13 Nonlinear and Multiple Regression

13.3 Polynomial Regression

Polynomial Regression

The nonlinear yet intrinsically linear models, involved functions of the independent variable x that were either strictly increasing or strictly decreasing.

In many situations, either theoretical reasoning or else a scatter plot of the data suggests that the true regression function $\mu_{Y \cdot x}$ has one or more peaks or valleys—that is, at least one relative minimum or maximum.

In such cases, a polynomial function $y = \beta_0 + \beta_1 x + \ldots + \beta_k x^k$ may provide a satisfactory approximation to the true regression function.

Polynomial Regression

Definition

The ***k*th-degree polynomial regression model equation** is

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + \epsilon \qquad \textbf{(13.6)}$$

where ϵ is a normally distributed random variable with

$$\mu_\epsilon = 0 \qquad \sigma_\epsilon^2 = \sigma^2 \qquad \textbf{(13.7)}$$

Polynomial Regression

From (13.6) and (13.7), it follows immediately that

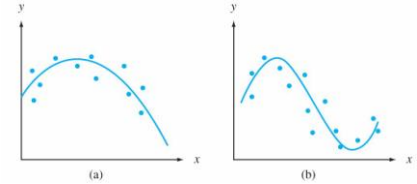
$$\mu_{Y\cdot x} = \beta_0 + \beta_1 x + \ldots + \beta_k x^k \qquad \sigma^2_{Y\cdot x} = \sigma^2 \qquad \textbf{(13.8)}$$

In words, the expected value of Y is a kth-degree polynomial function of x, whereas the variance of Y, which controls the spread of observed values about the regression function, is the same for each value of x. The observed pairs $(x_1, y_1), \ldots, (x_n, y_n)$ are assumed to have been generated independently from the model (13.6).

5

Polynomial Regression

Figure 13.10 illustrates both a quadratic and cubic model; very rarely in practice is it necessary to go beyond $k = 3$.



Quadratic regression model cubic regression model

Figure 13.10

6

Estimating Parameters

7

Estimating Parameters

To estimate the β_s, consider a trial regression function $y = b_0 + b_1 x + \ldots + b_k x^k$.

Then the goodness of fit of this function to the observed data can be assessed by computing the sum of squared deviations

$$f(b_0, b_1, \ldots, b_k) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i + b_2 x_i^2 + \cdots + b_k x_i^k)]^2 \qquad \textbf{(13.9)}$$

8

Estimating Parameters

According to the principle of least squares, the estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are those values of $b_0, b_1, \ldots, b_k$ that minimize Expression (13.9).

It should be noted that when $x_1, x_2, \ldots, x_n$ are all different, there is a polynomial of degree $n - 1$ that fits the data perfectly, so that the minimizing value of (13.9) is 0 when $k = n - 1$.

However, in virtually all applications, the polynomial model (13.6) with large k is quite unrealistic.

9

Estimating Parameters

To find the minimizing values in (13.9), take the $k + 1$ partial derivatives $\partial f/\partial b_0, \partial f/\partial b_1, \ldots, \partial f/\partial b_k$ and equate them to 0. This gives a system of normal equations for the estimates.

Because the trial function $b_0 + b_1x + \ldots + b_kx^k$ is linear in $b_0, \ldots, b_k$ (though not in x), the $k + 1$ normal equations are linear in these unknowns:

$$\begin{aligned} b_0 n + b_1\Sigma x_i + b_2\Sigma x_i^2 + \cdots + b_k\Sigma x_i^k &= \Sigma y_i \\ b_0\Sigma x_i + b_1\Sigma x_i^2 + b_2\Sigma x_i^3 + \cdots + b_k\Sigma x_i^{k+1} &= \Sigma x_i y_i \\ \vdots \qquad\quad \vdots \qquad\quad & \vdots \\ b_0\Sigma x_i^k + b_1\Sigma x_i^{k+1} + \cdots + b_k\Sigma x_i^{2k} &= \Sigma x_i^k y_i \end{aligned} \qquad \textbf{(13.10)}$$

10

Example 7

The article “Residual Stresses and Adhesion of Thermal Spray Coatings” (*Surface Engineering*, 2005: 35–40) considered the relationship between the thickness (μm) of NiCrAl coatings deposited on stainless steel substrate and corresponding bond strength (MPa).

