9

Regression: Basics

9.1 Introduction

Regression is one of the most widely used of all statistical methods. For univariate regression, the available data are one response variable and p predictor variables, all measured on each of n observations. We let Y denote the response variable and $X_1, \ldots, X_p$ be the predictor or explanatory variables. Also, Y_i and $X_{i,1}, \ldots, X_{i,p}$ are the values of these variables for the ith observation. The goals of regression modeling include the investigation of how Y is related to $X_1, \ldots, X_p$, estimation of the conditional expectation of Y given $X_1, \ldots, X_p$, and prediction of future Y values when the corresponding values of $X_1, \ldots, X_p$ are already available. These goals are closely connected.

The *multiple linear regression* model relating Y to the predictor or regressor variables is

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i, \tag{9.1}$$

where ϵ_i is called the noise, disturbances, or errors. The adjective "multiple" refers to the predictor variables. Multivariate regression, which has more than one response variable, is covered in Chap. 18. The ϵ_i are often called "errors" because they are the prediction errors when Y_i is predicted by $\beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}$. It is assumed that

$$E(\epsilon_i | X_{i,1}, \ldots, X_{i,p}) = 0, \tag{9.2}$$

which, with (9.1), implies that

$$E(Y_i | X_{i,1}, \ldots, X_{i,p}) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}.$$

The parameter β_0 is the intercept. The regression coefficients $\beta_1, \ldots, \beta_p$ are the slopes. More precisely, β_j is the partial derivative of the expected response with respect to the jth predictor:

D. Ruppert, D.S. Matteson, *Statistics and Data Analysis for Financial Engineering*, Springer Texts in Statistics,
DOI 10.1007/978-1-4939-2614-5_9

$$\beta_j = \frac{\partial\, E(Y_i|X_{i,1},\ldots,X_{i,p})}{\partial\, X_{i,j}}.$$

Therefore, β_j is the change in the expected value of Y_i when $X_{i,j}$ changes one unit. It is assumed that the noise is i.i.d. white so that

$$\epsilon_1,\ldots,\epsilon_n \text{ are i.i.d. with mean 0 and variance } \sigma_\epsilon^2. \tag{9.3}$$

Often the ϵ_is are assumed to be normally distributed, which with (9.3) implies Gaussian white noise.

For the reader's convenience, the assumptions of the linear regression model are summarized:

1. linearity of the conditional expectation: $E(Y_i|X_{i,1},\ldots,X_{i,p}) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}$;
2. independent noise: $\epsilon_1,\ldots,\epsilon_n$ are independent;
3. constant variance: $\text{Var}(\epsilon_i) = \sigma_\epsilon^2$ for all i;
4. Gaussian noise: ϵ_i is normally distributed for all i.

This chapter and, especially, the next two chapters discuss methods for checking these assumptions, the consequences of their violations, and possible remedies when they do not hold.

9.2 Straight-Line Regression

Straight-line regression is linear regression with only one predictor variable. The model is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \tag{9.4}$$

where β_0 and β_1 are the unknown intercept and slope of the line and ϵ_i is called the noise or error.

9.2.1 Least-Squares Estimation

The regression coefficients can be estimated by the *method of least squares*. The least-squares estimates are the values of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that minimize

$$\sum_{i=1}^{n} \left\{ Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i) \right\}^2. \tag{9.5}$$

Geometrically, we are minimizing the sum of the squared lengths of the vertical lines in Fig. 9.1. The data points are shown as asterisks. The vertical lines connect the data points and the predictions using the linear equation. The predictions themselves are called the *fitted values* or "y-hats" and shown as open circles. The differences between the Y-values and the fitted values are called the *residuals*. Using calculus to minimize (9.5), one can show that

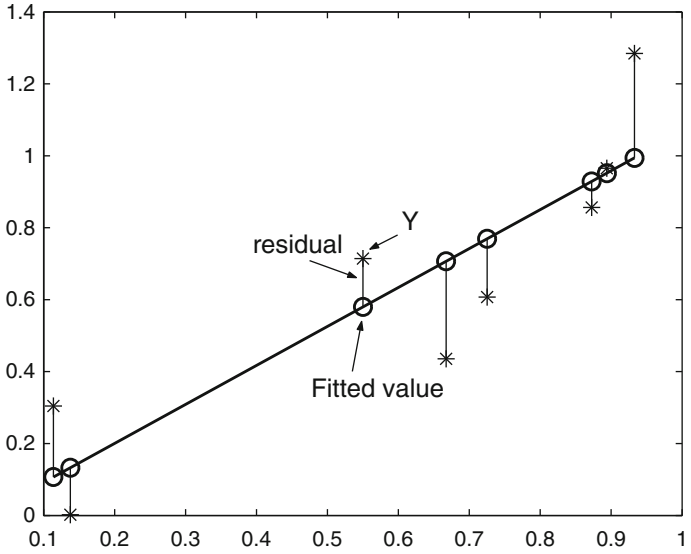

Fig. 9.1. *Least-squares estimation. The vertical lines connect the data (*) and the fitted values (o) represent the residuals. The least-squares line is defined as the line making the sum of the squared residuals as small as possible.*

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\sum_{i=1}^n Y_i (X_i - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2}. \tag{9.6}$$

and

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}. \tag{9.7}$$

The *least-squares line* is

$$\begin{aligned} \widehat{Y} &= \widehat{\beta}_0 + \widehat{\beta}_1 X = \overline{Y} + \widehat{\beta}_1 (X - \overline{X}) \\ &= \overline{Y} + \left\{ \frac{\sum_{i=1}^n (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2} \right\} (X - \overline{X}) \\ &= \overline{Y} + \frac{s_{XY}}{s_X^2} (X - \overline{X}), \end{aligned}$$

where $s_{XY} = (n-1)^{-1} \sum_{i=1}^n (Y_i - \overline{Y})(X_i - \overline{X})$ is the sample covariance between X and Y and s_X^2 is the sample variance of X.

Example 9.1. Weekly interest rates — least-squares estimates

Weekly interest rates from February 16, 1977, to December 31, 1993, were obtained from the Federal Reserve Bank of Chicago. Figure 9.2 is a plot of

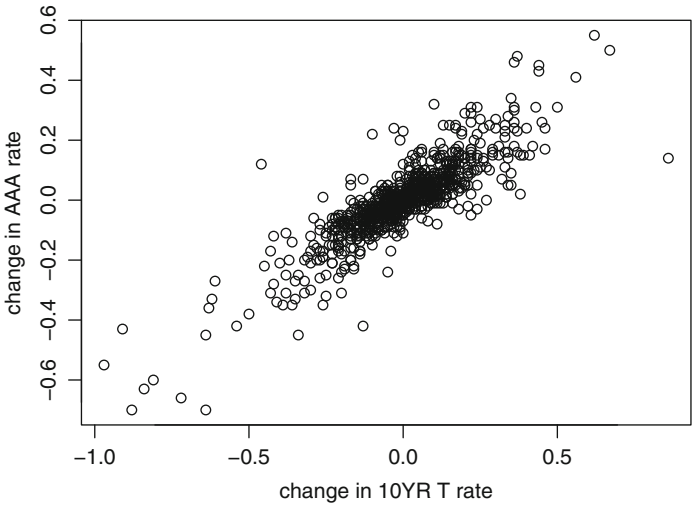

Fig. 9.2. *Changes in Moody's seasoned corporate AAA bond yields plotted against changes in* 10*-year Treasury constant maturity rate. Data from Federal Reserve Statistical Release H.*15 *and were taken from the Chicago Federal Bank's website.*

changes in the 10-year Treasury constant maturity rate and changes in the Moody's seasoned corporate AAA bond yield. The plot looks linear, so we try linear regression using R's `lm()` function. The code is:

```
options(digits = 3)
summary(lm(aaa_dif ~ cm10_dif))
```

The code `aaa_dif ~ cm10_dif` is an example of a formula in R with the outcome variable to the left of "~" and the explanatory variables to the right of "~." In this example, there is only one explanatory variable. In cases where there are multiple explanatory variables, they are separated by "+". Here is the output.

```
Call:
lm(formula = aaa_dif ~ cm10_dif)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.000109   0.002221   -0.05     0.96
cm10_dif     0.615762   0.012117   50.82   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 0.066 on 878 degrees of freedom
Multiple R-Squared: 0.746,      Adjusted R-squared: 0.746
F-statistic: 2.58e+03 on 1 and 878 DF,  p-value: <2e-16
```

From the output we see that the least-squares estimates of the intercept and slope are −0.000109 and 0.616. The `Residual standard error` is 0.066; this is what we call $\widehat{\sigma}_\epsilon$ or s, the estimate of σ_ϵ; see Sect. 9.3. The remaining items of the output are explained shortly. □

