7

Applications of Regression II

7.1 Overview

In Chapter 6 we discussed several practical applications of inference procedures, developed in Chapters 3 and 4, for straight line regression and multiple linear regression. In this chapter we consider two further important applications of regression: *subset analysis* and *growth curves*. Section 7.2 gives a brief introduction to subset analysis, which is often called *selection of variables*. In Sections 7.3 and 7.4 we discuss various methods of performing a subset analysis so that you can become acquainted with the terminology and some of the commonly used procedures. We do not discuss some of the complex theoretical and conceptual issues that underly model building. Section 7.5 introduces growth curve models and presents appropriate inference procedures associated with them.

7.2 Subset Analysis and Variable Selection

To motivate the material in this section, we consider Example 4.4.2. In that example we were interested in predicting the GPA (Y) of a student at the end of the freshman year based on the values of SATmath (X_1), SATverbal (X_2), HSmath(X_3), and HSenglish (X_4). Suppose that the regression function of Y on X_1, X_2, X_3, and X_4 is given by

$$\mu_Y(x_1, x_2, x_3, x_4) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \qquad (7.2.1)$$

and that the predictions based on (7.2.1) are sufficiently accurate for the needs of the admissions director. It may happen that other prediction functions, based on all or only some of the variables X_1, X_2, X_3, and X_4, may also be sufficiently accurate for the problem at hand. If a prediction function such as

$$P_Y(x_3, x_4) = \beta_0 + \beta_3 x_3 + \beta_4 x_4 \qquad (7.2.2)$$

based only on HSmath and HSenglish, is as good (or nearly as good) as the regression function in (7.2.1), which is based on all four predictors, then the director of admissions can base decisions solely on the high school grades of the applicants, who can then be exempted from having to take the SAT. Keep in mind the possibility that there may be other predictors, such as the ACT scores, which, together with the predictors in (7.2.1), might improve the predictions substantially. Here we are concerned only with the four predictors SATmath, SATverbal, HSmath, and HSenglish. The prediction function given in (7.2.2), which may not be the *best* function for predicting Y using X_3 and X_4, (i.e., it may not be the *regression* function of Y on X_3 and X_4), is called a **subset model** because it is obtained by using a *subset* of the *predictors in (7.2.1). Thus the director of admissions may be interested in examining* the performance of various prediction functions based on subsets of $\{X_1, X_2, X_3, X_4\}$ and eventually selecting one or more prediction functions that may be satisfactory for the problem. This process of examining subset models and selecting one or more suitable prediction functions is often called **subset selection, subset analysis**, or **selection of variables.**

The general problem of subset selection can be described as follows. We know from Chapter 4 that when the $(k+1)$-variable population $\{(Y, X_1, \ldots, X_k)\}$ satisfies population assumptions (A) or (B), the best function for predicting Y using $X_1, \ldots, X_k$ as predictors, is the regression function of Y on $X_1, \ldots, X_k$; i.e.,

$$\mu_Y(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{7.2.3}$$

If the investigator decides that σ is sufficiently small for the problem under study (where σ denotes $\sigma_{Y|X_1,\ldots,X_k}$), then the regression function in (7.2.3) will in fact be an adequate prediction function for this problem. However, investigators are often interested in examining prediction functions based on various *subsets* of the full set $\{X_1, \ldots, X_k\}$ for a variety of reasons. Two of the reasons are:

- To get some insight into how various subsets of predictors contribute to the prediction of Y.
- To find adequate prediction functions based on subsets of the predictor variables that are easier or less expensive to measure than the full set $\{X_1, \ldots, X_k\}$.

Suppose $\{(Y, X_1, \ldots, X_k)\}$ satisfies population assumptions (B). Then the regression function of Y on $X_1, \ldots, X_k$ is the one given in (7.2.3); i.e., it is a multiple *linear* regression function. In this case the regression function of Y on any subset $\{X_1, \ldots, X_m\}$ of m predictors $(0 < m < k)$ is also a multiple *linear* regression function of the form

$$\mu_Y^{(A)}(x_1, \ldots, x_m) = \beta_0^A + \beta_1^A x_1 + \cdots + \beta_m^A x_m \tag{7.2.4}$$

However, if $\{(Y, X_1, \ldots, X_k)\}$ satisfies population assumptions (A) but fails to satisfy population assumptions (B), then although the regression function of Y on the k predictors $X_1, \ldots, X_k$ is still of the form given in (7.2.3), the regression function of Y on the m predictors $X_1, \ldots, X_m$ $(m < k)$ is *not* necessarily of the form given in (7.2.4). Nevertheless, it is useful to consider **linear prediction functions** of the form

$$\beta_0^A + \beta_1^A x_1 + \cdots + \beta_m^A x_m \tag{7.2.5}$$

based on various subsets of predictors. We can relabel the predictor variables suitably so that the subset under consideration can always be written as the first m predictor variables $\{X_1, \ldots, X_m\}$.

In Sections 7.3 and 7.4 we discuss exploratory statistical procedures for finding linear prediction functions based on subsets of $\{X_1, \ldots, X_k\}$ that compare favorably with the full model regression function given in (7.2.3). The investigator may eventually choose a subset model (which may be the full model), or the investigator may choose several subset models, based on considerations such as cost of obtaining the values of the predictors, simplicity and interpretability of the prediction function, adequacy of the predictions, etc.

Commonly Used Procedures for Subset Analysis

Subset analysis generally requires a considerable amount of computing, and even with the availability of high-speed computers, the amount of computing must sometimes be taken into consideration in deciding which of the several available methods to use. A large amount of material has been written on this subject during the past three decades, but there are very few rigorous results of an inferential nature; most of the procedures are merely descriptive or exploratory. Four general procedures that are commonly used in subset analysis are:

1 All-subsets regression
2 Forward selection procedure
3 Backward elimination procedure
4 Stepwise regression procedure

The *all-subsets regression* procedure is discussed in Section 7.3, and the remaining three procedures are discussed in Section 7.4.

Problems 7.2

7.2.1 Give two examples from your field of specialization where the investigator might be interested in a subset analysis. In each case, explain clearly the purpose that would lead to the consideration of the subset selection problem.

7.2.2 Suppose, in a research problem, the form of the regression function $\mu_Y(x_1, x_2)$ of Y on X_1, X_2 is not known. However, the investigator thinks that the linear prediction function $P_Y(x_1, x_2) = \beta_0^A + \beta_1^A x_1 + \beta_2^A x_2$ is an adequate approximation to the regression function $\mu_Y(x_1, x_2)$, so $P_Y(x_1, x_2)$ is used in place of $\mu_Y(x_1, x_2)$. Explain in detail what this means.

7.3
All-Subsets Regression

The notation for subset analysis can get extremely cumbersome since there are many models to examine and each model may contain several β coefficients. For example, consider the two models

$$(1) \quad \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

and

$$(2) \quad \beta_0 + \beta_1 x_1 + \beta_3 x_3 \tag{7.3.1}$$

When studied together these should be written as

$$(3) \quad \beta_0^{(1)} + \beta_1^{(1)} x_1 + \beta_2^{(1)} x_2$$

and

$$(4) \quad \beta_0^{(2)} + \beta_1^{(2)} x_1 + \beta_3^{(2)} x_3 \tag{7.3.2}$$

to indicate that, in general, the β_i in (1) are not the same as the β_i in (2), etc. However, because there are many models to consider in subset analysis, and to avoid complicated notation, we simply write models (3) and (4) as in (7.3.1). *You should remember that the β_i in different models such as (1) and (2) are generally different.*

We begin with the premise that the investigator is interested in examining subsets of the predictors $X_1, \ldots, X_k$, to determine how good each subset is for predicting the response variable Y. Some of the predictor variables may be *derived quantities* based on other predictor variables. For example, X_3 may stand for X_1^2 and X_4 may stand for $X_1 X_2$, etc. We assume that the best prediction function (i.e., the regression function) using all of $X_1, \ldots, X_k$ is of the form

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{7.3.3}$$

Our objective is to determine how well Y may be predicted by each of the subset prediction functions of the form

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m \tag{7.3.4}$$

where $\{X_1, \ldots, X_m\}$ is a subset of $\{X_1, \ldots, X_k\}$. Although we have used the first m predictors $X_1, X_2, \ldots, X_m$ for notational simplicity, we could use any m predictors, not just the first m. Based on suitable criteria, the investigator will select one or more of these subset prediction functions for further consideration. The subset prediction functions considered are all linear in the predictor variables X_i as well as the parameters. However, since a predictor variable X_i may itself be a quantity computed from other predictor variables—for instance, X_3 might be X_1^2, etc.—this is not a serious limitation. As we pointed out earlier, you should bear in mind that the prediction function in (7.3.4) may not be the regression function of Y on $X_1, \ldots, X_m$; i.e., it may not be the *best* function for predicting Y using $X_1, \ldots, X_m$, but it may be an *adequate* prediction function for the problem.

When there are k predictor variables in all, the number of possible subset prediction functions of the form (7.3.4) is 2^k, which can be quite large. For instance, if $k = 10$, then the number of possible subset models is $2^{10} = 1{,}024$ (including the constant model that uses none of the k predictors). When $k = 20$, the number of subset models to be considered is $2^{20} = 1{,}048{,}576$. Clearly this presents a nontrivial computational problem. Fortunately, efficient computing methods exist, and many statistical software packages have the capability of performing the *all-subsets* regression analysis if k is not too large. The particular computer being used to perform the calculations will determine the feasibility of the computations for different values of k.

In principle the all-subsets regression analysis proceeds as follows:

1 Each of the 2^k subset models is fitted to the sample data by the method of least squares using the formulas in Chapter 4.

2 For each subset model, a *criterion measure* is calculated which summarizes how good that model is for predicting Y.

3 A table or a graph that exhibits the predictors in each model, along with the *criterion measure* calculated for that model, is constructed.

4 This table or graph is examined by the investigator, and one or more models are selected based on many considerations such as the adequacy of the predictions based on the model, costs of collecting data, subject matter knowledge regarding the reasonableness of the various models, etc. These models comprise the *short list*.

5 The models on this short list may be subjected to further scrutiny, which may include various residual analyses, the diagnostic procedures of Chapter 5, or validation based on additional sample data, etc. The investigator may finally select one or more of these models as a *tentative* model that is satisfactory for the problem under study. In some cases the investigator may not be interested in actually using any of these models for the purposes of prediction. Instead he or she may formulate some hypothesis or theory, or modify an existing hypothesis or theory, as a result of the examination of the subset models.

Steps (4) and (5) typically require the participation of both the investigator and the statistician (in some cases the same individual will play both roles) and involve subjective judgments. *No rigorous measure of confidence can be attached to the final conclusions at this stage, but a new study can be designed to validate or invalidate the conclusions arrived at as a result of the subset analysis.* Subset analysis is necessarily an iterative procedure, and only after considerable effort has been expended can an investigator reach the stage of being able to make valid probability or confidence statements regarding the conclusions. Thus, in many applications, subset analysis may be viewed as "data in search of a model."

Many different candidates have been proposed for the *criterion measure* mentioned in step (2) above. We discuss the most popular among these, which are: Root mean square of residuals, R-square, adjusted R-square, and Mallow's C_p criterion. Each one of these criteria may be given theoretical justification depending on the objective of the subset analysis and population and sample assumptions, but we do

not discuss the justifications here. If you are interested you should consult [6], and [13]. Because subset analysis is merely an empirical or descriptive analysis of all the subsets and involves no strict statistical inferences, any one of these criteria generally leads to essentially the same collection of subset models for further scrutiny.

Root Mean Squared Error (s)

To use this criterion we must calculate the square root of the mean squared error for each subset model. As before, consider a subset of *any* m predictors chosen from $X_1, \ldots, X_k$. For concreteness these m predictors will be denoted by $X_1, \ldots, X_m$ (after relabeling the predictor factors if necessary). For the model based on the subset of m predictors $X_1, \ldots, X_m$, the linear prediction function

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m \tag{7.3.5}$$

is fitted to the sample data using the least squares method of Chapter 4. The fitted model may be written as

$$\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m \tag{7.3.6}$$

The mean squared error $MSE(X_1, \ldots, X_m)$ corresponding to this model, which we simply write as MSE when there is no possibility of confusion, can then be calculated using the formula in (4.4.15), which gives

$$MSE(X_1, \ldots, X_m) = MSE = \frac{\sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_m x_{i,m})]^2}{(n-p)} \tag{7.3.7}$$

where p is the number of β parameters in the subset model (which is $m+1$ for the model in (7.3.5)) and n is the number of observations in the sample. The root mean square of the residuals for the subset model in (7.3.5), denoted by $s(X_1, \ldots, X_m)$ or simply as s when the subset model under consideration is unambiguous, is then

$$s(X_1, \ldots, X_m) = s = \sqrt{MSE(X_1, \ldots, X_m)} \tag{7.3.8}$$

Models that have small values of s are selected for inclusion in the *short list*.

In the special case when the $(m+1)$-variable population $\{(Y, X_1, \ldots, X_m)\}$ satisfies assumptions (A) or assumptions (B), the quantity s is an estimate of the subpopulation standard deviation $\sigma_{Y|X_1,\ldots,X_m}$; otherwise this may not be true. Throughout this book we have used subpopulation standard deviations as a measure of how good a regression function is for predicting Y, and in general, regression functions with smaller standard deviations are better than those with larger standard deviations. This is why s (which can be viewed as an estimate of σ when assumptions (A) or (B) are satisfied) is sometimes used to distinguish between prediction functions.

R-Square (R^2)

To use this criterion, we calculate the quantity called R-square (written R^2) for each subset model. For the model based on the subset of m predictors $X_1, \ldots, X_m$, the

linear prediction function in (7.3.5) is fitted to the sample data using the method of least squares. The fitted model may be written as in (7.3.6). The sum of squared errors $SSE(X_1, \ldots, X_m)$ for this subset model and the (corrected) total sum of squares of Y, denoted as usual by SSY, are calculated using the formulas

$$SSE(X_1, \ldots, X_m) = SSE = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_m x_{i,m})]^2 \qquad (7.3.9)$$

and

$$SSY = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad (7.3.10)$$

respectively. Then the value of R^2 corresponding to the subset of predictors $X_1, \ldots, X_m$ is calculated using

$$R^2(X_1, \ldots, X_m) = R^2 = \frac{SSY - SSE(X_1, \ldots, X_m)}{SSY} \qquad (7.3.11)$$

Models with large values of R^2 (equivalently, models with small values of SSE) are included in the short list. No subset model will have an R^2 larger than that of the full model, which includes all k predictor variables, but there may exist subset models with R^2 values that are nearly equal to that of the full model.

Note that when assumptions (B) hold (data must be obtained by simple random sampling), R^2 is in fact used to estimate the multiple coefficient of determination $\rho^2_{Y(X_1,\ldots,X_m)}$ of Y with the predictors $X_1, \ldots, X_m$. Otherwise R^2 may not be a valid estimate of $\rho^2_{Y(X_1,\ldots,X_m)}$. Throughout this book we have stated that $\rho^2_{Y(X_1,\ldots,X_m)}$ should not be used as a measure of how good $X_1, \ldots, X_m$ are for predicting Y. However, to compare one model (say, model A) with another (say, model B), we can compare ρ^2_A with ρ^2_B or we can compare σ_A with σ_B since ρ^2_A is less than ρ^2_B if and only if σ_A is greater than σ_B. This fact follows from the definition of $\rho^2_{Y(X_1,\ldots,X_k)}$ given in (4.9.8).

Adjusted R-Square ($Adj\text{-}R^2$)

To use this criterion we must calculate a quantity called adjusted R-square (written as $adj\text{-}R^2$) for each subset model. For the model based on the subset of predictors $X_1, \ldots, X_m$, the linear model in (7.3.5) is fitted to the sample data using least squares. The fitted model is in (7.3.6). The mean squared error MSE for this subset model is computed as in (7.3.7). Then the adjusted R-square for this model, denoted by $adj\text{-}R^2(X_1, \ldots, X_m)$, or simply by $adj\text{-}R^2$ when the subset model under consideration is unambiguous, is calculated using

$$adj\text{-}R^2(X_1, \ldots, X_m) = adj\text{-}R^2 = \frac{MSY - MSE(X_1, \ldots, X_m)}{MSY} \qquad (7.3.12)$$

where $MSY = SSY/(n-1)$. Subset models with large values of $adj\text{-}R^2$ (equivalently, models with small values of MSE) are included in the short list.

Note that when assumptions (B) are satisfied, *adj-*R^2 is in fact used as an alternative to R^2 to estimate the multiple coefficient of determination $\rho^2_{Y(X_1,\ldots,X_m)}$ of Y with the predictors $X_1, \ldots, X_m$; otherwise this may not be a valid estimate of $\rho^2_{Y(X_1,\ldots,X_m)}$.

Note The quantities *SSE*, *MSE*, and *SSY* can be obtained from an ANOVA table by regressing Y on $X_1, \ldots, X_m$. Also, you should observe that whenever one model is better than a second model according to the adjusted R-square criterion, then it is also better than the second model according to the root mean squared error (s) criterion, and vice versa. Thus the adjusted R-square criterion is equivalent to the root mean squared error (s) criterion in selecting models, but the R-square criterion is not equivalent to the s criterion or the adjusted R-square criterion. If n is very large relative to m, then all three are essentially equivalent.

Mallow's C_p Criterion

A measure that is quite widely used in subset selection analysis is the C_p criterion measure, originally proposed by C. Mallows [21]. Loosely speaking, this quantity is an estimate of the average squared prediction error relative to the estimate of $\sigma_{Y|X_1,\ldots,X_k}$ if the fitted model

$$\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m \tag{7.3.13}$$

is used to estimate $\mu_Y(X_1, \ldots, X_k)$ for values of $X_1, \ldots, X_k$ in the sample. The subscript p in C_p refers to the number of β parameters in the subset model, which is $m + 1$ in (7.3.13). The measure C_p is calculated for each of the 2^k possible subset models. For the model based on the predictors $X_1, \ldots, X_m$, the linear model in (7.3.5) is fitted to the sample data using least squares. The fitted model is given in (7.3.6). The sum of squared errors $SSE(X_1, \ldots, X_m)$ for this subset model, denoted simply by *SSE*, is calculated as in (7.3.9). Then the measure C_p corresponding to this subset is denoted by $C_p(X_1, \ldots, X_m)$, or simply by C_p when the set of predictors $X_1, \ldots, X_m$ does not need to be identified explicitly, and is calculated using the formula

$$C_p = C_p(X_1, \ldots, X_m) = \frac{SSE(X_1, \ldots, X_m)}{\hat{\sigma}^2_{Y|X_1,\ldots,X_k}} + 2p - n \tag{7.3.14}$$

where $\hat{\sigma}^2_{Y|X_1,\ldots,X_k}$ is the estimated variance of the subpopulation of Y values determined by using all of the predictors $X_1, \ldots, X_k$ and, as before, p is the number of β parameters in the subset model under consideration, and n is the sample size. For the model in (7.3.5) the value of p is $m + 1$. Models with small values of C_p are included in the short list.

When the population $\{(Y, X_1, \ldots, X_k)\}$ satisfies assumptions (A), it can be shown that the quantity $C_p - p$ is a measure of the tendency of the fitted model

$$\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m \tag{7.3.15}$$

corresponding to the subset of predictors $X_1, \ldots, X_m$, to underestimate or overestimate the subpopulation mean $\mu_Y(x_1, \ldots, x_k)$, which is used to predict Y when all k predictor factors are used. Negative values of $C_p - p$ should be regarded as zero because it can be shown that $C_p - p$ is an estimate of a population quantity that can never be negative. This tendency for underestimation or overestimation is referred to as bias. Thus $C_p - p$ is a measure of bias. Some authors recommend selection of models for which both C_p and $C_p - p$ are small.

C_p-Plot

As discussed above, in selecting subset models for further consideration we may want to examine the values of C_p as well as $C_p - p$. This is conveniently carried out by plotting the values of C_p for the different models and superimposing the straight line $C_p = p$ on this plot. The values of $C_p - p$ can be judged visually by examining how close the various points are to the line $C_p = p$. Good subset models should have small values for C_p and also should have $C_p - p$ close to zero (as stated earlier, negative values of $C_p - p$ are interpreted as zero).

