Polynomial Regression

8

8.1 Introduction

In each of the previous chapters we were concerned with experiments that were run as completely randomized designs for the purpose of investigating the effects of one or more treatment factors on a response variable. Analysis of variance and methods of multiple comparisons were used to analyze the data. These methods are applicable whether factor levels are qualitative or quantitative.

In this chapter, we consider an alternative approach for quantitative factors, when the set of possible levels of each factor is real-valued rather than discrete. We restrict attention to a single factor and denote its levels by x. The mean response $E[Y_{xt}]$ is modeled as a polynomial function of the level x of the factor, and the points $(x, E[Y_{xt}])$ are called the *response curve*. For example, if $E[Y_{xt}] = \beta_0 + \beta_1 x$ for unknown parameters β_0 and β_1, then the mean response is a linear function of x and the response curve is a line, called the *regression line*. Using data collected at various levels x, we can obtain estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of the intercept and slope of the line. Then $\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x$ provides an estimate of $E[Y_{xt}]$ as a function of x, and it can be used to estimate the mean response or to predict the values of new observations for any factor level x, including values for which no data have been collected. We call $\hat{y}_x$ the *fitted model* or the *estimated mean response* at the level x.

In Sect. 8.2, we look at polynomial regression and the fit of polynomial response curves to data. Estimation of the parameters in the model, using the method of least squares, is discussed in the optional Sect. 8.3. In Sect. 8.4, we investigate how well a regression model fits a given set of data via a "lack-of-fit" test. In Sect. 8.5, we look at the analysis of a simple linear regression model and test hypotheses about the values of the model parameters. Confidence intervals are also discussed. The general analysis of a higher-order polynomial regression model using a computer package is discussed in Sect. 8.6. Investigation of linear and quadratic trends in the data via orthogonal polynomials is the topic of optional Sect. 8.7. An experiment is examined in detail in Sect. 8.8, and analysis using the SAS and R software packages is done in Sects. 8.9 and 8.10, respectively.

Polynomial regression methods can be extended to experiments involving two or more quantitative factors. The mean response $E[Y_{xt}]$ is then a function of several variables and defines a *response surface* in three or more dimensions. Specialized designs are usually required for fitting response surfaces, and consequently, we postpone their discussion to Chap. 16.

A. Dean et al., *Design and Analysis of Experiments*,
Springer Texts in Statistics, DOI 10.1007/978-3-319-52250-0_8

8.2 Models

The standard model for polynomial regression is

$$\begin{aligned} Y_{xt} =& \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \epsilon_{xt}\,, \\ & \epsilon_{xt} \sim N(0, \sigma^2)\,, \\ & \epsilon_{xt}\text{'s are mutually independent} \\ & t = 1, \ldots, r_x; \quad x = x_1, \ldots, x_v. \end{aligned} \tag{8.2.1}$$

The treatment factor is observed at v different levels $x_1, \ldots, x_v$. There are r_x observations taken when the treatment factor is at level x, and Y_{xt} is the response for the tth of these. The responses Y_{xt} are modeled as independent random variables with mean

$$E[Y_{xt}] = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p\,,$$

which is a pth-degree polynomial function of the level x of the treatment factor. Since $\epsilon_{xt} \sim N(0, \sigma^2)$, it follows that

$$Y_{xt} \sim N(\beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p,\ \sigma^2)\,.$$

Typically, in a given experiment, the exact functional form of the true response curve is unknown. In polynomial regression, the true response curve is assumed to be well approximated by a polynomial function. If the true response curve is relatively smooth, then a low-order polynomial function will often provide a good model, at least for a limited range of levels of the treatment factor.

If $p = 1$ in the polynomial regression function, we have the case known as *simple linear regression*, for which the mean response is

$$E[Y_{xt}] = \beta_0 + \beta_1 x\,,$$

which is a linear function of x. This model assumes that an increase of one unit in the level of x produces a mean increase of β_1 in the response, and is illustrated in Fig. 8.1. At each value of x, there is a normal distribution of possible values of the response, the mean of which is the corresponding point, $E[Y_{xt}] = \beta_0 + \beta_1 x$, on the regression line and the variance of which is σ^2.

Consider now the data plotted in Fig. 8.2, for which polynomial regression might be appropriate. Envisage a normal distribution of possible values of Y_{xt} for each level x, and a smooth response curve connecting the distribution of their means, $E[Y_{xt}]$. It would appear that a quadratic response curve

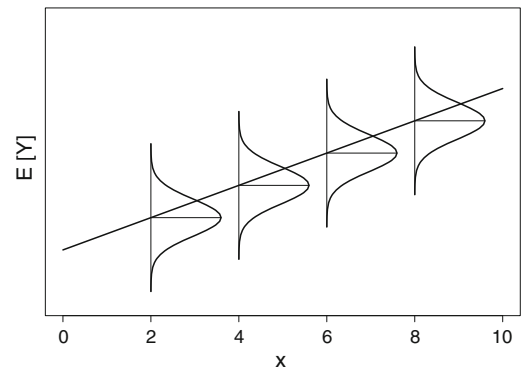


Fig. 8.1 Simple linear regression model

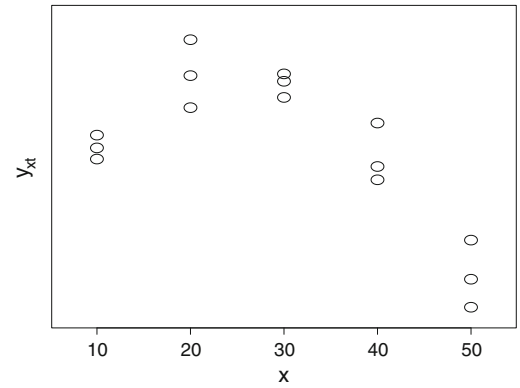


Fig. 8.2 Three hypothetical observations y_{xt} at each of five treatment factor levels

may provide a good fit to these data. This case, for which

$$E[Y_{xt}] = \beta_0 + \beta_1 x + \beta_2 x^2,$$

is called *quadratic regression*. If this model is adequate, the fitted quadratic model can be used to estimate the value of x for which the mean response is maximized, even though it may not occur at one of the x values for which data have been collected.

Although regression models can be used to estimate the mean response at values of x that have not been observed, estimation outside the range of observed x values must be done with caution. There is no guarantee that the model provides a good fit outside the observed range.

If observations are collected for v distinct levels x of the treatment factor, then any polynomial regression model of degree $p \leq v - 1$ (that is, with v or fewer parameters) can be fitted to the data. However, it is generally preferable to use the simplest model that provides an adequate fit. So for polynomial regression, lower-order models are preferred. Higher-order models are susceptible to *overfit*, a circumstance in which the model fits the data too well at the expense of having the fitted response curve vary or fluctuate excessively between data points. Over-fit is illustrated in Fig. 8.3, which contains plots for a simple linear regression model and a sixth-degree polynomial regression model, each fitted to the same set of data. The sixth-degree polynomial model provides the better fit in the sense of providing a smaller value for the sum of squared errors. However, since we may be looking at natural fluctuation of data around a true linear model, it is arguable that the simple linear regression model is actually a better model—better for predicting responses at new values of x, for example. Information concerning the nature of the treatment factor and the response variable may shed light on which model is more likely to be appropriate.

Least Squares Estimates

Once data are available, we can use the method of least squares to find estimates $\hat{\beta}_j$ of the parameters β_j of the chosen regression model. The fitted model is then

$$\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \cdots + \hat{\beta}_p x^p \,,$$

and the error sum of squares is

$$ssE = \sum_x \sum_t (y_{xt} - \hat{y}_x)^2 \,.$$

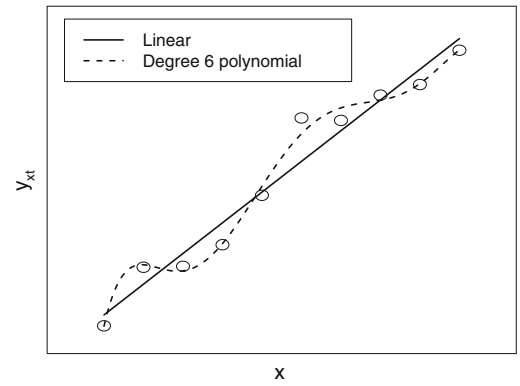


Fig. 8.3 Data and fitted linear and sixth-degree polynomial regression models

The number of error degrees of freedom is the number of observations minus the number of parameters in the model; that is, $n - (p + 1)$. The mean squared error,

$$msE = \sum_x \sum_t (y_{xt} - \hat{y}_x)^2/(n - p - 1)\,,$$

provides an unbiased estimate of σ^2.

In the following optional section, we obtain the least squares estimates of the parameters β_0 and β_1 in a simple linear regression model. However, in general we leave the determination of least squares estimates to a computer, since the formulae are not easily expressed without the use of matrices, and the hand computations are generally tedious. An exception to this occurs with the use of orthogonal polynomial models, discussed in Sect. 8.7.

Checking Model Assumptions

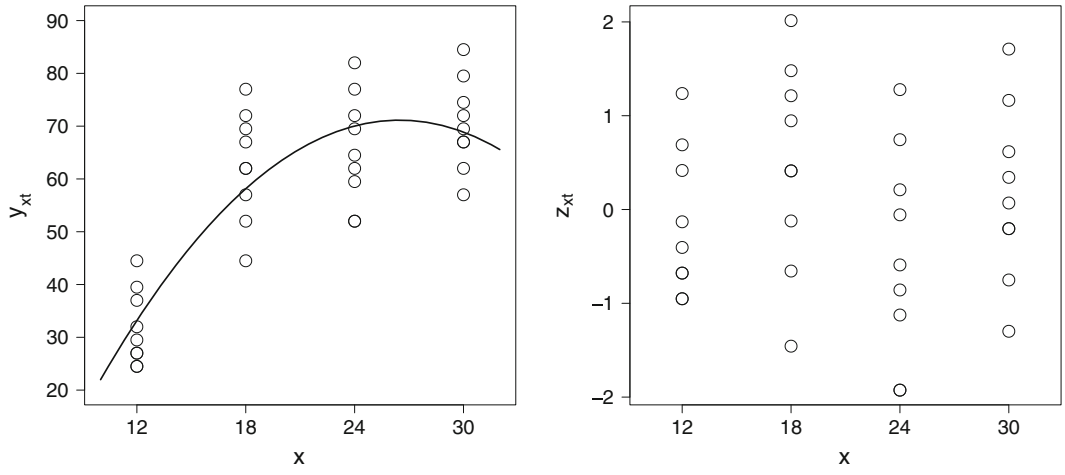
Having made an initial selection for the degree of polynomial model required in a given scenario, the model assumptions should be checked. The first assumption to check is that the proposed polynomial model for $E[Y_{xt}]$ is indeed adequate. This can done either by examination of a plot of the residuals versus x or by formally testing for model lack of fit. The standard test for lack of fit is discussed in Sect. 8.4.

If no pattern is apparent in a plot of the residuals versus x, this indicates that the model is adequate. Lack of fit is indicated if there is a clear function-like pattern. For example, suppose a quadratic model is fitted but a cubic model is needed. Any linear or quadratic pattern in the data would then be explained by the model and would not be evident in the residual plot, but the residual plot would show the pattern of a cubic polynomial function unexplained by the fitted model (see Fig. 8.4).

Residual plots can also be used to assess the assumptions on the random error terms in the model in the same way as discussed in Chap. 5. The residuals are plotted versus run order to evaluate independence of the error variables, plotted versus fitted values $\hat{y}_x$ to check the constant variance assumption and to check for outliers, and plotted versus the normal scores to check the normality assumption.

If the error assumptions are not valid, the fitted line still provides a model for mean response. However, the results of confidence intervals and hypothesis tests can be misleading. Departures from normality are generally serious problems only when the true error distribution has long tails or when prediction of a single observation is required. Nonconstant variance can sometimes be corrected via transformations, as in Chap. 5, but this may also change the order of the model that needs to be fitted.

If no model assumptions are invalidated, then analysis of variance can be used to determine whether or not a simpler model would suffice than the one postulated by the experimenter (see Sect. 8.6).