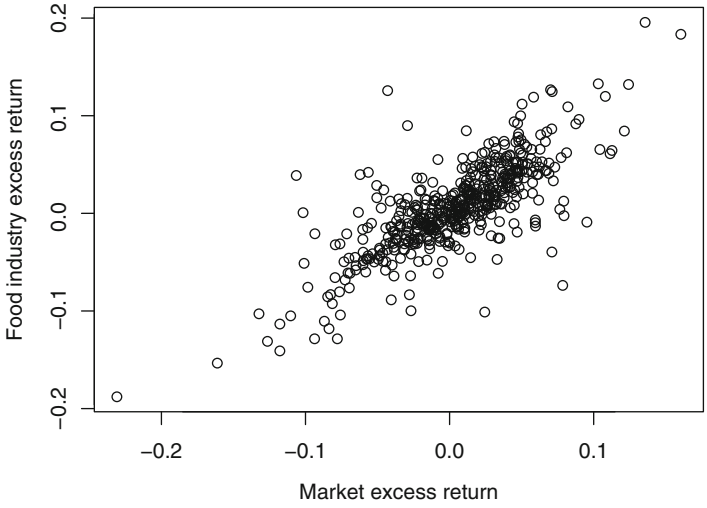

Fig. 9.3. *Plot of excess returns on the food industry versus excess returns on the market. Data from the data set* `Capm` *in* `R`*'s* `Ecdat` *package.*

Example 9.2. Excess returns on the food sector and the market portfolio

The excess return on a security or market index is the return minus the risk-free interest rate. An important application of linear regression in finance is the regression of the excess return of an asset or market sector on the excess return of the entire market. This type of application will be discussed much more fully in Chap. 17. In this example, we will regress the excess monthly return of the food sector (`rfood`) on the excess monthly return of the market portfolio (`rmrf`). The data are in `R`'s `Capm` data set in the `Ecdat` package and are plotted in Fig. 9.3. The returns are expressed as percentages in the data set but have been converted to fractions in this example. The output from `lm` is

```
Call:
lm(formula = rfood ~ rmrf)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  0.00339    0.00128    2.66   0.0081 **
rmrf         0.78342    0.02835   27.63   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.0289 on 514 degrees of freedom
Multiple R-Squared: 0.598,      Adjusted R-squared: 0.597
F-statistic:  763 on 1 and 514 DF,  p-value: <2e-16
```

Thus, the fitted regression equation is

$$\texttt{rfood} = 0.00339 + 0.78342\,\texttt{rmrf} + \epsilon,$$

and $\widehat{\sigma}_\epsilon = 0.0289$. □

9.2.2 Variance of $\widehat{\beta}_1$

It is useful to have a formula for the variance of an estimator to show how the estimator's precision depends on various aspects of the data such as the sample size and the values of the predictor variables. Fortunately, it is easy to derive a formula for the variance of $\widehat{\beta}_1$. By (9.6), we can write $\widehat{\beta}_1$ as a weighted average of the responses

$$\widehat{\beta}_1 = \sum_{i=1}^{n} w_i Y_i,$$

where w_i is the weight given by

$$w_i = \frac{X_i - \overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

We consider $X_1, \ldots, X_n$ as fixed, so if they are random we are conditioning upon their values. From the assumptions of the regression model, it follows that $\mathrm{Var}(Y_i|X_1, \ldots, X_n) = \sigma_\epsilon^2$ and $Y_1, \ldots, Y_n$ are conditionally uncorrelated. Therefore,

$$\mathrm{Var}(\widehat{\beta}_1|X_1, \ldots, X_n) = \sigma_\epsilon^2 \sum_{i=1}^{n} w_i^2 = \frac{\sigma_\epsilon^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\sigma_\epsilon^2}{(n-1)s_X^2}. \tag{9.8}$$

It is worth taking some time to examine this formula. First, the numerator σ_ϵ^2 is simply the variance of the ϵ_i. This is not surprising. More variability in the noise means more variable estimators. The denominator shows us that the variance of $\widehat{\beta}_1$ is inversely proportional to $(n-1)$ and to s_X^2. So the precision of $\widehat{\beta}_1$ increases as σ_ϵ^2 is reduced, n is increased, or s_X^2 is increased. Why does increasing s_X^2 decrease $\mathrm{Var}(\widehat{\beta}_1|X_1, \ldots, X_n)$? The reason is that increasing s_X^2 means that the X_i are spread farther apart, which makes the slope of the line easier to estimate.

Example 9.3. Optimal sampling frequencies for regression

Here is an important application of (9.8). Suppose that we have two stationary time series, X_t and Y_t, and we wish to regress Y_t on X_t. We have just seen examples of this. A significant practical question is whether one should use daily or weekly data, or perhaps even monthly or quarterly data. Does it matter which sampling frequency we use? The answer is "yes" and the highest possible sampling frequency gives the most precise estimate of the slope. To understand why this is so, we compare daily and weekly data. Assume that the X_t and Y_t are white noise sequences. Since a weekly log return is simply the sum of the five daily log returns within a week, σ_ϵ^2 and s_X^2 will each increase by a factor of five if we change from daily to weekly log returns, so the ratio σ_ϵ^2/s_X^2 will not change. However, by changing from daily to weekly log returns, $(n-1)$ is reduced by approximately a factor of five. The result is that $\mathrm{Var}(\widehat{\beta}_1|X_1,\ldots,X_n)$ is approximately five times smaller using daily rather than weekly log returns. Similarly, $\mathrm{Var}(\widehat{\beta}_1|X_1,\ldots,X_n)$ is about four times larger using monthly rather than weekly returns.

The obvious conclusion is that one should use the highest sampling frequency available, which is often daily returns. We have assumed that the X_t and Y_t are white noise in order to simplify the calculations, but this conclusion still holds if they are stationary but autocorrelated. (Autocorrelation is discussed in Chap. 12.) However, the noise series, that is $\epsilon_i,\ i=1,\ldots,$ in Eq. (9.4) needs to be uncorrelated. If the noise is autocorrelated and becomes more highly correlated as the sampling frequency increases, then this conclusion need not hold. There may be a point of diminishing returns where more frequent sampling does not improve estimation accuracy. □

9.3 Multiple Linear Regression

The multiple linear regression model is

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i.$$

The least-squares estimates are the values $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$ that minimize

$$\sum_{i=1}^{n} \left\{ Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_{i,1} + \cdots + \widehat{\beta}_p X_{i,p}) \right\}^2. \tag{9.9}$$

Calculation of the least-squares estimates is discussed in Sect. 11.1. For applications, the technical details are not important, since software for least-squares estimation is readily available.

The ith *fitted value* is

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{i,1} + \cdots + \widehat{\beta}_p X_{i,p} \tag{9.10}$$

and estimates $E(Y_i|X_{i,1},\ldots,X_{i,p})$. The ith residual is

$$\widehat{\epsilon}_i = Y_i - \widehat{Y}_i = Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_{i,1} + \cdots + \widehat{\beta}_p X_{i,p}) \tag{9.11}$$

and estimates ϵ_i. It is worth noting that (9.11) can be re-expressed as

$$Y_i = \widehat{Y}_i + \widehat{\epsilon}_i. \tag{9.12}$$

An unbiased estimate of σ_ϵ^2 is

$$\widehat{\sigma}_\epsilon^2 = \frac{\sum_{i=1}^n \widehat{\epsilon}_i^2}{n-1-p}. \tag{9.13}$$

The denominator in (9.13) is the sample size minus the number of regression coefficients that are estimated.

Example 9.4. Multiple linear regression with interest rates

As an example, we continue the analysis of the weekly interest-rate data but now with changes in the 30-year Treasury rate (`cm30_dif`) and changes in the Federal funds rate (`ff_dif`) as additional predictors. Thus $p = 3$. Figure 9.4 is a scatterplot matrix of the four time series. There is a strong linear relationship between all pairs of `aaa_dif`, `cm10_dif`, and `cm30_dif`, but `ff_dif` is not strongly related to the other series. The code is

```
summary(lm(aaa_dif ~ cm10_dif + cm30_dif + ff_dif))
```

The `lm()` output for this regression is

```
Call:
lm(formula = aaa_dif ~ cm10_dif + cm30_dif + ff_dif)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.07e-05   2.18e-03   -0.04     0.97
cm10_dif     3.55e-01   4.51e-02    7.86  1.1e-14 ***
cm30_dif     3.00e-01   5.00e-02    6.00  2.9e-09 ***
ff_dif       4.12e-03   5.28e-03    0.78     0.44
---

Residual standard error: 0.0646 on 876 degrees of freedom
Multiple R-Squared: 0.756,      Adjusted R-squared: 0.755
F-statistic:  906 on 3 and 876 DF,  p-value: <2e-16
```

We see that $\widehat{\beta}_0 = -9.07 \times 10^{-05}$, $\widehat{\beta}_1 = 0.355$, $\widehat{\beta}_2 = 0.300$, and $\widehat{\beta}_3 = 0.00412$. □

A commonly used special case of multiple regression is the polynomial regression model which uses powers of the predictors as well as the predictors

