Assignment 8 Outline Solution

A8 – 1. (a) From the $n = 10$ observations (x_j, y_j) given in the statement of the question, we find the following:

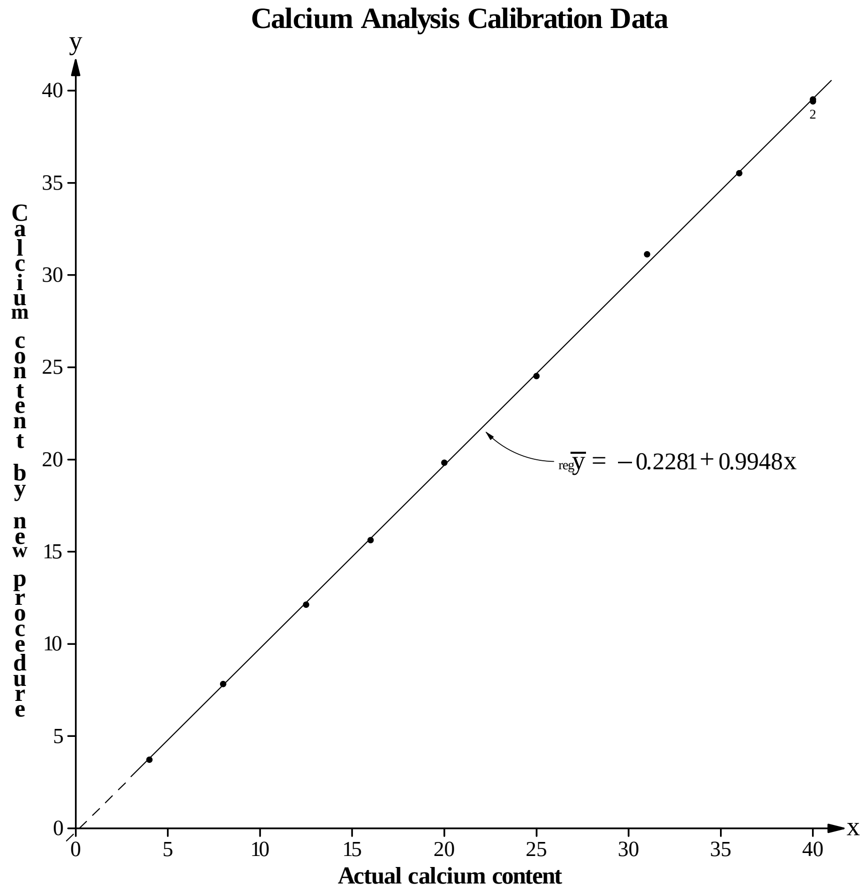
$\Sigma x_j = 232.5$,
$\Sigma x_j^2 = 6{,}974.25$;
$\Sigma y_j = 229$,
$\Sigma y_j^2 = 6{,}796.66$;
$\Sigma x_j y_j = 6{,}884.65$;
$\overline{x} = 23.25$,
$\overline{y} = 22.9$;
$SS_{xy} = 1{,}560.4$,
$SS_x = 1{,}568.625$,
$SS_y = 1{,}552.56$;

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

$b_1 = 0.994\,756\,554$,
$b_0 = -0.228\,089\,888$,

so the equation of the straight-line model is:

${}_{\text{reg}}\overline{y} = -0.2281 + 0.9948x$.



The scatter diagram of the the calcium analysis calibration data, 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	$b_1 SS_{xy}$ = 1,552.218 127 34	1	MSM = 1,552.218 127 34	$\frac{\text{MSM}}{\text{MSE}} = \frac{\text{MSM}}{\hat{\sigma}^2} \simeq 36{,}322.7$
Estimated residual	SSE = 0.341 872 66	8	MSE = 0.042 734 08	
Total	SS_y = 1,552.56	9	(s_y^2 = 172.50$\dot{6}$)	-----

Also, the coefficient of determination, correlation coefficient and estimate of σ for these data are:

$r^2 = 0.999\,779\,801$, $r = 0.999\,889\,894$, $\hat{\sigma} = \sqrt{\text{MSE}} = 0.206\,722\,235\,14$.

(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 $\mathbf{Y}$ on $\mathbf{X}$ superimposed on it shows the points lie *very* close to the line – the only estimated residual that is not *very* small is that of the *seventh* point;
  - 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 (36,323) is *extremely* high and so provides very highly statistically significant evidence against the hypothesis $\beta_1 = 0$, indicating a meaningful regression.
- The value of the coefficient of determination shows that well over 99.9% of the variation of the y_j's about their average has been accounted for by the estimated regression of $\mathbf{Y}$ on $\mathbf{X}$.

These considerations all show the straight-line model is an *excellent* fit to these data.

(d) We want a 95% confidence interval (CI) for the *intercept* β_0, representing the *average* of the response variate $\mathbf{Y}$ when the respondent (or study) population, specified by the explanatory variate $\mathbf{X}$, is an actual calcium content of *zero*.

(*continued overleaf*)

Assignment 8 Outline Solution (continued 1)

A8 – 1. (cont.) (d) We have: $\hat{\sigma} = 0.206\,722\,235\,14$, $b_0 = -0.228\,089\,888$, $\Pr[-2.30600 \le t_8 \le 2.30600] = 0.95$,

so that: $\hat{\sigma}\sqrt{\dfrac{\bar{x}^2}{SS_x} + \dfrac{1}{n}} = \hat{\sigma}\sqrt{\dfrac{23.25^2}{1{,}568.625} + \dfrac{1}{10}} \simeq 0.137\,840\,355$;

hence, a 95% CI for β_0 is: $-0.228\,089\,888 \pm 0.317\,859\,858 \implies (-0.545\,949\,746,\ 0.089\,769\,970)$ or about $(-0.55,\ 0.09)$ units.

Because zero lies with*in* this 95% CI, the data are consistent, at the 5% level, with the hypothesis $\beta_0 = 0$; *i.e.*, it seems that the respondent (or study) population regression of $\mathbf{Y}$ on $\mathbf{X}$ *can* be considered to pass through the origin.

(e) We want a 90% *confidence* interval (CI) for the *mean* $\mu_Y(x_j = 23.25)$ representing ${}_{\text{reg}}\bar{\mathbf{Y}}_i(\mathbf{X}_i)$, the respondent (or study) population regression *average* calcium content given by the new procedure when the actual calcium content is specified by explanatory variate $\mathbf{X}_i = 23.25$ units.

