Solution to Additional Exercise 2.4

1. Linear and polynomial regression

The data consist of measurements of height (m) and diameter (cm) of 18 trees (Corsican pines). The interest is in predicting height as a function of the diameter because the latter is much easier to measure. Therefore we take the diameter as our predictor and the height as the response (or dependent) variable. Using the notation,

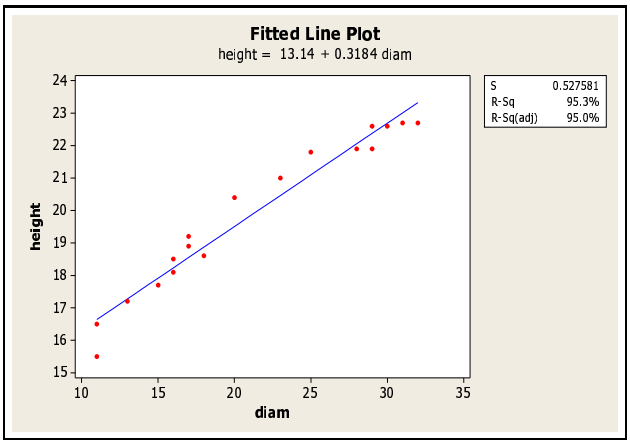
$$\left.\begin{array}{rcl} h_i & = & \text{height} \\ d_i & = & \text{diameter} \end{array}\right\} \text{ for tree } i,\ i = 1, \ldots, 18,$$

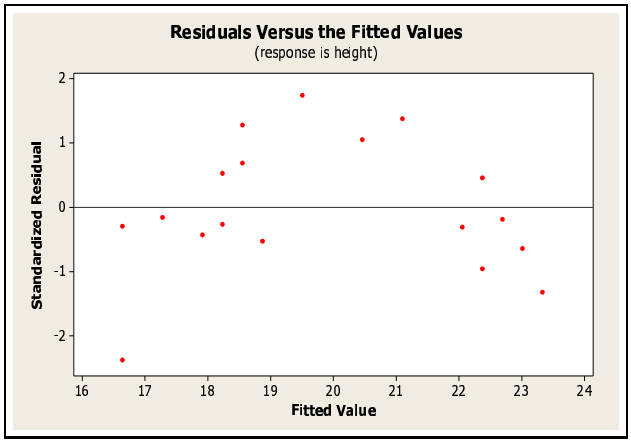
the linear, quadratic and cubic regression models take the forms (I)–(III),

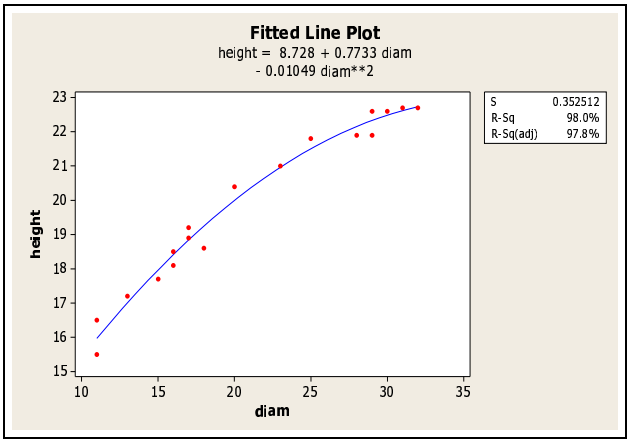
$$\begin{aligned} \text{(I):} \quad h_i &= \beta_0 + \beta_1 d_i + \varepsilon_i, \\ \text{(II):} \quad h_i &= \beta_0 + \beta_1 d_i + \beta_2 d_i^2 + \varepsilon_i, \\ \text{(III):} \quad h_i &= \beta_0 + \beta_1 d_i + \beta_2 d_i^2 + \beta_3 d_i^3 + \varepsilon_i, \end{aligned}$$

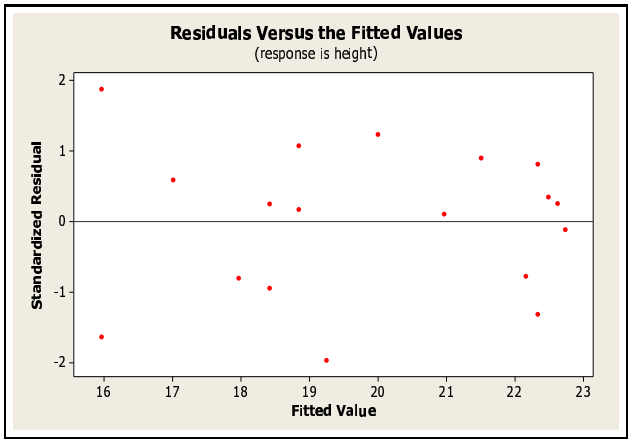
where in all models the errors $\varepsilon_1, \ldots, \varepsilon_{18}$ are assumed independent and identically distributed (i.i.d.) and normally distributed $\mathrm{N}(0, \sigma^2)$.

