THE ROYAL STATISTICAL SOCIETY

2009 EXAMINATIONS – SOLUTIONS

HIGHER CERTIFICATE

MODULE 4

LINEAR MODELS

The Society provides these solutions to assist candidates preparing for the examinations in future years and for the information of any other persons using the examinations.

The solutions should NOT be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids.

Users of the solutions should always be aware that in many cases there are valid alternative methods. Also, in the many cases where discussion is called for, there may be other valid points that could be made.

While every care has been taken with the preparation of these solutions, the Society will not be responsible for any errors or omissions.

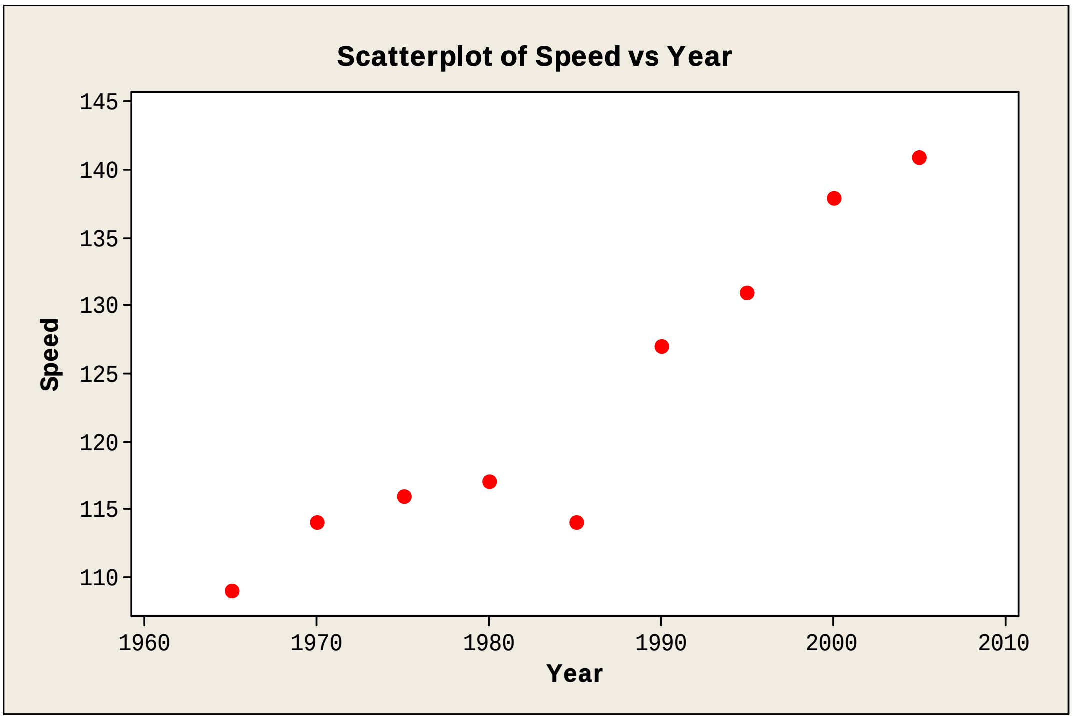
The Society will not enter into any correspondence in respect of these solutions.

Note. In accordance with the convention used in the Society's examination papers, the notation log denotes logarithm to base *e*. Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Higher Certificate, Module 4, 2009. Question 1

Part (i)

(a)



[Note the "false origin" of the scatterplot.]

The data show a clear and nearly linear trend, except for an unexpectedly low result in 1985. Linear regression analysis seems reasonable, with due caution regarding the 1985 observation.

(b) With the usual notation, the slope estimate is

$$\hat{\beta} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{\sum(x-\bar{x})y}{\sum(x-\bar{x})^2} = \frac{1200}{1500} = 0.8,$$

and the intercept is

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \frac{1107}{9} - (0.8 \times 1985) = 123 - 1588 = -1465.$$

The fitted line is $\hat{y} = 0.8x - 1465$.

Solution continued on next page

The total SS is

$$\Sigma(y-\bar{y})^2 = \Sigma y^2 - \frac{(\Sigma y)^2}{n} \text{ [or } \Sigma y^2 - n\bar{y}^2\text{]} = 137233 - (1107^2/9) = 1072.$$

The regression SS is

$$\hat{\beta}^2\Sigma(x-\bar{x})^2 = 0.8^2\times1500 = 960.$$

By subtraction, the error (or "residual") SS = 1072 − 960 = 112. This has 9 − 2 = 7 df. Hence the error mean square = 112/7 = 16.

