Unbalanced Analysis of Variance, Design, and Regression
Ronald Christensen, fletcher@stat.unm.edu
Copyright 2013

Chapter 8

Testing Lack of Fit

Most often in analyzing data we start with a relatively complicated model and look for simpler versions that still fit the data adequately. Lack of fit involves an initial model that does not fit the data adequately. Most often, we start with a full model and look at reduced models. When dealing with lack of fit, our initial model is the reduced model, and we look for models that fit significantly better than the reduced model. In this chapter, we introduce methods for testing lack of fit for a simple linear regression model. As with the chapter on model checking, these ideas translate with (relatively) minor modifications to testing lack of fit for other models. The issue of testing lack of fit will arise again in later chapters.

The full models that we create in order to test lack of fit are universally models that involve fitting more than one predictor variable; these are multiple regression models. Multiple regression was introduced in Section 6.9 and special cases were applied in Section 7.3. This chapter makes extensive use of special cases of multiple regression. The general topic, however, is considered in Chapter 9.

We illustrate lack-of-fit testing methods by testing for lack-of-fit in the simple linear regression on the Hooker data of Table 7.1 and Example 7.2.2. Figure 8.1 displays the data with the fitted line and we again provide the ANOVA table for this (reduced) model.

Analysis of variance: Hooker data SLR

Source	*df*	*SS*	*MS*	*F*	*P*
Regression	1	444.17	444.17	3497.89	0.000
Error	29	3.68	0.13		
Total	30	447.85			

8.1 Polynomial regression

With Hooker's data, the simple linear regression of pressure on temperature shows a lack of fit. The residual plot in Figure 7.7 clearly shows nonrandom structure. In Section 7.3, we used a power transformation to eliminate the lack of fit. In this section we introduce an alternative method called *polynomial regression*. Polynomial regression is a special case of the multiple regression model that was introduced in Section 6.9 and is discussed more fully in Chapter 9.

With a single predictor variable x, we can try to eliminate lack of fit in the simple linear regression $y_i = \beta_0 + \beta_2 x_i + \varepsilon_i$ by fitting larger models. In particular, we can fit the *quadratic* (parabolic) model

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

We could also try a *cubic* model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i,$$

the *quartic* model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \varepsilon_i,$$

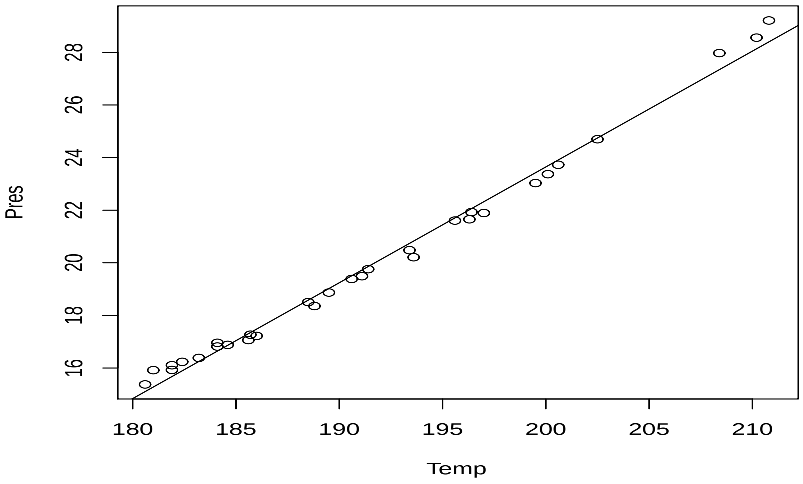

Figure 8.1: *Hooker data, linear fit.*

or higher degree polynomials. If we view our purpose as finding good, easily interpretable approximate models for the data, *high degree polynomials can behave poorly*. As we will see later, the process of fitting the observed data can cause high degree polynomials to give very erratic results in areas very near the observed data. A good approximate model should work well, not only at the observed data, but also near it. Thus, we focus on low degree polynomials. The problem of erratic fits is addressed in the next section. We now examine issues related to fitting polynomials.

EXAMPLE 8.1.1. Computer programs give output for polynomial regression that is very similar to that for simple linear regression. We fit a fifth degree (quintic) polynomial to Hooker's data,

$$y_i = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 + \gamma_3 x_i^3 + \gamma_4 x_i^4 + \gamma_5 x_i^5 + \varepsilon_i. \tag{8.1.1}$$

Actually, we tried fitting a cubic model to these data and encountered numerical instability. (Some computer programs object to fitting it.) This may be related to the fact that the R^2 is so high. To help with the numerical instability of the procedure, before computing the powers of the x variable we subtracted the mean $\bar{x}_\cdot = 191.787$. Thus, we actually fit,

$$y_i = \beta_0 + \beta_1 (x_i - \bar{x}_\cdot) + \beta_2 (x_i - \bar{x}_\cdot)^2 + \beta_3 (x_i - \bar{x}_\cdot)^3 + \beta_4 (x_i - \bar{x}_\cdot)^4 + \beta_5 (x_i - \bar{x}_\cdot)^5 + \varepsilon_i. \tag{8.1.2}$$

These two models are equivalent in that they always give the same fitted values, residuals, and degrees of freedom. Moreover, $\gamma_5 \equiv \beta_5$ but none of the other γ_js are equivalent to the corresponding β_js. (The equivalences are obtained by the rather ugly process of actually multiplying out the powers of $(x_i - \bar{x}_\cdot)$ in model (8.1.2) so that the model can be rewritten in the form of model (8.1.1).) The fitted model (8.1.2) is summarized by the table of coefficients and the ANOVA table.

Table of Coefficients: Model (8.1.2).

Predictor	$\hat{\beta}_k$	SE($\hat{\beta}_k$)	t	P
Constant	-59.911	2.337	-25.63	0.000
$(x-\bar{x}.)$	0.41540	0.01216	34.17	0.000
$(x-\bar{x}.)^2$	0.002179	0.002260	0.96	0.344
$(x-\bar{x}.)^3$	0.0000942	0.0001950	0.48	0.633
$(x-\bar{x}.)^4$	0.00001523	0.00001686	0.90	0.375
$(x-\bar{x}.)^5$	-0.00000080	0.00000095	-0.84	0.409

Analysis of variance: Model (8.1.2).

Source	df	SS	MS	F	P
Regression	5	447.175	89.435	3315.48	0.000
Error	25	0.674	0.027		
Total	30	447.850			

The most important things here are that we now know the SSE, dfE, and MSE from the fifth degree polynomial. The ANOVA table also provides an F test for comparing the fifth degree polynomial against the reduced model $y_i = \beta_0 + \varepsilon_i$, not a terribly interesting test.

Usually, *the only interesting t test for a regression coefficient in polynomial regression is the one for the highest term in the polynomial*. In this case the t statistic for the fifth degree term is -0.84 with a P value of 0.409, so there is little evidence that we need the fifth degree term in the polynomial. All the t statistics are computed as if the variable in question was the only variable being dropped from the fifth degree polynomial. For example, it usually makes little sense to have a quintic model that does not include a quadratic term, so there is little point in examining the t statistic for testing $\beta_2 = 0$. One reason for this is that simple linear transformations of the predictor variable change the roles of lower order terms. For example, something as simple as subtracting $\bar{x}.$ completely changes the meaning of γ_2 from model (8.1.1) to β_2 in model (8.1.2). Another way to think about this is that the Hooker data uses temperature measured in Fahrenheit as a predictor variable. The quintic model (8.1.2) for the Hooker data is consistent with $\beta_2 = 0$ with a P value of 0.344. If we changed to measuring temperature in Celsius, there is no reason to believe that the new quintic model would still be consistent with $\beta_2 = 0$. When there is a quintic term in the model, a quadratic term based on Fahrenheit measurements has a completely different meaning than a quadratic term based on Celsius measurements. The same is true for all the other terms except the highest order term, here the quintic term. On the other hand, the Fahrenheit and Celsius quintic models that include all lower order terms are equivalent, just as the simple linear regressions based on Fahrenheit and Celsius are equivalent. Of course these comments apply to all polynomial regressions. Exercise 8.6.7 explores the relationships among regression parameters for quadratic models that have and have not adjusted the predictor for its sample mean.

