CHAPTER 1

Basic Concepts for Experimental Design and Introductory Regression Analysis

Some basic concepts and principles in experimental design are introduced in this chapter, including the fundamental principles of replication, randomization, and blocking. A brief and self-contained introduction to regression analysis is also included. Commonly used techniques like simple and multiple linear regression, least squares estimation, and variable selection are covered.

1.1 INTRODUCTION AND HISTORICAL PERSPECTIVE

Experimentation is one of the most common activities that people engage in. It covers a wide range of applications from household activities like food preparation to technological innovation in material science, semiconductors, robotics, life science, and so on. It allows an investigator to find out what happens to the output or response when the settings of the input variables in a system are purposely changed. Statistical or often simple graphical analysis can then be used to study the relationship between the input and output values. A better understanding of how the input variables affect the performance of a system can thereby be achieved. This gain in knowledge provides a basis for selecting optimum input settings. Experimental design is a body of knowledge and techniques that enables an investigator to conduct better experiments, analyze data efficiently, and make the connections between the conclusions from the analysis and the original objectives of the investigation.

Experimentation is used to understand and/or improve a system. A system can be a product or process. A product can be one developed in engineering, biology, or the physical sciences. A process can be a manufacturing process, a

Experiments, Second Edition. By C. F. Jeff Wu and Michael S. Hamada

process that describes a physical phenomenon, or a nonphysical process such as those found in service or administration. Although most examples in the book are from engineering or the physical and biological sciences, the methods can also be applied to other disciplines, such as business, medicine, and psychology. For example, in studying the efficiency and cost of a payroll operation, the entire payroll operation can be viewed as a process with key input variables such as the number of supervisors, the number of clerks, method of bank deposit, level of automation, administrative structure, and so on. A computer simulation model can then be used to study the effects of changing these input variables on cost and efficiency.

Modern experimental design dates back to the pioneering work of the great statistician R. A. Fisher in the 1930s at the Rothamsted Agricultural Experimental Station in the United Kingdom. Fisher's work and the notable contributions by F. Yates and D. J. Finney were motivated by problems in agriculture and biology. Because of the nature of agricultural experiments, they tend to be large in scale, take a long time to complete, and must cope with variations in the field. Such considerations led to the development of blocking, randomization, replication, orthogonality, and the use of analysis of variance and fractional factorial designs. The theory of combinatorial designs, to which R. C. Bose has made fundamental contributions, was also stimulated by problems in block designs and fractional factorial designs. The work in this era also found applications in social science research and in the textile and woolen industries.

The next era of rapid development came soon after World War II. In attempting to apply previous techniques to solve problems in the chemical industries, G. E. P. Box and co-workers at Imperial Chemical Industries discovered that new techniques and concepts had to be developed to cope with the unique features of process industries. The new techniques focused on process modeling and optimization rather than on treatment comparisons, which was the primary objective in agricultural experiments. The experiments in process industries tend to take less time but put a premium on run size economy because of the cost of experimentation. These time and cost factors naturally favor sequential experimentation. The same considerations led to the development of new techniques for experimental planning, notably central composite designs and optimal designs. The analysis for these designs relies more heavily on regression modeling and graphical analysis. Process optimization based on the fitted model is also emphasized. Because the choice of design is often linked to a particular model (e.g., a second-order central composite design for a second-order regression model) and the experimental region may be irregularly shaped, a flexible strategy for finding designs to suit a particular model and/or experimental region is called for. With the availability of fast computational algorithms, optimal designs (which was pioneered by J. Kiefer) have become an important part of this strategy.