(a) Data and fitted model (b) Residual plot

Fig. 8.4 Plots for a quadratic polynomial regression model fitted to data from a cubic model

8.3 Least Squares Estimation (Optional)

In this section, we derive the normal equations for a general polynomial regression model. These equations can be solved to obtain the set of least squares estimates $\hat{\beta}_j$ of the parameters β_j. We illustrate this for the case of simple linear regression.

8.3.1 Normal Equations

For the pth-order polynomial regression model (8.2.1), the normal equations are obtained by differentiating the sum of squared errors

$$\sum_x \sum_t e_{xt}^2 = \sum_x \sum_t (y_{xt} - \beta_0 - \beta_1 x - \cdots - \beta_p x^p)^2$$

with respect to each parameter and setting each derivative equal to zero. For example, if we differentiate with respect to β_j, set the derivative equal to zero, and replace each β_i with $\hat{\beta}_i$, we obtain the jth normal equation as

$$\sum_x \sum_t x^j y_{xt} = \sum_x \sum_t x^j \left(\hat{\beta}_0 + x\hat{\beta}_1 + \cdots + x^p \hat{\beta}_p\right) . \tag{8.3.2}$$

We have one normal equation of this form for each value of j, $j = 0, 1, \ldots, p$. Thus, in total, we have $p + 1$ equations in $p + 1$ unknowns $\hat{\beta}_j$. Provided that the number of levels of the treatment factor exceeds the number of parameters in the model (that is, $v \geq p + 1$), there is a unique solution to the normal equations giving a unique set of least squares estimates, with the result that all parameters are estimable.

8.3.2 Least Squares Estimates for Simple Linear Regression

For the simple linear regression model, we have $p = 1$, and there are two normal equations obtained from (8.3.2) with $j = 0, 1$. These are

$$\sum_x \sum_t y_{xt} = n\hat{\beta}_0 + \sum_x \sum_t x\hat{\beta}_1 ,$$
$$\sum_x \sum_t x y_{xt} = \sum_x \sum_t x\hat{\beta}_0 + \sum_x \sum_t x^2\hat{\beta}_1 ,$$

where $n = \Sigma_x r_x$ denotes the total number of observations in the experiment. Dividing the first equation by n, we obtain

$$\hat{\beta}_0 = \overline{y}_{..} - \hat{\beta}_1 \overline{x}_{..} , \tag{8.3.3}$$

where $\overline{x}_{..} = \sum_x r_x x/n$. Substituting this into the second equation gives

$$\hat{\beta}_1 = \frac{\sum_x \sum_t xy_{xt} - n\overline{x}_{..}\overline{y}_{..}}{ss_{xx}} , \tag{8.3.4}$$

where $ss_{xx} = \sum_x r_x(x - \overline{x}_{..})^2$.

8.4 Test for Lack of Fit

We illustrate the lack-of-fit test via the quadratic regression model

$$E[Y_{xt}] = \beta_0 + \beta_1 x + \beta_2 x^2 .$$

If data have been collected for only three levels $x = x_1, x_2, x_3$ of the treatment factor, then the fitted model $\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$ will pass through the sample means $\overline{y}_{x.}$ computed at each value of x. This means that the predicted response $\hat{y}_x$ at the observed values of x is $\hat{y}_x = \overline{y}_{x.}$ (for $x = x_1, x_2, x_3$). This is the same fit as would be obtained using the one-way analysis of variance model, so we know that it is the best possible fit of a model to the data in the sense that no other model can give a smaller sum of squares for error, ssE.

If observations have been collected at more than three values of x, however, then the model is unlikely to fit the data perfectly, and in general, $\hat{y}_x \neq \overline{y}_{x.}$. If the values $\hat{y}_x$ and $\overline{y}_{x.}$ are too far apart relative to the amount of variability inherent in the data, then the model does not fit the data well, and there is said to be model *lack of fit*. In other words, in our example, the quadratic function is not sufficient to model the mean response $E[Y_{xt}]$.

If there is replication at one or more of the x-values, and if data are collected at more than three x-values, then it is possible to conduct a test for lack-of-fit of the quadratic model. The null hypothesis is that the quadratic model is adequate for modeling mean response; that is,

$$H_0^Q : E[Y_{xt}] = \beta_0 + \beta_1 x + \beta_2 x^2 .$$

The alternative hypothesis is that a more general model (the one-way analysis of variance model) is needed; that is,

$$H_A^Q : E[Y_{xt}] = \mu + \tau_x ,$$

where τ_x is the effect on the response of the treatment factor at level x. We fit the quadratic regression model and obtain ssE and $msE = ssE/(n-3)$. Now, MSE is an unbiased estimator of the error variance if the quadratic model is correct, but otherwise it has expected value larger than σ^2.

At each level x where more than one observation has been taken, we can calculate the sample variance s_x^2 of the responses. Each sample variance s_x^2 is an unbiased estimator of the error variance, σ^2, and these can be pooled to obtain the *pooled sample variance*,

$$s_p^2 = \left[\sum_x (r_x - 1)s_x^2\right]/(n - v)\,. \tag{8.4.5}$$

Provided that the assumption of equal error variances is valid, the pooled sample variance is an unbiased estimator of σ^2 even if the model does not fit the data well. This pooled sample variance is called the *mean square for pure error* and denoted by *msPE*. An alternative way to compute *msPE* is as the mean square for error obtained by fitting the one-way analysis of variance model.

The test of lack of fit, which is the test of H_0^Q versus H_A^Q, is based on a comparison of the two fitted models (the quadratic model and the one-way analysis of variance model), using the difference in the corresponding error sums of squares. We write *ssE* for the error sum of squares obtained from the quadratic regression model and *ssPE* for the error sum of squares from the one-way analysis of variance model. Then the *sum of squares for lack of fit* is

$$ssLOF = ssE - ssPE\,.$$

The sum of squares for pure error has $n - v$ degrees of freedom associated with it, whereas the sum of squares for error has $n - (p + 1) = n - 3$ (since there are $p + 1 = 3$ parameters in the quadratic regression model). The number of degrees of freedom for lack of fit is therefore $(n-3)-(n-v) = v-3$. The corresponding *mean square for lack of fit*,

$$msLOF = ssLOF/(v - 3),$$

measures model lack of fit because it is an unbiased estimator of σ^2 if the null hypothesis is true but has expected value larger than σ^2 otherwise.

Under the polynomial regression model (8.2.1) for $p = 2$, the decision rule for testing H_0^Q versus H_A^Q at significance level α is

$$\text{reject } H_0^Q \text{ if } msLOF/msPE > F_{v-3,n-v,\alpha}\,.$$

In general, a polynomial regression model of degree p can be tested for lack of fit as long as $v > p+1$ and there is replication for at least one of the x-levels. A test for lack of fit of the pth-degree polynomial regression model is a test of the null hypothesis

$$H_0^p : \{\, E[Y_{xt}] = \beta_0 + \beta_1 x + \cdots + \beta_p x^p;\;\; x = x_1, \ldots, x_v \,\}$$

versus the alternative hypothesis

$$H_A^p : \{\, E[Y_{xt}] = \mu + \tau_x;\;\; x = x_1, \ldots, x_v \,\}\,.$$

The decision rule at significance level α is

Table 8.1 Hypothetical data for one continuous treatment factor

x	y_{xt}			$\overline{y}_{x\cdot}$	s_x^2
10	69.42	66.07	71.70	69.0633	8.0196
20	79.91	81.45	85.52	82.2933	8.4014
30	88.33	82.01	84.43	84.9233	10.1681
40	62.59	70.98	64.12	65.8967	19.9654
50	25.86	32.73	24.39	27.6600	19.8189

Table 8.2 Test for lack of fit of quadratic regression model for hypothetical data

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
Lack of fit	2	30.0542	15.0271	1.13	0.3604
Pure error	10	132.7471	13.2747		
Error	12	162.8013			

$$\text{reject } H_0^p \text{ if } msLOF/msPE > F_{v-p-1,n-v,\alpha}\,,$$

where

$$msLOF = ssLOF/(v-p-1) \quad \text{and} \quad ssLOF = ssE - ssPE\,.$$

Here, ssE is the error sum of squares obtained by fitting the polynomial regression model of degree p, and $ssPE$ is the error sum of squares obtained by fitting the one-way analysis of variance model.

Example 8.4.1 `Lack-of-fit test for quadratic regression`

In this example we conduct a test for lack of fit of a quadratic polynomial regression model, using the hypothetical data that were plotted in Fig. 8.2 (p. 251). Table 8.1 lists the $r = 3$ observations for each of $v = 5$ levels x of the treatment factor, together with the sample mean and sample variance. The pooled sample variance (8.4.5) is

$$s_p^2 = msPE = \sum_x 2s_x^2/(15-5) = 13.2747\,,$$

and the sum of squares for pure error is therefore

$$ssPE = (15-5)msPE = 132.7471\,.$$

Alternatively, this can be obtained as the sum of squares for error from fitting the one-way analysis of variance model.

The error sum of squares ssE is obtained by fitting the quadratic polynomial regression model using a computer program (see Sects. 8.9 and 8.10 for achieving this via SAS and R software, respectively). We obtain $ssE = 162.8013$. Thus

$$ssLOF = ssE - ssPE = 162.8013 - 132.7471 = 30.0542$$

with

$$v - p - 1 = 5 - 2 - 1 = 2$$

degrees of freedom. The test for lack of fit is summarized in Table 8.2. Since the p-value is large, there is no significant lack of fit. The quadratic model seems to be adequate for these data. □

8.5 Analysis of the Simple Linear Regression Model

Suppose a linear regression model has been postulated for a given scenario, and a check of the model assumptions finds no significant violations including lack of fit. Then it is appropriate to proceed with analysis of the data.

It was shown in the optional Sect. 8.3 that the least squares estimates of the intercept and slope parameters in the simple linear regression model are

$$\hat{\beta}_0 = \overline{y}_{\cdot\cdot} - \hat{\beta}_1 \overline{x}_{\cdot\cdot} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_x \sum_t x y_{xt} - n\overline{x}_{\cdot\cdot}\overline{y}_{\cdot\cdot}}{ss_{xx}}, \tag{8.5.6}$$

where $\overline{x}_{\cdot\cdot} = \sum_x r_x x/n$ and $ss_{xx} = \sum_x r_x (x - \overline{x}_{\cdot\cdot})^2$. The corresponding estimators (random variables), which we also denote by $\hat{\beta}_0$ and $\hat{\beta}_1$, are normally distributed, since they are linear combinations of the normally distributed random variables Y_{xt}. In Exercise 1, the reader is asked to show that the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ are equal to

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}_{\cdot\cdot}^2}{ss_{xx}} \right) \quad \text{and} \quad \text{Var}(\hat{\beta}_1) = \sigma^2 \left(\frac{1}{ss_{xx}} \right). \tag{8.5.7}$$

If we estimate σ^2 by

$$msE = \frac{\sum_x \sum_t (y_{xt} - (\hat{\beta}_0 + \hat{\beta}_1 x))^2}{n-2}, \tag{8.5.8}$$

it follows that

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{msE\left(\frac{1}{n} + \frac{\overline{x}_{\cdot\cdot}^2}{ss_{xx}}\right)}} \sim t_{n-2} \quad \text{and} \quad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{msE\left(\frac{1}{ss_{xx}}\right)}} \sim t_{n-2}.$$

Thus, the decision rule at significance level α for testing whether or not the intercept is equal to a specific value a ($H_0^{\text{int}} : \{\beta_0 = a\}$ versus $H_A^{\text{int}} : \{\beta_0 \neq a\}$) is

$$\text{reject } H_0^{\text{int}} \quad \text{if} \quad \frac{\hat{\beta}_0 - a}{\sqrt{msE\left(\frac{1}{n} + \frac{\overline{x}_{\cdot\cdot}^2}{ss_{xx}}\right)}} > t_{n-2,\alpha/2} \quad \text{or} \quad < t_{n-2,1-\alpha/2}. \tag{8.5.9}$$

The decision rule at significance level α for testing whether or not the slope of the regression model is equal to a specific value b (H_0^{slp}:$\{\beta_1 = b\}$ versus H_A^{slp}:$\{\beta_1 \neq b\}$) is

$$\text{reject } H_0^{\text{slp}} \quad \text{if} \quad \frac{\hat{\beta}_1 - b}{\sqrt{msE\left(\frac{1}{ss_{xx}}\right)}} > t_{n-2,\alpha/2} \quad \text{or} \quad < t_{n-2,1-\alpha/2}. \tag{8.5.10}$$

Corresponding one-tailed tests can be constructed by choosing the appropriate tail of the t distribution and replacing $\alpha/2$ by α.