A lack of fit test is provided by testing the quintic model against the original simple linear regression model. The F statistic is

$$F = \frac{(3.68 - 0.674)/(29 - 25)}{0.027} = 27.83$$

which is much bigger than 1 and easily significant at the 0.01 level when compared to an $F(4,25)$ distribution. The test suggests lack of fit (or some other problem with the assumptions). □

8.1.1 Picking a polynomial

We now consider the problem of finding a small order polynomial that fits the data well.

The table of coefficients for the quintic polynomial on the Hooker data provides a t test for whether we can drop each variable out of the model, but for the most part these tests are uninteresting. The only t statistic that is of interest is that for x^5. It makes little sense, when dealing with a fifth degree polynomial to worry about whether you can drop out, say, the quadratic term. The only t statistic of interest is the one that tests whether you can drop x^5 so that you could get by with a quartic polynomial. If you are then satisfied with a quartic polynomial, it makes sense to test whether you can get by with a cubic. In other words, what we would really like to do is fit the sequence of models

$$y_i = \beta_0 + \varepsilon_i, \tag{8.1.3}$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{8.1.4}$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \tag{8.1.5}$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i, \tag{8.1.6}$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \varepsilon_i, \tag{8.1.7}$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \beta_5 x_i^5 + \varepsilon_i, \tag{8.1.8}$$

and find the smallest model that fits the data. It is equivalent to fit the sequence of polynomials with x adjusted for its mean, $\bar{x}_\cdot$. In subsequent discussion we refer to SSEs and other statistics for models (8.1.3) through (8.1.8) as $SSE(3)$ through $SSE(8)$ with other similar notations that are obvious. Recall that models (8.1.1), (8.1.2), and (8.1.8) are equivalent.

Many regression programs fit an overall model by fitting a sequence of models and provide key results from the sequence. Most often the results are the sequential sums of squares which are simply the difference in error sums of squares for consecutive models in the sequence. Note that you must specify the variables to the computer program in the order you want them fitted. For the Hooker data, sequential fitting of models (8.1.3) through (8.1.8) gives

	Model			
Source	Comparison	df	Seq SS	F
$(x-\bar{x}_\cdot)$	$SSE(3)-SSE(4)$	1	444.167	16450.7
$(x-\bar{x}_\cdot)^2$	$SSE(4)-SSE(5)$	1	2.986	110.6
$(x-\bar{x}_\cdot)^3$	$SSE(5)-SSE(6)$	1	0.000	0.0
$(x-\bar{x}_\cdot)^4$	$SSE(6)-SSE(7)$	1	0.003	0.1
$(x-\bar{x}_\cdot)^5$	$SSE(7)-SSE(8)$	1	0.019	0.7

Using these and statistics reported in Example 8.1.1, the F statistic for dropping the fifth degree term from the polynomial is

$$F = \frac{SSE(7)-SSE(8)}{MSE(8)} = \frac{0.019}{0.027} = 0.7 = (-0.84)^2.$$

The corresponding t statistic reported earlier for testing $H_0 : \beta_5 = 0$ in model (8.1.2) was -0.84. The data are consistent with a fourth degree polynomial.

The F test for dropping to a third degree polynomial from a fourth degree polynomial is

$$F = \frac{SSE(6)-SSE(7)}{MSE(8)} = \frac{0.003}{0.027} = 0.01.$$

In the denominator of the test we again use the MSE from the fifth degree polynomial. *When doing a series of tests on related models one generally uses the MSE from the largest model in the denominator of all tests*, cf. Subsection 3.1.1.

The t statistic corresponding to this F statistic is 0.11, not the value 0.90 reported for the fourth degree term in the table of coefficients for the fifth degree model (8.1.2). The t value of 0.11 is a statistic for testing $\beta_4 = 0$ in the fourth degree model. The value $t_{obs} = 0.11$ is not quite the t statistic you would get in the table of coefficients for fitting the fourth degree polynomial (8.1.7) because the table of coefficients would use the MSE from model (8.1.7) whereas this statistic is using the MSE from model (8.1.8). Nonetheless, t_{obs} provides a test for $\beta_4 = 0$ in a model that has already specified

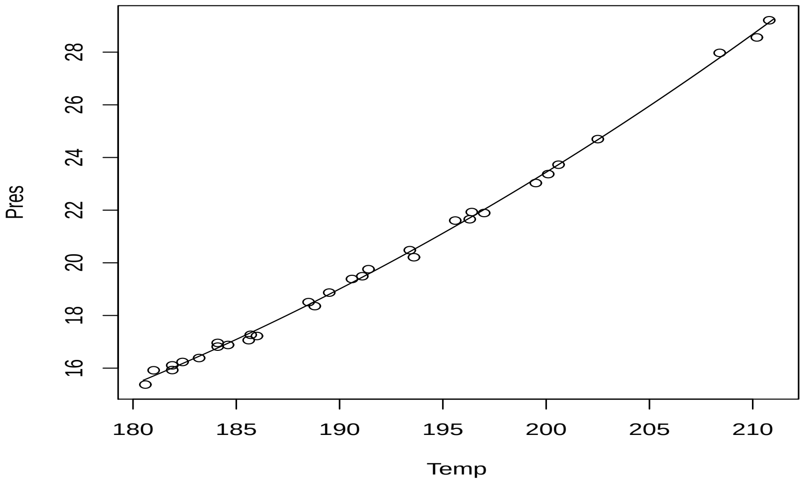

Figure 8.2: *Hooker data with quadratic fit.*

that $\beta_5 = 0$ whereas $t = 0.90$ from the table of coefficients for the fifth degree model (8.1.2) is testing $\beta_4 = 0$ without specifying that $\beta_5 = 0$.

The other F statistics listed are also computed as Seq $\text{SS}/MSE(8)$. From the list of F statistics, we can clearly drop any of the polynomial terms down to the quadratic term.

8.1.2 Exploring the chosen model

We now focus on the polynomial model that fits these data well: the quadratic model

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

We have switched to fitting the polynomial without correcting the predictor for its mean value. Summary tables for fitting the quadratic model are

Table of Coefficients: Hooker data, quadratic model.

Predictor	$\hat\beta_k$	SE($\hat\beta_k$)	t	P
Constant	88.02	13.93	6.32	0.000
x	-1.1295	0.1434	-7.88	0.000
x^2	0.0040330	0.0003682	10.95	0.000

Analysis of Variance: Hooker data, quadratic model.

