11.3 The coefficient of determination is given by:

$$R^2 = \frac{SS_{REG}}{SS_{TOT}} = \frac{6.4}{10.0} = 0.64$$

This gives the proportion of the total variance explained by the model. So 64% of the variance can be explained by the model, leaving 36% of the total variance unexplained.

11.6 (i) ***Regression line***

We are given:

$$s_{xx} = 60 \qquad s_{yy} = 925,262 \qquad s_{xy} = 7,087$$

So:

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}} = \frac{7,087}{60} = 118.117$$ [1]

Since $\overline{x} = \frac{36}{9} = 4$ and $\overline{y} = \frac{3{,}960}{9} = 440$, we get:

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} = 440 - 118.117 \times 4 = -32.47 \qquad [1]$$

So the regression line is:

$$\hat{y} = -32.47 + 118.117x$$

(ii)(a) ***Confidence interval for slope parameter***

The pivotal quantity is given by:

$$\frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / s_{xx}}} \sim t_{n-2}$$

A 99% confidence interval is given by:

$$\hat{\beta} \pm t_{n-2;0.005}\sqrt{\frac{\hat{\sigma}^2}{s_{xx}}}$$

From our data:

$$\hat{\sigma}^2 = \frac{1}{7}\left(925{,}262 - \frac{7{,}087^2}{60}\right) = 12{,}595.6 \qquad [1]$$

So the 99% confidence interval is given by:

$$118.117 \pm 3.499\sqrt{\frac{12{,}595.6}{60}} = 118.117 \pm 50.696 = (67.4, 169) \qquad [2]$$

(ii)(b) ***Confidence interval for variance***

The pivotal quantity is given by:

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2} \qquad [1]$$

A 99% confidence interval is given by:

$$0.99 = P\left(\chi^2_{n-2;0.995} < \frac{(n-2)\hat{\sigma}^2}{\sigma^2} < \chi^2_{n-2;0.005}\right)$$

which gives a confidence interval of:

$$\left(\frac{(n-2)\hat{\sigma}^2}{\chi^2_{n-2;0.005}}, \frac{(n-2)\hat{\sigma}^2}{\chi^2_{n-2;0.995}}\right)$$

Substituting in, the confidence interval (to 3 SF) is:

$$\left(\frac{7\times 12{,}595.6}{20.28}, \frac{7\times 12{,}595.6}{0.9893}\right) = (4350, 89100)$$ [1]

(iii)(a) ***Partition***

The total sum of squares, $SS_{TOT} = \sum (y_i - \overline{y})^2$ is given by s_{yy} which is 925,262. [1]

The partition given at the bottom of page 25 in the *Tables* is:

$$\sum (y_i - \overline{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \overline{y})^2$$

ie $SS_{TOT} = SS_{RES} + SS_{REG}$

Now, modifying the $\hat{\sigma}^2$ formula on page 24 of the *Tables*, we have:

$$SS_{RES} = \sum (y_i - \hat{y}_i)^2 = s_{yy} - \frac{s_{xy}^2}{s_{xx}} = 925{,}262 - \frac{7{,}087^2}{60} = 88{,}169$$ [1]

Alternatively, using $\hat{\sigma}^2$ *from part (ii), we get* $SS_{RES} = (n-2)\hat{\sigma}^2 = 7\times 12{,}595.6$.

Hence:

$$SS_{REG} = 925{,}262 - 88{,}169 = 837{,}093$$ [1]

Alternatively, this could be calculated as $SS_{REG} = \frac{s_{xy}^2}{s_{xx}} = \frac{7{,}087^2}{60} = 837{,}093$.