Part (ii)

The analysis with 1985 omitted is better because

- 1985 is plausibly an outlier and we have a good reason for omitting it
- the error mean square of 16 reduces to 3.48 when the plausible outlier at 1985 is omitted, i.e. we then have a far better fit.

Part (iii)

(a) ***Either*** substitute x = 1985 in the regression equation,

or note that 1985 = $\bar{x}$ and therefore $\hat{y}(1985) = \bar{y} = \frac{1107-114}{9-1} = \frac{993}{8} =$ 124.125.

(b) $\hat{y}(2010) = (0.8\times2010) - 1463.87 = 144.13$.

(c) We expect a winning speed of 160 mph when year x satisfies the equation $160 = 0.8x - 1463.87$, so

$$x = 1.25\times(160 + 1463.87) = 2029.845 \text{ or 2030 approximately.}$$

The answers to (iii)(b) and (iii)(c) involve extrapolating beyond the range of the data, i.e. assuming that the fitted trend continues to apply in the future, which may not be true.

Higher Certificate, Module 4, 2009. Question 2

(i) The null hypothesis is that there are no differences between the population mean breaking strains of steel wire manufactured by these four suppliers. The alternative hypothesis is that at least two of these means differ.

The grand total is 670 + 680 + 695 + 715 = 2760. The sum of squares of all 20 observations is 89810 + 92510 + 96659 + 102275 = 381254.

"Correction factor" is $\frac{2760^2}{20} = 380880$.

Therefore total SS = 318254 – 380880 = 374.

SS for suppliers = $\frac{670^2}{5} + \frac{680^2}{5} + \frac{695^2}{5} + \frac{715^2}{5} - 380880 = 230$.

The residual SS is obtained by subtraction.

Hence the analysis of variance table is as follows.

SOURCE	DF	SS	MS	*F* value
Suppliers	3	230	76.7	8.52 Compare $F_{3,16}$
Residual	16	144	9	$= \hat{\sigma}^2$
TOTAL	19	374		

A level of significance for a formal test is not specified in the question. However, the upper 0.5% point of $F_{3,16}$ is 6.30, and the F value from the analysis of variance exceeds this (indeed, it is not far short of the upper 0.1% point which is 9.01). So the suppliers effect is very highly significant. There is very strong evidence to reject the null hypothesis that all the suppliers produce steel wire with the same mean breaking strain.

The means of the breaking strains are 670/5 = 134, 680/5 = 136, 695/5 = 139 and 715/5 = 143. This suggests that supplier A is the worst (lowest population mean breaking strain) and D the best, though we cannot be certain at this stage that all differences between suppliers are significant.

(ii) The assumptions are that the residuals are independent identically distributed N(0, σ^2) random variables.

To check these (note that no details of formal tests are expected in this module), calculate the residuals and check that the within-suppliers variances appear equal; check, for example, the time sequence of the residuals for any pattern or serial correlation; check the distribution of the residuals for Normality.

Solution continued on next page

(iii) We have $\bar{y}_A - \bar{y}_B = -2.0$, and the standard error of this estimate is $\sqrt{\dfrac{\hat{\sigma}^2}{n_A} + \dfrac{\hat{\sigma}^2}{n_B}}$ $= \sqrt{9\left(\frac{1}{5}+\frac{1}{5}\right)}$ = 1.8974. The two-sided 10% critical value for t_{16} is 1.746, so a 90% confidence interval for the true population mean difference is given by

$$-2.0 \pm (1.746 \times 1.8974) \quad \text{or} \quad -2.0 \pm 3.31, \quad \text{i.e. } (-5.31, 1.31).$$

The interpretation is in terms of repeated sampling: 90% of all intervals calculated in this way from sets of experimental data would contain the true value of $\mu_A - \mu_B$.

