Chapter 9
Multiple Regression—More Than One Predictor

In Chapter 8 we introduce the algebra and geometry behind the fitting of a linear model relating a single response variable to one or more explanatory (predictor) variables using the criterion of least squares. In this chapter we consider in more detail situations where there are two or more predictors.

The two linear modeling techniques we have studied so far, regression in Chapter 8 and analysis of variance in Chapter 6, have much of their mathematics interpretation in common. In this chapter we explore the common mathematical features, with some examples of how they apply. In the following chapters we use this common structure.

We begin by extending the Chapter 8 discussion of regression with a single predictor (simple regression) to allow for two or more predictors. *Multiple* regression refers to regression analysis with at least two predictors. There is another term *multivariate regression* which refers to situations with more than one response variable. We do not discuss multivariate regression in this book.

9.1 Regression with Two Predictors—Least-Squares Geometry

The graphics for least squares with two x-variables, and in general for more than two x-variables, are similar to the graphics in Figure 8.2. We will work with two x-variables, `abdomin` and `biceps`, from the `data(fat)` dataset we used in Chapter 8. In the three snapshots of the basic 3-dimensional plot in Figure 9.1, `bodyfat` is plotted as y against the other two variables as x_1 and x_2.

The response variable is placed on the vertical dimension and the two x-variables `biceps` and `abdomin` define the horizontal plane. The red and green dots at the observations show the three-dimensional location of the observed points. Positive residuals are shown as green dots above the least-squares plane and are connected

R.M. Heiberger, B. Holland, *Statistical Analysis and Data Display*,
Springer Texts in Statistics, DOI 10.1007/978-1-4939-2122-5_9

to the fitted value on the plane by a green residual line. The green residual line forms one edge of the square. Negative residuals are shown as red dots below the least-squares plane and are connected to the fitted value on the plane by a red residual line. The red residual line forms one edge of the square.

The least-squares plane minimizes the sum of the squared areas. The displayed squares are the squares whose sum has been minimized by the least-squares process. The view in the right panel is from above the plane. It shows `biceps` coming out of the page and `abdomin` going into the page. The view in the center panel is from a point that is on the least-squares plane. The view in the left panel is from below the least-squares plane. Variable `biceps` is coming out of the page and variable `abdomin` is almost along the page. The code in file `HHscriptnames(9)` constructs an interactive 3-d version of this plot. We selected these specific static snapshots from the interactive plot.

We think of this plot as a point cloud in 3-space floating over the surface defined by the x-variables. Any plane other than the least-squares plane will show a larger sum of squared areas than the least-squares plane illustrated here.

9.2 Multiple Regression—Two-*X* Analysis

The specification of the analysis for two x-variables is similar to that for one x-variable. The sequential ANOVA table and the table of coefficients for a two x-variable analysis of the body fat data `data(fat)` are in Table 9.1.

Since both predictors are significantly different from 0, the arithmetic justifies the illustration in Figure 9.1, where we see from the regression plane that $\hat{y}$ changes linearly with changes in either x_1 and x_2. The table of coefficients tells us that on average for this population, percent body fat increases by 0.683 if abdomen circumference increases by one cm and biceps is unchanged, and percent body fat decreases by .922 if biceps increases by one cm while abdomin is unchanged.

The t-value for `biceps` (the second variable in the ANOVA table) is related to the F-value for `biceps`: $t^2 = (-2.946)^2 = 8.677 = F$. The t-value (8.693) for `abdomin` (the first variable in the ANOVA table) is not simply related to the correspondingly labeled F-value (101.172). We investigate this relationship in the discussion of Table 13.27.

Figure 9.2 shows the diagnostics from the two-X regression model of Section 9.2. Compare this to the similar plot for one-X regression in Figure 8.6.

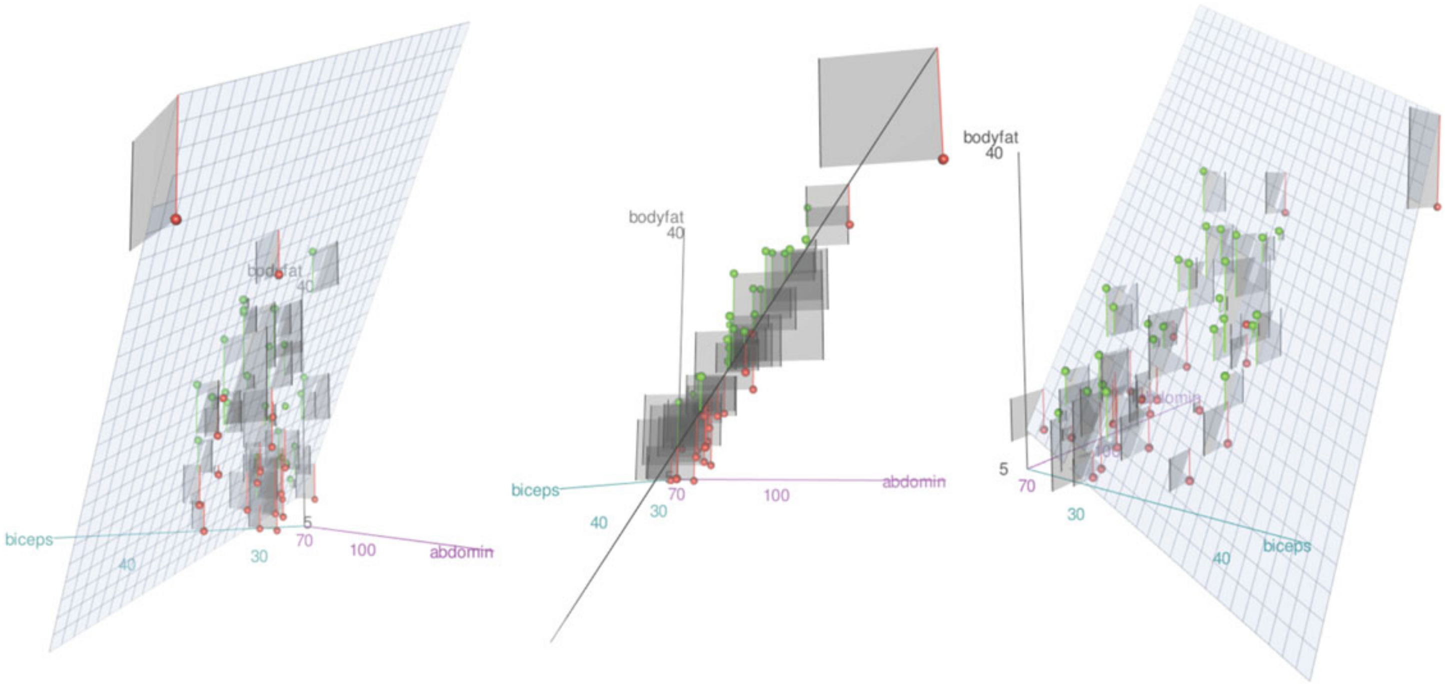

Fig. 9.1 Bodyfat data with one response variable `bodyfat` and two predictor variables `biceps` and `abdomin`. This figure is three static screenshots of an interactive 3-d plot. The plot is the two-*x* extension of the display of squared residuals we introduced in Figure 8.2. The center panel shows a viewpoint aligned with the least-squares plane. The left panel shows a rotation to the left and we see the negative residuals under the plane and the partially occluded positive residuals through the plane. The right panel shows a rotation to the right and we see the positive residuals above the plane and the partially occluded negative residuals through the plane. See the text for the detailed description of the three panels. See Table 9.1 for the corresponding ANOVA table.

Table 9.1 Sequential ANOVA table and table of regression coefficients from the two-x model with y=bodyfat, x_1=abdomin, and x_2=biceps. See Figure 9.1.

```
> fat2.lm <- lm(bodyfat ~ abdomin + biceps, data=fat)

> anova(fat2.lm)
Analysis of Variance Table

Response: bodyfat
          Df Sum Sq Mean Sq F value  Pr(>F)
abdomin    1   2440    2440  101.17 5.6e-13 ***
biceps     1    209     209    8.68  0.0051 **
Residuals 44   1061      24
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

> summary(fat2.lm)

Call:
lm(formula = bodyfat ~ abdomin + biceps, data = fat)

Residuals:
    Min      1Q  Median      3Q     Max
-11.252  -3.674   0.716   3.771  10.241

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.5937     6.6922   -2.18   0.0346 *
abdomin       0.6829     0.0786    8.69  4.2e-11 ***
biceps       -0.9222     0.3130   -2.95   0.0051 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 4.91 on 44 degrees of freedom
Multiple R-squared:  0.714,Adjusted R-squared:  0.701
F-statistic: 54.9 on 2 and 44 DF,  p-value: 1.1e-12
```

9.3 Multiple Regression—Algebra

Everything in simple regression analysis carries over to multiple regression. There are additional issues that arise because we must also study the relations among the predictor variables. The algebra for multiple regression is most easily expressed in matrix form. (A brief introduction to matrix algebra appears in Appendix I.) The formulas for simple regression can be derived as the special case of multiple regression with $p = 1$.

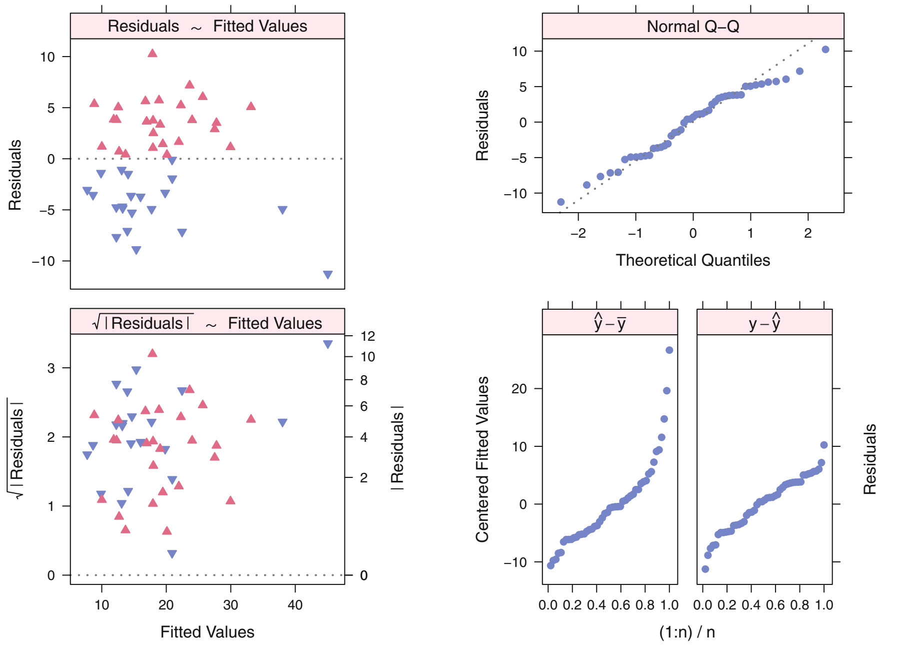

Fig. 9.2 Diagnostics for `lm(bodyfat ~ abdomin + biceps, data=fat)`. Compare this to the similar plot for one-X regression in Figure 8.6.

Assume

$$\underset{n\times1}{Y} = \underset{n\times(1+p)}{X}\ \underset{(1+p)\times1}{\beta} + \underset{n\times1}{\epsilon} \tag{9.1}$$

or equivalently

$$y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i \quad \text{for } i = 1, \ldots, n \tag{9.2}$$

where

- $\underset{n\times1}{Y}$ are observed values,
- $\underset{n\times(1+p)}{X} = [\mathbf{1}\, X_1 X_2 \ldots X_p\,]$ are observed values with $\underset{n\times1}{\mathbf{1}}$ representing the constant column with 1 in each row and $\underset{n\times1}{X_j}$ indicating the column with X_{ij} in the i^{th} row,
- $\underset{(1+p)\times1}{\beta}$ are unknown constants,
- $\underset{n\times1}{\epsilon} \sim N(0, \sigma^2 I)$ are independent.

Then the least-squares estimate $\widehat{\beta}$ is obtained by minimizing the sum of squared deviations

$$S = (Y - X\beta)'(Y - X\beta) = \sum_{i=1}^{n}\Big(y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip})\Big)^2$$

by taking the derivatives $(\partial S/\partial\beta_j)$ with respect to all the β_j and setting them to 0. The resulting set of equations, called the *Normal Equations* and generalizing Equation 8.6,

$$(X'X)\widehat{\beta} = (X'Y) \tag{9.3}$$

are solved for $\widehat{\beta}$. The solution [equivalent to Equation 8.7] is equal to

$$\widehat{\beta} = (X'X)^{-1}(X'Y) = \Big((X'X)^{-1}X'\Big)\, Y = (X^{+})Y \tag{9.4}$$

The symbol $X^{+} \overset{\text{def}}{=} (X'X)^{-1}X'$ is the notation for the Moore–Penrose *generalized inverse* of any rectangular matrix. In the special case of square invertible matrices the generalized inverse becomes the familiar matrix inverse. We introduce this notation here because it simplifies the appearance of the equations. We start with the model $Y = X\beta + \epsilon$ in Equation (9.1) and conclude with the estimate $\hat{\beta} = X^{+}Y$ in Equation (9.4). We effectively moved the X to the other side and replaced the ϵ with the hat on the β. Note that Equation (9.4) is an identity, but neither efficient nor numerically stable as a computing algorithm. An efficient algorithm uses Gaussian elimination to solve the equations directly. See Section I.4.7 for further discussion on efficient computation.