(iii)(b) ***Proportion of variability explained by the model***

This is the coefficient of determination, R^2, which is given by:

$$R^2 = \frac{SS_{REG}}{SS_{TOT}} = \frac{837{,}093}{925{,}262} = 90.5\%$$ [1]

This tells us that 90.5% of the variation in the prices is explained by the model. Since this leaves only 9.5% from other non-model sources, it would appear that the model is a very good fit to the data. [1]

(iv)(a) ***Residuals***

The residuals, e_i, be calculated from the actual prices, y_i, and the predicted prices, $\hat{y}_i$:

$$e_i = y_i - \hat{y}_i$$

Using our regression line $\hat{y}_i = -32.47 + 118.117x_i$ from part (i), we get:

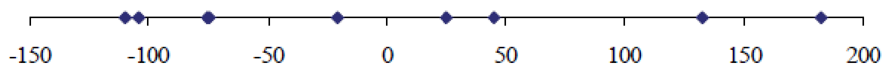
$$x = 1 \;\Rightarrow\; \hat{y} = -32.47 + 118.117 \times 1 \simeq 86 \;\Rightarrow\; \hat{e} = 131 - 86 \simeq 45 \qquad [1]$$

$$x = 4 \;\Rightarrow\; \hat{y} = -32.47 + 118.117 \times 4 \simeq 440 \;\Rightarrow\; \hat{e} = 330 - 440 \simeq -110 \qquad [1]$$

$$x = 8 \;\Rightarrow\; \hat{y} = -32.47 + 118.117 \times 8 \simeq 912 \;\Rightarrow\; \hat{e} = 1{,}095 - 912 \simeq 183 \qquad [1]$$

(iv)(b) ***Dotplot of residuals***

The dotplot is:

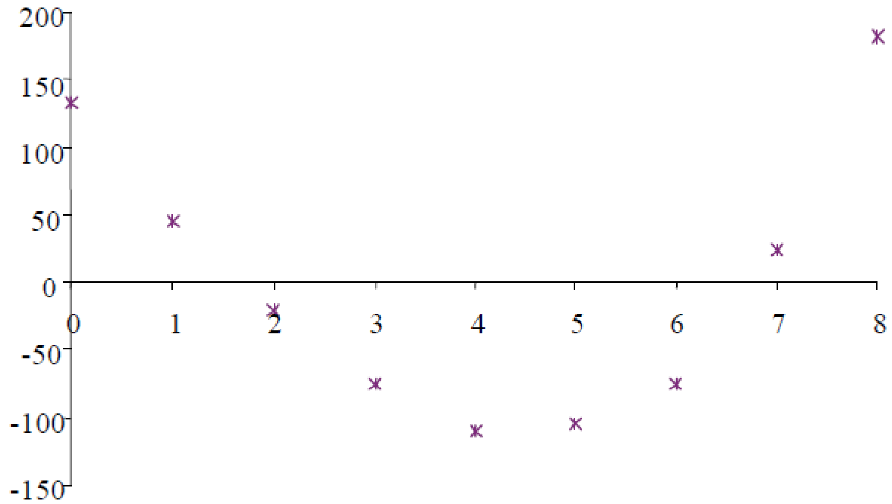


[1]

Since $e_i \sim N(0, \sigma^2)$ we would expect the dotplot to be normally distributed about zero. This does not appear to be the case, but it is difficult to tell with such a small data set. [1]

(iv)(c) ***Plot of residuals against time***

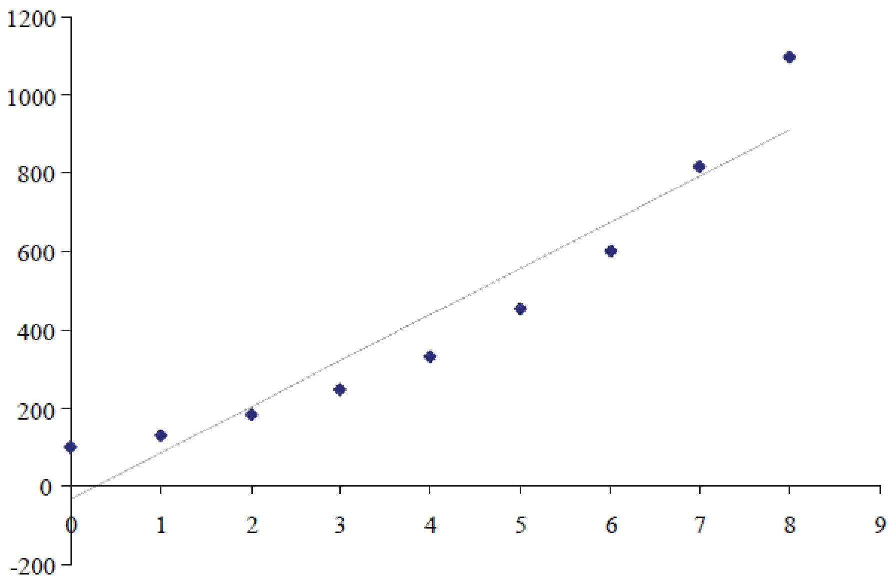
The graph is:



[1]

Clearly this is not patternless. The residuals are *not* independent of the time – this means that the linear model is definitely missing something and is not appropriate to these data. [1]

A plot of the original data (with the regression line) shows what's happening:



The price increases in an exponential (rather than linear) way. We should have used the log of the price against time instead.

11.7 (i) ***Obtain the fitted regression line***

The regression line for p on c is given by:

$$p = \hat{\alpha} + \hat{\beta}c$$

where $\hat{\beta} = \dfrac{S_{cp}}{S_{cc}}$ and $\hat{\alpha} = \overline{p} - \hat{\beta}\overline{c}$.

$$S_{cc} = \sum c^2 - \frac{\left(\sum c\right)^2}{n} = 6{,}884 - \frac{238^2}{9} = 590.2222$$

[1]

$$S_{cp} = \sum cp - \frac{\left(\sum c\right)\left(\sum p\right)}{n} = 983 - \frac{238 \times 33.4}{9} = 99.75556$$

So:

$$\hat{\beta} = \frac{99.75556}{590.2222} = 0.16901$$ [½]

$$\hat{\alpha} = \frac{33.4}{9} - 0.16901 \times \frac{238}{9} = -0.75836$$ [½]

Hence, the fitted regression line is:

$$p = 0.16901c - 0.75836$$ [1]

(ii) ***Estimate the GCSE score and its standard error***

The estimate of the average GCSE point score is obtained from the regression line:

$$\hat{P} = -0.75836 + 0.16901 \times 15 = 1.78 \qquad [1]$$

The standard error of this individual response is given by:

$$\sqrt{\left\{1 + \frac{1}{n} + \frac{(c_0 - \bar{c})^2}{S_{cc}}\right\}\hat{\sigma}^2} \qquad [1]$$

where $\hat{\sigma}^2 = \dfrac{1}{n-2}\left(S_{pp} - \dfrac{S_{cp}^2}{S_{cc}}\right) = \dfrac{1}{7}\left(25.66889 - \dfrac{99.75556^2}{590.2222}\right) = 1.25841$. [1]

Hence, the standard error is given by:

$$\sqrt{\left\{1 + \frac{1}{9} + \frac{(15 - \frac{238}{9})^2}{590.2222}\right\}1.25841}$$

$$= \sqrt{1.33302 \times 1.25841}$$

$$= \sqrt{1.67748}$$

$$= 1.29518 \qquad [1]$$

11.10 (i) ***Estimate parameters***

Now using x for i and y for $\ln P_i$, we get:

$$s_{xx} = \sum x^2 - n\bar{x}^2 = 16{,}799$$

$$s_{xy} = \sum xy - n\bar{x}\,\bar{y} = 237.39$$

$$s_{yy} = \sum y^2 - n\bar{y}^2 = 3.4322$$ [2]

So the estimates for a, b and σ^2 are:

$$\hat{b} = \frac{s_{xy}}{s_{xx}} = \frac{237.39}{16{,}799} = 0.01413$$ [1]

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = \frac{21.5953}{6} - 0.01413\left(\frac{475}{6}\right) = 2.4805$$ [1]

$$\hat{\sigma}^2 = \frac{1}{n-2}(s_{yy} - \frac{s_{xy}^2}{s_{xx}}) = \frac{1}{4}(3.4322 - \frac{237.39^2}{16{,}799}) = 0.01940$$ [1]