We have: $\hat{\sigma} = 0.206\,722\,235\,14$, $\hat{\mu}_Y(x_j = 23.25) = \hat{\mu}_Y(x_j = \bar{x}) = \bar{y} = 22.9$, $\Pr[-1.85955 \le t_8 \le 1.85955] = 0.90$,

so that: $\hat{\sigma}\sqrt{\dfrac{(x_j - \bar{x})^2}{SS_x} + \dfrac{1}{n}} = \hat{\sigma}\sqrt{\dfrac{(23.25 - 23.25)^2}{1{,}568.625} + \dfrac{1}{10}} \simeq 0.065\,371\,311$;

hence, a 90% CI for $\mu_Y(x_j = 23.25)$ is: $22.9 \pm 0.121\,561\,221 \implies (22.778\,438\,78,\ 23.021\,561\,22)$ or about $(22.8,\ 23.0)$ units.

A8 – 2. (a) From the $n = 8$ observations (x_j, y_j) given in the statement of the question, we find the following:

$\Sigma x_j = 36$,
$\Sigma x_j^2 = 204$,

$\Sigma y_j = 38.358$,
$\Sigma y_j^2 = 189.0325$,

$\Sigma x_j y_j = 187.16$;

$\bar{x} = 4.5$,
$\bar{y} = 4.79\,475$,

$SS_{xy} = 14.549$,
$SS_x = 42$,
$SS_y = 5.115\,4795$;

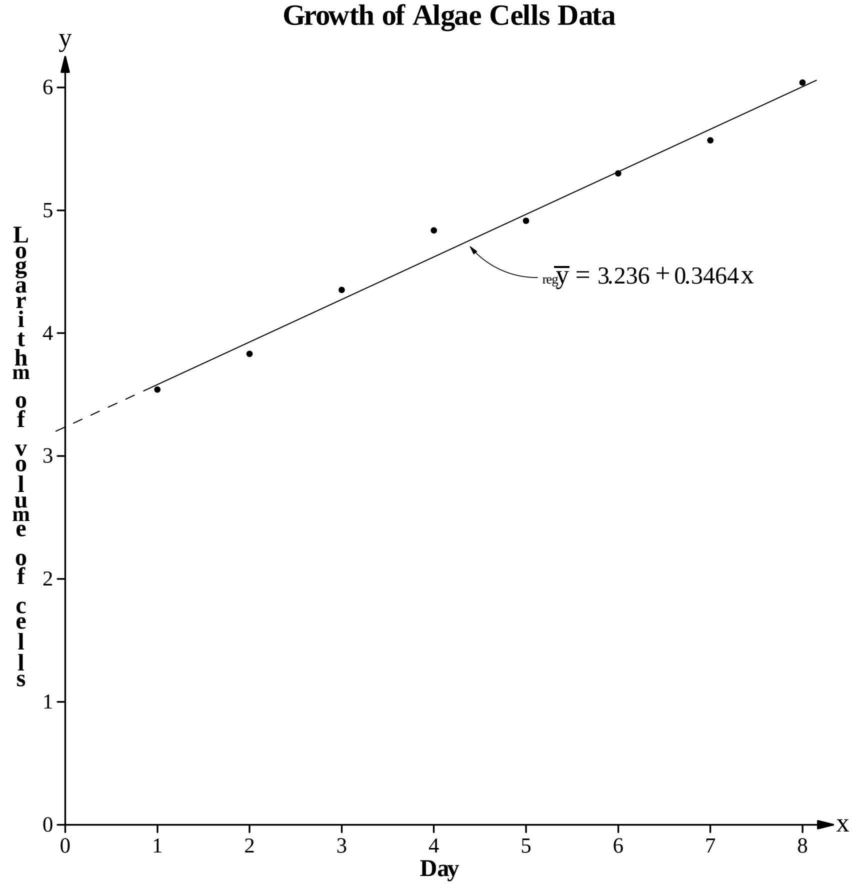
hence, the *estimates* of β_1 (the slope) and β_0 (the intercept) of the regression of $\mathbf{Y}$ on $\mathbf{X}$ are:

$b_1 = 0.346\,404\,762$,
$b_0 = 3.235\,928\,571$,

so the equation of the straight-line model is:

${}_{\text{reg}}\bar{y} = 3.236 + 0.3464x$.

Growth of Algae Cells Data



The scatter diagram of the growth of algae cells data, 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	$b_1 SS_{xy} = 5.039\,842\,881$	1	$MSM = 5.039\,842\,881$	$\dfrac{MSM}{MSE} = \dfrac{MSM}{\hat{\sigma}^2} \simeq 399.79$
Estimated residual	$SSE = 0.075\,636\,619$	6	$MSE = 0.012\,606\,103\,173$	
Total	$SS_y = 5.115\,479\,5$	7	$(s_y^2 = 0.730\,782\,786)$	-----

Assignment 8 Outline Solution (continued 2)

A8 – 2. (cont.) (b) Also, the coefficient of determination, the correlation coefficient and the estimate of σ are:

$r^2 = 0.985\ 214\ 168$, $r = 0.992\ 579\ 553$, $\hat{\sigma} = \sqrt{\text{MSE}} = 0.112\ 276\ 9040$.

(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 generally lie close to the line (the *fourth* point has the estimated residual of largest magnitude);
  - 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 (400) is very high and so provides highly statistically significant evidence against the hypothesis $\beta_1 = 0$, indicating a meaningful regression.
  - The value of the coefficient of determination shows that about 98½% of the variation of the y_j's about their average has been accounted for by the estimated regression of **Y** on **X**.

These considerations all show the straight-line model is a *very good* fit to the data.

(d) We want a 99% confidence interval (CI) for β_1, representing the *slope* of the respondent (or study) population regression of **Y** on **X**.

We have: $\hat{\sigma} = 0.112\ 276\ 9040$, $b_1 = 0.346\ 404\ 762$, $\Pr[-3.70743 \le t_6 \le 3.70743] = 0.99$,

so that: $\hat{\sigma}\sqrt{\frac{1}{SS_x}} = \hat{\sigma}\sqrt{\frac{1}{42}} \simeq 0.017\ 324\ 7024$;

hence, a 99% CI for β_1 is: $0.346\ 404\ 762 \pm 0.064\ 230\ 1215 \implies (0.282\ 174\ 64,\ 0.410\ 634\ 883)$
or about $(0.282,\ 0.411)$ units.

(e) The *labels* we attach to the days can make no fundamental difference to the nature of the regression; numbering the days 0, 1, ..., 7, instead of 1, 2, ..., 8, would leave the estimate of the *slope un*changed but would alter the estimate of the *intercept* (to about 3.5823).