Formulas for s, $adj\text{-}R^2$, and C_p in Terms of R^2

The quantities s, R^2, $adj\text{-}R^2$, and C_p for a subset model with p parameters β are interrelated, and the relationship of R^2 to the various measures is

$$adj\text{-}R^2(X_1, \ldots, X_m) = 1 - \frac{n-1}{n-p}[1 - R^2(X_1, \ldots, X_m)] \tag{7.3.16}$$

$$s(X_1, \ldots, X_m) = \sqrt{\frac{SSY[1 - R^2(X_1, \ldots, X_m)]}{n-p}} \tag{7.3.17}$$

$$C_p(X_1, \ldots, X_m) = \frac{(n-k-1)[1 - R^2(X_1, \ldots, X_m)]}{[1 - R^2(X_1, \ldots, X_k)]} + 2p - n \tag{7.3.18}$$

Recall that p is the number of β parameters in the subset model under consideration, and it equals $m + 1$ for the model with m predictor variables and an intercept term (see (7.3.5)). Also, in (7.3.18), $R^2(X_1, \ldots, X_k)$ is the value of R^2 corresponding to the model consisting of all k predictors $X_1, \ldots, X_k$; i.e.,

$$R^2(X_1, \ldots, X_k) = \frac{SSY - SSE(X_1, \ldots, X_k)}{SSY}$$

When comparing two or more subset models with the same number of predictor variables (i.e., the same value for p, the number of beta coefficients in the model), the four criteria—s, R^2, $adj\text{-}R^2$, C_p—will lead to the same ordering of the subset models. However, when comparing two or more subset models with *different* numbers of predictor variables (different values of p), these four criteria, in general, will not lead to the same ordering of the subset models (except, of

course, for the s and the *adj-*R^2 criteria, which will always lead to the same ordering of all subset models).

We illustrate the use of each of the preceding measures using the GPA data of Example 4.4.2.

EXAMPLE 7.3.1

To illustrate the results of this section, we carry out an all-subsets regression analysis for the GPA data of Example 4.4.2. Recall that Y = GPA, X_1 = SATmath, X_2 = SATverbal, X_3 = HSmath, and X_4 = HSenglish.

Table 7.3.1 displays the values of s, R^2, *adj-*R^2, C_p, and $C_p - p$ for each subset of the predictors $\{X_1, X_2, X_3, X_4\}$, and Figure 7.3.1 shows the C_p plot.

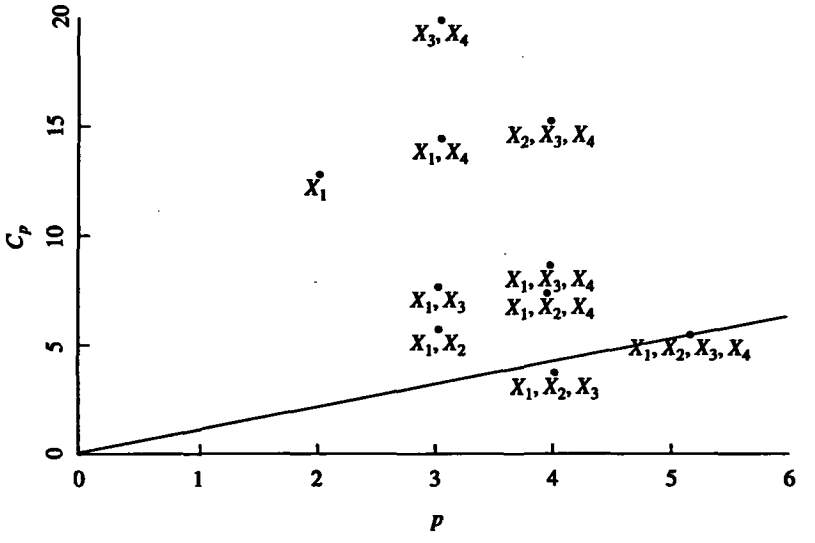
From Table 7.3.1 we see that if we use R^2 as the criterion measure, the best one-variable model uses X_1, which is SATmath; the second-best one-variable model uses X_3, which is HSmath; the best two-variable model uses X_1 and X_2—i.e., SATmath and SATverbal; the second-best two-variable model uses X_1 and X_3—i.e., SATmath and HSmath, etc.

The model with the smallest value of C_p is model (12) in Table 7.3.1, which uses X_1, X_2, and X_3. By the C_p criterion this model would be chosen as the best model. The value of $C_p - p$ computed for this model is -0.8, but recall that $C_p - p$ is estimating a population quantity that can never be negative, so we set $C_p - p = 0$

TABLE 7.3.1
s, R^2, *adj-*R^2, C_p, and $C_p - p$ for all 16 Subset Models for the GPA Data

Model	Predictors in the Model	p	s	R^2	*adj-*R^2	C_p	$C_p - p$
(1)	None	1	0.6218	not defined	not defined	83.9	82.9
(2)	X_1	2	0.3370	72.2	70.6	12.4	10.4
(3)	X_2	2	0.4837	42.7	39.5	42.4	40.4
(4)	X_3	2	0.4595	48.3	45.4	36.7	34.7
(5)	X_4	2	0.5079	36.8	33.3	48.4	46.4
(6)	X_1, X_2	3	0.2858	81.1	78.9	5.3	2.3
(7)	X_1, X_3	3	0.2998	79.2	76.7	7.2	4.2
(8)	X_1, X_4	3	0.3447	72.5	69.3	14.0	11.0
(9)	X_2, X_3	3	0.3971	63.5	59.2	23.2	20.2
(10)	X_2, X_4	3	0.4351	56.2	51.0	30.6	27.6
(11)	X_3, X_4	3	0.3771	67.1	63.2	19.5	16.5
(12)	X_1, X_2, X_3	4	0.2621	85.0	82.2	3.2	−0.8
(13)	X_1, X_2, X_4	4	0.2945	81.1	77.6	7.3	3.3
(14)	X_1, X_3, X_4	4	0.3014	80.2	76.5	8.2	4.2
(15)	X_2, X_3, X_4	4	0.3477	73.7	68.7	14.8	10.8
(16)	X_1, X_2, X_3, X_4	5	0.2685	85.3	81.4	5.0	0.0

FIGURE 7.3.1



for this model. The observed negative value of $C_p - p$ actually suggests that the bias for model (12) is close to zero if not equal to zero. This model also happens to be the model with the smallest s as well as the largest $adj\text{-}R^2$. Model (16) has the largest value of R^2, but as we mentioned earlier, no subset model can have a larger R^2. Based on Table 7.3.1 and Figure 7.3.1, the director of admissions may decide to restrict attention to models (6), (7), (12), (13), (14), and (16). ■

The entries in Table 7.3.1 can be calculated using formulas (7.3.8), (7.3.11), (7.3.12), and (7.3.14). However, we typically use a suitable statistical package—e.g., SAS, SPSS, BMDP, SPlus, MINITAB, etc.—for obtaining these quantities. In Section 7.3 of the laboratory manuals we demonstrate how to use the computer to obtain the quantities referred to above.

As stated earlier, computations for all-subsets regressions quickly become infeasible as the number of predictor variables increases. When it becomes infeasible or expensive to carry out an all-subsets regression analysis, we can resort to less expensive (but less desirable) methods. We discuss some such methods in the next section.

Problems 7.3

7.3.1 An investigator is interested in understanding the relationship, if any, between the analytical skills of young gifted children and the following variables:

X_1 = father's IQ

X_2 = mother's IQ

X_3 = age in months when the child first said "mommy" or "daddy"

X_4 = age in months when the child first counted to 10 successfully

X_5 = average number of hours per week the child's mother or father read to the child

X_6 = average number of hours per week the child watched an educational program on TV during the past 3 months

X_7 = average number of hours per week the child watched cartoons on TV during the past 3 months

The analytical skills are evaluated using a standard testing procedure, and the score Y on this test is used as the response variable. The model to be examined using all seven predictor variables is

$$\mu_Y(x_1, \ldots, x_7) = \beta_0 + \beta_1 x_1 + \cdots + \beta_7 x_7$$

Data were collected from schools in a large city on a set of thirty-six children who were identified as gifted children soon after they reached the age of four. These data are given in Table 7.3.2, and they are also stored in the file **gifted.dat** on the data disk.

TABLE 7.3.2
Gifted Children Data

Child	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	159	115	117	18	26	1.9	3.00	2.00
2	164	117	113	20	37	2.5	1.75	3.25
3	154	115	118	20	32	2.2	2.75	2.50
4	157	113	131	12	24	1.7	2.75	2.25
5	156	110	109	17	34	2.2	2.25	2.50
6	150	113	109	13	28	1.9	1.25	3.75
7	155	118	119	19	24	1.8	2.00	3.00
8	161	117	120	18	32	2.3	2.25	2.50
9	163	111	128	22	28	2.1	1.00	4.00
10	162	122	120	18	27	2.1	2.25	2.75
11	154	111	117	19	32	2.2	1.75	3.75
12	159	112	120	20	33	2.3	2.00	2.75
13	167	119	126	20	35	2.2	0.75	4.00
14	155	120	114	22	21	1.7	2.50	2.50
15	159	114	129	17	27	1.8	1.50	3.75
16	159	111	118	18	29	2.0	1.75	3.25
17	160	111	115	21	32	2.3	1.75	3.25
18	154	115	111	18	32	2.2	2.00	3.00

(Continued)

TABLE 7.3.2
(Continued)

Child	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
19	160	126	111	12	35	2.2	2.75	1.75
20	151	115	109	21	29	2.0	2.00	2.75
21	166	114	124	15	39	2.4	2.00	2.75
22	161	115	122	20	30	2.1	2.00	3.50
23	162	115	118	15	33	2.3	2.25	2.75
24	169	112	121	23	36	2.3	2.00	2.75
25	160	115	124	18	35	2.3	2.00	3.50
26	161	117	118	20	31	2.3	2.00	3.00
27	166	116	128	17	36	2.4	1.25	3.75
28	163	114	119	22	36	2.4	2.25	2.75
29	159	116	123	15	24	1.8	1.50	3.50
30	155	111	117	13	29	2.1	1.75	3.50
31	155	112	117	17	25	2.0	2.75	2.25
32	157	115	111	10	31	2.2	1.75	3.50
33	151	111	101	17	29	2.1	3.00	2.25
34	162	119	113	23	28	2.1	1.25	3.50
35	164	111	121	18	36	2.3	1.00	4.50
36	159	114	123	20	30	2.2	1.75	3.25

These data are considered to be a simple random sample from a study population of gifted children. In Exhibit 7.3.1 we give a MINITAB computer output for this problem containing the results from a subset analysis. Only the values of s, R^2, *adj-*R^2, and C_p are given for the best five subset models for each value of m, where m is the number of predictors in the subset model. This computer output is interpreted as follows.

When considering subset models with a *single* predictor, the best single variable is X_2 (using any of the four criterion functions s, R^2, *adj-*R^2, and C_p). Look across the line that contains the first 1 in the column labeled `Vars` (variables), where the 1 stands for the model with a single variable, until you come to the symbol `X`. Notice that `X` is below `X2`, so X_2 is the best single variable for predicting Y. The values of R^2, *adj-*R^2, C_p, and s for the subset model containing X_2 only are

$$R^2 = 32.6 \qquad \textit{adj-}R^2 = 30.7 \qquad C_p = 43.3 \qquad s = 3.8557$$

respectively. The second best variable for predicting Y when there is only one predictor is X_4. Look across the line that contains the second 1 under the column labeled `Vars` until you come to the symbol `X`. This `X` is under `X4`, so X_4 is the second best variable, among all single variable subset models, for predicting Y. The third best single variable is X_5, etc.

For the best subset models that contain two predictors, look under the column labeled `Vars` until you come to the first 2 (the 2 stands for models with two predictor variables). Look across that line until you come to two `X`'s. These `X`'s are under `X2` and `X5`, so X_2 and X_5 are the best two predictors in subset models that contain

EXHIBIT 7.3.1

MINITAB Output for All-Subsets Regression for Gifted Children Data

Best Subsets Regression of Y

Vars	R-sq	Adj. R-sq	C-p	s	X 1	X 2	X 3	X 4	X 5	X 6	X 7
1	32.6	30.7	43.3	3.8557		X					
1	29.6	27.5	46.7	3.9411				X			
1	27.6	25.5	49.0	3.9976					X		
1	13.7	11.2	64.5	4.3638						X	
1	7.2	4.4	71.8	4.5259			X				
2	62.9	60.7	11.5	2.9038		X			X		
2	60.8	58.4	13.9	2.9864		X		X			
2	37.8	34.0	39.5	3.7607		X	X				
2	36.7	32.9	40.7	3.7930	X	X					
2	36.3	32.5	41.2	3.8045		X				X	
3	68.7	65.8	7.0	2.7080	X	X			X		
3	66.7	63.5	9.3	2.7961	X	X		X			
3	64.6	61.3	11.5	2.8790		X	X	X			
3	64.5	61.2	11.7	2.8855		X	X		X		
3	63.7	60.3	12.6	2.9174		X			X	X	
4	70.8	67.0	6.7	2.6598	X	X	X	X			
4	70.7	66.9	6.8	2.6636		X			X	X	X
4	70.4	66.6	7.1	2.6765	X	X	X		X		
4	70.0	66.2	7.5	2.6933	X	X			X	X	
4	69.3	65.3	8.3	2.7257	X	X		X	X		
5	73.4	68.9	5.8	2.5805	X	X			X	X	X
5	71.8	67.1	7.5	2.6562	X	X	X	X	X		
5	71.6	66.9	7.7	2.6649		X	X		X	X	X
5	71.4	66.7	7.9	2.6733	X	X	X		X	X	
5	71.4	66.6	8.0	2.6759	X	X	X	X		X	
6	74.4	69.1	6.6	2.5726	X	X	X		X	X	X
6	73.5	68.0	7.6	2.6178	X	X		X	X	X	X
6	73.3	67.8	7.8	2.6270	X	X	X	X		X	X
6	72.5	66.8	8.7	2.6670	X	X	X	X	X	X	
6	71.9	66.1	9.4	2.6943		X	X	X	X	X	X
7	75.0	68.7	8.0	2.5905	X	X	X	X	X	X	X

exactly two predictors. You can look at the corresponding values of R^2, *adj*-R^2, C_p, and s to see how good that model is for predicting Y.

In Exhibit 7.3.2 we give a SAS computer output (edited for easier reading) for this problem containing the results from a subset analysis. To save space we have given the values of s, R^2, *adj*-R^2, and C_p for only the best five subset models for each value of m, where m is the number of predictors in the subset model.

EXHIBIT 7.3.2
SAS Output for All-Subsets Regression for Gifted Children Data

SAS Jan 1, Saturday, 1994

N = 36 Regression Models for Dependent Variable: Y

Number in Model	R-square	Adjusted R-square	C(p)	Root MSE	Variables in Model
1	0.32631738	0.30650318	43.32170	3.8557360	X2
1	0.29616080	0.27545965	46.69338	3.9410900	X4
1	0.27583227	0.25453322	48.96623	3.9975988	X5
1	0.13709230	0.11171266	64.47818	4.3637694	X6
1	0.07176563	0.04446462	71.78209	4.5259364	X3
2	0.62913424	0.60665753	11.46498	2.9038253	X2 X5
2	0.60775185	0.58397923	13.85566	2.9863628	X2 X4
2	0.37795937	0.34025994	39.54782	3.7607242	X2 X3
2	0.36724948	0.32890096	40.74525	3.7929609	X1 X2
2	0.36338752	0.32480494	41.17704	3.8045184	X2 X6
3	0.68725000	0.65792969	6.96730	2.7079632	X1 X2 X5
3	0.66655505	0.63529459	9.28112	2.7961223	X1 X2 X4
3	0.64649418	0.61335301	11.52404	2.8790047	X2 X3 X4
3	0.64490698	0.61161701	11.70150	2.8854607	X2 X3 X5
3	0.63700347	0.60297255	12.58516	2.9173956	X2 X5 X6
4	0.70770862	0.66999361	6.67990	2.6597832	X1 X2 X3 X4
4	0.70687611	0.66905367	6.77298	2.6635683	X2 X5 X6 X7
4	0.70401433	0.66582264	7.09295	2.6765390	X1 X2 X3 X5
4	0.70029934	0.66162829	7.50830	2.6932836	X1 X2 X5 X6
4	0.69303401	0.65342549	8.32061	2.7257333	X1 X2 X4 X5
5	0.73375503	0.68938087	5.76776	2.5804720	X1 X2 X5 X6 X7
5	0.71790384	0.67088781	7.54002	2.6561772	X1 X2 X3 X4 X5
5	0.71604225	0.66871596	7.74815	2.6649270	X2 X3 X5 X6 X7
5	0.71424574	0.66662002	7.94902	2.6733438	X1 X2 X3 X5 X6
5	0.71370085	0.66598432	8.00994	2.6758914	X1 X2 X3 X4 X6
6	0.74418888	0.69126244	6.60119	2.5726446	X1 X2 X3 X5 X6 X7
6	0.73512383	0.68032186	7.61472	2.6178305	X1 X2 X4 X5 X6 X7
6	0.73325500	0.67806638	7.82367	2.6270493	X1 X2 X3 X4 X6 X7
6	0.72508170	0.66820205	8.73749	2.6669932	X1 X2 X3 X4 X5 X6
6	0.71943293	0.66138457	9.36906	2.6942533	X2 X3 X4 X5 X6 X7
7	0.74956601	0.68695751	8.00000	2.5905185	X1 X2 X3 X4 X5 X6 X7

To read this table, suppose that you want to examine a model that contains m predictors. Select the number m and look under the column heading `Number in Model` for that value of m. Then the best five models (in order from first best, second best, ... , fifth best) are given under the column heading `Variables in Model`. For example, to find the variables in the third best model when the model contains $m = 2$ predictors, go down the column with heading `Number in Model` until you come to the number 2; then go down to the third 2 (because we want the third best model using two predictors); the value of R^2 is $0.378 = 37.8\%$. Go across this line to the column with heading `Variables in Model` and you see that the variables are X_2 and X_3. Hence the third best model that contains two predictor variables is $\beta_0 + \beta_2 x_2 + \beta_3 x_3$.

Answer (a)–(i) using both the MINITAB and the SAS exhibits and notice that both give the same answer.

a Find the predictor variable that results in the best prediction function (for the sample data at hand) among all subset models containing only *one* predictor.

b Find the two predictor variables that result in the best prediction function (for the sample data at hand) among all subset models containing exactly *two* predictors.

c Find the three predictor variables that result in the best prediction function (for the sample data at hand) among all subset models containing exactly *three* predictors.

d Find a short list of the three best models using the criterion R^2.

e Find a short list of the three best models using the criterion *adj-*R^2.

f Find a short list of the three best models using the criterion s.

g Find a short list of the three best models using the criterion C_p.

h Prepare a short list consisting of four models using C_p and $C_p - p$ together.

i Write a short report summarizing the results of parts (a) through (h).

7.3.2 Consider the GPA data of Example 4.4.2. Suppose that the model that we want to examine for this study is

$$\mu_Y(x_1, \ldots, x_6) = \beta_0 + \beta_1 x_1 + \cdots + \beta_6 x_6$$

where $X_1 = \text{SATmath}$, $X_2 = \text{SATverbal}$, $X_3 = \text{HSmath}$, $X_4 = \text{HSenglish}$, $X_5 = X_1X_2$, $X_6 = X_3X_4$. The raw data, which include the values of Y, X_1, X_2, X_3, and X_4, are in Table 7.3.3 along with the values of the derived variables $X_5 = X_1X_2$ and $X_6 = X_3X_4$. These data are also stored in the file **table733.dat** on the data disk.

A MINITAB output is given in Exhibit 7.3.3. It contains the results of an all-subsets regression analysis based on the six predictor variables $X_1, \ldots, X_6$. Only the results for the best five subset models for each subset size are given. A SAS output is given in Exhibit 7.3.4. It contains the results of an all-subsets regression analysis based on the six predictor variables $X_1, \ldots, X_6$. Only the results for the best five subset models for each subset size are given. We give a MINITAB output as well as a SAS output so that you can compare them against each other and become familiar with both. Answer (a)–(i) using either the MINITAB output or the SAS output. Of course your answers will be the same either way.

