12.7 Model Selection Procedures

Background

In a multiple linear regression model, several independent variables are used to describe the variation in a dependent variable. Frequently, there are many independent variable candidates for inclusion in a model. As we have seen in previous examples and exercises, if we reject the null hypothesis H_0: $\beta_i = 0$, then there is evidence to suggest that x_i should be included in the model. However, if there is no evidence to suggest that the regression coefficient is different from 0, the multiple regression model may be better without the corresponding independent variable candidate.

Consider a multiple regression model with a dependent variable of ocean water temperature (at a certain location). This value may be affected by (the independent variables) speed of the surrounding current, air temperature, salinity, humidity, air pressure, and cloud cover. Some of these independent variables might be highly correlated, or *move together*. Therefore, the final model may need to include only one of these variables. Still other variables may simply have no effect on water temperature.

This section presents methods for selecting the best multiple regression model. Given a collection of independent variables, we need to decide which ones contribute the most to the variation in y. Using the list of independent variable candidates, the goal is to construct (build) the best multiple regression model.

All Possible Subsets

Consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + E_i$$

in which there are k independent variable candidates for inclusion in the final model. One technique for building a model is to consider all possible subsets (or collections) of the independent variable candidates. We could then use some reasonable measure of model goodness and, based on this value, select the best model.

Using this all-possible-subsets approach, we need to consider all models with one independent variable, all models with two independent variables, and so on, and finally the model with all k independent variables. If we assume that β_0 is included in the final model, then there are 2^k possible models to consider. For example, if there are $k = 5$ independent variable candidates, then there are $2^5 = 32$ possible subset models.

We assume that the number of observations n is greater than k, and it is advantageous to have n a lot bigger than k.

One reasonable measure of model goodness is r^2, the coefficient of determination, a measure of the proportion of the variation in the data that is explained by the regression model. If model 1 has a larger r^2 than model 2, then model 1 can be used to explain more of the variation in the dependent variable. Recall, however, that adding another independent variable to a model will always increase r^2. Therefore, we want to consider significant increases in the value of r^2 when an additional independent variable is added to the model. What constitutes a significant increase is usually a subjective judgment.

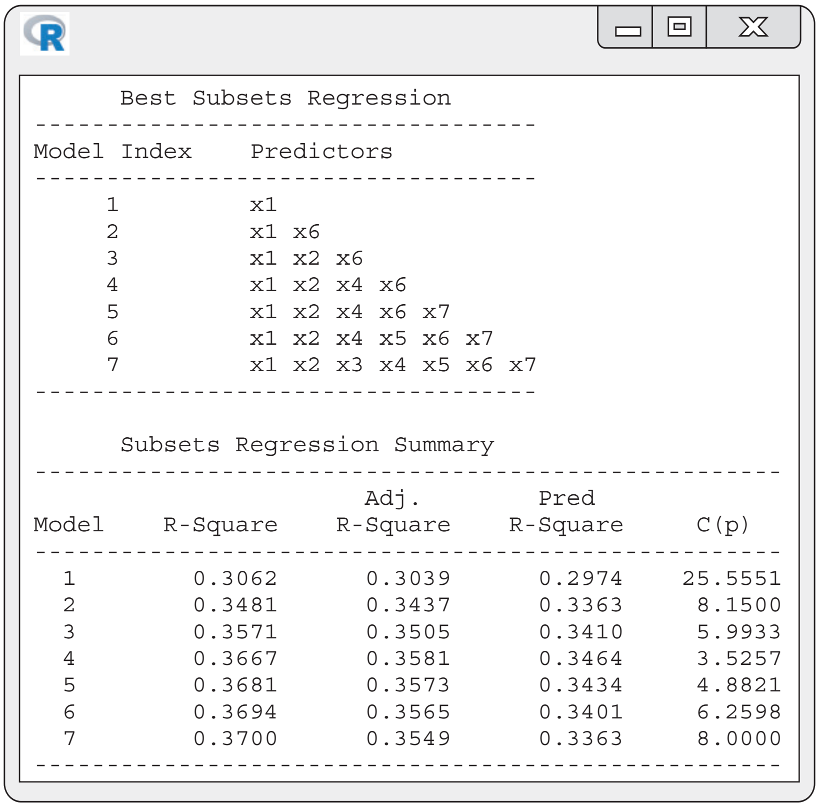
EXAMPLE 12.18 Exposure to Manganese

Manganese is an essential trace element. Found mostly in bones, the liver, kidneys, and the pancreas, manganese helps our bodies form connective tissue and bones. However, excessive exposure to manganese can be toxic, especially if it is inhaled. A study involving Mexican children living near mines suggested that intellectual function is inversely related to airborne manganese exposure. The goal was to build a multiple regression model to predict the total IQ of a child based on other independent observations.

Suppose a random sample of children living in Mexican mining towns was obtained, and the following variables were measured for each child:

Variable	Description
y	Total IQ level
x_1	Age of the child, in years
x_2	Concentration of manganese in the blood, in μg/L
x_3	Fuel used for cooking: 0, wood; 1, gas
x_4	Father a miner: 0, no; 1, yes
x_5	Mother's education in years
x_6	Number of miles from home to the nearest mine
x_7	Gender: 0, male; 1 female

The response variable is y and the predictor variables are x_1 through x_7. The results in **Figure 12.66** were obtained using R. Use the values of r^2 to select the best multiple linear regression model from all of the possible subsets of independent variables.



```
      Best Subsets Regression
-----------------------------------
Model Index    Predictors
-----------------------------------
     1         x1
     2         x1 x6
     3         x1 x2 x6
     4         x1 x2 x4 x6
     5         x1 x2 x4 x6 x7
     6         x1 x2 x4 x5 x6 x7
     7         x1 x2 x3 x4 x5 x6 x7
-----------------------------------

      Subsets Regression Summary
----------------------------------------------------
                        Adj.        Pred
Model    R-Square    R-Square    R-Square      C(p)
----------------------------------------------------
  1        0.3062      0.3039      0.2974   25.5551
  2        0.3481      0.3437      0.3363    8.1500
  3        0.3571      0.3505      0.3410    5.9933
  4        0.3667      0.3581      0.3464    3.5257
  5        0.3681      0.3573      0.3434    4.8821
  6        0.3694      0.3565      0.3401    6.2598
  7        0.3700      0.3549      0.3363    8.0000
----------------------------------------------------
```

Figure 12.66 The R function `ols_step_best_subset()` can be used to conduct a search for the best subsets of the predictor variables. This analysis produces various measures of model utility, including the value of r^2 for each model.

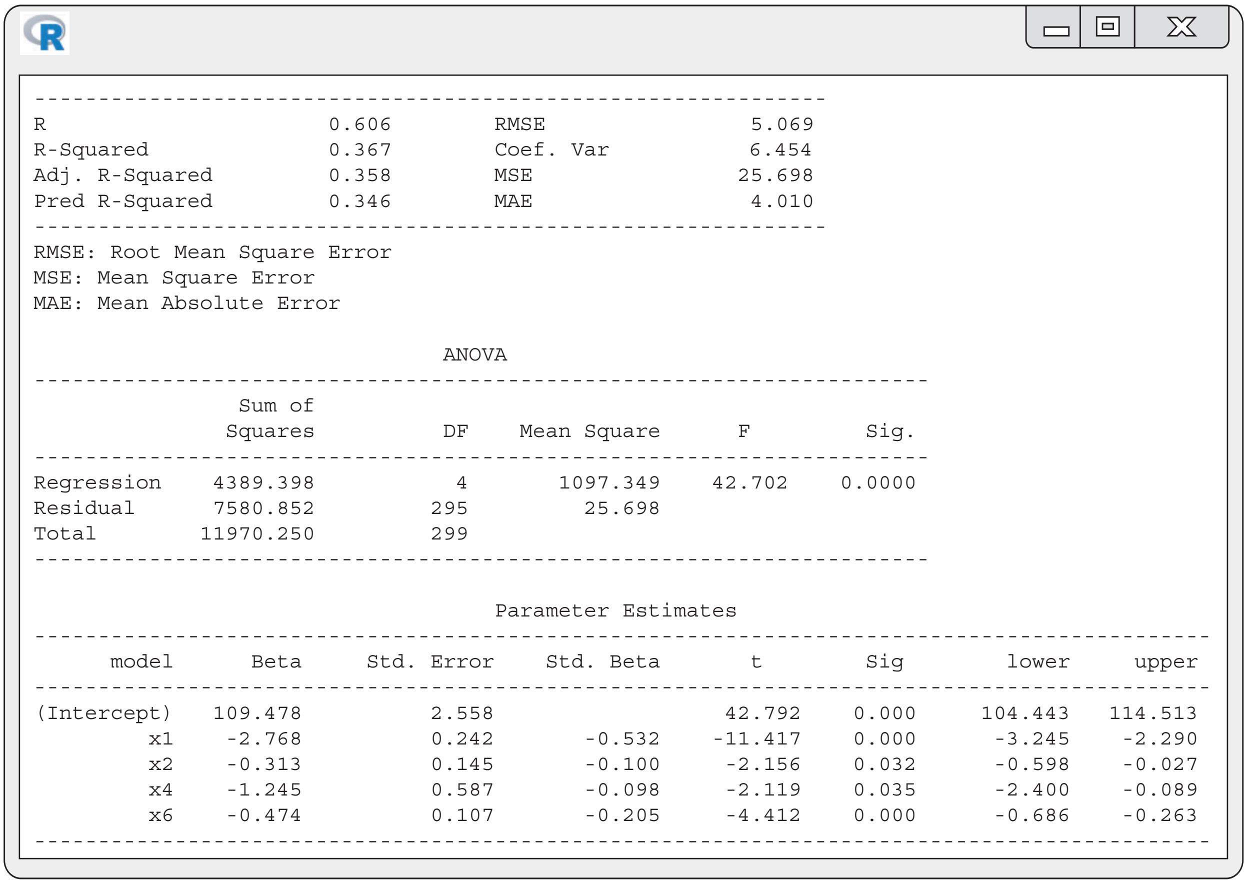
Solution

STEP 1 All possible subsets of the independent variables are considered when using technology. First, all models with a single independent variable are considered, then all models with two independent variables, and so on, until the final model with all independent variables included. The calculations are all completed in the background, and r^2 can be used to determine which models are the best. In Figure 12.66, the best models for each value of k (the number of predictor variables) are displayed, but more models for each value of k can be considered if desired.

STEP 2 Using the r^2 criterion to select a model, the model with four independent variables seems to be the best. In this model, $r^2 = 0.3667$, and this value does not increase significantly with the addition of any other independent variables. The estimated regression equation is

$$y = 109.478 - 2.768x_1 - 0.313x_2 - 1.245x_4 - 0.474x_6$$

The R output is shown in **Figure 12.67**.



```
--------------------------------------------------------------
R                       0.606       RMSE                 5.069
R-Squared               0.367       Coef. Var            6.454
Adj. R-Squared          0.358       MSE                 25.698
Pred R-Squared          0.346       MAE                  4.010
--------------------------------------------------------------
RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

                                ANOVA
----------------------------------------------------------------------
               Sum of
              Squares         DF    Mean Square      F          Sig.
----------------------------------------------------------------------
Regression    4389.398         4       1097.349    42.702     0.0000
Residual      7580.852       295         25.698
Total        11970.250       299
----------------------------------------------------------------------

                                    Parameter Estimates
--------------------------------------------------------------------------------------------
      model        Beta      Std. Error     Std. Beta        t          Sig        lower      upper
--------------------------------------------------------------------------------------------
(Intercept)     109.478          2.558                   42.792      0.000     104.443    114.513
         x1      -2.768          0.242        -0.532    -11.417      0.000      -3.245     -2.290
         x2      -0.313          0.145        -0.100     -2.156      0.032      -0.598     -0.027
         x4      -1.245          0.587        -0.098     -2.119      0.035      -2.400     -0.089
         x6      -0.474          0.107        -0.205     -4.412      0.000      -0.686     -0.263
--------------------------------------------------------------------------------------------
```

Figure 12.67 The ANOVA table and estimated regression coefficients.

The most important variables in predicting a child's IQ are age of the child, concentration of manganese in the blood, whether the child's father is a miner, and the number of miles from home to the nearest mine.