The following data was read from a plot in the paper:

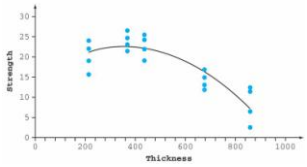
Thickness	220	220	220	220	370	370	370	370	440	440
Strength	24.0	22.0	19.1	15.5	26.3	24.6	23.1	21.2	25.2	24.0
Thickness	440	440	680	680	680	680	860	860	860	860
Strength	21.7	19.2	17.0	14.9	13.0	11.8	12.2	11.2	6.6	2.8

11

Example 7

cont'd

The scatter plot in Figure 13.11(a) supports the choice of the quadratic regression model.



Scatter plot of data

Figure 13.11(a)

12

Example 7

cont'd

Figure 13.11(b) contains Minitab output from a fit of this model.

```
The regression equation is
strength = 14.5 + 0.0432 thickness - 0.000060 thicksqd

Predictor              Coef          SE Coef           T         P
Constant             14.521            4.754        3.05     0.007
thickness           0.04323          0.01981        2.18     0.043
thicksqd         -0.00006001       0.00001786       -3.36     0.004

S = 3.26937      R-Sq = 78.0%      R-Sq(adj) = 75.4%

Analysis of Variance

Source              DF        SS          MS         F         P
Regression           2    643.29      321.65     30.09     0.000
Residual Error      17    181.71       10.69
Total               19    825.00

Predicted Values for New Observations

New
Obs        Fit      SE Fit           95% CI               95% PI
  1     21.136       1.167   (18.674, 23.598)     (13.812, 28.460)
  2     10.704       1.189   ( 8.195, 13.212)     ( 3.364, 18.043)

Values of Predictors for New Observations

New
Obs          thickness          thicksqd
  1                500            250000
  2                800            640000
```

Minitab output from fit of quadratic model

Figure 13.11(b)

Example 7

cont'd

The estimated regression coefficients are

$\hat{\beta}_0 = 14.521 \qquad \hat{\beta}_1 = .04323 \qquad \hat{\beta}_2 = -.00006001$

from which the estimated regression function is

$y = 14.521 + .04323x - .00006001x^2$

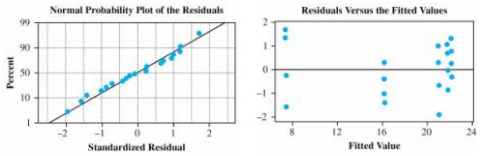
Example 7

Substitution of the successive x values 220, 220, . . . , 860, and 860 into this function gives the predicted values $\hat{y}_1 = 21.128, \ldots, \hat{y}_{20} = 7.321$, and the residuals $y_1 - \hat{y}_1 = 2.872, \ldots, y_{20} - \hat{y}_{20} = -4.521$ result from subtraction.

Example 7

cont'd

Figure 13.12 shows a plot of the standardized residuals versus $\hat{y}$ and also a normal probability plot of the standardized residuals, both of which validate the quadratic model.



Diagnostic plots for quadratic model fit to data of Example 7

Figure 13.12

$\hat{\sigma}^2$ and R^2

$\hat{\sigma}^2$ and R^2

To make further inferences, the error variance σ^2 must be estimated.

With $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \cdots + \hat{\beta}_k x_i^k$, the ith residual is $y_i - \hat{y}_i$, and the sum of squared residuals (error sum of squares) is SSE = $\Sigma(y_i - \hat{y}_i)^2$.

The estimate of σ^2 is then

$$\hat{\sigma}^2 = s^2 = \frac{\text{SSE}}{n - (k + 1)} = \text{MSE} \qquad \textbf{(13.11)}$$

where the denominator $n - (k + 1)$ is used because $k + 1$ df are lost in estimating $\beta_0, \beta_1, \ldots, \beta_k$.