TABLE 7.3.3

Subject	Y	X_1	X_2	X_3	X_4	$X_5 = X_1X_2$	$X_6 = X_3X_4$
1	1.97	321	247	2.30	2.63	79287	6.0490
2	2.74	718	436	3.80	3.57	313048	13.5660
3	2.19	358	578	2.98	2.57	206924	7.6586
4	2.60	403	447	3.58	2.21	180141	7.9118
5	2.98	640	563	3.38	3.48	360320	11.7624
6	1.65	237	342	1.48	2.14	81054	3.1672
7	1.89	270	472	1.67	2.64	127440	4.4088
8	2.38	418	356	3.73	2.52	148808	9.3996
9	2.66	443	327	3.09	3.20	144861	9.8880
10	1.96	359	385	1.54	3.46	138215	5.3284
11	3.14	669	664	3.21	3.37	444216	10.8177
12	1.96	409	518	2.77	2.60	211862	7.2020
13	2.20	582	364	1.47	2.90	211848	4.2630
14	3.90	750	632	3.14	3.49	474000	10.9586
15	2.02	451	435	1.54	3.20	196185	4.9280
16	3.61	645	704	3.50	3.74	454080	13.0900
17	3.07	791	341	3.20	2.93	269731	9.3760
18	2.63	521	483	3.59	3.32	251643	11.9188
19	3.11	594	665	3.42	2.70	395010	9.2340
20	3.20	653	606	3.69	3.52	395718	12.9888

a Find the predictor variable that results in the best one-variable subset model.

b Find the two predictor variables that result in the best two-variable subset model.

c Find the three predictor variables that result in the best three-variable subset model.

d Find a short list of the three best models using the R^2 criterion.

e Find a short list of the three best models using the *adj-*R^2 criterion.

f Find a short list of the three best models using the s criterion.

g Find a short list of the three best models using the C_p criterion.

h Find a short list of three models using C_p and $C_p - p$ together.

i Write a short report summarizing the results of parts (a) through (h).

EXHIBIT 7.3.3

MINITAB Output for Problem 7.3.2

Best Subsets Regression of Y

Vars	R-sq	Adj. R-sq	C-p	s	X 1	X 2	X 3	X 4	X 5	X 6
1	81.4	80.4	10.7	0.27515					X	
1	72.2	70.6	24.0	0.33700	X					
1	65.9	64.0	33.0	0.37286						X
1	48.3	45.4	58.4	0.45949			X			
1	42.7	39.5	66.4	0.48367		X				
2	86.3	84.7	5.6	0.24301			X		X	
2	85.9	84.3	6.3	0.24675					X	X
2	85.8	84.1	6.4	0.24757		X			X	
2	84.3	82.4	8.6	0.26075	X				X	
2	81.5	79.3	12.6	0.28271				X	X	
3	89.8	87.9	2.7	0.21638		X	X		X	
3	88.4	86.2	4.7	0.23106		X			X	X
3	88.0	85.8	5.2	0.23458	X		X		X	
3	87.1	84.7	6.5	0.24308	X				X	X
3	86.7	84.2	7.1	0.24728			X	X	X	
4	90.4	87.9	3.8	0.21657	X	X	X		X	
4	89.9	87.2	4.5	0.22264		X	X		X	X
4	89.8	87.1	4.6	0.22339		X	X	X	X	
4	89.3	86.4	5.4	0.22903		X		X	X	X
4	88.7	85.7	6.2	0.23518	X	X			X	X
5	90.6	87.2	5.6	0.22264	X	X	X		X	X
5	90.4	87.0	5.8	0.22407	X	X	X	X	X	
5	90.1	86.5	6.3	0.22837		X	X	X	X	X
5	89.7	86.1	6.8	0.23223	X	X		X	X	X
5	88.1	83.9	9.0	0.24944	X		X	X	X	X
6	91.0	86.8	7.0	0.22607	X	X	X	X	X	X

EXHIBIT 7.3.4
SAS Output for Problem 7.3.2

SAS 0:00 Saturday, Jan 1, 1994

N = 20 Regression Models for Dependent Variable: Y

Number in Model	R-square	Adjusted R-square	C(p)	Root MSE	Variables in Model
1	0.81449296	0.80418702	10.66302	0.27514656	X5
1	0.72170925	0.70624865	23.99888	0.33700262	X1
1	0.65934389	0.64041855	32.96268	0.37285672	X6
1	0.48264669	0.45390483	58.35946	0.45949153	X3
1	0.42677523	0.39492941	66.38990	0.48366690	X2
2	0.86333578	0.84725764	5.64282	0.24300941	X3 X5
2	0.85909950	0.84252297	6.25170	0.24674704	X5 X6
2	0.85815613	0.84146861	6.38729	0.24757169	X2 X5
2	0.84265718	0.82414626	8.61497	0.26074689	X1 X5
2	0.81503604	0.79327558	12.58497	0.28270874	X4 X5
3	0.89802358	0.87890300	2.65712	0.21637647	X2 X3 X5
3	0.88371683	0.86191373	4.71344	0.23105671	X2 X5 X6
3	0.88014407	0.85767108	5.22696	0.23457943	X1 X3 X5
3	0.87129761	0.84716591	6.49846	0.24308237	X1 X5 X6
3	0.86681811	0.84184651	7.14230	0.24727644	X3 X4 X5
4	0.90422221	0.87868146	3.76619	0.21657430	X1 X2 X3 X5
4	0.89878613	0.87179576	4.54752	0.22263556	X2 X3 X5 X6
4	0.89809996	0.87092662	4.64615	0.22338895	X2 X3 X4 X5
4	0.89288831	0.86432520	5.39522	0.22903030	X2 X4 X5 X6
4	0.88706066	0.85694350	6.23283	0.23517824	X1 X2 X5 X6
5	0.90553163	0.87179293	5.57799	0.22263802	X1 X2 X3 X5 X6
5	0.90430939	0.87013417	5.75366	0.22407365	X1 X2 X3 X4 X5
5	0.90060360	0.86510489	6.28630	0.22837126	X2 X3 X4 X5 X6
5	0.89721211	0.86050215	6.77376	0.23223470	X1 X2 X4 X5 X6
5	0.88141776	0.83906695	9.04389	0.24943992	X1 X3 X4 X5 X6
6	0.90955296	0.86780818	7.00000	0.22607141	X1 X2 X3 X4 X5 X6

7.4
Alternative Methods for Subset Selection

In this section we discuss three methods of variable selection that require fewer computations than the method of all-subsets regression. These methods are

1 Forward selection
2 Backward elimination
3 Stepwise regression

Unlike the all-subsets regression procedure, none of these methods generally examines all the possible subset models, and they should be used only when computing facilities are not adequate for handling the all-subsets regression procedure. But like the all-subsets regression procedure, these methods are to be regarded as only exploratory or descriptive in nature. No rigorous confidence statements regarding the conclusions are generally possible. Thus we regard these methods as mainly suitable for exploratory analysis whether or not assumptions (A) or (B) are satisfied. Nevertheless these three methods, particularly stepwise regression, have been found to be quite useful in subset analysis. The drawback of not being able to obtain confidence statements regarding the results may be overcome by conducting additional studies and by obtaining new data for validation (or invalidation) of the conclusions arrived at with the help of these procedures. Thus a tentative model can be obtained using the procedures discussed here and in the previous section, and this tentative model can then be evaluated by selecting and using a new sample from the same population.

The forward selection and the backward elimination procedures are special cases of the stepwise regression procedure. They are computationally somewhat less taxing than stepwise regression. However, for purely exploratory purposes, stepwise regression is the recommended procedure among the three being considered here. Nevertheless, we first discuss the forward selection and the backward elimination methods, mainly because they facilitate the understanding of the stepwise regression procedure. An additional reason for discussing them in this book is that some researchers occasionally use them and consequently journal articles contain references to these methods.

Forward Selection

This procedure begins with the simplest function, viz.,

$$\beta_0 \tag{7.4.1}$$

and successively adds one variable at a time to the model in such a way that at each step the variable added is the best variable that can be added. To make ideas concrete, we describe the forward selection algorithm using $k = 4$; i.e., the total number of predictors under consideration is four. We label them X_1, X_2, X_3, and X_4 as usual. We now describe the forward selection algorithm.

At each step of the algorithm we will have a current model, and we will choose a predictor variable not already included in the current model as the best candidate variable for adding to that model. Any criterion measure—s, R^2, $adj\text{-}R^2$, or C_p—described in Section 7.3 can be used, and each measure will select the same best candidate variable. Whether or not this candidate variable is actually added to the current model depends on whether a computed quantity, which we denote by F_C, exceeds a criterion value, which we denote by *F-in*. This criterion value is chosen by the investigator to somewhat correspond to a tabled F-value with 1 degree of freedom in the numerator and $n - (m + 1)$ degrees of freedom in the denominator ($m + 1$ is the number of βs in the model under consideration). You will note in Table T-5 in Appendix T, that if $n - m - 1$ is larger than 10, then $F_{0.95:1,n-m-1}$ is close to 4.0 so 4.0 is a commonly used value for *F-in*. More will be said about this later.

The criterion value *F-in* is denoted differently in different statistical computing packages. MINITAB uses the name FENTER to refer to the criterion value *F-in*. Other statistical packages use other names. Some statistical packages such as SAS use criteria related to P-values in place of F-table values.

The algorithm proceeds as follows. We start with the model

$$\beta_0 \tag{7.4.2}$$

as the current model. We calculate *SSY*, the sum of squared errors when using $\bar{y}$ to predict Y. Actually $\bar{y}$ is $\hat{\beta}_0$, the least squares estimate of β_0 for the model in (7.4.2).

We define a calculation (or set of calculations) as a step if a predictor variable is added to the current model.

Step 1 For each predictor variable X_j, $j = 1, \ldots, 4$, fit the model

$$\beta_0 + \beta_j x_j \tag{7.4.3}$$

by least squares and obtain $SSE(X_j)$, the sum of squared errors when using $\hat{\beta}_0 + \hat{\beta}_j x_j$ to predict Y. Choose the variable X_j that results in the smallest value for $SSE(X_j)$ as the best candidate variable to be added to the current model. Note that the variable X_j with the smallest value for $SSE(X_j)$ at this step is the same variable X_j with the largest value of $R^2(X_j)$, or $adj\text{-}R^2(X_j)$, or the smallest value of $s(X_j)$, or $C_p(X_j)$. Thus any of these criteria can be used in place of *SSE* to identify the best candidate variable at each step. They all give the same result. However, to simplify discussion, we use the smallest *SSE*. Suppose, for concreteness, this variable is X_1. Calculate

$$F_C = \frac{SSY - SSE(X_1)}{MSE(X_1)} \tag{7.4.4}$$

where $MSE(X_1) = SSE(X_1)/(n - 2)$. If F_C is less than or equal to *F-in*, then the algorithm stops and the original model in (7.4.2) is the final model.

If F_C is greater than *F-in*, then add X_1 to the current model. In this case in step 1 the variable X_1 is added and the resulting model, which contains variable X_1, is

$$\beta_0 + \beta_1 x_1 \tag{7.4.5}$$

Proceed to step 2.

Step 2 The revised current model is

$$\beta_0 + \beta_1 x_1$$

The predictor variables not in the current model at this step are X_2, X_3, and X_4. For $j = 2, 3, 4$, fit the model

$$\beta_0 + \beta_1 x_1 + \beta_j x_j \tag{7.4.6}$$

and obtain $SSE(X_1, X_j)$. Choose the variable X_j that results in the smallest value for $SSE(X_1, X_j)$ as the best candidate variable to be added to the current model. Suppose this variable is X_2. Calculate

$$F_C = \frac{SSE(X_1) - SSE(X_1, X_2)}{MSE(X_1, X_2)} \tag{7.4.7}$$

where $MSE(X_1, X_2) = SSE(X_1, X_2)/(n - 3)$.

If F_C is less than or equal to *F-in*, the algorithm stops and chooses the model in (7.4.5) as the final model (i.e., the algorithm chooses the model in step 1 as the final model). If F_C is greater than *F-in*, then add X_2 to the current model. In this case in step 2 the variable X_2 is added to the current model and the resulting model, which contains variables X_1, X_2, is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{7.4.8}$$

Proceed to step 3.

Step 3 The revised current model is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2$$

The predictor variables not in the current model at this stage are X_3 and X_4. For $j = 3, 4$, fit the model

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_j x_j \tag{7.4.9}$$

and obtain $SSE(X_1, X_2, X_j)$. Choose the variable X_j that results in the smallest value for $SSE(X_1, X_2, X_j)$ as the best candidate variable to be added to the current model. Suppose this variable is X_3. Calculate

$$F_C = \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{MSE(X_1, X_2, X_3)} \tag{7.4.10}$$

where $MSE(X_1, X_2, X_3) = SSE(X_1, X_2, X_3)/(n - 4)$. If F_C is less than or equal to *F-in*, then the algorithm stops and chooses the model in (7.4.8) as the final model (i.e., the algorithm chooses the model in step 2 as the final model).

If F_C is greater than *F-in*, then add X_3 to the current model. In this case in step 3 the variable X_3 is added and the resulting model, which contains variables X_1, X_2, X_3, is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{7.4.11}$$

Proceed to step 4.

Step 4 At step 4 the revised current model is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

The only predictor variable not in the current model at this stage is X_4. Fit the model

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \tag{7.4.12}$$

and obtain $SSE(X_1, X_2, X_3, X_4)$. Calculate

$$F_C = \frac{SSE(X_1, X_2, X_3) - SSE(X_1, X_2, X_3, X_4)}{MSE(X_1, X_2, X_3, X_4)} \tag{7.4.13}$$

where $MSE(X_1, X_2, X_3, X_4) = SSE(X_1, X_2, X_3, X_4)/(n-5)$.

If F_C is less than or equal to *F-in*, then the algorithm stops and chooses the model in (7.4.11) as the final model (i.e., the algorithm chooses the model in step 3 as the final model). If F_C is greater than *F-in*, then add X_4 to the current model and, since all the predictor variables are already included in the current model, there is no need to proceed further and the algorithm stops. In this case at step 4 the model is the one given in (7.4.14) and is the final model.

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \tag{7.4.14}$$

Although we used $k = 4$ for simplicity of explanation, the procedure can be extended in an obvious manner to handle any number of predictor variables. We now illustrate this procedure in Example 7.4.1.

EXAMPLE 7.4.1

Consider the GPA data of Example 4.4.2. We apply the forward selection procedure to choose a subset (which may turn out to be the full set in some problems) of the four predictor variables. The sums of squares and mean squares required during various steps of the forward selection algorithm may be obtained using a statistical package such as SAS, SPlus, BMDP, SPSS, MINITAB, etc. In Table 7.4.1 we give the error sums of squares and mean squares for all subsets, although in a real problem these would not all be computed (that is the advantage of the forward selection procedure over the all-subsets regression procedure). We select *F-in* = 4.0 and start with the model

$$\beta_0 \tag{7.4.15}$$

with no predictor variables as the current model.

Step 1 The variables not in the model are X_1, X_2, X_3, X_4. We regress Y on each of these variables and obtain

$$SSE(X_1) = 2.0443$$
$$SSE(X_2) = 4.2108$$
$$SSE(X_3) = 3.8004$$
$$SSE(X_4) = 4.6439$$

TABLE 7.4.1
Sums of Squares and Mean Squares for All 16 Subset Models for the GPA Data

Model	Predictors in the Model	Sum of Squares	Mean Square
(1)	None	7.3458	0.3866
(2)	X_1	2.0443	0.1136
(3)	X_2	4.2108	0.2339
(4)	X_3	3.8004	0.2111
(5)	X_4	4.6439	0.2580
(6)	X_1, X_2	1.3884	0.0817
(7)	X_1, X_3	1.5282	0.0899
(8)	X_1, X_4	2.0199	0.1188
(9)	X_2, X_3	2.6812	0.1577
(10)	X_2, X_4	3.2179	0.1893
(11)	X_3, X_4	2.4180	0.1422
(12)	X_1, X_2, X_3	1.0992	0.0687
(13)	X_1, X_2, X_4	1.3881	0.0868
(14)	X_1, X_3, X_4	1.4532	0.0908
(15)	X_2, X_3, X_4	1.9344	0.1209
(16)	X_1, X_2, X_3, X_4	1.0815	0.0721

The predictor variable that leads to the smallest value for SSE is X_1. We calculate

$$F_C = \frac{SSY - SSE(X_1)}{MSE(X_1)} = \frac{7.3458 - 2.0443}{0.1136} = 46.67 \qquad \textbf{(7.4.16)}$$

Since $F_C > F\text{-}in = 4.0$, we add X_1 to the current model. So in step 1 the variable X_1 is added to the current model and the resulting model, which contains variable X_1, is

$$\beta_0 + \beta_1 x_1 \qquad \textbf{(7.4.17)}$$

Proceed to step 2.

Step 2 The revised current model is

$$\beta_0 + \beta_1 x_1$$

The variables not in the current model are X_2, X_3, and X_4. We add each of these variables in turn to the current model in (7.4.17) and calculate the corresponding sums of squared errors.

$$SSE(X_1, X_2) = 1.3884$$
$$SSE(X_1, X_3) = 1.5282$$
$$SSE(X_1, X_4) = 2.0199$$

Hence the best predictor variable for adding to the current model is X_2. We calculate

$$F_C = \frac{SSE(X_1) - SSE(X_1, X_2)}{MSE(X_1, X_2)} = \frac{2.0443 - 1.3884}{0.0817} = 8.03 \qquad \textbf{(7.4.18)}$$

Because $F_C > F\text{-}in = 4.0$, we add X_2 to the current model. So in step 2 the variable X_2 is added and the resulting model, which contains the variables X_1, X_2, is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{7.4.19}$$

Proceed to step 3.

Step 3 The revised current model is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2$$

The variables not in the current model are X_3 and X_4. We add each of these variables in turn to the current model in (7.4.19) and calculate the corresponding *SSE*.

$$SSE(X_1, X_2, X_3) = 1.0992$$
$$SSE(X_1, X_2, X_4) = 1.3881$$

Hence the best predictor variable for adding to the current model is X_3. We calculate

$$F_C = \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{MSE(X_1, X_2, X_3)} = \frac{1.3884 - 1.0992}{0.0687} = 4.21 \tag{7.4.20}$$

Because $F_C > F\text{-}in = 4.0$, we add X_3 to the current model. So in step 3 the variable X_3 is added and the resulting model, which contains the variables X_1, X_2, X_3, is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{7.4.21}$$

Proceed to step 4.

Step 4 The revised current model is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

The only variable not in the current model is X_4. We add X_4 to the current model in (7.4.21) and calculate the corresponding *SSE*.

$$SSE(X_1, X_2, X_3, X_4) = 1.0815 \tag{7.4.22}$$

Also,

$$F_C = \frac{SSE(X_1, X_2, X_3) - SSE(X_1, X_2, X_3, X_4)}{MSE(X_1, X_2, X_3, X_4)} = \frac{1.0992 - 1.0815}{0.0721}$$
$$= 0.25 \tag{7.4.23}$$

Because $F_C < F\text{-}in = 4.0$, we do not add X_4 to the current model. Thus the model in (7.4.21) is the final model selected by the forward selection algorithm using $F\text{-}in = 4.0$.

Next we summarize the variable added and the variables in the model at each step.

Step	Variables Added	Variables in the Model
Start	—	None
1	X_1	X_1
2	X_2	X_1, X_2
3	X_3	X_1, X_2, X_3 ■

Some comments are in order.

a We regard the forward selection procedure as a purely exploratory analysis technique, whether or not assumptions (A) or (B) hold, because, even when these assumptions are satisfied, no measure of confidence is available that can be used to judge whether the final model is an adequate model.

b The decision to add or not add a candidate variable to the current model at any particular step is sometimes based on an F-test, discussed in Section 4.9, which compares two competing models where one model is nested in the other.

To see this in more detail, suppose that the current model in step 1 is

$$\beta_0 + \beta_1 x_1$$

Suppose that in step 2 the best candidate variable to add to this model is X_2. To determine if X_2 should be added to the current model, we decide whether or not the model

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2$$

is better than the model

$$\beta_0 + \beta_1 x_1$$

If we decide $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ is better than $\beta_0 + \beta_1 x_1$, then the variable X_2 is added. If we decide $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ is no better than $\beta_0 + \beta_1 x_1$, then the variable X_2 is not added. Many variable selection procedures make this decision based on a test of

$$\text{NH: } \beta_2 = 0 \qquad \text{against} \qquad \text{AH: } \beta_2 \neq 0$$

In Section 4.9 (under appropriate assumptions) we discussed an F-test for testing the preceding null hypothesis. According to this F-test, NH is rejected if a computed F, denoted by F_C, is larger than a tabled F value; viz., $F_{1-\alpha:1,dfd}$ where the degrees of freedom for the numerator is 1 and *dfd* is the degrees of freedom for the denominator. If NH is rejected, then based on the test the procedure will conclude that $\beta_2 \neq 0$ and that the model $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ is better than the model $\beta_0 + \beta_1 x_1$. Thus the variable X_2 is added to the model and the variables in the model at this step are X_1 and X_2; the resulting model, which is the new current model, is $\beta_0 + \beta_1 x_1 + \beta_2 x_2$. If F_C is less than or equal to the tabled F-value, then NH is not rejected and, based on the test, the procedure will conclude that the model $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ is no better than the model $\beta_0 + \beta_1 x_1$. Hence X_2 is not added to the current model. Thus the current model is still $\beta_0 + \beta_1 x_1$.