Confidence intervals at individual confidence levels of $100(1-\alpha)\%$ for β_0 and β_1 are, respectively,

$$\hat{\beta}_0 \pm t_{n-2,\alpha/2}\sqrt{msE\left(\frac{1}{n}+\frac{\overline{x}_{\cdot\cdot}^2}{ss_{xx}}\right)} \tag{8.5.11}$$

and

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2}\sqrt{msE\left(\frac{1}{ss_{xx}}\right)}. \tag{8.5.12}$$

We can use the regression line to estimate the expected mean response $E[Y_{xt}]$ at any particular value of x, say x_a; that is,

$$\widehat{E}[Y_{x_at}] = \hat{y}_{x_a} = \hat{\beta}_0 + \hat{\beta}_1 x_a\,.$$

The variance associated with this estimator is

$$\text{Var}(\hat{Y}_{x_a}) = \sigma^2\left(\frac{1}{n}+\frac{(x_a-\overline{x}_{\cdot\cdot})^2}{ss_{xx}}\right).$$

Since $\hat{Y}_{x_a}$ is a linear combination of the normally distributed random variables $\hat{\beta}_0$ and $\hat{\beta}_1$, it, too, has a normal distribution. Thus, if we estimate σ^2 by msE given in (8.5.8), we obtain a $100(1-\alpha)\%$ confidence interval for the expected mean response at x_a as

$$\hat{\beta}_0+\hat{\beta}_1 x_a \pm t_{n-2,\alpha/2}\sqrt{msE\left(\frac{1}{n}+\frac{(x_a-\overline{x}_{\cdot\cdot})^2}{ss_{xx}}\right)}. \tag{8.5.13}$$

A confidence "band" for the entire regression line can be obtained by calculating confidence intervals for the mean response at all values of x. Since this is an extremely large number of intervals, we need to use Scheffé's method of multiple comparisons. So, a $100(1-\alpha)\%$ confidence band for the regression line is given by

$$\hat{\beta}_0+\hat{\beta}_1 x_a \pm \sqrt{2F_{2,n-2,\alpha}}\sqrt{msE\left(\frac{1}{n}+\frac{(x_a-\overline{x}_{\cdot\cdot})^2}{ss_{xx}}\right)}. \tag{8.5.14}$$

The critical coefficient here is $w=\sqrt{2\,F_{2,n-2,\alpha}}$ rather than the value $w=\sqrt{(v-1)\,F_{v-1,n-v,\alpha}}$ that we had in the one-way analysis of variance model, since there are only two parameters of interest in our model (instead of linear combinations of $v-1$ pairwise comparisons) and the number of error degrees of freedom is $n-2$ rather than $n-v$.

Finally, we note that it is also possible to use the regression line to predict a future observation at a particular value x_a of x. The predicted value $\hat{y}_{x_a}$ is the same as the estimated mean response at x_a obtained from the regression line; that is,

$$\hat{y}_{x_a} = \hat{\beta}_0+\hat{\beta}_1 x_a\,.$$

The variance associated with this prediction is larger by an amount σ^2 than that associated with the estimated mean response, since the model acknowledges that the data values are distributed around their mean according to a normal distribution with variance σ^2. Consequently, we may adapt (8.5.13) to obtain a $100(1-\alpha)\%$ *prediction interval* for a future observation at x_a, as follows:

Table 8.3 Fluid flow in liters/minute for the heart–lung pump experiment

rpm	Liters per minute				
50	1.158	1.128	1.140	1.122	1.128
75	1.740	1.686	1.740		
100	2.340	2.328	2.328	2.340	2.298
125	2.868	2.982			
150	3.540	3.480	3.510	3.504	3.612

$$\hat{\beta}_0 + \hat{\beta}_1 x_a \ \pm \ t_{n-2,1-\alpha/2} \sqrt{msE \left(1 + \frac{1}{n} + \frac{(x_a - \overline{x}_{..})^2}{ss_{xx}}\right)}. \tag{8.5.15}$$

Alternatively, the prediction interval follows, because

$$\frac{\hat{Y}_{x_a} - Y_{x_a}}{\sqrt{msE \left(1 + \frac{1}{n} + \frac{(x_a - \overline{x}_{..})^2}{ss_{xx}}\right)}} \sim t(n-2)$$

under our model.

Example 8.5.1 `Heart-lung pump experiment, continued`

In Example 4.2.3, p. 73, a strong linear trend was discovered in the fluid flow rate as the number of revolutions per minute increases in a rotary pump head of an Olson heart–lung pump. Consequently, a simple linear regression model may provide a good model for the data. The data are reproduced in Table 8.3. It can be verified that

$$\overline{x}_{..} \ = \ \sum_x r_x x/n \ = \ [5(50) + 3(75) + 5(100) + 2(125) + 5(150)]/20 \ = \ 98.75\,,$$

and

$$\overline{y}_{..} = 2.2986 \ \text{ and } \sum_x \sum_t x y_{xt} = 5212.8\,.$$

So,

$$\begin{aligned} ss_{xx} \ &= \ [5(-48.75)^2 + 3(-23.75)^2 + 5(1.25)^2 + 2(26.25)^2 + 5(51.25)^2] \\ &= \ 28,093.75\,, \end{aligned}$$

giving

$$\begin{aligned} \hat{\beta}_1 \ &= \ [5212.8 \ - \ 20(98.75)(2.2986)]/[28{,}093.75] \\ &= \ 673.065/28{,}093.75 \ = \ 0.02396\,. \end{aligned}$$

The mean square for error (8.5.8) for the regression model is best calculated by a computer package. It is equal to $msE = 0.001177$, so the estimated variance of $\hat{\beta}_1$ is

$$\text{Var}(\hat{\beta}_1) \ = \ msE/ss_{xx} \ = \ (0.001177)/28{,}093.75 \ = \ 0.000000042\,.$$

A 95% confidence interval for β_1 is then given by (8.5.12), as

$$\begin{aligned} 0.02396 \ &\pm \ t_{18,.025}\sqrt{0.000000042}\,, \\ 0.02396 \ &\pm \ (2.101)(0.00020466)\,, \\ &(\,0.02353,\ 0.02439\,)\,. \end{aligned}$$

To test the null hypothesis $H_0^{\text{slp}} : \{\beta_1 = 0\}$, against the one-sided alternative hypothesis $H_A^{\text{slp}} : \{\beta_1 > 0\}$ that the slope is greater than zero at significance level $\alpha = 0.01$, we use a one-sided version of the decision rule (8.5.10) and calculate

$$\frac{\hat{\beta}_1 - 0}{\sqrt{msE\left(\frac{1}{ss_{xx}}\right)}} = \frac{0.02396}{0.00020466} = 117.07\,,$$

and since this is considerably greater than $t_{18,0.01} = 2.552$, we reject H_0^{slp}. We therefore conclude that the slope of the regression line is greater than zero, and the fluid flow increases as the revolutions per minute increase. □

8.6 Analysis of Polynomial Regression Models

8.6.1 Analysis of Variance

Suppose a polynomial regression model has been postulated for a given experiment, and the model assumptions appear to be satisfied, including no significant lack of fit. Then it is appropriate to proceed with analysis of the data. A common objective of the analysis of variance is to determine whether or not a lower-order model might suffice. One reasonable approach to the analysis, which we demonstrate for the quadratic model ($p = 2$), is as follows.

First, test the null hypothesis $H_0^L : \beta_2 = 0$ that the highest-order term $\beta_2 x^2$ is not needed in the model so that the simple linear regression model is adequate. If this hypothesis is rejected, then the full quadratic model is needed. Otherwise, testing continues and attempts to assess whether an even simpler model is suitable. Thus, the next step is to test the hypothesis $H_0 : \beta_1 = \beta_2 = 0$. If this is rejected, the simple linear regression model is needed and adequate. If it is not rejected, then x is apparently not useful in modeling the mean response.

Each test is constructed in the usual way, by comparing the error sum of squares of the full (quadratic) model with the error sum of squares of the reduced model corresponding to the null hypothesis being true. For example, to test the null hypothesis $H_0^L : \beta_2 = 0$ that the simple linear regression model is adequate versus the alternative hypothesis H_A^L that the linear model is not adequate, the decision rule at significance level α is

$$\text{reject } H_0^L \quad \text{if} \quad ms(\beta_2)/msE > F_{1,n-v,\alpha}\,,$$

where the mean square $ms(\beta_2) = ss(\beta_2)/1$ is based on one degree of freedom, and

$$ss(\beta_2) = ssE_1 - ssE_2\,,$$

where ssE_1 and ssE_2 are the error sums of squares obtained by fitting the models of degree one and two, respectively.

Table 8.4 Analysis of variance table for polynomial regression model of degree p. Here ssE_b denotes the error sum of squares obtained by fitting the polynomial regression model of degree b

Source of variation	Degrees of freedom	Sum of square	Mean squares	Ratio
β_p	1	$ssE_{p-1} - ssE$	$ms(\beta_p)$	$ms(\beta_p)/msE$
β_{p-1}, β_p	2	$ssE_{p-2} - ssE$	$ms(\beta_{p-1}, \beta_p)$	$ms(\beta_{p-1}, \beta_p)/msE$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\beta_2, \ldots, \beta_p$	$p-1$	$ssE_1 - ssE$	$ms(\beta_2, \ldots, \beta_p)$	$ms(\beta_2, \ldots, \beta_p)/msE$
Model	p	$ssE_0 - ssE$	$ms(\beta_1, \ldots, \beta_p)$ $ms(\beta_1, \ldots, \beta_p)/msE$	
Error	$n-p-1$	ssE	msE	
Total	$n-1$	$sstot$		

Similarly, the decision rule at significance level α for testing $H_0 : \beta_1 = \beta_2 = 0$ versus the alternative hypothesis that H_0 is false is

$$\text{reject } H_0 \quad \text{if} \quad ms(\beta_1, \beta_2)/msE > F_{2,n-v,\alpha}\,,$$

where the mean square $ms(\beta_1, \beta_2) = ss(\beta_1, \beta_2)/2$ is based on 2 degrees of freedom, and

$$ss(\beta_1, \beta_2) = (ssE_0 - ssE_2)/2\,,$$

and ssE_0 and ssE_2 are the error sums of squares obtained by fitting the models of degree zero and two, respectively.

The tests are generally summarized in an analysis of variance table, as indicated in Table 8.4 for the polynomial regression model of degree p. In the table, under sources of variability, "Model" is listed rather than "$\beta_1, \ldots, \beta_p$" for the test of $H_0 : \beta_1 = \cdots = \beta_p = 0$, since this is generally included as standard output in a computer package. Also, to save space, we have written the error sum of squares as ssE for the full model, rather than indicating the order of the model with a subscript p. Analysis of variance for quadratic regression ($p = 2$) is illustrated in the following example.

Example 8.6.1 `Analysis of variance for quadratic regression`

Consider the hypothetical data in Table 8.1, p. 256, with three observations for each of the levels $x = 10, 20, 30, 40, 50$. For five levels, the quartic model is the highest-order polynomial model that can be fitted to the data. However, a quadratic model was postulated for these data, and a test for lack of fit of the quadratic model, conducted in Example 8.4.1, suggested that this model is adequate.