$\hat{\sigma}^2$ and R^2

If we again let SST = $\Sigma(y_i - \bar{y})^2$, then SSE/SST is the proportion of the total variation in the observed y_i's that is not explained by the polynomial model.

The quantity 1 – SSE/SST, the proportion of variation explained by the model, is called the **coefficient of multiple determination** and is denoted by R^2.

Consider fitting a cubic model to the data in Example 7. Because this model includes the quadratic as a special case, the fit will be at least as good as the fit to a quadratic.

$\hat{\sigma}^2$ and R^2

More generally, with SSE_k = the error sum of squares from a kth-degree polynomial, $\text{SSE}_{k'} \le \text{SSE}_k$ and $R^2_{k'} \ge R^2_k$ whenever $k' > k$.

Because the objective of regression analysis is to find a model that is both simple (relatively few parameters) and provides a good fit to the data, a higher-degree polynomial may not specify a better model than a lower-degree model despite its higher R^2 value.

$\hat{\sigma}^2$ and R^2

To balance the cost of using more parameters against the gain in R^2, many statisticians use the **adjusted coefficient of multiple determination**

$$\text{adjusted } R^2 = 1 - \frac{n-1}{n-(k+1)} \cdot \frac{\text{SSE}}{\text{SST}} = \frac{(n-1)R^2 - k}{n-1-k} \quad \textbf{(13.12)}$$

$\hat{\sigma}^2$ and R^2

Adjusted R^2 adjusts the proportion of unexplained variation upward [since the ratio $(n-1)/(n-k-1)$ exceeds 1], which results in *adjusted* $R^2 < R^2$. For example, if $R_2^2 = .66$, $R_3^2 = .70$, and $n = 10$, then

$$\text{adjusted } R_2^2 = \frac{9(.66)-2}{10-3} = .563 \quad \text{adjusted } R_3^2 = \frac{9(.70)-3}{10-4} = .550$$

Thus the small gain in R^2 in going from a quadratic to a cubic model is not enough to offset the cost of adding an extra parameter to the model.

Example 8

Example 7 Continued…

SSE and SST are typically found on computer output in an ANOVA table.

Figure 13.11(b) gives SSE = 181.71 and SST = 825.00 for the bond strength data, from which
$R^2 = 1 - 181.71/825.00 = .780$
(alternatively, $R^2 = \text{SSR/SST} = 643.29/825.00 = .780$).

Example 8

cont'd

Thus 78.0% of the observed variation in bond strength can be attributed to the model relationship.

Adjusted $R^2 = .754$, only a small downward change in R^2. The estimates of σ^2 and σ are

$$\hat{\sigma}^2 = s^2 = \frac{\text{SSE}}{n-(k+1)} = \frac{181.71}{20-(2+1)} = 10.69$$

$$\hat{\sigma} = s = 3.27$$

Statistical Intervals and Test Procedures

Statistical Intervals and Test Procedures

Because the y_i's appear in the normal equations (13.10) only on the right-hand side and in a linear fashion, the resulting estimates $\hat{\beta}_0, \ldots, \hat{\beta}_k$ are themselves linear functions of the y_i's.

Thus the estimators are linear functions of the Y_i's, so each $\hat{\beta}_i$ has a normal distribution.

It can also be shown that each $\hat{\beta}_i$ is an unbiased estimator of β_i.

Statistical Intervals and Test Procedures

Let $\sigma_{\hat{\beta}_i}$ denote the standard deviation of the estimator $\hat{\beta}_i$. This standard deviation has the form

$$\sigma_{\hat{\beta}_i} = \sigma \cdot \left\{ \begin{array}{l} \text{a complicated expression involving } all \\ x_j\text{'s, } x_j^2\text{'s, } \ldots, \text{ and } x_j^k\text{'s} \end{array} \right\}$$

Fortunately, the expression in braces has been programmed into all of the most frequently used statistical software packages.

The estimated standard deviation of $\hat{\beta}_i$ results from substituting s in place of σ in the expression for $\sigma_{\hat{\beta}_i}$.

Statistical Intervals and Test Procedures

These estimated standard deviations $s_{\hat{\beta}_0}, s_{\hat{\beta}_1}, \ldots,$ and $s_{\hat{\beta}_k}$ appear on output from all the aforementioned statistical packages.