Whereas the forward selection procedure is *guided* by a statistical test, it is not a *valid* F-test, and the actual Type-I error is not known even though we use an α value of our choice. It is quite common (but arbitrary) to carry out this test using $\alpha = 0.05$. It may be verified that the tabled F-values for carrying out this test are roughly in the neighborhood of 4.0 when there are enough degrees of freedom for the denominator. For this reason *F-in* is often taken to be equal to 4.0. Some authors recommend using *F-in* = 2.0. In principle, any nonnegative value may be chosen as the value for *F-in*. However, the final model chosen by this procedure is generally dependent on the value of *F-in* that is used. It is clear that larger values of *F-in* tend to result in the selection of a final model with a smaller number of variables. You should convince yourself of this.

One reason why the preceding test is not valid, even if population assumptions (A) or (B) are satisfied for the population $\{(Y, X_1, \ldots, X_k)\}$, is that we look at all models $\beta_0 + \beta_1 x_1 + \beta_j x_j$ for $j = 2, 3, 4$ to find the best candidate model. Say it is $\beta_0 + \beta_1 x_1 + \beta_2 x_2$; then the forward selection procedure tests

$$\text{NH: } \beta_2 = 0 \qquad \text{against} \qquad \text{AH: } \beta_2 \neq 0$$

to decide whether the model $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ is better than the model $\beta_0 + \beta_1 x_1$. *The test is not a valid F-test because the data are used to determine which model to test.* In Section 4.9 the test is a valid F-test if the data are not used to determine which test to make, i.e., if we decide to test $\beta_2 = 0$ before looking at the data.

c Notice that the forward selection procedure does not examine each and every subset model, and it is quite possible that there exist other subsets of variables that predict Y as well as, or better than, the chosen subsets. This is the price we pay for the savings in computing that results from not examining every possible model.

d Some users of this procedure interpret the order in which the variables are entered into the model as the order of importance of the variables. This is an incorrect interpretation, particularly in view of the fact that the concept of importance is usually not precisely defined. For instance, the following situation occurs not infrequently. $\sigma_{Y|X_1}$ is the smallest and $\sigma_{Y|X_2}$ the second smallest among the quantities $\sigma_{Y|X_i}$, $i = 1, \ldots, k$, whereas $\sigma_{Y|X_2,X_3}$ is substantially smaller than $\sigma_{Y|X_1,X_3}$. Thus X_1 alone is a better predictor of Y than X_2 alone, but X_2 in the presence of X_3 is more useful as a predictor of Y than X_1 is in the presence of X_3. Hence it is usually not possible to precisely define a meaningful order of importance based on statistical considerations alone.

e Instead of selecting the candidate variable to add to the current model by using the sum of squared errors, we can use any of the criterion measures R^2, *adj-*R^2, s, or C_p discussed in Section 7.3. This will not alter the candidate variable to add at any step.

f Notice that some of the predictor variables can be derived variables, i.e., known functions of the *basic* predictor variables. For instance, we could have started the forward selection procedure in Example 7.4.1 with the

set $\{X_1, X_2, X_3, X_4, X_5, X_6\}$ as the full set of predictors, where $X_5 = X_1^2 =$ (SATmath)^2, $X_6 = X_3X_4 =$ (HSmath) $\times$ (HSenglish), etc.

Backward Elimination

Whereas the forward selection procedure begins with the constant model (i.e., the model β_0 that includes no predictor variables), the backward elimination procedure begins with the model that includes *all* of the available predictor variables, viz.,

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{7.4.24}$$

and proceeds by successively removing from the model one variable at a time in such a way that, at each step, the variable removed is the variable contributing the least to the prediction of Y at that step. The details of the algorithm are described next using $k = 4$ for simplicity. Thus the predictor variables are X_1, X_2, X_3, and X_4.

At each step of the algorithm (we define a step as a calculation or set of calculations in which a variable is deleted) we will have a current model, and we will label a predictor variable included in the current model as the best candidate variable for deletion from the model. Whether or not this candidate variable is actually deleted from the current model depends on whether a computed quantity, denoted by F_C, is smaller than a criterion value which we call *F-out*. This criterion value is chosen by the investigator to somewhat correspond to a tabled F-value with 1 degree of freedom in the numerator and $n - (m + 1)$ degrees of freedom in the denominator ($m + 1$ is the number of βs in the current model).

The criterion value *F-out* is denoted differently in different statistical computing packages. MINITAB uses the name FREMOVE to refer to the criterion value *F-out*. Other statistical packages use other names. Some statistical packages such as SAS use criteria related to P-values in place of F-table values.

We start with the model

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \tag{7.4.25}$$

as the current model. Fit this model by the method of least squares and calculate the quantity $SSE(X_1, X_2, X_3, X_4)$, the sum of squared errors when using $\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4$ to predict Y.

Step 1 In step 1 the variables in the model are X_1, X_2, X_3, X_4. For each predictor variable $X_j, j = 1, \ldots, 4$, fit the model obtained by deleting this predictor variable from the current model and calculate the corresponding *SSE*. In the case $k = 4$, this leads us to consider the following four models.

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \quad \text{i.e., } X_4 \text{ is omitted} \tag{7.4.26}$$

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 \quad \text{i.e., } X_3 \text{ is omitted} \tag{7.4.27}$$

$$\beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 \quad \text{i.e., } X_2 \text{ is omitted} \tag{7.4.28}$$

$$\beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \quad \text{i.e., } X_1 \text{ is omitted} \tag{7.4.29}$$

and the corresponding *SSE* are

$$SSE(X_1, X_2, X_3) \quad SSE(X_1, X_2, X_4) \quad SSE(X_1, X_3, X_4) \quad SSE(X_2, X_3, X_4)$$

respectively. Suppose the smallest among these *SSE* is $SSE(X_1, X_2, X_3)$. This means that if we want to delete one of the predictors in the current model, the best candidate to delete will be X_4 because the remaining three predictors, X_1, X_2, and X_3, are the best among all three predictor subset models of the current model. Calculate

$$F_C = \frac{SSE(X_1, X_2, X_3) - SSE(X_1, X_2, X_3, X_4)}{MSE(X_1, X_2, X_3, X_4)} \tag{7.4.30}$$

where $MSE(X_1, X_2, X_3, X_4) = SSE(X_1, X_2, X_3, X_4)/(n - 5)$.

If F_C is greater than *F-out*, then the algorithm stops and chooses the model in (7.4.25) as the final model. In this case no variables are deleted in step 1, and the variables in the model are X_1, X_2, X_3, and X_4.

If F_C is less than or equal to *F-out*, then delete X_4 from the current model. In this case the variable X_4 is deleted in step 1 and the model, which contains the variables X_1, X_2, X_3, is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{7.4.31}$$

Proceed to step 2.

Step 2 In step 2 the variables in the model are X_1, X_2, and X_3, so the revised current model is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

For each predictor variable $X_j, j = 1, 2, 3$, fit the model obtained by deleting this predictor variable from the current model and calculate the corresponding *SSE*. This step leads us to consider the following three models

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad \text{i.e., } X_3 \text{ is omitted from (7.4.31)} \tag{7.4.32}$$

$$\beta_0 + \beta_1 x_1 + \beta_3 x_3 \quad \text{i.e., } X_2 \text{ is omitted from (7.4.31)} \tag{7.4.33}$$

$$\beta_0 + \beta_2 x_2 + \beta_3 x_3 \quad \text{i.e., } X_1 \text{ is omitted from (7.4.31)} \tag{7.4.34}$$

and the corresponding *SSE* are $SSE(X_1, X_2)$, $SSE(X_1, X_3)$, and $SSE(X_2, X_3)$, respectively. Suppose the smallest among these *SSE* is $SSE(X_1, X_2)$. This means that if we want to delete one of the predictors in the current model, which is given in (7.4.31), the best candidate to delete will be X_3. Calculate

$$F_C = \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{MSE(X_1, X_2, X_3)} \tag{7.4.35}$$

where $MSE(X_1, X_2, X_3) = SSE(X_1, X_2, X_3)/(n - 4)$.

If F_C is greater than *F-out*, then the algorithm stops and chooses the model in (7.4.31) as the final model.

If F_C is less than or equal to *F-out*, then delete X_3 from the current model. In this case the variable X_3 is deleted in step 2 and the model, which contains the variables

X_1, X_2, is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{7.4.36}$$

Proceed to step 3.

Step 3 In step 3 the variables in the model are X_1, X_2, so the revised current model is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2$$

For each predictor variable X_j, $j = 1, 2$, fit the model obtained by deleting this predictor variable from the current model and calculate the corresponding *SSE*. At this step, we are led to consider the following two models.

$$\beta_0 + \beta_1 x_1 \quad \text{i.e., } X_2 \text{ is omitted from (7.4.36)} \tag{7.4.37}$$

$$\beta_0 + \beta_2 x_2 \quad \text{i.e., } X_1 \text{ is omitted from (7.4.36)} \tag{7.4.38}$$

and the corresponding *SSE* are $SSE(X_1)$ and $SSE(X_2)$, respectively. Suppose the smaller of these two *SSE* is $SSE(X_1)$. This means that if we want to delete one of the predictors in the current model, which is in (7.4.36), the best candidate to delete will be X_2. Calculate

$$F_C = \frac{SSE(X_1) - SSE(X_1, X_2)}{MSE(X_1, X_2)} \tag{7.4.39}$$

where $MSE(X_1, X_2) = SSE(X_1, X_2)/(n - 3)$.

If F_C is greater than *F-out*, then the algorithm stops and chooses the model in (7.4.36) as the final model.

If F_C is less than or equal to *F-out*, then delete X_2 from the current model. In this case the variable X_2 is deleted in step 3 and the model, which contains variable X_1, is

$$\beta_0 + \beta_1 x_1 \tag{7.4.40}$$

Proceed to step 4.

Step 4 In step 4 the variable in the model is X_1, so the revised current model is

$$\beta_0 + \beta_1 x_1$$

The only predictor variable in the current model at this step is X_1. If this variable is deleted, then the resulting model is β_0. The *SSE* when Y is predicted using $\hat{\beta}_0$ (which is equal to $\bar{y}$) is *SSY*. Calculate

$$F_C = \frac{SSY - SSE(X_1)}{MSE(X_1)} \tag{7.4.41}$$

where $MSE(X_1) = SSE(X_1)/(n - 2)$.

If F_C is greater than *F-out*, then the algorithm stops and chooses the model in (7.4.40) as the final model.

If F_C is less than or equal to *F-out*, then delete X_1 from the current model. In this case the variable X_1 is deleted in step 4 and there are no variables left in the model.

Because there are no predictor variables in the current model, there is no need to proceed further, and so the algorithm stops and declares the model

$$\beta_0 \tag{7.4.42}$$

to be the final model.

Although we used $k = 4$ for simplicity of explanation, the procedure can be extended in an obvious manner to handle any number of predictor variables.

We now illustrate this procedure with an example.

EXAMPLE 7.4.2

Consider the GPA data of Example 4.4.2. We apply the backward elimination procedure to choose a subset (which may turn out to be the full set in some problems) of the four predictor variables.

The required sums of squares and mean squares can be obtained using a statistical package. However, all of the sums of squares needed are given in Table 7.4.1.

We select *F-out* = 4.0 and start with the model

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \tag{7.4.43}$$

as the current model. The value of $SSE(X_1, X_2, X_3, X_4)$ is 1.0815.

Step 1 The variables in the current model are X_1, X_2, X_3, X_4. The models that would result when one of these four predictors is deleted from the current model are

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \quad \text{i.e., } X_4 \text{ is omitted} \tag{7.4.44}$$
$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 \quad \text{i.e., } X_3 \text{ is omitted} \tag{7.4.45}$$
$$\beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 \quad \text{i.e., } X_2 \text{ is omitted} \tag{7.4.46}$$
$$\beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \quad \text{i.e., } X_1 \text{ is omitted} \tag{7.4.47}$$

respectively. The *SSE* corresponding to these four models are

$$SSE(X_1, X_2, X_3) = 1.0992$$
$$SSE(X_1, X_2, X_4) = 1.3881$$
$$SSE(X_1, X_3, X_4) = 1.4532$$
$$SSE(X_2, X_3, X_4) = 1.9344$$

Thus we see that the model using X_1, X_2, X_3 to predict Y, which is obtained by deleting X_4 from the current model in (7.4.43), leads to the smallest *SSE* among the four models given in (7.4.44)–(7.4.47). Hence X_4 is the best candidate for deletion at this step. We calculate

$$F_C = \frac{SSE(X_1, X_2, X_3) - SSE(X_1, X_2, X_3, X_4)}{MSE(X_1, X_2, X_3, X_4)} = \frac{1.0992 - 1.0815}{0.0721}$$
$$= 0.25 \tag{7.4.48}$$

Since $F_C < F\text{-}out = 4.0$, we delete X_4 from the current model. Thus the variable X_4 is deleted in step 1 and the model, which contains variables X_1, X_2, X_3, is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{7.4.49}$$

Proceed to step 2.

Step 2 The revised current model is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

The variables in the model are X_1, X_2, and X_3. We delete each of these variables in turn from the current model in (7.4.49) and calculate the corresponding *SSE* for the resulting models.

$$SSE(X_1, X_2) = 1.3884 \qquad \text{i.e., } X_3 \text{ is omitted}$$
$$SSE(X_1, X_3) = 1.5282 \qquad \text{i.e., } X_2 \text{ is omitted}$$
$$SSE(X_2, X_3) = 2.6812 \qquad \text{i.e., } X_1 \text{ is omitted}$$

Hence the best predictor variable to delete from the current model is X_3. We calculate

$$F_C = \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{MSE(X_1, X_2, X_3)} = \frac{1.3884 - 1.0992}{0.0687} = 4.21 \tag{7.4.50}$$

Because $F_C > F\text{-}out = 4$, we do not delete X_3 from the current model.

Thus in step 2 no variable is deleted, so the algorithm stops and chooses the model in (7.4.49) as the final model. A summary follows:

Step	Variable Deleted	Variables in the Model
Start	—	X_1, X_2, X_3, X_4
1	X_4	X_1, X_2, X_3

■

Remarks

1 We regard the backward elimination procedure as a purely exploratory analysis technique whether or not assumptions (A) or (B) hold because, even if these assumptions are satisfied, no statistical measure of confidence is available that can be used to judge whether the final model is an adequate one.

2 Although in the example presented the forward and the backward selection procedures lead to the same final model, this is generally not the case. We illustrate this later with an example.

3 The decision to delete or not delete a candidate variable from the current model at any particular step is sometimes based on the computed F-value of a test that compares the two competing models (where one model is nested in the other) as discussed in Section 4.9. See the discussion in part (b) on page 526. While the procedure is *guided* by a statistical test, it is not a *valid* test (it could perhaps be called a *pseudo test*), and the actual Type-I error is not known even though

we use an α value of our choice. It is quite common (but arbitrary) to carry out this pseudo test using $\alpha = 0.05$. It can be verified that the corresponding table F-values for carrying out this test are roughly in the neighborhood of 4.0 when there are enough degrees of freedom for the denominator. For this reason *F-out* is often taken to be equal to 4.0. Some authors recommend using *F-out* = 2.0. In principle, any nonnegative value can be chosen as the value for *F-out*. However, the final model chosen by this procedure may depend on the value of *F-out* that is chosen. It is clear that larger values of *F-out* will tend to result in the selection of a final model with a smaller number of variables.

4 Notice that this procedure does not examine each and every subset model, and so it is quite possible that there exist other subset models that predict Y as well as, or better than, the chosen subset or subsets. This is the price we pay for the savings in computing that results from not examining every possible model.

5 Some users of this procedure interpret the order in which the variables are deleted as an indication of the order of importance of the variables, with the first variable deleted as the least important, etc. This interpretation is incorrect and could lead to erroneous conclusions. For example, it may be the case that $\sigma_{Y|X_i}$ is the smallest among $\sigma_{Y|X_i}$, $i = 1, 2, 3, 4$ and may actually be very close to $\sigma_{Y|X_1,X_2,X_3,X_4}$, and yet X_1 could be deleted first by the backward elimination procedure.

6 Instead of selecting the candidate variable to delete from the current model by using the sum of squared errors, we could use any of the criterion measures R^2, *adj*-R^2, s, or C_p discussed in Section 7.3. This will not alter the candidate variable to be deleted.

Stepwise Regression

The stepwise regression procedure is a combination of the forward selection procedure and the backward elimination procedure. In fact, there are several variations of this procedure, all of which are referred to by the name *stepwise regression*. Here we discuss only one of the versions in detail. Again we consider the case $k = 4$ for concreteness, but the generalization to any number of predictors is obvious.

Suppose the predictor variables are X_1, X_2, X_3, and X_4. We start with an initial or current model that may be the model with no predictors.

$$\beta_0 \tag{7.4.51}$$

or the full model

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \tag{7.4.52}$$

or any other subset model. The algorithm will proceed in two stages.

1 In stage 1 the backward elimination procedure is used to delete *as many* of the variables as possible.

2 In stage 2 the forward selection procedure is used *once* to add one variable if possible.

Then the stepwise procedure continues deleting variables and adding a variable (repeating stage 1, then stage 2, etc.) until a final model is obtained.

A step is defined as a calculation (or a set of calculations) when a predictor variable is added to, or deleted from, the current model.

The algorithm proceeds as follows. Select any model to start with and call it the current model. Investigators may be able to use their subject matter knowledge to decide on an initial model. Select two constants, one called *F-in* and the other called *F-out*, to be used in making decisions during various stages of the algorithm, regarding whether or not a predictor variable should be added to the current model or deleted from the current model. A word of caution applies here. *F-out must always be less than or equal to F-in; otherwise the procedure could end up in an infinite loop of adding and deleting the same predictor variable.* You should convince yourself of this possibility.

Stage 1 Start with the current model and perform the backward elimination procedure as many times as is necessary until no more variables can be deleted. If the current model is β_0, omit this stage and go to stage 2.

Stage 2 Start with the final model of stage 1 and perform the forward selection procedure *once*. If a predictor variable is added to the current model, then go back to stage 1 with this revised current model. If no predictor is added to the current model at this stage, then the procedure terminates because no variable can be added to the current model and no variable can be removed from the current model. In this case the current model is selected as the final model.

Comments

1. It is customary to choose as the initial model the model in (7.4.51) or the full model in (7.4.52), but any model can be the initial model. The investigator may feel that certain predictor variables should be together in the starting model, and so these variables can be included in the initial model. Both *F-in* and *F-out* must be specified by the analyst. They are both set equal to 4.0 by default in many statistical packages. We say more about this in Section 7.4 of the laboratory manuals.
2. If $F\text{-}out = 0$, then no variable can be deleted at any stage. You should verify this. So setting $F\text{-}out = 0$ and *F-in* equal to any other suitable value, say 4.0, and taking β_0 as the initial model is equivalent to performing the forward selection procedure. This is the way in which you would perform the forward selection procedure in MINITAB.
3. If $F\text{-}in = \infty$ (in practice, use a very large number, say 100,000), then no variable can enter the model at any stage. You should convince yourself of this. Hence setting *F-in* equal to a very large number and *F-out* equal to any suitable value, say 4.0, and taking as the initial model the one consisting of all the predictor variables is equivalent to performing the backward elimination procedure. In fact, this is the way in which you would perform the backward elimination procedure in MINITAB.

4 Because a stepwise regression procedure is a combination of the forward selection and backward elimination procedures, it almost always leads to a better model (or at least as good a model) than either of the other two procedures. There is usually some extra computation involved, but the potential benefits outweigh the additional cost involved.

5 As a final remark, we should point out that usually none of the three methods discussed (forward selection, backward elimination, or stepwise regression) examines all possible subsets of the predictors. This is in fact the most attractive feature of these methods—they are computationally less expensive than the all-subsets regression procedure that requires an examination of all possible subsets. The price we pay for using the computationally cheaper methods, however, is that there is no guarantee that they will find the best model from all the possible subset models according to any of the criteria discussed in Section 7.3.

All of the procedures, including the all-subsets regression procedure, should be used with caution and only in an exploratory sense, i.e., to gain insight into the way different subsets of predictor variables contribute to predicting Y for the data at hand.

Authors' Recommendation

For selection of variables, we recommend that the all-subsets regressions procedure be used whenever the computing facilities are adequate rather than the forward selection, the backward elimination, or the stepwise regression procedure.

EXAMPLE 7.4.3

We illustrate the stepwise regression procedure using the GPA data of Example 4.4.2. We shall use *F-in* $= 4.0$ and *F-out* $= 3.0$ (note that we have made sure *F-out* is not bigger than *F-in*). The sums of squares needed for this example are given in Table 7.4.1. Of course, in a real problem these would not all be computed. (This is the reason we use stepwise regression rather than the all-subsets regression procedure.) In a problem with only $k = 4$ predictors, you should always use the all-subsets regression procedure, but we use stepwise regression for illustration.