Source	*df*	*SS*	*MS*	*F*	*P*
Regression	2	447.15	223.58	8984.23	0.000
Error	28	0.70	0.02		
Total	30	447.85			

The MSE, regression parameter estimates, and standard errors are used in the usual way. The t statistics and P values are for the tests of whether the corresponding β parameters are 0. The t statistics for β_0 and β_1 are of little interest. The t statistic for β_2 is 10.95, which is highly significant,

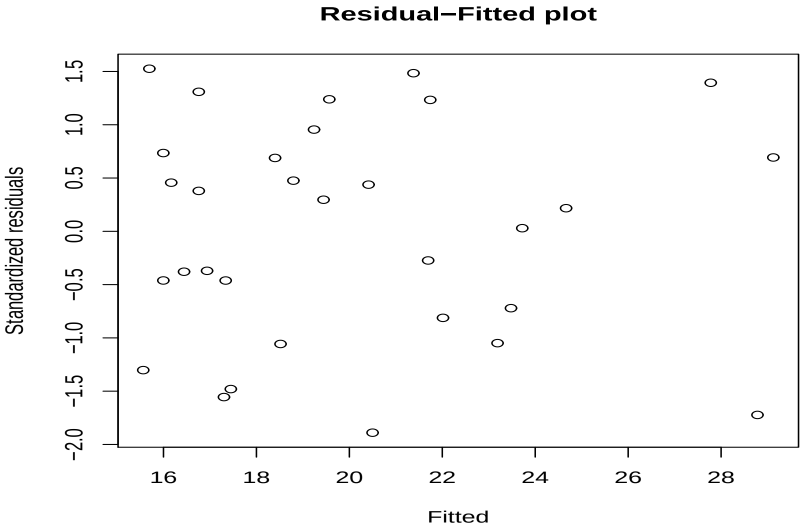

Figure 8.3: *Standardized residuals versus predicted values, quadratic model.*

so the quadratic model accounts for a significant amount of the lack of fit displayed by the simple linear regression model. Figure 8.2 gives the data with the fitted parabola.

We will not discuss the ANOVA table in detail, but note that with two predictors, x and x^2, there are 2 degrees of freedom for regression. In general, if we fit a polynomial of degree a, there will be a degrees of freedom for regression, one degree of freedom for every term other than the intercept. Correspondingly, when fitting a polynomial of degree a, there are $n-a-1$ degrees of freedom for error. *The ANOVA table F statistic provides a test of whether the polynomial (in this case quadratic) model explains the data better than the model with only an intercept.*

The fitted values are obtained by substituting the x_i values into

$$\hat{y} = 88.02 - 1.1295x + 0.004033x^2.$$

The residuals are $\hat{\varepsilon}_i = y_i - \hat{y}_i$.

The coefficient of determination is computed and interpreted as before. It is the squared correlation between the pairs $(\hat{y}_i, y_i)$ and also *SSReg* divided by the *SSTot*, so it measures the amount of the total variability that is explained by the predictor variables temperature and temperature squared. For these data, $R^2 = 99.8\%$, which is an increase from 99.2% for the simple linear regression model. It is not appropriate to compare the R^2 for this model to the R^2 from the log transformed model of the Section 7.4 because they are computed from data that use different scales. However, if we back transform the fitted log values to the original scale to give $\hat{y}_{i\ell}$ values and compute R_ℓ^2 as the squared correlation between the $(\hat{y}_{i\ell}, y_i)$ values, then R_ℓ^2 and R^2 are comparable.

The standardized residual plots are given in Figures 8.3 and 8.4. The plot against the predicted values looks good, just as it did for the transformed data examined in the Section 7.4. The normal plot for this model has a shoulder at the top but it looks much better than the normal plot for the simple linear regression on the log transformed data.

If we are interested in the mean value of pressure for a temperature of 205°F, the quadratic model estimate is (up to a little of round off error)

$$\hat{y} = 25.95 = 88.02 - 1.1295\,(205) + 0.004033\,(205)^2.$$

The standard error (as reported by the computer program) is 0.0528 and a 95% confidence interval is $(25.84, 26.06)$. This compares to a point estimate of 25.95 and a 95% confidence interval of

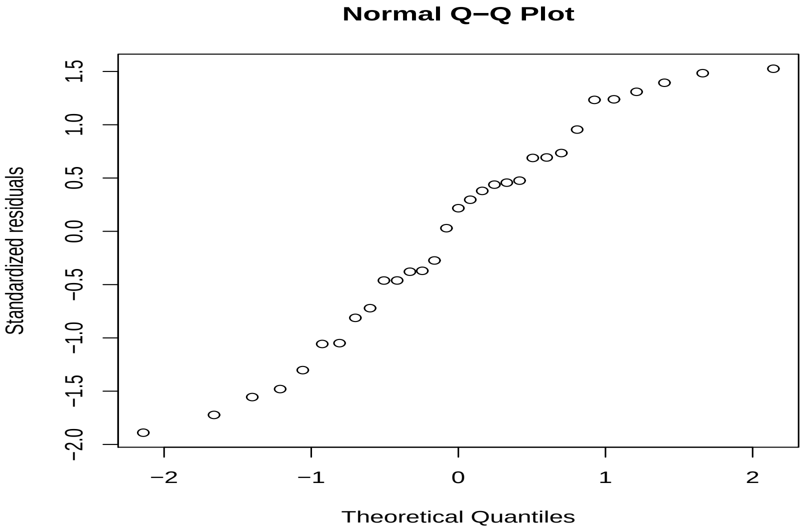

Figure 8.4: *Normal plot for quadratic model, $W' = 0.966$.*

$(25.80, 26.10)$ obtained in Section 7.3 from regressing the log of pressure on temperature and back transforming. The quadratic model *prediction* for a new observation at 205°F is again 25.95 with a 95% prediction interval of $(25.61, 26.29)$. The corresponding back transformed prediction interval from the log transformed data is $(25.49, 26.42)$. In this example, the results of the two methods for dealing with lack of fit are qualitatively very similar, at least at 205°F.

Finally, consider testing the quadratic model for lack of fit by comparing it to the quintic model (8.1.2). The F statistic is

$$F = \frac{(0.70 - 0.674)/(28 - 25)}{0.027} = 0.321$$

which is much smaller than 1 and makes no suggestion of lack of fit.

One thing we have not addressed is why we chose a fifth degree polynomial rather than a fourth degree or a sixth degree or a twelfth degree. The simplest answer is just to pick something that clearly turns out to be large enough to catch the important features of the data. If you start with too small a polynomial, go back and pick a bigger one. □

Computer commands

In what follows we illustrate Minitab commands for fitting quadratic, cubic, and quartic models. These include the prediction subcommand used with the quadratic model for $x = 205$. Note that the prediction subcommand requires us to enter both the value of x and the value of x^2 when using the quadratic model.

```
MTB > names c1 'y' c2 'x'
MTB > note     FIT QUADRATIC MODEL
MTB > let c22=c2**2
MTB > regress c2 2 c2 c22;
SUBC> pred 205 42025.
MTB > note     FIT CUBIC MODEL
MTB > let c23=c2**3
```

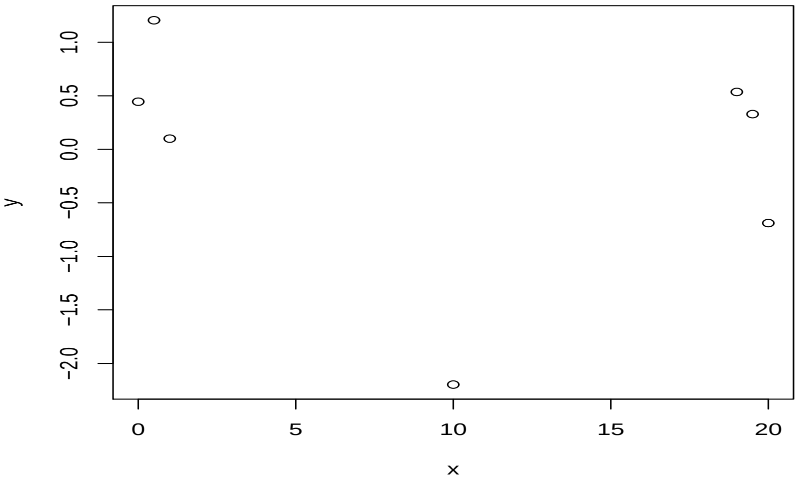

Figure 8.5: *Plot of y versus x..*

```
MTB > regress c1 3 c2 c22 c23
MTB > note     FIT QUARTIC MODEL
MTB > let c24=c2**4
MTB > regress c1 4 c2 c22-c24
```

8.2 Polynomial Regression and Leverages

We now present a simple example that illustrates two points: that leverages depend on the model and that high order polynomials can fit the data in very strange ways.