We construct the fitted values with

$$\widehat{Y} = X\widehat{\beta} = \Big(X(X'X)^{-1}X'\Big)Y = HY \tag{9.5}$$

where the matrix

$$H \overset{\text{def}}{=} X(X'X)^{-1}X' \tag{9.6}$$

is a projection matrix. The sum of squares (SS) for the regression is $\mathsf{SS}_{\text{Reg}} = Y'HY$. The projection matrix H is called the *hat matrix* because multiplying H by Y places a hat '$\widehat{\ }$' on Y. We can see that $H_{ij} = \partial\widehat{Y}_i/\partial Y_j$. We discuss the hat matrix in Section 9.3.1.

The *residuals* are defined as the difference

$$e = Y - \widehat{Y} = (I - H)Y \tag{9.7}$$

between the observed values Y and the fitted values $\widehat{Y}$. With least-squares fitting, the residuals are orthogonal to the observed x-values

$$e'X = 0 \tag{9.8}$$

and therefore to the fitted values

$$e'\hat{Y} = e'X\hat{\beta} = 0 \tag{9.9}$$

The variance–covariance matrix of the residuals e is $\sigma^2(I-H)$. Note in particular that $\text{var}(e_i) = \sigma^2(1 - H_{ii})$ is not constant for all i. As a consequence the confidence bands in Figure 8.5 are not parallel to the regression line, but instead have a minimum width at the mean of the x values.

An unbiased estimator of σ^2 is

$$s^2 = \frac{Y'(I-H)Y}{n-p-1} = \mathsf{MS}_{\text{Res}} = \mathsf{SS}_{\text{Res}}/\mathsf{df}_{\text{Res}} \tag{9.10}$$

Its square root, s, sometimes called the standard error of estimate, is an asymptotically unbiased estimator of σ. As in the case of simple regression, the sum of the residuals is zero, that is,

$$\sum_{i=1}^{n} e_i = \mathbf{1}'e = 0 \tag{9.11}$$

where $\mathbf{1}'$ is a row vector of ones. The proof of this assertion is requested in Exercise 9.1.

Both $\widehat{\beta}$ and $\widehat{Y}$ are linear combinations of y_i. The y_i are independent because the ϵ_i are independent. Hence the elementary theorems

$$E(a_1y_1 \pm a_2y_2) = a_1E(y_1) \pm a_2E(y_2) \tag{3.8}$$

and

$$\text{var}(a_1y_1 \pm a_2y_2) = a_1^2\,\text{var}(y_1) + a_2^2\,\text{var}(y_2) \tag{3.9}$$

are applicable. These are where we get Equation (8.15), the standard error for β_1, the corresponding formula

$$\text{var}(\hat{\beta}) = \sigma^2(X'X)^{-1} \tag{9.12}$$

for the estimator of β in Equation (9.4), and formulas (9.24) and (9.25) for tests and confidence intervals about $E(Y|X)$ and for prediction intervals about Y for new values of X.

9.3.1 The Hat Matrix and Leverage

The hat matrix in Equation (9.6) is called that because premultiplication by H places a hat '$\frown$' on Y: $\hat{Y} = HY$. The i^{th} diagonal of H is called the leverage of the i^{th} case because it tells how changes in Y_i affect the location of the fitted regression line, specifically:

$$\frac{\partial \widehat{Y_i}}{\partial Y_i} = H_{ii} \tag{9.13}$$

If $\left(H_{ii} \quad > \quad 2(p+1)/n\right)$, then the i^{th} point is called a high leverage point. See Section 11.3.1. Equation (9.13) shows that changes in the observed Y_i-value of high leverage points have a large effect on the predicted value $\hat{Y}_i$, that is, they have a large effect on the location of the fitted regression plane.

The hat matrix is used in regression diagnostics, that is, techniques for evaluating how the individual data points affect the regression analysis. Many diagnostics are discussed in Section 11.3.

Frequently these diagonals of H are denoted by $h_i = H_{ii}$. They are calculated in R with the command `hat(X)`.

A specific formula for the leverage h_i itself is almost simple:

$$h_i = X_{i\cdot}(X'X)^{-1}X'_{i\cdot} \qquad \text{where } X_{i\cdot} \text{ is the } i^{\text{th}} \text{ row of } X \tag{9.14}$$

In an alternate but common notation, the predictor matrix does not include the column **1**. To avoid excessive confusion, define Z to be all the columns of X except the initial column **1**:

$$\underset{n\times p}{Z} = [X_1 X_2 \dots X_p\,]$$

and let

$$\bar{Z} = (\bar{X}_1 \bar{X}_2 \dots \bar{X}_p)$$

In this notation the formula for leverage looks worse:

$$h_i = \frac{1}{n} + (Z_{i\cdot} - \bar{Z})\left((Z - \mathbf{1}\bar{Z})'(Z - \mathbf{1}\bar{Z})\right)^{-1}(Z_{i\cdot} - \bar{Z})' \tag{9.15}$$

The term $\frac{1}{n}$ in Equation (9.15), with the Z matrix which excludes the column **1**, is not needed in Equation (9.14), with the X matrix which includes the column **1**. In simple regression, with $Z = X_1 = x$, formula (9.15) simplifies to Equation (8.19)

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} \tag{8.19}$$

9.3.2 Geometry of Multiple Regression

Several types of pictures go along with multiple regression. We have already looked at the scatterplot matrix, drawn with the R command `splom(data.frame)`; for example, see Figure 8.1 for the `splom` of the body fat dataset `fat`.

The picture that goes best with the defining least-squares equations is the multidimensional point cloud. It is easiest to illustrate this with Y and two X-variables. See Figures 8.2 and 9.1 for one-X and two-X examples.

A similar construction is in principle possible for more X-variables. Illustrating the projection of four or more dimensions onto a two-dimensional graph is difficult at best.

9.4 Programming

9.4.1 Model Specification

We use several notations for the specification of a regression model to a computer program. How are the statements constructed in each notation, and what are their syntax and their semantics?

For specificity, let us look at a linear regression model with a response variable y and two predictor variables x_1 and x_2. We express this model in several equivalent notations. In the algebraic notation of Section 9.3, we have

$$\underset{n\times 1}{Y} = \underset{n\times(1+2)}{X} \quad \underset{(1+2)\times 1}{\beta} + \underset{n\times 1}{\epsilon} \tag{9.16}$$

or equivalently

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \quad \text{for } i = 1, \ldots, n \tag{9.17}$$

In R model formula notation, we have

```
y ~ x1 + x2                                   (9.18)
```

In SAS model statement notation (with the space character indicating the formulaic sum), we have

```
y = x1   x2                                   (9.19)
```

In both computer languages the statement is read, "y is modeled as a linear function of x_1 and x_2."

The four statements (9.16)–(9.19) are equivalent. Both computational specifications remove the redundancy in notation used by the traditional scalar algebra notation. The program knows that the variables (`y`, `x1`, and `x2`) have length `n`; there is no need to repeat that information. All linear model specifications have regression coefficients, and most have a constant term (we discuss models without a constant term in Section 9.8); there is no need to specify the obvious. There is always an error term because the model does not fit the data exactly; there is no need to specify the error term explicitly. The two pieces of information unknown to the program are

- Which variable is the response and which are the predictors. This is indicated positionally—the response is on the left, and notationally—the "~" or "=" separates the response from the predictors. A separation symbol is needed because the same notation can be generalized to express multiple response variables.
- The relationship between the predictors. R indicates summation explicitly with the "+" and SAS indicates it implicitly by leaving a space between the predictor variable names. Other relationships, for example crossing or nesting (to be discussed beginning in Section 13.5), are indicated by other algebraic symbols as indicated in Table 13.18.

The interpretation of operator symbols in the model specification notation is related to, but not identical to, the interpretation of the same symbols in an ordinary algebra statement. The model formulas (9.18) and (9.19) mean:

find the coefficients $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ that best fit

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\epsilon}_i \tag{9.20}$$

for the observed values (y_i, x_{i1}, x_{i2}) for all i: $1 \le i \le n$.

The "+" and space " " in formulas (9.18) and (9.19) do not have the ordinary arithmetic sense of $x_{i1} + x_{i2}$.

9.4.2 Printout Idiosyncrasies

The R `summary` and `anova` functions do not print the `Total` line in their ANOVA tables.

SAS `PROC GLM` uses the name "Type I Sum of Squares" for the sequential ANOVA table. See the discussion of sums of squares types in Section 13.6.1.

9.5 Example—Albuquerque Home Price Data

9.5.1 Study Objectives

Realtors can use a multiple regression model to justify a house selling price based on a list of desirable features the house possesses. Such data are commonly compiled by local boards of realtors. We consider a data file containing a random sample of 117 home sales in Albuquerque, New Mexico during the period February 15 through April 30, 1993, taken from Albuquerque Board of Realtors (1993).

9.5.2 Data Description

We use a subset of five of the eight variables for which data are provided, and 107 of the 117 houses that have information on all five of these variables.

`price:` Selling price in $100's

`sqft:` Square feet of living space

`custom:` Whether the house was built with custom features (1) or not (0)

`corner:` Whether the house sits on a corner lot (1) or not (0)

`taxes:` Annual taxes in $

We investigate models of `price` as a function of some or all of the candidate predictors `sqft`, `custom`, `corner`, and `taxes`. This example assumes that `taxes` potentially determine `price`. In some real estate contexts the causality could work in the opposite direction: selling prices can affect subsequent home appraisals and hence tax burden.

9.5.3 Data Input

The data are accessed with `data(houseprice)` and looked at initially with the scatterplot matrices in Figures 9.3 and 9.4. Two of the four candidate predictors, `custom` and `corner`, are dichotomous variables, and the panels involving them in Figure 9.3 are wasteful of space and not very informative. Figure 9.4, with separate superpanels for the two values of `corner` and separate plot symbols for the two values of `custom`, displays the information much more efficiently. We learn from these figures that custom houses tend to have higher prices than regular houses, and corner houses have different patterns of relationships between `price` and the continuous predictors than middle houses.

Figure 9.4 suggests that price is directly related to all four candidate predictors. We proceed with the analysis by regressing `price` on the four variables in Table 9.2. In this Table we examine the signs of the regression coefficients and the magnitudes of their p-values. We see that `price` is strongly positively associated with `sqft`, `taxes` and `custom` (as opposed to regular) houses. Such conclusions are consistent with common knowledge of house valuation. The predictor `corner` has a marginally significant negative coefficient. Hence there is moderate evidence that, on average, corner houses tend to be lower priced than middle houses.

The magnitudes of the regression coefficients also convey useful information. For example, on average, each additional square foot of living space corresponds to a $0.2076 \times \$100 = \$20.76$ increase in price, and on average custom houses sell

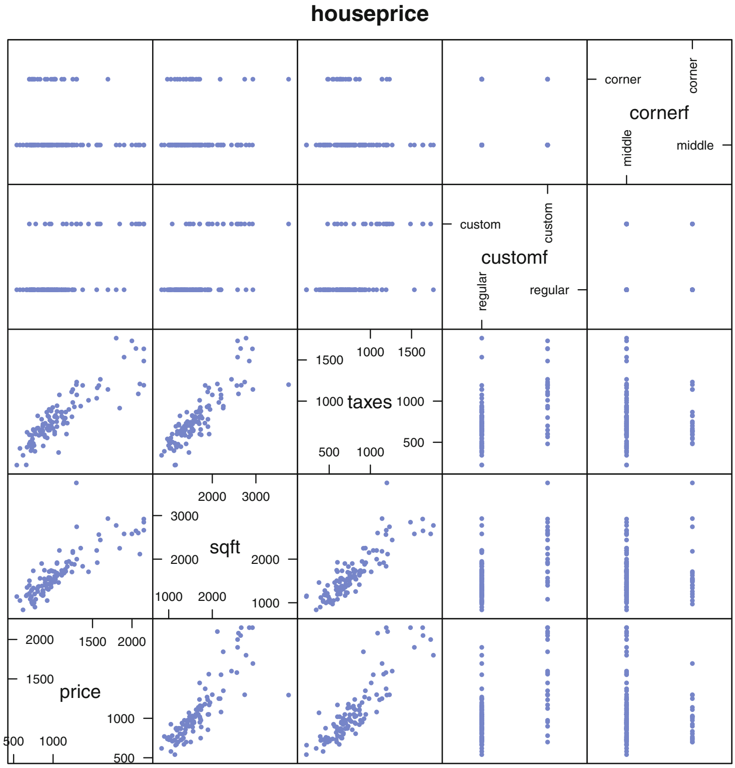

Fig. 9.3 House-price data. The discreteness of variables `customf` and `cornerf` decreases the informativeness of this splom, particularly the panel for this pair of variables. Figure 9.4 is a preferred splom presentation of these data.

for 156.81481 × \$100 = \$15,681.48 more than regular houses. The $R^2 = 0.8280$ says that in the population of houses from which `data(houseprice)` is a random sample, 82.8% of the variability in price is accounted for by these four predictors.