Let $S_{\hat{\beta}_i}$ denote the estimator of $\sigma_{\hat{\beta}_i}$ —that is, the random variable whose observed value is $s_{\hat{\beta}_i}$.

Then it can be shown that the standardized variable

$$T = \frac{\hat{\beta}_i - \beta_i}{S_{\hat{\beta}_i}} \qquad \textbf{(13.13)}$$

has a t distribution based on $n - (k + 1)$ df. This leads to the following inferential procedures.

Statistical Intervals and Test Procedures

A 100(1 – α)% CI for β_i, the coefficient of x^i in the polynomial regression function, is

$$\hat{\beta}_i \pm t_{\alpha/2,n-(k+1)} \cdot s_{\hat{\beta}_i}$$

A test of H_0: $\beta_i = \beta_{i0}$ is based on the t statistic value

$$t = \frac{\hat{\beta}_i - \beta_{i0}}{s_{\hat{\beta}_i}}$$

The test is based on $n – (k + 1)$ df and is upper-, lower-, or two-tailed according to whether the inequality in H_a is >, <, or ≠.

Statistical Intervals and Test Procedures

A point estimate of $\mu_{Y \cdot x}$—that is, of

$\beta_0 + \beta_1 x + \cdots + \beta_k x^k$—is $\hat{\mu}_{Y \cdot x} = \hat{\beta}_0 + \hat{\beta}_1 x + \cdots + \hat{\beta}_k x^k$.

The estimated standard deviation of the corresponding estimator is rather complicated.

Many computer packages will give this estimated standard deviation for *any* x value upon request.

Statistical Intervals and Test Procedures

This, along with an appropriate standardized t variable, can be used to justify the following procedures.

Let x^* denote a specified value of x. A 100(1 – α)% CI for $\mu_{Y \cdot x^*}$ is

$$\hat{\mu}_{Y \cdot x^*} \pm t_{\alpha/2,n-(k+1)} \cdot \left\{ \begin{matrix} \text{estimated SD of} \\ \hat{\mu}_{Y \cdot x^*} \end{matrix} \right\}$$

Statistical Intervals and Test Procedures

With $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^* + \cdots + \hat{\beta}_k (x^*)^k$, $\hat{y}$ denoting the calculated value of $\hat{Y}$ for the given data, and $s_{\hat{Y}}$ denoting the estimated standard deviation of the statistic $\hat{Y}$, the formula for the CI is much like the one in the case of simple linear regression:

$$\hat{y} \pm t_{\alpha/2,n-(k+1)} \cdot s_{\hat{Y}}$$

A 100(1 – α)% PI for a future y value to be observed when $x = x^*$ is

$$\hat{\mu}_{Y \cdot x^*} \pm t_{\alpha/2,n-(k+1)} \cdot \left\{ s^2 + \left(\begin{matrix} \text{estimated SD} \\ \text{of } \hat{\mu}_{Y \cdot x^*} \end{matrix} \right)^2 \right\}^{1/2} = \hat{y} \pm t_{\alpha/2,n-(k+1)} \cdot \sqrt{s^2 + s_{\hat{Y}}^2}$$

Example 9

Example 8 continued…

Figure 13.11(b) shows that $\hat{\beta}_2 = -.00006001$ and $s_{\hat{\beta}_2} = .00001786$ (from the SE Coef column at the top of the output).

The null hypothesis H_0: $\beta_2 = 0$ says that as long as the linear predictor x is retained in the model, the quadratic predictor x^2 provides no additional useful information.

The relevant alternative is H_a: $\beta_2 \neq 0$, and the test statistic is $T = \hat{\beta}_2 / S_{\hat{\beta}_2}$, with computed value –3.36.

33

Example 9

cont'd

The test is based on $n - (k + 1) = 17$ df. At significance level .05, the null hypothesis is rejected because $-3.36 \leq -2.110 = -t_{.025,\,17}$. Inclusion of the quadratic predictor is justified.

The same conclusion results from comparing the reported P-value .004 to the chosen significance level .05.