A8 – 3. (a) From the n = 9 observations $(x_j,\ y_j)$ given in the statement of the question, we find the following:

$\sum x_j = 6$,
$\sum x_j^2 = 5.4375$,
$\sum y_j = 1{,}494$,
$\sum y_j^2 = 311{,}558$;
$\sum x_j y_j = 1{,}293.25$;

$\bar{x} = 0.\dot{6}$,
$\bar{y} = 166$;

$SS_{xy} = 297.25$,
$SS_x = 1.4375$,
$SS_y = 63{,}554$;

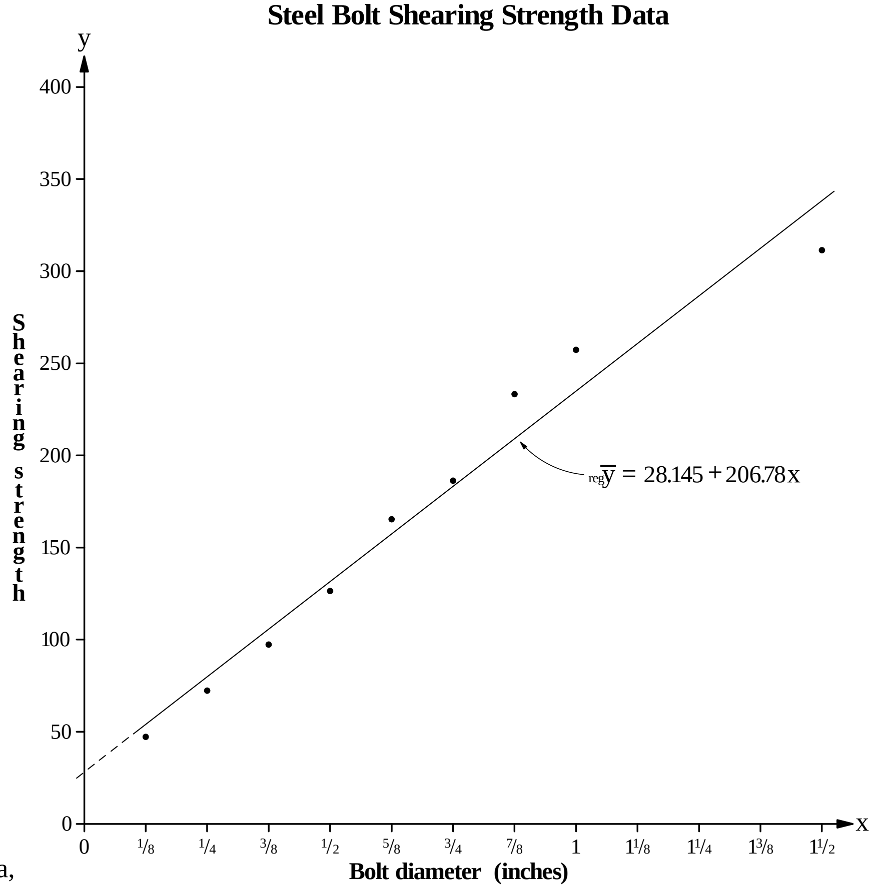
hence, the *estimates* of β_1 (the slope) and β_0 (the intercept) of the regression of **Y** on **X** are:

$b_1 = 206.782\ 6087$,
$b_0 = 28.144\ 927\ 54$,

so the equation of the straight-line model is:

$_{\text{reg}}\bar{y} = 28.145 + 206.78x$.

Steel Bolt Shearing Strength Data



The scatter diagram of the steel bolt shearing strength data,

(*continued overleaf*)

Assignment 8 Outline Solution (continued 3)

A8 – 3. (cont.) (a) with the estimated regression of **Y** on **X** superimposed on it, is shown at the lower right overleaf.

(b) The ANOVA table for these data is:

SOURCE	SUM of SQUARES	*Df*	MEAN SQUARE	*F*-RATIO
Model	$b_1 SS_{xy}$ = 61,466.130 43	1	MSM = 61,466.130 43	$\frac{MSM}{MSE} = \frac{MSM}{\hat{\sigma}^2} \simeq 206.08$
Estimated residual	SSE = 2,087.869 57	7	MSE = 298.267 080 74	
Total	SS_y = 63,554	8	(s_y^2 = 7,944.25)	-----

Also, the coefficient of determination, correlation coefficient and estimate of σ for these data are:

r^2 = 0.967 148 101, r = 0.983 436 882, $\hat{\sigma} = \sqrt{MSE}$ = 17.270 410 555.

(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 generally lie close to the line, although the estimated residuals are of noticably greater magnitude for the three largest bolts (diameters of seven-eighths of an inch and greater);
  - the points appear to be scattered without obvious pattern on both sides of the line;
  - although it is three *consecutive* points that have larger estimated residuals, this might reasonably occur by chance in as few as 9 points, so we are not unduly concerned about the assumption of constant σ.
- The *F*-ratio (206) is very high and so provides highly statistically significant evidence against the hypothesis $\beta_1 = 0$, indicating a meaningful regression.
- The value of the coefficient of determination shows that nearly 97% of the variation of the y_j's about their average has been accounted for by the estimated regression of **Y** on **X**.

These considerations all show the straight-line model is a *good* fit to the data.

(d) We want to carry out a test of significance for the *intercept* β_0, representing the *average* of the response variate **Y** when the respondent (or study) population, specified by the explanatory variate **X**, is bolts with diameter *zero*.

We have: $\hat{\sigma}$ = 17.270 410 555, b_0 = 28.144 927 54, $\frac{B_0 - \beta_0}{\hat{s.d.}(B_0)} \sim t_{n-2}$;

then: $\hat{s.d.}(B_0) = \hat{\sigma}\sqrt{\frac{\bar{x}^2}{SS_x} + \frac{1}{n}} = \hat{\sigma}\sqrt{\frac{0.\dot{6}^2}{1.4375} + \frac{1}{9}} \simeq 11.196\,366\,74$,

so that, under *H*: $\beta_0 = 0$, the value of the discrepancy measure is: $\frac{28.144\,927\,54 - 0}{11.196\,366\,75} \simeq 2.513\,755\,417$,

and the *P*-value is: $\Pr[|t_7| \geq 2.513\,755] = 2 \times \Pr[t_7 \geq 2.513\,755] \simeq 2 \times 0.020\,087\,835 \simeq 0.040\,176 \simeq 0.040$.

Hence on the basis of a two-sided *t*-test, the hypothesis $\beta_0 = 0$ is marginally *rejected* at the 5% level ($P \simeq 0.040$); thus, the data provide marginally statistically significant evidence the intercept of the regression of **Y** on **X** is *not* zero.

Alternatively, as in Question **A8 – 1**(d), we can approach this matter using a 95% confidence interval (CI) for β_0.

We have: $\hat{\sigma}$ = 17.270 410 555, b_0 = 28.144 927 54, $\Pr[-2.36462 \leq t_7 \leq 2.36462] = 0.95$;

hence, a 95% CI for β_0 is: 28.144 927 54 ± 26.475 152 72 ⟹ (1.669 774 82, 54.620 080 25)
or about (1.67, 54.6) units.

Because zero lies just *out*side this 95% CI, the data provide marginally statistically significant evidence *against* the hypothisis $\beta_0 = 0$; *i.e.*, the regression model for the relationship between **Y** and **X** does *not* appear to have zero ieeeeecept [as we previously concluded from the (equivalent) test of significance].

