Solution to Additional Exercise 2.1

1. Initial statistical model

The data consist of 17 pairs of measurements of boiling point of water and pressure, collected by the Scottish physicist James D. Forbes in the 19th century. The purpose of the data collection and analysis was to set up an equation to predict the pressure from the boiling point, which would enable estimation of mountain's heights from a (simple) measurement of boiling point. Therefore we take the boiling point as our predictor and pressure as the response (or dependent) variable. Using the notation,

$$\left.\begin{array}{rcl} y_i & = & \text{pressure} \\ x_i & = & \text{temp (boiling point)} \end{array}\right\} \text{ for observation } i,\ i = 1, \ldots, 17,$$

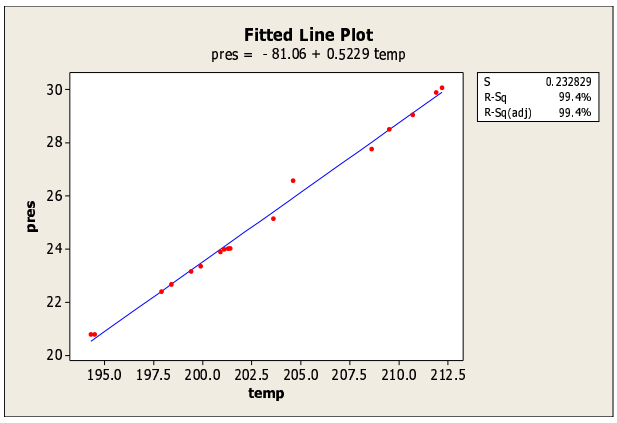
our initial statistical model is the simple linear regression,

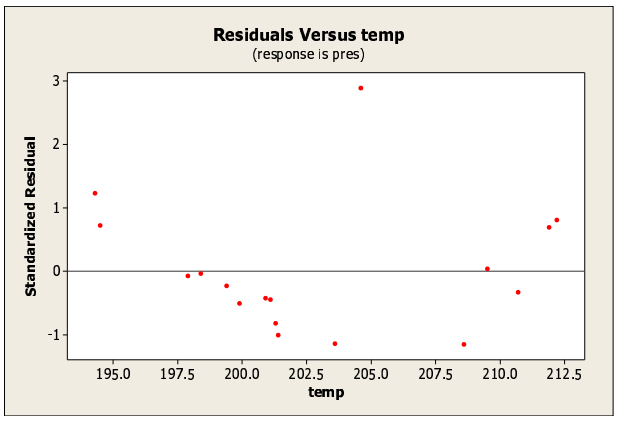
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where the errors $\varepsilon_1, \ldots, \varepsilon_{17}$ are assumed independent and identically distributed (i.i.d.) and normally distributed $\mathrm{N}(0, \sigma^2)$.

2. Analysis of simple linear regression model

The fitted line plot (from Minitab) below shows the estimated regression equation and the data points. Two features of the plot catch the eye: one of the data points (observation no. 12) is considerable above and off the fitted line, and the other points show a "smiling parabola" pattern — above the line for the lowest and highest x-values and below in-between (except for obs. 12), indicative of a curvature in the relation.

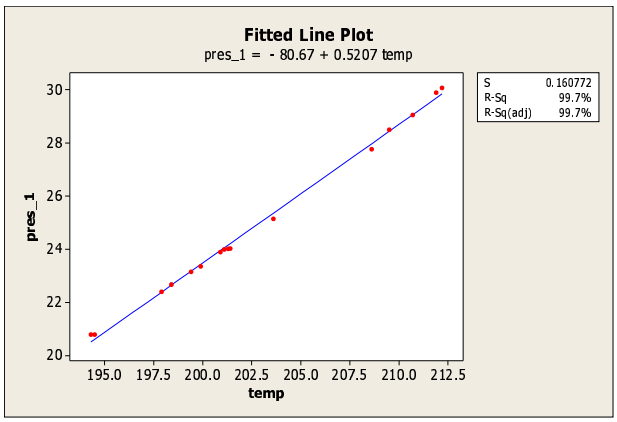


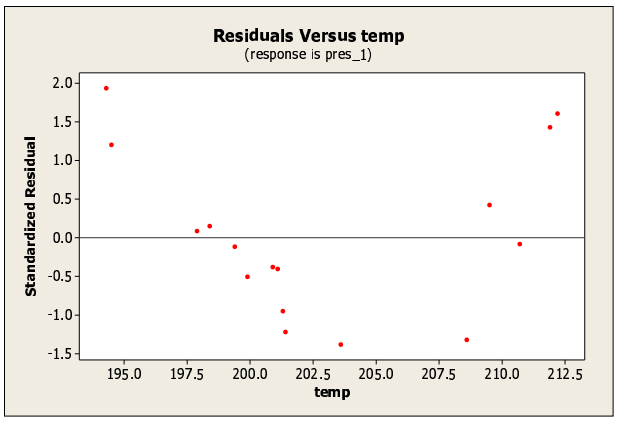


The plot of the standardised residuals against the predictor (temp) show these two features more clearly. On ground of these patterns in the residual plot, the linear statistical model is not satisfactory, despite the fact that it explains a huge proportion of the variation. The standardised residual of observation no. 12 is 2.89, and the deletion residual is 4.18, which is a very large value in such a small dataset. The critical value for a t-test at the standard 5% level based on the deletion residual is $t(1-0.025/17, 14) = 3.59$ (alternatively, $P = 2 \cdot 17 \cdot P(t(14) > 4.18) = 0.016$); therefore the observation can be considered a statistically significant outlier. Even without this test it is obvious