The relatively recent emphasis on variation reduction has provided a new source of inspiration and techniques in experimental design. In manufacturing, the ability to make many parts with few defects is a competitive advantage. Therefore variation reduction in the quality characteristics of these parts has become a major focus of quality and productivity improvement. G. Taguchi advocated the use of robust parameter design to improve a system (i.e., a product or process) by making it less sensitive to variation, which is hard to control during normal operating or use conditions of the product or process. The input variables of a system can be divided into two broad types: control factors, whose values remain fixed once they are chosen, and noise factors, which are hard to control during normal conditions. By exploiting the interactions between the control and noise factors, one can achieve robustness by choosing control factor settings that make the system less sensitive to noise variation. This is the motivation behind the new paradigm in experimental design, namely, modeling and reduction of variation. Traditionally, when the mean and variance are both considered, variance is used to assess the variability of the sample mean as with the *t* test or of the treatment comparisons as with the analysis of variance. The focus on variation and the division of factors into two types led to the development of new concepts and techniques in the planning and analysis of robust parameter design experiments. The original problem formulation and some basic concepts were developed by G. Taguchi. Other basic concepts and many sound statistical techniques have been developed by statisticians since the mid-1980s.

Given this historical background, we now classify experimental problems into five broad categories according to their objectives.

1. *Treatment Comparisons.* The main purpose is to compare several treatments and select the best ones. For example, in the comparison of six barley varieties, are they different in terms of yield and resistance to drought? If they are indeed different, how are they different and which are the best? Examples of treatments include varieties (rice, barley, corn, etc.) in agricultural trials, sitting positions in ergonomic studies, instructional methods, machine types, suppliers, and so on.

2. *Variable Screening.* If there is a large number of variables in a system but only a relatively small number of them is important, a screening experiment can be conducted to identify the important variables. Such an experiment tends to be economical in that it has few degrees of freedom left for estimating error variance and higher-order terms like quadratic effects or interactions. Once the important variables are identified, a follow-up experiment can be conducted to study their effects more thoroughly. This latter phase of the study falls into the category discussed next.

3. *Response Surface Exploration.* Once a smaller number of variables is identified as important, their effects on the response need to be explored. The relationship between the response and these variables is sometimes referred to as a response surface. Usually the experiment is based on a design that allows the linear and quadratic effects of the variables and some of the interactions between the variables to be estimated. This experiment tends to be larger (relative to the number of variables under study) than the screening experiment. Both parametric and semiparametric models may be considered. The latter is more computer-intensive but also more flexible in model fitting.

4. *System Optimization.* In many investigations, interest lies in the optimization of the system. For example, the throughput of an assembly plant or the yield of a chemical process is to be maximized; the amount of scrap or number of reworked pieces in a stamping operation is to be minimized; or the time required to process a travel claim reimbursement is to be reduced. If a response surface has been identified, it can be used for optimization. For the purpose of finding an optimum, it is, however, not necessary to map out the whole surface as in a response surface exploration. An intelligent sequential strategy can quickly move the experiment to a region containing the optimum settings of the variables. Only within this region is a thorough exploration of the response surface warranted.

5. *System Robustness.* Besides optimizing the response, it is important in quality improvement to make the system robust against noise (i.e., hard-to-control) variation. This is often achieved by choosing control factor settings at which the system is less sensitive to noise variation. Even though the noise variation is hard to control in normal conditions, it needs to be systematically varied during experimentation. The response in the statistical analysis is often the variance (or its transformation) among the noise replicates for a given control factor setting.

1.2 A SYSTEMATIC APPROACH TO THE PLANNING AND IMPLEMENTATION OF EXPERIMENTS

In this section, we provide some guidelines on the planning and implementation of experiments. The following seven-step procedure summarizes the important steps that the experimenter must address.

1. *State Objective.* The objective of the experiment needs to be clearly stated. All stakeholders should provide input. For example, for a manufactured product, the stakeholders may include design engineers who design the product, process engineers who design the manufacturing process, line engineers who run the manufacturing process, suppliers, lineworkers, customers, marketers, and managers.