The analysis of variance for the quadratic model is given in Table 8.5. The null hypothesis $H_0^L : \{\beta_2 = 0\}$ is rejected, since the p-value is less than 0.0001. So, the linear model is not adequate, and the quadratic model is needed. This is no surprise, based on the plot of the data shown in Fig. 8.5. Now, suppose the objective of the experiment was to determine how to maximize mean response. From the data plot, it appears that the maximum response occurs within the range of the levels x that were observed. The fitted quadratic regression model can be obtained from a computer program, as illustrated in Sects. 8.9 and 8.10 for the SAS and R programs, respectively. The fitted model is

Table 8.5 Analysis of variance for the quadratic model

Source of variation	Degrees of freedom	Sum of square	Mean squares	Ratio	p-value
β_2	1	3326.2860	3326.2860	245.18	0.0001
Model	2	6278.6764	3139.3382	231.40	0.0001
Error	12	162.8013	13.5668		
Total	14	6441.4777			

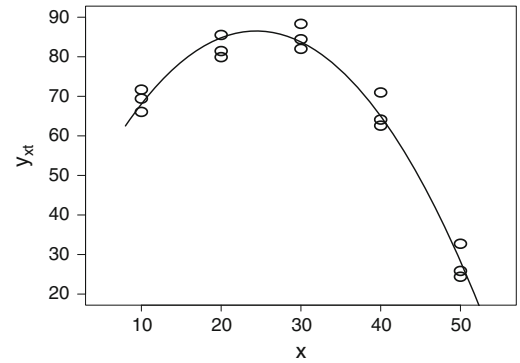


Fig. 8.5 Quadratic polynomial regression model fitted to hypothetical data

$$\hat{y}_x = 33.43333 + 4.34754x - 0.08899x^2\,,$$

and is plotted in Fig. 8.5 along with the raw data. The fitted curve achieves its maximum value when x is around 24.4, which should provide a good estimate of the level x that maximizes mean response. Further experimentation involving levels around this value could now be done. □

The adequacy of a regression model is sometimes assessed in terms of the proportion of variability in the response variable that is explained by the model. This proportion, which is the ratio of the model sum of squares to the sum of squares total, is called the *coefficient of multiple determination*, or the R^2*-value*. In the notation of Table 8.4,

$$R^2 = (ssE_0 - ssE)/sstot = ss(\beta_1, \ldots, \beta_p)/sstot\,. \tag{8.6.1}$$

For simple linear regression,

$$R^2 = ss(\beta_1)/sstot$$

is called the *coefficient of determination*, and in this case $R^2 = \mathrm{r}^2$, where

$$\mathrm{r} = ss_{xy}/\sqrt{ss_{xx}\,ss_{yy}}$$

is the *sample correlation coefficient*, or *Pearson product-moment correlation coefficient*.

8.6.2 Confidence Intervals

When the model is fitted via a computer program, the least squares estimates of $\hat{\beta}_j$ and their corresponding standard errors (estimated standard deviations) usually form part of the standard computer

output. If the model assumptions are satisfied, then

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\widehat{\text{Var}}\,(\hat{\beta}_j)}} \sim t_{n-p-1}\,.$$

Individual confidence intervals can be obtained for the model parameters, as we illustrated in Sect. 8.5 for the simple linear regression model. The general form is

$$\hat{\beta}_j \;\pm\; t_{n-2,\alpha/2}\sqrt{\widehat{\text{Var}}\,(\hat{\beta}_j)}\,.$$

Most programs will also allow calculation of the estimated mean response at any value of $x = x_a$ together with its standard error, and also calculation of the predicted response at $x = x_a$ plus its standard error. Confidence and prediction intervals for these can again be calculated using the t_{n-p-1} distribution. The confidence interval formula for mean response at $x = x_a$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x_a + \cdots + \hat{\beta}_p x_a^p \;\pm\; t_{n-p-1,\alpha/2}\sqrt{\widehat{\text{Var}}\,(\hat{Y}_{x_a})}$$

and the prediction interval formula for a new observation at $x = x_a$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x_a + \cdots + \hat{\beta}_p x_a^p \;\pm\; t_{n-p-1,\alpha/2}\sqrt{\widehat{\sigma}^2 + \widehat{\text{Var}}\,(\hat{Y}_{x_a})}\,.$$

The overall confidence level for all the intervals combined should be computed via the Bonferroni method as usual. A confidence band for the regression line is obtained by calculating confidence intervals for the estimated mean response at all values of x, using the critical coefficient for Scheffé's method; that is,

$$\hat{\beta}_0 + \hat{\beta}_1 x + \cdots + \hat{\beta}_p x^p \;\pm\; \sqrt{(p+1)\,F_{p+1,n-p-1,\alpha}}\sqrt{\widehat{\text{Var}}\,(\hat{Y}_x)}\,.$$

8.7 Orthogonal Polynomials and Trend Contrasts (Optional)

The normal equations for polynomial regression were presented in Eq. (8.3.2). It was noted that solving the equations can be tedious. However, the factor levels can be transformed in such a way that the least squares estimates have a simple algebraic form and are easily computed. Furthermore, the parameter estimators become uncorrelated and are multiples of the corresponding trend contrast estimators. This transformation is illustrated in this section for simple linear regression and for quadratic regression, when the factor levels x are equally spaced with equal numbers r of observations per level.

8.7.1 Simple Linear Regression

Consider the simple linear regression model, for which

$$Y_{xt} = \beta_0 + \beta_1 x + \epsilon_{xt}\,; \quad x = x_1, \ldots, x_v; \quad t = 1, \ldots, r\,. \tag{8.7.17}$$

When there are r observations on each of the v quantitative levels x of the treatment factor, the average value of x is $\overline{x}_{\cdot\cdot} = r\sum_x x/n = \sum_x x/v$. The transformation $z_x = x - \overline{x}_{\cdot\cdot}$ centers the levels x at zero, so that $\sum_x z_x = 0$. This makes the estimates of the slope and intercept parameters uncorrelated (or orthogonal). We can replace x in model (8.7.17) by z_x, so that the "centered" form of the model is

$$Y_{xt} = \beta_0^* + \beta_1^* z_x + \epsilon_{xt}\,; \quad x = x_1, \ldots, x_v; \quad t = 1, \ldots, r\,. \tag{8.7.18}$$

A transformation of the independent variable changes the interpretation of some of the parameters. For example, in the simple linear regression model (8.7.17), β_0 denotes mean response when $x = 0$, whereas in the transformed model (8.7.18), β_0^* denotes mean response when $z_x = 0$, which occurs when $x = \overline{x}_{\cdot\cdot}$.

The normal equations corresponding to $j = 0$ and $j = 1$ for the centered model are obtained from (8.3.2) with z_x in place of x. Thus, we have

$$\begin{aligned}\textstyle\sum_x\sum_t y_{xt} &= \textstyle\sum_x\sum_t \left(\hat\beta_0^* + z_x\hat\beta_1^*\right)\,, \;= vr\hat\beta_0^* \\ \textstyle\sum_x\sum_t z_x y_{xt} &= \textstyle\sum_x\sum_t z_x \left(\hat\beta_0^* + z_x\hat\beta_1^*\right) = \sum_x r z_x^2\hat\beta_1^*\,.\end{aligned}$$

Solving these equations gives the least squares estimates as

$$\hat\beta_0^* = \overline{y}_{\cdot\cdot} \quad \text{and} \quad \hat\beta_1^* = \frac{1}{r\sum_x z_x^2}\sum_x\sum_t z_x y_{xt}\,.$$

Now,

$$\mathrm{Cov}\left(Y_{\cdot\cdot}, \sum_x\sum_t z_x Y_{xt}\right) = \sum_x\sum_t z_x \mathrm{Cov}(Y_{xt}, Y_{xt}) = r\sigma^2\sum_x z_x = 0\,,$$

so the estimators $\hat\beta_0^*$ and $\hat\beta_1^*$ are uncorrelated.

We now consider a special case to illustrate the relationship of the slope estimator with the linear trend contrast that we used in Sect. 4.2.4. Suppose equal numbers of observations are collected at the three equally spaced levels

$$x_1 = 5, \quad x_2 = 7, \quad \text{and} \quad x_3 = 9\,.$$

Then $\overline{x}_{\cdot\cdot} = 7$, so

$$z_5 = -2, \quad z_7 = 0, \quad \text{and} \quad z_9 = 2\,.$$

These values are twice the corresponding linear trend contrast coefficients $(-1, 0, 1)$ listed in Appendix A.2. Now, $r = 2$, so $r\sum_x z_x^2 = 8r$, and

$$\begin{aligned}\hat\beta_1^* &= \frac{1}{r\sum_x z_x^2}\sum_x\sum_t z_x y_{z_x t} = \frac{1}{8r}(2y_{9\cdot} - 2y_{5\cdot}) \\ &= \frac{1}{4}(\overline{y}_{9\cdot} - \overline{y}_{5\cdot})\,,\end{aligned}$$

which is a quarter of the value of the linear trend contrast estimate. It follows that β_1^* and the linear trend contrast have the same normalized estimate and hence also the same sum of squares. Thus, testing $H_0 : \beta_1^* = 0$ under model (8.7.18) is analogous to testing the hypothesis $H_0 : \tau_3 - \tau_1 = 0$ of no linear

trend effect under the one-way analysis of variance model

$$Y_{it} = \mu + \tau_i + \epsilon_{it}\,; \quad i = 1, 2, 3\,; \quad t = 1, 2\,,$$

where τ_i is the effect on the response of the ith coded level of the treatment factor. The one difference is that in the first case, the model is the linear regression model ($p = 1$), while in the second case, the model is the one-way analysis of variance model, which is equivalent to a model of order $p = v-1 = 2$. Thus the two models will not yield the same mean squared error, so the F-statistics will not be identical.

8.7.2 Quadratic Regression

Consider the quadratic regression model, for which

$$Y_{xt} = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon_{xt}\,. \tag{8.7.19}$$

Assume that the treatment levels $x = x_1, \ldots, x_v$ are equally spaced, with r observations per level. To achieve orthogonality of estimates, it is necessary to transform both the linear and the quadratic independent variables.

Let $z_x = x - \overline{x}_{\cdot\cdot}$ as in the case of simple linear regression, so that again $\sum_x z_x = 0$. Similarly, define

$$z_x^{(2)} = z_x^2 - \sum_x z_x^2 / v\,.$$

Then $\sum_x z_x^{(2)} = 0$. Also, writing z_i for the ith value of z_x in rank order, we note that since the levels x are equally spaced,

$$z_i = -z_{v+1-i} \quad \text{and} \quad z_i^{(2)} = z_{v+1-i}^{(2)}\,,$$

so $\sum_x z_x z_x^{(2)} = 0$. These conditions give uncorrelated parameter estimators. To see this, consider the transformed model

$$Y_{xt} = \beta_0^* + \beta_1^* z_x + \beta_2^* z_x^{(2)} + \epsilon_{xt}\,. \tag{8.7.20}$$

The normal equations (8.3.2) become

$$\begin{aligned}
\textstyle\sum_x \sum_t y_{xt} &= \textstyle\sum_x \sum_t \left(\hat\beta_0^* + z_x \hat\beta_1^* + z_x^{(2)} \hat\beta_2^*\right) &&= vr\hat\beta_0^*\,,\\
\textstyle\sum_x \sum_t z_x y_{xt} &= \textstyle\sum_x \sum_t z_x \left(\hat\beta_0^* + z_x \hat\beta_1^* + z_x^{(2)} \hat\beta_2^*\right) &&= r\textstyle\sum_x z_x^2 \hat\beta_1^*\,,\\
\textstyle\sum_x \sum_t z_x^{(2)} y_{xt} &= \textstyle\sum_x \sum_t z_x^{(2)} \left(\hat\beta_0^* + z_x \hat\beta_1^* + z_x^{(2)} \hat\beta_2^*\right) &&= r\textstyle\sum_x (z_x^{(2)})^2 \hat\beta_2^*\,.
\end{aligned}$$

The least squares estimates, obtained by solving the normal equations, are

$$\hat\beta_0^* = \overline{y}_{\cdot\cdot}\,, \quad \hat\beta_1^* = \frac{\sum_x \sum_t z_x y_{xt}}{r \sum_x z_x^2}\,, \quad \hat\beta_2^* = \frac{\sum_x \sum_t z_x^{(2)} y_{xt}}{r \sum_x (z_x^{(2)})^2}\,.$$

The estimators $\hat\beta_0^*$ and $\hat\beta_1^*$ are unchanged from the simple linear regression model (8.7.18), so they remain uncorrelated. Similarly, $\hat\beta_0^*$ and $\hat\beta_2^*$ are uncorrelated, because

$$\text{Cov}\left(Y_{\cdot\cdot},\ \sum_x \sum_t z_x^{(2)} Y_{xt}\right) = r\sigma^2 \sum_x z_x^{(2)} = 0.$$