EXAMPLE 8.2.1. The data for the example follow. They were constructed to have most observations far from the middle.

Case	1	2	3	4	5	6	7
y	0.445	1.206	0.100	-2.198	0.536	0.329	-0.689
x	0.0	0.5	1.0	10.0	19.0	19.5	20.0

I selected the x values. The y values are a sample of size 7 from a $N(0,1)$ distribution. Note that with seven distinct x values, we can fit a polynomial of degree 6.

The data are plotted in Figure 8.5. Just by chance (honest folks), I observed a very small y value at $x = 10$, so the data appear to follow a parabola that opens up. The small y value at $x = 10$ totally dominates the impression given by Figure 8.5. If the y value at $x = 10$ had been near 3 rather than near -2, the data would appear to be a parabola that opens down. If the y value had been between 0 and 1, the data would appear to fit a line with a slightly negative slope. When thinking about fitting a parabola, the case with $x = 10$ is an extremely high leverage point.

Depending on the y value at $x = 10$, the data suggest a parabola opening up, a parabola opening down, or that we do not need a parabola to explain the data. Regardless of the y value observed at $x = 10$, the fitted parabola must go nearly through the point $(10, y)$. On the other hand, if we think only about fitting a line to these data, the small y value at $x = 10$ has much less effect. In fitting

Table 8.1: *Leverages.*

	Model					
x	Linear	Quadratic	Cubic	Quartic	Quintic	Hexic
0.0	0.33	0.40	0.64	0.87	0.94	1.00
0.5	0.31	0.33	0.33	0.34	0.67	1.00
1.0	0.29	0.29	0.55	0.80	0.89	1.00
10.0	0.14	0.96	0.96	1.00	1.00	1.00
19.0	0.29	0.29	0.55	0.80	0.89	1.00
19.5	0.31	0.33	0.33	0.34	0.67	1.00
20.0	0.33	0.40	0.64	0.87	0.94	1.00

a line, the value $y = -2.198$ will look unusually small (it will have a very noticeable standardized residual), but it will not force the fitted line to go nearly through the point $(10, -2.198)$.

Table 8.1 gives the leverages for all of the polynomial models that can be fitted to these data. Note that there are no large leverages for the simple linear regression model (the linear polynomial). For the quadratic (parabolic) model, all of the leverages are reasonably small except the leverage of 0.96 at $x = 10$ which very nearly equals 1. Thus, in the quadratic model, the value of y at $x = 10$ dominates the fitted polynomial. The cubic model has extremely high leverage at $x = 10$, but the leverages are also beginning to get large at $x = 0, 1, 19, 20$. For the quartic model, the leverage at $x = 10$ is 1 to two decimal places; the leverages for $x = 0, 1, 19, 20$ are also nearly 1. The same pattern continues with the quintic model but the leverages at $x = 0.5, 19.5$ are also becoming large. Finally, with the sixth degree (hexic) polynomial, all of the leverages are exactly one. This indicates that the sixth degree polynomial has to go through every data point exactly and thus every data point is extremely influential on the estimate of the sixth degree polynomial. (It is fortunate that there are only seven distinct x values. This discussion would really tank if we had to fit a seventh degree polynomial. [Think about it: quartic, quintic, hexic, ... tank])

As we fit larger polynomials, we get more high leverage cases (and more numerical instability). Actually, as in our example, this occurs when the size of the polynomial nears one less than the number of distinct x values and nearly all data points have distinct x values. *The estimated polynomials must go very nearly through all high leverage cases. To accomplish this the estimated polynomials may get very strange*. We now give all of the fitted polynomials for these data.

Model	Estimated polynomial
Linear	$\hat{y} = 0.252 - 0.029x$
Quadratic	$\hat{y} = 0.822 - 0.536x + 0.0253x^2$
Cubic	$\hat{y} = 1.188 - 1.395x + 0.1487x^2 - 0.0041x^3$
Quartic	$\hat{y} = 0.713 - 0.141x - 0.1540x^2 + 0.0199x^3$
	$- 0.00060x^4$
Quintic	$\hat{y} = 0.623 + 1.144x - 1.7196x^2 + 0.3011x^3$
	$- 0.01778x^4 + 0.000344x^5$
Hexic	$\hat{y} = 0.445 + 3.936x - 5.4316x^2 + 1.2626x^3$
	$- 0.11735x^4 + 0.004876x^5$
	$- 0.00007554x^6$

Figures 8.6 and 8.7 contain graphs of these estimated polynomials.

Figure 8.6 contains the estimated linear, quadratic, and cubic polynomials. The linear and quadratic curves fit about as one would expect from looking at the scatter plot Figure 8.5. For x values near the range 0 to 20, we could use these curves to predict y values and get reasonable, if not necessarily good, results. One could not say the same for the estimated cubic polynomial. The cubic curve takes on $\hat{y}$ values near -3 for some x values that are near 6. The y values in the data are between about -2 and 1.2; nothing in the data suggests that y values near -3 are likely to occur.

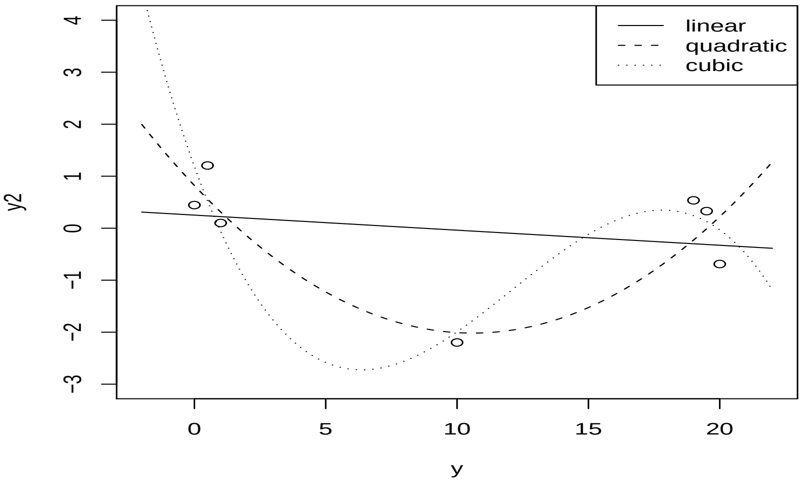

Figure 8.6: *Plots of linear (solid), quadratic (dashes), and cubic (dots) regression curves.*

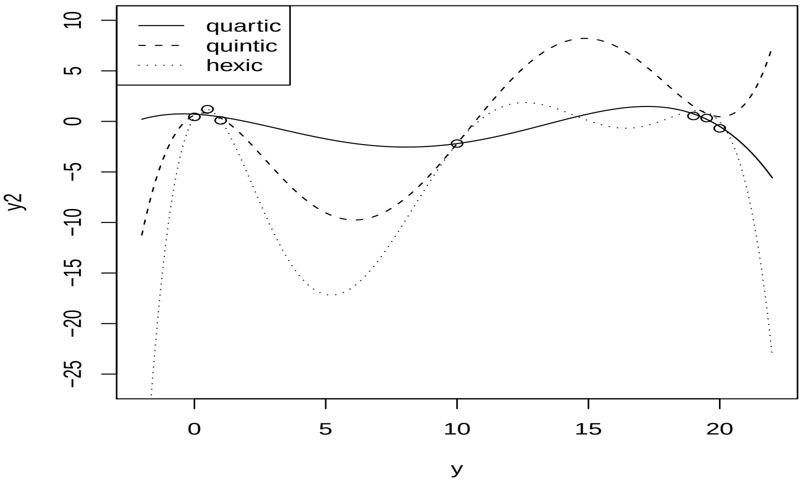

Figure 8.7: *Plots of quartic (solid), quintic (dashes), and hexic (dots) regression curves.*

Such predicted values are entirely the product of fitting a cubic polynomial. If we really knew that a cubic polynomial was correct for these data, the estimated polynomial would be perfectly appropriate. But most often we use polynomials to approximate the behavior of the data and for these data the cubic polynomial gives a poor approximation.