(ii) ***Correlation coefficient***

The correlation coefficient is:

$$r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = \frac{237.39}{\sqrt{16{,}799 \times 3.4322}} = 0.989$$ [1]

(iii) ***Confidence interval for slope parameter***

Using the result given on page 24 of the *Tables*, we have:

$$\hat{b} \pm t_{4;0.005}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = 0.01413 \pm 4.604\sqrt{\frac{0.01940}{16{,}799}}$$ [1]

This gives a confidence interval for b of $(0.00918, 0.0191)$. [1]

(iv)(a) ***Confidence interval for mean response***

If y_{365} denotes the log of the average price of a pint of lager in the country as a whole on day 365, the predicted value for y_{365} is:

$$\hat{y}_{365} = 2.4805 + 0.01413 \times 365 = 7.638$$ [1]

The distribution of $\dfrac{y_{365}-\hat{y}_{365}}{s_{365}}$ is t_4, where:

$$s_{365}^2=\left[\frac{1}{n}+\frac{(365-\bar{x})^2}{S_{xx}}\right]\hat{\sigma}^2=\left[\frac{1}{6}+\frac{[365-(475/6)]^2}{16{,}799}\right]\times 0.01940=0.09758 \qquad [1]$$

So a symmetrical 95% confidence interval for y_{365} is:

$$y_{365}=7.638\pm 2.776\sqrt{0.09758}=7.638\pm 0.867=(6.77,8.51) \qquad [1]$$

and the corresponding confidence interval for P_{365} is:

$$(e^{6.771},e^{8.505})=(870,4940) \qquad [1]$$

(iv)(b) ***Confidence interval for individual response***

If y_{365}^* denotes the log of the observed price of a pint of lager in a randomly selected bar on day 365, then $\dfrac{y_{365}^*-\hat{y}_{365}}{s_{365}^*}$ has a t_4 distribution, where:

$$s_{365}^{*2}=\left[1+\frac{1}{n}+\frac{(365-\bar{x})^2}{S_{xx}}\right]\hat{\sigma}^2=s_{365}^2+\hat{\sigma}^2=0.09758+0.01940=0.11698 \qquad [1]$$

This gives a confidence interval of:

$$y_{365}^*=7.638\pm 2.776\sqrt{0.11698}=7.638\pm 0.949=(6.69,8.59) \qquad [1]$$

So the confidence interval for P_{365}^* is:

$$(e^{6.689},e^{8.587})=(800,5360) \qquad [1]$$

Q10 (i) In bivariate data, the response variable is a random variable whose value may be influenced by the value of the explanatory variable. [2]

NOTE: This is not defined precisely in the core reading and should be marked on the basis of understanding rather than precision.

(ii) $S_{yy} = \left(270.16 - 38.4^2 / 9\right) = 106.32$ [1]

$S_{tt} = 4976 - 202^2 / 9 = 442.22$ [1]

$S_{yt} = 1011.2 - 38.4 \times 202 / 9 = 149.33$ [1]

$r = 149.33 / \sqrt{106.32 \times 442.22} = 0.6887$ [1]

(iii) $H_0 : \rho = 0 \text{ vs } H_1 : \rho \neq 0$ [1]

$z_r = \frac{1}{2} \log \frac{1+r}{1-r} = 0.8455$ [1]

$z_r \sim N\left(z_0, \frac{1}{n-3}\right) = N(0, 1/6)$ [1]

Test statistic $= 0.8455 / \left(\frac{1}{6}\right)^{0.5} = 2.071$ [1]

Compare with $Z_{0.975} = 1.96$ [1]

Therefore reject H_0 at 5% level (but note that we do not reject H_0 at 1% level). [1]

(iv) $\beta = S_{yt} / S_{yy} = 149.33 / 106.32 = 1.405$ [1]

$\alpha = 202 / 9 - 1.405 \times (38.4 / 9) = 16.45$ [1]

$t = 16.45 + 1.405y$ [1]

Q10 (i) $S_{gg} = \left(206.2462 - \frac{28.68^2}{9}\right) = 114.8526$ [1]

$S_{gd} = 15.55855 - \frac{2.97 * 28.68}{9} = 6.09415$ [1]