Observe that $\text{Cov}(\hat{\beta}_1^*, \hat{\beta}_2^*)$ is also zero, since it is proportional to

$$\text{Cov}\left(\sum_x \sum_t z_x Y_{xt}, \sum_x \sum_t z_x^{(2)} Y_{xt}\right) = r\sigma^2 \sum_x z_x z_x^{(2)} = 0.$$

The transformed variables z_x and $z_x^{(2)}$ are called *orthogonal polynomials*, because they are polynomial functions of the levels x and give rise to uncorrelated parameter estimators $\hat{\beta}_0^*$, $\hat{\beta}_1^*$, and $\hat{\beta}_2^*$. It was illustrated in the previous subsection on simple linear regression that the values z_x are multiples of the coefficients of the linear trend contrast. Likewise, the values $z_x^{(2)}$ are multiples of the coefficients of the quadratic trend contrast. For example, suppose we have $r = 17$ observations on the equally spaced levels

$$x_1 = 12, \quad x_2 = 18, \quad x_3 = 24, \quad x_4 = 30\,.$$

Then $z_x = x - \overline{x}_{\cdot\cdot}$, so

$$z_{12} = -9, \quad z_{18} = -3, \quad z_{24} = 3, \quad z_{30} = 9\,.$$

These are 3 times the linear trend contrast coefficients listed in Appendix A.2. Also, $\sum_x z_x^2/v = 45$, so

$$z_{12}^{(2)} = 36, \quad z_{18}^{(2)} = -36, \quad z_{24}^{(2)} = -36, \quad z_{30}^{(2)} = 36\,,$$

which are 36 times the quadratic trend contrasts.

As in the simple linear regression case, one can likewise show that the least squares estimates $\hat{\beta}_1^*$ and $\hat{\beta}_2^*$ are constant multiples of the corresponding linear and quadratic trend contrast estimates $\hat{\tau}_3 - \hat{\tau}_1$ and $\hat{\tau}_1 - 2\hat{\tau}_2 + \hat{\tau}_3$ that would be used in the one-way analysis of variance model. Consequently, the sums of squares for testing no quadratic trend and no linear trend are the same, although again, the error mean square will differ.

8.7.3 Comments

We have illustrated via two examples the equivalence between the orthogonal trend contrasts in analysis of variance and orthogonal polynomials in regression analysis for the case of equispaced, equireplicated treatment levels. While both are convenient tools for data analysis, identification of orthogonal trend contrasts and orthogonal polynomials can be rather complicated for higher-order trends, unequally spaced levels, or unequal numbers of observations per level. Fortunately, analogous testing information can also be generated by fitting appropriate full and reduced models, as was discussed in Sect. 8.6.1. This is easily accomplished using computer regression software. Use of SAS and R software for such tests will be illustrated in Sects. 8.9 and 8.10.

8.8 A Real Experiment—Bean-Soaking Experiment

The bean-soaking experiment was run by Gordon Keeler in 1984 to study how long mung bean seeds ought to be soaked prior to planting in order to promote early growth of the bean sprouts. The experiment was run using a completely randomized design, and the experimenter used a one-way

analysis of variance model and methods of multiple comparisons to analyze the data. In Sect. 8.8.2, we present the one-way analysis of variance, and then in Sect. 8.8.3, we reanalyze the data using polynomial regression methods.

8.8.1 Checklist

The following checklist has been drawn from the experimenter's report.

(a) **Define the objectives of the experiment.**

The objective of the experiment is to determine whether the length of the soaking period affects the rate of growth of mung bean seed sprouts. The directions for planting merely advise soaking overnight, and no further details are given.
As indicated in Fig. 8.6, I expect to see no sprouting whatsoever for short soaking times, as the water does not have sufficient time to penetrate the bean coat and initiate sprouting. Then, as the soaking time is increased, I would expect to see a transition period of sprouting with higher rates of growth as water begins to penetrate the bean coat. Eventually, the maximum growth rate would be reached due to complete saturation of the bean. A possible decrease in growth rates could ensue from even longer soaking times due to bacterial infection and "drowning" the bean.

(b) **Identify all sources of variation**.

(i) Treatment factors and their levels.
There is just one treatment factor in this experiment, namely soaking time. A pilot experiment was run to obtain an indication of suitable times to be examined in the main experiment. The pilot experiment examined soaking times from 0.5 to 16 h. Many beans that had been soaked for less than 6 h failed to germinate, and at 16 h the saturation point had not yet been reached. Consequently, the five equally spaced soaking times of 6, 12, 18, 24 and 30 h will be selected as treatment factor levels for the experiment.

(ii) Experimental units.
The experimental units are the mung bean seeds selected at random from a large sack of approximately 10,000 beans.

(iii) Blocking factors, noise factors, and covariates.
Sources of variation that could affect growth rates include: individual bean differences; protozoan, bacterial, fungal, and viral parasitism; light; temperature; humidity; water quality.
Differences between beans will hopefully balance out in the random assignment to soaking times. Light, temperature, humidity, and water quality will be kept constant for all beans in the experiment. Thus, no blocking factors or covariates will be needed in the model.
Bacterial infection could differ from one treatment factor level to another due to soaking the beans in different baths. However, if the beans assigned to different treatment factor levels are soaked in the same bath, this introduces the possibility of a chemical signal from beans ready to germinate to the still dormant beans that sprouting conditions are prime. Consequently, separate baths will be used.

(c) **Choose a rule by which to assign experimental units to treatments**.

A completely randomized design will be used with an equal number of beans assigned to each soaking time.

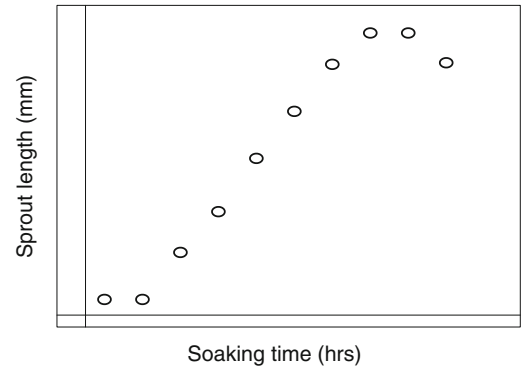


Fig. 8.6 Anticipated results from the bean-soaking experiment

(d) **Specify the measurements to be made, the experimental procedure, and the anticipated difficulties**.

The soaking periods will be started at 6-h intervals, so that the beans are removed from the water at the same time. They will then be allowed to grow in the same environmental conditions for 48 h, when the lengths of the bean sprouts will be measured (in millimeters).
The main difficulty in running the experiment is in controlling all the factors that affect growth. The beans themselves will be randomly selected and randomly assigned to soaking times. Different soaking dishes for the different soaking times will be filled at the same time from the same source. On removal from the soaking dishes, the beans will be put in a growth chamber with no light but high humidity. During the pilot experiment, the beans were rinsed after 24 h to keep them from dehydrating. However, the procedure cannot be well controlled from treatment to treatment, and will not be done in the main experiment.
A further difficulty is that of accurately measuring the shoot length.

(e) **Run a pilot experiment**.

A pilot study was run and the rest of the checklist was completed. As indicated in step (b), the results were used to determine the soaking times to be included in the experiment.

(f) **Specify the model**.

The one-way analysis of variance model (3.3.1) will be used, and the assumptions will be checked after the data are collected.

(g) **Outline the analysis**.

Confidence intervals for the pairwise differences in the effects of soaking time on the 48-h shoot lengths will be calculated. Also, in view of the expected results, linear, quadratic and cubic trends in the shoot length will be examined. Tukey's method will be used for the pairwise comparisons with $\alpha_1 = 0.01$, and Bonferroni's method will be used for the three trend contrasts with overall level $\alpha_2 \leq 0.01$. The experimentwise error rate will then be at most 0.02.

Table 8.6 Length of shoots of beans after 48 h for the bean-soaking experiment

Soaking time (h)	r	Length (mm)						Average length	Sample variance
12	17	5	11	8	11	4	4	5.9412	7.0588
		8	3	6	4	7	3		
		5	4	6	9	3			
18	17	11	16	18	24	18	18	18.4118	12.6324
		21	14	21	19	17	24		
		14	20	16	20	22			
24	17	17	16	26	18	14	24	19.5294	15.6397
		18	14	24	26	21	21		
		22	19	14	19	19			
30	17	20	18	22	20	21	17	21.2941	8.5956
		16	23	25	19	21	20		
		27	25	22	23	23			

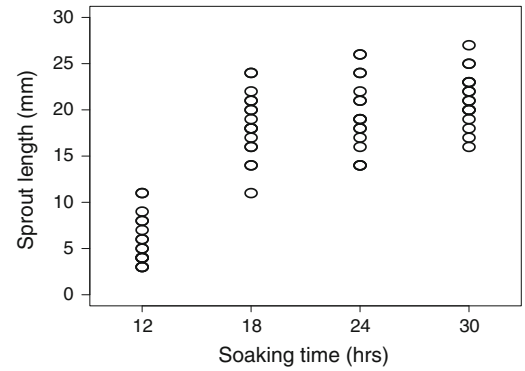


Fig. 8.7 Plot of sprout length y_{xt} against soaking time x for the bean-soaking experiment

(h) **Calculate the number of observations that need to be taken**.

Using the results of the pilot experiment, a calculation showed that 17 observations should be taken on each treatment (see Example 4.5.1, p. 93).

(i) **Review the above decisions. Revise, if necessary**.

Since 17 observations could easily be taken for the soaking time, there was no need to revise the previous steps of the checklist.

The experiment was run, and the resulting data are shown in Table 8.6. The data for soaking time 6 h have been omitted from the table, since none of these beans germinated.

The data are plotted in Fig. 8.7 and show that the trend expected by the experimenter is approximately correct. For the soaking times included in the study, sprout length appears to increase with soaking time, with soaking times of 18, 24, and 30 h yielding similar results, but a soaking of time of only 12 h yielding consistently shorter sprouts.

8.8.2 One-Way Analysis of Variance and Multiple Comparisons

The experimenter used Tukey's method with a 99% simultaneous confidence level to compare the effects of soaking the beans for 12, 18, 24, or 30 h. The formula for Tukey's method for the one-way analysis of variance model was given in (4.4.28) as

$$\tau_i - \tau_s \in \left(\overline{y}_{i.} - \overline{y}_{s.} \pm w_T \sqrt{\left(\frac{2}{r}\right) msE} \right),$$

where $w_T = q_{v,n-v,\alpha}/\sqrt{2}$.

The treatment sample means are shown in Table 8.6. There are $r = 17$ observations on each of the $v = 4$ levels of the treatment factor. The formula for the sum of squares for error in the one-way analysis of variance model was given in (3.4.5), p. 39. Using the data in Table 8.6 we have

$$msE = ssE/(n - v) = 10.9816\,.$$

From Table A.8, $q_{4,64,0.01} = 4.60$. Thus, in terms of the coded factor levels, the 99% simultaneous confidence intervals for pairwise comparisons are

$$\begin{aligned}
&\tau_4 - \tau_3 \in (-1.93,\ \ 5.46)\,, \quad \tau_3 - \tau_2 \in (-2.58,\ \ 4.81)\,,\\
&\tau_4 - \tau_2 \in (-0.81,\ \ 6.58)\,, \quad \tau_3 - \tau_1 \in \ \ (9.89, 17.29)\,,\\
&\tau_4 - \tau_1 \in \ (11.66, 19.05)\,, \quad \tau_2 - \tau_1 \in \ \ (8.77, 16.17)\,.
\end{aligned}$$

From these, we can deduce that soaking times of 18, 24, and 30 h yield significantly longer sprouts on average after 48 h than does a soaking time of only 12 h. The three highest soaking times are not significantly different in their effects on the sprout lengths, although the plot (Fig. 8.7) suggests that the optimum soaking time might approach or even exceed 30 h.