Figure 8.7 gives the estimated quartic, quintic, and hexic polynomials. Note that the scale on the *y* axis has changed drastically from Figure 8.6. Qualitatively, the fitted polynomials behave like the cubic except their behavior is even worse. These polynomials do very strange things everywhere except near the observed data.

Table 8.2: *Analysis of variance tables.*

Simple linear regression					
Source	*df*	*SS*	*MS*	*F*	*P*
Regression	1	0.457	0.457	0.33	0.59
Error	5	6.896	1.379		
Total	6	7.353			
Quadratic model					
Source	*df*	*SS*	*MS*	*F*	*P*
Regression	2	5.185	2.593	4.78	0.09
Error	4	2.168	0.542		
Total	6	7.353			
Cubic model					
Source	*df*	*SS*	*MS*	*F*	*P*
Regression	3	5.735	1.912	3.55	0.16
Error	3	1.618	0.539		
Total	6	7.353			
Quartic model					
Source	*df*	*SS*	*MS*	*F*	*P*
Regression	4	6.741	1.685	5.51	0.16
Error	2	0.612	0.306		
Total	6	7.353			
Quintic model					
Source	*df*	*SS*	*MS*	*F*	*P*
Regression	5	6.856	1.371	2.76	0.43
Error	1	0.497	0.497		
Total	6	7.353			
Hexic model					
Source	*df*	*SS*	*MS*	*F*	*P*
Regression	6	7.353	1.2255	—	—
Error	0	0.000	—		
Total	6	7.353			

Another phenomenon that sometimes occurs when fitting large models to data is that the mean squared error gets unnaturally small. Table 8.2 gives the analysis of variance tables for all of the polynomial models. Our original data were a sample from a $N(0,1)$ distribution. The data were constructed with no regression structure so the best estimate of the variance comes from the total line and is $7.353/6 = 1.2255$. This value is a reasonable estimate of the true value 1. The *MSE* from the simple linear regression model also provides a reasonable estimate of $\sigma^2 = 1$. The larger models do not work as well. Most have variance estimates near 0.5, while the hexic model does not even allow an estimate of σ^2 because it fits every data point perfectly. By fitting models that are too large one can often make the *MSE* artificially small. For example, the quartic model has a *MSE* of 0.306 and an F statistic of 5.51; if it were not for the small value of *dfE*, such an F value would be highly significant. *If you find a large model that has an unnaturally small MSE with a reasonable number of degrees of freedom, everything can appear to be significant even though nothing you look at is really significant.*

Just as the mean squared error often gets unnaturally small when fitting large models, R^2 gets unnaturally large. As we have seen, there can be no possible reason to use a larger model than the quadratic with its R^2 of 0.71 for these 7 data points, but the cubic, quartic, quintic, and hexic models have R^2s of 0.78, 0.92, 0.93, and 1, respectively. □

8.3 Other Basis Functions

In a SLR, one method for testing lack of fit was to fit a larger polynomial model. In particular, for the Hooker data we fit a fifth degree polynomial,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \beta_5 x_i^5 + \varepsilon_i.$$

There was no particularly good reason to fit a fifth degree, rather than a third degree or seventh degree polynomial. We just picked a polynomial that we hoped would be larger than we needed.

Rather than expanding the SLR model by adding polynomial terms, we can add other functions of x to the model. Most commonly used functions are simplified if we rescale x into a new variable taking values between 0 and 1, say, $\tilde{x}$. Commonly used functions are trig. functions, so we might fit a full model consisting of

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 \cos(\pi \tilde{x}_i) + \beta_3 \sin(\pi \tilde{x}_i) + \beta_4 \cos(\pi 2 \tilde{x}_i) + \beta_5 \sin(\pi 2 \tilde{x}_i) + \varepsilon_i \tag{8.3.1}$$

or a full model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 \cos(\pi \tilde{x}_i) + \beta_3 \cos(\pi 2 \tilde{x}_i) + \beta_4 \cos(\pi 3 \tilde{x}_i) + \beta_5 \cos(\pi 4 \tilde{x}_i) + \varepsilon_i. \tag{8.3.2}$$

As with the polynomial models, the number of additional predictors to add depends on how complicated the data are. For the purpose of testing lack of fit, we simply need the number to be large enough to find any salient aspects of the data that are not fitted well by the SLR model.

Another approach is to add a number of indicator functions. An *indicator function* of a set A is defined as

$$I_A(\theta) = \begin{cases} 1 & \text{if } \theta \in A \\ 0 & \text{if } \theta \notin A \end{cases}. \tag{8.3.3}$$

We can fit models like

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 I_{[0,.25)}(\tilde{x}_i) + \beta_3 I_{[.25,.5)}(\tilde{x}_i) + \beta_4 I_{[.5,.75)}(\tilde{x}_i) + \beta_5 I_{[.75,1]}(\tilde{x}_i) + \varepsilon_i.$$

Adding indicator functions of length 2^{-j} defined on $\tilde{x}$ is equivalent to adding *Haar wavelets* to the model, cf. Christensen (2001). Unfortunately, no regression programs will fit this model because it is no longer a regression model. It is no longer a regression model because there is a redundancy in the predictor variables. The model includes an intercept, which corresponds to a predictor variable that always takes on the value 1. However, if we add together our four indicator functions, their sum is also a variable that always takes on the value 1. To evade this problem, we need either to delete one of the indicator functions (doesn't matter which one) or remove the intercept from the model. Dropping the last indicator is convenient, so we fit

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 I_{[0,.25)}(\tilde{x}_i) + \beta_3 I_{[.25,.5)}(\tilde{x}_i) + \beta_4 I_{[.5,.75)}(\tilde{x}_i) + \varepsilon_i. \tag{8.3.4}$$

Any continuous function defined on an interval $[a,b]$ can be approximated arbitrarily well by a sufficiently large polynomial. Similar statements can be made about the other classes of functions introduced here. Because of this, these classes of functions are known as *basis functions*.

EXAMPLE 8.3.1. We illustrate the methods on the Hooker data. With x the temperature, we defined $\tilde{x} = (x - 180.5)/30.5$. Fitting model (8.3.1) gives

Analysis of variance

Source	*df*	*SS*	*MS*	*F*	*P*
Regression	5	447.185	89.437	3364.82	0.000
Residual Error	25	0.665	0.027		
Total	30	447.850			

A test of whether model (8.3.1) fits significantly better than SLR has statistic

$$F = \frac{(3.68 - 0.665)/(29 - 25)}{0.027} = 27.9$$

Clearly the reduced model of a simple linear regression fits worse than the model with two additional sine and cosine terms.

Fitting model (8.3.2) gives

Analysis of variance

Source	*df*	*SS*	*MS*	*F*	*P*
Regression	5	447.208	89.442	3486.60	0.000
Residual Error	25	0.641	0.026		
Total	30	447.850			

A test of whether the cosine model fits significantly better than SLR has statistic

$$F = \frac{(3.68 - 0.641)/(29 - 25)}{0.026} = 29.2.$$

Clearly the reduced model of a simple linear regression fits worse than the model with four additional cosine terms.