We start with the initial model

$$\beta_0 \tag{7.4.53}$$

Stage 1 There are no variables to delete and so we proceed to stage 2.

Stage 2 We perform the forward selection procedure *once*. The best candidate to add to the current model is X_1 because $SSE(X_1)$ is smaller than $SSE(X_2)$, $SSE(X_3)$, and $SSE(X_4)$. To determine whether or not X_1 should be added to the current model

we compute

$$F_C = \frac{SSY - SSE(X_1)}{MSE(X_1)} = \frac{7.3458 - 2.0443}{0.1136} = 46.67 \tag{7.4.54}$$

Because $F_C > F\text{-}in = 4.0$, we add X_1 to the current model. So in step 1 X_1 is added and the model, which contains variable X_1, is

$$\beta_0 + \beta_1 x_1 \tag{7.4.55}$$

We now go back to stage 1.

Stage 1 (Repeated) The revised current model is

$$\beta_0 + \beta_1 x_1$$

We examine the possibility of removing one or more variables from the current model by applying the backward elimination procedure. The only candidate for removal is X_1 because it is the only predictor variable in the current model. To decide whether or not X_1 should be removed from the current model, we need to compare the quantity

$$F_C = \frac{SSY - SSE(X_1)}{MSE(X_1)} = \frac{7.3458 - 2.0443}{0.1136} = 46.67 \tag{7.4.56}$$

with $F\text{-}out$. Because $F_C > F\text{-}out = 3$, we conclude that X_1 cannot be deleted from the current model and the backward elimination algorithm terminates. Thus we go to stage 2 again.

Stage 2 (Repeated) We now examine the possibility of adding a variable to the current model by applying the forward selection algorithm once. The best candidate for addition is X_2 because $SSE(X_1, X_2) = 1.3884$ is smaller than $SSE(X_1, X_3)$ and $SSE(X_1, X_4)$. To decide whether or not X_2 should actually be added to the current model we compute

$$F_C = \frac{SSE(X_1) - SSE(X_1, X_2)}{MSE(X_1, X_2)} = \frac{2.0443 - 1.3884}{0.0817} = 8.03 \tag{7.4.57}$$

Because F_C is greater than $F\text{-}in = 4.0$, we add X_2 to the current model. Thus in step 2 the variable X_2 is added and the model, which contains variables X_1, X_2, is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{7.4.58}$$

Now we return to stage 1.

Stage 1 (Repeated) The revised current model is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2$$

We examine the possibility of removing one or more variables from the current model. The candidate variable for removal is X_2 because $SSE(X_1)$ is smaller than $SSE(X_2)$. To decide whether or not X_2 should actually be removed from the current

model we compute

$$F_C = \frac{SSE(X_1) - SSE(X_1, X_2)}{MSE(X_1, X_2)} = 8.03 \tag{7.4.59}$$

as in (7.4.57). Because $F_C > F\text{-}out = 3.0$, we cannot remove X_2 from the current model and the backward elimination algorithm terminates. This sends us back to stage 2.

Stage 2 (Repeated) We now apply the forward selection algorithm *once*. The predictor variables not in the current model are X_3 and X_4. Because $SSE(X_1, X_2, X_3) = 1.0992$ is smaller than $SSE(X_1, X_2, X_4) = 1.3881$, the candidate variable for addition is X_3. To decide whether or not X_3 should actually be added to the current model we compute

$$F_C = \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{MSE(X_1, X_2, X_3)} = \frac{1.3884 - 1.0992}{0.0687} = 4.21 \tag{7.4.60}$$

Because $F_C > F\text{-}in = 4.0$, we add X_3 to the current model. Thus in step 3 the variable X_3 is added and the model, which contains variables X_1, X_2, X_3, is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{7.4.61}$$

We now go back to stage 1.

Stage 1 (Repeated) The revised current model is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Again we examine the possibility of deleting one or more variables from the current model. The candidate variable for deletion is X_3 because $SSE(X_1, X_2)$ is smaller than either of $SSE(X_1, X_3)$ or $SSE(X_2, X_3)$. To decide whether or not X_3 should actually be deleted from the current model we compute

$$F_C = \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{MSE(X_1, X_2, X_3)} = 4.21 \tag{7.4.62}$$

as in (7.4.60). Because $F_C > F\text{-}out = 3.0$, we do not delete X_3 from the current model and the backward elimination algorithm terminates. This sends us back to stage 2.

Stage 2 (Repeated) Now apply the forward selection procedure *once*. The only variable not in the current model is X_4. To decide whether or not X_4 should be added to the current model we compute

$$F_C = \frac{SSE(X_1, X_2, X_3) - SSE(X_1, X_2, X_3, X_4)}{MSE(X_1, X_2, X_3, X_4)} = \frac{1.0992 - 1.0815}{0.0721} \tag{7.4.63}$$
$$= 0.25$$

Because $F_C < F\text{-}in = 4.0$, we cannot add X_4 to the current model and the forward selection algorithm terminates. Because both the backward elimination and the forward selection algorithms have terminated, the current model given in (7.4.61)

becomes the final model chosen by the stepwise regression procedure, using *F-in* = 4.0 and *F-out* = 3.0 and starting with β_0 given in (7.4.53) as the initial model.

A summary of the results of each step follows:

Step	Variable Added or Deleted	Variables in the Model
Start	—	None
1	X_1 added	X_1
2	X_2 added	X_1, X_2
3	X_3 added	X_1, X_2, X_3

■

EXAMPLE 7.4.4

We now again apply the stepwise regression procedure to the GPA data, this time using as our initial model

$$\beta_0 + \beta_3 x_3 + \beta_4 x_4 \tag{7.4.64}$$

This initial model may have been chosen based on the investigator's experience and knowledge. Let us again use *F-in* = 4 and *F-out* = 3.0.

Stage 1 The variables X_3, X_4 are in the model. The candidate for deletion is X_4 because $SSE(X_3)$ is smaller than $SSE(X_4)$. To decide whether or not X_4 should actually be deleted from the current model we compute

$$F_C = \frac{SSE(X_3) - SSE(X_3, X_4)}{MSE(X_3, X_4)} = \frac{3.8004 - 2.4180}{0.1422} = 9.72 \tag{7.4.65}$$

Because $F_C > F\text{-}out = 3.0$, we do not delete X_4 from the current model. Instead we proceed to stage 2.

Stage 2 We now apply the forward selection algorithm *once*. The variables not in the current model are X_1 and X_2. The candidate variable for adding to the current model is X_1 because $SSE(X_1, X_3, X_4)$ is smaller than $SSE(X_2, X_3, X_4)$. We compute

$$F_C = \frac{SSE(X_3, X_4) - SSE(X_1, X_3, X_4)}{MSE(X_1, X_3, X_4)} = \frac{2.4180 - 1.4532}{0.0908} = 10.63 \tag{7.4.66}$$

Because F_C is greater than *F-in* = 4.0, we add X_1 to the current model. Thus in step 1 the variable X_1 is added and the model, which contains variables X_1, X_3, X_4, is

$$\beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 \tag{7.4.67}$$

We go to stage 1 and apply the backward elimination procedure.

Stage 1 (Repeated) The revised current model is

$$\beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4$$

In this case the candidate for deletion is X_4 because $SSE(X_1, X_3)$ is smaller than $SSE(X_1, X_4)$ and $SSE(X_3, X_4)$. We compute

$$F_C = \frac{SSE(X_1, X_3) - SSE(X_1, X_3, X_4)}{MSE(X_1, X_3, X_4)} = \frac{1.5282 - 1.4532}{0.0908} = 0.83 \qquad (7.4.68)$$

Because F_C is less than $F\text{-}out = 3.0$, we delete X_4 from the model. Thus in step 2 the variable X_4 is deleted and the model, which contains X_1, X_3, is

$$\beta_0 + \beta_1 x_1 + \beta_3 x_3 \qquad (7.4.69)$$

We continue to apply the backward elimination algorithm to see if we can delete any more variables.

Stage 1 (Continued) The revised current model is

$$\beta_0 + \beta_1 x_1 + \beta_3 x_3$$

The candidate variable for deletion is X_3 because $SSE(X_1)$ is smaller than $SSE(X_3)$. We compute

$$F_C = \frac{SSE(X_1) - SSE(X_1, X_3)}{MSE(X_1, X_3)} = \frac{2.0443 - 1.5282}{0.0899} = 5.74 \qquad (7.4.70)$$

Because F_C is greater than $F\text{-}out = 3.0$, X_3 cannot be deleted from the current model. We now apply the forward selection algorithm *once*.

Stage 2 (Repeated) Variables not in the current model are X_2 and X_4. The candidate variable for addition is X_2 because $SSE(X_1, X_2, X_3)$ is smaller than $SSE(X_1, X_3, X_4)$. We compute

$$F_C = \frac{SSE(X_1, X_3) - SSE(X_1, X_2, X_3)}{MSE(X_1, X_2, X_3)} = \frac{1.5282 - 1.0992}{0.0687} = 6.24 \qquad (7.4.71)$$

Because F_C is greater than $F\text{-}in = 4.0$, X_2 is added to the model. Thus in step 3 the variable X_2 is added and the model, which contains variables X_1, X_2, X_3, is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \qquad (7.4.72)$$

We now see whether variables can be deleted by the backward elimination procedure.

Stage 1 (Repeated) The revised current model is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

The candidate variable for deletion is X_3 because $SSE(X_1, X_2)$ is smaller than $SSE(X_1, X_3)$ and $SSE(X_2, X_3)$. We compute

$$F_C = \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{MSE(X_1, X_2, X_3)} = \frac{1.3884 - 1.0992}{0.0687} = 4.21 \qquad (7.4.73)$$

Because F_C is greater than $F\text{-}out = 3.0$, X_3 cannot be deleted from the current model. We proceed to apply the forward selection algorithm *once*.

Stage 2 (Repeated) The only candidate for addition is X_4. We compute

$$F_C = \frac{SSE(X_1, X_2, X_3) - SSE(X_1, X_2, X_3, X_4)}{MSE(X_1, X_2, X_3, X_4)} = \frac{1.0992 - 1.0815}{0.0721} \quad (7.4.74)$$
$$= 0.25$$

Because F_C is smaller than $F\text{-}in = 4.0$, we cannot add X_4 to the current model. Since no more variables can be added or deleted from the current model given in (7.4.72), the stepwise regression algorithm terminates and chooses the model

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \quad (7.4.75)$$

as the final model.

A summary of the steps of the stepwise procedure for this example follows:

Step	Variable Added or Deleted	Variables in the Model
Start	—	X_3, X_4
1	X_1 added	X_1, X_3, X_4
2	X_4 deleted	X_1, X_3
3	X_2 added	X_1, X_2, X_3

Note In this particular problem the stepwise regression procedure results in the same final model whether we use β_0 or $\beta_0 + \beta_3 x_3 + \beta_4 x_4$ as the initial model. In general this will not be the case because the final model may depend on which initial model is used. Also, for the GPA data, we find that all three variable selection procedures—forward selection, backward elimination, and stepwise regression—lead to the same final model in the examples discussed if the same values of *F-in* and *F-out* are used. This is not always the case as you can see from the following example. ■

EXAMPLE 7.4.5

To illustrate the computations we use the small data set given in Table 7.4.2. Here Y is the response variable, and X_1, X_2, and X_3 are the predictor variables. The total number of observations is 10. (In real problems we do not recommend selection procedures with such a small data set.)

We will carry out a stepwise regression analysis for this data using $F\text{-}in = 3.0$ and $F\text{-}out = 3.0$. For convenience, we list all the pertinent sums of squares and mean squares in Table 7.4.3.

We start with the initial model

$$\beta_0 \quad (7.4.76)$$

Stage 1 There are no variables to delete and so we proceed to stage 2.

Stage 2 We perform the forward selection procedure *once*. The best candidate to add to the current model is X_1 because $SSE(X_1)$ is smaller than $SSE(X_2)$ and $SSE(X_3)$. To determine whether or not X_1 should be added to the current model we compute

$$F_C = \frac{SSY - SSE(X_1)}{MSE(X_1)} = \frac{8.5090 - 5.3710}{0.6714} = 4.67 \tag{7.4.77}$$

Because $F_C > F\text{-}in = 3.0$, we add X_1 to the current model. Thus in step 1 the variable X_1 is added and the model, which contains variable X_1, is

$$\beta_0 + \beta_1 x_1 \tag{7.4.78}$$

We now go back to stage 1.

TABLE 7.4.2

Observation Number	Y	X_1	X_2	X_3
1	12.5	7.0	1.7	5.7
2	11.4	6.8	2.0	5.0
3	9.7	1.7	2.1	3.8
4	11.4	3.8	2.1	4.7
5	10.7	3.8	3.3	2.7
6	12.9	3.3	4.1	3.0
7	10.6	3.3	2.6	4.3
8	10.7	3.2	2.5	3.5
9	10.5	2.2	4.0	2.4
10	11.7	5.2	2.9	4.1

TABLE 7.4.3

Sums of Squares and Mean Squares for All 8 Subset Models for the Data in Table 7.4.2

Model	Predictors in the Model	Sum of Squares	Mean Square
(1)	None	8.5090	0.9454
(2)	X_1	5.3710	0.6714
(3)	X_2	8.4160	1.0520
(4)	X_3	7.5759	0.9470
(5)	X_1, X_2	3.6315	0.5188
(6)	X_1, X_3	5.1229	0.7318
(7)	X_2, X_3	2.4325	0.3475
(8)	X_1, X_2, X_3	2.0764	0.3461

Stage 1 (Repeated) The revised current model is

$$\beta_0 + \beta_1 x_1$$

In this stage we examine the possibility of removing one or more variables from the current model by applying the backward elimination procedure. The only candidate for removal is X_1 because it is the only predictor variable in the current model. To decide whether or not X_1 should be removed from the current model we need to compare the quantity

$$F_C = \frac{SSY - SSE(X_1)}{MSE(X_1)} = \frac{8.5090 - 5.3710}{0.6714} = 4.67 \qquad \textbf{(7.4.79)}$$

with *F-out*. Since $F_C > F\text{-}out$, we conclude that X_1 cannot be deleted from the current model, and the backward elimination algorithm terminates. Thus we go to stage 2 again.

Stage 2 (Repeated) We examine the possibility of adding a variable to the current model by applying the forward selection algorithm *once*. The best candidate for addition is X_2 because $SSE(X_1, X_2) = 3.6315$ is smaller than $SSE(X_1, X_3) = 5.1229$. To decide whether or not X_2 should actually be added to the current model we compute

$$F_C = \frac{SSE(X_1) - SSE(X_1, X_2)}{MSE(X_1, X_2)} = \frac{5.3710 - 3.6315}{0.5188} = 3.35 \; . \qquad \textbf{(7.4.80)}$$

Because F_C is greater than $F\text{-}in = 3.0$, we add X_2 to the current model.

Thus in step 2 the variable X_2 is added and the model, which contains variables X_1, X_2, is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad \textbf{(7.4.81)}$$

Now we return to stage 1.

Stage 1 (Repeated) The revised current model is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2$$

We now attempt to remove one or more variables from the current model. The candidate variable for removal is X_2 because $SSE(X_1)$ is smaller than $SSE(X_2)$. To decide whether or not X_2 should actually be removed from the current model we compute

$$F_C = \frac{SSE(X_1) - SSE(X_1, X_2)}{MSE(X_1, X_2)} = \frac{5.3710 - 3.6315}{0.5188} = 3.35 \qquad \textbf{(7.4.82)}$$

as in (7.4.80). Because $F_C > F\text{-}out = 3.0$, we cannot remove X_2 from the current model, and the backward elimination algorithm terminates. This sends us back to stage 2.

Stage 2 (Repeated) We now apply the forward selection algorithm *once*. The only predictor variable not in the current model is X_3. To decide whether or not X_3 should

actually be added to the current model we compute

$$F_C = \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{MSE(X_1, X_2, X_3)} = \frac{3.6315 - 2.0764}{0.3461} = 4.49 \qquad (7.4.83)$$

Since $F_C > F\text{-}in = 3.0$, we add X_3 to the current model. Thus in step 3 the variable X_3 is added and the model, which contains variables X_1, X_2, X_3, is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \qquad (7.4.84)$$

We now go back to stage 1.

Stage 1 (Repeated) The revised current model is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Again we examine the possibility of deleting one or more variables from the current model. The candidate variable for deletion is X_1 because $SSE(X_2, X_3)$ is smaller than either $SSE(X_1, X_3)$ or $SSE(X_1, X_2)$. To decide whether or not X_1 should actually be deleted from the current model we compute

$$F_C = \frac{SSE(X_2, X_3) - SSE(X_1, X_2, X_3)}{MSE(X_1, X_2, X_3)} = \frac{2.4325 - 2.0764}{0.3461} = 1.03 \qquad (7.4.85)$$

Because $F_C < F\text{-}out = 3.0$, we delete X_1 from the current model. Thus in step 4 the variable X_1 is deleted and the model, which contains the variables X_2, X_3, is

$$\beta_0 + \beta_2 x_2 + \beta_3 x_3 \qquad (7.4.86)$$

We continue with stage 1 to see whether any more variables can be removed from the model.

Stage 1 (Continued) The revised current model is

$$\beta_0 + \beta_2 x_2 + \beta_3 x_3$$

The candidate variable for deletion is X_2 since $SSE(X_3)$ is smaller than $SSE(X_2)$. To decide whether or not X_2 should actually be deleted from the current model we compute

$$F_C = \frac{SSE(X_3) - SSE(X_2, X_3)}{MSE(X_2, X_3)} = \frac{7.5759 - 2.4325}{0.3475} = 14.80 \qquad (7.4.87)$$

Because $F_C > F\text{-}out = 3.0$, we cannot delete X_2 from the current model. This sends us back to stage 2.

Stage 2 (Repeated) The current model is

$$\beta_0 + \beta_2 x_2 + \beta_3 x_3$$

Apply the forward selection procedure *once*. The only variable not in the current model is X_1. To decide whether or not X_1 should be added to the current model we

compute

$$F_C = \frac{SSE(X_2, X_3) - SSE(X_1, X_2, X_3)}{MSE(X_1, X_2, X_3)} = \frac{2.4325 - 2.0764}{0.3461} = 1.03 \qquad (7.4.88)$$

Since $F_C < F\text{-}in = 3.0$, we cannot add X_1 to the current model, and the forward selection algorithm terminates. Because both the backward elimination and the forward selection algorithms have terminated, the current model given in (7.4.86) becomes the final model chosen by the stepwise regression procedure, using $F\text{-}in = 3.0$ and $F\text{-}out = 3.0$ and starting with (7.4.76) as the initial model.

A summary of the steps of the stepwise procedure for this example follows:

Step	Variable Added or Deleted	Variables in the Model
Start	—	None
1	X_1 added	X_1
2	X_2 added	X_1, X_2
3	X_3 added	X_1, X_2, X_3
4	X_1 deleted	X_2, X_3

You can easily verify that the forward selection method with $F\text{-}in = 3.0$ will lead to the full model as the final model, whereas the backward elimination method with $F\text{-}out = 3.0$ will pick the model in (7.4.86) as the final model. Hence the backward elimination and stepwise regression procedures agree for this problem, but the forward selection algorithm leads to a different model. ■

Caution

Note that even though the variable selection procedures discussed in this section automatically pick a best final subset model, this final model is not chosen after examining *all* the possible subset models. Typically it is chosen after examining only a fraction of all the possible subset models. Moreover, the model chosen by the algorithm very much depends on the choice of values for $F\text{-}in$ and $F\text{-}out$. Smaller values for $F\text{-}in$ (larger values for $F\text{-}out$) generally lead to final models with a larger number of predictor variables, whereas larger values of $F\text{-}in$ (smaller values of $F\text{-}out$) generally lead to final models with a smaller number of predictor variables. In practice, it is advisable to try different values for $F\text{-}in$ and $F\text{-}out$. Moreover, you should examine not only the final models from these procedures but also the models chosen at each stage of the algorithm. The bottom line is this: When conducting a variable selection analysis, you should examine as many subset models as possible—when it is not practical to examine every subset model—and construct a short list of subset models for further consideration.

Realistically, additional data need to be collected to validate the results of subset analyses. The additional data may or may not support the results of such analyses,

and it is quite possible that the chosen models will need to be revised or discarded in favor of other models. This iterative nature of scientific inquiry should come as no surprise.

When we have a large amount of data, the validity of a tentative model can be examined by applying **data splitting** methods whereby the data are divided (randomly) into two sets, with one of the sets used to carry out variable selection procedures and the other set used to assess the validity of the models in the chosen short list.