STEP 3 A graph of r^2 versus k for various models may be helpful in selecting the best regression model. The point at which the graph tends to level off suggests the best model. See **Figure 12.68**. A graph of Mallows C_p, defined below, versus k may also help. A value of C_p near k suggests the best model. See **Figure 12.69**.

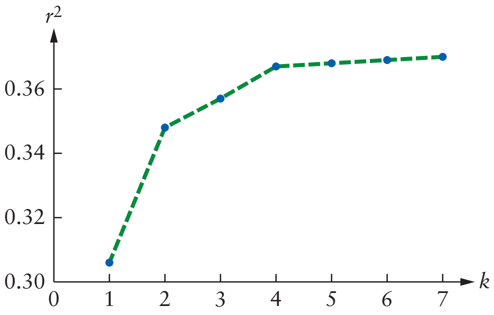


Figure 12.68 Graph of r^2 versus k for various models.

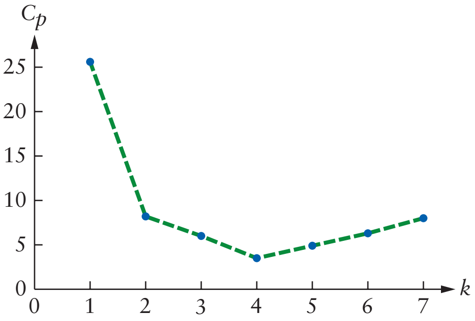


Figure 12.69 Graph of C_p versus k for various models.

In addition to r^2, other statistics can be used to evaluate each subset of predictor variables.

R-Sq(adj): Adjusted r^2. This modification of r^2 takes into account the number of independent variables in the model. The adjusted r^2 can actually be negative, it will always be less than or equal to r^2, and it does not have the same interpretation as r^2.

R-Sq(pred): Predicted r^2 measures how well a model predicts new responses. A regression model could fit the given data very well but predict the value of new responses inadequately. The predicted r^2 can be negative, and a value much less than r^2 suggests that there are too many terms in the model.

Mallows C_p: This measure involves the mean square due to error (MSE). Values of C_p that indicate a good regression model include small values around k, the number of independent variables in the model.

s: The estimate of the standard deviation of the error terms. Small values of s are desirable, indicating that the observations do not vary much from the estimated regression line. ■

TRY IT NOW **Go to Exercise 12.231**

Forward Selection

Although the model selection procedure just described is reasonable, it has several disadvantages. Selecting the best model is very subjective. There is no clear, concise method for deciding whether an increase in r^2 is really significant. In addition, if the number of possible independent variables is large—for example, 30—it would take lots of computer power and time to check all possible models.

Several very prescriptive algorithms have been developed that produce a best multiple regression model. They may or may not all produce the same final model. One widely used procedure is **forward selection**. In this technique, the single most significant independent variable, based on the t tests for significant regression coefficients, is added to the model at each step. **Backward elimination** is a similar procedure.

The next example demonstrates the method of forward selection.

EXAMPLE 12.19 Forward Selection

Suppose there are five possible independent variables in a multiple linear regression model: x_1, x_2, x_3, x_4, x_5.

Solution

Recall: In a simple linear regression model, the t test of H_0: $\beta_1 = 0$ is equivalent to the F test for an overall significant regression.

STEP 1 Consider all five simple linear regression models: $Y_i = \beta_0 + \beta_1 x_{ki} + E_i$ for $k = 1, 2, 3, 4, 5$. In each case, conduct the hypothesis test to determine whether the predictor variable is significant. That is, consider the test of H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$. The following table shows the t statistic and the p value for each of these tests.

Model Variable	t Statistic	p Value
x_1	3.34	0.0016
x_2	3.47	0.0011
x_3	−0.18	0.8579
x_4	−3.90	0.0003
x_5	−1.20	0.2359

In some cases, no predictor variable produces a significant regression. Also, the traditional p value cutoff value of 0.05 is often relaxed to 0.10 in forward selection.

Add the most significant predictor variable to the model, the independent variable that produces the smallest p value (below some threshold value, for example, ≤ 0.05). In this example, add the variable x_4 to the model.

STEP 2 Consider the four regression models with x_4 (already in the model) to determine whether any other predictor helps explain the variation in the dependent variable: $Y_i = \beta_0 + \beta_1 x_{4i} + \beta_2 x_{ki} + E_i$ for $k = 1, 2, 3, 5$. Conduct the hypothesis test to determine if the additional predictor variable is significant (H_0: $\beta_2 = 0$ versus H_a: $\beta_2 \neq 0$). The table shows the t statistic and p value for each of these tests.

Added Variable	*t* Statistic	*p* Value
x_1	1.64	0.1075
x_2	4.24	0.0001
x_3	−0.38	0.7056
x_5	−0.67	0.5061

Add the variable with the smallest p value (≤ 0.05). Therefore, add x_2 to the model.

STEP 3 Continue in this manner until no additional variable is significant.
Consider the three regression models with x_4 and x_2 (already in the model) to determine if any other predictor helps to explain the variation in the dependent variable: $Y_i = \beta_0 + \beta_1 x_{4i} + \beta_2 x_{2i} + \beta_3 x_{ki} + E_i$ for $k = 1, 3, 5$. Conduct the hypothesis test to determine whether the additional predictor variable is significant (H_0: $\beta_3 = 0$ versus H_a: $\beta_3 \neq 0$). The table shows the t statistic and p value for each of these tests.

Added Variable	*t* Statistic	*p* Value
x_1	1.41	0.1651
x_3	−1.22	0.2286
x_5	0.05	0.9603

All p values are greater than 0.05. No additional variable is added to the model. The forward selection procedure yields the final multiple regression model: $Y_i = \beta_0 + \beta_1 x_{4i} + \beta_2 x_{2i} + E_i$

EXAMPLE 12.20 Exposure to Manganese (Continued)

Use the forward selection procedure to find the best multiple linear regression model to predict IQ level.

Solution

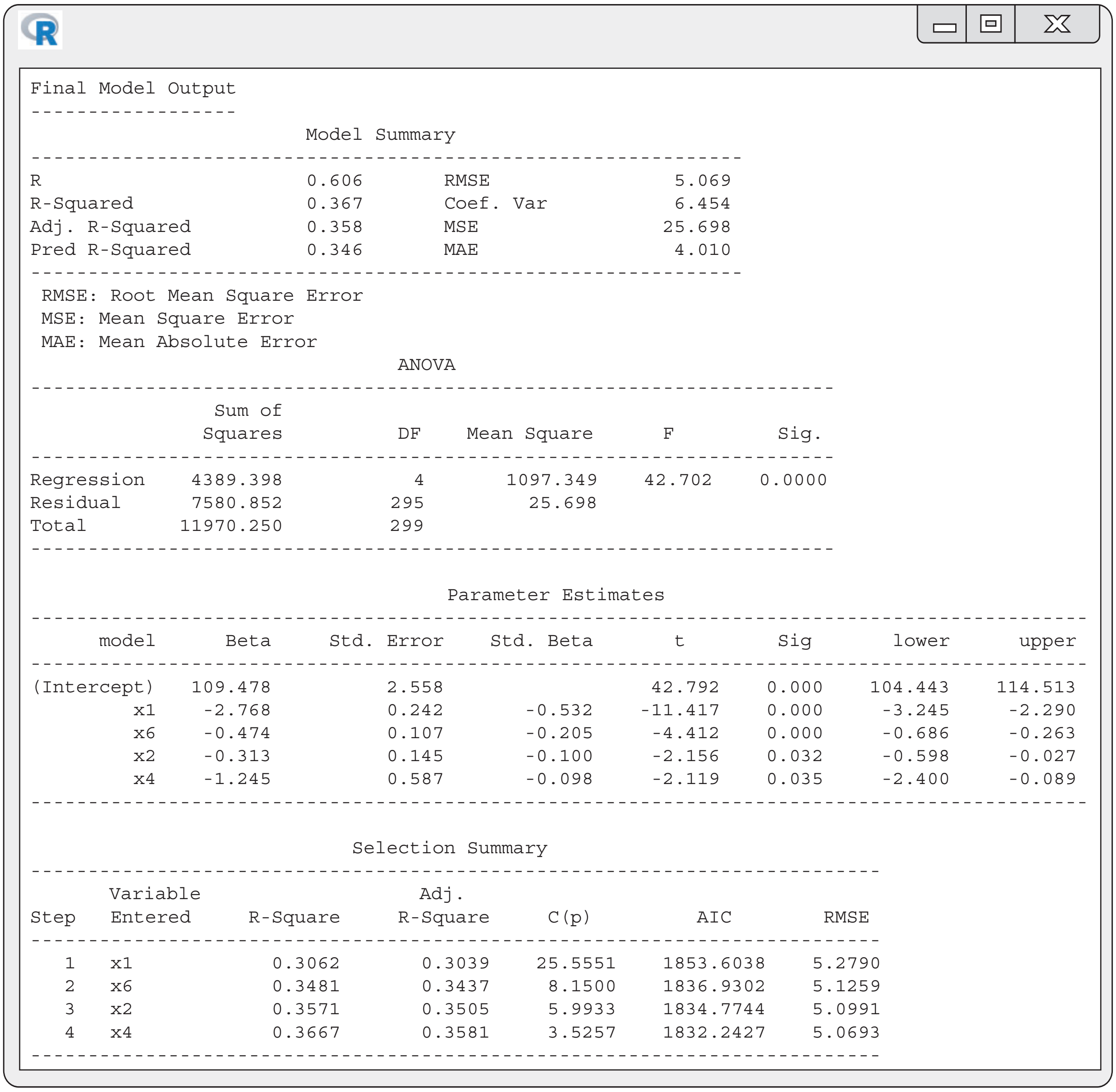
STEP 1 **Figure 12.70** shows the forward selection results using R.

STEP 2 The first variable included in the model is x_1. The variables added in order via forward selection are x_6, x_2, x_4. Each p value to enter the model is less than 0.05. At each step, the value of r^2, adjusted r^2, Mallows C_p, AIC (Akaike Information Criteria), and s (RMSE) are given.

STEP 3 The final estimated regression equation is

$$y = 109.478 - 2.768x_1 - 0.313x_2 - 1.245x_4 - 0.474x_6$$

Note that increasing the value of the p value required to enter the model may add predictor variables to the final estimated regression equation.



```
Final Model Output
------------------
                        Model Summary
--------------------------------------------------------------
R                       0.606       RMSE               5.069
R-Squared               0.367       Coef. Var          6.454
Adj. R-Squared          0.358       MSE               25.698
Pred R-Squared          0.346       MAE                4.010
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
                                ANOVA
----------------------------------------------------------------------
                Sum of
               Squares           DF    Mean Square      F         Sig.
----------------------------------------------------------------------
Regression    4389.398            4       1097.349    42.702    0.0000
Residual      7580.852          295         25.698
Total        11970.250          299
----------------------------------------------------------------------

                                    Parameter Estimates
--------------------------------------------------------------------------------------------
      model        Beta     Std. Error    Std. Beta        t         Sig        lower      upper
--------------------------------------------------------------------------------------------
(Intercept)    109.478         2.558                  42.792     0.000    104.443    114.513
         x1     -2.768         0.242       -0.532    -11.417     0.000     -3.245     -2.290
         x6     -0.474         0.107       -0.205     -4.412     0.000     -0.686     -0.263
         x2     -0.313         0.145       -0.100     -2.156     0.032     -0.598     -0.027
         x4     -1.245         0.587       -0.098     -2.119     0.035     -2.400     -0.089
--------------------------------------------------------------------------------------------

                            Selection Summary
--------------------------------------------------------------------------
       Variable                   Adj.
Step   Entered      R-Square    R-Square     C(p)         AIC         RMSE
--------------------------------------------------------------------------
   1   x1             0.3062      0.3039    25.5551    1853.6038    5.2790
   2   x6             0.3481      0.3437     8.1500    1836.9302    5.1259
   3   x2             0.3571      0.3505     5.9933    1834.7744    5.0991
   4   x4             0.3667      0.3581     3.5257    1832.2427    5.0693
--------------------------------------------------------------------------
```

Figure 12.70 The results using forward selection.