Fitting model (8.3.4) gives

Analysis of variance

Source	*df*	*SS*	*MS*	*F*	*P*
Regression	4	446.77	111.69	2678.37	0.000
Residual Error	26	1.08	0.042		
Total	30	447.85			

A test of whether this Haar wavelet model fits significantly better than SLR has statistic

$$F = \frac{(3.68 - 1.08)/(29 - 26)}{0.042} = 20.6.$$

Clearly the reduced model of a simple linear regression fits worse than the model with three additional indicator functions.

8.3.1 High order models

For continuous basis functions like the trig functions, high order models can behave as strangely between the data points as polynomials. For example, Figure 8.8 contains a plot of the 7 data points discussed in Section 8.2 and, using $\tilde{x} = x/20$, a fitted cosine model with 5 terms and an intercept,

$$y_i = \beta_0 + \beta_1 \cos(\pi\tilde{x}_i) + \beta_2 \cos(\pi 2\tilde{x}_i) + \beta_3 \cos(\pi 3\tilde{x}_i) + \beta_4 \cos(\pi 4\tilde{x}_i) + \beta_5 \cos(\pi 5\tilde{x}_i) + \varepsilon_i.$$

The fit away from the data is even worse than for 5th and 6th order polynomials.

8.4 Partitioning Methods

The basic idea of the partitioning method is quite simple. Suppose we are fitting a simple linear regression but that the actual relationship between x and y is a quadratic. If you can split the x values into two parts near the maximum or minimum of the quadratic, you can get a much better approximate fit using two lines instead of one. More generally, the idea is that an approximate model should work better on a smaller set of data that has predictor variables that are more similar. Thus, if the original model is wrong, we should get a better approximation to the truth by fitting the original

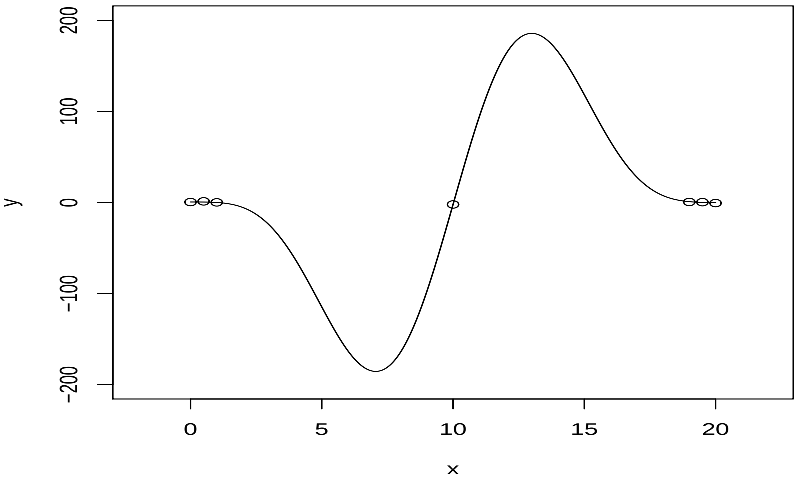

Figure 8.8: *Plot of fifth order cosine model.*

model on a series of smaller subsets of the data. Of course if the original model is correct, it should work about the same on each subset as it does on the complete data. The statistician partitions the data into disjoint subsets, fits the original model on each subset, and compares the overall fit of the subsets to the fit of the original model on the entire data. The statistician is free to select the partitions, including the number of distinct sets, but the subsets need to be chosen based on the predictor variable(s) alone.

EXAMPLE 8.4.1. We illustrate the partitioning method by splitting the Hooker data into two parts. Our partition sets are the data with the 16 smallest temperatures and the data with the 15 largest temperatures. We then fit a separate regression line to each partition. The two fitted lines are given in Figure 8.9. The ANOVA table is

Analysis of variance

Source	*df*	*SS*	*MS*	*F*	*P*
Regression	3	446.66	148.89	3385.73	0.000
Error	27	1.19	0.04		
Total	30	447.85			

A test of whether this partitioning fits significantly better than SLR has statistic

$$F = \frac{(3.68-1.19)/(29-27)}{0.04} = 31.125.$$

Clearly the reduced model of a simple linear regression fits worse than the model with two SLRs. Note that this is a simultaneous test of whether the slopes and intercepts are the same in each partition.

Fitting the partitioned model

We now consider three different ways to fit this partitioned model. Our computations will be subject to some round-off error. One way to fit this model is simply to divide the data into two parts and fit a simple linear regression to each one. Fitting the lowest 16 x (temperature) values gives

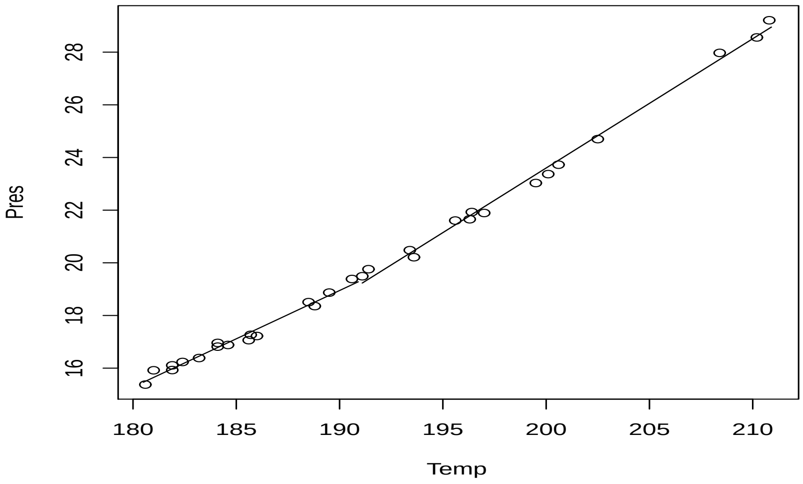

Figure 8.9: *Hooker data, partition method.*

Table of Coefficients: low x values.

Predictor	$\hat{\beta}_k$	SE($\hat{\beta}_k$)	t	P
Constant	-50.725	2.596	-19.54	0.000
x-low	0.36670	0.01404	26.13	0.0001

Analysis of variance: low x values.

Source	df	SS	MS	F	P
Regression	2	4687.1	2342.5	81269.77	0.000
Error	14	0.4	0.0		
Total	16	4687.5			

To get some extra numerical accuracy, from the F statistic we can compute $MSE = 2342.5/81269.77 = 0.028836$ so $SSE = 0.4037$. Fitting the highest 15 x values gives

Table of Coefficients: high x values.

Predictor	$\hat{\beta}_k$	SE($\hat{\beta}_k$)	t	P
Constant	-74.574	2.032	-36.70	0.000
x-high	0.49088	0.01020	48.12	0.000

Analysis of variance: high x values.