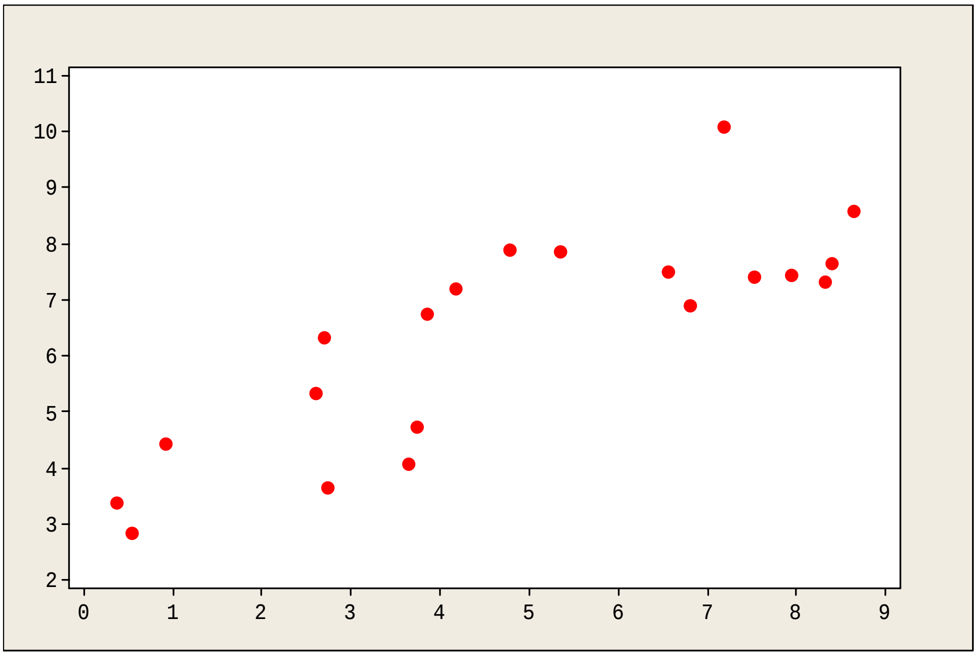
The stated null hypothesis, that $\mu_A = \mu_B$, is accepted or rejected with respect to the one-sided alternative $\mu_B > \mu_A$, at the 5% significance level according as the upper end-point of the confidence interval above is greater or less than 0.

We note that this end-point is > 0, so we do not reject the null hypothesis. There is no evidence of a difference in population mean breaking strain between suppliers A and B.

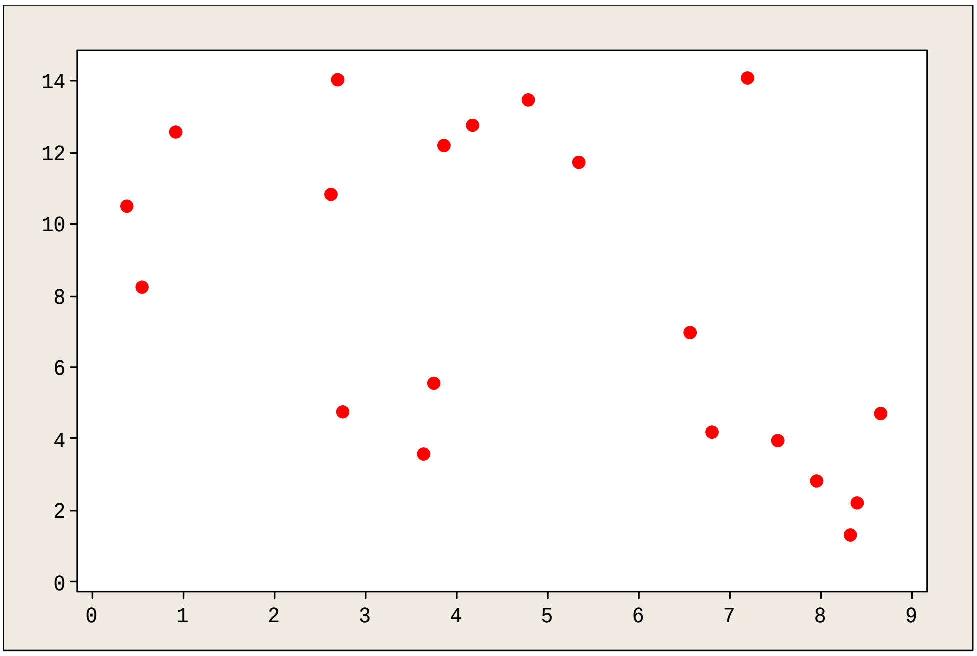
<u>Higher Certificate, Module 4, 2009. Question 3</u>

<u>Part (a)</u>   [Other appropriate diagrams were rewarded accordingly in the examination.]