Backward Elimination

The method of backward elimination begins with the maximum model, a regression equation with all independent variables included. At each step, the single least significant variable, based on t tests for significant regression coefficients, is eliminated from the model. The next example demonstrates the method of backward elimination.

EXAMPLE 12.21 Backward Elimination

Suppose there are four possible independent variables in a multiple linear regression model: x_1, x_2, x_3, x_4. (Assume the intercept term is in the model.)

Solution

STEP 1 Consider the multiple linear regression model with all four predictors:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + E_i$$

Consider each hypothesis test, H_0: $\beta_i = 0$ versus H_a: $\beta_i \neq 0$, for $i = 1, 2, 3, 4$. The table shows the t statistic and the p value for each of these tests.

Added Variable	*t* Statistic	*p* Value
x_1	2.24	0.0309
x_2	1.36	0.1816
x_3	0.25	0.8039
x_4	−4.30	0.0001

Eliminate the least significant predictor variable, the independent variable associated with the largest p value greater than 0.05. In this example, delete x_3 from the model. Note that all p values could be less than 0.05. This suggests that all predictors are significant, and the final regression equation will include all of the independent variables.

STEP 2 Consider the multiple linear regression model with the remaining three predictors:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_4 x_{4i} + E_i$$

Conduct each hypothesis test, H_0: $\beta_i = 0$ versus H_a: $\beta_i \neq 0$, for $i = 1, 2, 4$. The table shows the t statistic and the p value for each of these tests.

Added Variable	*t* Statistic	*p* Value
x_1	2.56	0.0146
x_2	1.02	0.3142
x_4	−4.34	0.0001

Eliminate the variable with the largest p value (> 0.05). Therefore, eliminate x_2 from the model.

STEP 3 Continue in this manner until no variable can be eliminated from the model. Consider the multiple linear regression model with x_1 and x_4:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_4 x_{4i} + E_i$$

Conduct each hypothesis test, H_0: $\beta_i = 0$ versus H_a: $\beta_i \neq 0$, for $i = 1, 4$. The table shows the t statistic and the p value for each of these tests.

Added Variable	*t* Statistic	*p* Value
x_1	3.01	0.0047
x_4	−4.57	0.0000

There are no p values greater than 0.05. No other variable is eliminated. The backward elimination procedure suggests that the best model includes only the variables x_1 and x_4. ■

TRY IT NOW **Go to Exercise 12.233**

EXAMPLE 12.22 Exposure to Manganese (Continued)

Use the backward elimination procedure to find the best multiple linear regression model to predict IQ level.

Solution

STEP 1 **Figure 12.71** shows the backward elimination results using R.

STEP 2 The first variable eliminated from the model is x_3. The p value was the largest among the seven, and was greater than 0.05.

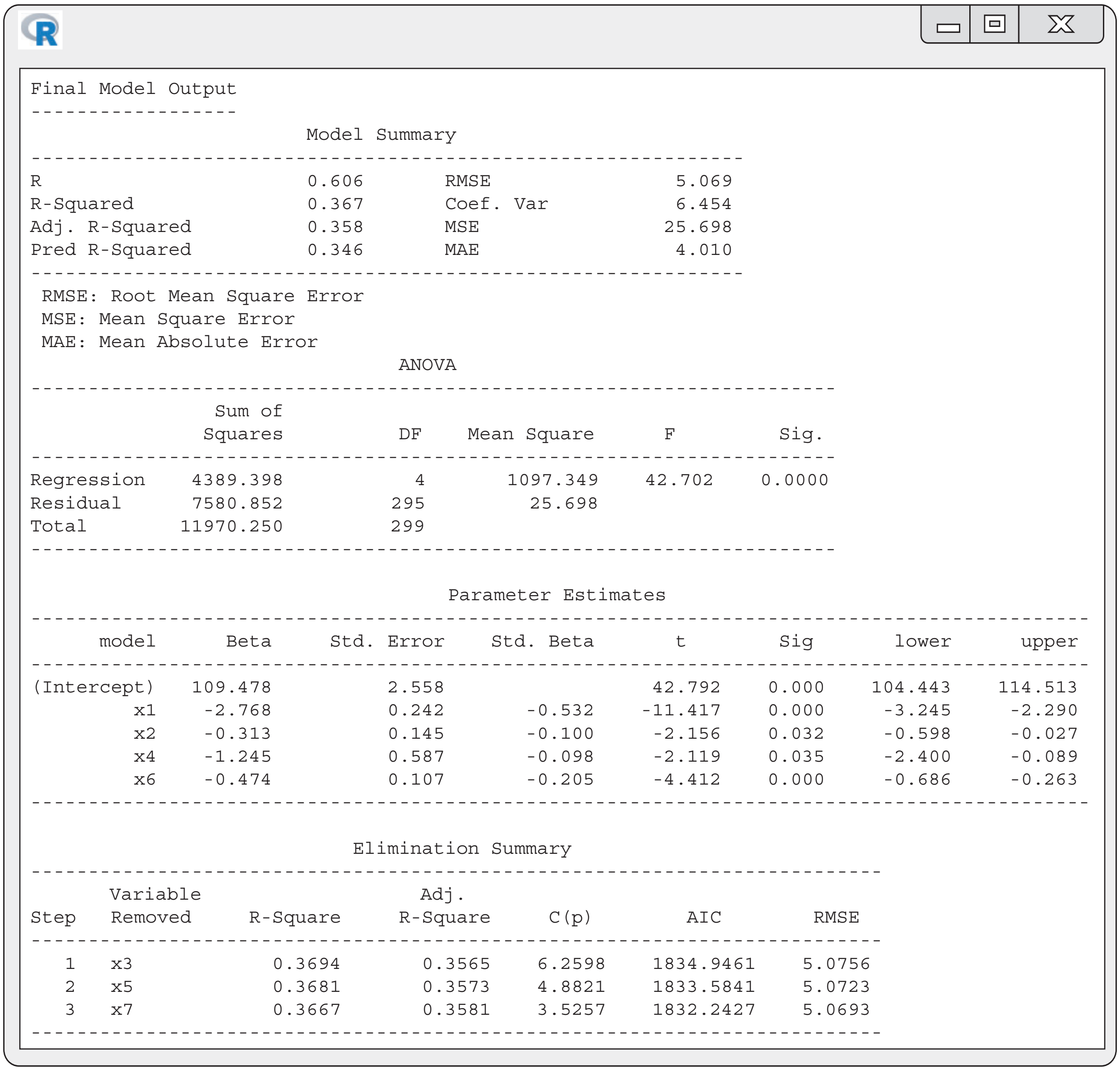
STEP 3 The other variables eliminated from the model, in order, are x_5 and x_7.

STEP 4 The final estimated regression equation is

$$y = 109.478 - 2.768x_1 - 0.313x_2 - 1.245x_4 - 0.474x_6$$

Note that this is the same model produced using forward selection (and best subsets). This is not always the case. Forward selection and backward elimination may result in different models. ■

TRY IT NOW **Go to Exercises 12.235 and 12.237**



```
Final Model Output
------------------
                        Model Summary
--------------------------------------------------------------
R                       0.606       RMSE                5.069
R-Squared               0.367       Coef. Var           6.454
Adj. R-Squared          0.358       MSE                25.698
Pred R-Squared          0.346       MAE                 4.010
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
                                ANOVA
----------------------------------------------------------------------
                Sum of
               Squares          DF    Mean Square      F         Sig.
----------------------------------------------------------------------
Regression    4389.398           4       1097.349    42.702    0.0000
Residual      7580.852         295         25.698
Total        11970.250         299
----------------------------------------------------------------------

                                    Parameter Estimates
--------------------------------------------------------------------------------------------
      model      Beta     Std. Error    Std. Beta       t        Sig       lower      upper
--------------------------------------------------------------------------------------------
(Intercept)   109.478          2.558                  42.792    0.000    104.443    114.513
         x1    -2.768          0.242       -0.532    -11.417    0.000     -3.245     -2.290
         x2    -0.313          0.145       -0.100     -2.156    0.032     -0.598     -0.027
         x4    -1.245          0.587       -0.098     -2.119    0.035     -2.400     -0.089
         x6    -0.474          0.107       -0.205     -4.412    0.000     -0.686     -0.263
--------------------------------------------------------------------------------------------

                            Elimination Summary
--------------------------------------------------------------------------
       Variable                   Adj.
Step   Removed     R-Square    R-Square     C(p)        AIC        RMSE
--------------------------------------------------------------------------
   1   x3            0.3694       0.3565    6.2598    1834.9461    5.0756
   2   x5            0.3681       0.3573    4.8821    1833.5841    5.0723
   3   x7            0.3667       0.3581    3.5257    1832.2427    5.0693
--------------------------------------------------------------------------
```

Figure 12.71 The results using backward elimination.

Stepwise Regression

If all possible subsets of predictor variables are considered to construct the best multiple linear regression model, this analysis could take a lot of computer power and time. Forward selection and backward elimination greatly reduce the number of models considered and, therefore, the computation time. However, there is a chance of missing a better model, especially in cases with many possible predictor variables.

Using forward selection, once a variable is included in the model, it remains in the model regardless of other variables added later. Similarly, using backward elimination, if a variable is eliminated from the model, it can never be included in a later step.

Stepwise regression is a modification of forward selection or backward elimination. At each step in the procedure, the entire model is reevaluated. Here is how it works applied to forward selection.

Suppose the variable x_i is added to the model in the usual way at a certain step. Before another variable is added, compute the current estimated regression equation and test the significance of each variable in the model. Consider each hypothesis test for a significant regression coefficient. Delete the variable with the highest p value (above a threshold, such as 0.05). Recompute the estimated regression equation if necessary, and proceed with the next step in forward selection.

There is an analogous procedure for stepwise regression applied to backward elimination. Stepwise regression is generally better than either forward selection or backward elimination for selecting the best model because it considers more models (but still not all possible subsets of predictor variables). Most statistical software packages have an option for stepwise regression applied to forward selection and/or backward elimination.

Section 12.7 Exercises

Concept Check

12.221 **True or False** Adding any additional independent variable to a regression model will always increase the value of r^2.

12.222 **True or False** Forward selection and backward elimination will always produce the same multiple linear regression model.

12.223 **True or False** Using forward selection, increasing the value of p required to enter the model may add predictor variables to the final estimated regression equation.

12.224 **True or False** Using backward elimination, at each step, eliminate the variable with the largest t statistic greater than 1.96 in absolute value.

12.225 **True or False** None of the predictors can be indicator variables when using forward selection or backward elimination.

12.226 **Short Answer** Name one disadvantage of using the best-subsets method to select the best multiple linear regression model.

12.227 **Short Answer** Name two reasonable measures of multiple linear regression model goodness.

Practice

12.228 Describe stepwise regression applied to backward elimination.

12.229 Suppose a statistician wants to find the best model in which there are 10 possible predictor variables.

(a) How many possible models are there if the researcher considers all possible subsets?

(b) Suppose forward selection is used, and the final model contains all 10 independent variables. How many models were considered in the process?

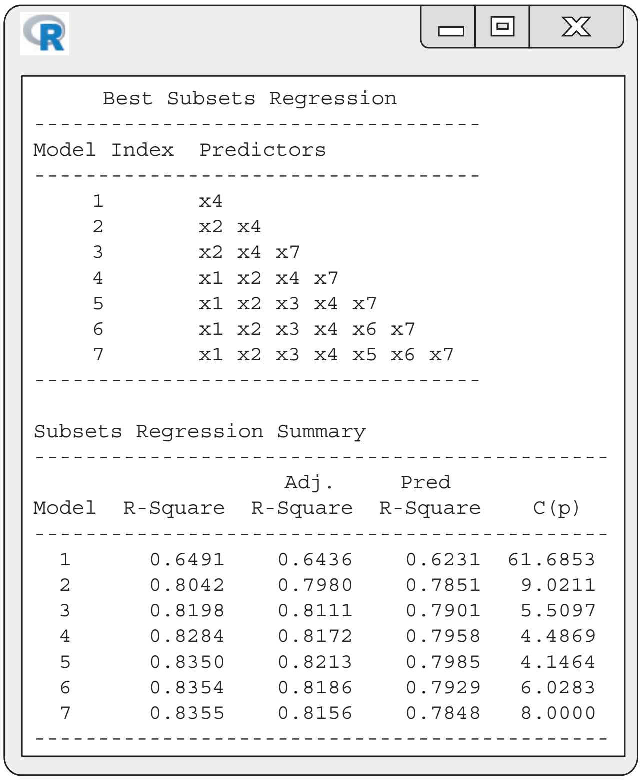
(c) Suppose backward elimination is used, and the final model contains a single independent variable. How many models were considered in the process?