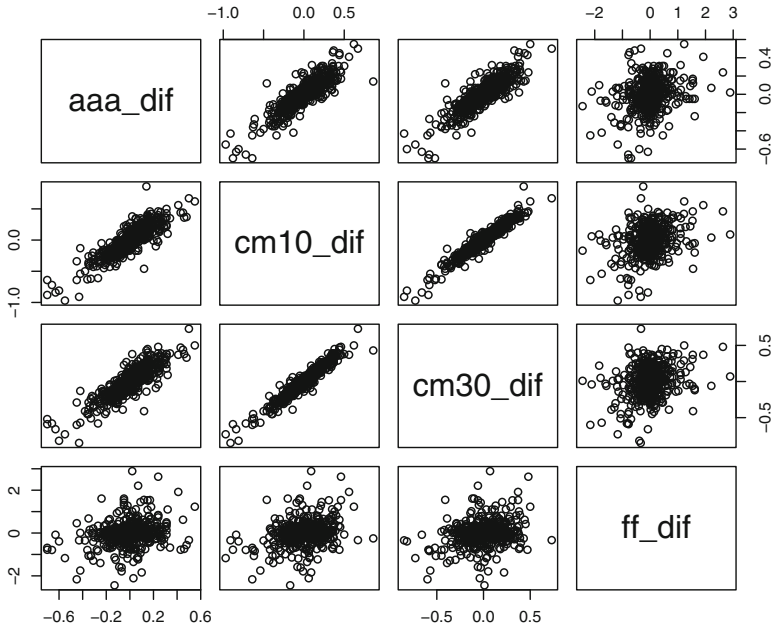

Fig. 9.4. *Scatterplot matrix of the changes in four weekly interest rates. The variable* `aaa_dif` *is the response in Example* 9.4.

themselves. For example, when there is one X-variable, the p-degree polynomial regression model is

$$Y_i = \beta_0 + \beta_1 X_i + \cdots + \beta_p X_i^p + \epsilon_i.$$

As another example, the quadratic regression model with two predictors is

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,1}^2 + \beta_3 X_{i,1} X_{i,2} + \beta_4 X_{i,2} + \beta_5 X_{i,2}^2 + \epsilon_i.$$

9.3.1 Standard Errors, *t*-Values, and *p*-Values

In this section we explain the use of several statistics included in regression output. We use the output in Example 9.4 as an illustration.

As noted before, the estimated coefficients are $\widehat{\beta}_0 = -9.07 \times 10^{-05}$, $\widehat{\beta}_1 = 0.355$, $\widehat{\beta}_2 = 0.300$, and $\widehat{\beta}_3 = 0.00412$. Each of these coefficients has three other statistics associated with it.

- The standard error (SE), which is the estimated standard deviation of the least-squares estimator, tells us the precision of the estimator.
- The t-value, is the t-statistic for testing that the coefficient is 0. The t-value is the ratio of the estimate to its standard error. For example, for `cm10_dif`, the t-value is $7.86 = 0.355/0.0451$.

- The p-value (`Pr > |t|` in the `lm()` output), associated with testing the null hypothesis that the coefficient is 0 versus the alternative that it is not 0. If a p-value for a slope parameter is small, as it is here for β_1, then this is evidence that the corresponding coefficient is *not* 0, which means that the predictor has a *linear* relationship with the response.

It is important to keep in mind that the p-value only tells us if there is a linear relationship. The existence of a linear relationship between Y_i and $X_{i,j}$ means only that the linear predictor of Y_i has a nonzero slope on $X_{i,j}$, or, equivalently, that partial correlation between $X_{i,j}$ and Y_i is not zero. (The partial correlation between two variables is their correlation when all other variables are held fixed.) When the p-value is small (so a linear relationship exists), there could also be a strong nonlinear deviation from the linear relationship as in Fig. A.4g. Moreover, when the p-value is large (so no linear relationship exists), there could still be a strong nonlinear relationship in Fig. A.4f. Because of the potential for nonlinear relationships to go undetected in a linear regression analysis, graphical analysis of the data (e.g., Fig. 9.4) and residual analysis (see Chap. 10) are essential.

The p-values for β_1 and β_2 are *very* small, so we can conclude that these slopes are *not* 0. The p-value is large (0.97) for β_0, so we would not reject the hypothesis that the intercept is 0.

Similarly, we would not reject the null hypothesis that β_3 is zero. Stated differently, we can accept the null hypothesis that, conditional on `cm10_dif` and `cm30_dif`, `aaa_dif` and `ff_dif` are not linearly related. This result should *not* be interpreted as stating that `aaa_dif` and `ff_dif` are unrelated, but only that `ff_dif` is not useful for predicting `aaa_dif` when `cm10_dif` and `cm30_dif` are included in the regression model. (In fact, `aaa_dif` and `ff_dif` have a correlation of 0.25 (this is the full, not partial, correlation) and the linear regression of `aaa_dif` on `ff_dif` alone is highly significant; the p-value for testing that the slope is zero is 5.158×10^{-14}.)

Since the Federal Funds rate is a short-term (overnight) rate, it is not surprising that `ff_dif` is less useful than changes in the 10- and 30-year Treasury rates for predicting `aaa_dif`.

For regression with one predictor variable, by (9.8) the standard error of $\widehat{\beta}_1$ is $\widehat{\sigma}_\epsilon / \sqrt{\sum_{i=1}^n (X_i - \overline{X})^2}$. When there are more than two predictor variables, formulas of standard errors are more complex and are facilitated by the use of matrix notation. Because standard errors can be computed with standard software such as `lm`, the formulas are not needed for applications and so are postponed to Sect. 11.1.

9.4 Analysis of Variance, Sums of Squares, and R^2

9.4.1 ANOVA Table

Certain results of a regression fit are often displayed in an *analysis of variance table*, also called the ANOVA or AOV table. The idea behind the ANOVA table is to describe how much of the variation in Y is predictable if one knows $X_1, \ldots, X_p$. Here is the ANOVA table for the model in Example 9.4.

```
> anova(lm(aaa_dif ~ cm10_dif + cm30_dif + ff_dif))
Analysis of Variance Table

Response: aaa_dif
           Df Sum Sq Mean Sq F value  Pr(>F)
cm10_dif    1  11.21   11.21 2682.61 < 2e-16 ***
cm30_dif    1   0.15    0.15   35.46 3.8e-09 ***
ff_dif      1 0.0025  0.0025    0.61    0.44
Residuals 876   3.66  0.0042
---
```

The total variation in Y can be partitioned into two parts: the variation that can be predicted by $X_1, \ldots, X_p$ and the variation that cannot be predicted. The variation that can be predicted is measured by the regression sum of squares, which is

$$\text{regression SS} = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2.$$

The regression sum of squares for the model that uses only `cm10_dif` is in the first row of the ANOVA table and is 11.21. The entry, 0.15, in the second row is the increase in the regression sum of squares when `cm30_dif` is added to the model. Similarly, 0.0025 is the increase in the regression sum of squares when `ff_dif` is added. Thus, rounding to two decimal places, 11.36 = 11.21 + 0.15 + 0.00 is the regression sum of squares with all three predictors in the model.

The amount of variation in Y that cannot be predicted by a linear function of $X_1, \ldots, X_p$ is measured by the residual error sum of squares, which is the sum of the squared residuals; i.e.,

$$\text{residual error SS} = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2.$$

In the ANOVA table, the residual error sum of squares is in the last row and is 3.66. The total variation is measured by the total sum of squares (total SS), which is the sum of the squared deviations of Y from its mean; that is,

$$\text{total SS} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2. \tag{9.14}$$

It can be shown algebraically that

$$\text{total SS} = \text{regression SS} + \text{residual error SS}. \tag{9.15}$$

Therefore, in Example 9.4, the total SS is 11.36 + 3.66 = 15.02.

R-squared, denoted by R^2, is

$$R^2 = \frac{\text{regression SS}}{\text{total SS}} = 1 - \frac{\text{residual error SS}}{\text{total SS}}$$

and measures the proportion of the total variation in Y that can be linearly predicted by X. In the example, R^2 is $0.746 = 11.21/15.02$ if only `cm10_dif` is the model and is $11.36/15.02 = 0.756$ if all three predictors are in the model. This value can be found in the output displayed in Example 9.4.

When there is only a single X variable, then $R^2 = r^2_{XY} = r^2_{\widehat{Y}Y}$, where r_{XY} and $r_{\widehat{Y}Y}$ are the sample correlations between Y and X and between Y and the predicted values, respectively. Put differently, R^2 is the squared correlation between Y and X and also between Y and $\widehat{Y}$. When there are multiple predictors, then we still have $R^2 = r^2_{\widehat{Y}Y}$. Since $\widehat{Y}$ is a linear combination of the X variables, R can be viewed as the "multiple" correlation between Y and many Xs. The residual error sum of squares is also called the error sum of squares or sum of squared errors and is denoted by SSE.

It is important to understand that sums of squares in an ANOVA table depend upon the order of the predictor variables in the regression, because the sum of squares for any variable is the increase in the regression sum of squares when that variable is added to the predictors already in the model.