9.6 Partial *F*-Tests

Sometimes we wish to examine whether two or more predictor variables *acting together* have a significant impact on the response variable. For example, suppose we consider the house-price data of Section 9.5 with four candidate predictors, `sqft`, `custom`, `corner`, and `taxes`, and wish to examine if `custom` and `corner` together have a significant impact on `price`, above and beyond the impacts of `sqft` and `taxes`. R (in Table 9.3) approaches this by direct comparison of two models. The *full model* contains all predictors under consideration. The *reduced*

Table 9.2 Analysis of variance table for house-price data.

```
> houseprice.lm2 <- lm(price ~ sqft + taxes + custom + corner,
+                       data=houseprice)

> anova(houseprice.lm2)
Analysis of Variance Table

Response: price
           Df   Sum Sq  Mean Sq F value  Pr(>F)
sqft        1 11102445 11102445  421.34 < 2e-16 ***
taxes       1  1374474  1374474   52.16 9.5e-11 ***
custom      1   350716   350716   13.31 0.00042 ***
corner      1   114215   114215    4.33 0.03985 *
Residuals 102  2687729    26350
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

> summary(houseprice.lm2)

Call:
lm(formula = price ~ sqft + taxes + custom + corner,
   data = houseprice)

Residuals:
   Min     1Q Median     3Q    Max
-544.6  -99.5   -4.8   64.8  510.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  175.166     56.312    3.11  0.00242 **
sqft           0.208      0.061    3.40  0.00096 ***
taxes          0.677      0.101    6.70  1.2e-09 ***
custom       156.815     44.495    3.52  0.00064 ***
corner       -83.401     40.059   -2.08  0.03985 *
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 162 on 102 degrees of freedom
Multiple R-squared:  0.828,Adjusted R-squared:  0.821
F-statistic:  123 on 4 and 102 DF,  p-value: <2e-16
```

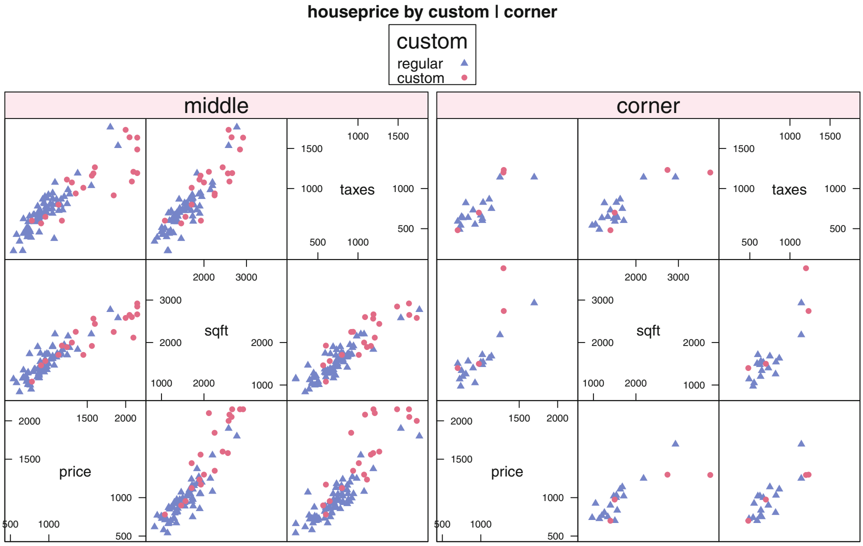

Fig. 9.4 Albuquerque house-price data. Custom houses go for higher prices than regular houses. Corner houses have a different pattern than middle houses.

Table 9.3 Partial F-tests of H_0: $\beta_{\text{custom}} = \beta_{\text{corner}} = 0$ using the R `anova()` function with two linear models as arguments.

```
> houseprice.lm1 <- lm(price ~ sqft + taxes, data=houseprice)

> anova(houseprice.lm1, houseprice.lm2)
Analysis of Variance Table

Model 1: price ~ sqft + taxes
Model 2: price ~ sqft + taxes + custom + corner
  Res.Df     RSS Df Sum of Sq    F  Pr(>F)
1    104 3152660
2    102 2687729  2    464931 8.82 0.00029 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

model contains all predictors apart from the ones we test in order to see if then can be eliminated from the model. *Partial F* refers to the fact that we are simultaneously testing *part* of the model’s predictors, not all predictors but perhaps more than just one of them. The idea behind this test is apparent from Table 9.3. The F-test examines whether the reduction in residual sum of squares as a result of fitting the more elaborate model is a significant reduction. This assessment is performed by measuring the *extra sum of squares*, defined as

$$\text{(residual SS from reduced model)} - \text{(residual SS from full model)} \tag{9.21}$$

against the residual sum of squares from the full model. The degrees of freedom associated with the extra sum of squares equals the number of parameters being tested for possible elimination.

The general form of the test is

$$F = \frac{\text{(extra SS)/(df associated with extra SS)}}{\text{(full model residual SS)/(df associated with full model residual SS)}} \tag{9.22}$$

The strategy of this approach is used whenever one wishes to compare the fits of two linear models, one of which has the same terms as the other plus at least one more term.

For testing the hypothesis that the population regression coefficients of `custom` and `corner` are both equal to 0, we see that the F-statistic is 8.82 on 2 and 102 degrees of freedom. There are two numerator degrees of freedom because the null hypothesis involves constraints on two model parameters. The very small p-value strongly suggests that this null hypothesis is false. We conclude that at least one of `custom` and `corner` is needed in the model.

The preceding discussion assumes that `sqft` and `taxes` were already in the model. It is also possible to test the combined effect on `price` of `custom` and `corner` compared with no other predictors, or exactly one of the predictors `sqft` and `taxes`. However, we do not pursue these possibilities here.

9.7 Polynomial Models

If the relationship between a response Y and an explanatory variable X is believed to be nonlinear, it is sometimes possible to model the relationship by adding an X^2-term to the model in addition to an X-term. For example, if Y is product demand and X is advertising expenditure on the product, an analyst might feel that beyond some value of X there is "diminishing marginal returns" on this expenditure. Then the analyst would model Y as a function of X, X^2, and possibly other predictors, and anticipate a significant negative coefficient for X^2. Occasionally a need is encountered for higher-order polynomial terms.

An example from Hand et al. (1994), original reference Williams (1959), is `data(hardness)` which we first encountered in Exercise 4.5. In this section we investigate the modeling of `hardness` as a quadratic function of `density`. We pursue this analysis in Exercise 11.2 from another angle, a transformation of the response variable `hardness`.

Hardness of wood is more difficult to measure than density. Modeling hardness in terms of density is therefore desirable. These data come from a sample of Australian Janka timbers. The Janka hardness test measures the resistance of a sample of wood to denting and wear. A quadratic model fits these data better than a linear model. An additional virtue of the quadratic model is that its intercept term differs insignificantly from zero; this is not true of a model for these data containing only a linear term. (If wood has zero hardness, it certainly has zero density.)

The fitted quadratic model in Table 9.4 is

$$\texttt{density} = -118.007 + 9.4340\,\texttt{hardness} + 0.5091\,\texttt{hardness}^2$$

The regression coefficient for the quadratic term is significantly greater than zero, indicating that the plot is a parabola opening upwards as shown in Figure 9.5. The p-value for the quadratic regression coefficient is identical to the p-value for the quadratic term in the ANOVA table because both tests are for the marginal effect of the quadratic term assuming the linear term is already in the model. The two p-values for the linear term differ because they are testing the linear coefficient in two different models. The p-value for linear regression coefficient assumes the presence of a quadratic term in the model, but the linear p-value in the sequential ANOVA table addresses a model with only a linear component.

When fitting a truly quadratic model, it is necessary to include the linear term in the model even if its coefficient does not significantly differ from zero unless there is subject area theory stating that the relationship between the response and predictor lacks a linear component.

The regression coefficients of the x^2 term are difficult to interpret. An interpretation should be done with the coefficients of the orthogonal polynomials shown in Table 9.5, not the simple polynomials of Table 9.4. See Section 10.4 for further discussion.

Table 9.4 Quadratic regression of hardness data. The quadratic term, with p=.0027, is very important in explaining the curvature of the observations. See Figure 9.5 to compare this fit with the linear fit. Compare the regression coefficients here with the regression coefficients in Table 9.5 where we use the orthogonal quadratic polynomial, rather than the simple square, for the quadratic regressor.

```
> data(hardness)

> hardness.lin.lm  <- lm(hardness ~ density,
+                        data=hardness)

> anova(hardness.lin.lm)
Analysis of Variance Table

Response: hardness
          Df   Sum Sq  Mean Sq F value Pr(>F)
density    1 21345674 21345674     637 <2e-16 ***
Residuals 34  1139366    33511
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

> hardness.quad.lm <- lm(hardness ~ density + I(density^2),
+                        data=hardness)

> anova(hardness.quad.lm)
Analysis of Variance Table

Response: hardness
             Df   Sum Sq  Mean Sq F value Pr(>F)
density       1 21345674 21345674   815.9 <2e-16 ***
I(density^2)  1   276041   276041    10.6 0.0027 **
Residuals    33   863325    26161
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

> coef(summary.lm(hardness.quad.lm))
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -118.0074   334.9669 -0.3523 0.726857
density         9.4340    14.9356  0.6316 0.531970
I(density^2)    0.5091     0.1567  3.2483 0.002669
```

Table 9.5 Quadratic regression of hardness data with orthogonal polynomials. The quadratic term, with p=.0027, is very important in explaining the curvature of the observations. See Figure 9.5 to compare this fit with the linear fit. In this fit with the orthogonal polynomial for the quadratic term, the regression coefficient for the linear term is identical to the regression coefficient in the simple linear regression. Compare to the very different regression coefficients in Table 9.4. The ANOVA tables are identical.

```
> data(hardness)

> hardness.lin.lm <- lm(hardness ~ density,
+                       data=hardness)

> anova(hardness.lin.lm)
Analysis of Variance Table

Response: hardness
          Df   Sum Sq  Mean Sq F value Pr(>F)
density    1 21345674 21345674     637 <2e-16 ***
Residuals 34  1139366    33511
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

> coef(summary.lm(hardness.lin.lm))
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) -1160.50    108.580  -10.69 2.066e-12
density        57.51      2.279   25.24 1.333e-23

> h2 <- data.frame(density=hardness$density, poly(hardness$density, 2))

> xyplot(X1 + X2 ~ density, data=h2)  ## graph not shown in book

> hardness.quad.orth.lm <- lm(hardness ~ density + h2$X2,
+                             data=hardness)

> anova(hardness.quad.orth.lm)
Analysis of Variance Table

Response: hardness
          Df   Sum Sq  Mean Sq F value Pr(>F)
density    1 21345674 21345674   815.9 <2e-16 ***
h2$X2      1   276041   276041    10.6 0.0027 **
Residuals 33   863325    26161
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

> coef(summary.lm(hardness.quad.orth.lm))
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) -1160.50     95.937 -12.096 1.125e-13
density        57.51      2.013  28.564 7.528e-25
h2$X2         525.40    161.745   3.248 2.669e-03
```

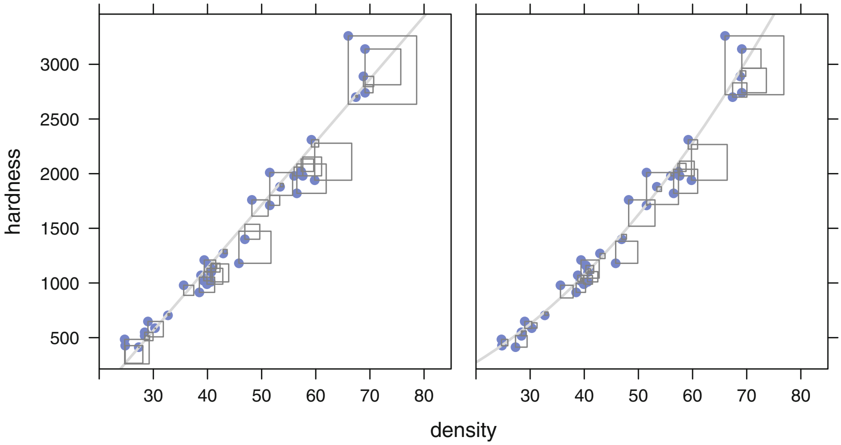

Fig. 9.5 Linear $y \sim x$ and quadratic $y \sim x + x^2$ fits of y=hardness to x=density. The quadratic curve fits much better as can be seen from the much smaller squares (leading to smaller residual sum of squares) at the left and right ends of the `density` range in the quadratic fit. See Table 9.4 for the numerical comparison.

9.8 Models Without a Constant Term

Sometimes it is desired that the statistical model for a response not contain a constant (i.e., vertical intercept) term because the response is necessarily equal to zero if all predictors are zero. An example is the modeling of the body fat data discussed in Section 9.1. Obviously, if a "subject" has zero measurements for `abdomin` and `biceps`, then the response `bodyfat` is necessarily zero also. Similarly, if we wish to model the volume of trees in a forest as a function of trees' diameters and heights, a "tree" having zero diameter and height must have no volume.

An advantage to explicitly recognizing the zero intercept constraint is that a degree of freedom that would be used to estimate the intercept is instead used to estimate the model residual. This results in slightly increased power of tests and decreased sizes of interval estimates of model parameters.

Figure 9.6 and Table 9.6 are for regressions of `bodyfat` on `biceps`, both with and without a constraint that the regression pass through the origin. Note the appreciably smaller slope of the no-intercept regression and that the no-intercept model has 46 df for residual as compared with 45 df for the unconstrained model.

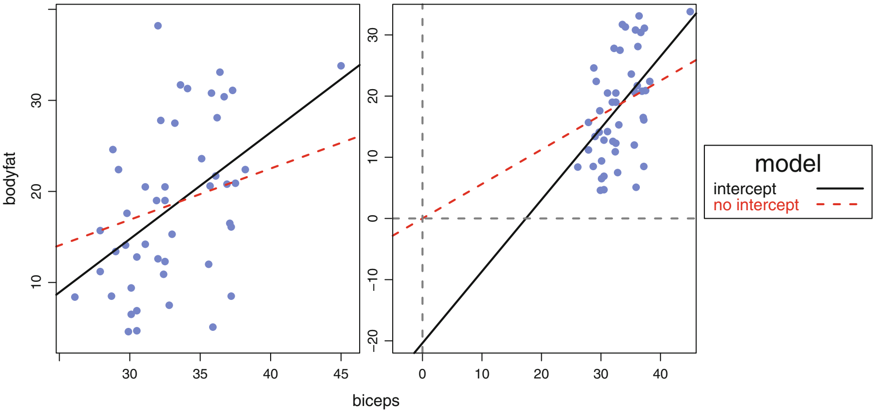

Fig. 9.6 Regressions with and without a constant term for a portion of the body fat data. See Table 9.6. The left panel is limited to the range of the data. The right panel extends the range to include both intercepts. The dotted line through the origin at (0,0) makes an unwarranted extrapolation outside the range of the data.

Table 9.6 Body fat data: Regressions of `bodyfat` on `biceps`, with an intercept term (here) and without an intercept term (in Table 9.7). See Figure 9.6. As compared with the intercept model, the no-intercept model has larger values of both the regression sum of squares and the total sum of squares, and hence also a larger value of R^2.