that the observation does not fit to the relation of the other data points, and when adjusting for the curvature in the relation things only get worse (e.g., in the quadratic regression model, the standardised and deletion residuals are 3.57 and 11.43, respectively). We therefore decide to remove observation no. 12.

When refitting the linear regression model without this observation, the curvature in the relation is even more obvious in the plots (below). In the following we examine two ways to deal with this problem: adding a quadratic term and transforming the y-variable.





3. Parameter estimates of the quadratic regression model

The quadratic regression model is

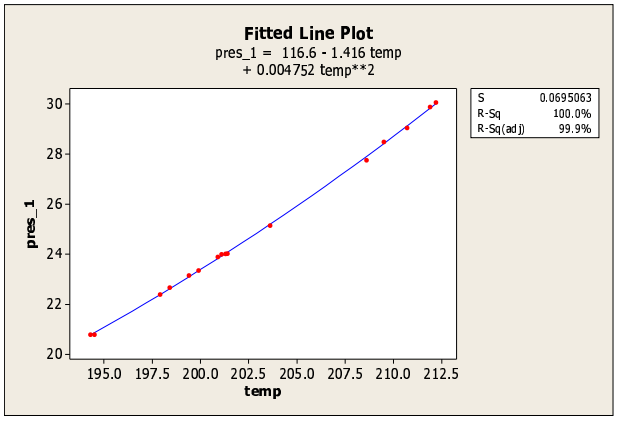
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i,$$

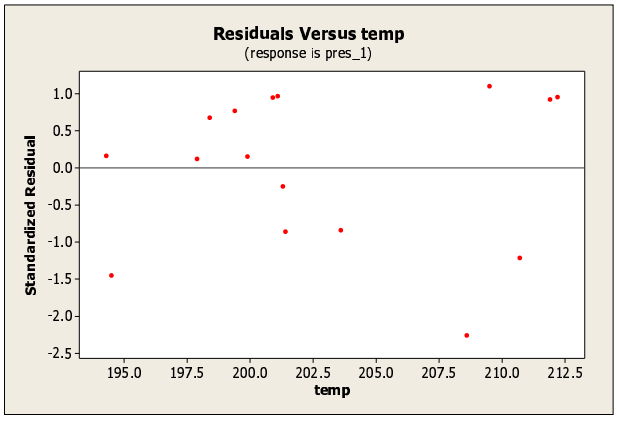
for the remaining 16 data points, and with the same assumptions on the error terms as above. The least squares estimates for the parameters are

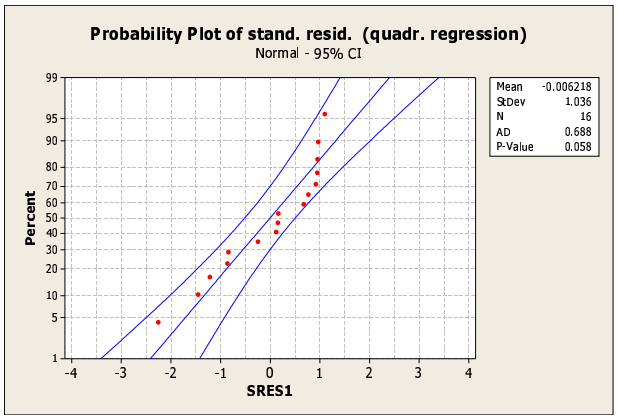
$$\begin{aligned} \hat{\beta}_2 &= 0.00475, \quad \mathrm{SE}(\hat{\beta}_2) = 0.00060, \\ \hat{\beta}_1 &= -1.417, \quad \mathrm{SE}(\hat{\beta}_1) = 0.246, \\ \hat{\beta}_0 &= 116.6, \quad \mathrm{SE}(\hat{\beta}_0) = 25.1, \\ \hat{\sigma} &= 0.0695. \end{aligned}$$

For this model, the residual plot (standardised residuals versus temp) looks satisfactory but the normal plot shows a strange pattern of points to the far right, and tests for normality give P-values close to 0.05. It may be due to the fact that the very accurate data fit leaves only little room for deviations from the fitted curve. There is one somewhat extreme residual (observation no. 14), but its deletion residual of -2.78 is nowhere near significance. It is possible to extend the model with a cubic polynomial term, which turns out weakly significant. It is not clear that the improvement warrants the increased complexity of the model.