12.230 Consider the R output using the best subset approach to determine a multiple linear regression model. EX12.230

(a) Construct a graph of r^2 versus k.

(b) Construct a graph of C_p versus k.

(c) Which predictor variables do you think should be included in the model? Justify your answer.



```
     Best Subsets Regression
-----------------------------------
Model  Index  Predictors
-----------------------------------
     1       x4
     2       x2 x4
     3       x2 x4 x7
     4       x1 x2 x4 x7
     5       x1 x2 x3 x4 x7
     6       x1 x2 x3 x4 x6 x7
     7       x1 x2 x3 x4 x5 x6 x7
-----------------------------------

Subsets Regression Summary
---------------------------------------------
                    Adj.      Pred
Model  R-Square  R-Square  R-Square    C(p)
---------------------------------------------
  1      0.6491    0.6436    0.6231  61.6853
  2      0.8042    0.7980    0.7851   9.0211
  3      0.8198    0.8111    0.7901   5.5097
  4      0.8284    0.8172    0.7958   4.4869
  5      0.8350    0.8213    0.7985   4.1464
  6      0.8354    0.8186    0.7929   6.0283
  7      0.8355    0.8156    0.7848   8.0000
---------------------------------------------
```

12.231 Consider a data set for which y is the dependent variable and x_1 through x_5 are possible predictor variables. EX12.231

(a) Use the value of r^2 to select the best multiple linear regression model from all possible subsets of independent variables.

(b) Construct a graph of r^2 versus k to support your conclusion in part (a).

(c) Use the best model to find an estimate of the mean value of Y for $x = (60, 51, 2.2, 35.67, 56)$.

12.232 Consider a data set for which y is the dependent variable and x_1 through x_4 are possible predictor variables. EX12.232

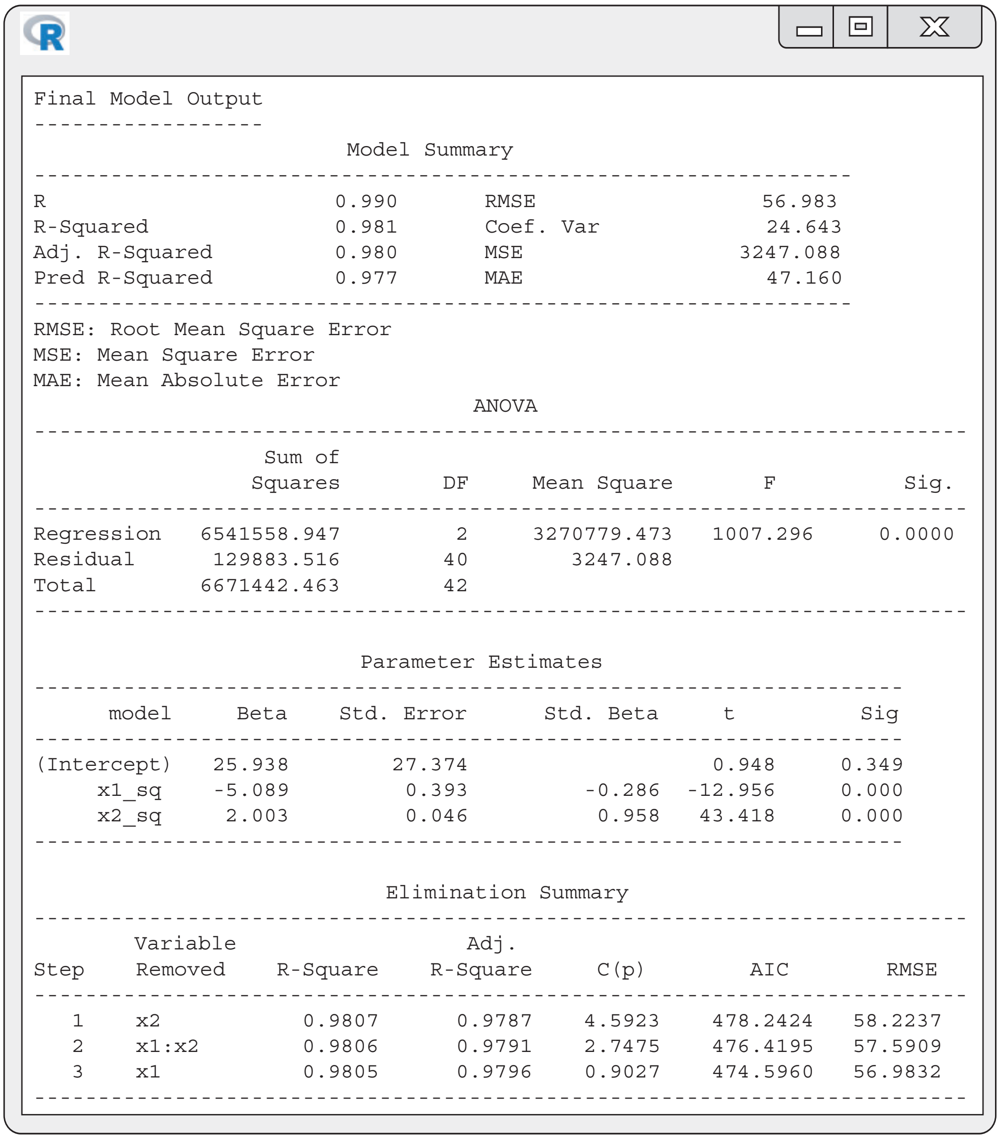
(a) Use forward selection to find the best multiple linear regression model.

(b) Use stepwise regression applied to forward selection to find the best multiple linear regression model. Is the resulting model any different from part (a)? If so, why?

12.233 Consider the R output using the backward elimination approach to determine a multiple linear regression model. The possible predictor variables are x_1, x_1^2, x_2, x_2^2, and x_1x_2. The first model considered in this process is quadratic in x_1 and x_2, and includes an interaction term. EX12.233

(a) Which variable is eliminated first? Second?

(b) Which variables are included in the best multiple linear regression model?



```
Final Model Output
------------------
                        Model Summary
----------------------------------------------------------------
R                       0.990       RMSE                  56.983
R-Squared               0.981       Coef. Var             24.643
Adj. R-Squared          0.980       MSE                 3247.088
Pred R-Squared          0.977       MAE                   47.160
----------------------------------------------------------------
RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error
                                  ANOVA
-------------------------------------------------------------------------
                 Sum of
                Squares        DF    Mean Square       F          Sig.
-------------------------------------------------------------------------
Regression   6541558.947        2    3270779.473    1007.296     0.0000
Residual      129883.516       40       3247.088
Total        6671442.463       42
-------------------------------------------------------------------------

                          Parameter Estimates
---------------------------------------------------------------------
      model      Beta    Std. Error      Std. Beta      t         Sig
---------------------------------------------------------------------
(Intercept)    25.938        27.374                   0.948     0.349
     x1_sq     -5.089         0.393         -0.286  -12.956     0.000
     x2_sq      2.003         0.046          0.958   43.418     0.000
---------------------------------------------------------------------

                            Elimination Summary
-------------------------------------------------------------------------
        Variable                  Adj.
Step    Removed     R-Square    R-Square     C(p)        AIC         RMSE
-------------------------------------------------------------------------
   1    x2            0.9807      0.9787    4.5923    478.2424    58.2237
   2    x1:x2         0.9806      0.9791    2.7475    476.4195    57.5909
   3    x1            0.9805      0.9796    0.9027    474.5960    56.9832
-------------------------------------------------------------------------
```

Applications

12.234 Fuel Consumption and Cars Research suggests that it can cost significantly more to repair luxury cars than non-luxury cars when they are involved in low-speed crashes. Accidents that occur in parking lots or commuter traffic usually cause damage to bumpers, grills, and headlights. To build a model to predict the cost of repairs, the Insurance Institute for Highway Safety obtained records from random luxury cars involved in low-speed crashes. The following variables were considered: REPAIRS

y = repair cost as a result of the accident damage
x_1 = list price of the automobile
x_2 = speed of the automobile at the time of the accident
x_3 = indicator variable: 0 = forward, 1 = reverse
x_4 = wheelbase, in inches
x_5 = curb weight, in pounds
x_6 = engine power, in horsepower, HP

(a) Use backward elimination to find the best multiple linear regression model.

(b) Suppose a 2019 BMW 528i is involved in a 3-mph crash while backing into a parking spot at a shopping mall. The other specifications include MSRP, $49,500; HP, 240; wheelbase, 116.9 in.; curb weight, 3814 lb. Find a 95% confidence interval for the mean repair cost.

12.235 Public Health and Nutrition Suppose a two-year study was conducted to investigate the effect of nutrient intake on changes in body composition. Thirty-six women, 18–31 years of age, were selected at random. A diet survey was used to determine daily nutrient intake. The following variables were considered: INTAKE

y = percent change in body weight

x_1 = daily fat intake, in grams

x_2 = daily protein intake, in grams

x_3 = daily calcium/energy intake, in mg/kcal

x_4 = daily sodium intake, in milligrams

x_5 = daily vitamin A intake, in International Units, IU

x_6 = daily carbohydrate intake, in grams

(a) Use backward elimination to find the best model.

(b) Use forward selection to find the best model. Compare this model with the one in part (a).

(c) Add another possible predictor variable, an interaction term, x_3x_5. Use backward elimination again to find the best model.

12.236 Medicine and Clinical Studies Physicians at the Emergency Medicine Research (EMR) Group in Coventry, England, conducted a study to predict the number of days a patient stays in the hospital based on emergency department data. The following observations were recorded for each patient: EMR

y = number of days in the hospital

x_1 = elapsed time from arrival in the emergency department until seen by a doctor

x_2 = elapsed time from initial evaluation until decision to admit/not admit

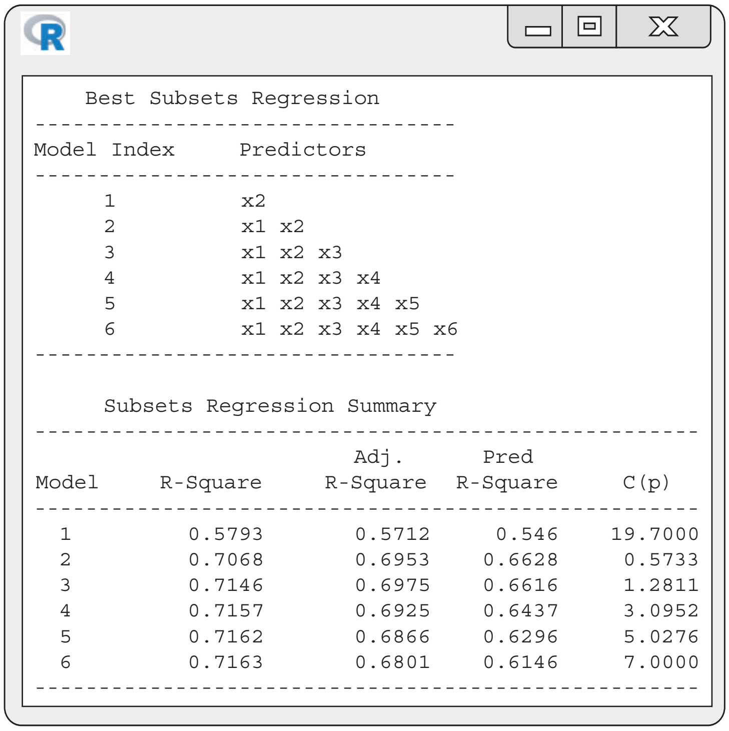
x_3 = severity of the injury, using the EMR scale