The one-way analysis of variance for the data is given in Table 8.7 and includes the information for testing for linear, quadratic, and cubic trends. The coefficients for the trend contrasts, when there are $v = 4$ equally spaced levels and equal sample sizes, are listed in Table A.2. The linear contrast is $[-3, -1,\ \ 1,\ \ 3]$, and the hypothesis of no linear trend is $H_0^L : \{-3\tau_1 - \tau_2 + \tau_3 + 3\tau_4 = 0\}$. Obtaining the treatment sample means from Table 8.6, the estimate of the linear trend is

$$\sum_i c_i \overline{y}_i = -3\overline{y}_{1.} - \overline{y}_{2.} + \overline{y}_{3.} + 3\overline{y}_{4.} = 47.1765\,,$$

with associated variance

$$\Sigma_i (c_i^2/r)\sigma^2 = (1/17)(9 + 1 + 1 + 9)\sigma^2 = (20/17)\sigma^2\,.$$

The sum of squares is calculated from (4.3.14), p. 77; that is,

$$ssc = \left(\sum_i c_i \overline{y}_{i.} \right)^2 \Big/ \left(\sum_i c_i^2/17 \right).$$

So, the sum of squares for the linear trend is

Table 8.7 One-way ANOVA for the bean-soaking experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
Soaking time	3	2501.29	833.76	75.92	0.0001
Linear trend	1	1891.78	1891.78	172.27	0.0001
Quadratic trend	1	487.12	487.12	44.36	0.0001
Cubic trend	1	122.40	122.40	11.15	0.0014
Error	64	702.82	10.98		
Total	67	3204.12			

$$ssc = (47.1765)^2/(20/17) = 1891.78.$$

The quadratic and cubic trends correspond to the contrasts $[\ 1, -1, -1,\ \ 1\ ]$ and $[-1,\ \ 3, -3,\ \ 1\ ]$, respectively, and their corresponding sums of squares are calculated in a similar way and are listed in Table 8.7. If we test the hypotheses that each of these three trends is zero with an overall significance level of $\alpha = 0.01$ using the Bonferroni method, then, using (4.4.24) on p. 84 for each trend, the null hypothesis that the trend is zero is rejected if $ssc/msE > F_{1,64,0.01/3}$. This critical value is not tabulated, but since $F_{1,64,0.0033} = t^2_{1,64,0.00166}$, it can be approximated using (4.4.22) as follows:

$$t_{1,64,0.00166} \approx 2.935 + (2.935^3 + 2.935)/(4 \times 64) = 3.0454,$$

so the critical value is $F_{1,64,0.0033} \approx 9.2747$. (Alternatively, the critical value could be obtained from a computer package using the "inverse cumulative distribution function" of the F-distribution.)

To test the null hypothesis H_0^L that the linear trend is zero against the alternative hypothesis H_A^L : $-3\tau_1 - \tau_2 + \tau_3 + 3\tau_4 \neq 0$ that the linear trend is nonzero, the decision rule is to

$$\text{reject } H_0 \text{ if } ssc/msE = 172.27 > F_{1,64,.0033} \approx 9.2747\,.$$

Thus, using a simultaneous significance level $\alpha = 0.01$ for the three trends, the linear trend is determined to be nonzero.

The corresponding test ratios for the quadratic and cubic trends are given in Table 8.7. There is sufficient evidence to conclude that the linear, quadratic, and cubic trends are all significantly different from zero. The probability that one or more of these hypotheses would be incorrectly rejected by this procedure is at most $\alpha = 0.01$.

8.8.3 Regression Analysis

In the previous subsection, the bean-soaking experiment was analyzed using the one-way analysis of variance and multiple comparison methods. In this subsection, we reanalyze the experiment using regression analysis. Since there are four levels of the treatment factor "soaking time," the highest-order polynomial regression model that can be (uniquely) fitted to the data is the cubic regression model, namely,

Table 8.8 Cubic regression ANOVA for the bean-soaking experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
β_3	1	122.40	122.40	11.15	0.0014
β_2, β_3	2	609.52	304.76	27.76	0.0001
Model	3	2501.29	833.76	75.92	0.0001
Error	64	702.82	10.98		
Total	67	3204.12			

$$
\begin{gathered}
Y_{xt} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon_{xt}\,, \\
\epsilon_{xt} \sim N(0, \sigma^2)\,, \\
\epsilon_{xt}\text{'s are mutually independent}\,, \\
x = 12, 18, 24, 30; \quad t = 1, \ldots, 17.
\end{gathered}
$$

Using the data given in Table 8.6, the fitted model can be obtained from a computer program (see Sects. 8.9 and 8.10) as

$$\hat{y}_x = -101.058824 + 15.475490x - 0.657680x^2 + 0.009259x^3\,.$$

Table 8.8 contains the analysis of variance for the bean experiment data based on the cubic regression model. The cubic model provides the same fit as does the one-way analysis of variance model, since $p+1 = v = 4$. Thus, $\hat{y}_x = \overline{y}_{x\cdot}$ for $x = 12, 18, 24, 30$, and the number of degrees of freedom, the sum of squares, and the mean square for the major sources of variation—the treatment factor ("Model"), error, and total—are the same in the regression analysis of variance as in the one-way analysis of variance. It is not possible to test for model lack of fit, since the postulated model is of order $p = 3 = v - 1$. We can, however, test to see whether a lower-order model would suffice.

We first test the null hypothesis $H_0^Q : \beta_3 = 0$, or equivalently, that the quadratic regression model $E[Y_{xt}] = \beta_0 + \beta_1 x + \beta_2 x^2$ would provide an adequate fit to the data. The result of the test is summarized in Table 8.8. The test ratio is 11.15 with a p-value of 0.0014. So, we reject H_0^Q and conclude that the cubic model is needed. Since the cubic regression model provides the same fit as the analysis of variance model, this test is identical to the test that the cubic trend contrast is zero in the one-way analysis of variance, shown in Table 8.7.

If $H_0^Q : \beta_3 = 0$ had not been rejected, then the next step would have been to have tested the null hypothesis $H_0^L : \beta_2 = \beta_3 = 0$, or equivalently, that the simple linear regression model is adequate. If neither $H_0^Q : \beta_3 = 0$ nor $H_0^L : \beta_2 = \beta_3 = 0$ had been rejected, the next step would have been to have tested $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$.

Based on the previous analysis, the cubic model is needed to provide an adequate fit to the data. Figure 8.8 illustrates the cubic model fitted to the data. We may now see the dangers of using a model to predict the value of the response beyond the range of observed x values. The cubic model predicts that mean sprout length will increase rapidly as soaking time is increased beyond 30 h! Clearly, this model is extremely unlikely to be reliable for extrapolation beyond 30 h.

Recall that Tukey's method of multiple comparisons did not yield any significant differences in mean response between the soaking times of 18, 24, and 30 h. Yet the plot of the data in Fig. 8.8 suggests that a trend over these levels might well exist. There is a lot of variability inherent in the data that prevents significant differences between the soaking times from being detected. Nevertheless, a

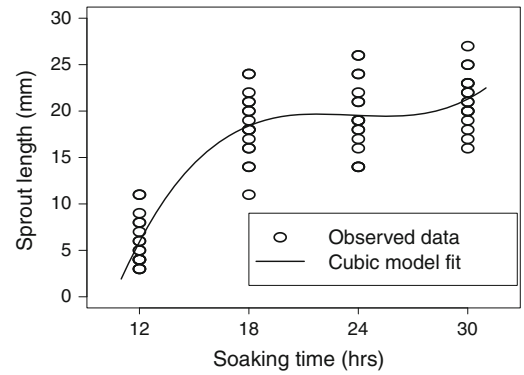


Fig. 8.8 Plot of data and fitted cubic polynomial regression model for the bean-soaking experiment

followup experiment examining soaking times from 18 to, say, 48 h might provide the information needed to determine the best range of soaking times.

8.9 Using SAS Software

Polynomial regression models can be fitted using the SAS regression procedure `PROC REG`. The procedure provides least squares estimates of the regression parameters. Predicted (fitted) values and residuals can be saved to an output data set, as can 95% confidence limits for mean response, 95% prediction limits for new observations for given treatment levels x, and corresponding standard errors.

A sample SAS program to analyze the data from the bean-soaking experiment of Sect. 8.8 is shown in Table 8.9 . In the first `DATA` statement, the variables x^2 and x^3 are created for the cubic regression model. `PROC REG` is used to fit the cubic regression model, and the output is shown in Fig. 8.9.

An analysis of variance table is automatically generated and includes information needed for testing the hypothesis that the treatment factor "soaking time" has no predictive value for mean growth length, namely, $H_0 : \{\beta_1 = \beta_2 = \beta_3 = 0\}$. The information for this test is listed with source of variation "Model". We see that the p-value is less than 0.0001, so H_0 would be rejected.

Below the analysis of variance table, parameter estimates for the fitted model are given. Using these, we have the fitted cubic regression model

$$\hat{y}_x = -101.05882 + 15.47549x - 0.65768x^2 + 0.00926x^3 .$$

The standard error of each estimate is also provided, together with the information for conducting a t-test of each individual hypothesis $H_0 : \{\beta_i = 0\}$, $i = 1, 2, 3$.

Inclusion of the option `SS1` in the `MODEL` statement of `PROC REG` causes printing of the Type I (sequential) sums of squares in the output. Each Type I sum of squares is the variation explained by entering the corresponding variable into the model, given that the previously listed variables are already in the model. For example, the Type I sum of squares for `X` is $ssE_0 - ssE_1$, where ssE_0 is the error sum of squares for the model with $E[Y_{xt}] = \beta_0$, and ssE_1 is the error sum of squares for the simple linear regression model $\mathrm{E}[Y_{xt}] = \beta_0 + \beta_1 x$; that is,

$$ss(\beta_1|\beta_0) = ssE_0 - ssE_1 = 1891.77647 .$$

Likewise, the Type I sum of squares for `X2` is the difference in error sums of squares for the linear and quadratic regression models; that is,

Fig. 8.9 Output generated by PROC REG

Results Viewer - SAS Output

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: LENGTH

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2501.29412	833.76471	75.92	<.0001
Error	64	702.82353	10.98162		
Corrected Total	67	3204.11765			

Root MSE	3.31385	R-Square	0.7806
Dependent Mean	16.29412	Adj R-Sq	0.7704
Coeff Var	20.33772		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	-101.05882	21.87851	-4.62	<.0001	18054
X	1	15.47549	3.49667	4.43	<.0001	1891.77647
X2	1	-0.65768	0.17508	-3.76	0.0004	487.11765
X3	1	0.00926	0.00277	3.34	0.0014	122.40000

Test QUAD Results for Dependent Variable LENGTH

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	122.40000	11.15	0.0014
Denominator	64	10.98162		

Test LINEAR Results for Dependent Variable LENGTH

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	304.75882	27.75	<.0001
Denominator	64	10.98162		

$$ss(\beta_2|\beta_0, \beta_1) = ssE_1 - ssE_2 = 487.11765\,,$$

and for X3, the Type I sum of squares is the difference in error sums of squares for the quadratic and cubic regression models; that is,

$$ss(\beta_3|\beta_0, \beta_1, \beta_2) = ssE_2 - ssE = 122.40000\,,$$

where we have written ssE for the error sum of squares for the full cubic model (rather than ssE_3). Thus, the ratio used to test the null hypothesis $H_0^Q : \{\beta_3 = 0\}$ versus $H_A^Q : \{\beta_3 \neq 0\}$ is

$$ss(\beta_3)/msE = ss(\beta_3|\beta_0, \beta_1, \beta_2)/msE = 122.4/10.98162 = 11.1459\,.$$