```
> data(fat)

> ## usual model with intercept
> xy.int.lm <- lm(bodyfat ~ biceps, data=fat)

> summary(xy.int.lm)

Call:
lm(formula = bodyfat ~ biceps, data = fat)

Residuals:
    Min      1Q  Median      3Q     Max
-16.580  -5.443  -0.846   5.255  21.088

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -20.364     10.855   -1.88  0.06715 .
biceps         1.171      0.326    3.59  0.00081 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 8.01 on 45 degrees of freedom
Multiple R-squared:  0.223,Adjusted R-squared:  0.206
F-statistic: 12.9 on 1 and 45 DF,  p-value: 0.00081


> anova(xy.int.lm)
Analysis of Variance Table

Response: bodyfat
          Df Sum Sq Mean Sq F value  Pr(>F)
biceps     1    827     827    12.9 0.00081 ***
Residuals 45   2884      64
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Table 9.7 Body fat data: Regressions of `bodyfat` on `biceps`, without an intercept term. See Table 9.6 for the model with an intercept term. See Figure 9.6. R uses the notation `- 1` in the formula to indicate that the column of **1** is to be suppressed from the dummy variable matrix. As compared with the intercept model, the no-intercept model has larger values of both the regression sum of squares and the total sum of squares, and hence also a larger value of R^2. The no-intercept model has a very high regression sum of squares and corresponding F-value because it includes the contribution from the constant term.

```
> data(fat)

> ## model without a constant term
> xy.noint.lm <- lm(bodyfat ~ biceps - 1, data=fat)

> summary(xy.noint.lm)

Call:
lm(formula = bodyfat ~ biceps - 1, data = fat)

Residuals:
    Min      1Q  Median      3Q     Max
-15.110  -6.145  -0.006   6.841  20.185

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
biceps    0.563      0.036    15.6   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 8.22 on 46 degrees of freedom
Multiple R-squared:  0.841,Adjusted R-squared:  0.838
F-statistic:  244 on 1 and 46 DF,  p-value: <2e-16


> anova(xy.noint.lm)
Analysis of Variance Table

Response: bodyfat
          Df Sum Sq Mean Sq F value Pr(>F)
biceps     1  16506   16506     244 <2e-16 ***
Residuals 46   3110      68
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

9.9 Prediction

Generalizing the discussion in Section 8.3.5 for simple regression, the multiple regression model equation, with regression coefficients estimated by the least-squares analysis, is commonly used for two distinct but related problems.

1. Find a confidence interval on the conditional mean of the population of $Y|x$. That is, estimate a range of mean $E(Y|x)$-values that (with high confidence) bracket the true mean for the specified values of the predictors x.
2. Find a prediction interval for a new observed response Y_0 from these values of the predictors x; i.e., an interval within which a particular new observation will fall with a certain probability.

We continue the analysis of `data(fat)` to illustrate the distinction between these two problems. Using R we continue with `fat2.lm` displayed in Table 9.1. For specificity, we work with x_1 = `abdomin` = 93 and x_2 = `biceps` = 33.

The algebraic setup begins from the model in Equation (9.1), from which it follows that

$$s_e^2 = \frac{Y'Y - \hat{\beta}'X'Y}{n-p-1} = \frac{(Y-\hat{Y})'(Y-\hat{Y})}{n-p-1}$$

Let $x_0 = (93\ 33)$ denote the vector of predictor values for which we wish to construct these two intervals. Define

$$h_0 = x_0(X'X)^{-1}x_0' \tag{9.23}$$

Let $t_{\frac{\alpha}{2},n-p-1}$ denote the $100(1-\frac{\alpha}{2})$ percentage point of the t distribution with $n-p-1$ degrees of freedom. The expected response $E(y|x_0)$ (the center of the confidence interval) and the predicted response $\hat{y}_{x_0}$ for a new observation (the center of the prediction interval) are both equal to $x_0'\hat{\beta}$. Then the $100(1-\alpha)\%$ confidence interval is

$$x_0'\hat{\beta} \pm t_{\frac{\alpha}{2},n-p-1}\, s_e \sqrt{h_0} \tag{9.24}$$

and the $100(1-\alpha)\%$ prediction interval is

$$x_0'\hat{\beta} \pm t_{\frac{\alpha}{2},n-p-1}\, s_e \sqrt{1+h_0} \tag{9.25}$$

The prediction interval is wider than the confidence interval because we are predicting one particular y corresponding to x_0, but estimating with confidence the mean $E(y|x_0)$ of all possible y's that could arise from x_0. A particular y could be much smaller or larger than the mean, and hence there is more uncertainty about y than about the mean. This is captured in the distinction between the two preceding formulas: the "1+" inside the square root. The "1+" arises from the fact that we must predict the ϵ_0 part of the model, but in the estimation problem, we estimate that ϵ_0

Table 9.8 95% Confidence and prediction intervals for the body-fat example. See Tables 9.1 and 13.27 for the ANOVA table and the regression coefficients. The `predict` function produces $s_e\sqrt{h_0}$=`se.fit`, s_e=`residual.scale` and the confidence and prediction intervals.

```
> fat2.lm <- lm(bodyfat ~ abdomin + biceps, data=fat)

> pi.fit <- predict(fat2.lm,
+                   newdata=data.frame(abdomin=93:94, biceps=33:34),
+                   se.fit=TRUE, interval="prediction")

> ci.fit <- predict(fat2.lm,
+                   newdata=data.frame(abdomin=93:94,
+                   biceps=33:34),
+                   se.fit=TRUE, interval="confidence")

> pi.fit
$fit
    fit   lwr   upr
1 18.49 8.485 28.49
2 18.25 8.236 28.26

$se.fit
     1      2
0.7171 0.7518

$df
[1] 44

$residual.scale
[1] 4.911


> ci.fit$fit
    fit   lwr   upr
1 18.49 17.04 19.93
2 18.25 16.73 19.76
```

is zero. As a result, the prediction interval for a given set of explanatory variables is always wider than the corresponding confidence interval.

The confidence and prediction intervals for this example are shown in Table 9.8. The confidence interval (17.0, 19.9) is for the mean percentage body fat of a population of individuals each having `abdomin` circumference 93 cm and `biceps` circumference 33 cm. The prediction interval (8.5, 28.5) is for one particular individual with this combination of `abdomin` and `biceps`. Observe that the prediction interval is wider than the confidence interval. This is because a single person can have atypically low or high body fat, but “many” people includes those with both atypically low and high body-fat percentages in comparison to their `abdomin` and `biceps`, and the lows and highs tend to cancel out when averaging. See Table 8.8 for an illustration of this in the more familiar setting of estimation of a sample mean.

9.10 Example—Longley Data

9.10.1 Study Objectives

The Longley data is a classic small set containing 16 years of annual macroeconomic data that Longley (1967) used to illustrate difficulties arising in computations involving highly intercorrelated variables. R does accurately calculate the regression coefficients for these data. Less numerically sophisticated statistical software packages, including most in existence at the time Longley wrote his article, produce incorrect analyses because the high intercorrelation, or ill-conditioning of the data, is a computational challenge for the numerical solution of linear equations and related matrix operations. Please see the computational discussion in Section I.4.7 for details.

We use `data(longley)`, distributed with R, a subset of all variables in Longley's original data set. Our intent here is to develop a parsimonious model to explain the response variable `Employed` as a function of the remaining variables as candidate predictors. The extreme collinearity arises in this data set because all of its economic variables tend to increase as time progresses. We acknowledge that these are really time series data, and if more than 16 years were involved, it would be appropriate to use time series techniques such as those in Chapter 18 for a proper analysis. We use this example because it is now a classical dataset for investigating a set of poorly conditioned linear equations. Our intention in this section is to analyze these data using multiple regression, demonstrating ways to bypass or confront the difficulties collinearity presents for regression modeling. In contrast, time series analyses specifically seek to model the interdependence caused by time.

9.10.2 Data Description

`GNP.deflator:` GNP adjusted for inflation based on year 1954 = 100

`GNP:` Gross National Product, 1964 Economic Report of the President

`Unemployed:` 1964 Economic Report of the President

`Armed.Forces:` Number serving in the U.S. Armed Forces

`Population:` Noninstitutional, aged at least 14

`Year:` 1947 through 1962

`Employed:` Total employment, U.S. Department of Labor, March 1963

9.10.3 Discussion

Figure 9.7 contains a scatterplot matrix of the Longley data. Here the response variable `Employed` appears in the top (last) row and last column. (In general, for ease of interpretation, response variables should appear in this way or in the bottom (first) row and first column. Remember from Section 4.7 and Figure 4.12 that we strongly recommend that sploms have the main diagonal in the SW–NE direction.)

We see that `Employed` is highly positively correlated with four of the six predictors and mildly positively correlated with the others. In addition, the predictors (including `Year`) that are highly correlated with `Employed` are also highly correlated with one another. This suggests that these four predictors carry redundant information and therefore some of them are unnecessary for modeling the response.

Consider the listing in Table 9.9 for a model containing all six candidate predictors. The proportion of variability in the response `Employed` that is collectively explained by all six predictors is given by R^2, the proportion of the `Sum of Squares`

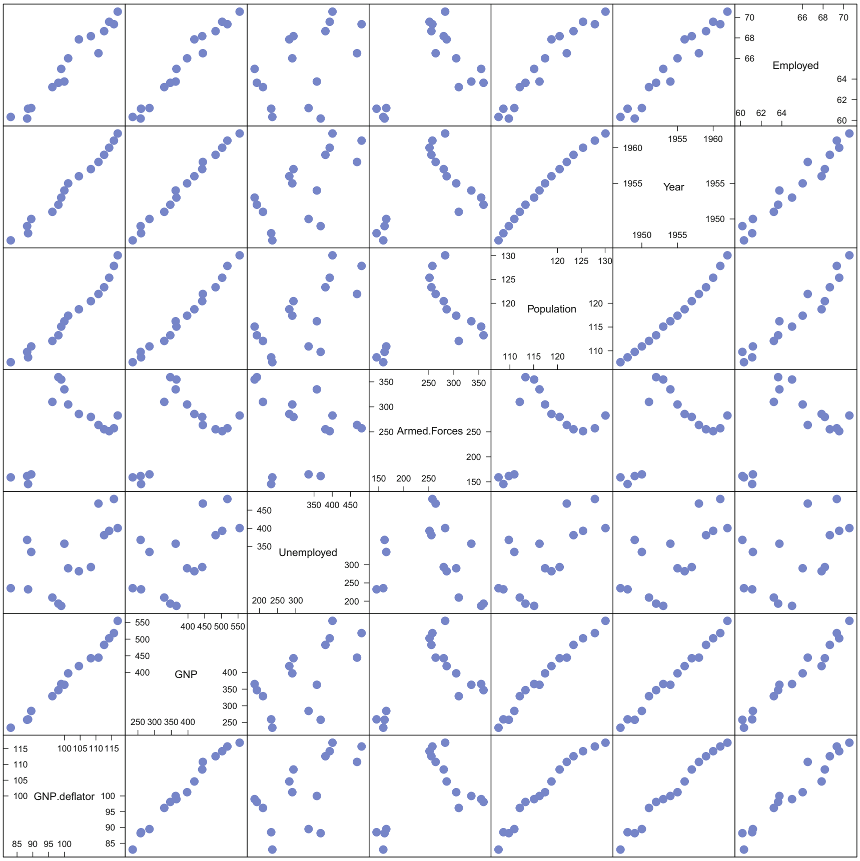

Fig. 9.7 Longley data splom. Notice the high positive correlations of four predictors (including `Year`) with one another and with the response variable `Employed`.

Table 9.9 Longley data regression using all six original predictors.