x_4 = elapsed time from accident to arrival at the emergency department

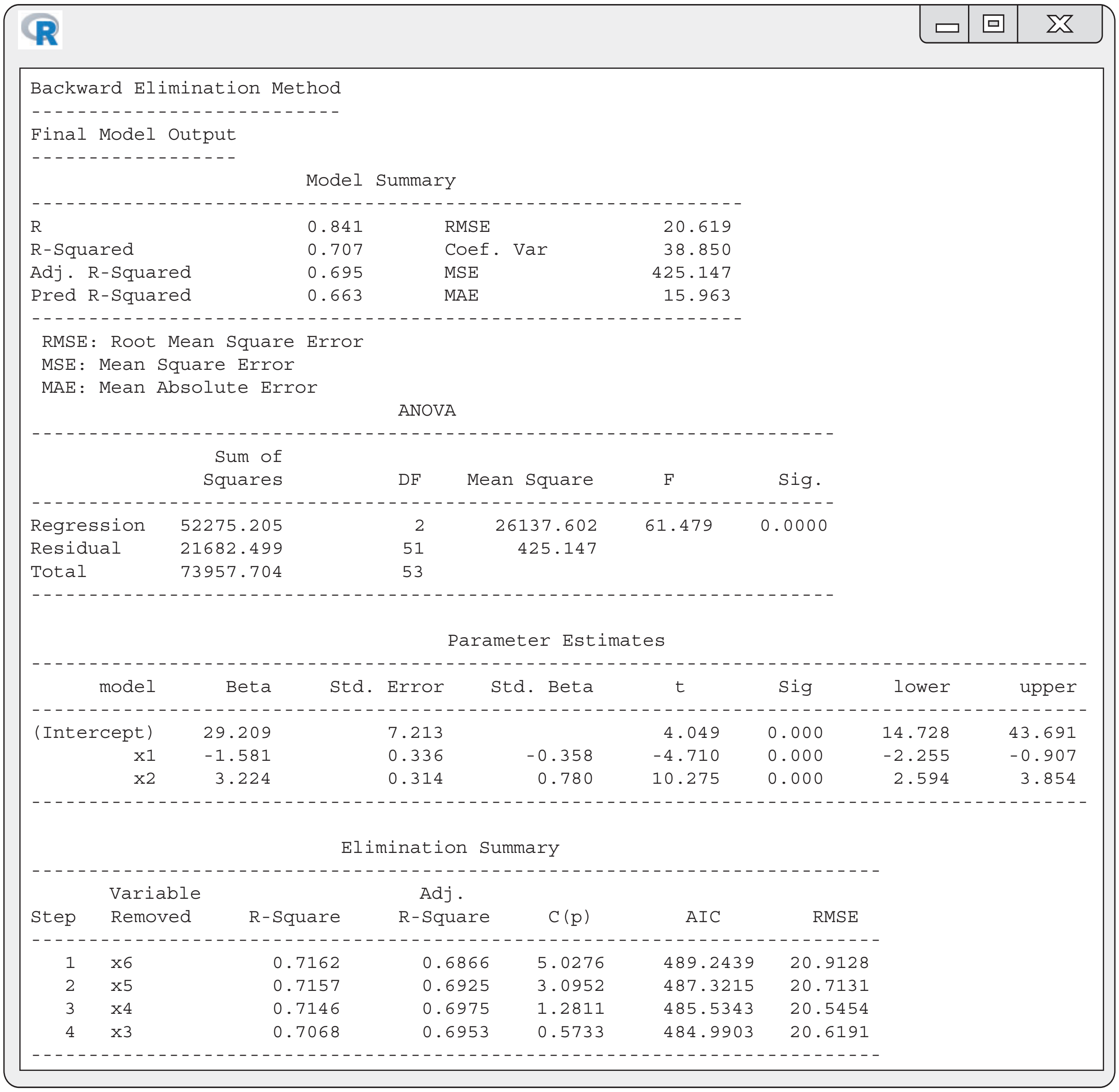
x_5 = initial pulse rate

x_6 = initial respiratory rate

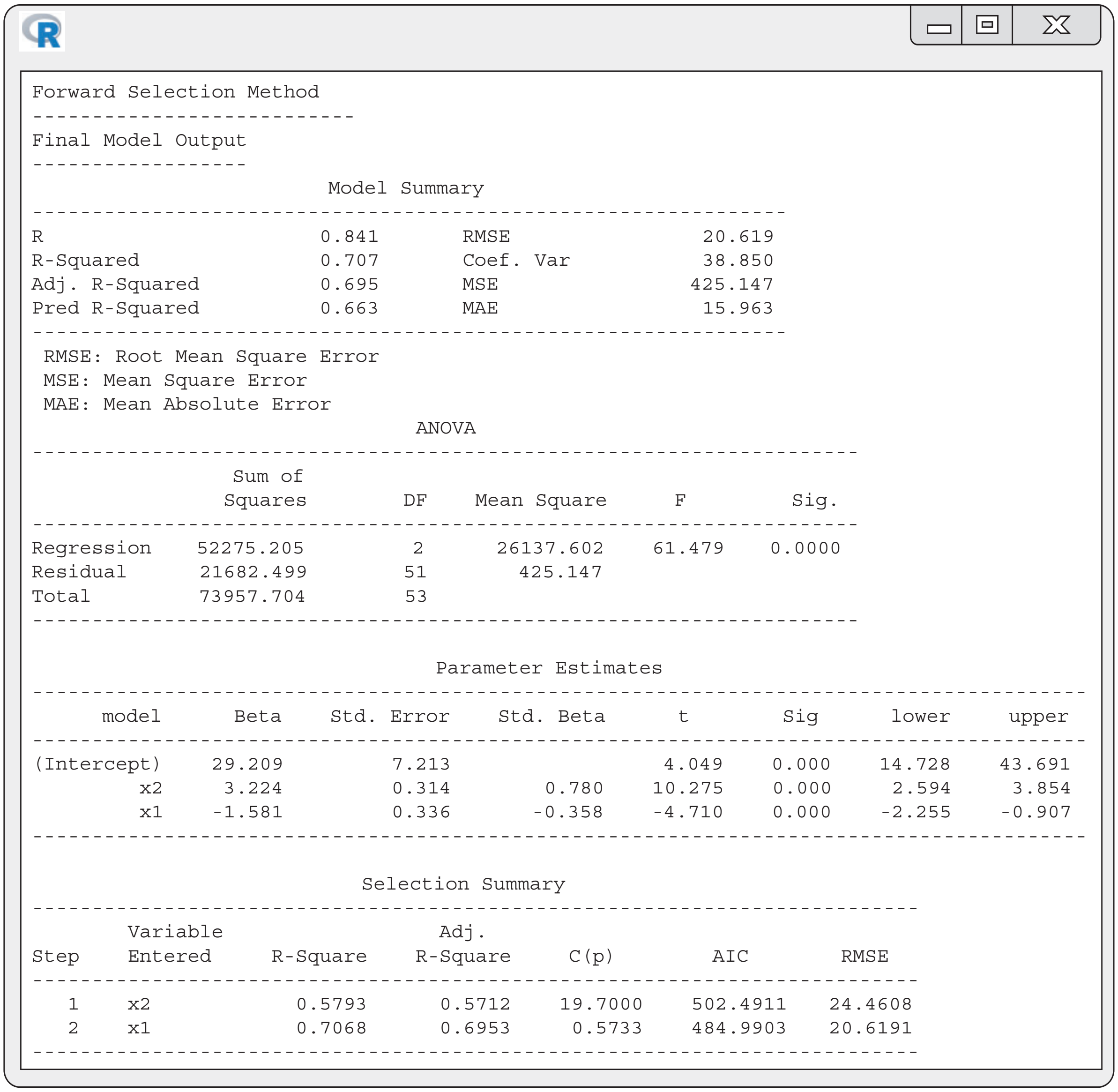
The R output from best subsets, backward elimination, and forward selection are shown in the figures that follow.



```
    Best Subsets Regression
--------------------------------
Model  Index      Predictors
--------------------------------
     1          x2
     2          x1 x2
     3          x1 x2 x3
     4          x1 x2 x3 x4
     5          x1 x2 x3 x4 x5
     6          x1 x2 x3 x4 x5 x6
--------------------------------

     Subsets Regression Summary
----------------------------------------------------
                         Adj.       Pred
Model     R-Square     R-Square  R-Square      C(p)
----------------------------------------------------
  1         0.5793        0.5712      0.546    19.7000
  2         0.7068        0.6953     0.6628     0.5733
  3         0.7146        0.6975     0.6616     1.2811
  4         0.7157        0.6925     0.6437     3.0952
  5         0.7162        0.6866     0.6296     5.0276
  6         0.7163        0.6801     0.6146     7.0000
----------------------------------------------------
```



```
Backward Elimination Method
---------------------------
Final Model Output
------------------
                        Model Summary
--------------------------------------------------------------
R                       0.841       RMSE               20.619
R-Squared               0.707       Coef. Var          38.850
Adj. R-Squared          0.695       MSE               425.147
Pred R-Squared          0.663       MAE                15.963
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
                                ANOVA
----------------------------------------------------------------------
                Sum of
               Squares          DF    Mean Square      F         Sig.
----------------------------------------------------------------------
Regression   52275.205           2      26137.602    61.479     0.0000
Residual     21682.499          51        425.147
Total        73957.704          53
----------------------------------------------------------------------

                                    Parameter Estimates
--------------------------------------------------------------------------------------------
      model        Beta      Std. Error      Std. Beta         t          Sig         lower        upper
--------------------------------------------------------------------------------------------
(Intercept)     29.209           7.213                       4.049      0.000        14.728       43.691
         x1     -1.581           0.336          -0.358      -4.710      0.000        -2.255       -0.907
         x2      3.224           0.314           0.780      10.275      0.000         2.594        3.854
--------------------------------------------------------------------------------------------

                           Elimination Summary
--------------------------------------------------------------------------
       Variable                  Adj.
Step   Removed     R-Square     R-Square     C(p)         AIC         RMSE
--------------------------------------------------------------------------
   1   x6            0.7162       0.6866     5.0276     489.2439    20.9128
   2   x5            0.7157       0.6925     3.0952     487.3215    20.7131
   3   x4            0.7146       0.6975     1.2811     485.5343    20.5454
   4   x3            0.7068       0.6953     0.5733     484.9903    20.6191
--------------------------------------------------------------------------
```



```
Forward Selection Method
---------------------------
Final Model Output
------------------
                        Model Summary
---------------------------------------------------------------
R                        0.841       RMSE               20.619
R-Squared                0.707       Coef. Var          38.850
Adj. R-Squared           0.695       MSE               425.147
Pred R-Squared           0.663       MAE                15.963
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
                               ANOVA
---------------------------------------------------------------------
                Sum of
               Squares        DF    Mean Square      F         Sig.
---------------------------------------------------------------------
Regression    52275.205        2      26137.602    61.479    0.0000
Residual      21682.499       51        425.147
Total         73957.704       53
---------------------------------------------------------------------

                                Parameter Estimates
----------------------------------------------------------------------------------------
      model       Beta     Std. Error     Std. Beta      t         Sig       lower     upper
----------------------------------------------------------------------------------------
(Intercept)     29.209         7.213                   4.049     0.000     14.728    43.691
         x2      3.224         0.314       0.780      10.275     0.000      2.594     3.854
         x1     -1.581         0.336      -0.358      -4.710     0.000     -2.255    -0.907
----------------------------------------------------------------------------------------

                             Selection Summary
--------------------------------------------------------------------------
        Variable                  Adj.
Step    Entered      R-Square    R-Square     C(p)        AIC        RMSE
--------------------------------------------------------------------------
   1    x2             0.5793      0.5712    19.7000    502.4911    24.4608
   2    x1             0.7068      0.6953     0.5733    484.9903    20.6191
--------------------------------------------------------------------------
```

(a) Using the best subsets output, construct a graph of r^2 versus various models. Which predictor variables would you recommend for inclusion in the model? Why?

(b) Using the backward elimination output, what predictor variables would you recommend for inclusion in the model? Which variable was excluded first in this process? Second?

(c) Using the forward selection output, what predictor variables would you recommend for inclusion in the model?

(d) Consider all three methods. What predictor variables would you recommend for inclusion in the model? Do these predictor variables seem reasonable? Are any variables left out of the model that might be good predictors?