Table 8.9 SAS program for analysis of the bean-soaking experiment

```
DATA BEAN;
  INPUT X LENGTH;
  X2 = X**2; X3 = X**3;
  LINES;
  12   5
  12  11
  12   8
   :   :
  30  23
  30  23
;
* create extra x-values for plotting the fitted curve;
DATA TOPLOT;
  DO X = 8 TO 34;  X2 = X**2; X3 = X**3;
    LENGTH = .; * "." denotes a missing value;
    OUTPUT;
  END; * X loop;
* concatenate data sets BEAN and TOPLOT;
DATA; SET BEAN TOPLOT;
* do the analysis;
PROC REG;  MODEL LENGTH = X X2 X3 / SS1;
  QUAD:   TEST X3 = 0; * test adequacy of quadratic model;
  LINEAR: TEST X2 = 0, X3 = 0; * test adequacy of linear model;
  OUTPUT PREDICTED = LHAT RESIDUAL = E
     L95M = L95M U95M = U95M STDP = STDM
     L95 = L95I U95 = U95I STDI = STDI;
* plot the data and fitted model, overlayed on one plot;
PROC SGPLOT;
  SCATTER Y = LENGTH X = X / LEGENDLABEL = 'Observed data'
   MARKEREATTRS = (SIZE = 0.25cm COLOR = BLACK);
  SCATTER Y = LHAT X = X / LEGENDLABEL = 'Cubic model fit'
   MARKEREATTRS = (SYMBOL = SQUARE SIZE = 0.25cm COLOR = BLACK);
  YAXIS LABEL = "Sprout length (mm)" VALUES = (-20 TO 30 by 5);
  XAXIS LABEL = "Soaking time (hrs)" VALUES = (8 TO 36 by 4);
* 95% confidence intervals and standard errors for mean response;
PROC PRINT; VAR X L95M LHAT U95M STDM;
* 95% prediction intervals and standard errors for new observations;
PROC PRINT; VAR X L95I LHAT U95I STDI;
* generate residual plots;
PROC RANK NORMAL = BLOM; VAR E; RANKS NSCORE;
PROC SGPLOT; SCATTER Y = E X = X;
PROC SGPLOT; SCATTER Y = E X = LHAT;
PROC SGPLOT; SCATTER Y = E X = NSCORE;
```

The output of the `TEST` statement labeled `QUAD` provides the same information, as well as the p-value 0.0014. The null hypothesis H_0^Q is thus rejected, so the quadratic model is not adequate—the cubic model is needed. Hence, there is no reason to test further reduced models, but the information for such tests will be discussed for illustrative purposes.

To test $H_0^L : \beta_2 = \beta_3 = 0$, the full model is the cubic model and the reduced model is the linear model, so the numerator sum of squares of the test statistic is

$$ss(\beta_2, \beta_3) = ssE_1 - ssE = ss(\beta_2|\beta_0, \beta_1) + ss(\beta_3|\beta_0, \beta_1, \beta_2)$$
$$= 487.117647 + 122.400000 = 609.517647\,,$$

and the decision rule for testing H_0^L against the alternative hypothesis H_A^L that the cubic model is needed is

$$\text{reject } H_0^L \text{ if } ms(\beta_2, \beta_3)/msE > F_{2,64,\alpha}\,,$$

where

$$ms(\beta_2, \beta_3) = ss(\beta_2, \beta_3)/2\,.$$

The information for this test of adequacy of the linear model is also generated by the `TEST` statement labeled `LINEAR`.

The `OUTPUT` statement in `PROC REG` saves into an output data set the upper and lower 95% confidence limits for mean response and the corresponding standard error under the variable names `L95M`, `U95M` and `STDM`. This is done for each x-value in the input data set for which all regressors are available. Similarly, the upper and lower 95% prediction limits for a new individual observation and the corresponding standard error are saved under the variable names `L95I`, `U95I` and `STDI`. These could be printed or plotted, though we do not do so here.

The plot produced by `PROC SGPLOT` is not shown but is similar to the plot in Fig. 8.8. Overlaid on the same axes are plots of the raw data and the fitted cubic polynomial regression curve. A trick was used to generate data to plot the fitted curve. Actual x values range from 12 to 30. In the `DATA TOPLOT` step in Table 8.9, additional observations were created in the data set corresponding to the integer x values ranging from 8 to 34 but with missing values for the dependent variable length. While observations with missing length values cannot be used to fit the model, the regression procedure does compute the corresponding predicted values `LHAT`. The `OUTPUT` statement includes these fitted values in the newly created output data set, so they can be plotted to show the fitted model.

In this example, it is not possible to test for lack of fit of the cubic model, since data were collected at only four x-levels. If we had been fitting a quadratic model, then a lack-of-fit test would have been possible. An easy way to generate the relevant output using the SAS software is as follows. In line 4 of the program, add a classification variable `A`, using the statement "`A = X;`". Then insert a `PROC GLM` procedure before `PROC REG` as follows.

```
PROC GLM;
  CLASS A;
  MODEL LENGTH = X X2 A;
```

Then the Type I sum of squares for `A` is the appropriate numerator *ssLOF* for the test ratio.

Statements for generation of residual plots for checking the error assumptions are included in the sample SAS program in Table 8.9, but the output is not shown here.

8.10 Using R Software

Polynomial regression models can be fitted using the R function `lm` that fits linear models. The function provides least squares estimates of the regression parameters. Predicted (fitted) values and residuals are available, as are 95% confidence limits for mean response, 95% prediction limits for new observations for given treatment levels x, and corresponding standard errors.

A sample R program to analyze the data from the bean-soaking experiment of Sect. 8.8 is shown in Table 8.10. In the first block of code, the data are read from file into the data set `bean.data`.

Subsequently, though the results are not shown here, the `head` command would display the first six rows of data, showing for example that the data set contains the two variables `x` and `Length`, then the dimension command would reveal that the data set contains 68 observations, and a scatterplot of the data would be generated.

In the second block of code, the linear model function `lm` is used to fit the cubic, quadratic, and linear regression models, saving the respective results as `model3`, `model2` and `model1`, and related commands are used to generate the output shown in Tables 8.11 and 8.12. Since the data set contains the soaking time, x, the syntax `I(x^2)` allows inclusion of the quadratic term x^2 as a predictor variable in a model without creating a corresponding variable in the data set, and likewise `I(x^3)` for the cubic term x^3. The command `summary(model3)` displays the parameter least squares estimates shown in the middle of Table 8.11. From these, we have the fitted cubic regression model

$$\hat{y}_x = -101.05882 + 15.47549x - 0.65768x^2 + 0.00926x^3 \,.$$

The standard error of each estimate is also provided, together with the information for conducting a t-test of each individual hypothesis $H_0 : \{\beta_i = 0\}$, $i = 1, 2, 3$.

The `summary` command also generates the analysis of variance F-test of the hypothesis that the treatment factor "soaking time" has no predictive value for mean growth length, namely, $H_0 : \{\beta_1 = \beta_2 = \beta_3 = 0\}$. The information for this test is listed after "`F-statistic`". We see that the p-value is very small, only 2×10^{-16}, so H_0 would be rejected.

Having saved the results of the cubic fit as `model3`, the statement `anova(model3)` causes display of the Type I (sequential) sums of squares, provided in an analysis of variance table in the bottom of Table 8.11. Each Type I sum of squares is the variation explained by entering the corresponding variable into the model, given that the previously listed variables are already in the model. For example, the Type I sum of squares for `x` is $ssE_0 - ssE_1$, where ssE_0 is the error sum of squares for the model with $E[Y_{xt}] = \beta_0$, and ssE_1 is the error sum of squares for the simple linear regression model $\mathrm{E}[Y_{xt}] = \beta_0 + \beta_1 x$; that is,

$$ss(\beta_1|\beta_0) = ssE_0 - ssE_1 \approx 1892 \,.$$

Likewise, the Type I sum of squares for `x^2` is the difference in error sums of squares for the linear and quadratic regression models; that is,

$$ss(\beta_2|\beta_0, \beta_1) = ssE_1 - ssE_2 \approx 487 \,,$$

and for `x^3`, the Type I sum of squares is the difference in error sums of squares for the quadratic and cubic regression models; that is,

$$ss(\beta_3|\beta_0, \beta_1, \beta_2) = ssE_2 - ssE \approx 122 \,,$$

where we have written ssE for the error sum of squares for the full cubic model (rather than ssE_3). Thus, the ratio used to test the null hypothesis $H_0^Q : \{\beta_3 = 0\}$ versus $H_A^Q : \{\beta_3 \neq 0\}$ is

$$ss(\beta_3)/msE = ss(\beta_3|\beta_0, \beta_1, \beta_2)/msE \approx 122/11 \approx 11.2 \,,$$

with corresponding p-value 0.0014. The null hypothesis H_0^Q is thus rejected, so the quadratic model is not adequate—the cubic model is needed. The same information is generated by the statement `anova(model2, model3)`, which compares the reduced quadratic and full cubic models, with

Table 8.10 R program for analysis of the bean-soaking experiment

```
bean.data = read.table("data/bean.txt", header=T)
head(bean.data); dim(bean.data); plot(Length ~ x, data=bean.data)

# Fit regression models and generate ANOVA info
model3 = lm(Length ~ x + I(x^2) + I(x^3), data=bean.data) # Fit cubic model
summary(model3) # Display least squares estimates, overall F test
anova(model3) # Display type 1 SS
# Would a lower order model suffice?
model2 = lm(Length ~ x + I(x^2), data=bean.data) # Fit quadratic model
model1 = lm(Length ~ x, data=bean.data) # Fit simple linear reg model
anova(model2, model3) # Can cubic term be dropped?
anova(model1, model3) # Can both cubic and quadratic terms be dropped?

# Compute predicted values, CIs, PIs, and std errors for x=8, 8.01, ..., 34
# Set up a grid of x's for prediction: x=8, 8.01, 8.02, ..., 34
xPred = data.frame(x=seq(8, 34, 0.01))
# Calculate fitted values, 95% CIs for mean response, se.fit
preds = predict(model3, xPred, se.fit=T, interval=c("confidence"))
# Calculate 95% PIs for new observations
preds2 = predict(model3, xPred, interval = c("prediction"))
# preds; preds2 # (Reader: display preds and preds2 to see contents)
se.fit = preds$se.fit # to remove "preds$" from column header name
# Compute standard error for prediction
rmse = preds$residual.scale # used to compute se.pred
se.pred = sqrt(preds$se.fit^2 + rmse^2)
# Consolidate results for display
stats = cbind(xPred, preds$fit, se.fit, preds2[,2:3], se.pred)
head(stats) # display first six rows of results

# Plot data (length vs x), plus fitted model for x=8:34
plot(Length ~ x, xlim = c(8, 34), ylim = c(-10, 30), data=bean.data)
lines(xPred$x, preds$fit[, 1])

# Some plots to check model assumptions
bean.data$e = residuals(model3); # Obtain residuals
bean.data$pred = fitted(model3) # Obtain predicted values
# Plot residuals vs x
plot(e ~ x, ylab = "Residuals", las=1, xaxt="n", data=bean.data)
axis(1, at = c(12,18,24,30)); abline(h=0)
# Plot residuals vs predicted values
plot(e ~ pred, xlim=c(5,25), las=1, xaxt="n", data=bean.data,
     xlab="Predicted Values", ylab = "Residuals")
axis(1, at=seq(5,25,5)); abline(h=0)
# Normal probability plot of residuals
qqnorm(model3$res, ylim=c(-10,10), xlim=c(-4,4)); abline(h=0, v=0)
```

output shown in the top of Table 8.12. Since the cubic model is needed, there is no reason to test further reduced models, but the information for such tests will be discussed for illustrative purposes.