Problems 7.4

7.4.1 Consider the GPA problem in Example 4.4.2. The sums of squares and mean squares are in Table 7.4.1.

- **a** Apply the forward selection procedure with $F\text{-}in = 2.0$. What variables are in the final model?
- **b** Repeat (a) with $F\text{-}in = 5.0$.
- **c** Apply the backward elimination procedure with $F\text{-}out = 2.0$. What variables are in the final model?
- **d** Repeat (c) with $F\text{-}out = 5.0$.
- **e** Apply the stepwise regression procedure with $F\text{-}in = F\text{-}out = 2.0$. List the variables in the model at each step if the initial model is β_0.
- **f** In (e) what variables are in the final model if the initial model is $\beta_0 + \beta_1 x_1$?
- **g** In (e) what variables are in the final model if the initial model is $\beta_0 + \beta_4 x_4$?
- **h** Repeat (e) with $F\text{-}in = F\text{-}out = 4.0$.
- **i** Repeat (e) with $F\text{-}in = 2.0$ and $F\text{-}out = 4.0$. What's wrong?
- **j** Repeat (e) with $F\text{-}in = 4.0$ and $F\text{-}out = 2.0$.

7.4.2 Consider Example 7.4.5. The data are given in Table 7.4.2 and the sums of squares and mean squares are given in Table 7.4.3.

- **a** Apply the forward selection procedure with $F\text{-}in = 5.0$. What variables are in the final model?
- **b** Repeat (a) with $F\text{-}in = 2.0$. What variables are in the final model?
- **c** Apply the backward elimination procedure with $F\text{-}out = 4.0$. What variables are in the final model?
- **d** Repeat (c) with $F\text{-}out = 2.0$. What variables are in the final model?
- **e** Apply the stepwise regression procedure with $F\text{-}in = F\text{-}out = 2.0$. List the variables in the model at each step if the initial model is $\beta_0 + \beta_1 x_1$.
- **f** In (e) list the variables in the model at each step if the initial model is $\beta_0 + \beta_3 x_3$.
- **g** Repeat (e) with $F\text{-}in = F\text{-}out = 4.0$.
- **h** Repeat (e) with $F\text{-}in = 2.0$ and $F\text{-}out = 4.0$. What's wrong?
- **i** Repeat (f) with $F\text{-}in = 4.0$ and $F\text{-}out = 2.0$.

7.4.3 In Exercise 4.12.4 an investigator is interested in studying how Y, the height at age 18 years of males belonging to a group of people who have lived in mountain isolation for several generations, is related to the following variables.

X_1 = length at birth
X_2 = mother's height at age 18
X_3 = father's height at age 18
X_4 = maternal grandmother's height at age 18
X_5 = maternal grandfather's height at age 18
X_6 = paternal grandmother's height at age 18
X_7 = paternal grandfather's height at age 18

All heights and lengths are in inches. A simple random sample of 20 males of age 18 or more was drawn, and all the preceding information was recorded. The data are given in Table 7.4.4 and are also stored in the file **age18.dat** on the data disk. In Exhibit 7.4.1 is a SAS computer output (edited for ease of presentation) for all-subsets regressions and includes R^2, $adj\text{-}R^2$, C_p, s, and SSE.

TABLE 7.4.4
Heights at Age 18 of a Sample of 20 males

Observation Number	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	67.2	19.7	60.5	70.3	65.7	69.3	65.7	67.3
2	69.1	19.6	64.9	70.4	62.6	69.6	64.6	66.4
3	67.0	19.4	65.4	65.8	66.2	68.8	64.0	69.4
4	72.4	19.4	63.4	71.9	60.7	68.0	64.9	67.1
5	63.6	19.7	65.1	65.1	65.5	65.5	61.8	70.9
6	72.7	19.6	65.2	71.1	63.5	66.2	67.3	68.6
7	68.5	19.8	64.3	67.9	62.4	71.4	63.4	69.4
8	69.7	19.7	65.3	68.8	61.5	66.0	62.4	67.7
9	68.4	19.7	64.5	68.7	63.9	68.8	62.3	68.8
10	70.4	19.9	63.4	70.3	65.9	69.0	63.7	65.1
11	67.5	18.9	63.3	70.4	63.7	68.2	66.2	68.5
12	73.3	20.8	66.2	70.2	65.4	66.6	61.7	64.0
13	70.0	20.3	64.9	68.8	65.2	70.2	62.4	67.0
14	69.8	19.7	63.5	70.3	63.1	64.4	65.1	67.0
15	63.6	19.9	62.0	65.5	64.1	67.7	62.1	66.5
16	64.3	19.6	63.5	65.2	63.9	70.0	64.2	64.5
17	68.5	21.3	66.1	65.4	64.8	68.4	66.4	70.8
18	70.5	20.1	64.8	70.2	65.3	65.5	63.7	66.9
19	68.1	20.2	62.6	68.6	63.7	69.8	66.7	68.0
20	66.1	19.2	62.2	67.3	63.6	70.9	63.6	66.7

EXHIBIT 7.4.1
SAS Output for Problem 7.4.3

SAS 0:00 Saturday, Jan 1, 1994

N = 20 Regression Models for Dependent Variable: Y

In	R-square	Adj Rsq	C(p)	Root MSE	SSE	Variables in Model
1	0.626215	0.605449	38.1	1.73969	54.477	X3
1	0.159555	0.112864	105.6	2.60865	122.491	X2
1	0.067723	0.015930	118.9	2.74748	135.875	X1
1	0.063297	0.011258	119.5	2.75399	136.520	X5
1	0.053640	0.001064	120.9	2.76815	137.928	X7
1	0.045767	-.007246	122.1	2.77964	139.075	X4
1	0.037081	-.016414	123.3	2.79226	140.341	X6
2	0.850862	0.833316	7.5778	1.13075	21.736	X2 X3
2	0.793939	0.769697	15.8136	1.32914	30.032	X1 X3
2	0.635011	0.592072	38.8077	1.76894	53.195	X3 X5
2	0.628645	0.584956	39.7289	1.78430	54.123	X3 X6
2	0.627145	0.583280	39.9459	1.78790	54.342	X3 X7
2	0.626973	0.583087	39.9708	1.78831	54.367	X3 X4
2	0.274930	0.189628	90.9055	2.49323	105.676	X2 X7
2	0.242256	0.153109	95.6	2.54879	110.438	X2 X6
2	0.216180	0.123965	99.4	2.59228	114.238	X2 X4
2	0.176363	0.079465	105.2	2.65730	120.041	X2 X5
2	0.169445	0.071733	106.2	2.66844	121.050	X1 X2
2	0.169122	0.071372	106.2	2.66896	121.097	X1 X4
2	0.125925	0.023092	112.5	2.73746	127.393	X5 X7
2	0.121524	0.018174	113.1	2.74434	128.034	X1 X7
2	0.120306	0.016812	113.3	2.74624	128.212	X1 X5
2	0.116103	0.012116	113.9	2.75279	128.824	X6 X7
2	0.113808	0.009550	114.2	2.75637	129.158	X1 X6
2	0.110820	0.006211	114.6	2.76101	129.594	X4 X5
2	0.107410	0.002400	115.1	2.76630	130.091	X5 X6
2	0.099637	-.006288	116.3	2.77832	131.224	X4 X7
2	0.071884	-.037306	120.3	2.82081	135.269	X4 X6

EXHIBIT 7.4.1
(Continued)

3	0.904959	0.887139	1.7509	0.93045	13.852	X1 X2 X3
3	0.857221	0.830450	8.6578	1.14043	20.809	X2 X3 X7
3	0.856767	0.829911	8.7234	1.14224	20.876	X2 X3 X5
3	0.853147	0.825612	9.2471	1.15659	21.403	X2 X3 X6
3	0.851003	0.823067	9.5573	1.16500	21.716	X2 X3 X4
3	0.803560	0.766727	16.4216	1.33768	28.630	X1 X3 X4
3	0.797093	0.759048	17.3573	1.35952	29.573	X1 X3 X7
3	0.795603	0.757278	17.5729	1.36450	29.790	X1 X3 X5
3	0.795326	0.756950	17.6129	1.36543	29.830	X1 X3 X6
3	0.636360	0.568177	40.6126	1.82001	52.999	X3 X5 X6
3	0.635405	0.567043	40.7509	1.82240	53.138	X3 X4 X5
3	0.635290	0.566907	40.7675	1.82268	53.155	X3 X5 X7
3	0.631290	0.562156	41.3462	1.83265	53.738	X3 X6 X7
3	0.629232	0.559713	41.6439	1.83776	54.038	X3 X4 X6
3	0.628118	0.558390	41.8051	1.84052	54.200	X3 X4 X7
3	0.437918	0.332528	69.3239	2.26275	81.921	X2 X6 X7
3	0.334408	0.209610	84.3000	2.46230	97.007	X2 X4 X7
3	0.291041	0.158111	90.5746	2.54125	103.328	X2 X5 X7
3	0.281658	0.146968	91.9322	2.55802	104.695	X2 X4 X6
3	0.278528	0.143252	92.3849	2.56358	105.151	X1 X2 X7
3	0.258781	0.119803	95.2	2.59843	108.029	X2 X5 X6
3	0.252943	0.112870	96.1	2.60864	108.880	X1 X2 X4
3	0.251404	0.111042	96.3	2.61133	109.105	X1 X2 X6
3	0.232792	0.088940	99.0	2.64359	111.817	X2 X4 X5
3	0.223356	0.077735	100.4	2.65980	113.192	X1 X4 X7
3	0.219631	0.073312	100.9	2.66617	113.735	X1 X4 X5
3	0.200984	0.051168	103.6	2.69783	116.453	X5 X6 X7
3	0.199202	0.049053	103.9	2.70084	116.713	X1 X4 X6
3	0.196018	0.045271	104.3	2.70621	117.177	X1 X6 X7
3	0.187802	0.035515	105.5	2.72000	118.374	X1 X2 X5
3	0.182330	0.029017	106.3	2.72915	119.172	X1 X5 X7
3	0.173828	0.018921	107.5	2.74330	120.411	X4 X5 X7
3	0.172748	0.017638	107.7	2.74509	120.568	X1 X5 X6
3	0.147581	-.012248	111.3	2.78653	124.236	X4 X6 X7
3	0.142786	-.017941	112.0	2.79436	124.935	X4 X5 X6

EXHIBIT 7.4.1
(Continued)

4	0.910211	0.886267	2.9910	0.93404	13.086	X1 X2 X3 X5
4	0.908997	0.884729	3.1667	0.94033	13.263	X1 X2 X3 X4
4	0.906468	0.881526	3.5325	0.95330	13.632	X1 X2 X3 X7
4	0.906074	0.881027	3.5895	0.95531	13.689	X1 X2 X3 X6
4	0.865298	0.829377	9.4892	1.14404	19.632	X2 X3 X6 X7
4	0.862259	0.825529	9.9288	1.15687	20.075	X2 X3 X5 X7
4	0.858558	0.820840	10.4643	1.17231	20.615	X2 X3 X5 X6
4	0.857225	0.819151	10.6572	1.17782	20.809	X2 X3 X4 X7
4	0.857085	0.818974	10.6775	1.17840	20.829	X2 X3 X4 X5
4	0.853370	0.814269	11.2149	1.19361	21.371	X2 X3 X4 X6
4	0.805732	0.753927	18.1073	1.37389	28.314	X1 X3 X4 X7
4	0.805623	0.753789	18.1231	1.37428	28.330	X1 X3 X4 X5
4	0.805446	0.753564	18.1488	1.37490	28.355	X1 X3 X4 X6
4	0.800836	0.747726	18.8157	1.39110	29.027	X1 X3 X6 X7
4	0.798134	0.744303	19.2066	1.40050	29.421	X1 X3 X5 X7
4	0.796628	0.742396	19.4245	1.40571	29.640	X1 X3 X5 X6
4	0.637474	0.540801	42.4514	1.87681	52.837	X3 X5 X6 X7
4	0.636675	0.539788	42.5671	1.87888	52.953	X3 X4 X5 X6
4	0.635780	0.538655	42.6965	1.88119	53.083	X3 X4 X5 X7
4	0.632148	0.534054	43.2221	1.89055	53.613	X3 X4 X6 X7
4	0.472904	0.332345	66.2620	2.26306	76.822	X2 X4 X6 X7
4	0.453379	0.307613	69.0870	2.30460	79.668	X2 X5 X6 X7
4	0.439511	0.290047	71.0935	2.33365	81.689	X1 X2 X6 X7
4	0.357523	0.186196	82.9557	2.49851	93.638	X1 X2 X4 X7
4	0.350314	0.177064	83.9988	2.51249	94.689	X2 X4 X5 X7
4	0.312278	0.128886	89.5019	2.58499	100.232	X1 X2 X4 X6
4	0.298047	0.110860	91.5608	2.61160	102.306	X2 X4 X5 X6
4	0.295590	0.107747	91.9164	2.61616	102.665	X1 X2 X5 X7
4	0.281952	0.090473	93.8895	2.64137	104.652	X1 X4 X5 X7
4	0.276871	0.084037	94.6	2.65069	105.393	X1 X4 X6 X7
4	0.272666	0.078711	95.2	2.65839	106.006	X1 X2 X4 X5
4	0.269408	0.074584	95.7	2.66434	106.480	X1 X2 X5 X6
4	0.268330	0.073218	95.9	2.66630	106.638	X1 X5 X6 X7
4	0.254977	0.056305	97.8	2.69052	108.584	X1 X4 X5 X6
4	0.232757	0.028159	101.0	2.73035	111.822	X4 X5 X6 X7

EXHIBIT 7.4.1
(Continued)

5	0.913517	0.882630	4.5126	0.94885	12.604	X1 X2 X3 X4 X5
5	0.911339	0.879674	4.8278	0.96073	12.922	X1 X2 X3 X5 X7
5	0.911009	0.879227	4.8755	0.96251	12.970	X1 X2 X3 X5 X6
5	0.910902	0.879082	4.8909	0.96309	12.986	X1 X2 X3 X4 X7
5	0.909781	0.877559	5.0532	0.96913	13.149	X1 X2 X3 X4 X6
5	0.909639	0.877368	5.0737	0.96989	13.170	X1 X2 X3 X6 X7
5	0.869013	0.822232	10.9516	1.16774	19.091	X2 X3 X5 X6 X7
5	0.865307	0.817202	11.4879	1.18415	19.631	X2 X3 X4 X6 X7
5	0.862326	0.813157	11.9191	1.19718	20.065	X2 X3 X4 X5 X7
5	0.858969	0.808601	12.4048	1.21169	20.555	X2 X3 X4 X5 X6
5	0.809813	0.741889	19.5169	1.40710	27.719	X1 X3 X4 X6 X7
5	0.807197	0.738339	19.8954	1.41674	28.100	X1 X3 X4 X5 X7
5	0.807047	0.738135	19.9171	1.41729	28.122	X1 X3 X4 X5 X6
5	0.801236	0.730249	20.7578	1.43847	28.969	X1 X3 X5 X6 X7
5	0.637946	0.508641	44.3832	1.94142	52.768	X3 X4 X5 X6 X7
5	0.488252	0.305485	66.0414	2.30814	74.585	X2 X4 X5 X6 X7
5	0.485651	0.301955	66.4177	2.31400	74.964	X1 X2 X4 X6 X7
5	0.455610	0.261185	70.7642	2.38061	79.342	X1 X2 X5 X6 X7
5	0.375894	0.152999	82.2978	2.54896	90.961	X1 X2 X4 X5 X7
5	0.345401	0.111615	86.7096	2.61049	95.405	X1 X4 X5 X6 X7
5	0.331501	0.092752	88.7206	2.63805	97.431	X1 X2 X4 X5 X6
6	0.914988	0.875751	6.2999	0.97626	12.390	X1 X2 X3 X4 X5 X7
6	0.914078	0.874422	6.4314	0.98147	12.523	X1 X2 X3 X4 X5 X6
6	0.913690	0.873855	6.4876	0.98368	12.579	X1 X2 X3 X5 X6 X7
6	0.913647	0.873792	6.4938	0.98393	12.586	X1 X2 X3 X4 X6 X7
6	0.869082	0.808658	12.9417	1.21151	19.081	X2 X3 X4 X5 X6 X7
6	0.810470	0.722995	21.4218	1.45769	27.623	X1 X3 X4 X5 X6 X7
6	0.502828	0.273364	65.9325	2.36091	72.461	X1 X2 X4 X5 X6 X7
7	0.917060	0.868679	8.0000	1.00366	12.088	X1 X2 X3 X4 X5 X6 X7

a Find the predictor variable that results in the best prediction function (for the sample data at hand) among all subset models containing only one predictor.

b Find the two predictor variables that result in the best prediction function (for the sample data at hand) among all subset models containing exactly two variables.

c Find the three predictor variables that result in the best prediction function (for the sample data at hand) among all subset models containing exactly three variables.

d Find a short list of the three best models using the criterion R^2 in Section 7.3.

e Find a short list of the three best models using the criterion *adj-*R^2 in Section 7.3.

f Find a short list of the three best models using the criterion s in Section 7.3.

g Find a short list of the three best models using the criterion C_p in Section 7.3.

h Find a short list of four best models using C_p and $C_p - p$ together.

i Write a short report giving your findings using parts (a) through (h).

7.4.4 a In Problem 7.4.3 apply the forward selection procedure with *F-in* = 5.0. What variables are in the final model?

b Repeat (a) with *F-in* = 4.0. What variables are in the final model?

c Apply the backward elimination procedure with *F-out* = 5.0. What variables are in the final model?

d Repeat (c) with *F-out* = 4.0. What variables are in the final model?

e Apply the stepwise regression procedure with *F-in* = *F-out* = 4.0. List the variables that are in the model in each step if the initial model is $\beta_0 + \beta_1 x_1$.

f In (e) list the variables that are in the model at each step if the initial model is β_0.

g Repeat (e) with *F-in* = *F-out* = 3.0.

h Repeat (e) with *F-in* = 4.0 and *F-out* = 2.0.

7.5 Growth Curves

In this section we discuss an important class of statistical models known as **growth curve models** or **longitudinal models**. These models may be viewed as modifications of the multiple regression model. We introduce this topic with some examples.

EXAMPLE 7.5.1

Consider the population of all premature babies (let M represent the number of such babies) born in the state of New York over the past 5 years. An investigator is interested in studying how premature babies grow (how their weight Y changes) as a function of time t from day 5 to day 50 after birth. Let us consider one baby in this population, whom we denote as baby I. The observed weight of this baby at time t will be denoted by $Y_I(t)$, and its 'true' weight at time t will be denoted by $\mu_{Y_I}(t)$, or simply $\mu_I(t)$ for ease of notation. Suppose that the 'true' weight of baby I as a function of time t is given by

$$\mu_I(t) = \alpha_I + \beta_I t \tag{7.5.1}$$

and that $Y_I(t)$, the observed weight at time t, is related to the 'true' weight according to the model

$$Y_I(t) = \mu_I(t) + E_{I,t} = \alpha_I + \beta_I t + E_{I,t} \tag{7.5.2}$$

The quantity $E_{I,t}$ represents the difference between the observed weight and the 'true' weight at time t for baby I.

In many applications it is reasonable to regard $E_{I,t}$ as a randomly chosen number from a Gaussian population with mean zero and standard deviation σ_E. In such a case, the function $\mu_I(t)$ is the regression function of the weight of baby I on the predictor variable t, and it is called the *growth curve* of baby I. The reason that $Y_I(t)$, the observed weight of baby I at time t, does not equal $\mu_I(t)$, the 'true' weight of baby I at time t, is that the 'true' weight may not be observed exactly but is observed with error; this error includes measurement error as well as other uncontrollable random fluctuations. In practice the 'true' weight of baby I as a function of time t may not be of the form $\mu_I(t) = \alpha_I + \beta_I t$ *exactly*, but such a function may be an adequate approximation to the growth curve for baby I.

Clearly, there is a growth curve for each of the M babies in the population, so there are M growth curves

$$\mu_I(t) = \alpha_I + \beta_I t \qquad \text{for } I = 1, \ldots, M \tag{7.5.3}$$

The investigator may be interested in the growth curve of some of the individual babies, or in the average of the growth curves of all M babies, or both. The average of the growth curves of all M babies is denoted by $\mu_Y(t)$ and is called the **population growth curve**. It is given by

$$\mu_Y(t) = \frac{1}{M}\sum_{I=1}^{M} \mu_I(t) = \alpha + \beta t \tag{7.5.4}$$

where

$$\alpha = \frac{1}{M}\sum_{i=1}^{M} \alpha_I \qquad \text{and} \qquad \beta = \frac{1}{M}\sum_{i=1}^{M} \beta_I \tag{7.5.5}$$

Note To be consistent with our notation for regression functions thus far, we should write Equation (7.5.1) using β_0, β_1, etc. However, we need one set of β coefficients for each item (baby) in the population, thus making it necessary to use double subscripts. Because this would lead to unnecessarily complicated notation, we have used α_I and β_I.