2. *Choose Response.* The response is the experimental outcome or observation. There may be multiple responses in an experiment. Several issues arise in choosing a response. Responses may be *discrete* or *continuous*. Discrete responses can be counts or categories—for example, binary (good, bad) or ordinal (easy, normal, hard). Continuous responses are generally preferable. For example, a continuous force measurement for opening a door is better than an ordinal (easy, normal, hard to open) judgment; the recording of a continuous characteristic is preferred to the recording of the percent that the characteristic is within its specifications. Trade-offs may need to be made. For example, an ordinal measurement of force to open a door may be preferable to delaying the experiment until a device to take continuous measurements can be developed. Most importantly, there should be a good measurement system for measuring the response. In fact, an experiment called a *gauge repeatability and reproducibility* (R&R) *study* can be performed to assess a continuous measurement system (AIAG, 1990). When

there is a single measuring device, the variation due to the measurement system can be divided into two types: variation between the operators and variation within the operators. Ideally, there should be no between-operator variation and small within-operator variation. The gauge R&R study provides estimates for these two components of measurement system variation. Finally, the response should be chosen to increase understanding of mechanisms and physical laws involved in the problem. For example, in a process that is producing underweight soap bars, soap bar weight is the obvious choice for the response in an experiment to improve the underweight problem. By examining the process more closely, there are two subprocesses that have a direct bearing on soap bar weight: the mixing process that affects the soap bar density and the forming process that impacts the dimensions of the soap bars. In order to better understand the mechanism that causes the underweight problem, soap bar density and soap bar dimensions are chosen as the responses. Even though soap bar weight is not used as a response, it can be easily determined from its density and dimensions. Therefore, no information is lost in studying the density and dimensions. Such a study may reveal new information about the mixing and forming subprocesses, which can in turn lead to a better understanding of the underweight problem. Further discussions on and other examples of the choice of responses can be found in Phadke (1989) and León, Shoemaker, and Tsui (1993).

The chosen responses can be classified according to the stated objective. Three broad categories will be considered in this book: **nominal-the-best, larger-the-better**, and **smaller-the-better**. The first one will be addressed in Section 4.10, and the last two will be discussed in Section 6.2.

3. *Choose Factors and Levels.* A **factor** is a variable that is studied in the experiment. In order to study the effect of a factor on the response, two or more values of the factor are used. These values are referred to as **levels** or settings. A **treatment** is a combination of factor levels. When there is a single factor, its levels are the treatments. For the success of the experiment, it is crucial that potentially important factors be identified at the planning stage. There are two graphical methods for identifying potential factors. First, a **flow chart** of the process or system is helpful to see where the factors arise in a multistage process. In Figure 1.1, a rough sketch of a paper pulp manufacturing process is given which involves raw materials from suppliers, a chemical process to make a slurry which is passed through a mechanical process to produce the pulp. Involving all the stakeholders is invaluable in capturing an accurate description of the process or system. Second, a **cause-and-effect diagram** can be used to list and organize the potential factors that may impact the response. In Figure 1.2, a cause-and-effect diagram is given which lists the factors thought to affect the product quality of an injection molding process. Traditionally, the factors are organized under the headings: Man, Machine, Measurement, Material, Method, and Environment (Mother Nature for those who like M's). Because of their appearance, cause-and-effect diagrams are also called *fishbone diagrams*. Different characteristics of the factors need to be recognized because they can affect the choice of the experimental design. For example, a factor such as furnace temperature is *hard to change*. That

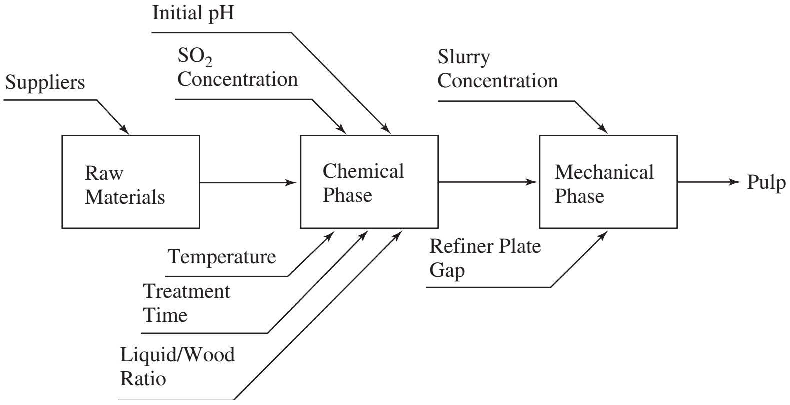

Figure 1.1. Flow chart, pulp manufacturing process.