The table below has the same variables as before, but the order of the predictor variables is reversed. Now that `ff_dif` is the first predictor, its sum of squares is much larger than before and its p-value is highly significant; before it was nonsignificant, only 0.44. The sum of squares for `cm30_dif` is now much larger than that of `cm10_dif`, the reverse of what we saw earlier, since `cm10_dif` and `cm30_dif` are highly correlated and the first of them in the list of predictors will have the larger sum of squares.

```
> anova(lm(aaa_dif ~ ff_dif + cm30_dif + cm10_dif))
Analysis of Variance Table

Response: aaa_dif
           Df Sum Sq Mean Sq F value  Pr(>F)
ff_dif      1   0.94    0.94   224.8 < 2e-16 ***
cm30_dif    1  10.16   10.16  2432.1 < 2e-16 ***
cm10_dif    1   0.26    0.26    61.8 1.1e-14 ***
Residuals 876   3.66  0.0042
```

The lesson here is that an ANOVA table is most useful for assessing the effects of adding predictors in some natural order. Since AAA bonds have maturities closer to 10 than to 30 years, and since the Federal Funds rate is an overnight rate, it made sense to order the predictors as `cm10_dif`, `cm30_dif`, and `ff_dif` as done initially.

9.4.2 Degrees of Freedom (DF)

There are degrees of freedom (DF) associated with each of these sources of variation. The degrees of freedom for regression is p, which is the number of predictor variables. The total degrees of freedom is $n-1$. The residual error degrees of freedom is $n-p-1$. Here is a way to think of degrees of freedom. Initially, there are n degrees of freedom, one for each observation. Then one degree of freedom is allocated to estimation of the intercept. This leaves a total of $n-1$ degrees of freedom for estimating the effects of the X variables and σ_ϵ^2. Each regression parameter uses one degree of freedom for estimation. Thus, there are $(n-1)-p$ degrees of freedom remaining for estimation of σ_ϵ^2 using the residuals. There is an elegant geometrical theory of regression where the responses are viewed as lying in an n-dimensional vector space and degrees of freedom are the dimensions of various subspaces. However, there is not sufficient space to pursue this subject here.

9.4.3 Mean Sums of Squares (MS) and *F*-Tests

As just discussed, every sum of squares in an ANOVA table has an associated degrees of freedom. The ratio of the sum of squares to the degrees of freedom is the mean sum of squares:

$$\text{mean sum of squares} = \frac{\text{sum of squares}}{\text{degrees of freedom}}.$$

The residual mean sum of squares is the unbiased estimate σ_ϵ^2 given by (9.13); that is,

$$\begin{aligned}\widehat{\sigma}_\epsilon^2 &= \frac{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}{n-1-p} \\ &= \text{residual mean sum of squares} \\ &= \frac{\text{residual error SS}}{\text{residual degrees of freedom}}.\end{aligned} \tag{9.16}$$

Other mean sums of squares are used in testing. Suppose we have two models, I and II, and the predictor variables in model I are a subset of those in model II, so that model I is a submodel of II. A common null hypothesis is that the data are generated by model I. Equivalently, in model II the slopes are zero for variables not also in model I. To test this hypothesis, we use the excess regression sum of squares of model II relative to model I:

$$\begin{aligned}\text{SS(II}\,|\,\text{I)} &= \text{regression SS for model II} - \text{regression SS for model I} \\ &= \text{residual SS for model I} - \text{residual SS for model II.}\end{aligned} \tag{9.17}$$

Equality (9.17) holds because (9.15) is true for all models and, in particular, for both model I and model II. The degrees of freedom for SS(II | I) is the number

of extra predictor variables in model II compared to model I. The mean square is denoted as MS(II | I). Stated differently, if p_{I} and p_{II} are the number of parameters in models I and II, respectively, then $\mathrm{df}_{\mathrm{II}|\,\mathrm{I}} = p_{\mathrm{II}} - p_{\mathrm{I}}$ and MS(II | I) = SS(II | I)/$\mathrm{df}_{\mathrm{II}|\,\mathrm{I}}$. The F-statistic for testing the null hypothesis is

$$F = \frac{\mathrm{MS(II|I)}}{\widehat{\sigma}_{\epsilon}^2},$$

where $\widehat{\sigma}_{\epsilon}^2$ is the mean residual sum of squares for model II. Under the null hypothesis, the F-statistic has an F-distribution with $\mathrm{df}_{\mathrm{II}|\,\mathrm{I}}$ and $n - p_{\mathrm{II}} - 1$ degrees of freedom and the null hypothesis is rejected if the F-statistic exceeds the α-upper quantile of this F-distribution.

Example 9.5. Weekly interest rates—Testing the one-predictor versus three-predictor model

In this example, the null hypothesis is that, in the three-predictor model, the slopes for `cm30_dif` and `ff_dif` are zero. The F-test can be computed using `R`'s `anova` function. The output is

```
Analysis of Variance Table

Model 1: aaa_dif ~ cm10_dif
Model 2: aaa_dif ~ cm10_dif + cm30_dif + ff_dif
  Res.Df  RSS  Df Sum of Sq    F  Pr(>F)
1    878 3.81
2    876 3.66   2      0.15 18.0 2.1e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1
```

In the last row, the entry 2 in the "`Df`" column is the difference between the two models in the number of parameters and 0.15 in the "`Sum of Sq`" column is the difference between the residual sum of squares (RSS) for the two models.

The very small p-value (2.1×10^{-8}) leads us to reject the null hypothesis and say that the result is "highly significant." It is important to be aware that this phrase refers to statistical significance. When the sample size is as large as it is here, it is common to reject the null hypothesis. The reason for this is that the null hypothesis is rarely true exactly, and with a large sample size it is highly likely that even a small deviation from the null hypothesis will be detected. Statistically significance must be distinguished from practical significance. The adjusted R^2 values for the two- and three-variable models are very similar, 0.746 and 0.755, respectively. Therefore, the rejection of the two-variable model may not be of practical importance. □

Example 9.6. Weekly interest rates—Testing a two-predictor versus three-predictor model

In this example, the null hypothesis is that, in the three predictor model, the slope `ff_dif` is zero. The F-test is again computed using `R`'s `anova` function with output:

```
Analysis of Variance Table

Model 1: aaa_dif ~ cm10_dif + cm30_dif
Model 2: aaa_dif ~ cm10_dif + cm30_dif + ff_dif
  Res.Df  RSS  Df Sum of Sq    F Pr(>F)
1    877 3.66
2    876 3.66   1    0.0025 0.61   0.44
```

The large p-value (0.44) leads us to accept the null hypothesis. Notice that this is the same as the p-value for `ff_dif` in the ANOVA table in Sect. 9.4.1. This is not a coincidence. Both p-values are the same because they are testing the same hypothesis. □

9.4.4 Adjusted R^2

R^2 is biased in favor of large models, because R^2 is always increased by adding more predictors to the model, even if they are independent of the response. Recall that

$$R^2 = 1 - \frac{\text{residual error SS}}{\text{total SS}} = 1 - \frac{n^{-1}\text{residual error SS}}{n^{-1}\text{total SS}}.$$

The bias in R^2 can be reduced by using the following "adjustment," which replaces both occurrences of n by the appropriate degrees of freedom:

$$\text{adjusted } R^2 = 1 - \frac{(n-p-1)^{-1}\text{residual error SS}}{(n-1)^{-1}\text{total SS}} = 1 - \frac{\text{residual error MS}}{\text{total MS}}.$$

The presence of p in the adjusted R^2 penalizes the criterion for the number of predictor variables, so adjusted R^2 can either increase or decrease when predictor variables are added to the model. Adjusted R^2 increases if the added variables decrease the residual sum of squares enough to compensate for the increase in p.

9.5 Model Selection

When there are many potential predictor variables, often we wish to find a subset of them that provide a parsimonious regression model. F-tests are not very suitable for model selection. One problem is that there are many possible F-tests and the joint statistical behavior of all of them is not known. For model selection, it is more appropriate to use a model selection criterion such as AIC or BIC. For linear regression models, AIC is

$$\text{AIC} = n\log(\widehat{\sigma}^2) + 2(1+p),$$

where $1+p$ is the number of parameters in a model with p predictor variables; the intercept gives us the final parameter. BIC replaces $2(1+p)$ in AIC by $\log(n)(1+p)$. The first term, $n\log(\widehat{\sigma}^2)$, is equal to, up to an additive constant that does not affect model comparisons, -2 times the log-likelihood evaluated at the MLE, assuming that the noise is Gaussian.

In addition to AIC and BIC, there are two model selection criteria specialized for regression. One is adjusted R^2, which we have seen before. Another is C_p. C_p is related to AIC and usually C_p and AIC are minimized by the same model. The primary reason for using C_p instead of AIC is that some regression software computes only C_p, not AIC—this is true of the `regsubsets()` function in `R`'s `leaps` package which will be used in the following example.

To define C_p, suppose there are M predictor variables. Let $\widehat{\sigma}^2_{\epsilon,M}$ be the estimate of σ^2_ϵ using all of them, and let SSE(p) be the sum of squares for residual error for a model with some subset of only $p \leq M$ of the predictors. As usual, n is the sample size. Then C_p is

$$C_p = \frac{SSE(p)}{\widehat{\sigma}^2_{\epsilon,M}} - n + 2(p+1). \tag{9.18}$$

Of course, C_p will depend on which particular model is used among all of those with p predictors, so the notation "C_p" may not be ideal.

With C_p, AIC, and BIC, smaller values are better, but for adjusted R^2, larger values are better.