$\hat{\beta} = \frac{S_{gd}}{S_{gg}} = \frac{6.09415}{114.8526} = 0.05306$ [1]

$\hat{\alpha} = \bar{d} - \beta\bar{g} = \frac{2.97 - 0.05306 * 28.68}{9} = 0.1609$ [1]

So $d = 0.1609 + 0.05306g$ [1]

(ii) $S_{dd} = 1.33525 - \frac{2.97^2}{9} = 0.35515$ [1]

$\widehat{\sigma^2} = \frac{1}{7}\left(S_{dd} - \frac{S_{dg}^2}{S_{gg}}\right) = \frac{1}{7}\left(0.35515 - \frac{6.09415^2}{114.8526}\right) = 0.004542$ [1]

test statistic $= \hat{\beta} / \sqrt{\frac{\widehat{\sigma^2}}{S_{gg}}} = 0.05306 / \sqrt{\frac{0.004542}{114.8526}} = 8.438$ [1]

$t_{7;0.975} = 2.365$ (two sided) so reject $H_0 : \beta = 0$ [1]

(iii) If g_0=3 then $\hat{d}_0 = 0.1609 + 0.05306 * 3 = 0.32008$ [1]

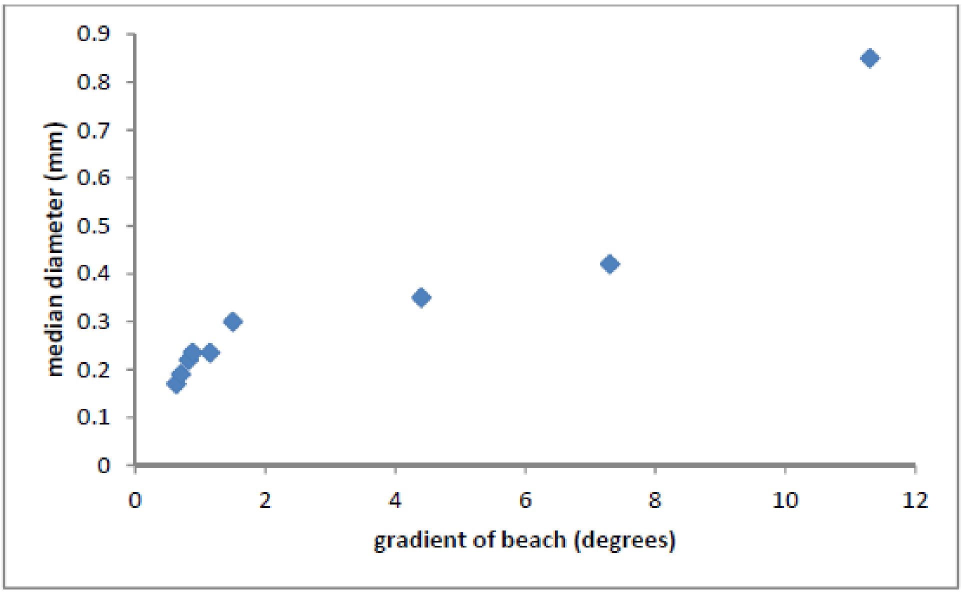
$$\mathrm{Var}\left(\hat{d}_0\right) = \left\{\frac{1}{n} + \frac{(g_0 - \bar{g})^2}{S_{gg}}\right\}\hat{\sigma}^2 = \left\{\frac{1}{9} + \frac{(3-3.187)^2}{114.8526}\right\} * 0.004542$$

$$= 5.060 \times 10^{-4}$$ [2]

$$\text{C.I.} = \hat{d}_0 \pm t_{7;0.975} * \mathrm{Var}\left(\hat{d}_0\right)^{\frac{1}{2}} = 0.32008 \pm 2.365 * \left(5.060 \times 10^{-4}\right)^{\frac{1}{2}}$$

$$= (0.267, 0.373)$$ [2]

(iv) (a)



[2]

(b) With only three observations for g>1.5, the slope is determined by a small amount of data. Getting more observations in that range would give a better analysis. [2]