34

Example 9

cont'd

The output in Figure 13.11(b) also contains estimation and prediction information both for $x = 500$ and for $x = 800$. In particular, for $x = 500$,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1(500) + \hat{\beta}_2(500)^2 = \text{Fit} = 21.136$$
$$s_{\hat{Y}} = \text{estimated SD of } \hat{Y} = \text{SE Fit} = 1.167$$

from which a 95% CI for mean strength when thickness = 500 is

$$21.136 \pm (2.110) \times (1.167) = (18.67,\ 23.60).$$

35

Example 9

A 95% PI for the strength resulting from a single bond when thickness = 500 is

$$21.136 \pm (2.110)[(3.27)^2 + (1.167)^2]^{1/2} = (13.81,\ 28.46).$$

As before, the PI is substantially wider than the CI because s is large compared to SE Fit.

36

Centering *x* Values

Centering *x* Values

For the quadratic model with regression function $\mu_{Y \cdot x} = \beta_0 + \beta_1 x + \beta_2 x^2$ the parameters β_0, β_1 and β_2 characterize the behavior of the function near $x = 0$.

For example, β_0 is the height at which the regression function crosses the vertical axis $x = 0$, whereas β_1 is the first derivative of the function at $x = 0$ (instantaneous rate of change of $\mu_{Y \cdot x}$ at $x = 0$). If the x_i's all lie far from 0, we may not have precise information about the values of these parameters.

Centering *x* Values

Let $\bar{x}$ = the average of the x_i's for which observations are to be taken, and consider the model

$$Y = \beta_0^* + \beta_1^*(x - \bar{x}) + \beta_2^*(x - \bar{x})^2 + \epsilon \qquad \textbf{(13.14)}$$

In the model (13.14),

$$\mu_{Y \cdot x} = \beta_0^* + \beta_1^*(x - \bar{x}) + \beta_2^*(x - \bar{x})^2$$

and the parameters now describe the behavior of the regression function near the center $\bar{x}$ of the data.

Centering *x* Values

To estimate the parameters of (13.14), we simply subtract $\bar{x}$ from each x_i to obtain $x'_i = x_i - \bar{x}$ and then use the x'_i's in place of the x_i's.

An important benefit of this is that the coefficients of $b_0, \ldots, b_k$ in the normal equations (13.10) will be of much smaller magnitude than would be the case were the original x_i's used. When the system is solved by computer, this centering protects against any round-off error that may result.

Example 10

The article "A Method for Improving the Accuracy of Polynomial Regression Analysis" (*J. of Quality Tech.*, 1971: 149–155) reports the following data on x = cure temperature (°F) and y = ultimate shear strength of a rubber compound (psi), with $\bar{x} = 297.13$:

x	280	284	292	295	298	305	308	315
x'	−17.13	−13.13	−5.13	−2.13	.87	7.87	10.87	17.87
y	770	800	840	810	735	640	590	560

Example 10

cont'd

A computer analysis yielded the results shown in Table 13.4.

Parameter	Estimate	Estimated SD	Parameter	Estimate	Estimated SD
β_0	−26,219.64	11,912.78	β_0^*	759.36	23.20
β_1	189.21	80.25	β_1^*	−7.61	1.43
β_2	−.3312	.1350	β_2^*	−.3312	.1350

Estimated Coefficients and Standard Deviations for Example 10

Table 13.4

Example 10

cont'd

The estimated regression function using the original model is $y = -26{,}219.64 + 189.21x - .3312x^2$, whereas for the centered model the function is
$y = 759.36 - 7.61(x - 297.13) - .3312(x - 297.13)^2$

These estimated functions are identical; the only difference is that different parameters have been estimated for the two models.

The estimated standard deviations indicate clearly that β_0^* and β_1^* have been more accurately estimated than β_0 and β_1.

Example 10

cont'd

The quadratic parameters are identical $(\beta_2 = \beta_2^*)$, as can be seen by comparing the x^2 term in (13.14) with the original model.

We emphasize again that a major benefit of centering is the gain in computational accuracy, not only in quadratic but also in higher-degree models.