One should not use model selection criteria blindly. Model choice should be guided by economic theory and practical considerations, as well as by model selection criteria. It is important that the final model makes sense to the user. Subject-matter expertise might lead to adoption of a model not optimal according to the criterion being used but, instead, to a model slightly below optimal but more parsimonious or with a better economic rationale.

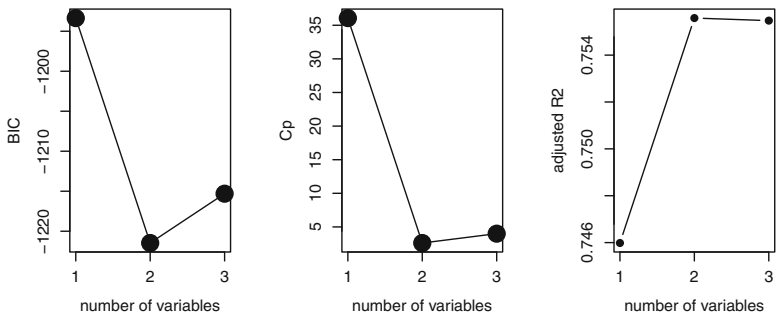

Fig. 9.5. *Changes in weekly interest rates. Plots for model selection.*

Example 9.7. Weekly interest rates—Model selection by AIC and BIC

Figure 9.5 contains plots of the number of predictors in the model versus the optimized value of a selection criterion. By "optimized value," we mean the best value among all models with the given number of predictor variables. "Best" means smallest for BIC and C_p and largest for adjusted R^2. There are three plots, one for each of BIC, C_p, and adjusted R^2. All three criteria are optimized by two predictor variables.

There are three models with two of the three predictors. The one that optimized the criteria[1] is the model with `cm10_dif` and `cm30_dif`, as can be seen in the following output from `regsubsets`. Here `"*"` indicates a variable in the model and `" "` indicates a variable not in the model, so the three rows of the table indicate that the best one-variable model is `cm10_dif` and the best two-variable model is `cm10_dif` and `cm30_dif`—the third row does not contain any real information since, with only three variables, there is only one possible three-variable model.

```
Selection Algorithm: exhaustive
         cm10_dif cm30_dif ff_dif
1  ( 1 ) "*"      " "      " "
2  ( 1 ) "*"      "*"      " "
3  ( 1 ) "*"      "*"      "*"
```

□

9.6 Collinearity and Variance Inflation

If two or more predictor variables are highly correlated with one another, then it is difficult to estimate their separate effects on the response. For example, `cm10_dif` and `cm30_dif` have a correlation of 0.96 and the scatterplot in Fig. 9.4 shows that they are highly related to each other. If we regress `aaa_dif` on `cm10_dif`, then the adjusted R^2 is 0.7460, but adjusted R^2 only increases to 0.7556 if we add `cm30_dif` as a second predictor. This suggests that `cm30_dif` might not be related to `aaa_dif`, but this is not the case. In fact, the adjusted R^2 is 0.7376 when `cm30_dif` is the only predictor, which indicates that `cm30_dif` is a good predictor of `aaa_dif`, nearly as good as `cm10_dif`.

Another effect of the high correlation between the predictor variables is that the regression coefficient for each variable is very sensitive to whether the other variable is in the model. For example, the coefficient of `cm10_dif` is 0.616 when `cm10_dif` is the sole predictor variable but only 0.360 if `cm30_dif` is also included.

[1] When comparing models with the same number of parameters, all three criteria are optimized by the same model.

The problem here is that `cm10_dif` and `cm30_dif` provide redundant information because of their high correlation. This problem is called *collinearity* or, in the case of more than two predictors, *multicollinearity*. Collinearity increases standard errors. The standard error of the β of `cm10_dif` is 0.01212 when only `cm10_dif` is in the model, but increases to 0.0451, a 372 % increase, if `cm30_dif` is added to the model.

The *variance inflation factor* (*VIF*) of a variable tells us how much the squared standard error, i.e., the variance of $\widehat{\beta}$, of that variable is increased by having the other predictor variables in the model. For example, if a variable has a VIF of 4, then the variance of its $\widehat{\beta}$ is four times larger than it would be if the other predictors were either deleted or were not correlated with it. The standard error is increased by a factor of 2.

Suppose we have predictor variables $X_1, \ldots, X_p$. Then the VIF of X_j is found by regressing X_j on the $p-1$ other predictors. Let R_j^2 be the R^2-value of this regression, so that R_j^2 measures how well X_j can be predicted from the other Xs. Then the VIF of X_j is

$$\text{VIF}_j = \frac{1}{1 - R_j^2}.$$

A value of R_j^2 close to 1 implies a large VIF. In other words, the more accurately that X_j can be predicted from the other Xs, the more redundant it is and the higher its VIF. The minimum value of VIF_j is 1 and occurs when R_j^2 is 0. There is, unfortunately, no upper bound to VIF_j. Variance inflation becomes infinite as R_j^2 approaches 1.

When interpreting VIFs, it is important to keep in mind that VIF_j tells us nothing about the relationship between the response and jth predictor. Rather, it tells us only how correlated the jth predictor is with the other predictors. In fact, the VIFs can be computed without knowing the values of the response variable.

The usual remedy to collinearity is to reduce the number of predictor variables by using one of the model selection criteria discussed in Sect. 9.5.

Example 9.8. Variance inflation factors for the weekly interest-rate example.

The function `vif()` in `R`'s `faraway` library returned the following VIF values for the changes in weekly interest rates:

```
> library(faraway)
> options(digits = 2)
> vif(lm(aaa_dif ~ cm10_dif + cm30_dif + ff_dif))
cm10_dif cm30_dif   ff_dif
    14.4     14.1      1.1
```

`cm10_dif` and `cm30_dif` have large VIFs due to their high correlation with each other. The predictor `ff_dif` is not highly correlated with `cm10_dif` and `cm30_dif` and has a lower VIF.

VIF values give us information about linear relationships between the predictor variables, but not about their relationships with the response. In this example, `ff_dif` has a small VIF value but is not an important predictor because of its low correlation with the response. Despite their high VIF values, `cm10_dif` and `cm30_dif` are important predictors. The high VIF values tell us only that the regression coefficients for `cm10_dif` and `cm30_dif` are impossible to estimate with high precision.

The question is whether VIF values of 14.4 and 14.1 are so large that the number of predictor variables should be reduced to 1, that is, whether we should use only `cm10_dif`. The answer is "perhaps not" because the model with both `cm10_dif` and `cm30_dif` minimizes BIC. BIC generally selects a parsimonious model because of the high penalty BIC places on the number of predictor variables. Therefore, a model that minimizes BIC is unlikely to need further deletion of predictor variables simply to reduce VIF values. However, we saw earlier that adding `cm30_dif` to the model with `cm10_dif` offers only a minor increase in adjusted R^2, so the issue of whether or not to include `cm30_dif` is not clear. □

Example 9.9. Nelson–Plosser macroeconomic variables

To illustrate model selection, we now turn to an example with more predictors. We will start with six predictors but will find that a model with only two predictors fits rather well.

This example uses a subset of the well-known Nelson–Plosser data set of U.S. yearly macroeconomic time series. These data are available in the file `nelsonplosser.csv`. The variables we will use are:

1. sp-Stock Prices, [Index; 1941-43 = 100], [1871–1970].
2. gnp.r-Real GNP, [Billions of 1958 Dollars], [1909–1970],
3. gnp.pc-Real Per Capita GNP, [1958 Dollars], [1909–1970],
4. ip-Industrial Production Index, [1967 = 100], [1860–1970],
5. cpi-Consumer Price Index, [1967 = 100], [1860–1970],
6. emp-Total Employment, [Thousands], [1890–1970],
7. bnd-Basic Yields 30-year Corporate Bonds, [% pa], [1900–1970].

Since two of the time series start in 1909, we use only the data from 1909 until the end of the series in 1970, a total of 62 years. The response will be the differences of `log(sp)`, the log returns on the stock prices. The regressors will be the differences of variables 2 through 7, with variables 4 and 5 log-transformed before differencing. A differenced log-series contains the approximate relative changes in the original variable, in the same way that a log return approximates a return that is the relative change in price.

How does one decide whether to difference the original series, the log-transformed series, or some other function of the series? Usually the aim is to stabilize the fluctuations in the differenced series. The top row of Fig. 9.6 has time series plots of changes in `gnp.r`, `log(gnp.r)`, and `sqrt(gnp.r)` and the bottom row has similar plots for `ip`. For `ip` the fluctuations in the differenced series increase steadily over time, but this is less true if one uses the square roots or logs of the series. This is the reason why `diff(log(ip))` is used here as a regressor. For `gnp.r`, the fluctuations in changes are more stable and we used `diff(gnp.r)` rather than `diff(log(gnp.r))` as a regressor. In this analysis, we did not consider using square-root transformations, since changes in the square roots are less interpretable than changes in the original variable or its logarithm. However, the changes in the square roots of both series are reasonably stable, so square-root transformations might be considered. Another possibility would be to use the transformation that gives the best-fitting model. One could, for example, put all three variables, `diff(ip)`, `diff(log(ip))`, and `diff(sqrt(ip))`, into the model and use model selection to decide which gives the best fit. The same could be done with `gnp.r` and the other regressors.