Source	df	SS	MS	F	P
Regression	2	8193.9	4096.9	67967.66	0.000
Error	13	0.8	0.1		
Total	15	8194.7			

Again, from the F statistic $MSE = 4096.9/67967.66 = 0.060277$, so $SSE = 0.7836$. The fit of the overall model is obtained by pooling the two Error terms to give $dfE(F) = 14+13 = 27$, $SSE(F) = 0.4037+0.7836 = 1.1873$, with $MSE(F) = 0.044$

A more efficient way to proceed is to fit both simple linear regressions at once. Construct a variable h that identifies the 15 high values of x. In other words, h is 1 for the 15 highest temperature values and 0 for the 16 lowest values. Define $x_1 = h \times x$, $h_2 = 1 - h$, and $x_2 = h_2 \times x$. Fitting these four variables in a *regression through the origin* gives

Table of Coefficients

Predictor	$\hat{\beta}_k$	SE($\hat{\beta}_k$)	t	P
h_2	−50.725	3.205	−15.82	0.000
x_2	0.36670	0.01733	21.16	0.000
h	−74.574	1.736	−42.97	0.000
x_1	0.490875	0.008712	56.34	0.000

Analysis of variance

Source	df	SS	MS	F	P
Regression	4	12881.0	3220.2	73229.01	0.000
Error	27	1.2	0.0		
Total	31	12882.2			

Note that these regression estimates agree with those obtained from fitting each set of data separately. The standard errors differ because here we are pooling the information in the error rather than using separate estimates of σ^2 from each set of data.

The way the model was originally fitted was regressing on x, h, and x_1. The ANOVA table is as given earlier and the table of regression coefficients is

Table of Coefficients

Predictor	$\hat{\beta}_k$	SE($\hat{\beta}_k$)	t	P
Constant	−50.725	3.205	−15.82	0.000
x	0.36670	0.01733	21.16	0.000
h	−23.849	3.645	−6.54	0.000
x_1	0.12418	0.01940	6.40	0.000

The slope for the low group is 0.36670 and for the high group it is $0.36670 + 0.12418 = 0.49088$. The t test for whether the slopes are different in a model that retains separate intercepts is based on the x_1 row of this table and has $t = 6.40$. The intercepts also look different. The intercept for the low group is -50.725 and for the high group it is $-50.725 + -23.849 = -74.574$. The t test for whether the intercepts are different in a model that retains separate slopes is based on the h row and has $t = -6.54$.

Computer commands

We give three different sets of Minitab commands that provide the last of our analyses. In the text we discussed the output from fitting the regression command. Alternatively, one can fit this model using a general linear model procedure like Minitab's glm.

```
MTB > regress Pres 3 x h x1
MTB > glm c2 = h h*x;
SUBC> covar x.
MTB > glm Pres = h x h*x;
SUBC> covar x.
```

Note that in the glm commands, there does not exist a variable called 'h*x'. This is a term that we are telling glm to construct out of two variables that do exist. The only difference between the two "glm" commands, is that when the model does not specify an x term, glm fits x before fitting h. The table of coefficients provided by the two glm commands are the same but different from that provided by regress. The glm command by default only gives results for the intercept and terms that involve the covariate x.

Table of Coefficients

Predictor	$\hat{\beta}_k$	SE($\hat{\beta}_k$)	t	P
Constant	-62.650	1.823	-34.37	0.000
x	0.428787	0.009700	44.21	0.000
x*h				
0	-0.062089	0.009700	-6.40	0.000

The “constant” value of -62.650 is the average of the two intercept estimates that were reported earlier. Similarly, the “x” value 0.428787 is the average of the two slope estimates reported earlier. Although the table does not give enough information to reconstruct the two intercepts, the slope for the low group ($h = 0$) is $0.428787 + (-0.062089)$ and the slope for the high group is $0.428787 - (-0.062089)$. Note that the t test for “x*h 0” is the same -6.40 as that reported earlier for testing whether the slopes were different.

Similar things would occur if using SAS’s “proc glm” except the exact relationship between the estimates reported and the slopes and intercepts of the lines would change.

Our discussion used the variable h that partitions the data into the smallest 16 observations and the largest 15 observations. Minitab provides a test that partitions the data into the 18 observations below $\bar{x}_\cdot = 191.79$ and the 13 observations larger than the mean. Their test gets considerably more complicated when there is more than one predictor variable. They perform both this test (in more complicated situations, these tests) and a version of the test described in the next subsection, and combine the results from the various tests.

8.4.1 Utts’ method

Utts (1982) proposed a lack of fit test based on comparing the original (reduced) model to a full model that consists of fitting the original model on a subset of the original data. In other words, you fit the model on all the data and test this against a full model that consists of fitting the model on a subset of the data. The subset is chosen to contain the points closest to $\bar{x}_\cdot$. Although it seems like fitting the model to a reduced set of points should create a reduced model, just the opposite is true. To fit a model to a reduced set of points, we can think of fitting the original model and then adding a separate parameter for every data point that we want to exclude from the fitting procedure. In fact, that is what makes this a partitioning method. There is one subset which consists of the central data and the rest of the partition has every data point in a separate set.

The central subset is chosen to be a group of points close to $\bar{x}_\cdot$. With only one predictor variable, it is easy to determine a group of central points. It turns out that for models with an intercept, the leverages are really measures of distance from $\bar{x}_\cdot$, see Christensen (2011, Section 13.1), so even with more predictor variables, one could choose a group of points that have the lowest leverages in the original model.

EXAMPLE 8.4.1. We consider first the use the 15 central points with leverages below 0.05; about half the data. We then consider a group of 6 central points; about a quarter of the data.

The ANOVA table when fitting a simple linear regression to 15 central points is

Analysis of variance

Source	df	SS	MS	F	P
Regression	1	40.658	40.658	1762.20	0.000
Error	13	0.300	0.023		
Total	14	40.958			

The lack of fit test against a reduced model of simple linear regression on the entire data has

$$F = \frac{(3.68 - 0.300)/(29 - 13)}{0.023} = 9.18,$$

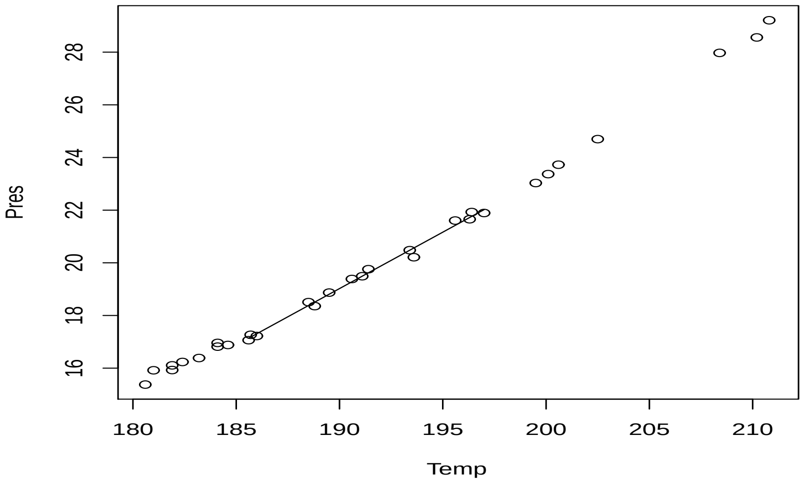

Figure 8.10: *Hooker data, Utts method with 15 points.*

which is highly significant. Figure 8.10 illustrates the fitting method.

When using 6 central points having leverages below 0.035, the ANOVA table is

Analysis of variance

Source	*df*	*SS*	*MS*	*F*	*P*
Regression	1	1.6214	1.6214	75.63	0.001
Error	4	0.0858	0.0214		
Total	5	1.7072			

and the F statistic is

$$F = \frac{(3.68 - 0.0858)/(29-4)}{0.0214} = 6.72.$$

This is much bigger than 1 and easily significant at the 0.01 level. Both tests suggest lack of fit. Figure 8.11 illustrates the fitting method. □

My experience is that Utt's test tends to work better with relatively small groups of central points. (Even though the F statistic here was smaller for the smaller group.) Minitab incorporates a version of Utt's test that defines the central region as those points with leverages less than $1.1p/n$ where p is the number of regression coefficients in the model, so for a simple linear regression $p = 2$. For these data, their central region consists of the 22 observations with temperature between 183.2 and 200.6.