(e) We want a 95% *confidence* (CI) interval for the *mean* $\mu_Y(x_j = 1¼)$ representing ${}_{reg}\bar{\mathbf{Y}}_i(\mathbf{X}_i)$, the respondent (or study) population regression *average* shearing strength of bolts with diameter specified by explanatory variate $\mathbf{X}_i$ = 1¼ inches.

We have: $\hat{\sigma}$ = 19.270 410 555, $\hat{\mu}_Y(x_j = 1¼)$ = 286.623 1884, $\Pr[-2.36462 \leq t_7 \leq 2.36462] = 0.95$,

so that: $\hat{\sigma}\sqrt{\frac{(x_j - \bar{x})^2}{SS_x} + \frac{1}{n}} = \hat{\sigma}\sqrt{\frac{(1.25 - 0.\dot{6})^2}{1.4375} + \frac{1}{9}} \simeq 10.185\,532\,46$;

hence, a 95% CI for $\mu_Y(x_j = 1¼)$ is: 286.623 1884 ± 24.084 9138 ⟹ (262.538 2746, 310.708 1022)
or about (262.5, 310.7) units.

(f) We want a 95% *prediction* interval (PI) for the *random variable* $Y(x_j = 1¼)$ representing ${}_{reg}\mathbf{Y}_i(\mathbf{X}_i)$, the respondent (or study) population regression shearing strength of an equiprobably-selected *individual* bolt with diameter specified by explanatory variate $\mathbf{X}_i$ = 1¼ inches.

(*continued*)

Assignment 8 Outline Solution (continued 4)

A8 – 3. (cont.) (f) We have: $\hat{\sigma} = 17.270\,410\,555$, $\hat{\mu}_Y(\mathrm{x}_j = 1¼) = 286.623\,1884$, $\Pr[-2.36462 \le t_7 \le 2.36462] = 0.95$,

so that: $\hat{\sigma}\sqrt{\frac{(\mathrm{x}_j - \overline{\mathrm{x}})^2}{\mathrm{SS_x}} + \frac{1}{\mathrm{n}} + 1} = \hat{\sigma}\sqrt{\frac{(1.25 - 0.\dot{6})^2}{1.4375} + \frac{1}{9} + 1} \simeq 20.050\,240\,70$;

hence, a 95% PI for $Y(\mathrm{x}_j = 1¼)$ is: $286.623\,1884 \pm 47.411\,200\,17 \implies (239.211\,9882,\ 334.034\,3886)$
or about $(239.2,\ 334.0)$ units

(g) The solution of the **BONUS** part of question **A8 - 3** is given at the end of this solution, statring on page 0.12h.

A8 – 4. (a) From the $\mathrm{n} = 8$ observations $(\mathrm{x_j},\ \mathrm{y_j})$ given in the statement of the question, we find the following:

$\Sigma \mathrm{x_j} = 900$,
$\Sigma \mathrm{x_j^2} = 126{,}550$,
$\Sigma \mathrm{y_j} = 638$,
$\Sigma \mathrm{y_j^2} = 51{,}236$;
$\Sigma \mathrm{x_j y_j} = 74{,}445$;
$\overline{\mathrm{x}} = 112.5$,
$\overline{\mathrm{y}} = 79.75$,
$\mathrm{SS_{xy}} = 2{,}670$,
$\mathrm{SS_x} = 25{,}300$,
$\mathrm{SS_y} = 355.5$;

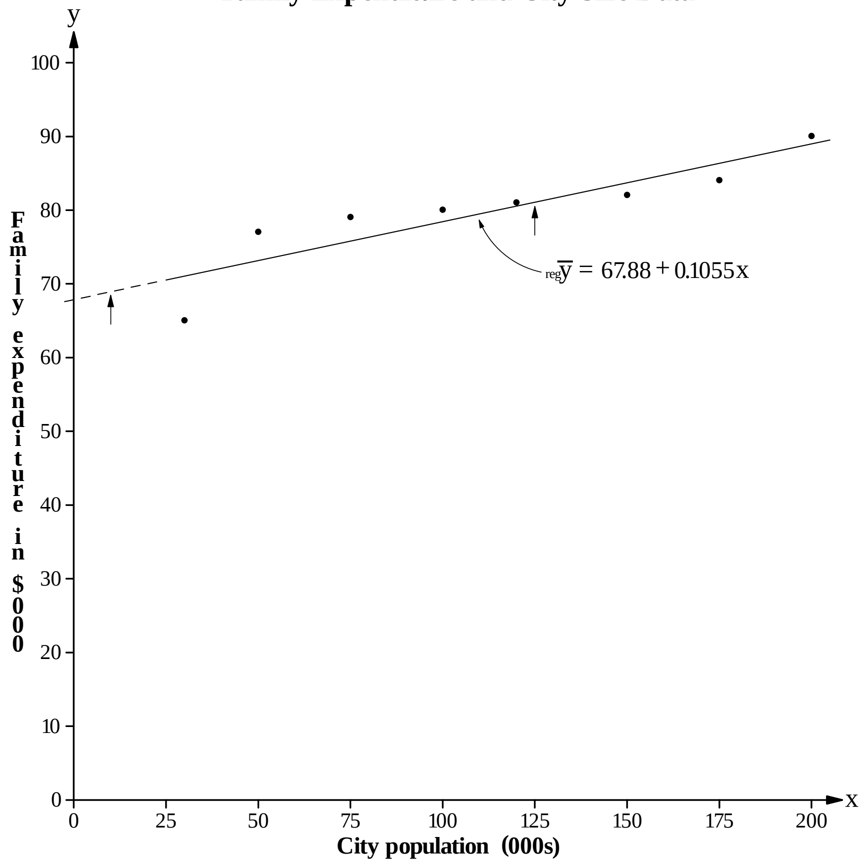
hence, the *estimates* of β_1 (the slope) and β_0 (the intercept) of the regression of **Y** on **X** are:

$b_1 = 0.105\,533\,5968$,
$b_0 = 67.877\,470\,36$,

so the equation of the straight-line model is:

${}_{\mathrm{reg}}\overline{\mathrm{y}} = 67.88 + 0.1055\mathrm{x}$.

Family Expenditure and City Size Data



The scatter diagram of the family expenditure and city size data, with the estimated regression of **Y** on **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	$b_1\mathrm{SS_{xy}} = 281.774\,7036$	1	MSM = 281.774 7036	$\frac{\mathrm{MSM}}{\mathrm{MSE}} = \frac{\mathrm{MSM}}{\hat{\sigma}^2} \simeq 22.93$
Estimated residual	SSE = 73.725 2964	6	MSE = 12.287 549 407	
Total	$\mathrm{SS_y} = 355.5$	7	$(\mathrm{s_y^2} = 50.785\,714)$	-----

Also, the coefficient of determination, correlation coefficient and estimate of σ for these data are:

$\mathrm{r}^2 = 0.792\,615\,200$, $\mathrm{r} = 0.890\,289\,391$, $\hat{\sigma} = \sqrt{\mathrm{MSE}} = 3.505\,360\,0966$.

