5.37 Using glass to encapsulate waste. Since glass is not subject to radiation damage, encapsulation of waste in glass is considered to be one of the most promising solutions to the problem of low-level nuclear waste in the environment. However, glass undergoes chemical changes when exposed to extreme environmental conditions, and certain of its constituents can leach into the surroundings. In addition, these chemical reactions may weaken the glass. These concerns led to a study undertaken jointly by the Department of Materials Science and Engineering at the University of Florida and the U.S. Department of Energy to assess the utility of glass as a waste encapsulant material.† Corrosive chemical solutions (called corrosion baths) were prepared and applied directly to glass samples containing one of three types of waste (TDS-3A, FE, and AL); the chemical reactions were observed over time. A few of the key variables measured were

y = Amount of silicon (in parts per million) found in solution at end of experiment. (This is both a measure of the degree of breakdown in the glass and a proxy for the amount of radioactive species released) into the environment.

x_1 = Temperature (°C) of the corrosion bath

$$x_2 = \begin{cases} 1 & \text{if waste type TDS-3A} \\ 0 & \text{if not} \end{cases} \qquad x_3 = \begin{cases} 1 & \text{if waste type FE} \\ 0 & \text{if not} \end{cases}$$

Waste type AL is the base level. Suppose we want to model amount y of silicon as a function of temperature (x_1) and type of waste (x_2, x_3).

(a) Write a model that proposes parallel straight-line relationships between amount of silicon and temperature, one line for each of the three waste types.
(b) Add terms for the interaction between temperature and waste type to the model of part a.
(c) Refer to the model of part b. For each waste type, give the slope of the line relating amount of silicon to temperature.
(d) Explain how you could test for the presence of temperature–waste type interaction.

5.11 External Model Validation (Optional)

Regression analysis is one of the most widely used statistical tools for estimation and prediction. All too frequently, however, a regression model deemed to be an adequate predictor of some response y performs poorly when applied in practice. For example, a model developed for forecasting new housing starts, although found to be statistically useful based on a test for overall model adequacy, may fail to take into account any extreme changes in future home mortgage rates generated by new government policy. This points out an important problem. *Models that fit the sample data well may not be successful predictors of y when applied to new data*. For this reason, it is important to assess the **validity** of the regression model in addition to its **adequacy** before using it in practice.

In Chapter 4, we presented several techniques for checking *model adequacy* (e.g., tests of overall model adequacy, partial F-tests, R_a^2, and s). In short, checking model adequacy involves determining whether the regression model adequately fits the *sample data*. **Model validation**, however, involves an assessment of how the fitted regression model will perform in practice—that is, how successful it will be when applied to new or future data. A number of different model validation techniques have been proposed, several of which are briefly discussed in this section. You will need to consult the references for more details on how to apply these techniques.

1. *Examining the predicted values*: Sometimes, the predicted values $\hat{y}$ of the fitted regression model can help to identify an invalid model. Nonsensical or unreasonable predicted values may indicate that the form of the model is incorrect or that the β coefficients are poorly estimated. For example, a model for a binary response y, where y is 0 or 1, may yield predicted probabilities that

† The background information for this exercise was provided by Dr. David Clark, Department of Materials Science and Engineering, University of Florida.

are negative or greater than 1. In this case, the user may want to consider a model that produces predicted values between 0 and 1 in practice (One such model, called the *logistic regression model*, is covered in Chapter 9.) On the other hand, if the predicted values of the fitted model all seem reasonable, the user should refrain from using the model in practice until further checks of model validity are carried out.

2. *Examining the estimated model parameters*: Typically, the user of a regression model has some knowledge of the relative size and sign (positive or negative) of the model parameters. This information should be used as a check on the estimated β coefficients. Coefficients with signs opposite to what is expected or with unusually small or large values or unstable coefficients (i.e., coefficients with large standard errors) forewarn that the final model may perform poorly when applied to new or different data.
3. *Collecting new data for prediction*: One of the most effective ways of validating a regression model is to use the model to predict y for a new sample. By directly comparing the predicted values to the observed values of the new data, we can determine the accuracy of the predictions and use this information to assess how well the model performs in practice.

   Several measures of model validity have been proposed for this purpose. One simple technique is to calculate the percentage of variability in the new data explained by the model, denoted $R^2_{\text{prediction}}$, and compare it to the coefficient of determination R^2 for the least squares fit of the final model. Let $y_1, y_2, \ldots, y_n$ represent the n observations used to build and fit the final regression model and $y_{n+1}, y_{n+2}, \ldots, y_{n+m}$ represent the m observations in the new data set. Then

$$R^2_{\text{prediction}} = 1 - \left\{ \frac{\sum_{i=n+1}^{n+m} (y_i - \hat{y}_i)^2}{\sum_{i=n+1}^{n+m} (y_i - \bar{y})^2} \right\}$$

   where $\hat{y}_i$ is the predicted value for the ith observation using the β estimates from the fitted model and $\bar{y}$ is the sample mean of the original data.* If $R^2_{\text{prediction}}$ compares favorably to R^2 from the least squares fit, we will have increased confidence in the usefulness of the model. However, if a significant drop in R^2 is observed, we should be cautious about using the model for prediction in practice.