12.237 **Physical Sciences** Coal is still an abundant energy source and is used in many electricity-generating plants in the United States. For example, the coal-burning Montour Power Plant in Washingtonville, Pennsylvania, has a generating capacity of 1504 MW. To manage production and purchasing, plant officials studied coal production at various mines. A random sample of mines in the United States was obtained, and the following observations were obtained for each one: COAL

$y =$ output of raw coal, in metric tons

$x_1 =$ area of the mine, in km^2

$x_2 =$ total reserve, in metric tons

$x_3 =$ recoverable reserve, in metric tons

$x_4 =$ depth of the shaft, in meters

$x_5 =$ designed annual output of raw coal

$x_6 =$ average thickness of the coal seam, in meters

(a) Use backward elimination to find the best multiple linear regression model. Find the estimated regression coefficients.

(b) Estimate the true mean value of Y when $x = (35.2, 132.85, 28.17, 297, 2.4, 8.6)$.

(c) Find the residuals associated with your model. Construct a normal probability plot for the residuals. Is there any evidence to suggest that the random errors are not normal? Justify your answer.

(d) Carefully sketch a graph of the residuals versus each predictor variable in your model. Is there any evidence of a violation of the regression assumptions? Justify your answer.

12.238 Travel and Transportation Ferryboats are a major source of public transportation in some cities. For example, New York City and Miami operate ferryboats, and Seattle, Washington, has approximately 28 ferryboats in operation. A study was conducted to develop a model to predict total yearly passenger miles. The following variables were considered: FERRY

$y =$ total passenger miles, in thousands

$x_1 =$ transportation zone population

$x_2 =$ maximum number of vehicles operated

$x_3 =$ maximum number of vehicles available

$x_4 =$ vehicle revenue miles, in thousands

$x_5 =$ vehicle revenue hours, in thousands

$x_6 =$ passenger unlinked trips in thousands

A random sample of primary cities was obtained.

(a) Use backward elimination with $\alpha = 0.05$ to find the best model.

(b) Use forward selection with $\alpha = 0.05$ to find the best model.

(c) Repeat parts (a) and (b) with $\alpha = 0.10$. Does either procedure produce a different model?

(d) Of all four models, which would you recommend as the best? Justify your answer.

12.239 Fuel Consumption and Cars Zipcar is a self-serve hourly car rental agency. After paying an annual membership fee, customers use a swipe ID card to rent available cars, usually in urban areas. There are no lines, no paperwork, and no service people. The company owner decided to conduct a study to predict the number of hours each car is rented. A random sample of rentals was obtained, and observations were recorded for the following variables: ZIPCAR

$y =$ number of hours rented

$x_1 =$ time of day for rental, in hours after 12:00 A.M.

$x_2 =$ indicator variable, $0 =$ weekday, $1 =$ weekend

$x_3 =$ renter's annual income, in thousands of dollars

$x_4 =$ hourly rate for the rental

$x_5 =$ annual membership fee

(a) Use backward elimination with $\alpha = 0.10$ to find the best model. Find the estimated regression coefficients.

(b) Use forward selection with stepwise regression, $\alpha = 0.10$, to find the best model. Compare this model with the one found in part (a).

(c) Use the estimated regression coefficients to explain the relationship between the indicator variable and the hours a car is rented, and between the hourly rate and the hours a car is rented.

(d) Suppose the hourly rate is $10 on a weekend day. Find a 95% confidence interval for the true mean number of hours the car will be rented.

Extended Applications

12.240 Business and Management Whenever a consumer makes a purchase with a credit card, it must be verified and authenticated. Usually the card is swiped, and the sale is approved within a few seconds. A consumer group is trying to develop a model to predict the amount of time it takes to verify a purchase. A random sample of sales was obtained, and observations were recorded for the following variables: CREDIT

$y =$ time until the purchase is approved, in seconds

$x_1 =$ amount of the purchase, in dollars

$x_2 =$ type of store, $0 =$ convenience, $1 =$ restaurant, $2 =$ retail

$x_3 =$ distance from home, in miles

$x_4 =$ time of day, in hours after 12:00 A.M.

$x_5 =$ number of days until Christmas

(a) Use backward elimination to find the best regression model ($\alpha = 0.10$). Use forward selection to find the best regression model ($\alpha = 0.10$). Don't forget to use the appropriate indicator variables.

(b) One member of the research team believes that the time of day raised to the fourth power should be in the model. Add this as a possible predictor, and use the method of your choice to find the best model.

(c) Using the model found in part (b), find an estimate of the mean time for credit card verification when $x = (152.00, 1, 20, 18.5, 120)$.

(d) Consider a model with x_4, x_4^2, x_4^3, and x_4^4. What happens when you try any method with these as possible predictors? Why?

12.241 Biology and Environmental Science Reliable measurements of river beds are important for water management, shipping, and flood predictions. A random sample of rivers in the United States was obtained, and observations were recorded for the following variables: RIVBED

$y =$ total river-bed material load, in kg/s

$x_1 =$ discharge, in m^3/s

$x_2 =$ average velocity, in m/s

$x_3 =$ bottom width, in meters

$x_4 =$ flow depth, in meters

$x_5 =$ area, in m^2

$x_6 =$ longitudinal slope

(a) Use forward selection with $\alpha = 0.05$ to find the best model. Use the sign of each regression coefficient to explain the relationship between each predictor in the model and the total river-bed material load.

(b) Use backward elimination with $\alpha = 0.05$ to find the best model. Compare this model with the model in part (a). Carefully sketch a normal probability plot of the residuals. Is there any evidence to suggest a violation of the regression assumptions?

(c) Suppose that the river bed will be dredged if the total river-bed material load is greater than 50 kg/s. Measurements for the Kissimmee River in Florida showed $x = (40, 1.25, 27, 1.05, 31.35, 0.015)$. Using the model in part (b), find a 95% confidence interval for the mean total river-bed material load. Use this confidence interval to determine if there is any evidence to suggest that this river should be dredged. Justify your answer.

12.242 Technology and the Internet A recent research study examined the effects of economic and information/communications technology (ICT) on Internet purchases by individuals in European Union member states. The following variables were considered: ICT

y = Internet purchases in the last 12 months, expressed as a percentage

x_1 = level of Internet access, expressed as a percentage

x_2 = fixed broadband penetration rate per 100 inhabitants

x_3 = level of computer skills, as a percentage of individuals aged 16–74

x_4 = public expenditure on education as a percentage of GDP

x_5 = GDP per capita in Purchasing Power Standards

x_6 = individuals using the Internet for finding information, as a percentage of individuals aged 16–74

x_7 = individuals' level of Internet skills, as a percentage of all individuals aged 16–74

x_8 = individuals who ordered/bought goods or services over the Internet, as a percentage of individuals aged 16–74

x_9 = concern about possible problems related to Internet usage, as a percentage of all individuals

A random sample of individuals was obtained.

(a) Use forward selection and backward elimination to determine the best multiple linear regression model. Use the sign of each estimated coefficient to explain the effect of each predictor variable on the percentage of Internet purchases.
(b) What would you say is the most important variable in predicting the percentage of Internet purchases? Why?
(c) Construct a normal probability plot of the residuals. Is there any evidence to suggest that the random errors are not normal? Justify your answer.
(d) Carefully sketch a graph of the residuals versus each predictor variable in the model. Is there any evidence of a violation of the regression assumptions?

12.243 Economics and Finance Researchers A. Q. Do and G. Grudnitski have written several articles concerning the prediction of the selling price of a residential home. They indicate that some of the important variables are the age of the home and the lot size. A real estate agent in Napa Valley would like to develop a model to predict selling price in her area. She would like to consider the following variables: REPRICE

y = selling price of the home, in thousands of dollars

x_1 = age of the home, in years

x_2 = number of bedrooms

x_3 = number of bathrooms

x_4 = square footage of living area

x_5 = number of garage stalls

x_6 = number of fireplaces

x_7 = number of stories

x_8 = lot size, in acres

x_9 = indicator variable, abuts golf course $(0 = \text{no}, 1 = \text{yes})$

A random sample of home sales was obtained.

(a) Use any method you wish to develop the best multiple linear regression model. What would you say is the most important variable in predicting the selling price of a home in Napa Valley? Why?
(b) Find the estimated regression equation. Explain the relationship between each predictor in the model and the selling price.
(c) Construct a normal probability plot of the residuals. Is there any evidence to suggest that the random errors are not normal? Justify your answer.
(d) Carefully sketch a graph of the residuals versus each predictor variable in the model. Is there any evidence of a violation of the regression assumptions?
(e) Are there any other variables that you believe should be considered in the model? If so, why?

12.244 Public Health and Nutrition Several factors affect the quality of espresso coffee, especially the amount of caffeine in a cup. An official from the U.S. Food and Drug Administration would like to determine the best model for predicting the amount of caffeine in a 12-oz cup of espresso. Here are the variables for consideration: CAFF

y = amount of caffeine, in milligrams

x_1 = temperature of the steam, in °C

x_2 = brewing pressure, in bars

x_3 = dose of ground coffee, in grams

x_4 = brewing time, in seconds

x_5 = grinding level of coffee, 0 = fine, 1 = fine-coarse, 2 = coarse

x_6 = filter holder, 0 = 1-cup, 1 = 2-cups

x_7 = tamping, 0 = no, 1 = yes

x_8 = coffee beans, 0 = light, 1 = dark

A random sample of 12-oz espresso drinks was obtained.

(a) Use any method you wish to develop the best multiple linear regression model for predicting the amount of caffeine in a 12-oz cup of espresso. What would you say is the most important variable in predicting the amount of caffeine? Justify your answer.
(b) Find the estimated regression equation and explain the relationship between each predictor in the model and the amount of caffeine.
(c) Find the residuals. Use the four methods presented in Section 6.3 to determine whether there is any evidence to suggest the residuals are from a non-normal population.
(d) Find the expected amount of caffeine in a 12-oz cup of espresso for $x = (90, 15, 6, 30, 0, 1, 1, 1)$

Chapter 12 Summary

Multiple Linear Regression Model

Let $(x_{11}, x_{21}, \ldots, x_{k1}, y_1), (x_{12}, x_{22}, \ldots, x_{k2}, y_2), \ldots, (x_{1n}, x_{2n}, \ldots, x_{kn}, y_n)$ be n sets of observations such that y_i is an observed value of the random variable Y_i. We assume that there exist constants $\beta_0, \beta_1, \ldots, \beta_k$ such that

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + E_i$$

where $E_1, E_2, \ldots, E_n$ are independent, normal random variables with mean 0 and variance σ^2. That is,

1. The E_i's are normally distributed, which implies that the Y_i's are normally distributed.
2. The expected value of E_i is 0, which implies that
$$E(Y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$$
3. $\text{Var}(E_i) = \sigma^2$, which implies that $\text{Var}(Y_i) = \sigma^2$.
4. The E_i's are independent, which implies that the Y_i's are independent.

The E_i's are the random deviations or random error terms.

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ is the true regression line.

Principle of least squares

The estimated regression equation is obtained by minimizing the sum of the squared deviations between the observations and the estimated values.

The estimated regression equation is $y = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_k x_k$.

The ith predicted (fitted) value is $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} + \cdots + \widehat{\beta}_k x_{ki}, (i = 1, 2, \ldots, n)$

The ith residual is $\widehat{e}_i = y_i - \widehat{y}_i$.

The sum of squares:

$$\underbrace{\Sigma(y_i - \overline{y})^2}_{\text{SST}} = \underbrace{\Sigma(\widehat{y}_i - \overline{y})^2}_{\text{SSR}} + \underbrace{\Sigma(y_i - \widehat{y}_i)^2}_{\text{SSE}}$$

ANOVA summary table for multiple linear regression

Source of variation	Sum of squares	Degrees of freedom	Mean square	F	p Value
Regression	SSR	k	$\text{MSR} = \dfrac{\text{SSR}}{k}$	$\dfrac{\text{MSR}}{\text{MSE}}$	p
Error	SSE	$n-k-1$	$\text{MSE} = \dfrac{\text{SSE}}{n-k-1}$		
Total	SST	$n-1$			

Coefficient of determination: $r^2 = \text{SSR/SST}$

Estimate of variance: $s^2 = \text{SSE}/(n-k-1)$

F test for a significant multiple linear regression

$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$

(None of the predictor variables helps explain the variation in y.)