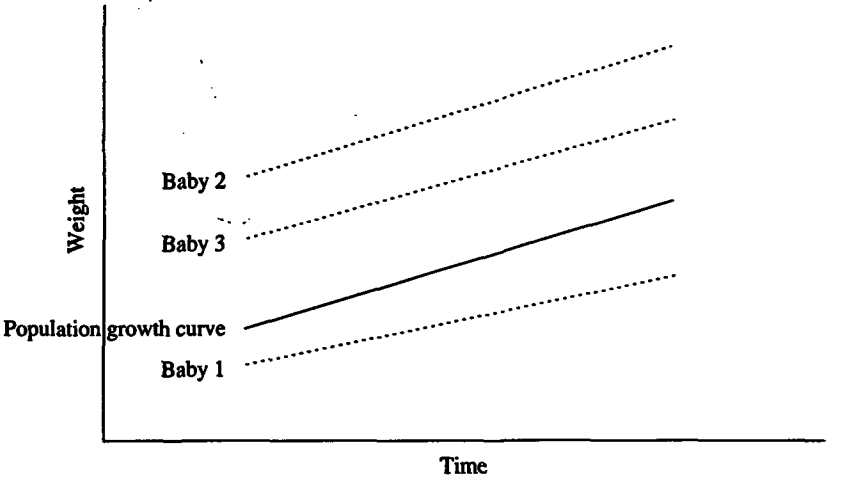
To estimate the population intercept and slope, α and β in (7.5.5), we select a simple random sample of m babies from the population of M babies and measure the weight of each baby at k preselected times $t_1, \ldots, t_k$. From these data we calculate point estimates and confidence intervals for α_i and β_i, the coefficients in the growth curve for the ith baby in the sample, for $i = 1, \ldots, m$, and from these we obtain point estimates and confidence intervals for the population regression coefficients α and β.

To summarize, there is a population of M growth curves (corresponding to the M babies), and a simple random sample of m of them is selected (they correspond to the m babies in the sample). Values of each of these m regression functions are measured at the same k preselected times $t_1, \ldots, t_k$. From these data we compute point estimates and confidence intervals for parameters in each of the m growth

curves (i.e., for each of the m babies in the sample) and for the parameters in the population growth curve in (7.5.4).

To graphically illustrate the situation, we have plotted, in Figure 7.5.1, the growth curves, $\mu_i(t)$, $i = 1, 2, 3$, for three babies in the population, and also the population growth curve $\mu_Y(t)$, i.e., the average of all M growth curves. ■

FIGURE 7.5.1



EXAMPLE 7.5.2

Suppose an investigator is interested in the change in blood pressure (Y) of individuals, in a population of M individuals, as a function of the number of minutes (t) after they are placed in a certain stressful situation, such as being subjected to very loud noise. In this example the growth curves are the regression functions relating blood pressures of the individuals in the population to elapsed time since the introduction of stress. Suppose that the ith individual's true blood pressure as a function of time t is given by

$$\mu_I(t) = \alpha_I + \beta_I t + \gamma_I t^2 \qquad 0 \le t \le 30$$

Each individual in the population has a growth curve that is a quadratic function of t but perhaps with different coefficients α_I, β_I, and γ_I. The average of all M growth curves is called the population growth curve and it is given by

$$\mu_Y(t) = \alpha + \beta t + \gamma t^2$$

where

$$\alpha = \frac{1}{M}\sum_{I=1}^{M}\alpha_I, \quad \beta = \frac{1}{M}\sum_{I=1}^{M}\beta_I, \text{ and } \gamma = \frac{1}{M}\sum_{I=1}^{M}\gamma_I \quad ■$$

In Examples 7.5.1 and 7.5.2 the predictor variable is time. The response variable Y is measured on each subject in the sample at several preselected times. However, the distinguishing feature of growth curve models is not that the predictor variable is time but that the Y values are measured or observed on each of the m sample subjects more than once, usually under different conditions.

This type of model is sometimes referred to as a *repeated measurements* model. The different conditions are different times in Examples 7.5.1 and 7.5.2. This is not the case in the following example, which involves a growth curve model but time is not the predictor variable.

EXAMPLE 7.5.3

A company evaluates Y, the number of miles per gallon (mpg) that different makes and models of new cars get when X milliliters/gallon (ml/g for short) of a gasoline additive is used. The company wants to make statements such as, "On the average this make of automobile gets $\mu_Y(x)$ miles per gallon when x ml/g of the additive is used." Suppose that the mpg for the ith car in the population, when x ml/g of the additive is used, is given by the regression function $\mu_{Y_i}(x)$, or $\mu_I(x)$ for short, where

$$\mu_I(x) = \alpha_I + \beta_I x \qquad 50 \le x \le 90$$

The population growth curve is the average of M regression functions of all M cars in the population, and it is given by

$$\mu_Y(x) = \alpha + \beta x$$

where

$$\alpha = \frac{1}{M}\sum_{I=1}^{M}\alpha_I \qquad \text{and} \qquad \beta = \frac{1}{M}\sum_{I=1}^{M}\beta_I$$

To obtain point estimates and confidence intervals for α and β, the company selects a simple random sample of m automobiles of this make and model. The relationship of Y and X for the ith car sampled is given by the regression function

$$\mu_i(x) = \alpha_i + \beta_i x$$

The number of miles per gallon of the ith car in the sample is measured by driving it a specified distance on k different occasions with k preselected amounts $x_1, \ldots, x_k$ of the additive. From these data, estimates of α_i and β_i are computed for each i (each car in the sample). Using these we then compute point and confidence interval estimates for the population parameters α, β, and $\mu_Y(x)$.

Observe that the number of miles per gallon is measured on each car in the sample under k different conditions—viz., with k different amounts of the additive—and so a growth curve model is appropriate here. However, note that the predictor variable here is the amount of the additive used, not time. ■

EXAMPLE 7.5.4

An agronomist is interested in studying the average height of a new variety of wheat as a function of time t after planting. It is assumed that the relationship between height Y and time t in days after planting, where $20 \leq t \leq 60$ for plant I in the population of M plants, is given by

$$\mu_I(t) = \alpha_I + \beta_I t + \gamma_I t^2$$

The corresponding population growth curve is given by

$$\mu_Y(t) = \alpha + \beta t + \gamma t^2$$

where

$$\alpha = \frac{1}{M}\sum_{I=1}^{M}\alpha_I \qquad \beta = \frac{1}{M}\sum_{I=1}^{M}\beta_I \qquad \gamma = \frac{1}{M}\sum_{I=1}^{M}\gamma_I$$

To obtain point estimates and confidence intervals for the quantities α, β, γ, and $\mu_Y(t)$, the agronomist selects a simple random sample of m wheat plants and measures the height of the ith plant in the sample, $i = 1, \ldots, m$, at k specified times $t_1, \ldots, t_k$. From these data, estimates of the growth curves are obtained for each of the m plants in the sample. These in turn lead to point and confidence interval estimates of the population growth curve. ▪

Models of the type discussed in Examples 7.5.1–7.5.4 are called **random coefficient growth curve models**. In this section we give formulas for estimating $\mu_i(t)$, the **individual growth curves** for each of the m individuals in the sample, and also for estimating the population growth curve $\mu_Y(t)$. Keep in mind that we will use the symbol t for the predictor variable and refer to it as *time* for simplicity of explanations, and also because it is in this context that growth curve models are commonly used. However, as pointed out in Example 7.5.3, predictor variables other than time may also make sense in certain situations. We now state the assumptions underlying random coefficient growth curve models.

Assumptions for the Growth Curve Model

Notation There is a population consisting of M items (in the preceding examples the items are babies, cars, wheat plants, or individuals). For each item in the population there is a regression function of the response variable Y on the predictor variable t, and this is the growth curve for that item. The regression function for the ith item in the population is denoted by $\mu_{Y_I}(t)$, or $\mu_I(t)$ for short. The average of the M individual regression functions in the population is denoted by $\mu_Y(t)$, and it is given by

$$\mu_Y(t) = \frac{1}{M}\sum_{I=1}^{M}\mu_I(t)$$

The function $\mu_Y(t)$ is called the *population growth curve* or the *population growth function*. In the mathematical theory of growth curves, M is infinite.

Any linear regression function is a candidate for modeling growth curves of population items. Some examples with the number of parameters $p = 2, 3$, and 4 are

$$\mu_I(t) = \alpha_I + \beta_I t \quad (p = 2)$$
$$\mu_I(t) = \alpha_I + \beta_I t + \gamma_I t^2 \quad (p = 3)$$
$$\mu_I(t) = \alpha_I + \beta_I t + \gamma_I t^2 + \delta_I t^3 \quad (p = 4)$$
$$\mu_I(t) = \alpha_I + \beta_I t + \gamma_I \log(t) \quad (p = 3)$$
$$\mu_I(t) = \alpha_I + \beta_I e^t \quad (p = 2)$$

BOX 7.5.1 Assumptions for Growth Curve Models

(Population) Assumption 1 The regression function for the ith population item is a *linear* regression function with p unknown parameters. We choose a second-degree polynomial model (quadratic model) for illustration; i.e., for population item I the growth curve is

$$\mu_I(t) = \alpha_I + \beta_I t + \gamma_I t^2 \tag{7.5.6}$$

Note that $\mu_I(t)$ has the same form for each item in the population, but with perhaps different α_I, β_I, and γ_I values, because each item grows at its own rate. The population growth curve at time t is

$$\mu_Y(t) = \alpha + \beta t + \gamma t^2 \tag{7.5.7}$$

where $\alpha = \frac{1}{M}\sum_{I=1}^{M} \alpha_I$, $\beta = \frac{1}{M}\sum_{I=1}^{M} \beta_I$, and $\gamma = \frac{1}{M}\sum_{I=1}^{M} \gamma_I$.

(Population) Assumption 2 The population of regression coefficients $\{(\alpha_I, \beta_I, \gamma_I), I = 1, \ldots, M\}$ is (approximately) a Gaussian population.

(Sample) Assumption 3 A simple random sample of m items is selected from the population of M items.

(Sample) Assumption 4 For each of the m chosen items, the value of Y is measured at k preselected times $t_1, \ldots, t_k$. The recorded Y value for sample item i at time t_j is denoted by $y_{i,j}$. A schematic representation of the sample data is given in Table 7.5.1.

The observed responses $y_{i,j}$ $(j = 1, \ldots, k)$, for sample item i, are given by

$$y_{i,j} = \mu_i(t_j) + e_{i,j}$$

i.e.,

$$y_{i,j} = \alpha_i + \beta_i t_j + \gamma_i t_j^2 + e_{i,j} \tag{7.5.8}$$

where $\mu_i(t_j) = \alpha_i + \beta_i t_j + \gamma_i t_j^2$ is the 'true' response of sample item i at time t_j, and $e_{i,j}$ is an error that includes measurement error and other random errors due to unknown and uncontrolled disturbances.

(Sample) Assumption 5 $\{e_{i,j}, i = 1, \ldots, m; j = 1, \ldots, k\}$ is a simple random sample from a Gaussian population with zero mean and standard deviation σ_E. As usual the $\{e_{i,j}\}$ are not observable.

(Sample) Assumption 6 The quantities $t_1, \ldots, t_k$ are measured without error.

Even though we are using a quadratic model for explaining the methods for analyzing growth curves, the methods presented in this section are easily adapted for application to any linear regression function described in Chapters 3 and 4.

Inferences for Growth Curve Models

One of the main objectives in a growth curve study is to make inferences about the population growth curve. This includes inferences about the unknown parameters α, β, γ, etc. in the population growth curve, as well as certain functions of them. We begin our discussion by examining the data for each item in the sample (see Table 7.5.1). As stated, we work with a quadratic growth curve model although the procedures given can be adapted very easily for any linear growth curve model (linear in the unknown parameters).

TABLE 7.5.1

A Schematic Representation of the Sample Data for a Growth Curve Study

Item	Response at Time t_1	⋯ ⋯	Response at Time t_j	⋯ ⋯	Response at Time t_k
1	$y_{1,1}$	⋯	$y_{1,j}$	⋯	$y_{1,k}$
2	$y_{2,1}$	⋯	$y_{2,j}$	⋯	$y_{2,k}$
⋮	⋮	⋮	⋮	⋮	⋮
i	$y_{i,1}$	⋯	$y_{i,j}$	⋯	$y_{i,k}$
⋮	⋮	⋮	⋮	⋮	⋮
m	$y_{m,1}$	⋯	$y_{m,j}$	⋯	$y_{m,k}$

The observed value $y_{1,j}$ for sample item 1 (say baby 1 in Example 7.5.1) at time t_j is given by

$$y_{1,j} = \alpha_1 + \beta_1 t_j + \gamma_1 t_j^2 + e_{1,j} \qquad j = 1, \ldots, k \tag{7.5.9}$$

where $e_{1,j}$ are random errors from a Gaussian population with zero mean and standard deviation σ_E. This is a multiple linear regression model with $X_1 = t$ and $X_2 = t^2$, so using the methods in Chapter 4 we can estimate α_1, β_1, and γ_1. The

data for sample item 1 are

$$\begin{array}{ccc} y_1 & t & t^2 \\ \hline y_{1,1} & t_1 & t_1^2 \\ y_{1,2} & t_2 & t_2^2 \\ \vdots & \vdots & \vdots \\ y_{1,k} & t_k & t_k^2 \end{array} \tag{7.5.10}$$

and the model equations in (7.5.9) for the first sample item can be written in matrix notation as

$$\boldsymbol{y}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{e}_1 \tag{7.5.11}$$

where

$$\boldsymbol{y}_1 = \begin{bmatrix} y_{1,1} \\ y_{1,2} \\ \vdots \\ y_{1,k} \end{bmatrix} \quad \boldsymbol{X}_1 = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ \vdots & \vdots & \vdots \\ 1 & t_k & t_k^2 \end{bmatrix} \quad \boldsymbol{\beta}_1 = \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \gamma_1 \end{bmatrix} \quad \boldsymbol{e}_1 = \begin{bmatrix} e_{1,1} \\ e_{1,2} \\ \vdots \\ e_{1,k} \end{bmatrix} \tag{7.5.12}$$

The point estimate of $\boldsymbol{\beta}_1$, using (4.4.8), is given by

$$\hat{\boldsymbol{\beta}}_1 = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\beta}_1 \\ \hat{\gamma}_1 \end{bmatrix} = (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{y}_1 \tag{7.5.13}$$

Hence the estimated growth curve for sample item 1 is given by

$$\hat{\mu}_1(t) = \hat{\alpha}_1 + \hat{\beta}_1 t + \hat{\gamma}_1 t^2 \tag{7.5.14}$$

The observed value $y_{2,j}$ for sample item 2 at time t_j is given by

$$y_{2,j} = \alpha_2 + \beta_2 t_j + \gamma_2 t_j^2 + e_{2,j} \qquad j = 1, \ldots, k \tag{7.5.15}$$

where $e_{2,j}$ are random errors from a Gaussian population with zero mean and standard deviation σ_E. This is a multiple linear regression model, so using the methods of Chapter 4 we can estimate α_2, β_2, and γ_2. The data for sample item 2 are

$$\begin{array}{ccc} y_2 & t & t^2 \\ \hline y_{2,1} & t_1 & t_1^2 \\ y_{2,2} & t_2 & t_2^2 \\ \vdots & \vdots & \vdots \\ y_{2,k} & t_k & t_k^2 \end{array} \tag{7.5.16}$$

and the model equations in (7.5.15) for the 2nd sample item can be written in matrix notation as

$$\boldsymbol{y}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}_2 \tag{7.5.17}$$

where

$$y_2 = \begin{bmatrix} y_{2,1} \\ y_{2,2} \\ \vdots \\ y_{2,k} \end{bmatrix} \quad X_2 = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ \vdots & \vdots & \vdots \\ 1 & t_k & t_k^2 \end{bmatrix} \quad \beta_2 = \begin{bmatrix} \alpha_2 \\ \beta_2 \\ \gamma_2 \end{bmatrix} \quad e_2 = \begin{bmatrix} e_{2,1} \\ e_{2,2} \\ \vdots \\ e_{2,k} \end{bmatrix} \tag{7.5.18}$$

The point estimate of β_2, using (4.4.8), is given by

$$\hat{\beta}_2 = \begin{bmatrix} \hat{\alpha}_2 \\ \hat{\beta}_2 \\ \hat{\gamma}_2 \end{bmatrix} = (X_2^T X_2)^{-1} X_2^T y_2 \tag{7.5.19}$$

Hence the estimated growth curve for sample item 2 is given by

$$\hat{\mu}_2(t) = \hat{\alpha}_2 + \hat{\beta}_2 t + \hat{\gamma}_2 t^2 \tag{7.5.20}$$

We proceed along similar lines for each of the m sample items. The observed value $y_{m,j}$ for sample item m at time t_j is given by

$$y_{m,j} = \alpha_m + \beta_m t_j + \gamma_m t_j^2 + e_{m,j} \qquad j = 1, \ldots, k \tag{7.5.21}$$

The data for sample item m are

$$\begin{array}{ccc} y_m & t & t^2 \\ \hline y_{m,1} & t_1 & t_1^2 \\ y_{m,2} & t_2 & t_2^2 \\ \vdots & \vdots & \vdots \\ y_{m,k} & t_k & t_k^2 \end{array} \tag{7.5.22}$$

and the model equations in (7.5.21) for sample item m can be written in matrix notation as

$$y_m = X_m \beta_m + e_m \tag{7.5.23}$$

where

$$y_m = \begin{bmatrix} y_{m,1} \\ y_{m,2} \\ \vdots \\ y_{m,k} \end{bmatrix} \quad X_m = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ \vdots & \vdots & \vdots \\ 1 & t_k & t_k^2 \end{bmatrix} \quad \beta_m = \begin{bmatrix} \alpha_m \\ \beta_m \\ \gamma_m \end{bmatrix} \quad e_m = \begin{bmatrix} e_{m,1} \\ e_{m,2} \\ \vdots \\ e_{m,k} \end{bmatrix} \tag{7.5.24}$$

The point estimate of β_m, using (4.4.8), is given by

$$\hat{\beta}_m = \begin{bmatrix} \hat{\alpha}_m \\ \hat{\beta}_m \\ \hat{\gamma}_m \end{bmatrix} = (X_m^T X_m)^{-1} X_m^T y_m \tag{7.5.25}$$

Hence the estimated growth curve for sample item m is given by

$$\hat{\mu}_m(t) = \hat{\alpha}_m + \hat{\beta}_m t + \hat{\gamma}_m t^2 \tag{7.5.26}$$

Note Because of the way the sample observations are obtained, the responses for each item in the sample are observed at the same k times $t_1, \ldots, t_k$. Hence $X_1 = X_2 = \cdots = X_m$. We denote this common matrix by X. Thus

$$X = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ \vdots & \vdots & \vdots \\ 1 & t_k & t_k^2 \end{bmatrix} \quad \text{a } k \times 3 \text{ matrix} \tag{7.5.27}$$

Point Estimates for the Population Growth Curve

The parameters determining the population growth curve are in the vector $\boldsymbol{\beta}$ where $\boldsymbol{\beta} = [\alpha, \beta, \gamma]^T$ and α, β, and γ are as in (7.5.7). Their estimates are

$$\hat{\boldsymbol{\beta}}^T = [\hat{\alpha}, \hat{\beta}, \hat{\gamma}], \hat{\alpha} = \frac{1}{m}\sum_{i=1}^{m}\hat{\alpha}_i, \hat{\beta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\beta}_i, \text{ and } \hat{\gamma} = \frac{1}{m}\sum_{i=1}^{m}\hat{\gamma}_i \tag{7.5.28}$$

with $\hat{\alpha}_i$, $\hat{\beta}_i$, and $\hat{\gamma}_i$ being the estimated values of the coefficients in the growth curve for sample item i. Hence the estimated population growth curve is

$$\hat{\mu}_Y(t) = \hat{\alpha} + \hat{\beta}t + \hat{\gamma}t^2 \tag{7.5.29}$$

Confidence intervals

Many questions about the population growth curve can be answered using point estimates and confidence intervals for θ where

$$\theta = \boldsymbol{a}^T\boldsymbol{\beta} = a_0\alpha + a_1\beta + a_2\gamma \tag{7.5.30}$$

by appropriately specifying the elements of the vector

$$\boldsymbol{a}^T = [a_0, a_1, a_2] \tag{7.5.31}$$

A $1 - \alpha$ confidence interval for $\theta = \boldsymbol{a}^T\boldsymbol{\beta}$ for a specified vector $\boldsymbol{a}$ is given by

$$\boldsymbol{a}^T\hat{\boldsymbol{\beta}} - t_{1-\alpha/2:m-1}SE(\boldsymbol{a}^T\hat{\boldsymbol{\beta}}) \le \boldsymbol{a}^T\boldsymbol{\beta} \le \boldsymbol{a}^T\hat{\boldsymbol{\beta}} + t_{1-\alpha/2:m-1}SE(\boldsymbol{a}^T\hat{\boldsymbol{\beta}}) \tag{7.5.32}$$

with

$$SE(\boldsymbol{a}^T\hat{\boldsymbol{\beta}}) = \sqrt{\frac{\boldsymbol{a}^T\boldsymbol{G}\boldsymbol{a}}{m}} \tag{7.5.33}$$

and

$$\boldsymbol{G} = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}})^T \tag{7.5.34}$$

where

$$\hat{\boldsymbol{\beta}}_i = [\hat{\alpha}_i, \hat{\beta}_i, \hat{\gamma}_i]^T \tag{7.5.35}$$

for $i = 1, 2, \ldots, m$ and

$$\hat{\boldsymbol{\beta}} = [\hat{\alpha}, \hat{\beta}, \hat{\gamma}]^T \tag{7.5.36}$$

In particular, the formula in (7.5.32) can be used for obtaining confidence intervals for the following quantities:

a for α, by choosing $\boldsymbol{a}^T = [1, 0, 0]$

b for β, by choosing $\boldsymbol{a}^T = [0, 1, 0]$

c for γ, by choosing $\boldsymbol{a}^T = [0, 0, 1]$

d for $\mu_Y(t)$, by choosing $\boldsymbol{a}^T = [1, t, t^2]$

e for $\mu_Y(t_1) - \mu_Y(t_2)$, by choosing $\boldsymbol{a}^T = [0, t_1 - t_2, t_1^2 - t_2^2]$

We illustrate the use of the preceding formulas in Example 7.5.5.