(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 generally lie fairly close to the line (the *first* point has the estimated residual of largest magnitude);
  – the points appear to be scattered on both sides of the line but there may be a *pattern* in the values of the estimated residuals;
  – there does not appear to be any systematic change in the magnitude of the estimated residuals with increas-

Assignment 8 Outline Solution (continued 5)

A8 – 4. (cont.) (c) ing values of x, which is consistent with the assumption of constant σ.

- The *F*-ratio (23) is high and so provides statistically significant evidence against the hypothesis $\beta_1 = 0$, indicating a meaningful regression.
- The value of the coefficient of determination shows that just over 79% of the variation of the y_j's about their average has been accounted for by the estimated regression of $\mathbf{Y}$ on $\mathbf{X}$.

These considerations all show the straight-line model is a *fair* fit to the data; however, the pattern in the values of the estimated residuals suggests an *extension* of the straight-line model (*e.g.*, a *cubic* polynomial) would be a *better* fit.

(d) We want a *point* estimate of each of the random variables $Y(x_j = 125)$ and $Y(x_j = 10)$; these are obtained by substituting the relevant value of x in the equation of the line: ${}_{\text{reg}}\overline{y} = 67.88 + 0.1055x$. The results are:

$\hat{\mu}_Y(x_j = 125)$ = \$81.069 or about \$81,000;

- this prediction should be fairly close to the true value because it is based on a model that is a fair fit to the data and the value of x is with*in* (actually around the *centre*) of the range of the observed values of x.

$\hat{\mu}_Y(x_j = 10)$ = \$68.932 or about \$69,000;

- this prediction may *not* be close to the true value because the value of x is appreciably *out*side the range of the observed values; in addition, the data strongly suggest the *straight-line* model is irrelevant for values of x around 10.

NOTE: Although not asked for in the question, the corresponding 95% prediction *intervals* (in \$000) are:
for x = 125: (71.946, 90.192) or about (72, 90);
for x = 10: (58.286, 79.577) or about (58, 80).

A8 – 5. (a) From the n = 10 observations (t_j, w_j) given in the statement of the question, we find the following:

$\Sigma t_j = 20$,
$\Sigma t_j^2 = 60$,
$\Sigma w_j = 1{,}294$,
$\Sigma w_j^2 = 429{,}191$,
$\Sigma t_j w_j = 4{,}466$;

$\overline{t} = 2$,
$\overline{w} = 124.9$,
$SS_{tw} = 1{,}968$,
$SS_t = 20$,
$SS_w = 273{,}190.9$;

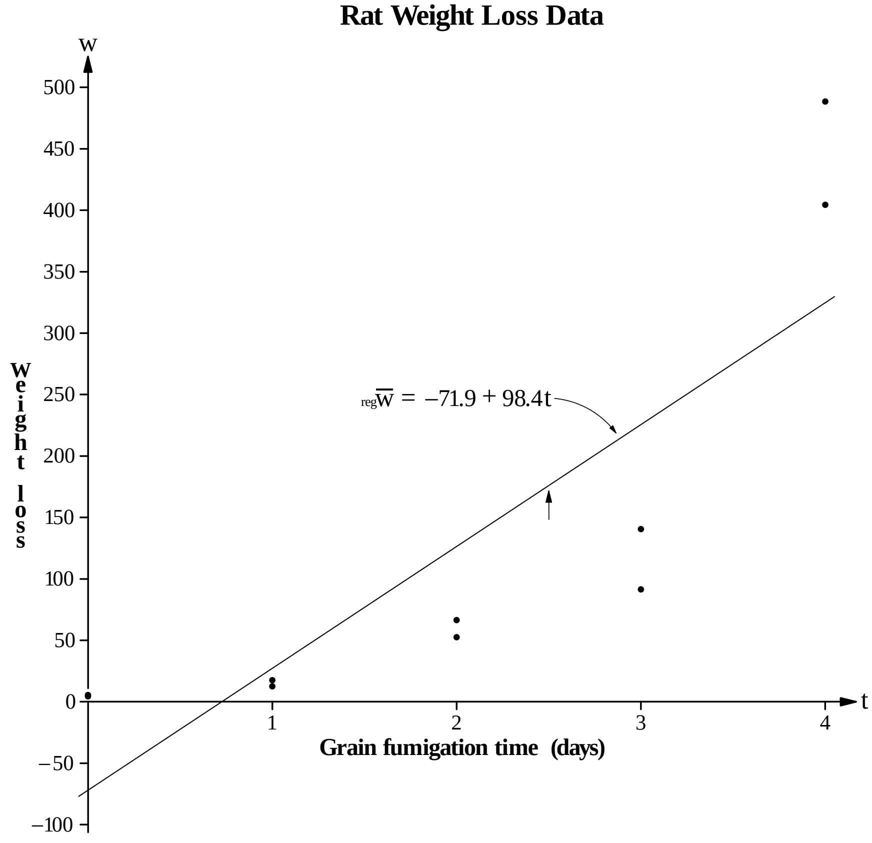
hence, the *estimates* of β_1 (the slope) and β_0 (the intercept) of the regression of $\mathbf{W}$ on $\mathbf{T}$ are:

$b_1 = 98.4$,
$b_0 = -71.9$,

so the equation of the straight-line model is:

${}_{\text{reg}}\overline{w} = -71.9 + 98.4t$.

Rat Weight Loss Data



The scatter diagram of the rat weight loss data, with the estimated regression of $\mathbf{W}$ on $\mathbf{T}$ 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	$b_1 SS_{tw}$ = 193,651.2	1	MSM = 193,651.2	$\frac{MSM}{MSE} = \frac{MSM}{\hat{\sigma}^2} \simeq 19.48$
Estimated residual	SSE = 79,539.7	8	MSE = 9,942.4625	
Total	SS_w = 273,190.9	9	(s_w^2 = 30,354.5$\dot{4}$)	-----

(*continued*)

Assignment 8 Outline Solution (continued 6)

A8 – 5. (cont.) (b) Also, the coefficient of determination, correlation coefficient and estimate of σ for these data are:

$r^2 = 0.708\,849\,380$, $r = 0.841\,931\,933$, $\hat{\sigma} = \sqrt{MSE} = 99.711\,897\,485$.