$H_a: \beta_i \neq 0$ for at least one i

(At least one predictor variable helps explain the variation in y.)

$$\text{TS: } F = \frac{\text{MSR}}{\text{MSE}}$$

RR: $F \geq F_{\alpha,\,k,\,n-k-1}$

Hypothesis test concerning β_i

H_0: $\beta_i = \beta_{i0}$,
H_a: $\beta_i > \beta_{i0}$, $\quad \beta_i < \beta_{i0}$, $\quad$ or $\quad \beta_i \neq \beta_{i0}$

TS: $T = \dfrac{B_i - \beta_{i0}}{S_{B_i}}$

RR: $T \geq t_{\alpha,\,n-k-1}$, $\quad T \leq -t_{\alpha,\,n-k-1}$, $\quad$ or $\quad |T| \geq t_{\alpha/2,\,n-k-1}$

A $100(1-\alpha)\%$ confidence interval for β_i has the following values as endpoints:

$\widehat{\beta}_i \pm t_{\alpha/2,\,n-k-1} \cdot s_{B_i}$

*Hypothesis test concerning the mean value of Y for x=x**

H_0: $y^* = y_0^*$,
H_a: $y^* > y_0^*$, $\quad y^* < y_0^*,$ $\quad$ or $\quad y^* \neq y_0^*$

TS: $T = \dfrac{(B_0 + B_1 x_1^* + \cdots + B_k x_k^*) - y_0^*}{S_{Y^*}}$

RR: $T \geq t_{\alpha,\,n-k-1}$, $\quad T \leq -t_{\alpha,\,n-k-1}$, $\quad$ or $\quad |T| \geq t_{\alpha/2,\,n-k-1}$

A $100(1-\alpha)\%$ confidence interval for $\mu_{Y|x^*}$, the mean value of Y for $x = x^*$, has the following values as endpoints:

$$(\widehat{\beta}_0 + \widehat{\beta}_1\ x_1^* + \cdots + \widehat{\beta}_k x_k^*) \pm t_{\alpha/2,\,n-k-1} \cdot s_{Y^*}$$

Prediction interval for an observed value of Y

A $100(1-\alpha)\%$ prediction interval for an observed value of Y when $x = x^*$ has the following values as endpoints:

$$(\widehat{\beta}_0 + \widehat{\beta}_1\ x_1^* + \cdots + \widehat{\beta}_k x_k^*) \pm t_{\alpha/2,\,n-k-1} \cdot \sqrt{s^2 + s_{Y^*}^2}$$

Regression diagnostics

1. Construct a histogram, stem-and-leaf plot, scatter plot, and/or normal probability plot of the residuals. These graphs are all used to check the normality assumption.
2. Construct a scatter plot of the residuals versus each independent variable. If there are no violations in assumptions, each scatter plot should appear as a horizontal band around 0. There should be no recognizable pattern.

Polynomial model

A polynomial regression model includes quadratic or higher-degree terms.

Interaction term

An interaction term is the product of two (or more) predictor variables—for example, $x_1 x_2$.

Indicator variables

An indicator variable takes on only the value 0 or 1. If there are c categories to account for in a regression model, then the model is adjusted by adding $c-1$ indicator variables.

Model selection procedures

1. r^2: A larger r^2 indicates that the model can be used to explain more of the variation in the dependent variable. Mallows C_p: Small values, near k, indicate a good regression model.
2. Forward selection: The single most significant independent variable is added to the model at each step.
3. Backward elimination: The single least significant independent variable is eliminated from the model at each step.
4. Stepwise regression: A modification to forward selection or backward elimination. At each step in the procedure, the entire model is re-evaluated. Applied to forward selection, this method allows variables already in the model to be eliminated. Applied to backward elimination, it allows variables already eliminated from the model to be added.

Sections 12.6 and 12.7 Summary Exercises

Applications

12.245 Demographics and Population Statistics As urban areas expand and residents build more secluded homes outside of standard city and town infrastructure, owners often face inadequate fire protection due to the distance to the nearest hydrant. Many zoning ordinances and the National Fire Protection Association recommend a fire hydrant within 1000 ft of a residential dwelling. An insurance company conducted a study to investigate the relationship between claims due to fire and fire hydrants. A random sample of residential home fires was obtained, and observations were recorded for the following variables: 📊 HYDRANT

$y =$ insurance claim due to the fire, in thousands of dollars
$x_1 =$ distance to the nearest fire hydrant, in feet
$x_2 =$ water pressure at the fire hydrant, in psi

(a) Estimate the regression coefficients in the model $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + E_i$. Interpret each estimated regression coefficient.

(b) Conduct the model utility test. Use $\alpha = 0.01$. Use technology to find the exact p value associated with this test.

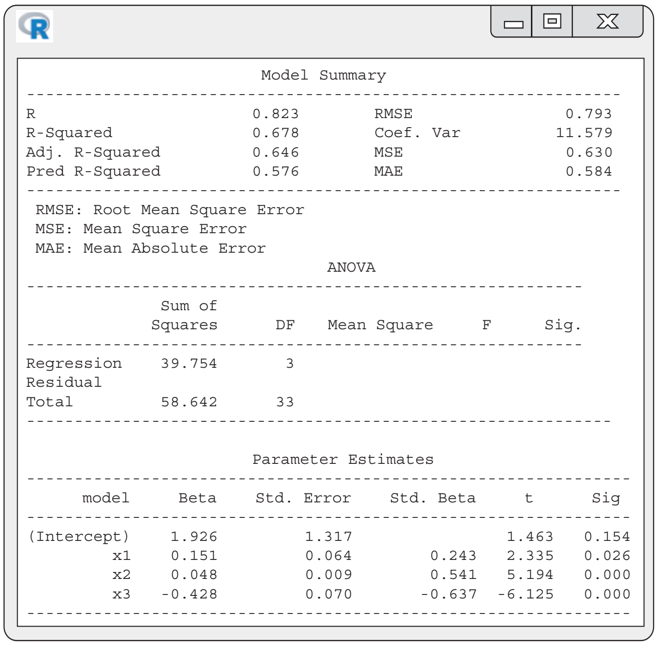
(c) Find the value of r^2 and interpret this value.

(d) Suppose a new home is constructed 750 ft away from the nearest fire hydrant and the water pressure is approximately 50 psi. If there is a fire in this home, what is the expected loss?

12.246 Manufacturing and Product Development A manufacturer of metal sheets would like to predict the springback angle from a given punch stroke in order to determine the final dimensions of the sheet accurately. A random sample of metal sheets was obtained, and observations were recorded for the following variables: 📊 METAL

$y =$ springback angle, in degrees
$x_1 =$ punch stroke, in millimeters
$x_2 =$ initial length of the sheet, in millimeters
$x_3 =$ sheet strength coefficient

The following multiple linear regression model was used: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + E_i$. R was used to analyze the data, and a portion of the output is shown here.



```
                          Model Summary
-----------------------------------------------------------------
R                          0.823       RMSE                 0.793
R-Squared                  0.678       Coef. Var           11.579
Adj. R-Squared             0.646       MSE                  0.630
Pred R-Squared             0.576       MAE                  0.584
-----------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
                                  ANOVA
---------------------------------------------------------
              Sum of
             Squares      DF    Mean Square     F      Sig.
---------------------------------------------------------
Regression    39.754       3
Residual
Total         58.642      33
-------------------------------------------------------------


                        Parameter Estimates
------------------------------------------------------------------
      model       Beta    Std. Error      Std. Beta       t      Sig
------------------------------------------------------------------
(Intercept)      1.926         1.317                  1.463    0.154
         x1      0.151         0.064          0.243   2.335    0.026
         x2      0.048         0.009          0.541   5.194    0.000
         x3     -0.428         0.070         -0.637  -6.125    0.000
------------------------------------------------------------------
```

(a) Complete the ANOVA table and conduct a model utility test. Find the exact p value.

(b) Explain the relationship between each predictor variable and the springback angle.

(c) Suppose $x^* = (5, 125, 9)$. Estimate the mean springback angle for this value of x^*.

12.247 Physical Sciences Because of movies such as *Deep Impact* and *Armageddon*, and recent meteor strikes (such as the one that occurred in Greenland in July 2018), the public has become more aware and concerned about objects striking Earth. An astronomer selected several well-documented meteor impacts on Earth at random, and, using sophisticated scientific equipment, measured values for the following variables: 📊 METEOR

$y =$ diameter of the crater created by the impact, in meters
$x_1 =$ diameter of the object, in meters
$x_2 =$ density of the object, in kg/m^3
$x_3 =$ velocity of the object, in km/s
$x_4 =$ angle of the impact, in degrees
$x_5 =$ elevation of the impact, in kilometers

(a) Use forward selection, with $\alpha = 0.05$, to find the best model. Find the estimated regression coefficients. Explain the meaning of each estimated regression coefficient.

(b) Estimate the mean crater diameter for $x^* = (50, 8, 60, 45)$. (*Note:* Your model should include x_1, x_2, x_3, and x_4.)

(c) Carefully sketch a normal probability plot. Is there any evidence to suggest that the random errors are not normal? Justify your answer.

12.248 Medicine and Clinical Studies Most prescription and over-the-counter medications have expiration dates. Under reasonable storage conditions, medication should retain at least 90% of its potency, or remain stable, if used prior to the expiration date. A consumer group recently conducted a study to predict the potency of prescription medications. A random sample of drugs was obtained, and the following variables were measured for each: MEDS

$y =$ potency, a percentage
$x_1 =$ temperature at which the drug was stored, in °F
$x_2 =$ time since the drug was manufactured, in months

(a) Estimate the regression coefficients in the model $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + E_i$.
(b) Explain the relationship between each independent variable and the potency.
(c) Construct the ANOVA table and conduct the model utility test. Find the exact p value.
(d) Conduct the necessary hypothesis tests to determine whether each regression coefficient is significantly different from 0.
(e) Check the model assumptions by constructing a normal probability plot of the residuals and the appropriate scatter plots.

12.249 Marketing and Consumer Behavior ATM cash machines are readily available throughout the United States and around the world. To satisfy customers, banks must carefully plan when to restock ATM machines with cash, and with how much. A large U.S. bank conducted a study to predict the amount of cash withdrawn by its customers at ATMs. A random sample of withdrawal transactions was obtained, and the following variables were measured for each: ATM

$y =$ amount of cash withdrawn
$x_1 =$ number of visits to an ATM in the previous month
$x_2 =$ amount of money in the user's account, in thousands of dollars
$x_3 =$ indicator variable, 1 if Friday, 0 otherwise

(a) Estimate the regression coefficients in the model $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + E_i$.
(b) Use the estimated regression coefficients to explain the relationship between each independent variable and the amount of cash withdrawn. Why do you think an indicator variable for Friday should be included in the model?
(c) Construct the ANOVA table and conduct the model utility test. Find the exact p value associated with this test.
(d) Conduct the necessary hypothesis tests to determine whether each regression coefficient is significantly different from 0.
(e) Suppose a customer has $20,000 in her account, used an ATM four times last month, and withdraws cash on a Friday evening. Find a 95% prediction interval for an observed value of the amount of cash withdrawn.
(f) Carefully sketch a normal probability plot of the residuals. Is there any evidence to suggest that the residuals are not normally distributed? Justify your answer.

12.250 Psychology and Human Behavior The bounty hunter profession is very risky, but the reward for apprehending a criminal who has jumped bail can be huge. The Miami Dade County Bail Bond Company recently conducted a study to predict the amount of time needed to apprehend a criminal based on the amount of the reward. A random sample of criminals who jumped bail was obtained, and the reward (y, in thousands of dollars) and time until apprehension (x, in days) were recorded for each. BOUNTY

(a) Carefully sketch a scatter plot of the data. Write an appropriate polynomial regression model.
(b) Find the estimated regression equation.
(c) Conduct the model utility test using $\alpha = 0.01$.
(d) Find an estimate of the time until apprehension for a criminal who has jumped bail with a reward of $50,000.
(e) Find an estimate of the time until apprehension for a criminal who has jumped bail with a reward of $200,000. Why does this estimate seem inconsistent? What error is made in using the estimated regression equation to find this estimate?

