J. 1997: 39–44) reported on a study carried out to relate yarn tenacity (y, in g/tex) to yarn count (x_1, in tex), percentage polyester (x_2), first nozzle pressure (x_3, in kg/cm^2), and second nozzle pressure (x_4, in kg/cm^2). The estimate of the constant term in the corresponding multiple regression equation was 6.121. The estimated coefficients for the four predictors were −.082, .113, .256, and −.219, respectively, and the coefficient of multiple determination was .946. Assume that $n = 25$.

a. State and test the appropriate hypotheses to decide whether the fitted model specifies a useful linear relationship between the response variable and at least one of the four model predictors.
b. Calculate the value of adjusted R^2 and comment.
c. Calculate a 99% confidence interval for true mean yarn tenacity when yarn count is 16.5, yarn contains 50% polyester, first nozzle pressure is 3, and second nozzle pressure is 5 if the estimated standard deviation of predicted tenacity under these circumstances is .350.

12.8 Quadratic, Interaction, and Indicator Terms

The fit of a multiple regression model can often be improved by creating new predictors from the original explanatory variables. In this section we discuss the two primary examples: *quadratic terms* and *interaction terms*. We also explain how to incorporate categorical predictor variables into the multiple regression model through the use of *indicator variables*.

Polynomial Regression

Let's return for a moment to the case of bivariate data consisting of n (x, y) pairs. Suppose that a scatterplot shows a parabolic rather than linear shape. Then it is natural to specify a **quadratic regression model**:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

The corresponding population regression function $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ gives the mean or expected value of Y for any particular x.

What does this have to do with multiple regression? Re-write the quadratic model equation as follows:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \quad \text{where } x_1 = x \text{ and } x_2 = x^2$$

Now this looks exactly like a multiple regression equation with two predictors. Although we interpret this model as a quadratic function of x, the multiple linear regression model (12.12) only requires that the response be a linear function of the β_j's and ε. Nothing precludes one predictor being a mathematical function of another one. So, from a modeling perspective, *quadratic regression is a special case of multiple regression*. Thus any software package capable of carrying out a multiple regression analysis can fit the quadratic regression model. The same is true of cubic regression and even higher-order polynomial models, although in practice very rarely are such higher-order predictors needed.

The coefficient β_1 on the linear predictor x_1 cannot be interpreted as the change in expected Y when x_1 increases by one unit while x_2 is held fixed. This is because it is impossible to increase x without also increasing x^2. A similar comment applies to β_2. More generally, the interpretation of regression coefficients requires extra care when some predictor variables are mathematical functions of others.

Example 12.25 Reconsider the solar cell data of Example 12.2. Figure 12.2 clearly shows a parabolic relationship between x = sheet resistance and y = cell efficiency. To calculate the "least squares parabola" for this data, software fits a multiple regression model with two predictors: $x_1 = x$ and $x_2 = x^2$. The first few rows of data for this scenario are as follows:

y	$x_1 = x$	$x_2 = x^2$
13.91	43.58	1898.78
13.50	50.94	2594.63
13.59	60.03	3603.60
13.86	66.82	4464.91
⋮	⋮	⋮

(In most software packages, it is not necessary to calculate x^2 for each observation; rather, the user can merely instruct the software to fit a quadratic model.) The coefficients that minimize the residual sum of squares are $\hat{\beta}_0 = 4.008$, $\hat{\beta}_1 = .3617$, and $\hat{\beta}_2 = -.003344$, so the estimated regression equation is

$$y = 4.008 + .3617x_1 - .003344x_2 = 4.008 + .3617x - .003344x^2$$

Figure 12.30 shows this parabola superimposed on a scatterplot of the original (x, y) data. Notice that the negative coefficient on x^2 matches the concave-downward contour of the data.

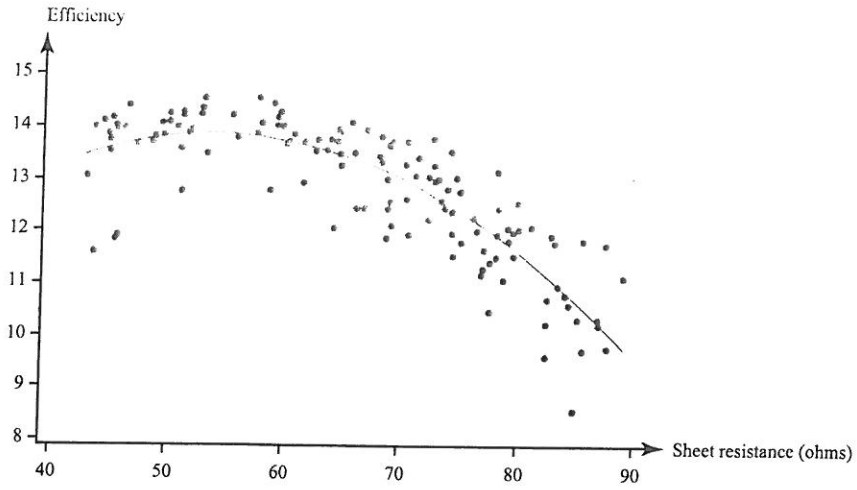

Figure 12.30 Scatterplot for Example 12.25 with a best-fit parabola

The estimated equation can now be used to make estimates and predictions at any particular x value. For example, the predicted efficiency at $x = 60$ ohms is determined by substituting $x_1 = x = 60$ and $x_2 = x^2 = 60^2 = 3600$:

$$y = 4.008 + .3617(60) - .003344(60)^2 = 13.67 \text{ percent}$$

Using software, a 95% CI for the mean efficiency of all 60-ohm solar panels is (13.50, 13.84), while a 95% PI for the efficiency of a single future 60-ohm panel is (12.32, 15.03). As always, the prediction interval is substantially wider than the confidence interval. ■

$$s_e = 33.124(\text{df} = 41), R^2 = 71.25\%, R_a^2 = 67.74\%$$

Variable utility tests indicate that all predictors except x_1 are useful; it's possible that x_1 is redundant with x_2. At the other extreme, a complete second-order model here involves 20 predictor variables: the original x_j's (five), their squares (another five), and all $\binom{5}{2} = 10$ possible interaction terms: x_1x_2, x_1x_3, ..., x_4x_5. Summary quantities from fitting this enormous model include $s_e = 24.878$ (df = 26), $R^2 = 89.72\%$, and $R_a^2 = 81.80\%$. The reduced standard deviation and greatly increased adjusted R^2 both suggest that at least some of the 15 second-order terms are useful, and so it was wise to incorporate these additional terms.

By considering the relative importance of the terms based on P-values, the researchers reduced their model to "just" 12 terms: all first-order terms, two of the quadratic terms, and five of the ten interactions. Based on the resulting estimated regression equation (shown in the article, but not here) researchers were able to determine the values of $x_1, \ldots, x_5$ that minimize residual Cd concentration. (Optimization is one of the added benefits of quadratic terms. For example, in Figure 12.30, we can see there is an x value for which solar cell efficiency is maximized. A linear model has no such local maxima or minima.)

It's worth noting that while x_1 was not considered useful in the first-order model, several second-order terms involving x_1 *were* significant. When fitting second-order models, it is recommended to fit the complete model first and delete useless terms rather than building up from the simpler first-order model; using the latter approach, important quadratic and interaction effects can be missed. ■

One issue that arises with fitting a model with an abundance of terms, as in Example 12.26, is the potential to commit many type I errors when performing variable utility t tests on every predictor. Exercise 94 presents a method called the *partial F test* for determining whether a group of predictors can all be deleted while controlling the overall type I error rate.

Models with Categorical Predictors

Thus far we have explicitly considered the inclusion of only quantitative (numerical) predictor variables in a multiple regression model. Using simple numerical coding, categorical variables such as sex, type of college (private or state), or type of wood (pine, oak, or walnut) can also be incorporated into a model. Let's first focus on the case of a dichotomous variable, one with just two possible categories—alive or dead, US or foreign manufacture, and so on. With any such variable, we associate an **indicator** (or **dummy**) **variable** whose possible values 0 and 1 indicate which category is relevant for any particular observation.

Example 12.27 Is it possible to predict graduation rates from freshman test scores? Based on the median SAT score of entering freshmen at a university, can we predict the percentage of those freshmen who will get a degree there within six years? To investigate, let y = six-year graduation rate, x_2 = median freshman SAT score, and x_1 = a variable defined to indicate private or public status:

$$x_1 = \begin{cases} 1 & \text{if the university is private} \\ 0 & \text{if the university is public} \end{cases}$$

The corresponding multiple regression model is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

The mean graduation rate depends on whether the university is public or private:

$$\begin{aligned} &\text{mean graduation rate} = \beta_0 + \beta_2 x_2 && \text{when } x_1 = 0 \text{ (public)} \\ &\text{mean graduation rate} = \beta_0 + \beta_1 + \beta_2 x_2 && \text{when } x_1 = 1 \text{ (private)} \end{aligned}$$

Thus there are two parallel lines with vertical separation β_1, as shown in Figure 12.32a. The coefficient β_1 is the *difference* in mean graduation rates between private and public universities, after adjusting for median SAT score. If $\beta_1 > 0$ then, on average, for a given SAT, private universities will have a higher graduation rate.