(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 $\mathbf{W}$ on $\mathbf{T}$ superimposed on it shows that, except for 1 day fumigation time, the remaining eight points deviate *substantially* from the line; in fact, the data suggest weight loss increases roughly *exponentially* (not *linearly*) with grain fumigation time.
  – although the points are fall on both sides of the fitted line, there is a pattern in the estimated residuals which reflects the curvelinear relationship between weight loss and fumigation time.
- The F-ratio (19.5) is high and so provides statistically significant evidence against the hypothesis $\beta_1 = 0$, indicating a meaningful regression.
- The value of the coefficient of determination shows that nearly 71% of the variation of the w_j's about their average has been accounted for by the estimated regression of $\mathbf{W}$ on $\mathbf{T}$.

These considerations show there *is* a clear *linear* tend in the data over the range of the observed values of t, but the form of the dependence of w on t is *curvi*linear and so a *transformation* of the straight-line model (*e.g.*, $ln_{reg}\overline{w} = \beta_0 + \beta_1 t$) would be a *better* fit [see (e) below].

(d) We want a *point* estimate of each of the random variables $W(t_j = 2½)$ and $W(t_j = 10)$; these are obtained by substituting the relevant value of t in the equation of the fitted line: $ln_{reg}\overline{w} = -71.9 + 98.4t$. The results are:

$\hat{\mu}_W(t_j = 2½) = 174.1$ or about 174 units of weight loss;

- this prediction may *not* be close to the true value, although it may be of the correct *order of magnitude*, because it is in about the *centre* of the range of observed values of t and the model shows *rough* correspondence with the data around such values of t.

$\hat{\mu}_W(t_j = 10) = 912.1$ or about 900 units of weight loss;

- this prediction is essentially *meaningless* because the value of t is substantially *out*side the range of the observed values, and the data suggest the *straight-line* model will be irrelevant for such values of t.

(e) The transformed data (with the logarithms to 4 significant figures) are given at the top of the page overleaf.

From these $n = 10$ observations (t_j, w_j) we find the following, where $l = ln\,w$:

$\Sigma t_j = 20$,
$\Sigma t_j^2 = 60$,

$\Sigma l_j = 35.1391$,
$\Sigma l_j^2 = 163.003\,451\,61$,

$\Sigma t_j l_j = 97.809$;

$\overline{t} = 2$,
$\overline{l} = 3.513\,91$,

$SS_{tl} = 27.5308$,
$SS_t = 20$,
$SS_l = 39.527\,816\,729$;

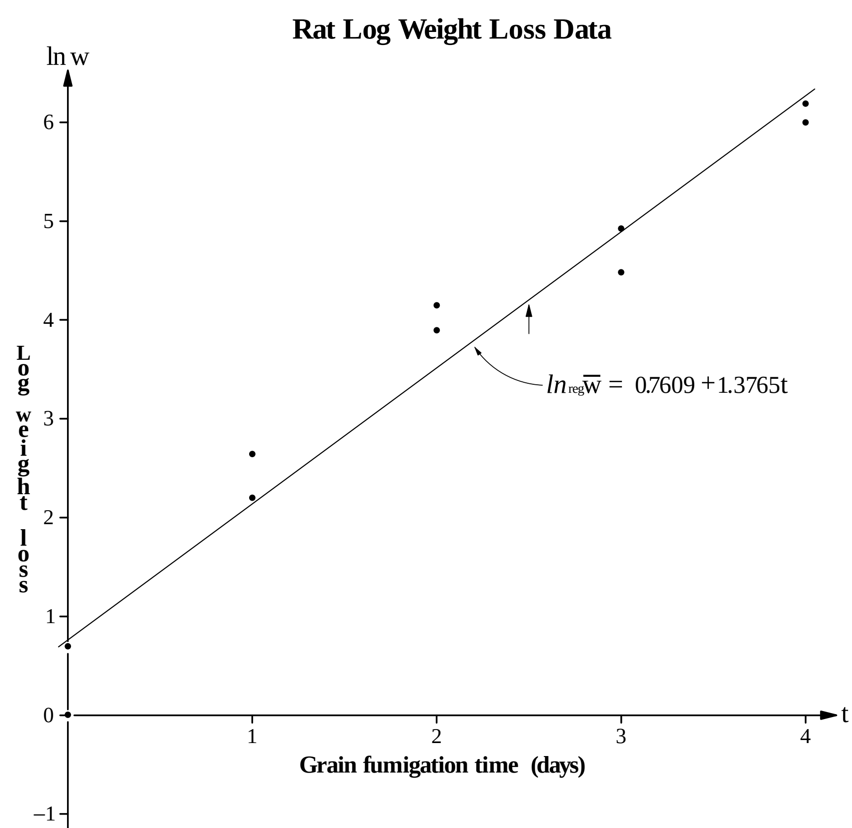
hence, the *estimates* of β_1 (the slope) and β_0 (the intercept) of the regression of $ln\mathbf{W}$ on $\mathbf{T}$ are:

$b_1 = 1.376\,54$,
$b_0 = 0.760\,83$,

so the equation of the straight-line model is:

$ln_{reg}\overline{w} = 0.7609 + 1.3765t$.

The scatter diagram of the rat log weight loss data, with



(*continued overleaf*)

Assignment 8 Outline Solution (continued 7)

A8 – 5. (cont.) (e) the (transformed) estimated regression of $ln\mathbf{W}$ on $\mathbf{T}$ superimposed on it, is shown at the lower right overleaf.

Fumigation time (t)	0	0	1	1	2	2	3	3	4	4
Log weight loss (ln w)	0	0.6931	2.197	2.639	3.892	4.143	4.477	4.920	6.184	5.994.

The ANOVA table for these data is:

SOURCE	SUM of SQUARES	*Df*	MEAN SQUARE	*F*-RATIO
Model	$b_1 SS_{t1}$ = 37.897 247 432	1	MSM = 37.897 247 432	$\frac{\text{MSM}}{\text{MSE}} = \frac{\text{MSM}}{\hat{\sigma}^2} \simeq 185.9$
Estimated residual	SSE = 1.630 569 297	8	MSE = 0.203 821 162 125	
Total	SS_1 = 39.527 816 73	9	(s_1^2 = 4.391 979 637)	-----

Also, the coefficient of determination, correlation coefficient and estimate of σ for these data are:

$r^2 = 0.958\,748\,815$, $r = 0.979\,157\,196$, $\hat{\sigma} = \sqrt{\text{MSE}} = 0.451\,465\,571\,362$.

Three matters provide an assessment of how well the straight-line regression model fits the transformed data:

- Visual inspection of the scatter diagram with the estimated regression of $ln\mathbf{W}$ on $\mathbf{T}$ superimposed on it shows the points generally lie close, or reasonably close, to the line;
  - the points appear to be scattered on both sides of the line, although there is some suggestion of a *pattern* in that the two observations for the same fumigation time tend to lie on the *same* side of the line;
  - there does not appear to be any systematic change in the magnitude of the estimated residuals with increasing values of t, which is consistent with the assumption of constant σ.
- The *F*-ratio (186, compared with 19.5 previously) is very high and so provides *highly* statistically significant evidence against the hypothesis $\beta_1 = 0$, indicating a meaningful regression.
- The value of the coefficient of determination shows that 95.9% (compared with 70.9% previously) of the variation of the l_j's about their average has been accounted for by the estimated regression of $ln\mathbf{W}$ on $\mathbf{T}$.