```
> longley.lm <- lm( Employed ˜ . , data=longley)

> summary(longley.lm)

Call:
lm(formula = Employed ˜ ., data = longley)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4101 -0.1577 -0.0282  0.1016  0.4554

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.48e+03   8.90e+02   -3.91  0.00356 **
GNP.deflator  1.51e-02   8.49e-02    0.18  0.86314
GNP          -3.58e-02   3.35e-02   -1.07  0.31268
Unemployed   -2.02e-02   4.88e-03   -4.14  0.00254 **
Armed.Forces -1.03e-02   2.14e-03   -4.82  0.00094 ***
Population   -5.11e-02   2.26e-01   -0.23  0.82621
Year          1.83e+00   4.55e-01    4.02  0.00304 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.305 on 9 degrees of freedom
Multiple R-squared:  0.995,Adjusted R-squared:  0.992
F-statistic:  330 on 6 and 9 DF,  p-value: 4.98e-10


> anova(longley.lm)
Analysis of Variance Table

Response: Employed
             Df Sum Sq Mean Sq F value  Pr(>F)
GNP.deflator  1   174.4   174.4 1876.53 9.3e-12 ***
GNP           1     4.8     4.8   51.51 5.2e-05 ***
Unemployed    1     2.3     2.3   24.36 0.00081 ***
Armed.Forces  1     0.9     0.9    9.43 0.01334 *
Population    1     0.3     0.3    3.75 0.08476 .
Year          1     1.5     1.5   16.13 0.00304 **
Residuals     9     0.8     0.1
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

> vif(longley.lm)
GNP.deflator          GNP   Unemployed Armed.Forces
     135.532     1788.513       33.619        3.589
  Population         Year
     399.151      758.981
```

column *not* in the `Residuals` row: more than 0.99. So the predictors can be used to adequately explain `Employed`. In this model, three predictors that seem to be closely correlated with the response `Employed` in Figure 9.7, `Population`, `GNP`, and `GNP.deflator`, are not statistically significant in Table 9.9. We continue to discuss the Longley data, focusing on the selection of an appropriate subset of the predictors, in Sections 9.11 and 9.12.

9.11 Collinearity

Collinearity, also called multicollinearity, is a condition where the model's predictors variables are highly intercorrelated. A consequence of this situation is the inability to estimate the model's regression coefficients with acceptable precision. Therefore, models with this problem are not considered useful. It is unacceptable to reach a final model that has this condition to an appreciable extent.

Collinearity arises when investigators include predictors carrying redundant information in the model. A symptom is a model with a high R^2, showing that collectively the predictors bear heavily on the response, but paradoxically, few or none of the predictors have regression coefficients significantly different from zero.

Consider the case of a single response Y and two predictors X_1 and X_2. The fitted model plots as a plane in the 3-dimensional space of (Y, X_1, X_2). A near-collinear situation exists if the correlation between X_1 and X_2 is close to ± 1. Geometrically, this occurs when the data points congregate close to a (2-dimensional) straight line when plotted in the 3-dimensional space. When this happens, the points can be fitted fairly well by any plane containing this straight line. Since each of these many planes is a candidate for the best model, the model decided upon as being *the* best will be similar to other model candidates. Therefore, declaring any model to be best will be a tentative decision. This tentativeness is expressed by large standard errors of the estimated regression coefficients that comprise the coefficients of the plane corresponding to the best model.

Figure 9.8, based on a portion of the Longley data introduced in Section 9.10, illustrates these ideas. Here the variables `GNP` and `Year` are almost perfectly correlated and so the scattering of points falls close to a line in 3-dimensional space. Many planes fit this line approximately equally well. The uncertainty about the best fitting of these many planes causes the coefficients of the estimated plane, the regression coefficients, to have large standard errors.

When there are more than two predictors, the geometric argument extends to discussions of hyperplanes. The consequence is again unacceptably large standard errors of regression coefficients.

Although collinearity limits our ability to model the relationship between the predictors and the response accurately, it does not necessarily impede our ability to use the predictors to predict the response. In the context of the example associated

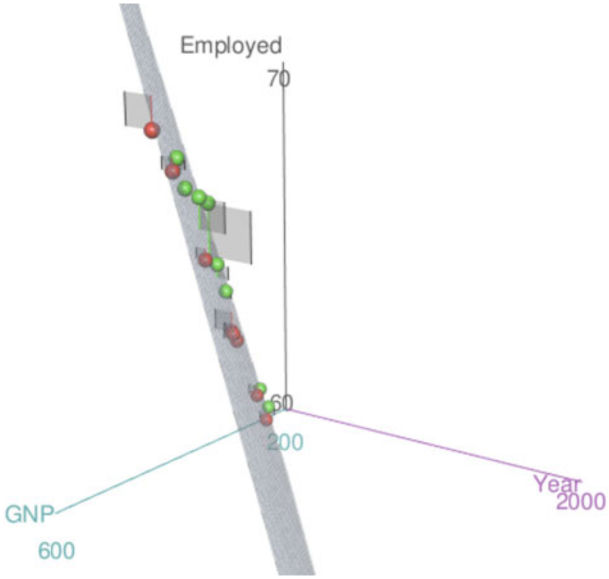

Fig. 9.8 The two X-variables, `Year` and `GNP`, are highly collinear. See model

```
longley2.lm <- lm(Employed ~ Year + GNP, data=longley)
```

in file `HHscriptnames(9)`. The response variable `Employed` is essentially on a straight line in the three-dimensional space of the figure. The specific plane displayed is almost arbitrary. Any plane that goes through the straight line of the observed points on the plane we see would work just as well.

with Figure 9.8, if we want to predict the response for values of the predictors near the straight line in 3-dimensional space, many planes that are good fits to this straight line will yield roughly the same prediction.

A simple diagnostic of collinearity is the *variance inflation factor*, VIF, one for each regression coefficient (other than the intercept). Since the condition of collinearity involves the predictors but not the response, this measure is a function of the X's but not of Y. The VIF for predictor i is

$$\text{VIF}_i = 1/(1 - R_i^2) \tag{9.26}$$

where R_i^2 is the R^2 from a regression of predictor i against the remaining predictors. If R_i^2 is close to 1, this means that predictor i is well explained by a linear function of the remaining predictors, and, therefore, the presence of predictor i in the model is redundant. Values of VIF exceeding 5 are considered evidence of collinearity: The information carried by a predictor having such a VIF is contained in a subset of the remaining predictors. If, however, all of a model's regression coefficients differ significantly from 0 (p-value $< .05$), a somewhat larger VIF may be tolerable.

VIF is an imperfect measure of collinearity. Occasionally the condition can be attributable to more complicated relationships among the predictors than VIF can detect.

The best approach for alleviating collinearity is to reduce the set of predictors to a noncollinear subset. Methods for accomplishing this are presented in Section 9.12. An ad hoc (manual) procedure, presented in Section 9.12.1, involves eliminating predictors one at a time, at each stage deleting the predictor having the highest VIF. If two predictors are almost tied for highest, then subject area information should be used to choose between them. Proceed until all remaining predictors have VIF ≤ 5. Other approaches (not discussed in this book) include ridge regression and regression on principal components Gunst and Mason (1980).

For the regression analysis of the Longley data, evidence of collinearity appears in Table 9.9 in the variance inflation factors (VIF) for the six predictors. Five of these exceed 33. The next section discusses an approach for dealing with multicollinearity.

Collinearity often arises in polynomial regression models discussed in Section 9.7 because polynomials can be approximated by linear functions within a restricted domain. To avoid both collinearity in polynomial models and numerical instability caused by working with variables of greatly differing orders of magnitude, it is recommended to recenter the response variable to have mean = 0 prior to initiating a polynomial modeling.

9.12 Variable Selection

In building a regression model the analyst should consider for use any explanatory variable that is likely to bear upon the response while avoiding the use of two explanatory variables that carry essentially the same information. For example, in modeling the monthly cost of energy needed to heat a 2000-square-foot home, one should avoid using both the mean monthly exterior temperature and the heating degree days (a measure used by heating fuel suppliers) in the same model. The use of redundant explanatory variables is likely to lead to a model with unacceptable collinearity having large standard errors for the predictor regression coefficients.

When subject area theory does not suggest a parsimonious model (i.e., one with relatively few predictors), it is tempting to construct a model using all possibly relevant predictors for which data are available. However, doing so is again likely to result in a collinearity problem. In such circumstances, how can the analyst decide on an appropriate subset of the candidate predictors for a regression model?

Stepwise regression is a tool for answering this question. But this mechanical technique should not be used in order to avoid careful thought about potentially useful predictor variables. Careless use of stepwise regression can, to some extent, distort the significance and confidence levels of inferences in the ultimately specified model, potentially leading to erroneous conclusions. In addition, a model that makes reasonable subject area sense to the client is much preferred to an equally well fitting one that is less intuitive and harder to understand and explain.

In our experience, a careful systematic approach can often be used to develop a more interpretable model than one produced by a mechanical stepwise algorithm. The starting point is a scatterplot matrix that, along with examination of variance inflation factors, can be used to identify redundant predictors. If two predictors are seen to be highly correlated, we prefer to avoid using the one that has a less obvious subject matter connection to the response variable. An algorithm cannot make such a judgment. Inspection of sploms invite the analyst to consider whether an original variable should be transformed before inclusion in the model. Nevertheless, stepwise approaches to model selection continue to be commonly used, particularly when there are a large number of potential predictors and the analyst has minimal feel for which variables should be or need not be included in the model.

We discuss in turn two systematic methods for model selection, a manual approach and an automated approach, and apply both methods to the Longley data.

9.12.1 Manual Use of the Stepwise Philosophy

The first approach involves manual inspections of the VIFs, the p-values associated with the t-tests on the regression coefficients, and any available subject matter information to eliminate variables one at a time until a final model is reached with all predictors significant and all VIFs under 5. This approach is viable if the number of predictors is small as in this example. It would be too cumbersome in a situation with more than 12 to 15 predictors.

The three largest VIFs belong to `GNP`, `Year`, and `Population`. The splom implies that they carry almost identical information. We begin by removing one of them from the model. We choose to eliminate `Population` because the t-test that its regression coefficient is zero has a larger p-value than the tests for either `GNP` or `Year`.

The analysis with all predictors except `population` appears in Table 9.10.

The outstanding feature of this model is the high p-value associated with variable `GNP.deflator`. Its VIF is well in excess of 5. We proceed with an analysis eliminating `GNP.deflator` in Table 9.11.

All four predictors in this model have significant regression coefficients. However, two of the VIFs are still large, and one of the predictors corresponding to them must be eliminated. We choose to eliminate `GNP` because its p-value, while small, is larger than those of the three other remaining predictors.

The results of the analysis with the remaining predictors `Unemployed`, `Year`, and `Armed.Forces` are in Table 9.12.

This is our tentative final model. The collinearity has been eliminated (all VIFs are below 5), and all regression coefficients differ significantly from zero. In addition, $R^2 = 0.993$, so these three predictors account for virtually all of the variability in `Employed`.

Table 9.10 Longley data regression. Best five-predictor model after eliminating one predictor using the manual stepwise approach.

```
> longley3.lm <- lm( Employed ˜
+          GNP.deflator + GNP + Unemployed + Armed.Forces + Year,
+          data=longley)

> summary(longley3.lm)

Call:
lm(formula = Employed ˜ GNP.deflator + GNP + Unemployed +
    Armed.Forces + Year, data = longley)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3901 -0.1434 -0.0356  0.0973  0.4614

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.56e+03   7.72e+02   -4.62  0.00096 ***
GNP.deflator  2.77e-02   6.07e-02    0.46  0.65798
GNP          -4.21e-02   1.76e-02   -2.39  0.03789 *
Unemployed   -2.10e-02   3.03e-03   -6.95    4e-05 ***
Armed.Forces -1.04e-02   2.00e-03   -5.21  0.00040 ***
Year          1.87e+00   3.99e-01    4.68  0.00087 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.29 on 10 degrees of freedom
Multiple R-squared:  0.995,Adjusted R-squared:  0.993
F-statistic:  438 on 5 and 10 DF,  p-value: 2.27e-11


> anova(longley3.lm)
Analysis of Variance Table

Response: Employed
             Df Sum Sq Mean Sq F value  Pr(>F)
GNP.deflator  1  174.4   174.4  2073.3 6.3e-13 ***
GNP           1    4.8     4.8    56.9 2.0e-05 ***
Unemployed    1    2.3     2.3    26.9 0.00041 ***
Armed.Forces  1    0.9     0.9    10.4 0.00905 **
Year          1    1.8     1.8    21.9 0.00087 ***
Residuals    10    0.8     0.1
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

> vif(longley3.lm)
GNP.deflator          GNP   Unemployed Armed.Forces         Year
      76.641      546.870       14.290        3.461      644.626
```

Table 9.11 Longley data regression. Best four-predictor model after eliminating two predictors using the manual stepwise approach.

```
> longley4.lm <- lm(Employed ˜
+                   GNP + Unemployed + Armed.Forces + Year,
+                   data=longley)

> summary(longley4.lm)

Call:
lm(formula = Employed ˜ GNP + Unemployed + Armed.Forces + Year,
    data = longley)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4217 -0.1246 -0.0242  0.0837  0.4527

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.60e+03   7.41e+02   -4.86  0.00050 ***
GNP          -4.02e-02   1.65e-02   -2.44  0.03283 *
Unemployed   -2.09e-02   2.90e-03   -7.20  1.7e-05 ***
Armed.Forces -1.01e-02   1.84e-03   -5.52  0.00018 ***
Year          1.89e+00   3.83e-01    4.93  0.00045 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.279 on 11 degrees of freedom
Multiple R-squared:  0.995,Adjusted R-squared:  0.994
F-statistic:  590 on 4 and 11 DF,  p-value: 9.5e-13


> anova(longley4.lm)
Analysis of Variance Table

Response: Employed
             Df Sum Sq Mean Sq F value  Pr(>F)
GNP           1  179.0   179.0  2292.7   4e-14 ***
Unemployed    1    2.5     2.5    31.5 0.00016 ***
Armed.Forces  1    0.8     0.8    10.5 0.00779 **
Year          1    1.9     1.9    24.3 0.00045 ***
Residuals    11    0.9     0.1
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

> vif(longley4.lm)
         GNP   Unemployed Armed.Forces         Year
     515.124       14.109        3.142      638.128
```

Table 9.12 Longley data regression. Best three-predictor model after eliminating three predictors using the manual stepwise approach.