12.251 Travel and Transportation Advances in technology have made turboprop airplanes quieter and more efficient. The noise inside the cabin is caused by the aerodynamic noise, engine exhaust, engine vibration, auxiliary systems, and, most of all, propellers. A random sample of turboprop airplanes was obtained, and tests were conducted to measure the resonant frequency of the propellers (x, in thousands of rpm) and the noise level (y, in decibels, dB). The data are given in the table. TURBO

x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
60	54.5	52	56.7	44	57.9	88	52.2
91	51.8	69	54.5	96	51.4	58	46.8
89	48.3	81	51.6	75	49.1	51	43.6
61	55.6	64	53.7	81	52.9	67	51.7
80	45.2	72	53.8	70	46.3	68	47.3
46	55.9						

(a) Carefully sketch a scatter plot of the data. Consider a quadratic regression model and find the estimated regression line.
(b) Conduct the hypothesis tests with H_0: $\beta_i = 0, i = 1, 2$ and $\alpha = 0.05$. Are both regression coefficients significantly different from 0?
(c) Find a 95% confidence interval for the mean sound level when the frequency is 57,000 rpm.
(d) Estimate the frequency at which the maximum sound level occurs.
(e) Construct a normal probability plot of the residuals and a scatter plot of the residuals versus resonant frequency. Is there any evidence of a violation of the regression assumptions? Justify your answer.

12.252 Economics and Finance A study was conducted by a research team at Fidelity Investments concerning the amount of money individuals have saved for retirement. A random sample of working adults was obtained, and the percentage of income saved for retirement last year (y), age (x_1, in years), and yearly salary (x_2, in thousands of dollars) were recorded for each. The data are given in the table. RETIRE

y	x_1	x_2	y	x_1	x_2
11	38	87	11	46	64
22	29	119	5	38	49
11	59	46	17	47	114
9	27	103	12	45	76
12	29	82	13	25	110
12	48	84	14	26	102
18	60	87	12	45	41
14	40	50	13	36	60
18	35	98	15	30	99
14	36	93	14	26	79
12	47	67	17	30	75
22	57	116	16	48	69

(a) Estimate the regression coefficients in the model $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + E_i$. Conduct the model utility test with $\alpha = 0.05$.

(b) Compute the residuals and construct a normal probability plot for the residuals. Is there any evidence to suggest that the random error terms are not normal?

(c) Construct a graph of the residuals versus each predictor variable. Is there any evidence of a violation of the regression assumptions? Justify your answer.

(d) Suppose a 45-year-old is earning $80,000 per year. Find a 95% confidence interval for the mean percentage of yearly salary saved. Fidelity suggests an individual at this age and salary should save 17% of his or her yearly income. Using the confidence interval, is there any evidence to suggest that adults in this situation are not saving enough for retirement? Justify your answer.

12.253 The Eyes Have It A person's intraocular pressure (IOP) is an indicator of risk of glaucoma. Some evidence suggests that other health factors may also be related to IOP. A random sample of American male adults was obtained. Each person was subject to a variety of medical tests, and the following variables were measured: IOP

y = IOP, right eye, in mm Hg
x_1 = age, in years
x_2 = total cholesterol, in mg/dL
x_3 = high-density lipoprotein, in mg/dL
x_4 = triglyceride, in mg/dL
x_5 = body mass index, in kg/m^2

(a) Use backward elimination with $\alpha = 0.05$ to find the best multiple linear regression model.

(b) Use the estimated regression coefficients to explain the relationship between each independent variable in the final model and the IOP in the right eye of American males.

(c) Construct the ANOVA table and conduct the model utility test. Find the exact p value associated with this test.

(d) Carefully sketch a normal probability plot of the residuals. Is there any evidence to suggest that the residuals are not normally distributed? Justify your answer.

(e) Suppose an American male has the following measurements: $x = (40, 280, 42, 114, 24)$. Use your model to find a 95% confidence interval for the mean IOP in his right eye. Suppose an IOP value of 16 mm Hg or greater is a good indicator of glaucoma. Use the confidence interval to determine whether there is evidence that this person has glaucoma in his right eye.

Extended Applications

12.254 Manufacturing and Product Development Noise-canceling headphones/earphones have become very popular for ordinary use as well as in cars, in planes, and even at the office. These devices block out external noise by using active noise-reduction technology. A study was conducted to predict the percentage of sound eliminated. A random sample of headphones was selected. Each was subject to a noise-canceling test, and the following variables were measured: NOISE

y = percentage of sound eliminated
x_1 = impedance, in ohms
x_2 = sensitivity, in decibels, dB
x_3 = driver units, in millimeters
x_4 = noise control range at 300 Hz, in dB
x_5 = form factor, 0 = earbud, 1 = headphone

(a) Use forward selection to find the best multiple linear regression model.

(b) If you were to advise someone interested in buying a set of noise-canceling headphones, which characteristics would you recommend? Why?

(c) Estimate the true mean percentage of sound reduction when $x = (18, 121, 40, 15, 1)$.

(d) Construct a normal probability plot for the residuals. Is there any evidence to suggest that the random error terms are not normally distributed? Justify your answer.

(e) Delete observation 21 (82, 25, 122, 40, 21, 0) from the data set. Use forward selection on this reduced data set to find the best multiple linear regression model. Explain any differences between this model and the one in part (a).

12.255 Public Health and Nutrition Many breakfast cereal companies advertise brands with high fiber, which can lower the risk of heart disease, cancer, and diabetes. The U.S. Food and Drug Administration recently conducted research to predict the amount of fiber in one cup of various breakfast cereals. A random sample of cereals was obtained, and the following variables were measured for each one-cup serving: CEREAL

y = fiber, in grams
x_1 = calories
x_2 = fat, in grams
x_3 = protein, in grams

x_4 = carbohydrates, in grams
x_5 = sodium, in milligrams
x_6 = calcium, in milligrams

(a) Use backward elimination to find the best multiple linear regression model.
(b) Use forward selection to find the best multiple linear regression model. Compare this model with the one in part (a). Which do you think is better? Why?
(c) Consider a cup of Frosted Mini-Wheats with $x = (151.5, 1, 4, 36, 1.5, 15)$. Estimate the mean amount of fiber in one cup of this cereal using both models. If the true amount of fiber is 4 g, which model is better?
(d) Use the best subsets approach to find a multiple linear regression model. Construct a graph of r^2 versus various models.

12.256 Manufacturing and Product Development Aardvark, with stores in Santa Ana, California, and Las Vegas, Nevada, sells a wide variety of equipment and supplies used to make ceramics. Researchers at this company recently collected data in an attempt to predict the fired shrinkage of various types of clay. The following measurements were recorded for each randomly selected clay sample: CLAY

y = percent fired shrinkage
x_1 = water absorption, percent
x_2 = pH
x_3 = indicator variable: 0 = dark clay, 1 = light clay

(a) Estimate the regression coefficients in the model $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + E_i$.
(b) Conduct a model utility test and the other appropriate tests to determine which variables are the most important predictors of fired shrinkage.
(c) Consider a new, reduced regression model as a result of the hypothesis tests in part (b). Estimate the regression coefficients in this model.
(d) Using the reduced model, estimate the mean fired shrinkage for $x_1 = 3.5$ in dark clay, and for $x_1 = 5.5$ in light clay.

12.257 Biology and Environmental Science When the last ice age ended, retreating glaciers in Norway created deep valleys that became filled with sea water. These fjords are often very deep and characterized by steep sides. In a recent study, researchers investigated the density of *Calanus finmarchicus* (the dominant species in the zooplankton biomass) in fjords as a function of the distance from the coast. A random sample of west coast Norwegian fjords was obtained. The following measurements were recorded for each fjord:

y = zooplankton density, in mg/m^2
x_1 = distance from the outer coast, in kilometers
x_2 = indicator variable for season

(a) Consider a regression model with the appropriate number of indicator variables. Conduct the model utility test ($\alpha = 0.05$).
(b) Consider a regression model to predict $\ln(y)$ using $\ln(x_1)$ and the appropriate number of indicator variables. Find the estimated regression equation. Conduct the model utility test ($\alpha = 0.05$) and explain the meaning of each estimated regression coefficient.
(c) Construct a normal probability plot of the residuals from the model in part (b). Is there any evidence to suggest that the residuals are not normally distributed?
(d) Using the model in part (b), find an estimate of the mean density of zooplankton during the summer in a fjord that is 150 km from the coast.

12.258 Travel and Transportation As towns and municipalities add or improve streets, one of the biggest concerns is safety. There is an undocumented theory that narrow streets are safer than wider streets. Research was conducted to predict the safety of town streets as a function of several characteristics. A random sample of streets from all across the country was obtained, and yearly accident reports were examined. To focus on street characteristics, accidents caused by road conditions (wet, icy, or snow covered), substance abuse, or traffic volume were eliminated from the study. The following variables were recorded for each street: STREET

y = accidents per mile per year
x_1 = degree of curvature of the street
x_2 = street width, in feet
x_3 = curb type: 0 = none, 1 = 6 inch vertical, 2 = modified
x_4 = tree density: trees per 1000 feet along the street
x_5 = number of traffic lights per day, in thousands
x_6 = mean number of vehicles per day, in thousands
x_7 = parking density: parking spaces per mile

(a) Consider a regression model with six predictors and two indicator variables (for curb type). Find the estimated regression equation. Which variables do you think are significant? Justify your answer.
(b) Use forward selection to find the best model ($\alpha = 0.10$). Compare the significant predictors with those identified in part (a).
(c) Use backward elimination to find the best model ($\alpha = 0.10$)
(d) Use the best model to explain the relationship between each significant predictor variable and the accidents per mile per year.
(e) Construct a normal probability plot of the residuals. Is there any evidence that the residuals are not normal? Justify your answer.

12.259 Psychology and Human Behavior Many factors affect how optimistic a person is about life in general. Family situation, relationships, and even the weather probably all have an effect. Recent research suggests that a person's optimism might be affected by the ceiling height in the home. Contractors and real estate agents frequently find it easier to sell homes with higher ceilings. A random sample of American adults living in the northeast was obtained, and each was asked to complete a detailed survey that resulted in an Optimism Score (y), a number from 1 to 100 that measures the level of optimism people feel about themselves and the future. A large number suggests increased optimism. In addition to this score, the following variables were recorded for each individual: OPTIM

x_1 = height of home ceiling, in feet
x_2 = total living area of home, in ft^2
x_3 = temperature of the thermostat in the winter, in °F
x_4 = color of the walls in the main family room, 0 = dark, 1 = light

(a) Use the techniques discussed in these two sections to find the best model to predict the optimism score.

(b) Construct a normal probability plot of the residuals. Is there any evidence that the residuals are not normal?

(c) Based on your model, how does an increase in ceiling height of 1 ft change the optimism score?

Challenge Problems

12.260 Biology and Environmental Science Consumption advisories all over the country warn people about eating certain species of fish because of elevated levels of mercury and other contaminants. Some state fishing regulations suggest that individuals should not have more than one or two meals per month of walleye or trout caught in certain lakes and rivers, because of high levels of mercury in these species. The U.S. Geological Survey (USGS) is conducting a study to predict the mercury level in smallmouth bass in the Susquehanna River. Five locations (USGS stations) were used, and the following measurements were recorded for each bass: FISH

y = level of mercury, in parts per million, ppm
x_1 = river flow rate, in thousands of cubic feet per second
x_2 = temperature of the water, in °C
x_3 = length of the fish, in inches
x_4 = weight of the fish, in pounds
x_5 = pH of the water
x_6 = USGS station number

(a) Use the techniques discussed in this chapter to find the best model to predict the level of mercury in a smallmouth bass. Consider interaction terms and polynomial models, and be sure to use the appropriate indicator variables.

(b) Use your model to suggest the best location to fish for smallmouth bass—that is, the spot on the river where fish tend to have the lowest levels of mercury.

Notes

1 Fluoridealert.org, accessed on December 26, 2018, http://fluoridealert.org/studies/brain01/

2 The Conversation, accessed on December 26, 2018, http://theconversation.com/diet-soda-may-be-hurting-your-diet-96181

3 Healthline, accessed on December 26, 2018, https://www.healthline.com/health-news/heres-how-nicotine-affects-the-body#1

4 Medical Press, accessed on December 26, 2018, https://medicalxpress.com/news/2018-03-combating-childhood-obesity-fatty-liver.html