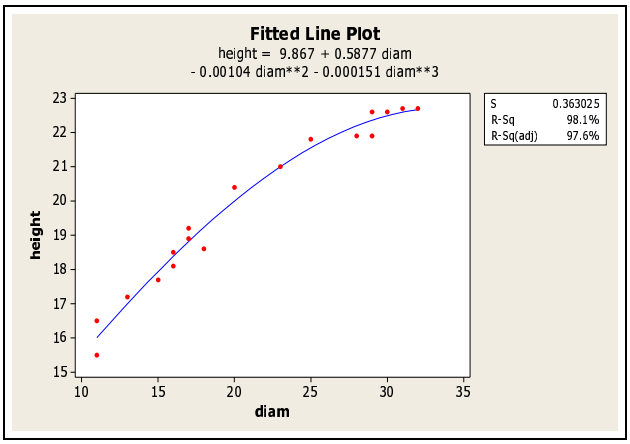
The left part of the three sets of plots below gives the estimated regressions equations and shows the fitted curves overlaid the observed values. The right part shows the usual residual plot — the standardized residuals plotted against the fitted values.

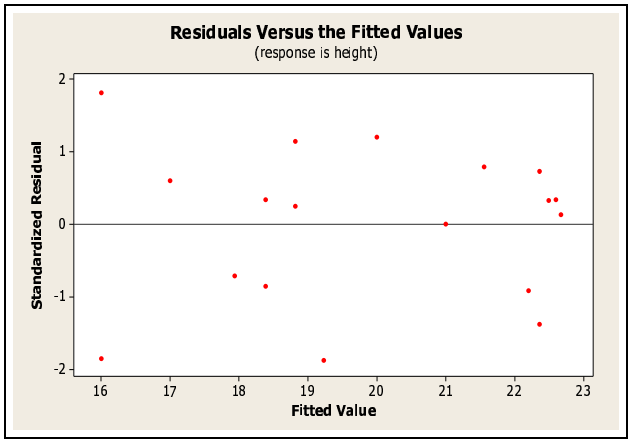












The figures show that a linear regression is inadequate because the relationship between height and diameter is not linear. This is seen most clearly in the residual plot which has a clear parabola shape. The quadratic and cubic regressions seem to fit the data well, and the figures don't lead us to prefer one over the other. The residual standard deviation and the R^2 value of the two models are similar, indicating that cubic regression does not not improve the fit much. As all three models are nested, we can test them against each other. In the quadratic regression model, the coefficient for the quadratic term (β_2) is highly significant ($t = -4.56, P < 0.0005$), further demonstrating the inadequacy of the linear regression model. In the cubic regression model, the coefficient for the cubic term (β_3) is clearly non-significant ($t = -0.38, P = 0.71$), demonstrating that the cubic regression model gives no improvement over the quadratic regression. Finally, a table of predicted values and prediction intervals (for a new tree) from the three models for diameters 11, 20 and 30 *cm*.

	Observed	Linear regression		Quadratic regression		Cubic regression	
Diameter	h	$\hat{h}$	95% PI	$\hat{h}$	95% PI	$\hat{h}$	95% PI
11	15.5,16.5	16.64	15.43–17.85	15.97	15.09–16.84	16.01	15.07–16.94
20	20.4	19.51	18.36–20.66	20.00	19.19–20.80	20.00	19.16–20.83
30	22.6	22.69	21.49–23.89	22.49	21.68–23.30	22.49	21.65–23.33

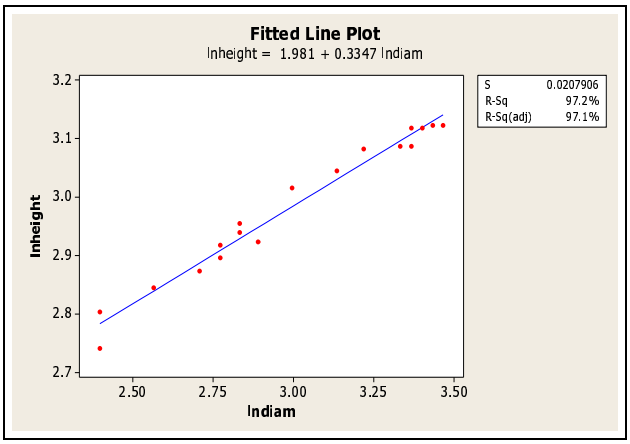
It is seen that the predictions for the quadratic and cubic models are closest to the observed values. Also, the prediction intervals are considerably shorter than in the linear regression model. Estimates and intervals are similar for the quadratic and cubic models, the intervals being slightly larger with the cubic regression model.