```
> longley5.lm <- lm(Employed ~
+                   Unemployed + Armed.Forces + Year,
+                   data=longley)

> summary(longley5.lm)

Call:
lm(formula = Employed ~ Unemployed + Armed.Forces + Year,
   data = longley)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5729 -0.1199  0.0409  0.1398  0.7530

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.80e+03   6.86e+01  -26.18  5.9e-12 ***
Unemployed   -1.47e-02   1.67e-03   -8.79  1.4e-06 ***
Armed.Forces -7.72e-03   1.84e-03   -4.20   0.0012 **
Year          9.56e-01   3.55e-02   26.92  4.2e-12 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.332 on 12 degrees of freedom
Multiple R-squared:  0.993,Adjusted R-squared:  0.991
F-statistic:  555 on 3 and 12 DF,  p-value: 3.92e-13


> anova(longley5.lm)
Analysis of Variance Table

Response: Employed
             Df Sum Sq Mean Sq F value  Pr(>F)
Unemployed    1   46.7    46.7     424 1.0e-10 ***
Armed.Forces  1   57.0    57.0     517 3.1e-11 ***
Year          1   79.9    79.9     725 4.2e-12 ***
Residuals    12    1.3     0.1
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

> vif(longley5.lm)
  Unemployed Armed.Forces         Year
       3.318        2.223        3.891
```

9.12.2 Automated Stepwise Regression

The second approach to model selection is stepwise regression. This automated approach is recommended when the number of predictors is so large that the manual approach becomes unacceptably laborious. We illustrate here how it is used to reach the same model that we found with the manual procedure. A stepwise approach that examines all subsets of predictors is viable if the number of predictors p is less than 10 to 12. If $p > 12$, then forward selection or backward elimination is preferred.

The three basic methods for automated stepwise regression are

forward selection: Predictors are added to the model one at a time until a stopping rule is satisfied.

backward elimination: All predictors are initially placed in the model. Predictors are removed from the model one at a time until a stopping rule is satisfied.

all subsets: All $2^p - 1$ possible models, where p is the number of predictors, are attempted and the best is identified. This method is viable only for "small" values of p. Efficient algorithms exist that avoid actually examining every such model.

The literature contains many hybrids and refinements of these basic methods.

Each of the automated stepwise methods uses a criterion for choosing the next step or stopping the algorithm.

Such criteria may relate to appreciable R^2_{adj} or F-statistic improvement or detriment, substantial mean square error decrease or increase, or size of change in Daniel–Mallows' C_p statistic discussed below. Another possibility is to look, at each step, at the p-value for the variables already in the model and for the potential next variable to be brought in to the model. If the largest p-value of the variables already in the model is larger than the threshold, then remove it. If the smallest p-value of the potential variables is larger than the threshold, then stop. Otherwise, bring in a new variable and repeat the process.

Computer algorithms allow the option of accepting or overriding default criterion values or thresholds for appreciable change.

Each of the automated stepwise methods uses one or more criteria for choosing among competing models. Here is a list of possible criteria.

p Models containing fewer predictors are easier to interpret and understand. It is desirable that the number of predictors p be as small as possible.

$\hat\sigma^2$ We also require that the predictors account for most of the variability in the response. Equivalently, we wish that the residual mean square, MSE = $\hat\sigma^2$, be as small as possible, preferably not much larger than for the model containing all candidate predictors. This criterion is easier to meet with more predictors rather than few; hence it asks that the number of predictors p be as large as possible and competes with the goal of minimizing p.

The above criteria address one of the two competing objectives at a time. Other criteria jointly address the two objectives.

R^2_{adj} Unadjusted R^2 is not used as a model selection criterion because it necessarily increases as the number of predictors increases. A model can have R^2 close to 1 but be unacceptable due to severe collinearity. Instead we use R^2_{adj}, which is R^2 adjusted downward for the number of predictors,

$$R^2_{\text{adj}} = 1 - \left(\frac{n-1}{n-p-1}\right)(1 - R^2) \tag{9.27}$$

which increases as R^2 increases but provides a penalty for an excessive number of predictors p. Models with higher R^2_{adj} are preferred to ones with lower R^2_{adj}.

C_p Daniel–Mallows' C_p statistic is another criterion that addresses both the fit of the model and the number of predictors used. Consistent with customary notation, in the context of the C_p statistic but nowhere else in this chapter, p is the number of regression coefficient *parameters*, equal to the number of predictors *plus 1*. The original definition is

$$C_p = (\mathsf{SS}_{\text{Res}}/\hat{\sigma}^2_{\text{full}}) + 2p - n \tag{9.28}$$

where SS_{Res} is the residual sum of squares for the reduced model under discussion (fewer X-variables than the full model) and the $\hat{\sigma}^2_{\text{full}}$ is the error mean square for the full model containing all candidate predictors. If the extra X-variables are noise, rather than useful, then the ratio $\mathsf{SS}_{\text{Res}}/\hat{\sigma}^2_{\text{full}} \approx ((n-p)\sigma^2)/\sigma^2_{\text{full}} \approx n - p$. If the extra X-variables are useful, then the numerator $\sigma^2 \gg \sigma^2_{\text{full}}$ and the ratio will be much larger than $n - p$. The extra terms $2p - n$ make the entire C_p approximate p when the extra X-variables are not needed.

A desirable model has $C_p \approx p$ for a small number of parameters p. (If p_{max} denotes p for a model containing all candidate predictors, then necessarily $C_{p_{\text{max}}} = p_{\text{max}}$, but such a model is almost never acceptable.) C_p results are often conveyed with a C_p plot, that is, a plot of C_p vs p, with each point labeled with an identifier for its model and the diagonal line having equation $C_p = p$ added to the plot. Desirable models are those close to or under this diagonal line.

AIC The Akaike information criterion is proportional to the C_p statistic. The AIC is scaled in sum of squares units.

F At each step we can look at the p-value associated with the F-statistic for the variables already in the model and for the potential next variable to be brought in to the model. If the largest p-value of the variables already in the model is larger than the threshold, then remove it. If the smallest p-value of the potential variables is larger than the threshold, then stop. Otherwise, bring in a new variable and repeat the process.

9.12.3 Automated Stepwise Modeling of the Longley Data

Table 9.13 contains the results of an R stepwise regression analysis considering all subsets of the predictors, with printouts of the properties of two models of each size having smallest residual sum of squares among models having $C_p < 10$. Figure 9.9 is a plot of the C_p-values for all models with $C_p < 10$. The acronymic plot symbols in Figure 9.9 are decoded in Table 9.13. According to Table 9.13, the best parsimonious model is the one with the four predictors `GNP`, `Unemployed`, `Armed.Forces`, and `Year` displayed in Table 9.11. This model has C_p close to p, and a smaller AIC and larger adjusted R^2 than any of the other models in Table 9.13. Unlike the model we selected with our manual approach, this one includes the predictor `GNP`. The algorithm underlying Table 9.13 suggests inclusion of `GNP` despite its high correlation with `Year` and high VIF shown in Table 9.11.

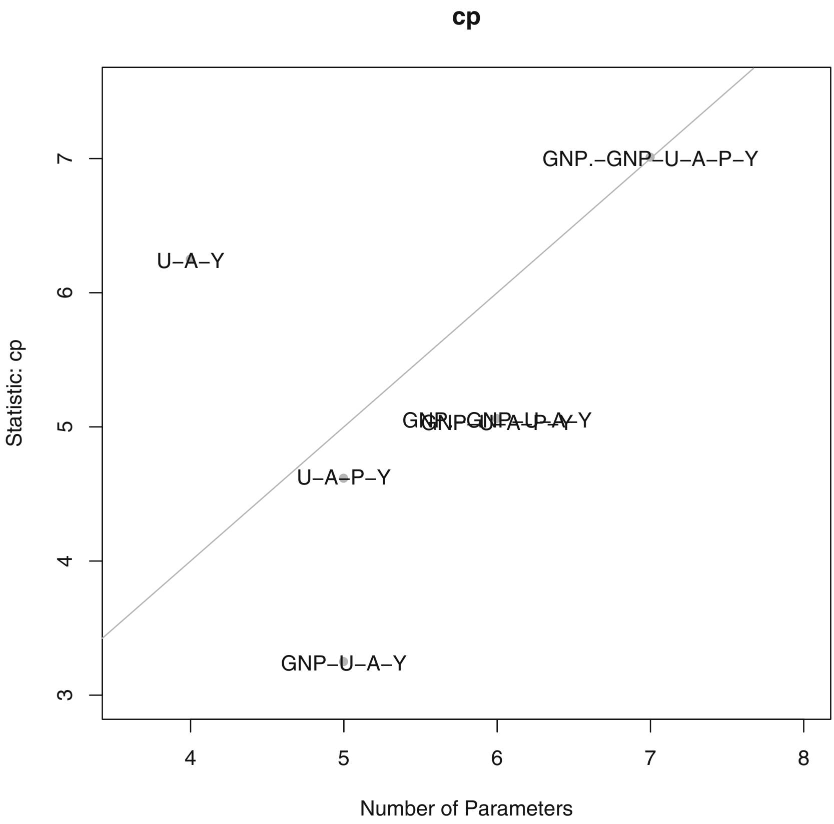

Fig. 9.9 C_p Plot for Longley data. See Table 9.13 for interpretations of the acronyms used to label points. The overplotting occurs because, as seen in Table 9.13, two models have almost identical values of C_p.

Table 9.13 Longley data regression. Model 7 with the four predictors GNP, Unemployed, Armed.Forces and Year is competitive with respect to C_p and other criteria. This model has the largest adjusted R^2 and the smallest C_p. This is the same model we found in Table 9.11.

```
> longley.subsets <-
+   leaps::regsubsets(Employed ~ GNP.deflator + GNP +
+                       Unemployed +
+                       Armed.Forces + Population + Year,
+                       data=longley, nbest=2)

> longley.subsets.Summary <- summaryHH(longley.subsets)

> ## longley.subsets.Summary
> tmp <- (longley.subsets.Summary$cp <= 10)

> longley.subsets.Summary[tmp,]
              model p   rsq   rss adjr2   cp   bic stderr
5             U-A-Y 4 0.993 1.323 0.991 6.24 -68.0  0.332
7         GNP-U-A-Y 5 0.995 0.859 0.994 3.24 -72.1  0.279
8           U-A-P-Y 5 0.995 0.986 0.993 4.61 -69.9  0.299
9       GNP-U-A-P-Y 6 0.995 0.839 0.993 5.03 -69.7  0.290
10   GNP.-GNP-U-A-Y 6 0.995 0.841 0.993 5.05 -69.7  0.290
11 GNP.-GNP-U-A-P-Y 7 0.995 0.836 0.992 7.00 -67.0  0.305

Model variables with abbreviations

                                                                    model
GNP                                                                   GNP
Y                                                                    Year
U-Y                                                       Unemployed-Year
GNP-U                                                      GNP-Unemployed
U-A-Y                                        Unemployed-Armed.Forces-Year
GNP-U-A                                       GNP-Unemployed-Armed.Forces
GNP-U-A-Y                                GNP-Unemployed-Armed.Forces-Year
U-A-P-Y                           Unemployed-Armed.Forces-Population-Year
GNP-U-A-P-Y                   GNP-Unemployed-Armed.Forces-Population-Year
GNP.-GNP-U-A-Y              GNP.deflator-GNP-Unemployed-Armed.Forces-Year
GNP.-GNP-U-A-P-Y GNP.deflator-GNP-Unemployed-Armed.Forces-Population-Year