The above parameter estimates agree reasonably well with the estimates for the Hooker data set in the RC textbook (R. Christensen: *Analysis of Variance, Design and Regression*; page 204). One notable difference is that the residual standard deviation in the present data is less than half of that in the Hooker data. Predictions from the two models are very similar within the common range of the two datasets.





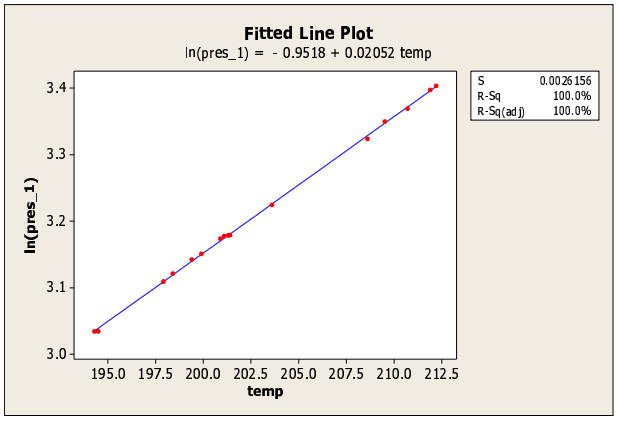


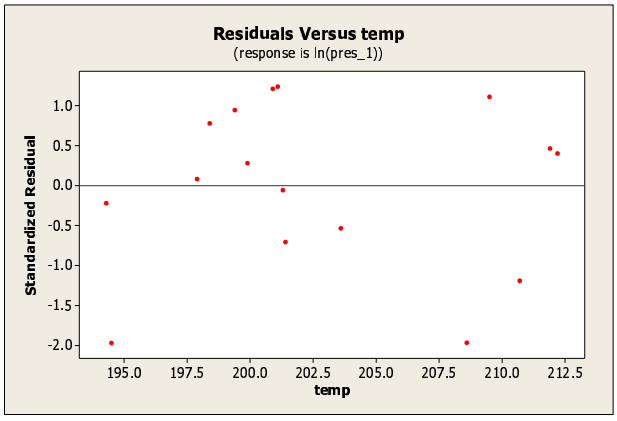
4. Analysis of log(pressure)

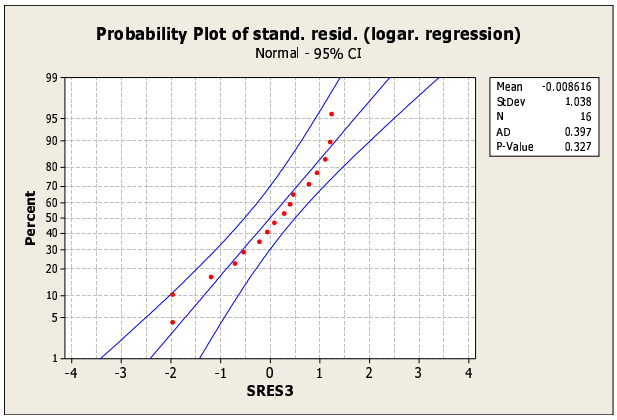
A linear regression for the (natural) log-transformed values of pressure, still excluding observation no. 12, gives the parameter estimates

$$\begin{aligned} \hat{\beta}_1 &= \quad 0.0205, \quad \mathrm{SE}(\hat{\beta}_1) = 0.0001, \\ \hat{\beta}_0 &= -0.952, \quad \mathrm{SE}(\hat{\beta}_0) = 0.023, \\ \hat{\sigma} &= \quad 0.026. \end{aligned}$$

which are also in nice agreement with the analysis of the Hooker data (RC, page 200). The residual plot and normal plot both look good (next page), and there are no standardised residuals outside the range (-2,2).







As with the quadratic regression model above, the proportion of variance explained is close to 100%. Therefore, contrary to the conclusion in the textbook, for the Forbes data there seems to be no reason to prefer the quadratic regression to log-transformation of data; on the contrary, the log-transformation seems preferable.

So far, model choice has been based solely on the diagnostics of the different models fitted. As an addendum, we examine the alternative approach of carrying out a Box-Cox analysis for transformation of the outcome in either the initial straight line model, or in a quadratic regression model. The table below gives optimal λ-values with approximate 95% confidence intervals (using the default method in Stata) based on a dataset without observation no. 12.

Statistical model	$\hat{\lambda}$ with 95% CI	
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$	0.10	$(-0.08, 0.29)$
$y_i = \beta_0 + \beta_1 x_i + + \beta_2 x_i^2 + \varepsilon_i$	-0.17	$(-1.65, 1.31)$

For the linear model, the Box-Cox analysis points towards a transformation close the log-transform (because $\hat{\lambda} \approx 0$), and there is clear evidence of the need of a transformation. For the quadratic model, the optimal λ-value is also close to zero, however in this case there is no evidence of the need of a transformation (because the 95% CI includes the value 1, and the test given by the `boxcox` command in Stata is also non-significant). When there seems to be no pressing need to transform the data, one usually prefers the analysis on the original scale. Hence, the Box-Cox analysis can be said to confirm that the principal choice for these data is between a quadratic regression on original scale and a linear regression on logarithmic scale.