8.5 Fisher's Lack-of-fit Test

We now introduce Fisher's lack-of-fit test for the Hooker data. The test is discussed in much more detail in Chapter 12 and extended in Chapter 15. For now, notice that the predictor variable includes two replicate temperatures: $x = 181.9$ with y values 15.106 and 15.928 and $x = 184.1$ with y values 16.959 and 16.817. In this case, the computation for Fisher's lack-of-fit test is quite simple. We use the replicated x values to obtain a measure of pure error. First, compute the sample variance of the y_is at each replicated x value. There are 2 observations at each replicated x, so the sample variance

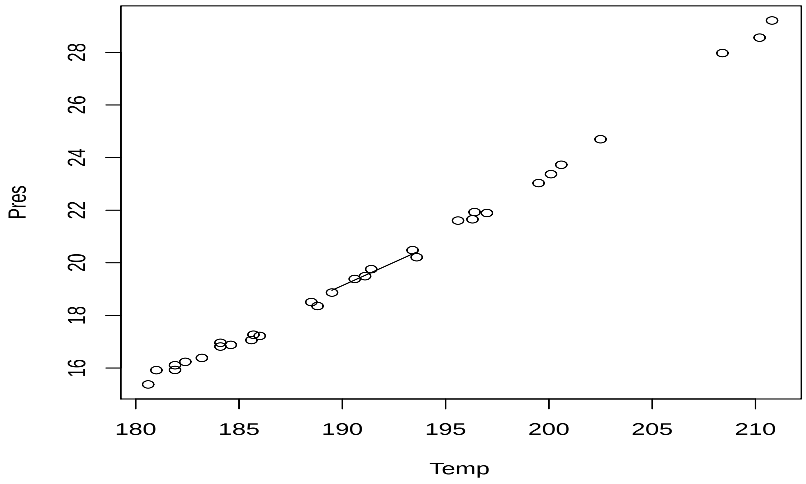

Figure 8.11: *Hooker data, Utts method with 6 points.*

computed at each x has 1 degree of freedom. Since there are two replicated xs, the pure error has $1+1=2$ degrees of freedom. To compute the sum of squares for pure error, observe that when $x=181.9$, the mean y is 15.517. The contribution to the sum of squares pure error from this x value is $(15.106-15.517)^2+(15.928-15.517)^2$. A similar contribution is computed for $x=184.1$ and they are added to get the sum of squares pure error. The degrees of freedom and sum of squares for lack of fit are found by taking the values from the original error and subtracting the values for the pure error. The F test for lack of fit examines the mean square lack of fit divided by the mean square pure error.

Analysis of variance

Source	*df*	*SS*	*MS*	*F*	*P*
Regression	1	444.17	444.17	3497.89	0.000
Error	29	3.68	0.13		
(Lack of Fit)	27	3.66	0.14	10.45	0.091
(Pure Error)	2	0.03	0.01		
Total	30	447.85			

The F statistic for lack of fit, 10.45, seems substantially larger than 1, but because there are only 2 degrees of freedom in the denominator, the P value is a relatively large 0.09. This method is closely related to one-way analysis of variance as discussed in Chapter 12.

8.6 Exercises

EXERCISE 8.6.1. Dixon and Massey (1969) presented data on the relationship between IQ scores and results on an achievement test in a general science course. Table 8.3 contains a subset of the data. Fit the simple linear regression model of achievement on IQ and the quadratic model of achievement on IQ and IQ squared. Evaluate both models and decide which is the best.

EXERCISE 8.6.2. In Exercise 7.4.2 we considered data on the relationship between farm sizes

Table 8.3: *IQs and achievement scores.*

IQ	Achiev.	IQ	Achiev.	IQ	Achiev.	IQ	Achiev.	IQ	Achiev.
100	49	105	50	134	78	107	43	122	66
117	47	89	72	125	39	121	75	130	63
98	69	96	45	140	66	90	40	116	43
87	47	105	47	137	69	132	80	101	44
106	45	95	46	142	68	116	55	92	50
134	55	126	67	130	71	137	73	120	60
77	72	111	66	92	31	113	48	80	31
107	59	121	59	125	53	110	41	117	55
125	27	106	49	120	64	114	29	93	50

Table 8.4: *Weights for various heights.*

Ht.	Wt.	Ht.	Wt.
65	120	63	110
65	140	63	135
65	130	63	120
65	135	72	170
66	150	72	185
66	135	72	160

and the acreage in corn. Fit the linear, quadratic, cubic, and quartic polynomial models to the logs of the acreages in corn. Find the model that fits best. Check the assumptions for this model.

EXERCISE 8.6.3. Use two methods other than fitting polynomial models to test for lack of fit in Exercise 8.6.1

EXERCISE 8.6.4. Based on the height and weight data given in Table 8.4, Fit a simple linear regression of weight on height for these data and check the assumptions. Give a 99% confidence interval for the mean weight of people with a 72 inch height. Test the lack of fit of the simple linear regression model.

EXERCISE 8.6.5. Jensen (1977) and Weisberg (1985, p. 101) considered data on the outside diameter of crank pins that were produced in an industrial process. The diameters of batches of crank pins were measured on various days; if the industrial process is 'under control' the diameters should not depend on the day they were measured. A subset of the data is given in Table 8.5 in a format consistent with performing a regression analysis on the data. The diameters of the crank pins are actually $.742 + y_{ij}10^{-5}$ inches, where the y_{ij}s are reported in Table 8.5. Perform polynomial regressions on the data. Give two lack of fit tests for the simple linear regression not based on polynomial regression.

EXERCISE 8.6.6. Beineke and Suddarth (1979) and Devore (1991, p. 380) consider data on roof

Table 8.5: *Jensen's crank pin data.*

Days	Diameters	Days	Diameters	Days	Diameters	Days	Diameters
4	93	10	93	16	82	22	90
4	100	10	88	16	72	22	92
4	88	10	87	16	80	22	82
4	85	10	87	16	72	22	77
4	89	10	87	16	89	22	89

Table 8.6: *Axial stiffness index data.*

Plate	ASI	Plate	ASI	Plate	ASI	Plate	ASI	Plate	ASI
4	309.2	6	402.1	8	392.4	10	346.7	12	407.4
4	409.5	6	347.2	8	366.2	10	452.9	12	441.8
4	311.0	6	361.0	8	351.0	10	461.4	12	419.9
4	326.5	6	404.5	8	357.1	10	433.1	12	410.7
4	316.8	6	331.0	8	409.9	10	410.6	12	473.4
4	349.8	6	348.9	8	367.3	10	384.2	12	441.2
4	309.7	6	381.7	8	382.0	10	362.6	12	465.8

supports involving trusses that use light gauge metal connector plates. Their dependent variable is an axial stiffness index (ASI) measured in kips per inch. The predictor variable is the length of the light gauge metal connector plates. The data are given in Table 8.6.

Fit linear, quadratic, cubic, and quartic polynomial regression models using powers of x, the plate length, and using powers of $x - \bar{x}_\cdot$, the plate length minus the average plate length. Compare the results of the two procedures. If your computer program will not fit some of the models, report on that in addition to comparing results for the models you could fit.

EXERCISE 8.6.7. Consider fitting quadratic models $y_i = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 + \varepsilon_i$ and $y_i = \beta_0 + \beta_1 (x_i - \bar{x}_\cdot) + \beta_2 (x_i - \bar{x}_\cdot)^2 + \varepsilon_i$. Show that $\gamma_2 = \beta_2$, $\gamma_1 = \beta_1 + \beta_2 \bar{x}_\cdot$, and $\gamma_0 = \beta_0 - \beta_1 \bar{x}_\cdot + \beta_2 \bar{x}_\cdot^2$.