Notice that the variables are transformed first and then differenced. Differencing first and then taking logarithms or square roots would result in complex-valued variables, which would be difficult to interpret, to say the least.

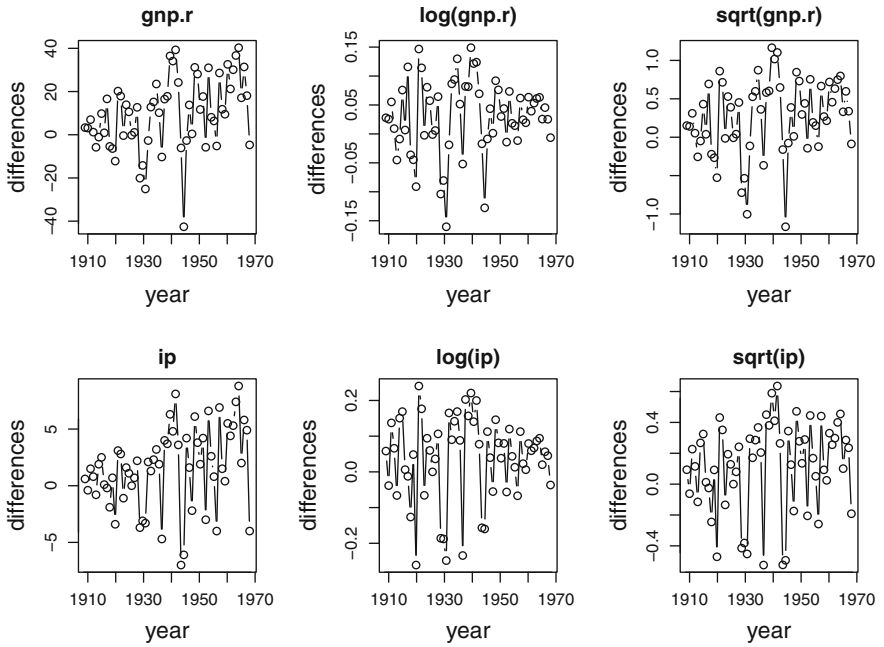

Fig. 9.6. *Differences in* `gnp.r` *and* `ip` *with and without transformations.*

There are additional variables in this data set that could be tried in the model. The analysis presented here is only an illustration and much more exploration is certainly possible with this rich data set.

Time series and normal plots of all eight differenced series did not reveal any outliers. The normal plots were only used to check for outliers, not to check for normal distributions. There is no assumption in a regression analysis that the regressors are normally distributed or that the response has a marginal normal distribution. It is only the conditional distribution of the response given the regressors that is assumed to be normal, and even that assumption can be weakened.

A linear regression with all of the regressors shows that only two, `diff(log(ip))` and `diff(bnd)`, are statistically significant at the 0.05 level and some have very large p-values:

```
Call:
lm(formula = diff(log(sp)) ~ diff(gnp.r) + diff(gnp.pc)
   + diff(log(ip)) + diff(log(cpi))
   + diff(emp) + diff(bnd), data = new_np)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.766e-02  3.135e-02  -0.882   0.3815
diff(gnp.r)     8.384e-03  4.605e-03   1.821   0.0742
diff(gnp.pc)   -9.752e-04  9.490e-04  -1.028   0.3087
diff(log(ip))   6.245e-01  2.996e-01   2.085   0.0418
diff(log(cpi))  4.935e-01  4.017e-01   1.229   0.2246
diff(emp)      -9.591e-06  3.347e-05  -0.287   0.7756
diff(bnd)      -2.030e-01  7.394e-02  -2.745   0.0082
```

A likely problem here is multicollinearity, so variance inflation factors were computed:

```
   diff(gnp.r)   diff(gnp.pc)  diff(log(ip)) diff(log(cpi))
          16.0           31.8            3.3            1.3
     diff(emp)       diff(bnd)
          10.9            1.5
```

We see that `diff(gnp.r)` and `diff(gnp.pc)` have high VIF values, which is not surprising since they are expected to be highly correlated. In fact, their correlation is 0.96.

Next, we search for a more parsimonious model using `stepAIC()`, a variable selection procedure in `R` that starts with a user-specified model and adds or deletes variables sequentially. At each step it either makes the addition or deletion that most improves AIC. It this example, `stepAIC()` will start with all six predictors.

Here is the first step:

```
Start:  AIC=-224.92
diff(log(sp)) ~ diff(gnp.r) + diff(gnp.pc) + diff(log(ip)) +
    diff(log(cpi)) + diff(emp) + diff(bnd)

                 Df Sum of Sq      RSS      AIC
- diff(emp)       1     0.002    1.216 -226.826
- diff(gnp.pc)    1     0.024    1.238 -225.737
- diff(log(cpi))  1     0.034    1.248 -225.237
<none>                         1.214 -224.918
- diff(gnp.r)     1     0.075    1.289 -223.284
- diff(log(ip))   1     0.098    1.312 -222.196
- diff(bnd)       1     0.169    1.384 -218.949
```

The listed models have either zero or one variable removed from the starting model with all regressors. The models are listed in order of their AIC values. The first model, which has `diff(emp)` removed (the minus sign indicates a variable that has been removed), has the best (smallest) AIC. Therefore, in the first step, `diff(emp)` is removed. Notice that the fourth-best model has no variables removed.

The second step starts with the model without `diff(emp)` and examines the effect on AIC of removing additional variables. The removal of `diff(log(cpi))` leads to the largest improvement in AIC, so in the second step this variable is removed:

```
Step:  AIC=-226.83
diff(log(sp)) ~ diff(gnp.r) + diff(gnp.pc) + diff(log(ip)) +
    diff(log(cpi)) + diff(bnd)

                 Df Sum of Sq      RSS      AIC
- diff(log(cpi))  1     0.032    1.248 -227.236
<none>                         1.216 -226.826
- diff(gnp.pc)    1     0.057    1.273 -226.025
- diff(gnp.r)     1     0.084    1.301 -224.730
- diff(log(ip))   1     0.096    1.312 -224.179
- diff(bnd)       1     0.189    1.405 -220.032
```

On the third step no variables are removed and the process stops:

```
Step:  AIC=-227.24
diff(log(sp)) ~ diff(gnp.r) + diff(gnp.pc) + diff(log(ip)) +
    diff(bnd)

                 Df Sum of Sq      RSS      AIC
<none>                         1.248 -227.236
- diff(gnp.pc)   1     0.047    1.295 -227.001
- diff(gnp.r)    1     0.069    1.318 -225.942
- diff(log(ip))  1     0.122    1.371 -223.534
- diff(bnd)      1     0.157    1.405 -222.001
```

Notice that the removal of `diff(gnp.pc)` would cause only a very small increase in AIC. We should investigate whether this variable might be removed. The new model was refit to the data.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.018664   0.028723   -0.65    0.518
diff(gnp.r)    0.007743   0.004393    1.76    0.083
diff(gnp.pc)  -0.001029   0.000712   -1.45    0.154
diff(log(ip))  0.672924   0.287276    2.34    0.023
diff(bnd)     -0.177490   0.066840   -2.66    0.010

Residual standard error: 0.15 on 56 degrees of freedom
Multiple R-squared: 0.347,      Adjusted R-squared:  0.3
F-statistic: 7.44 on 4 and 56 DF,  p-value: 7.06e-05
```

Now three of the four variables are statistically significant at 0.1, though `diff(gnp.pc)` has a rather large p-value, and it seems to be worth exploring other possible models.

The `R` function `leaps()` in the `leaps` package will compute C_p for all possible models. To reduce the amount of output, only the `nbest` models with k regressors [for each $k = 1, \ldots, \dim(\boldsymbol{\beta})$] are printed. The value of `nbest` is selected by the user and in this analysis `nbest` was set at 1, so only the best model is given for each value of k. The following table gives the value of C_p (last column) for the best k-variable models, for $k = 1, \ldots, 6$ (k is in the first column). The remaining columns indicate with a "1" which variables are in the models. All predictors have been differenced, but to save space "`diff`" has been omitted from the variable names heading the columns.

	gnp.r	gnp.pc	log(ip)	log(cpi)	emp	bnd	Cp
1	0	0	1	0	0	0	6.3
2	0	0	1	0	0	1	3.8
3	1	0	1	0	0	1	4.6
4	1	1	1	0	0	1	4.5
5	1	1	1	1	0	1	5.1
6	1	1	1	1	1	1	7.0

We see that `stepAIC` stopping at the four-variable model was perhaps premature. The model selection process was stopped at the four-variable model because the three-variable model had a slightly larger C_p-value. However, if one continues to the best two-variable model, the minimum of C_p is obtained. Here is the fit to the best two-variable model:

```
Call:
lm(formula = diff(log(sp)) ~ +diff(log(ip)) + diff(bnd),
          data = new_np)

Residuals:
```

```
      Min       1Q   Median       3Q      Max
-0.44254 -0.09786  0.00377  0.10525  0.28136

