${}_{\mathrm{reg}}\overline{\mathrm{y}} = 64.247 - 1.01300\mathrm{x} = 45 - 1.01300(\mathrm{x} - 19).$$

Percent Wormy vs. Crop Size for 12 Apple Trees



The scatter diagram of the data for the percentage of wormy apples against the number of apples on the tree, with the estimated regression of $\mathbf{Y}$ on $\mathbf{X}$ superimposed on it, is shown above at the right.

(b) The ANOVA table for these data is:

SOURCE	SUM of SQUARES	*Df*	MEAN SQUARE	*F*-RATIO
Model	$b_1\mathrm{SS_{xy}} = 948.155\ 844$	1	MSM = 948.155 844	$\frac{\mathrm{MSM}}{\mathrm{MSE}} = \frac{\mathrm{MSM}}{\hat{\sigma}^2} \simeq 34.62$
Estimated residual	SSE = 273.844 156	10	MSE = 27.384 415 584	
Total	$\mathrm{SS_y} = 1{,}222$	11	$(\mathrm{s_y^2} = 111.0\dot{9})$	-----

(*continued overleaf*)

Example 13.8.1: (continued) (b) Also, the coefficient of determination, the correlation coefficient and the estimate of σ are:

$r^2 = 0.775\,904\,946$, $r = -0.880\,854\,668$, $\hat{\sigma} = \sqrt{\text{MSE}} = 5.233\,012\,0948$.

(c) *Three* matters provide an assessment of how well the straight-line regression model fits the data:

- Visual inspection of the scatter diagram with the estimated regression of **Y** on **X** superimposed on it shows the points lie close, or reasonably close, to the line – the point with the estimated residual of largest magnitude is for the tree with 2,200 apples, which lies well *above* the line;
  - the points appear to be scattered without obvious pattern on both sides of the line;
  - there does not appear to be any systematic change in the magnitude of the estimated residuals with increasing values of x, which is consistent with the assumption of constant σ.
- The *F*-ratio (34.6) is very high and so provides (very) highly statistically significant evidence against the hypothesis $\beta_1 = 0$, indicating a meaningful regression – from page 1 of Figure 13.13 (page 13.61) and Figure 13.13k (page 13.77), $\Pr(F_{1,10} \geq 21.04) = 0.001$, $\Pr(F_{1,10} \geq 25.43) = 0.0005$, $\Pr(F_{1,10} \geq 38.58) = 0.0001$, so $P \simeq 0.0004$ for a two-sided test.
- The value of the coefficient of determination shows that nearly 78% of the variation of the y_js about their average has been accounted for by the estimated regression of **Y** on **X**.

These matters show the straight-line model is quite a good fit to these data.

(d) We want a 95% confidence interval (CI) for β_1, representing the *slope* of the regression of **Y** on **X**.

We have: $\hat{\sigma} = 5.233\,012\,0948$, $b_1 = -1.012\,987\,01$, $\Pr[-2.22814 \leq t_{10} \leq 2.22814] = 0.95$,

so that: $\hat{\sigma}\sqrt{\frac{1}{\text{SS}_\text{x}}} = \hat{\sigma}\sqrt{\frac{1}{924}} \simeq 0.172\,153\,458$;

hence, a 95% CI for β_1 is: $-1.012\,987\,01 \pm 0.383\,582\,007 \implies (-1.396\,569\,021, -0.629\,405)$
or about $(-1.40, -0.63)$ % wormy/100 apples.

(e) We want a 95% confidence interval (CI) for the *intercept* β_0, representing the *average* of the response variate **Y** when the respondent population, specified by the explanatory variate **X**, is trees with *zero* apples.

We have: $\hat{\sigma} = 5.233\,012\,0948$, $b_0 = 64.246\,753\,25$, $\Pr[-2.22814 \leq t_{10} \leq 2.22814] = 0.95$,

so that: $\hat{\sigma}\sqrt{\frac{\bar{\text{x}}^2}{\text{SS}_\text{x}} + \frac{1}{\text{n}}} = \hat{\sigma}\sqrt{\frac{19^2}{924} + \frac{1}{12}} \simeq 3.602\,904\,977$;

hence, a 95% CI for β_0 is: $64.246\,753\,25 \pm 8.027\,776\,695 \implies (56.218\,9765, 72.274\,529\,94)$
or about $(56.2, 72.3)$ % wormy.

(f) We want a 95% *confidence* interval (CI) for the *mean* $\mu_Y(\text{x}_j = 30)$ representing ${}_{\text{reg}}\bar{\mathbf{Y}}_\text{i}(\mathbf{X}_\text{i})$, the respondent population regression *average* percentage of wormy apples for trees with a crop size specified by explanatory variate $\mathbf{X}_\text{i} = 3{,}000$ (*i.e.*, 30 hundred) apples.

We have: $\hat{\sigma} = 5.233\,012\,0948$, $\hat{\mu}_Y(\text{x}_j = 30) = 33.857\,142\,86$, $\Pr[-2.22814 \leq t_{10} \leq 2.22814] = 0.95$,

so that: $\hat{\sigma}\sqrt{\frac{(\text{x}_j - \bar{\text{x}})^2}{\text{SS}_\text{x}} + \frac{1}{\text{n}}} = \hat{\sigma}\sqrt{\frac{(30-19)^2}{924} + \frac{1}{12}} \simeq 2.422\,413\,89$;

hence, a 95% CI for $\mu_Y(\text{x}_j = 30)$ is: $33.857\,142\,86 \pm 5.397\,477\,285 \implies (28.459\,6655, 39.254\,620\,14)$
or about $(28.4, 39.3)$ % wormy.

(g) We want a 95% *prediction* interval (PI) for the *random variable* $Y(\text{x}_j = 30)$ representing ${}_{\text{reg}}\mathbf{Y}_\text{i}(\mathbf{X}_\text{i})$, the respondent population regression percentage of wormy apples of an equiprobably-selected *individual* tree with a crop size specified by explanatory variate $\mathbf{X}_\text{i} = 3{,}000$ (*i.e.*, 30 hundred) apples.

We have: $\hat{\sigma} = 5.233\,012\,0948$, $\hat{\mu}_Y(\text{x}_j = 30) = 33.857\,142\,86$, $\Pr[-2.22814 \leq t_{10} \leq 2.22814] = 0.95$,

so that: $\hat{\sigma}\sqrt{\frac{(\text{x}_j - \bar{\text{x}})^2}{\text{SS}_\text{x}} + \frac{1}{\text{n}} + 1} = \hat{\sigma}\sqrt{\frac{(30-19)^2}{924} + \frac{1}{12} + 1} \simeq 5.766\,498\,474$;

hence, a 95% PI for $Y(\text{x}_j = 30)$ is: $33.857\,142\,86 \pm 12.848\,565\,91 \implies (21.008\,5769, 46.705\,708\,77)$
or about $(21.0, 46.7)$ % wormy.

REFERENCE: Snedecor, G.W. and Cochran, W.G.: *Statistical Methods*. Seventh edition, The Iowa State University Press, Ames, Iowa, 1980, page 162.

Snedecor and Cochran comment that *Apparently the density of the flying moths is unrelated to the size of the crop on a tree, so the chance of attack for any particular fruit is augmented if few fruits are on the tree.*