These considerations all show the straight-line model is a good fit to the *transformed* data, clearly a *better* fit than in (a) and (b) *before* transformation.

(f) We want a *point* estimate of each of the random variables $lnW(t_j = 2½)$ and $lnW(t_j = 10)$; these are obtained by substituting the relevant value of t in the equation of the transformed fitted line: $ln_{\text{reg}}\overline{w} = 0.7609 + 1.3765t$.
The results are:

$ln\hat{\mu}_w(t_j = 2½) = 4.2022$ so that $\hat{\mu}_w(t_j = 2½) = 66.83$ or about 67 units of weight loss;

- this prediction should be close to the true value because it is based on a model that is a good fit to the data and it is for a value of t that is in about the *centre* of the range of the observed values of t.

$ln\hat{\mu}_w(t_j = 10) = 14.53$ so that $\hat{\mu}_w(t_j = 10) = 2{,}035{,}455$ or about 2 million units of weight loss;

- this prediction is, as before, essentially *meaningless* because the value of t is substantially *out*side the range of the observed values of t, and we have no information about whether the transformed *straight-line* model applies for such values of t; one reflection of this meaninglessness is that a weight loss as high as this is not realizable in practice.

A8 – 3. (g) **BONUS:** The analysis (which starts on the facing page 0.12i) of the bolt shearing strength data, omitting the observation for a bolt of diameter 1½ inches, illustrates five matters of statistical interest:

- We can properly discard observations from a data set *only* for adequate reason(s) – lack of fit to our model is *not*, in itself, an adequate reason. To justify the omission here, we would need to discuss the applicability of the theory over varying bolt diameters with a person with subject matter knowledge. *We* pursue the analysis for pedagogic reasons.
- We see again the importance of *visual* assessment of the fit of a regression model; in this instance, the more technical criteria of the increases in the values of r^2 and the *F*-ratio are less compelling than the clear improvement of fit apparent from the scatter diagram.
- Comparing the fitted regression lines for the full and truncated data sets on the scatter diagram on the facing page 0.12i, we see that the observation at x = 1½ inches is an *influential* observation – recall Data Set 3 of Anscombe's four regression data sets, presented on the last two sides of Figure 13.5.
- The test of significance in (d) on the *last* side (page 0.12j) of this Outline Solution shows that, for the *truncated* data set, the regression of $\mathbf{Y}$ on $\mathbf{X}$ *can* be considered to pass through the origin, a *different* Answer from that based on the *full* data set.

(*continued*)

Assignment 8 Outline Solution (continued 8)

A8 – 3. (g) **(cont.)**

- The 95% CI for $\mu_Y(x_j = 1\frac{1}{4})$ differs substantially between the two models, with the interval based on the *truncated* data being higher and narrower; the difference between the two PI's for $Y(x_j = 1\frac{1}{4})$ is somewhat less striking because of the greater widths of the intervals.

Model quantity	Interval n = 9	Interval n = 8
$\mu_Y(x_j = 1\frac{1}{4})$	(262.5, 310.7) 48.2	(302.3, 331.4) 29.1
$Y(x_j = 1\frac{1}{4})$	(239.2, 334.0) 94.8	(295.4, 338.3) 42.9

From the first n = 8 observations (x_j, y_j) given in the statement of the question, we find the following:

$\Sigma x_j = 4.5$,
$\Sigma x_j^2 = 3.1875$,
$\Sigma y_j = 1{,}183$,
$\Sigma y_j^2 = 214.837$;
$\Sigma x_j y_j = 826.75$;

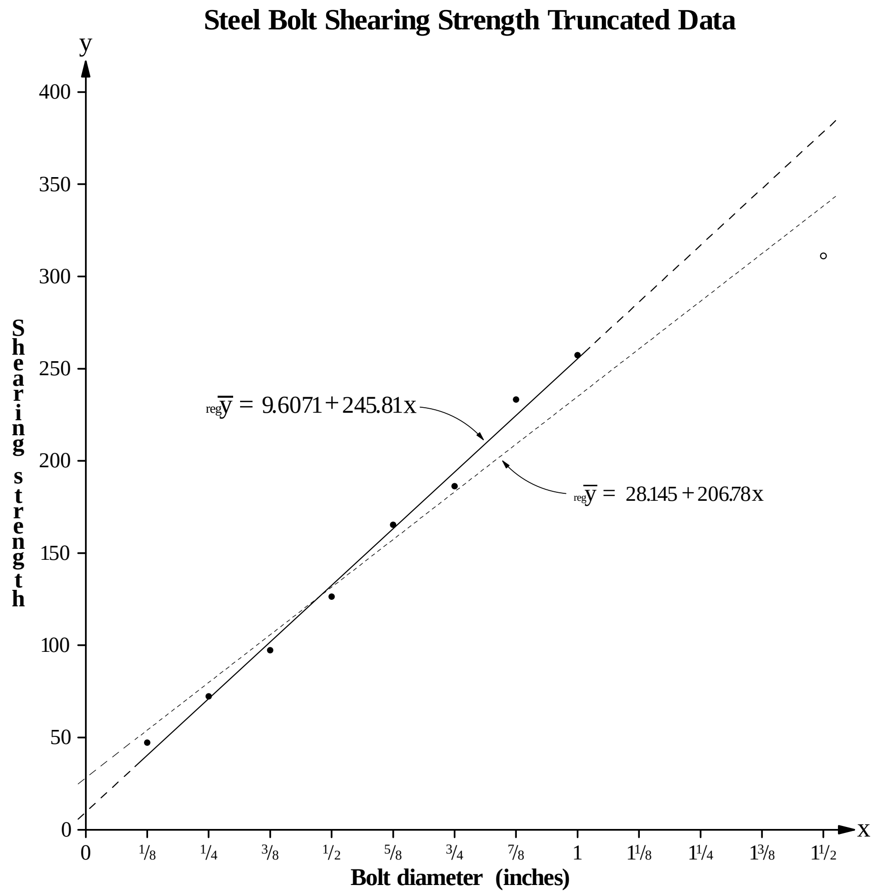
$\bar{x} = 0.5625$,
$\bar{y} = 147.875$,
$SS_{xy} = 161.3125$,
$SS_x = 0.65625$,
$SS_y = 39{,}900.875$;

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

$b_1 = 245.809\,5238$,
$b_0 = 9.607\,142\,857$,

so the equation of the straight-line model is:

${}_{reg}\bar{y} = 9.6071 + 245.81x$.