   A similar type of comparison can be made between the mean square error, MSE, for the least squares fit and the mean squared prediction error

$$\text{MSE}_{\text{prediction}} = \frac{\sum_{i=n+1}^{n+m} (y_i - \hat{y}_i)^2}{m - (k+1)}$$

   where k is the number of β coefficients (excluding β_0) in the model. Whichever measure of model validity you decide to use, the number of observations in the new data set should be large enough to reliably assess the model's prediction performance. Montgomery, Peck, and Vining (2006), for example, recommend 15–20 new observations, *at minimum*.

* Alternatively, the sample mean of the new data may be used.

4. *Data-splitting (cross-validation)*: For those applications where it is impossible or impractical to collect new data, the original sample data can be split into two parts, with one part used to estimate the model parameters and the other part used to assess the fitted model's predictive ability. **Data-splitting** (or **cross-validation**, as it is sometimes known) can be accomplished in a variety of ways. A common technique is to randomly assign half the observations to the estimation data set and the other half to the prediction data set.† Measures of model validity, such as $R^2_{\text{prediction}}$ or $\text{MSE}_{\text{prediction}}$, can then be calculated. Of course, a sufficient number of observations must be available for data-splitting to be effective. For the estimation and prediction data sets of equal size, it has been recommended that the entire sample consist of *at least* $n = 2k + 25$ observations, where k is the number of β parameters in the model [see Snee (1977)].
5. *Jackknifing*: In situations where the sample data set is too small to apply data-splitting, a method called the **jackknife** can be applied. Let $y_{(i)}$ denote the predicted value for the ith observation obtained when the regression model is fit with the data point for y_i omitted (or deleted) from the sample. The jackknife method involves leaving each observation out of the data set, one at a time, and calculating the difference, $y_i - \hat{y}_{(i)}$, for all n observations in the data set. Measures of model validity, such as R^2 and MSE, are then calculated:

$$R^2_{\text{jackknife}} = 1 - \frac{\Sigma(y_i - \hat{y}_{(i)})^2}{\Sigma(y_i - \bar{y})^2}$$

$$\text{MSE}_{\text{jackknife}} = \frac{\Sigma(y_i - \hat{y}_{(i)})^2}{n - (k+1)}$$

The numerator of both $R^2_{\text{jackknife}}$ and $\text{MSE}_{\text{jackknife}}$ is called the **prediction sum of squares**, or **PRESS**. In general, PRESS will be larger than the SSE of the fitted model. Consequently, R^2_{jacknife} will be smaller than the R^2 of the fitted model and $\text{MSE}_{\text{jackknife}}$ will be larger than the MSE of the fitted model. These jackknife measures, then, give a more conservative (and more realistic) assessment of the ability of the model to predict future observations than the usual measures of model adequacy.

Example 5.18

In Chapter 4 (Example 4.10), we presented a model for executive salary (y) developed by Towers, Perrin, Forster & Crosby, an international management consulting firm. The multiplicative model fit was

$$E\{\ln(y)\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_1^2 + \beta_7 x_3 x_4$$

where $x_1 =$ years of experience, $x_2 =$ years of education, $x_3 = \{1$ if male, 0 if female$\}$, $x_4 =$ number of employees supervised, and $x_5 =$ corporate assets. Since the consulting firm intends to use the model in evaluating executive salaries at a variety of companies, it is important to validate the model externally. Apply one of the model validation techniques discussed in this section to the data for the $n = 100$ executives saved in the EXECSAL file.

EXECSAL

Solution

With $n = 100$ observations, a data-splitting method could be employed to validate the model. For example, 80 observations (randomly selected from the sample) could

† Random splits are usually applied in cases where there is no logical basis for dividing the data. Consult the references for other, more formal data-splitting techniques.

be used to estimate the prediction equation, and the remaining 20 observations used to validate the results. Ideally, we would like to have more observations (say, 50) in the validation subset. However, this would dramatically reduce the sample size for the estimation subset and possibly lead to less reliable results.