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.0166     0.0210    0.79  0.43332
diff(log(ip))   0.6975     0.1683    4.14  0.00011
diff(bnd)      -0.1322     0.0623   -2.12  0.03792

Residual standard error: 0.15 on 58 degrees of freedom
Multiple R-squared: 0.309,      Adjusted R-squared: 0.285
F-statistic: 12.9 on 2 and 58 DF,  p-value: 2.24e-05
```

Both variables are significant at 0.05. However, it is not crucial that all regressors be significant at 0.05 or at any other predetermined level. Other models could be used, especially if there were good economic reasons for doing so. One cannot say that the two-variable model is best, except in the narrow sense of minimizing C_p, and choosing instead the best three- or four-predictor model would not increase C_p by much. Also, which model is best depends on the criterion used. The best four-predictor model has a better adjusted R^2 than the best two-predictor model. □

9.7 Partial Residual Plots

A partial residual plot is used to visualize the effect of a predictor on the response while removing the effects of the other predictors. The partial residual for the jth predictor variable is

$$Y_i - \left(\widehat{\beta}_0 + \sum_{j' \neq j} X_{i,j'} \widehat{\beta}_{j'}\right) = \widehat{Y}_i + \widehat{\epsilon}_i - \left(\widehat{\beta}_0 + \sum_{j' \neq j} X_{i,j'} \widehat{\beta}_{j'}\right) = X_{i,j} \widehat{\beta}_j + \widehat{\epsilon}_i, \tag{9.19}$$

where the first equality uses (9.12) and the second uses (9.10). Notice that the left-hand side of (9.19) shows that the partial residual is the response with the effects of all predictors but the jth subtracted off. The right-hand side of (9.19) shows that the partial residual is also equal to the residual with the effect of the jth variable added back. The partial residual plot is simply the plot of the responses against these partial residuals.

Example 9.10. Partial residual plots for the weekly interest-rate example

Partial residual plots for the weekly interest-rate example are shown in Fig. 9.7a, b. For comparison, scatterplots of `cm10_dif` and `cm30_dif` versus `aaa_dif` with the corresponding one-variable fitted lines are shown in panels (c) and (d). The main conclusion from examining the plots is that the slopes in (a) and (b) are shallower than the slopes in (c) and (d). What does this tell

us? It says that, due to collinearity, the effect of `cm10_dif` on `aaa_dif` when `cm30_dif` is in the model [panel (a)] is less than when `cm30_dif` is not in the model [panel (c)], and similarly when the roles of `cm10_dif` and `cm30_dif` are reversed.

The same conclusion can be reached by looking at the estimated regression coefficients. From Examples 9.1 and 9.4, we can see that the coefficient of `cm10_dif` is 0.615 when `cm10_dif` is the only variable in the model, but the coefficient drops to 0.355 when `cm30_dif` is also in the model. There is a similar decrease in the coefficient for `cm30_dif` when `cm10_dif` is added to the model. □

Example 9.11. Nelson–Plosser macroeconomic variables—Partial residual Plots

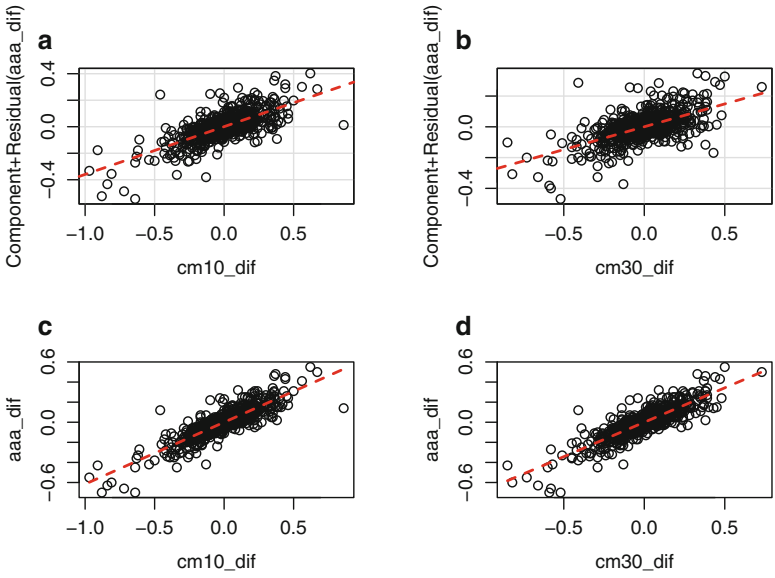

Fig. 9.7. *Partial residual plots for the weekly interest rates [panels (**a**) and (**b**)] and scatterplots of the predictors and the response [panels (**c**) and (**d**)].*

This example continues the analysis of the Nelson–Plosser macroeconomic variables. Partial residual plots for the four-variable model selected by `stepAIC` in Example 9.9 are shown in Fig. 9.8. One can see that all four variables have explanatory power, since the partial residuals have linear trends in the variables.

One puzzling aspect of this model is that the slope for `gnp.pc` is negative. However, the p-value for this regressor is large and the minimum C_p model

does not contain either `gnp.r` or `gnp.pc`. Often, a regressor that is highly correlated with other regressors has an estimated slope that is counterintuitive. If used alone, both `gnp.r` and `gnp.pc` have positive slopes. The slope of `gnp.pc` is negative only when `gnp.r` is in the model. □

9.8 Centering the Predictors

Centering or, more precisely, *mean-centering* a variable means expressing it as a deviation from its mean. Thus, if $X_{1,k}, \ldots, X_{n,k}$ are the values of the kth predictor and $\overline{X}_k$ is their mean, then $(X_{1,k} - \overline{X}_k), \ldots, (X_{n,k} - \overline{X}_k)$ are values of the centered predictor.

Centering is useful for two reasons:

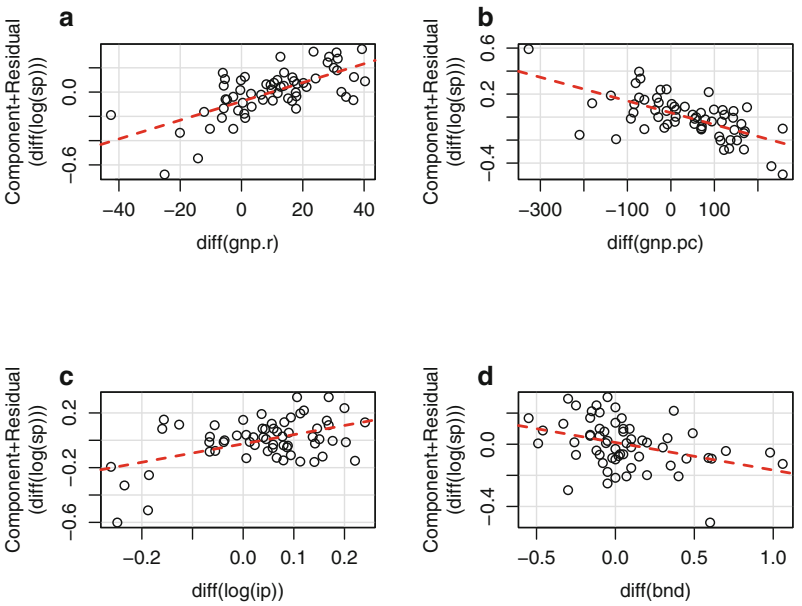

Fig. 9.8. *Partial residual plots for the Nelson–Plosser U.S. economic time series. (**a**) Change in* `gnp.r`. *(**b**) Change in* `gnp.pc`. *(**c**) Change in* `log(ip)`. *(**d**) Change in* `bnd`.

- centering can reduce collinearity in polynomial regression;
- if all predictors are centered, then β_0 is the expected value of Y when each of the predictors is equal to its mean. This gives β_0 an interpretable meaning. In contrast, if the variables are not centered, then β_0 is the expected value of Y when all of the predictors are equal to 0. Frequently, 0 is outside the range of some predictors, making the interpretation of β_0 of little real interest unless the variables are centered.

9.9 Orthogonal Polynomials

As just mentioned, centering can reduce collinearity in polynomial regression because, for example, if X is positive, then X and X^2 will be highly correlated but $X - \overline{X}$ and $(X - \overline{X})^2$ will be less correlated.

Orthogonal polynomials can eliminate correlation entirely, since they are defined in a way so that they are uncorrelated. This is done using the Gram–Schmidt orthogonalization procedure discussed in textbooks on linear algebra. Orthogonal polynomials can be created easily in most software packages, for instance, by using the `poly()` function in `R`. Orthogonal polynomials are particularly useful for polynomial regression of degree higher than 2 where centering is less successful at reducing collinearity. However, the use of polynomial models of degree 4 and higher is discouraged and nonparametric regression (see Chap. 21) is recommended instead. Even cubic regression can be problematic because cubic polynomials have only a limited range of shapes.

9.10 Bibliographic Notes

Harrell (2001) , Ryan (1997), Neter et al. (1996) and Draper and Smith (1998) are four of the many good introductions to regression. Faraway (2005) is an excellent modern treatment of linear regression with `R`. See Nelson and Plosser (1982) for information about their data set.

9.11 R Lab

9.11.1 U.S. Macroeconomic Variables

This section uses the data set `USMacroG` in `R`'s `AER` package. This data set contains quarterly times series on 12 U.S. macroeconomic variables for the period 1950–2000. We will use the variables `consumption` = real consumption expenditures, `dpi` = real disposable personal income, `government` = real government expenditures, and `unemp` = unemployment rate. Our goal is to predict changes in `consumption` from changes in the other variables.

Run the following `R` code to load the data, difference the data (since we wish to work with changes in these variables), and create a scatterplot matrix.