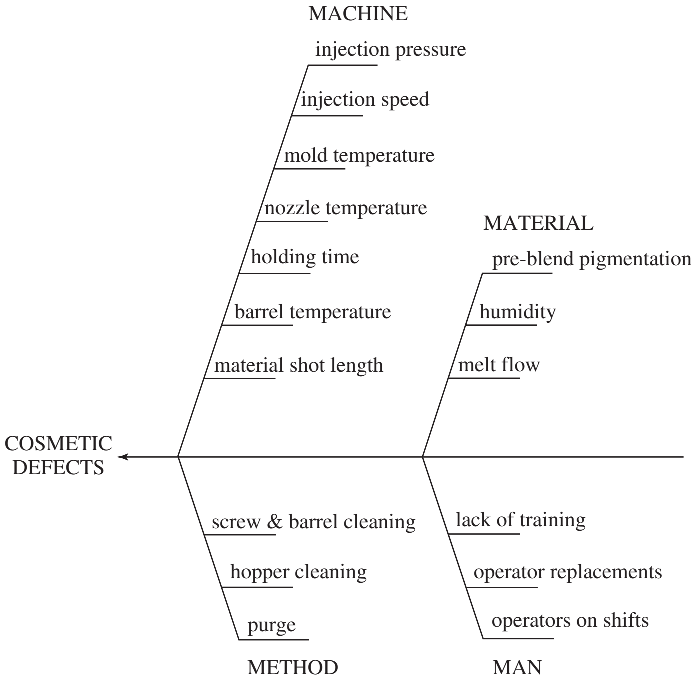

Figure 1.2. Cause-and-effect diagram, injection molding experiment.

is, after changing the temperature setting, it may take a considerable amount of time before the temperature stabilizes at the new setting. A factor may also be *hard to set* so that the actual level used in the experiment may be different than the intended level. For example, the actual impact force of a pellet projected at an automobile windshield can only be set within 3 psi of the intended impact force. Other factors that may be hard or impossible to control are referred to as *noise* factors. Examples of noise factors include environmental and customer use conditions. (An in-depth discussion of noise factors will be given in Section 11.3.)

Factors may be *quantitative* and *qualitative*. Quantitative factors like temperature, time, and pressure take values over a continuous range. Qualitative factors take on a discrete number of values. Examples of qualitative factors include operation mode, supplier, position, line, and so on. Of the two types of factors, there is more freedom in choosing the levels of quantitative factors. For example, if temperature (in degrees Celsius) is in the range 100–200°C, one could choose 130°C and 160°C for two levels or 125°C, 150°C, and 175°C for three levels. If only a linear effect is expected, two levels should suffice. If curvature is expected, then three or more levels are required. In general, the levels of quantitative factors must be chosen far enough apart so that an effect can be detected but not too far so that different physical mechanisms are involved (which would make it difficult to do statistical modeling and prediction). There is less flexibility in choosing the levels of qualitative factors. Suppose there are three testing methods under comparison. All three must be included as three levels of the factor "testing method," unless the investigator is willing to postpone the study of one method so that only two methods are compared in a two-level experiment.

When there is flexibility in choosing the number of levels, the choice may depend on the availability of experimental plans for the given combination of factor levels. In choosing factors and levels, *cost* and *practical constraints* must be considered. If two levels of the factor "material" represent expensive and cheap materials, a negligible effect of material on the response will be welcomed because the cost can be drastically reduced by replacing the expensive material by the cheap alternative. Factor levels must be chosen to meet practical constraints. If a factor level combination (e.g., high temperature and long time in an oven) can potentially lead to disastrous results (e.g., burned or overbaked), it should be avoided and a different plan should be chosen.

4. *Choose Experimental Plan.* Use the fundamental principles discussed in Section 1.3 as well as other principles presented throughout the book. The choice of the experimental plan is crucial. A poor design may capture little information which no analysis can rescue. On the other hand, if the experiment is well planned, the results may be obvious so that no sophisticated analysis is needed.