 model with largest adjr2
 7

 Number of observations
 16
```

Which model is preferred, the one in Table 9.11 containing four predictors including GNP or the three predictor model in Table 9.12 that excludes GNP? Our answer to this question demonstrates our preference for the manual approach. The coefficient of GNP in Table 9.11 is negative. This model says that holding Unemployed, Armed.Forces and Year constant, GNP and Employed are *negatively* associated. This statement conflicts with our expectation that this association is positive, and is a strong argument against the four-predictor model in Table 9.11.

9.13 Residual Plots

Partial residual plots and *added variable plots* are visual aids for interpreting relationships between variables used in regression. They can serve as additional components of our manual approach for variable selection.

Figure 9.10 shows four different types of plots.

- Row 1 shows the response variable Y=`Employed` against each of the six predictors X_j.
- Row 2 shows the ordinary residuals $e = Y - \hat{Y}$ from the regression on all six variables against each of the six predictors.
- Row 3 shows the "partial residual plots", the partial residuals e^j for each predictor against that predictor. See Section 9.13.1 for construction of the partial residuals and Section 9.13.2 for construction of the partial residual plots.
- Row 4 shows the "added variable plots", the partial residuals e^j against the partial residuals $X_{j|1,2,\ldots,j-1,j+1,\ldots,p}$ of X_j regressed on the other five predictors. See Section 9.13.3 for the definition of partial correlation, and Section 9.13.4 for construction of the $X_{j|1,2,\ldots,j-1,j+1,\ldots,p}$ and the added variable plots.

We discuss the interpretation of the all four types of plots in Section 9.13.5. We recommend the discussions of partial residual plots and added variable plots in Weisberg (1985) and Hamilton (1992).

9.13.1 Partial Residuals

The partial residuals e^j for variable X_j in a model with p predictor variables X_j are defined

$$e^j = Y - \hat{Y}_{1,2,\ldots,j-1,j+1,\ldots,p} \tag{9.29}$$

and calculated with

$$e^j = X_j\hat{\beta}_j + e \tag{9.30}$$

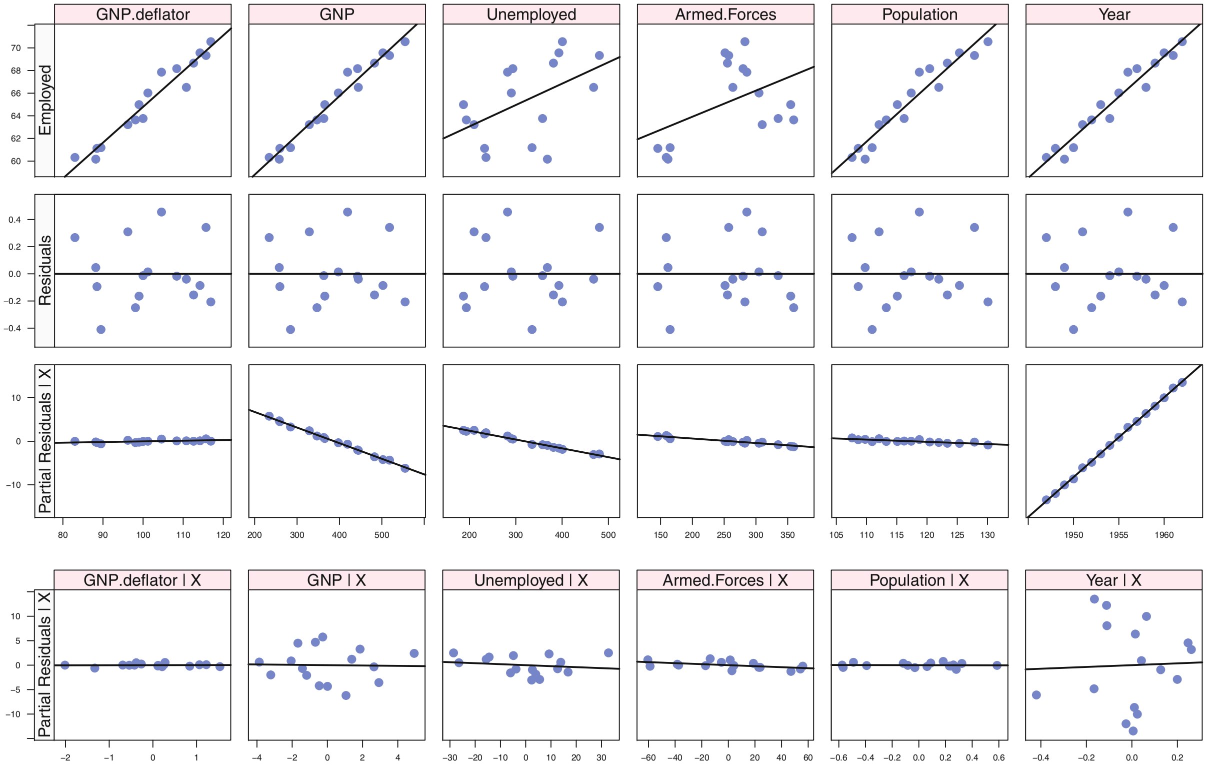

Fig. 9.10 Four types of plots for the regression of the Longley data against all six potential predictors: Response variable Y against each X_j, residuals e against each X_j, partial residuals plots of e^j against each X_j, added variable plots of e^j against the residuals of each X_j adjusted for the other X columns. The slopes shown in the panels of both bottom rows are equal to the regression coefficients from the regression shown in Table 9.9.

or equivalently

$$e_i^j = X_{ij}\hat{\beta}_j + e_i \quad \text{for } i = 1, \ldots, n \tag{9.31}$$

where $e = (e_i)$ are the ordinary residuals from the model with all p predictors

$$e = Y - \hat{Y}_{1,2,\ldots,p} \tag{9.32}$$

The partial residuals are interpreted as the additional information available for X_j to pick up after all X except X_j have been included in the model.

9.13.2 Partial Residual Plots

Partial residual plots are the set of plots of e^j against X_j for all j. Each panel's slope has exactly the numerical value of the corresponding regression coefficient.

We show the partial residual plots for the Longley data in Row 3 of Figure 9.10.

9.13.3 Partial Correlation

The partial correlation $r(X_1, X_2 | X_3, X_4, X_5)$ between X_1 and X_2, after correction for the effect of X_3, X_4, X_5, is the correlation coefficient between X_1 and X_2 after the (linear) effects of X_3, X_4, X_5 have been removed from both X_1 and X_2. When X_1 through X_5 are multivariate data, we can compute the sample partial correlation coefficient as follows:

- Regress X_1 on X_3, X_4, X_5. Get the residuals E_1.
- Regress X_2 on X_3, X_4, X_5. Get the residuals E_2.
- Find the (usual) correlation coefficient between E_1 and E_2. This turns out to be $r(X_1, X_2 | X_3, X_4, X_5)$.

In R, we use

```
partial.corr(cbind(X1,X2),
             cbind(X3,X4,X5))
```

using the function `partial.corr` defined in the **HH** package.

9.13.4 Added Variable Plots

The added variable plots are the set of plots of $E_1 = e^j$ against $E_2 = X_{j|1,2,\ldots,j-1,j+1,\ldots,p}$ for all j. We define $\hat{X}_{1,2,\ldots,j-1,j+1,\ldots,p}$ to be the predicted value of X_j after regressing X_j against all the other X-variables in the model. We define the residual

$$X_{j|1,2,\ldots,j-1,j+1,\ldots,p} = X_j - \hat{X}_{1,2,\ldots,j-1,j+1,\ldots,p} \tag{9.33}$$

to be the additional information in X_j after removing the information provided by all the other X in the model. Thus the added variable plots are the plots of the E_1 and E_2 defined by regressing Y and X_j against all the other X-variables. Each panel's slope has exactly the numerical value of the corresponding regression coefficient.

We show the added variable plots for the Longley data in Row 4 of Figure 9.10.

9.13.5 Interpretation of Residual Plots

9.13.5.1 Response Variable Against Each of the Predictors

Row 1 of Figure 9.10, the plots of the response variable Y=`Employed` against each of the six predictors X_j, is almost identical to the top row of the splom in Figure 9.7. The only difference is the explicit one-x regression line in Figure 9.10. If there is no visible slope in any of these panels, then we can effectively eliminate that x-variable from further consideration as a potential explanatory variable. This row is essentially the same as the first step of a stepwise-forward procedure. In this example, we cannot eliminate any of the potential predictors at this stage.

9.13.5.2 Residuals Against Each of the Predictors

Row 2 of Figure 9.10, the plots of the ordinary residuals $e = Y - \hat{Y}$ (from the complete regression of the response on all six potential predictors X_j), against each of the X_j shows horizontal slopes. This is by construction, as the least-squares residuals are orthogonal to all X-variables. In this example, we see no structure in the plots. The types of structure we look for are

Curvature. Plot the residuals from the quadratic fit in the left side of Figure 9.5 against the predictor `density` and note that the residuals are predominantly above the $y = 0$ axis at the left and right ends of the range and predominantly below the axis in the middle of the range. Curvature in the residual plots often suggests that additional predictors, possibly powers of existing predictors, are needed in the model.

Nonuniformity of variance. The `life.exp ~ ppl.per.tv` panel of Figure 4.14 shows high variability in `life.exp` for low values of `ppl.per.tv` and very low variability for high values of `ppl.per.tv`. Nonuniformity of variance in the residual plots often suggests power transformations of one or more of the variables. Transformations of both the response and predictor variables need to be considered.

Bunching or granularity. See the `residuals ~ lime` panel of Figure 11.11 where we see that `lime` has only two levels and there are different variances for each.

9.13.5.3 Partial Residuals

Both Rows 3 and 4 use the partial residuals of the response as the y-variable of each plot. Since "partial" means "adjusted for all the other x-variables", each column of Rows 3 and 4 is different. Column 1 is adjusted for $X_2, X_3, \dots, X_6$. Column 2 is adjusted for X_1, $X_3, \dots, X_6$. Similarly through Column 6, which is adjusted for $X_1, \dots, X_5$.

In Row 3, the *partial residual plots*, the x-variables are the observed x-variables X_j.

In Row 4, the *added residual plots*, the x-variables are the adjusted-x variables, that is, "adjusted for all the other x-variables". Thus the x-variable in Column 1 of Row 4 is $X_{1|2,\dots,6}$, that is, X_1 adjusted for $X_2, \dots, X_6$.

In both Rows 3 and 4 the slope of the two-dimensional least-squares line in panel j is exactly the value of the regression coefficient β_j for the complete regression of Y on all the X-variables in the model.

9.13.5.4 Partial Residual Plots

In Row 3, the partial residuals e^j are plotted against the observed x-variables X_j. Since the partial residuals e^j are specific to each X_j, the values for the y-range are unique to each panel. The x-range of the x-variables in Row 3 is the same as it is in Rows 1 and 2 of this display.

We look for the tightness of the points in each plot around their least-squares line. High variability around the two-dimensional least-squares line indicates low significance for the corresponding regression coefficient. Low variability around the least-squares line indicates a significant regression coefficient.

In Row 3 of Figure 9.10, we see that Columns 1 (`GNP.deflator`) and 5 (`Population`) have high variability around their least-squares lines. This is a reflection of the high p-value that we see for those regression coefficients in Table 9.9. The remaining four columns all look like their points are tightly placed against their least-squares lines, an indication of possible significance. Note that Column 2 (`GNP`) looks tight, even though its p-value is the nonsignificant 0.3127. We really do need the tabular results to completely understand what the graph is showing us.

9.13.5.5 Added Variable Plots

In Row 4, the partial residuals e^j are plotted against the adjusted x-variables $X_{j|1,2,\ldots,j-1,j+1,\ldots,p}$. In Row 4, both the x- and y-variables in each column have been adjusted for all the other X-variables. Therefore, both the x- and y-ranges are unique to each panel. The partial residuals, the y-variables in the added variable plots, are identical to the y-variables in the partial residual plots; hence the y-ranges are identical for corresponding columns of Rows 3 and 4.

We look at the slope of the two-dimensional least-squares line in each plot. A nearly horizontal line indicates low significance for the corresponding regression coefficient. A nonzero slope indicates a significant regression coefficient.

The three x-variables with significant regression coefficients in Table 9.9 have visible nonzero slopes to their least-squares lines in Row 4 of Figure 9.10. The three x-variables with nonsignificant regression coefficients have almost horizontal least-squares lines.

9.14 Example—U.S. Air Pollution Data

Exercise 4.2 introduces the data set `data(usair)` on causes of air pollution in U.S. cities. A scatterplot matrix of these data appears in Figure 9.11. Here we seek to develop a model to explain the response `SO2`, SO_2 content of air, using a subset of six available explanatory variables.

In Figure 9.11 we see that the three variables `SO2`, `mfgfirms`, and `popn` are all pushed against their minimum value with a long tail toward the maximum value. This pattern suggests a log transformation to bring these three distributions close to symmetry. Following these transformations, Figure 9.12 shows the new response variable `lnSO2` and the revised list of six potential explanatory variables.

For pedagogical purposes we approach this problem in two different ways. We first use the automated stepwise regression approach and then consider the manual approach.

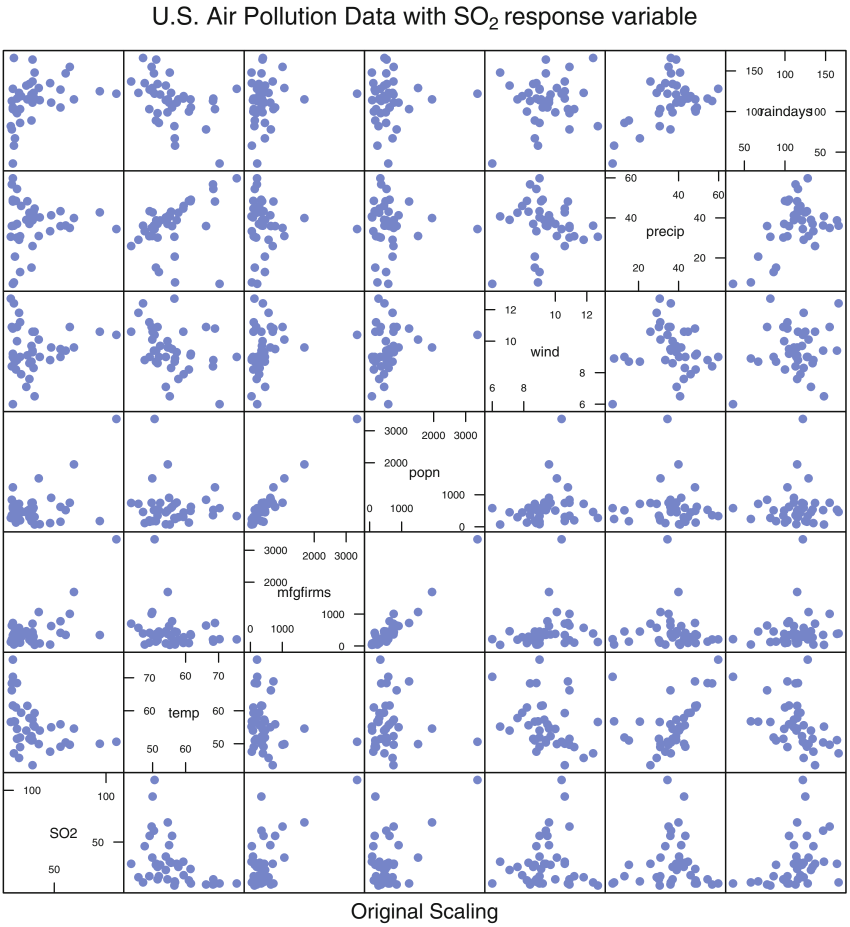

Fig. 9.11 Scatterplot matrices for air pollution data with the original scaling.

We illustrate the automated approach with the `leaps::regsubsets` function in R, using the `exhaustive` method that considers all subsets. In this problem there are only a small number, $2^6 - 1 = 31$, of subsets to consider, so this method is viable. We request the best two subsets for each possible value of the number of included explanatory variables. The tabular and graphical results of the stepwise analysis are displayed in Table 9.14 and Figure 9.13. The model with the four predictors `temp`, `lnmfg`, `wind`, and `precip` seems best. It has $C_p \approx p$, the smallest AIC of contenders, the largest R^2_{adj}, and one of the smallest values of SS_{Res}.

In Table 9.15 we look at the detail for the selected model. We observe that all VIFs are small and the p-values are below 0.01 for all model coefficients. The signs of the estimated coefficients are reasonable or defensible. United States cities with high average annual temperature are located in the Sunbelt and tend to have less

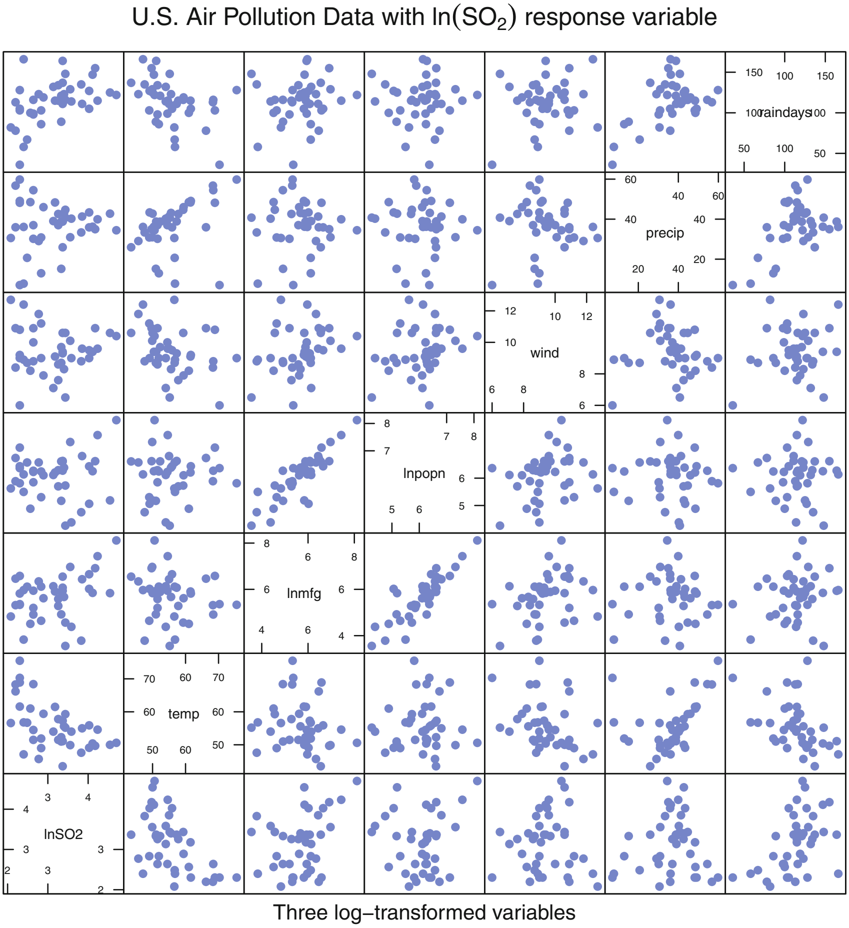

Fig. 9.12 Scatterplot matrices for air pollution data with improved symmetry after a log transformation of the three variables: `SO2`, `mfgfirms`, `popn`.

pollution-causing heavy industry than colder temperature cities well north of the Sunbelt. We are not surprised that greater amounts of manufacturing are associated with more pollution or that wind dissipates pollution.

We can arrive at the same model without a formal stepwise approach. We notice from Figure 9.12 that `lnmfg` and `lnpopn` are highly correlated, so it would be redundant to include both in the model. The variables `precip` and `raindays` seem quite similar, so again, it is unlikely that both are needed. Inspection of the C_p plot in Figure 9.13 indicates that the model with `temp`, `lnmfg`, `wind`, and `precip` has C_p close to p and only one member of each pair of similar predictors.

Table 9.14 Stepwise regression analysis of U.S. air pollution data. See also Figure 9.13.