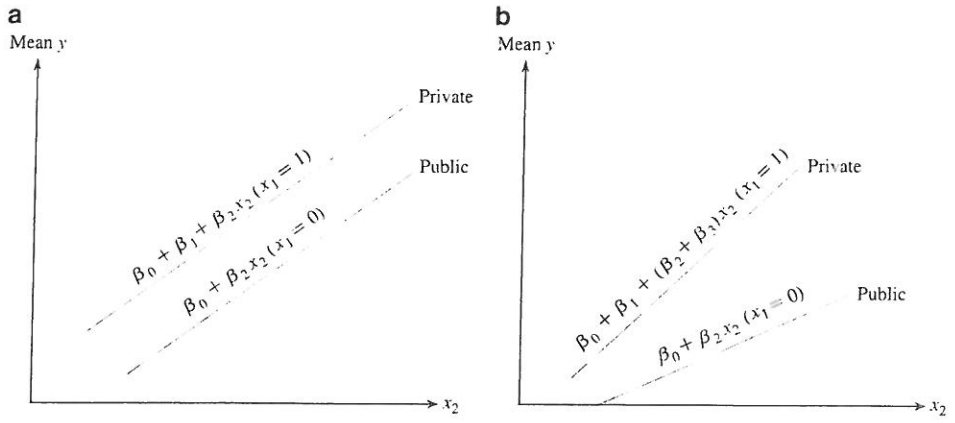

Figure 12.32 Regression functions for models with one indicator variable (x_1) and one quantitative variable (x_2): **(a)** no interaction; **(b)** interaction

A second possibility is a model with an interaction term:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

Now the mean graduation rates for the two types of university are

$$\begin{aligned} &\text{mean graduation rate} = \beta_0 + \beta_2 x_2 && \text{when } x_1 = 0 \text{ (public)} \\ &\text{mean graduation rate} = \beta_0 + \beta_1 + (\beta_2 + \beta_3) x_2 && \text{when } x_1 = 1 \text{ (private)} \end{aligned}$$

Here we have two lines, where β_1 is the difference in intercepts and β_3 is the difference in slopes, as shown in Figure 12.32b. Unless $\beta_3 = 0$, the lines will not be parallel and there will be interaction effect, meaning that the separation between public and private graduation rates depends on SAT.

To make inferences, we obtained a random sample of 20 Master's level universities from the 2017 data file available on www.collegeresults.org.

University	Grad rate	Median SAT	Sector
Appalachian State University	73.4	1140	Public
Brenau University	49.4	973	Private
Campbellsville University	34.1	1008	Private
Delta State University	39.6	1028	Public
DeSales University	70.1	1072	Private
Lasell College	54.1	966	Private
Marshall University	49.3	1010	Public
Medaille College	43.9	906	Private
Mount Saint Joseph University	60.7	1011	Private
Mount Saint Mary College	53.8	991	Private
Muskingum University	48.2	1009	Private
Pacific University	64.4	1122	Private
Simpson University	56.7	985	Private
SUNY Oneonta	70.9	1082	Public
Texas A&M University-Texarkana	29.7	1016	Public
Truman State University	74.9	1224	Public
University of Redlands	77.0	1101	Private
University of Southern Indiana	39.6	1005	Public
University of Tennessee-Chattanooga	45.2	1088	Public
Western State Colorado University	41.0	1026	Public

First of all, does the interaction predictor provide useful information over and above what is contained in x_1 and x_2? To answer this question, we should test the hypothesis H_0: $\beta_3 = 0$ versus H_a: $\beta_3 \neq 0$ first. If H_0 is not rejected (meaning interaction is not informative) then we can use the parallel lines model to see if there is a separation (β_1) between lines. Of course, it does not make sense to estimate the difference between lines if the difference depends on x_2, which is the case when there is interaction.

Figure 12.33 shows R output for these two tests. The coefficient for interaction has a P-value of roughly .42, so there is no reason to reject the null hypothesis H_0: $\beta_3 = 0$. Since we fail to reject the "no-interaction" hypothesis, we drop the interaction term and re-run the analysis. The estimated regression equation specified by R is

$$y = -124.56039 + 13.33553x_1 + 0.16474x_2$$

The t ratio values and P-values indicate that both H_0: $\beta_1 = 0$ and H_0: $\beta_2 = 0$ should be rejected at the .05 significance level. The coefficient on x_1 indicates that a private university is estimated to have a graduation rate about 13 percentage points higher than a state university with the same median SAT.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -152.37773   49.18039  -3.098 0.006904 **
SectorPrivate             70.06920   69.28401   1.011 0.326908
Median.SAT                 0.19077    0.04592   4.154 0.000746 ***
SectorPrivate:Median.SAT  -0.05457    0.06649  -0.821 0.423857
```

Testing without interaction

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -124.56039   35.29279  -3.529 0.002575 **
SectorPrivate  13.33553    4.64033   2.874 0.010530 *
Median.SAT      0.16474    0.03289   5.009 0.000108 ***
```

Figure 12.33 R output for an interaction model and a "parallel lines" model