```
library(AER)
data("USMacroG")
MacroDiff = as.data.frame(apply(USMacroG, 2, diff))
attach(MacroDiff)
pairs(cbind(consumption, dpi, cpi, government, unemp))
```

Problem 1 *Describe any interesting features, such as outliers, seen in the scatterplot matrix. Keep in mind that the goal is to predict changes in* `consumption`*. Which variables seem best suited for that purpose? Do you think there will be collinearity problems?*

Next, run the code below to fit a multiple linear regression model to `consumption` using the other four variables as predictors.

```
fitLm1 = lm(consumption ~ dpi + cpi + government + unemp)
summary(fitLm1)
confint(fitLm1)
```

Problem 2 *From the summary, which variables seem useful for predicting changes in* `consumption`*?*

Next, print an ANOVA table.

```
anova(fitLm1)
```

Problem 3 *For the purpose of variable selection, does the ANOVA table provide any useful information not already in the summary?*

Upon examination of the p-values, we might be tempted to drop several variables from the regression model, but we will not do that since variables should be removed from a model one at a time. The reason is that, due to correlation between the predictors, when one is removed the significance of the others changes. To remove variables sequentially, we will use the function `stepAIC()` in the `MASS` package.

```
library(MASS)
fitLm2 = stepAIC(fitLm1)
summary(fitLm2)
```

Problem 4 *Which variables are removed from the model, and in what order?*

Now compare the initial and final models by AIC.

```
AIC(fitLm1)
AIC(fitLm2)
AIC(fitLm1) - AIC(fitLm2)
```

Problem 5 *How much of an improvement in AIC was achieved by removing variables? Was the improvement large? Is so, can you suggest why? If not, why not?*

The function `vif()` in the `car` package will compute variance inflation factors. A similar function with the same name is in the `faraway` package. Run

```
library(car)
vif(fitLm1)
vif(fitLm2)
```

Problem 6 *Was there much collinearity in the original four-variable model? Was the collinearity reduced much by dropping two variables?*

Partial residual plots, which are also called *component plus residual* or *cr* plots, can be constructed using the function `crPlot()` in the `car` package. Run

```
par(mfrow = c(2, 2))
sp = 0.8
crPlot(fitLm1, dpi, span = sp, col = "black")
crPlot(fitLm1, cpi, span = sp, col = "black")
crPlot(fitLm1, government, span = sp, col = "black")
crPlot(fitLm1, unemp, span = sp, col = "black")
```

Besides dashed least-squares lines, the partial residual plots have solid lowess smooths through them unless this feature is turned off by specifying `smooth=F`, as was done in Fig. 9.8. Lowess is an earlier version of loess. The smoothness of the lowess curves is determined by the parameter span, with larger values of span giving smoother plots. The default is span = 0.5. In the code above, span is 0.8 but can be changed for all four plots by changing the variable `sp`. Lowess, loess, and span are described in Sect. 21.2.1. A substantial deviation of the lowess curve from the least-squares line is an indication that the effect of the predictor is nonlinear. The default color of the `crPlot` figure is red, but this can be changed as in the code above.

Problem 7 *What conclusions can you draw from the partial residual plots?*

9.12 Exercises

1. Suppose that $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where ϵ_i is $N(0, 0.3)$, $\beta_0 = 1.4$, and $\beta_1 = 1.7$.
   (a) What are the conditional mean and standard deviation of Y_i given that $X_i = 1$? What is $P(Y_i \leq 3 | X_i = 1)$?
   (b) A regression model is a model for the conditional distribution of Y_i given X_i. However, if we also have a model for the marginal distribution of X_i, then we can find the marginal distribution of Y_i. Assume that X_i is $N(1, 0.7)$. What is the marginal distribution of Y_i? What is $P(Y_i \leq 3)$?

2. Show that if $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $N(0, \sigma_\epsilon^2)$, then in straight-line regression the least-squares estimates of β_0 and β_1 are also the maximum likelihood estimates.
   Hint: This problem is similar to the example in Sect. 5.9. The only difference is that in that section, $Y_1, \ldots, Y_n$ are independent $N(\mu, \sigma^2)$, while in this exercise $Y_1, \ldots, Y_n$ are independent $N(\beta_0 + \beta_1 X_i, \sigma_\epsilon^2)$.
3. Use (7.11), (9.3), and (9.2) to show that (9.8) holds.
4. It was stated in Sect. 9.8 that centering reduces collinearity. As an illustration, consider the example of quadratic polynomial regression where X takes 30 equally spaced values between 1 and 15.
   (a) What is the correlation between X and X^2? What are the VIFs of X and X^2?
   (b) Now suppose that we center X before squaring. What is the correlation between $(X - \overline{X})$ and $(X - \overline{X})^2$? What are the VIFs of $(X - \overline{X})$ and $(X - \overline{X})^2$?
5. A linear regression model with three predictor variables was fit to a data set with 40 observations. The correlation between Y and $\widehat{Y}$ was 0.65. The total sum of squares was 100.
   (a) What is the value of R^2?
   (b) What is the value of the residual error SS?
   (c) What is the value of the regression SS?
   (d) What is the value of s^2?
6. A data set has 66 observations and five predictor variables. Three models are being considered. One has all five predictors and the others are smaller. Below is residual error SS for all three models. The total SS was 48. Compute C_p and R^2 for all three models. Which model should be used based on this information?

Number of predictors	Residual error SS
3	12.2
4	10.1
5	10.0

7. The quadratic polynomial regression model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$

   was fit to data. The p-value for β_1 was 0.67 and for β_2 was 0.84. Can we accept the hypothesis that β_1 and β_2 are both 0? Discuss.
8. Sometimes it is believed that β_0 is 0 because we think that $E(Y|X = 0) = 0$. Then the appropriate model is

$$y_i = \beta_1 X_i + \epsilon_i.$$

   This model is usually called "regression through the origin" since the regression line is forced through the origin. The least-squares estimator of β_1 minimizes

$$\sum_{i=1}^{n}\{Y_i - \beta_1 X_i\}^2.$$

Find a formula that gives $\hat{\beta}_1$ as a function of the Y_is and the X_is.

9. Complete the following ANOVA table for the model $Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i$:

Source	df	SS	MS	F	P
Regression	?	?	?	?	0.04
Error	?	5.66	?		
Total	15	?			

R-sq = ?

10. Pairs of random variables (X_i, Y_i) were observed. They were assumed to follow a linear regression with $E(Y_i|X_i) = \theta_1 + \theta_2 X_i$ but with t-distributed noise, rather than the usual normally distributed noise. More specifically, the assumed model was that conditionally, given X_i, Y_i is t-distributed with mean $\theta_1 + \theta_2 X_i$, standard deviation θ_3, and degrees of freedom θ_4. Also, the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are mutually independent. The model could also be expressed as

$$Y_i = \theta_1 + \theta_2 X_i + \epsilon_i$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. t with mean 0 and standard deviation θ_3 and degrees of freedom θ_4. The model was fit by maximum likelihood. The R code and output are

```
#(Code to input x and y not shown)
library(fGarch)
start = c(lmfit$coef, sd(lmfit$resid), 4)
loglik = function(theta)
{
-sum(log(dstd(y, mean = theta[1] + theta[2] * x, sd = theta[3],
   nu = theta[4])))
}
mle = optim(start, loglik, hessian = TRUE)
InvFishInfo = solve(mle$hessian)
mle$par
mle$value
mle$convergence
sqrt(diag(InvFishInfo))
qnorm(0.975)

> mle$par
[1] 0.511 1.042 0.152 4.133
> mle$value
[1] -188
```

```
> mle$convergence
[1] 0
> sqrt(diag(InvFishInfo))
[1] 0.00697 0.11522 0.01209 0.93492
>
> qnorm(.975)
[1] 1.96
>
```

(a) What is the MLE of the slope of Y_i on X_i?
(b) What is the standard error of the MLE of the degrees-of-freedom parameter?
(c) Find a 95 % confidence interval for the standard deviation of the noise.
(d) Did `optim` converge? Why or why not?

References

Draper, N. R. and Smith, H. (1998) *Applied Regression Analysis,* 3rd ed., Wiley, New York.

Faraway, J. J. (2005) *Linear Models with R*, Chapman & Hall, Boca Raton, FL.

Harrell, F. E., Jr. (2001) *Regression Modeling Strategies*, Springer-Verlag, New York.

Nelson C.R., and Plosser C.I. (1982) Trends and random walks in macroeconomic time series. *Journal of Monetary Economics*, **10**, 139–162.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996) *Applied Linear Statistical Models*, 4th ed., Irwin, Chicago.

Ryan, T. P. (1997) *Modern Regression Methods*, Wiley, New York.