The scatter diagram of the *truncated* steel bolt shearing strength data, with the estimated regression of **Y** on **X** superimposed on it, is shown above as the *solid* line; the longer line in *short* dashes is the estimated regression of **Y** on **X** based on the *full* data set of 9 observations. (The observation omitted from the truncated data set is the *circle* at the upper right of the scatter diagram, for a bolt of diameter 1½ inches)

BONUS (b) The ANOVA table for these truncated data is:

SOURCE	SUM of SQUARES	*Df*	MEAN SQUARE	*F*-RATIO
Model	$b_1 SS_{xy}$ = 39,652.148 809	1	MSM = 39,652.148 809	$\frac{MSM}{MSE} = \frac{MSM}{\hat{\sigma}^2} \simeq 956.5$
Estimated residual	SSE = 248.726 191	6	MSE = 41.454 3651	
Total	SS_y = 39,900.875	7	(s_y^2 = 5,700.107 143)	-----

Also, the coefficient of determination, correlation coefficient and estimate of σ for these data are:

$r^2 = 0.993\,766\,397$, $r = 0.996\,878\,326$, $\hat{\sigma} = \sqrt{MSE} = 6.438\,506\,433\,95$.

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

- Visual inspection of the scatter diagram with the estimated regression of **Y** on **X** superimposed on it shows the points genereally lie close to 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.

(*continued overleaf*)

Assignment 8 Outline Solution (continued 9)

A8 – 3. (cont.) (c)
- The *F*-ratio (956, compared with 206 previously) is very high and so provides more highly statistically significant evidence against the hypothesis $\beta_1 = 0$, indicating a meaningful regression.
- The value of the coefficient of determination shows that 99.4% (compared with 96.7% previously) of the variation of the y_j's about their average has been accounted for by the estimated regression of **Y** on **X**.

These considerations all show the straight-line model is a *very good* fit to the truncated data, an appreciably *better* fit than previously.

NOTE: An additional indication of the improved fit is the estimate of residual variability about the regression line of $\hat{\sigma} = \sqrt{\text{MSE}} = 6.439$ for the *truncated* data set of 8 observations is roughly 2.7 times *smaller* than the estimate $\hat{\sigma} = \sqrt{\text{MSE}} = 17.270$ based on the *full* data set of 9 observations.

BONUS (d) We want to carry out a test of significance for the *intercept* β_0, representing the *average* of the response variate **Y** when the respondent (or study) population, specified by the explanatory variate **X**, is bolts with diameter *zero*.

We have: $\hat{\sigma} = 6.438\,506\,433\,95$, $b_0 = 9.607\,142\,857$, $\dfrac{B_0 - \beta_0}{\widehat{s.d.}(B_0)} \sim t_{n-2}$;

then: $\widehat{s.d.}(B_0) = \hat{\sigma}\sqrt{\dfrac{\bar{x}^2}{SS_x} + \dfrac{1}{n}} = \hat{\sigma}\sqrt{\dfrac{0.5625^2}{0.65625} + \dfrac{1}{8}} \simeq 5.016\,843\,795$,

so that, under H: $\beta_0 = 0$, the value of the discrepancy measure is: $\dfrac{9.607\,142\,857 - 0}{5.016\,843\,82} \simeq 1.914\,977\,474$,

and the *P*-value is: $\Pr[|t_6| \geq 1.914\,977] = 2 \times \Pr[t_6 \geq 1.914\,977] \simeq 2 \times 0.051\,993\,887 \simeq 0.104$.

Hence on the basis of a two-sided *t*-test, the hypothesis $\beta_0 = 0$ is *not* rejected at the 5% level ($P \simeq 0.104$); thus, the data suggest the intercept of the regression of **Y** on **X** *can* be considered to be zero.

Alternatively, as in Question **A8 – 1**(d), we can approach this matter using a 95% confidence interval (CI) for β_0.

We have: $\hat{\sigma} = 6.438\,506\,433\,95$, $b_0 = 9.607\,142\,857$, $\Pr[-2.44691 \leq t_6 \leq 2.44691] = 0.95$;

hence, a 95% CI for β_0 is: $9.607\,142\,857 \pm 12.275\,765\,252 \implies (-2.668\,622\,394,\ 21.882\,908\,11)$
or about $(-2.66,\ 21.9)$ units.

Because zero lies *in*side this 95% CI, the data provide no statistically significant evidence *against* the hypothisis $\beta_0 = 0$; *i.e.*, the regression model for the relationship between **Y** and **X** *can* be considered to have zero intercept [as we previously concluded from the (equivalent) test of significance].

BONUS (e) We want a 95% *confidence* (CI) for the *mean* $\mu_Y(x_j = 1¼)$ representing ${}_{\text{reg}}\bar{\mathbf{Y}}_i(\mathbf{X}_i)$, the respondent (or study) population regression *average* shearing strength of bolts with diameter specified by yexplanatory variate $\mathbf{X}_i = 1¼$ inches.

We have: $\hat{\sigma} = 6.438\,506\,433\,95$, $\hat{\mu}_Y(x_j = 1¼) = 316.869\,0476$, $\Pr[-2.44691 \leq t_6 \leq 2.44691] = 0.95$,

so that: $\hat{\sigma}\sqrt{\dfrac{(x_j - \bar{x})^2}{SS_x} + \dfrac{1}{n}} = \hat{\sigma}\sqrt{\dfrac{(1.25 - 0.5625)^2}{0.65625} + \dfrac{1}{8}} \simeq 5.919\,358\,799$;

hence, a 95% CI for $\mu_Y(x_j = 1¼)$ is: $316.869\,047\,62 \pm 14.484\,138\,24 \implies (302.384\,9094,\ 331.353\,1859)$
or about $(302.3,\ 331.4)$ units.

BONUS (f) We want a 95% *prediction* interval (PI) for the *random variable* $Y(x_j = 1¼)$ representing ${}_{\text{reg}}\mathbf{Y}_i(\mathbf{X}_i)$, the respondent (or study) population regression shearing strength of an equiprobably-selected *individual* bolt with diameter specified by explanatory variate $\mathbf{X}_i = 1¼$ inches.

We have: $\hat{\sigma} = 6.438\,506\,433\,95$, $\hat{\mu}_Y(x_j = 1¼) = 316.869\,0476$, $\Pr[-2.44691 \leq t_6 \leq 2.44691] = 0.95$,

so that: $\hat{\sigma}\sqrt{\dfrac{(x_j - \bar{x})^2}{SS_x} + \dfrac{1}{n} + 1} = \hat{\sigma}\sqrt{\dfrac{(1.25 - 0.5625)^2}{0.65625} + \dfrac{1}{8} + 1} \simeq 8.746\,037\,600$;

hence, a 95% PI for $Y(x_j = 1¼)$ is: $316.869\,047\,62 \pm 21.400\,766\,86 \implies (295.468\,2808,\ 338.269\,8145)$
or about $(295.4,\ 338.3)$ units.