```
> usair.regsubset <- leaps::regsubsets(
+      lnSO2 ~ lnmfg + lnpopn + precip + raindays + temp + wind,
+      data=usair, nbest=2)

> usair.subsets.Summary <- summaryHH(usair.regsubset)

> tmp <- (usair.subsets.Summary$cp <= 10)

> usair.subsets.Summary[tmp,]
             model p   rsq   rss adjr2   cp    bic stderr
5          lnm-t-w 4 0.456 10.74 0.412 8.15 -10.09  0.539
6            p-t-w 4 0.446 10.94 0.401 8.93  -9.33  0.544
7        lnm-p-t-w 5 0.543  9.02 0.492 3.58 -13.51  0.501
8        lnm-r-t-w 5 0.513  9.61 0.459 5.82 -10.93  0.517
9    lnm-lnp-p-t-w 6 0.550  8.88 0.486 5.03 -10.46  0.504
10     lnm-p-r-t-w 6 0.543  9.02 0.477 5.58  -9.80  0.508
11 lnm-lnp-p-r-t-w 7 0.550  8.87 0.471 7.00  -6.78  0.511

Model variables with abbreviations
                                                  model
t                                                  temp
r                                              raindays
p-t                                         precip-temp
r-t                                       raindays-temp
lnm-t-w                                 lnmfg-temp-wind
p-t-w                                  precip-temp-wind
lnm-p-t-w                        lnmfg-precip-temp-wind
lnm-r-t-w                      lnmfg-raindays-temp-wind
lnm-lnp-p-t-w               lnmfg-lnpopn-precip-temp-wind
lnm-p-r-t-w               lnmfg-precip-raindays-temp-wind
lnm-lnp-p-r-t-w lnmfg-lnpopn-precip-raindays-temp-wind

model with largest adjr2
7

Number of observations
41
```

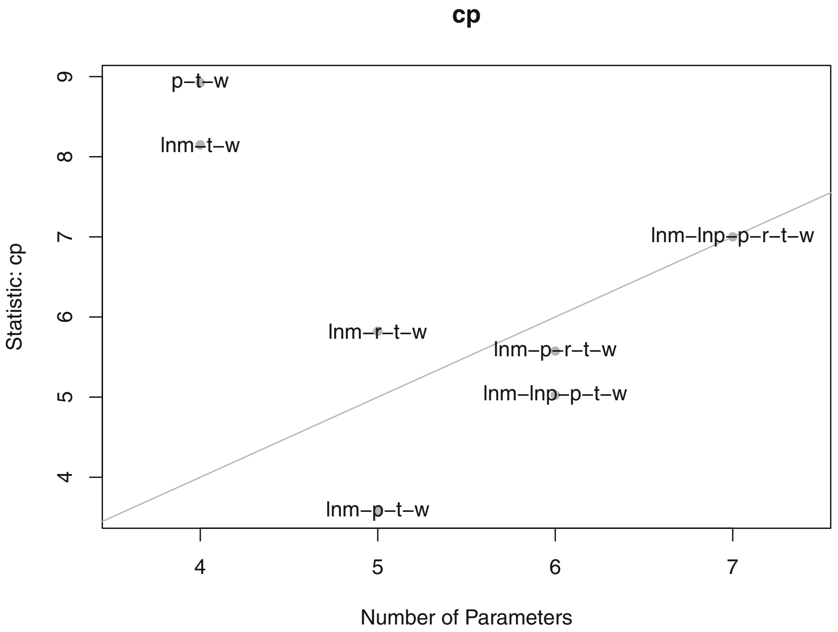

Fig. 9.13 C_p plot. Model "`lnm-p-t-w`" (`lnmfg, precip, temp, wind`) has the smallest C_p value and the largest R^2_{adj}. See also Table 9.14.

9.15 Exercises

We recommend that for all exercises involving a data set, you begin by examining a scatterplot matrix of the variables.

9.1. Use matrix algebra to prove the assertion in Equation (9.11) that the sum of the calculated residuals is also zero in multiple regression. We proved the assertion for simple linear regression in Exercise 8.9.

Hint: Write the vector of residuals as $e = (I - H)Y$, verify that $X = HX$, and use the fact that in a model with a nonzero intercept coefficient, as in Equation (9.1) and following, the first column of X is a column of ones.

9.2. Davies and Goldsmith (1972), reprinted in Hand et al. (1994), investigated the relationship between the `abrasion` loss of samples of rubber (in grams per hour) as a function of `hardness` and tensile `strength` (kg/cm^2). Higher values of `hardness` indicate harder rubber. The data are accessed as `data(abrasion)`.

Table 9.15 Fit of recommended model for U.S. air pollution data.

```
> usair.lm7 <- lm.regsubsets(usair.regsubset, 7)

> anova(usair.lm7)
Analysis of Variance Table

Response: lnSO2
          Df Sum Sq Mean Sq F value  Pr(>F)
lnmfg      1   2.26    2.26    9.00  0.0049 **
precip     1   0.03    0.03    0.11  0.7396
temp       1   6.21    6.21   24.77 1.6e-05 ***
wind       1   2.21    2.21    8.84  0.0052 **
Residuals 36   9.02    0.25
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

> summary(usair.lm7)

Call:
lm(formula = lnSO2 ~ lnmfg + precip + temp + wind, data = usair)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8965 -0.3405 -0.0854  0.2963  1.0321

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.89138    1.07009    6.44  1.8e-07 ***
lnmfg        0.23999    0.08677    2.77   0.0089 **
precip       0.01930    0.00738    2.62   0.0129 *
temp        -0.07304    0.01283   -5.69  1.8e-06 ***
wind        -0.18437    0.06203   -2.97   0.0052 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.501 on 36 degrees of freedom
Multiple R-squared:  0.543,Adjusted R-squared:  0.492
F-statistic: 10.7 on 4 and 36 DF,  p-value: 8.23e-06


> vif(usair.lm7)
 lnmfg precip   temp   wind
 1.115  1.204  1.373  1.253
```

a. Produce a scatterplot matrix of these data. Based on this plot, does it appear that `strength` would be helpful in explaining `abrasion`?

b. Calculate the fitted regression equation.

c. Find a 95% prediction interval for the abrasion corresponding to a new rubber sample having hardness 60 and strength 200.

9.3. Narula and Wellington (1977) provide data on the sale price of 28 houses in Erie, Pennsylvania, in the early 1970s, along with 11 possible predictors of these prices. The data are accessed as `data(hpErie)`. The variables are:

`price:` price in $100's

`taxes:` taxes in dollars

`bathrm:` number of bathrooms

`lotsize:` lot size in square feet

`sqfeet:` square footage of living space

`garage:` number of cars for which there is garage space

`rooms:` number of rooms

`bedrm:` number of bedrooms

`age:` age in years

`type:` type of house
brick, brick and frame, aluminum and frame, frame

`style:` 2 story, 1.5 story, ranch

`fireplac:` number of fireplaces

In parts a–d, exclude factors `type` and `style` from the analysis.

a. Produce a scatterplot matrix for these data. Notice that two houses had a sale price much higher than the others.

b. Use a stepwise regression technique to formulate a parsimonious model for sale price. Do the arithmetic signs of your model's regression coefficients make economic sense?

c. Redo part a with the two large-priced houses excluded. Compare your answer with that of part a.

d. Add a new variable `sqfeetsq` (defined as the square of `sqfeet`) to the list of variables. Perform the stepwise regression allowing for this new variable. Does its presence change the preferred model?

e. For the model you found in part d, provide plots of the residuals vs the fitted response for each of the 12 combinations of `type` and `style`. Use Figure 13.1 and its code included in `HHscriptnames(13)` as a template for constructing these plots. Based on these plots, does it appear that including either of the variables `type` or `style` would contribute to the model fit?

9.4. World Almanac and Book of Facts (2001) lists the winning `times` for the men's 1500-meter sprint event for the Olympics from `years` 1900 through 2000. The data are accessed as `data(sprint)`.

a. Plot the data.

b. Use linear regression to fit the winning times to the year, producing a plot of the residuals vs the fitted values.

c. The residual plot suggests that an additional predictor should be added to the model. Refit this expanded model and compare it with the model you found in part b.

d. Interpret the sign of the coefficient of this additional predictor.

9.5. A company wished to model the number of `minutes` required to unload shipments of drums of chemicals at its warehouse as a function of the number of `drums` and the total shipment `weight` in hundreds of pounds. The data from 20 consecutive shipments, from Neter et al. (1996), are accessed as `data(shipment)`.

a. Regress `minutes` on `drums` and `weight`, storing the residuals.

b. Interpret the regression coefficients of `drum` and `weight`.

c. Provide and discuss plots of the residuals against the fitted values and both predictors, and a normality plot.

d. Provide a 90% prediction interval for the time it would take to unload a new shipment of 10 drums weighing 1000 pounds.

9.6. The dataset `data(uscrime)` is introduced in Exercise 4.3. Use a stepwise regression approach to develop a model to explain `R`. Your solution should not have a collinearity problem, all predictor regression coefficients should be significantly different from zero and have an arithmetic sign consistent with common knowledge of the model variables, and no standard residual plots should display a problem.

9.7. It is desired to model the `manhours` needed to operate living quarters for U.S. Navy bachelor officers. Candidate explanatory variables are listed below. The data in `data(manhours)` are from Freund and Littell (1991) and Myers (1990), and originally from Navy (1979). Perform a thorough regression analysis, including relevant plots. Note that at least initially, there is a minor collinearity problem to be addressed. Show that, no matter how the collinearity is addressed, the predictions are similar. Only the interpretation of the effects of the x-variables is affected.

`manhours`: monthly manhours needed to operate the establishment

`occupanc`: average daily occupancy

`checkins`: average monthly number of check-ins

`svcdesk`: weekly hours of service desk operation

`common`: common use area, in square feet

`wings`: number of building wings

`berthing`: operational berthing capacity

`rooms`: number of rooms

9.A Appendix: Computation for Regression Analysis

`regr2.plot`

The `regr2.plot` function does the same type of plot for bivariate regression, one *y*-variable and two *x*-variables. The function is based on the `persp` perspective plotting function in R. We designed the `regr2.plot` function with options to display grids for the base plane and the two back planes in addition to the observed points and the regression plane and the fitted points. We turned off the default plot of the 3-dimensional box. The function `regr2.plot` uses the functions defined in our function `persp.hh.s`.