EXAMPLE 7.5.5

A scientist working for a pharmaceutical company is interested in the concentrations of a certain drug at various times after it is injected into the bloodstream. In particular, she wants to determine the amount of time it takes for the body to eliminate the drug so that the drug concentrations reach near-zero levels in the circulatory system. To understand this, she selects a simple random sample of 24 human subjects from the study population. A fixed dose of the drug is injected into the bloodstream of each subject, and blood samples are collected at the end of 1 hour, 2 hours, 3 hours, and 4 hours after injection. The blood samples are analyzed in a laboratory, and the drug concentrations are determined in milligrams per liter (mg/l for short). The data are presented in Table 7.5.2 and they are also stored in the file **drugconc.dat** on the data disk.

Assume that in the range from 1 to 4 hours after injection, the growth curve for each subject in the population is a quadratic function, so for subject i in the sample the growth curve is given by

$$\mu_i(t) = \alpha_i + \beta_i t + \gamma_i t^2$$

We regress Y (concentration) on t (time) and t^2 and obtain estimates $\hat{\alpha}_i$, $\hat{\beta}_i$, and $\hat{\gamma}_i$ for $i = 1, \ldots, 24$ (i.e., for each subject). One of the objectives in this study is to estimate the population growth curve, $\mu_Y(t) = \alpha + \beta t + \gamma t^2$, and obtain a 95% confidence interval for $\mu_Y(3) = \alpha + 3\beta + 9\gamma$, the (population) average concentration of the drug at the end of 3 hours after injection.

We assume that the 24 subjects ($m = 24$) constitute a simple random sample from a well-defined population of subjects of interest (e.g., population of persons for whom the symptoms might suggest the use of this particular drug). We further suppose that assumptions in Box 7.5.1 for the random coefficient growth model are satisfied. We have $k = 4$, $p = 3$, $m = 24$, and the population growth curve is

$$\mu_Y(t) = \alpha + \beta t + \gamma t^2 \qquad 1 \le t \le 4 \tag{7.5.37}$$

TABLE 7.5.2
Drug Concentration Data (in milligrams/liter)

Subject	*t*			
	1 Hour	2 Hours	3 Hours	4 Hours
1	10.55	4.11	2.00	1.02
2	10.47	4.30	2.15	1.11
3	9.46	3.81	1.78	0.94
4	9.27	3.72	1.92	0.95
5	9.37	3.75	1.95	0.97
6	9.67	4.28	1.96	1.04
7	10.58	3.95	2.30	1.08
8	9.96	3.73	1.86	1.01
9	9.84	3.92	2.00	1.05
10	10.20	4.20	1.96	1.03
11	9.45	4.18	2.18	1.02
12	9.64	4.04	2.08	0.96
13	10.03	4.01	2.08	1.04
14	9.81	3.65	1.97	0.97
15	10.74	4.41	2.07	1.03
16	10.08	3.80	1.86	0.99
17	10.00	3.84	2.07	0.95
18	9.73	3.94	1.93	0.96
19	9.64	4.24	2.11	1.06
20	10.40	4.11	2.07	1.01
21	10.34	4.20	2.21	1.14
22	10.09	4.35	1.91	1.07
23	9.51	3.74	1.87	0.99
24	9.63	3.77	1.96	1.01

Note that the X matrix is given by

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{bmatrix} \tag{7.5.38}$$

Also

$$y_1^T = [10.55, 4.11, 2.00, 1.02]$$

which is the first row of Table 7.5.2. The data for subject 1 are in this first row, and

$$y_2^T = [10.47, 4.30, 2.15, 1.11]$$

which is the second row of Table 7.5.2. The data for subject 2 are in this second row, etc. Thus we perform the regressions of y on t and t^2 for each sample subject and get $\hat{\alpha}_i$, $\hat{\beta}_i$, and $\hat{\gamma}_i$ for $i = 1, \ldots, 24$. From these, we compute the quantities $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{G}$ using (7.5.28) and (7.5.34), respectively. The results are

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 17.6035 \\ -9.0499 \\ 1.2350 \end{bmatrix} \qquad \boldsymbol{G} = \begin{bmatrix} 0.695830 & -0.436192 & 0.067454 \\ -0.436192 & 0.298748 & -0.048488 \\ 0.067454 & -0.048488 & 0.008067 \end{bmatrix} \tag{7.5.39}$$

Thus

$$\hat{\mu}_Y(t) = 17.6035 - 9.0499t + 1.2350t^2 \qquad 1 \le t \le 4$$

To compute a 95% confidence interval for the (population) average drug concentration 3 hours after injection, we let $\boldsymbol{a} = [1, 3, 9]^T$ and use (7.5.32). We get $t_{0.975:23} = 2.069$, $\hat{\mu}_Y(3) = 1.569$, and $SE(\boldsymbol{a}^T\hat{\boldsymbol{\beta}}) = 0.02642$. This leads to the confidence statement

$$C[1.569 - 2.069(0.02642) \le \mu_Y(3) \le 1.569 + 2.069(0.02642)] = 0.95$$

which, when simplified, is

$$C[1.514 \le \mu_Y(3) \le 1.623] = 0.95$$

From this we have 95% confidence that the *average* drug concentration in the blood for the study population of subjects, 3 hours after injection, is between 1.514 and 1.623 mg/l. ■

Remarks There are many other questions that are often of interest in growth curve studies. For instance, in Example 7.5.5 the investigator may be interested in the *distribution* of the drug concentrations 3 hours after injection. This involves not only the estimation of the *mean* drug concentration 3 hours after injection (which is computed using (7.5.29) with $t = 3$) but also the estimation of the standard deviation of the drug concentration 3 hours after injection. In some applications the investigator may be interested in a *tolerance interval* for this distribution. In other applications the investigator may want to predict a future value of the response of an item in the sample, to predict the value of the response of an item not in the sample, and so on. The procedures for answering these and other interesting questions may be found in more advanced textbooks.

In Section 7.5 of the laboratory manuals we discuss a macro we have supplied on the data disk that can be used to perform the calculations needed in this section to obtain point estimates and confidence intervals for $\boldsymbol{a}^T\boldsymbol{\beta}$ for specified vectors $\boldsymbol{a}$. In particular, this program can be used to compute point estimates and confidence intervals for $\alpha, \beta, \gamma, \mu_Y(t), \mu_Y(t_1) - \mu_Y(t_2)$, etc.

We pointed out earlier that even though the procedures for growth curves are explained using a quadratic growth curve model, they can be easily adapted for other situations involving multiple linear regression models. In the next example we illustrate the procedure when the growth curve for each subject in the population is a straight line as a function of time.

EXAMPLE 7.5.6

To establish a growth curve for the ramus bone (a bone in the jaw) in young boys, a random sample of 20 boys was selected and the ramus height measured in millimeters at the ages of 8, $8\frac{1}{2}$, 9, and $9\frac{1}{2}$ years. See [10]. The data are given in Table 7.5.3 and are also stored in the file **ramus.dat** on the data disk. Assume that in the range from 8 to 10 years of age, the growth curve for the ramus bone of each boy in the population is a straight line, so for boy i in the sample the growth curve is given by $\mu_i(t) = \alpha_i + \beta_i t$. We regress Y (ramus height) on t (age) and obtain estimates $\hat{\alpha}_i$ and $\hat{\beta}_i$ for $i = 1, \ldots, 20$ (i.e., for each boy). The principle objective in this study is to estimate the population growth curve and to obtain a 95% confidence interval for β, the average growth rate of the ramus bone. We assume the 20 boys ($m = 20$) constitute a simple random sample from a well-defined population of boys in this age group. We further suppose that assumptions in Box 7.5.1 for the random coefficient growth curve model are satisfied. We have $k = 4, p = 2, m = 20$, and the population growth curve is

$$\mu_Y(t) = \alpha + \beta t \qquad 8 \le t \le 10 \tag{7.5.40}$$

TABLE 7.5.3

Height of 20 Boys' Ramus Bones

	t			
Boy	**Age 8**	**Age $8\frac{1}{2}$**	**Age 9**	**Age $9\frac{1}{2}$**
1	47.8	48.8	49.0	49.7
2	46.4	47.3	47.7	48.4
3	46.3	46.8	47.8	48.5
4	45.1	45.3	46.1	47.2
5	47.6	48.5	48.9	49.3
6	52.5	53.2	53.3	53.7
7	51.2	53.0	54.3	54.5
8	49.8	50.0	50.3	52.7
9	48.1	50.8	52.3	54.4
10	45.0	47.0	47.3	48.3
11	51.2	51.4	51.6	51.9
12	48.5	49.2	53.0	55.5
13	52.1	52.8	53.7	55.0
14	48.2	48.9	49.3	49.8
15	49.6	50.4	51.2	51.8
16	50.7	51.7	52.7	53.3
17	47.2	47.7	48.4	49.5
18	53.3	54.6	55.1	55.3
19	46.2	47.5	48.1	48.4
20	46.3	47.6	51.3	51.8

Also note that the X matrix is given by

$$X = \begin{bmatrix} 1 & 8.0 \\ 1 & 8.5 \\ 1 & 9.0 \\ 1 & 9.5 \end{bmatrix} \tag{7.5.41}$$

Furthermore

$$y_1^T = [47.8, 48.8, 49.0, 49.7]$$

which is the first row of Table 7.5.3; i.e., the data for boy 1, and

$$y_2^T = [46.4, 47.3, 47.7, 48.4]$$

which is the second row of Table 7.5.3; i.e., the data for boy 2, etc. So we regress Y on t for each sample subject and get $\hat{\alpha}_i$, $\hat{\beta}_i$ for $i = 1, \ldots, 20$. From these we compute the quantities $\hat{\beta}$ and G using (7.5.28) and (7.5.34), respectively. The results are

$$\hat{\beta} = \begin{bmatrix} 33.7475 \\ 1.8660 \end{bmatrix} \qquad G = \begin{bmatrix} 103.959 & -11.525 \\ -11.525 & 1.358 \end{bmatrix}$$

Thus

$$\hat{\mu}_Y(t) = 33.7475 + 1.8660t \qquad 8 \leq t \leq 10$$

To compute a 95% confidence interval for β, we let $a = [0, 1]^T$ and use (7.5.32). We get $a^T\hat{\beta} = \hat{\beta} = 1.866$, and $SE(\hat{\beta}) = 0.2606$. This leads to the confidence statement (rounding to two decimals)

$$C[1.32 \leq \beta \leq 2.41] = 0.95$$

Thus we have 95% confidence that the average growth rate of the ramus bone for this population of boys is between 1.32 and 2.41 millimeters per year for $8 \leq t \leq 10$. ■

Problems 7.5

7.5.1 An agricultural experiment station investigator is interested in studying the growth pattern of pumpkins as a function of time. He monitors the weights (in pounds) of 24 pumpkins (selected by simple random sampling from a large pumpkin patch) each week, from the time they are 4 weeks old until they are 12 weeks old. The study population is all pumpkins in this patch, and the target population is the set of all pumpkins in similar patches. The data are given in Table 7.5.4 and are also stored in the file **pumpkin.dat** on the data disk.

Suppose that a quadratic growth curve model holds for each pumpkin growing on the given patch; i.e., (7.5.8) holds. In this case the population growth curve has the form

$$\mu_Y(t) = \alpha + \beta t + \gamma t^2$$

where α, β, and γ are unknown parameters. We assume that all the conditions in Box 7.5.1 are satisfied.

a What is the value of k? What is the value of m? What is the value of p?

b What are the values of t_i for $i = 1, 2, \ldots, k$?

c Display the X matrix.

d Display the vectors y_3 and y_{10}.

e Write out the regression model for pumpkin 15.

f We have computed $\hat{\beta} = [\hat{\alpha}, \hat{\beta}, \hat{\gamma}]^T$ and $SE(\hat{\gamma})$. They are

$$\hat{\beta} = [-18.9433, 6.1739, -0.2459]^T$$

and

$$SE(\hat{\gamma}) = 0.006821$$

Use these values to obtain a 90% confidence interval for γ.

g Find $\hat{\mu}_Y(t)$.

TABLE 7.5.4
Pumpkin Growth Data

Pumpkin	4 Weeks	6 Weeks	8 Weeks	10 Weeks	12 Weeks
1	2.1	9.6	15.7	20.3	21.6
2	3.1	11.3	18.5	23.6	26.6
3	3.5	11.3	18.4	22.5	25.3
4	0.2	5.9	9.5	12.0	12.1
5	2.8	11.8	17.8	22.8	24.0
6	2.1	8.7	13.3	16.8	17.4
7	2.9	11.3	15.7	19.6	20.4
8	0.3	6.3	11.6	14.6	16.7
9	0.8	7.5	13.2	16.6	16.8
10	3.6	11.4	18.3	21.6	23.9
11	1.6	8.8	13.5	16.0	16.3
12	2.2	9.8	15.6	19.3	20.3
13	1.5	8.5	13.8	16.9	17.9
14	3.0	10.0	16.3	20.2	21.1
15	0.2	6.6	11.5	15.5	17.4
16	1.3	8.9	13.0	17.5	18.4
17	2.9	11.0	16.6	22.6	24.7
18	1.8	9.6	15.6	19.2	21.3
19	0.5	7.5	13.0	16.2	17.8
20	2.0	8.7	12.8	14.6	14.7
21	2.3	10.6	15.9	20.4	20.6
22	0.2	6.8	11.7	14.6	16.2
23	2.1	9.7	15.1	18.2	20.6
24	1.6	8.8	15.4	18.5	20.1

h Find $\hat{\mu}_Y(8)$.

i Display $\boldsymbol{a}$ if $\boldsymbol{a}^T\boldsymbol{\beta} = \beta$.

j Display $\boldsymbol{a}$ if $\boldsymbol{a}^T\boldsymbol{\beta} = \mu_Y(t)$.

k An investigator wants to determine what the change in the average weight (in pounds) of pumpkins is from week 4 to week 12. What population parameter(s) must be estimated to estimate this change in average weight?

7.6 Exercises

7.6.1 An investigator wants to study the relationship between height and mineral composition of foliage in Japanese larch (trees). She has obtained a simple random sample of 26 trees from a plantation of Japanese larch and recorded the following information on each tree. (See Leyton, *Plant and Soil* 7: 167–177, 1956.)

Y = height of the tree in centimeters
X_1 = percentage content of nitrogen
X_2 = percentage content of phosporus
X_3 = percentage content of potassium
X_4 = percentage content of residual ash

The mineral composition data were obtained from dried, ground, new needles collected immediately below the terminal shoot. The data are given in Table 7.6.1 and are also stored in the file **larch.dat** on the data disk.

The sums of squared errors, *SSE*, and mean squared errors *MSE* for each subset model follow:

Variables in model	Sum of Squares	Mean Squares
none	227954	9118
X1	75363	3140
X2	91404	3809
X3	85123	3547
X4	93614	3901
X1,X2	47090	2047
X1,X3	36771	1599
X1,X4	53503	2326
X2,X3	63875	2777
X2,X4	65810	2861

X3,X4	61929	2693
X1,X2,X3	31858	1448
X1,X2,X4	40782	1854
X1,X3,X4	33403	1518
X2,X3,X4	52828	2401
X1,X2,X3,X4	30122	1434

a Carry out a variable selection analysis by the method of all-subsets regression using each of the following criteria: s, R^2, and $adj\text{-}R^2$. Find a short list of three best subset models in each case.

b Carry out a variable selection analysis by the method of all-subsets regression using the C_p criterion.

i Display in a table the value of C_p that corresponds to each subset model.

TABLE 7.6.1
Japanese Larch Tree Data

Observation Number	Y	X_1	X_2	X_3	X_4
1	351	2.20	0.417	1.35	1.79
2	249	2.10	0.354	0.90	1.08
3	171	1.52	0.208	0.71	0.47
4	373	2.88	0.335	0.90	1.48
5	321	2.18	0.314	1.26	1.09
6	191	1.87	0.271	1.15	0.99
7	225	1.52	0.164	0.83	0.85
8	291	2.37	0.302	0.89	0.94
9	284	2.06	0.373	0.79	0.80
10	213	1.84	0.265	0.72	0.77
11	138	1.89	0.192	0.46	0.46
12	213	2.45	0.221	0.76	0.95
13	151	1.88	0.186	0.52	0.95
14	130	1.93	0.207	0.60	0.92
15	93	1.80	0.157	0.67	0.60
16	95	1.81	0.195	0.47	0.57
17	147	1.49	0.165	0.66	0.80
18	88	1.53	0.226	0.68	0.66
19	65	1.43	0.224	0.44	0.45
20	120	1.54	0.271	0.51	0.95
21	72	1.13	0.187	0.38	0.63
22	160	1.63	0.200	0.62	1.10
23	72	1.36	0.211	0.71	0.47
24	252	1.76	0.283	0.96	0.96
25	310	2.53	0.284	0.85	1.39
26	336	2.59	0.303	1.02	0.95

ii Exhibit a plot of the C_p values. On this plot show the line through the origin with unit slope for comparing the values of $C_p - p$ for the various subset models.

iii Exhibit the best subset model according to the C_p criterion.

iv Exhibit any other models that are almost as good as the best model.

c Carry out a variable selection analysis for the preceding data using the forward selection procedure. Use *F-in* = 3.0. List the variables in the model at each step.

d Carry out a variable selection analysis for the preceding data using the backward elimination procedure with *F-out* = 3.0. List the variables in the model at each step.

e Carry out a variable selection analysis for the preceding data using stepwise regression with *F-in* = *F-out* = 3.0. Use β_0 as the initial model. List the variables in the model at each step.

f Summarize the results of parts (a)–(e) in a short written report.

7.6.2 Consider the data for the heights of the ramus bone in Table 7.5.3. Assume that in the range from 8 to 10 years of age, the growth curve for each boy (and for the population) is a quadratic function of time; i.e., $\mu_Y(t) = \alpha + \beta t + \gamma t^2, 8 \leq t \leq 10$. Suppose assumptions in Box 7.5.1 for a random coefficients growth curve model are satisfied. The results of some computations that you need follow:

$$\hat{\beta}^T = [\hat{\alpha}, \hat{\beta}, \hat{\gamma}] = [27.8742, 3.3863, -0.0916]$$
$$SE(\hat{\beta}) = 3.668$$
$$SE(\hat{\gamma}) = 0.2114$$

a What are the values of m, k, and p?

b Display the X matrix.

c Display the vector a such that $a^T\beta = \alpha$.

d Estimate the population growth curve.

e Construct 95% two sided confidence intervals for β and γ.

f Express $\mu_Y(t)$ in terms of α, β, γ, and t.

g Express $\mu_Y(8.5)$ in terms of α, β, and γ.

h Based on the confidence interval for γ obtained in (e), is it reasonable to conclude that the population growth curve is, for all practical purposes, a straight line for the ages of interest $8 \leq t \leq 10$? You may suppose that if $|\gamma| < 0.002$, it may be considered negligible for this problem.