(i) Dominant upward linear trend with little scatter, eg:-

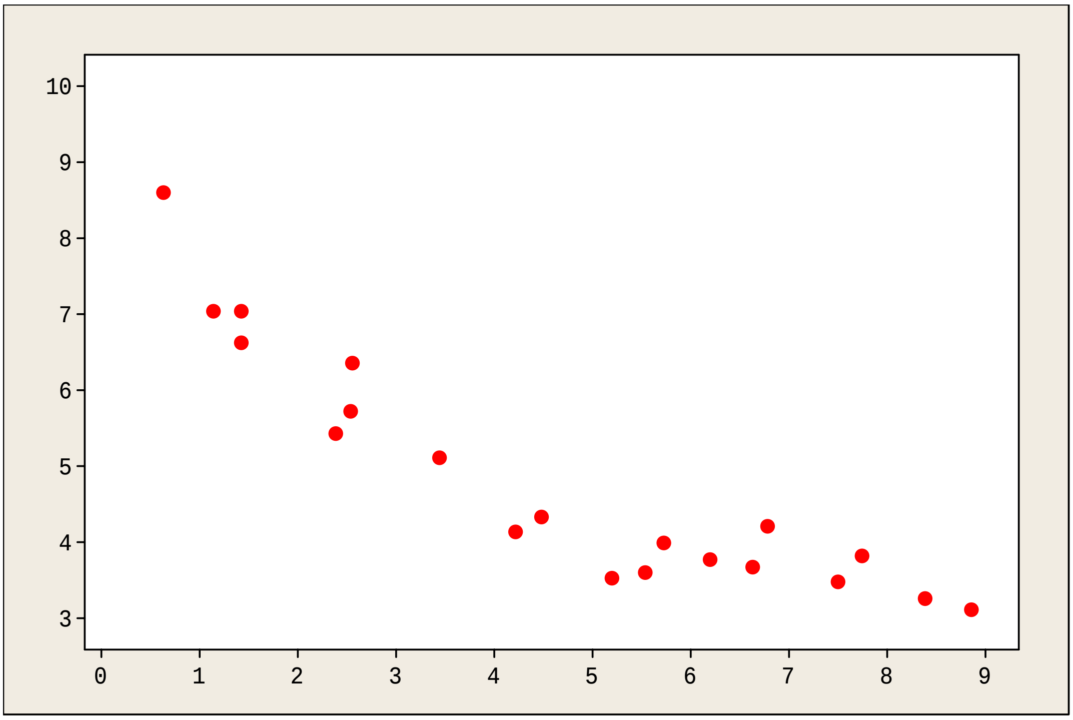


(ii) Weak downward linear trend with wide scatter, eg:-

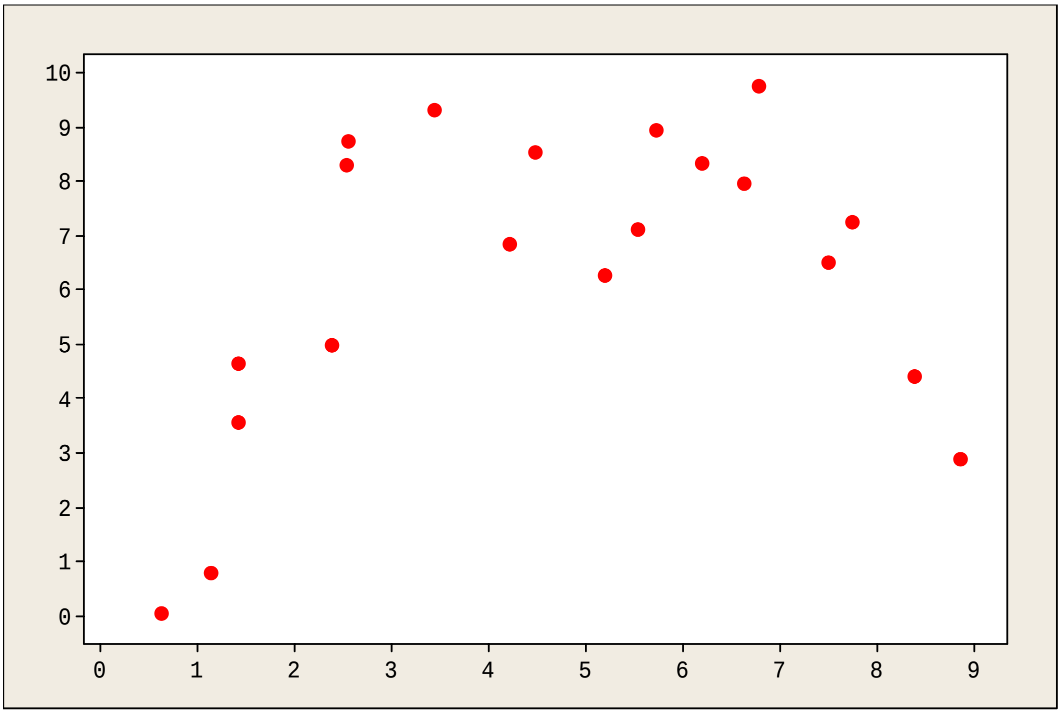


Solution continued on next page

(iii) Strong downward nonlinear trend with little scatter, eg:-



(iv) Dominant U-trend (convex or concave), eg:-



Solution continued on next page

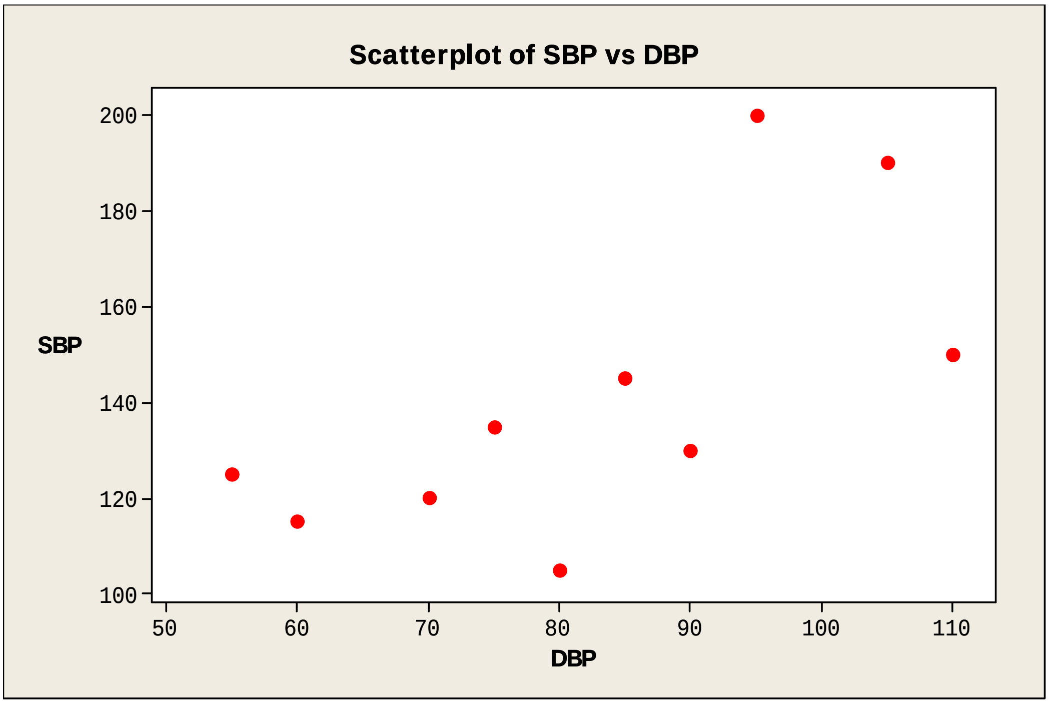
Part (b)

(i) Standard assumptions underlying the test based on the product-moment correlation coefficient are that both x- and y-variables are stochastic, and that the conditional expectation of either given the other is a linear function. (Bivariate) Normality is also required for the usual test.

(ii) Standard assumptions underlying the test of slope in simple linear regression are that the (independent) x-variable is preset without error (or analysis is conditioned on the set of x-values as fixed), the y-variable is stochastic (Normal with constant variance) and $E(Y \mid x)$ is linear in x.

Part (c)

(i) [Note the "false origin" of the scatterplot. The axes could of course be interchanged.]



The trend looks reasonably monotonic but perhaps curvilinear. This suggests using rank correlation.

Solution continued on next page

(ii) Consider first the product-moment correlation coefficient and the hypotheses H_0: $\rho = 0$ and H_1: $\rho > 0$, where ρ is the population product-moment correlation coefficient between DBP and SBP. The Society's *Statistical tables for use in examinations* show that, for a sample size of $n = 10$, the critical value to be exceeded for rejection of H_0 in favour of H_1 at the 1% significance level is 0.7155. So, as $r = 0.673$, H_0 is not rejected at this level. We conclude that the value of the population product-moment correlation coefficient between DBP and SBP may be assumed to be zero.