As an alternative, we use the jackknifing (one-observation-out-at-a-time) approach to model validation. Most statistical software packages have routines that automatically perform the jackknifing and produce the PRESS statistic. Figure 5.31 is a MINITAB printout for the multiplicative model fit to the data in the EXECSAL file. The value of PRESS (highlighted) is .482665. We compare this jackknifed value of SSE to the total sum of squares (also highlighted on the printout) value of 6.68240 as follows:

$$R^2_{\text{jackknife}} = 1 - (\text{PRESS}/\text{SSTotal}) = 1 - (.482665/6.68240) = .92777$$

(Note that $R^2_{\text{jackknife}}$ is also highlighted in Figure 5.31.) Since $R^2_{\text{jackknife}}$ is only slightly smaller than the R^2 of the original model fit (.94), the consulting firm has increased confidence in using the model for evaluation of executive salaries. ■

Figure 5.31 MINITAB printout for the multiplicative model of executive salary

Regression Analysis: LNSAL versus EXP, EDUC, ...

```
The regression equation is
LNSAL = 9.86 + 0.0436 EXP + 0.0309 EDUC + 0.117 GENDER + 0.000326 NUMSUP
        + 0.00239 ASSETS - 0.000635 EXPSQ + 0.000302 GEN_SUP


Predictor        Coef     SE Coef       T      P
Constant      9.86182     0.09703  101.64  0.000
EXP          0.043643    0.003761   11.60  0.000
EDUC         0.030936    0.002950   10.49  0.000
GENDER        0.11661     0.03696    3.16  0.002
NUMSUP     0.00032594  0.00007850    4.15  0.000
ASSETS      0.0023911   0.0004439    5.39  0.000
EXPSQ      -0.0006348   0.0001383   -4.59  0.000
GEN_SUP    0.00030196  0.00009238    3.27  0.002


S = 0.0659583   R-Sq = 94.0%   R-Sq(adj) = 93.6%

PRESS = 0.482665   R-Sq(pred) = 92.78%


Analysis of Variance

Source           DF      SS       MS       F      P
Regression        7  6.28215  0.89745  206.29  0.000
Residual Error   92  0.40025  0.00435
Total            99  6.68240
```

The appropriate model validation technique(s) you employ will vary from application to application. Keep in mind that a favorable result is still no guarantee that the model will always perform successfully in practice. However, we have much greater confidence in a validated model than in one that simply fits the sample data well.

Quick Summary/Guides

Key Formulas

Coding Quantitative x's

$u = (x - \bar{x})/s_x$, where $\bar{x}$ and s are the mean and standard deviation of x

Cross-validation

$$R^2_{\text{prediction}} = 1 - \sum_{i=n+1}^{n+m} (y_i - \hat{y}_i)^2 \Big/ \sum_{i=n+1}^{n+m} (y_i - \bar{y})^2$$

$$\text{MSE}_{\text{prediction}} = \sum_{i=n+1}^{n+m} (y_i - \hat{y}_i)^2 \Big/ [m - (k+1)]$$

$$R^2_{\text{jackknife}} = 1 - \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2 \Big/ \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$\text{MSE}_{\text{jackknife}} = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2 \Big/ [n - (k+1)]$$

Key Ideas

Steps in Model Building

1. Identify the response (*dependent*) variable y
2. Classify each potential predictor *(independent)* variable as *quantitative* or *qualitative*
3. Define *dummy variables* to represent the qualitative independent variables
4. Consider *higher-order* terms (e.g., x^2, x^3) for quantitative variables
5. Possibly *code the quantitative variables* in higher-order polynomials
6. Consider *interaction* terms for both quantitative and qualitative independent variables
7. Compare *nested* models using *partial F-tests* to arrive at a final model
8. Consider validation of the final model using data-splitting or jackknifing

Procedure for Writing a Complete Second-order Model

1. Enter terms for all *quantitative* x's, including interactions and second-order terms
2. Enter terms for all *qualitative* x's, including main effects, two-way, three-way, . . . , and k-way interactions
3. Enter terms for all possible *quantitative by qualitative interactions,* that is interact all terms in Step 1 with all terms in Step 2

Models with One Quantitative x

First-order: $E(y) = \beta_0 + \beta_1 x$
Second-order: $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$
pth-order: $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p$

Models with Two Quantitative x's

First-order: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
Second-order, interaction only: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$
Complete second-order: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$

Models with Three Quantitative x's

First-order: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
Second-order, interaction only: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3$
Complete second-order: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \beta_7 x_1^2 + \beta_8 x_2^2 + \beta_9 x_3^2$

Model with One Qualitative x (k levels)

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_{k-1} x_k$,

where $x_1 = \{1$ if level 1, 0 if not$\}$, $x_2 = \{1$ if level 2, 0 if not$\}$, $\ldots x_{k-1} = \{1$ if level $k-1$, 0 if not$\}$

Models with Two Qualitative x's (one at two levels, one at three levels)

Main effects: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$,

where x_1 represents the dummy variable for the qualitative variable at two levels;

x_2 and x_3 represent the dummy variables for the qualitative variable at three levels

Interaction: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3$

Models with One Quantitative x and One Qualitative x (at three levels)

First-order, no interaction: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$,

where x_2 and x_3 represent the dummy variables for the qualitative variable at three levels