To test $H_0^L : \beta_2 = \beta_3 = 0$, the full model is the cubic model and the reduced model is the simple linear regression model, so the numerator sum of squares of the test statistic computed from the Type I sums of squares is

Table 8.11 Output generated for the cubic model

```
> # Fit regression models and generate ANOVA info
> model3 = lm(Length ~ x + I(x^2) + I(x^3), data=bean.data) # Fit cubic model
> summary(model3) # Display least squares estimates, overall F test

Call:
lm(formula = Length ~ x + I(x^2) + I(x^3), data = bean.data)

Residuals:
   Min     1Q Median     3Q    Max
-7.412 -2.029 -0.412  2.059  6.471

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -101.05882   21.87851   -4.62 0.000019
x             15.47549    3.49667    4.43 0.000038
I(x^2)        -0.65768    0.17508   -3.76  0.00037
I(x^3)         0.00926    0.00277    3.34  0.00141

Residual standard error: 3.31 on 64 degrees of freedom
Multiple R-squared:  0.781, Adjusted R-squared:  0.77
F-statistic: 75.9 on 3 and 64 DF,  p-value: <2e-16

> anova(model3) # Display type 1 SS
Analysis of Variance Table

Response: Length
          Df Sum Sq Mean Sq F value  Pr(>F)
x          1   1892    1892   172.3 < 2e-16
I(x^2)     1    487     487    44.4 7.3e-09
I(x^3)     1    122     122    11.2  0.0014
Residuals 64    703      11
```

$$ss(\beta_2, \beta_3) = ssE_1 - ssE = ss(\beta_2|\beta_0, \beta_1) + ss(\beta_3|\beta_0, \beta_1, \beta_2)$$
$$\approx 487 + 122 = 609\,,$$

and the decision rule for testing H_0^L against the alternative hypothesis H_A^L that the cubic model is needed is

$$\text{reject } H_0^L \text{ if } ms(\beta_2, \beta_3)/msE > F_{2,64,\alpha}\,,$$

where

$$ms(\beta_2, \beta_3) = ss(\beta_2, \beta_3)/2\,.$$

The information for this test of adequacy of the simple linear regression model is also generated by the statement `anova(model1, model3)`, with results shown in the bottom of Table 8.12. Here, the numerator sum of squares is rounded to 610, yielding $F = 27.8$ and $p = 2.1 \times 10^{-09}$.

The third block of code in Table 8.10 saves a grid of x values from 8 to 34 in step of 0.01 in a data set `xPred`, in order to compute confidence and prediction intervals at these x values. The `predict` function is called twice, each time using the results of the cubic fit saved as `model3`. For each x value in the grid, the first call of `predict` computes the predicted length, the standard error for predicting mean length, and the 95% confidence interval for mean response, saving these results asvadjust `preds`. By

Table 8.12 Output generated for the cubic model (continued)

```
> # Would a lower order model suffice?
> model2 = lm(Length ~ x + I(x^2), data=bean.data) # Fit quadratic model
> model1 = lm(Length ~ x, data=bean.data) # Fit simple linear reg model
> anova(model2, model3) # Can cubic term be dropped?
Analysis of Variance Table

Model 1: Length ~ x + I(x^2)
Model 2: Length ~ x + I(x^2) + I(x^3)
  Res.Df RSS Df Sum of Sq    F Pr(>F)
1     65 825
2     64 703  1       122 11.2 0.0014
> anova(model1, model3) # Can both cubic and quadratic terms be dropped?
Analysis of Variance Table

Model 1: Length ~ x
Model 2: Length ~ x + I(x^2) + I(x^3)
  Res.Df  RSS Df Sum of Sq    F  Pr(>F)
1     66 1312
2     64  703  2       610 27.8 2.1e-09
```

displaying `preds`, one would see that the predicted values and confidence limits are saved as the three columns of the object `preds$fit`, the standard error is saved as the lone column of `preds$se.fit`, and the root mean squared error is saved as a scalar as `preds$residual.scale`. Due to the `"prediction"` option, the second call of `predict` computes the predicted length and the 95% prediction interval for a new observation for each $x = 8, \ldots, 34$, saving these results as the three columns of `preds2`. The standard error for prediction is not provided directly, but is subsequently computed for each x from the standard error for estimation of mean length for the given x value and from the common root mean squared error value, both available from `preds`. The column bind command `cbind` is used to combine the desired information into the columns of the object `stats`. These could be printed or plotted, though we do not do so here.

The output of the `plot` function in the fourth block of code is not shown here, but is similar to the plot show in Fig. 8.8. The `plot` command causes the raw data to be plotted. Then the `lines` subcommand augments the plot with the line corresponding to the predicted values at the grid points $x = 8, 8.01, 8.02, \ldots, 34$, giving a sense of the fitted cubic polynomial regression curve.

In this example, it is not possible to test for lack of fit of the cubic model, since data were collected at only four x-levels. If we had been fitting a quadratic model, then a lack-of-fit test would have been possible. An easy way to generate the relevant output using the R software is as follows. Anyplace after saving the fitted quadratic model as `model2` in the second block of code, add the following code.

```
bean.data$fA = factor(bean.data$x)
modelA = lm(Length ~ fA, data=bean.data)
anova(model2, modelA))
```

The first command adds a factor variable `fA` to the data set, the second fits the one-way model with a different mean for each x value (i.e. for each level of `fA`), and the third generates the F-test for lack of fit by comparing the reduced quadratic model to the full one-way model, which is the fullest model one can fit here.

Statements for generation of residual plots for checking the error assumptions are included in the sample R program in Table 8.10, but the output is not shown here.

Table 8.13 Data for the bicycle experiment

Treatment (mph)	Crank rates y_{xt}		
x	y_{xt}		
5	15	19	22
10	32	34	27
15	44	47	44
20	59	61	61
25	75	73	75

Exercises

1. For the simple linear regression model

$$\mathrm{E}[Y_{xt}] = \beta_0 + \beta_1 x\,,$$

the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ for the parameters β_0 and β_1 are given in (8.5.6), p. 257. Show that their variances are

$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}_{\cdot\cdot}^2}{ss_{xx}}\right) \quad \text{and} \quad \mathrm{Var}(\hat{\beta}_1) = \sigma^2\left(\frac{1}{ss_{xx}}\right),$$

where $ss_{xx} = \sum_x r_x(x - \overline{x}_{\cdot\cdot})^2$, as given in (8.5.7).

2. **Bicycle experiment, continued**

The bicycle experiment was run to compare the crank rates required to keep a bicycle at certain speeds, when the bicycle (a Cannondale SR400) was in twelfth gear on flat ground. The speeds chosen were $x = 5$, 10, 15, 20, and 25 mph. The data are given in Table 8.13. (See also Exercise 6 of Chap. 5.)

(a) Fit the simple linear regression model to the data, and use residual plots to check the assumptions of the simple linear regression model.
(b) If a transformation of the data is needed, choose a transformation, refit the simple linear regression model, and check for lack of fit.
(c) Using your results from parts (a) and (b), select a model for the data. Use this model to obtain an estimate for the mean crank rate needed to maintain a speed of 18 mph in twelfth gear on level ground.
(d) Calculate a 95% confidence interval for the mean crank rate needed to maintain a speed of 18 mph in twelfth gear on level ground.
(e) Find the 95% confidence band for the regression line. Draw a scatter plot of the data and superimpose the regression line and the confidence band on the plot.
(f) Would you be happy to use your model to estimate the mean crank rate needed to maintain a speed of 35 mph in twelfth gear on level ground. Why or why not?

Table 8.14 Systolic blood pressure measurements—(order of collection in parentheses)

Jogging time in seconds (order of collection)					
10	20	25	30	40	50
120 (1)	125 (2)	127 (10)	128 (3)	137 (5)	143 (6)
118 (9)	126 (4)		131 (7)		
	123 (8)				

Table 8.15 Data for the trout experiment

x	Hemoglobin (grams per 100 ml)									
00	6.7	7.8	5.5	8.4	7.0	7.8	8.6	7.4	5.8	7.0
05	9.9	8.4	10.4	9.3	10.7	11.9	7.1	6.4	8.6	10.6
10	10.4	8.1	10.6	8.7	10.7	9.1	8.8	8.1	7.8	8.0
15	9.3	9.3	7.2	7.8	9.3	10.2	8.7	8.6	9.3	7.2

Source Gutsell, J.S. (1951). Copyright© 1951 International Biometric Society. Reprinted with permission

3. **Systolic blood pressure experiment**

A pilot experiment was run by John Spitak in 1987 to investigate the effect of jogging on systolic blood pressure. Only one subject was used in the pilot experiment, and a main experiment involving a random sample of subjects from a population of interest would need to be run in order to draw more general conclusions. The subject jogged in place for a specified number of seconds and then his systolic blood pressure was measured. The subject rested for at least 5 min, and then the next observation was taken.

The data and their order of observation are given in Table 8.14.

(a) Fit a simple linear regression model to the data and test for model lack of fit.
(b) Use residual plots to check the assumptions of the simple linear regression model.
(c) Give a 95% confidence interval for the slope of the regression line.
(d) Using the confidence interval in part (c), test at significance level $\alpha = 0.05$ whether the linear term is needed in the model.
(e) Repeat the test in part (d) but using the formula for the orthogonal polynomial linear trend coefficients for unequally spaced levels and unequal sample sizes given in Sect. 4.2.4. Do these two tests give the same information?
(f) Estimate the mean systolic blood pressure of the subject after jogging in place for 35 sec and calculate a 99% confidence interval.
(g) The current experiment was only a pilot experiment. Write out a checklist for the main experiment.

4. **Trout experiment, continued**

The data in Table 8.15 show the measurements of hemoglobin (grams per 100 ml) in the blood of brown trout. (The same data were used in Exercise 15 of Chap. 3.) The trout were placed at random in four different troughs. The fish food added to the troughs contained, respectively, $x = 0$, 5, 10, and 15 grams of sulfamerazine per 100 pounds of fish. The measurements were made on ten randomly selected fish from each trough after 35 days.

(a) Fit a quadratic regression model to the data.
(b) Test the quadratic model for lack of fit.

(c) Use residual plots to check the assumptions of the quadratic model.
(d) Test whether the quadratic term is needed in the model.
(e) Use the fitted quadratic model to estimate the number of grams of sulfamerazine per 100 pounds of fish to maximize the mean amount of hemoglobin in the blood of the brown trout.

5. **Bean-soaking experiment, continued**

Use residual plots to check the assumptions of the cubic regression model for the data of the bean-soaking experiment. (The data are in Table 8.6, p. 269).

6. **Bean-soaking experiment, continued**

Suppose the experimenter in the bean-soaking experiment of Sect. 8.8 had presumed that the quadratic regression model would be adequate for soaking times ranging from 12 to 30 h.

(a) Figure 8.8, p. 273, shows the fitted response curve and the standardized residuals each plotted against soaking time. Based on these plots, discuss model adequacy.
(b) Test the quadratic model for lack of fit.

7. **Orthogonal polynomials**

Consider an experiment in which an equal number of observations are collected for each of the treatment factor levels $x = 10, 20, 30, 40, 50$.

(a) Compute the corresponding values z_x for the linear orthogonal polynomial, and determine the rescaling factor by which the z_x differ from the coefficients of the linear trend contrast.
(b) Compute the values $z_x^{(2)}$ for the quadratic orthogonal polynomial, and determine the rescaling factor by which the $z_x^{(2)}$ differ from the coefficients of the quadratic trend contrast.
(c) Use the data of Table 8.1 and the orthogonal polynomial coefficients to test that the quadratic and linear trends are zero.
(d) Using the data of Table 8.1 and a statistical computing package, fit a quadratic model to the original values. Test the hypotheses

$$H_0^L : \{\beta_2 = 0\} \quad \text{and} \quad H_0 : \{\beta_1 = \beta_2 = 0\}$$

against their respective two-sided alternative hypotheses. Compare the results of these tests with those in (c).

8. **Orthogonal polynomials**

Consider use of the quadratic orthogonal polynomial regression model (8.7.20), p. 265, for the data at levels 18, 24, and 30 of the bean-soaking experiment—the data are in Table 8.6, p. 269.

(a) Compute the least squares estimates of the parameters.
(b) Why is it not possible to test for lack of fit of the quadratic model?
(c) Give an analysis of variance table and test the hypothesis that a linear model would provide an adequate representation of the data.

9. **Heart–lung pump experiment, continued**

In Example 8.5.1, p. 259, we fitted a linear regression model to the data of the heart–lung pump experiment. We rejected the null hypothesis that the slope of the line is zero.

(a) Show that the numerator sum of squares for testing $H_0 : \{\beta_1 = 0\}$ against the alternative hypothesis $H_A : \{\beta_1 \neq 0\}$ is the same as the sum of squares *ssc* that would be obtained for testing that the linear trend is zero in the analysis of variance model (the relevant calculations were done in Example 4.2.3, p. 73).
(b) Obtain a 95% confidence band for the regression line.
(c) Calculate a 99% prediction interval for the fluid flow rate at 100 revolutions per minute.
(d) Estimate the intercept β_0. This is not zero, which suggests that the fluid flow rate is not zero at 0 rpm. Since this should not be the case, explain what is happening.