The calculations for Spearman's rank correlation coefficient are as follows.

DBP	55	60	70	75	80	85	90	95	105	110
DBP rank	1	2	3	4	5	6	7	8	9	10
SBP	125	115	120	135	105	145	130	200	190	150
SBP rank	4	2	3	6	1	7	5	10	9	8
Difference d_i	−3	0	0	−2	4	−1	2	−2	0	2

$\Sigma d_i^2 = 9 + 0 + \ldots + 4 = 42$.

Spearman's coefficient $r_S = 1 - \dfrac{6\Sigma d^2}{n\left(n^2-1\right)} = 1 - \dfrac{252}{990} = 0.7455$.

From the tables, this is (exactly – note that the null distribution is of course discrete) significant at the 1% level (one-sided). So we (marginally) reject the null hypothesis in this case. Thus, at the 1% level, the rank correlation test (just) picks up association between DBP and SBP, but this is not detected by the product-moment test which is based on linear association.

Higher Certificate, Module 4, 2009. Question 4

(i) The "residual error" has 7 df, so the test statistics are referred to the t_7 distribution.

The test statistic for the intercept ("constant" in the output) is 78.33/29.01 = 2.70.

The p-value is $P(|t_7| > 2.70)$. This is not directly tabulated in the Society's *Statistical tables for use in examinations* but it is certainly smaller than $P(|t_7| > 2.365)$ which, from the tables, is 0.05. Thus the value is significant at the 5% level (but, from the tables, not at the conventional stricter levels).

Similarly, the test statistic for the intercept ("x" in the output) is 54.000/5.155 = 10.48. The p-value is $P(|t_7| > 10.48)$ and this is (much) smaller than $P(|t_7| > 5.408)$ which, from the tables, is 0.001. So the slope is significant at the 0.1% level and thus also at the 5% level.

> [**Note.** The solution set out above is in terms of p-values, but these were not explicitly asked for in the question. The test statistic values 2.70 and 10.48 could therefore have been referred directly to the 5% critical points of t_7.]

There is evidence that the intercept is non-zero. There is (very strong) evidence that the slope is non-zero.

The assumptions are that the data can be regarded as a random sample from an underlying Normal distribution.

(ii) Model 1 is simple linear regression. Plot 1 shows a smooth trend with a shallow "faster than linear" curve. The straight line in the plot shows that simple linear regression will achieve a quite good explanation but is likely to overestimate sales in the middle part of the range of x and underestimate sales towards the ends of the range (perhaps seriously so at the upper end).

(iii) The p-value for the (partial, in the presence of x^2) test of the coefficient of x in Model 2 is 0.785, so this is not significant at any of the usual levels. x was strongly significant in Model 1 but is not at all significant in Model 2.

R^2 = 98.1% means that this regression model explains 98.1% of the variation in y,

or, equivalently, $R^2 = \dfrac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \dfrac{182660}{186120} = 0.981 \quad (98.1\%)$,

or, equivalently, the square of the (product-moment) correlation between the observed and fitted values of y is 0.981.

Solution continued on next page

(iv) Plot 2 appears to show that the scatter of the data (y) increases with the fitted value, and this suggests that the scatter of the data increases with x, contradicting the standard constant variance assumption.

(v) Plot 3 shows that trend of $\log_{10}$(sales) is closer to linear than the trend of raw sales in Plot 1.

This is backed up by a higher R^2 of 99.1% for Model 3.

The log data ($\log_{10}(y)$) used in Model 3 are much less variable than the original data (y). This can be seen in the very much smaller total sum of squares, which is $(n - 1) \times$(sample variance of dependent variable). In Model 3, the dependent variable is $\log_{10}(y)$, but in Model 1 it is simply y. The log data, as well as having smaller values, have much reduced spread.

(vi) The predictions (given to 3 significant figures) for $x = 10$ from the three models are as follows.

Model 1	$78.33 + (54.00\times10) = 618$
Model 2	$170 + (4\times10) + (5\times100) = 710$
Model 3	$10^{(2.16086+(10\times0.06891))} = 10^{2.84996} = 708$

Model 1 does not fit the nonlinear trend of the data. Model 2 includes a non-significant parameter; there is also the point that Plot 2 suggests that the scatter of the data appears to increase with x. Model 3 achieves the best explanation (highest R^2), has all parameters significant and plausibly constant scatter. So Model 3 seems the best model, and we would choose the estimate of 708.