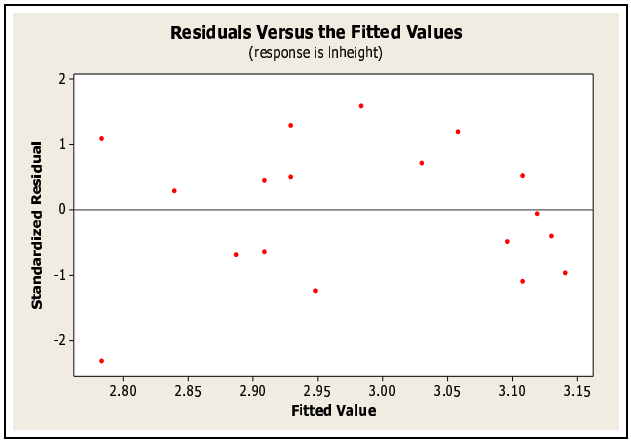
2. Log transformation of both height and diameter

The power relation: $h = \alpha\, d^{\beta}$, corresponds to a linear relation between log-transformed heights and diameters: $\log(h) = \log(\alpha) + \beta \log(d)$ (no matter whether using the natural logarithm or the base 10 logarithm), and we formulate the statistical model

$$\text{(IV):} \quad \ln(h_i) = \beta_0 + \beta_1 \ln(d_i) + \varepsilon_i,$$

where the errors $\varepsilon_1, \ldots, \varepsilon_{18}$ are subject to the same assumptions as before (although now on log-transformed scala and therefore not at all comparable to the previous errors).

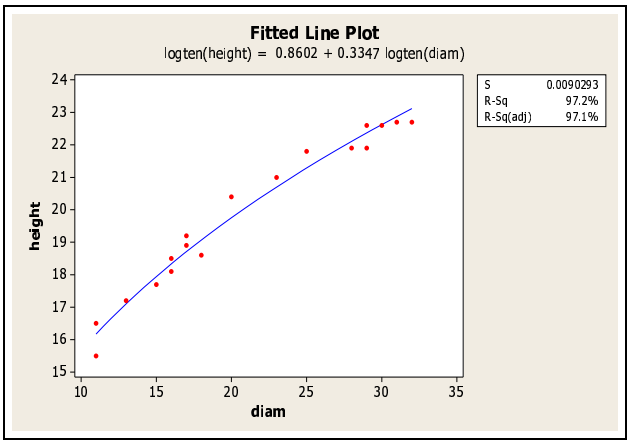




The graphs show a very nice fit of the log-scale model. The last observation (11,15.5), which is really the first and lowest value on the plot, has a slightly elevated standardized residual - the value is -2.31. The corresponding deletion residual is -2.738, corresponding to a $P = 2 \cdot 18 \cdot \mathrm{P}(t_{15} > 2.738) = 0.27$; that is, there is no evidence to say that it is outlying, and it does not seem strongly off the curve. The estimated equation on log-scale is converted to original scale as follows,

$$\ln(\hat{h}) = 1.98 + 0.335 \ln(d) \quad \Rightarrow \quad \hat{h} = \exp(1.98)\, d^{0.335} = 7.25\, d^{0.335}.$$

Our interpretation is that height is proportional to the cubic root of the diameter. The fitted curve on original scale is shown in the graph below.



The quadratic regression curve seems more curved than the power relation curve. To compare the model on log-scale with the quadratic model is not so easy, but one method is to compare the predictions from the two models. (Note that a comparison of models by their respective R^2 values does not make sense, because the models are on different scales).

	Observed	Log-scale regression		Quadratic regression	
Diameter	h	$\hat{h}$	95% PI	$\hat{h}$	95% PI
11	15.5,16.5	16.17	15.40–16.98	15.97	15.09–16.84
20	20.4	19.75	18.88–20.67	20.00	19.19–20.80
30	22.6	22.63	21.59–23.71	22.49	21.68–23.30

Some differences are seen in predictions and the intervals, but it seems difficult to say which are preferable. The two models give slightly different fitted curves, but another difference between them

is that the quadratic regression model has constant variance for the heights across all diameters, whereas power model has constant variance on the log scale, and therefore increasing variance for the heights with increasing diameters. This is also visible in the prediction intervals listed in the table. However, the range of diameters is too narrow to use this to make a preference for one model over the other one. In general, analyses of this type in forestry are usually conducted on logarithmic scale.