5. *Perform the Experiment.* The use of a *planning matrix* is recommended. This matrix describes the experimental plan in terms of the actual values or settings of the factors. For example, it lists the actual levels such as 50 or 70 psi if the factor is pressure. To avoid confusion and eliminate potential problems of running the wrong combination of factor levels in a multifactor experiment, each of the treatments, such as temperature at 30°C and pressure at 70 psi, should be put on a separate piece of paper and given to the personnel performing the experiment. It is also worthwhile to perform a *trial run* to see if there will be difficulties in running the experiment, namely, if there are problems with setting the factors and measuring the responses. Any deviations from the planned

experiment need to be recorded. For example, for hard-to-set factors, the actual values should be recorded.

6. *Analyze the Data.* An analysis appropriate for the design used to collect the data needs to be carried out. This includes model fitting and assessment of the model assumptions through an analysis of residuals. Many analysis methods will be presented throughout the book.

7. *Draw Conclusions and Make Recommendations.* Based on the data analysis, conclusions are presented which include the important factors and a model for the response in terms of the important factors. Recommended settings or levels for the important factors may also be given. The conclusions should refer back to the stated objectives of the experiment. A *confirmation experiment* is worthwhile, for example, to confirm the recommended settings. Recommendations for further experimentation in a *follow-up experiment* may also be given. For example, a follow-up experiment is needed if two models explain the experimental data equally well and one must be chosen for optimization.

For further discussion on the planning of experiments, see Coleman and Montgomery (1993), Knowlton and Keppinger (1993), and Barton (1997).

1.3 FUNDAMENTAL PRINCIPLES: REPLICATION, RANDOMIZATION, AND BLOCKING

There are three fundamental principles that need to be considered in the design of an experiment: **replication, randomization**, and **blocking**. Other principles will be introduced later in the book as they arise.

An *experimental unit* is a generic term that refers to a basic unit such as material, animal, person, machine, or time period, to which a treatment is applied. By *replication*, we mean that each treatment is applied to experimental units that are representative of the population of units to which the conclusions of the experiment will apply. It enables the estimation of the magnitude of experimental error (i.e., the error variance) against which the differences among treatments are judged. Increasing the number of replications, or *replicates*, decreases the variance of the treatment effect estimates and provides more power for detecting differences in treatments. A distinction needs to be made between replicates and *repetitions*. For example, three readings from the same experimental unit are repetitions, while the readings from three separate experimental units are replicates. The error variance from the former is less than that from the latter because repeated readings only measure the variation due to errors in reading while the latter also measures the unit-to-unit variation. Underestimation of the true error variance can result in the false declaration of an effect as significant.

The second principle is that of *randomization*. It should be applied to the allocation of units to treatments, the order in which the treatments are applied in performing the experiment, and the order in which the responses are measured. It provides protection against variables that are unknown to the experimenter but may impact the response. It reduces the unwanted influence of subjective judgment in treatment allocation. Moreover, randomization ensures validity of the estimate of experimental error and provides a basis for inference in analyzing the experiments. For an in-depth discussion on randomization, see Hinkelmann and Kempthorne (1994).

A prominent example of randomization is its use in clinical trials. If a physician were free to assign a treatment or control (or a new treatment versus an old treatment) to his/her patients, there might be a tendency to assign the treatment to those patients who are sicker and would not benefit from receiving a control. This would bias the outcome of the trial as it would create an unbalance between the control and treatment groups. A potentially effective treatment like a new drug may not even show up as promising if it is assigned to a larger proportion of "sick" patients. A random assignment of treatment/control to patients would prevent this from happening. Particularly commonplace is the use of the *double-blind trial*, in which neither the patient nor the doctor or investigator has access to the information about the actual treatment assignment. More on clinical trials can be found in Rosenberger and Lachin (2002).

A group of homogeneous units is referred to as a *block*. Examples of blocks include days, weeks, morning vs. afternoon, batches, lots, sets of twins, and pairs of kidneys. For *blocking* to be effective, the units should be arranged so that the within-block variation is much smaller than the between-block variation. By comparing the treatments within the same block, the block effects are eliminated in the comparison of the treatment effects, thereby making the experiment more efficient. For example, there may be a known day effect on the response so that if all the treatments can be applied within the same day, the day-to-day variation is eliminated.

If blocking is effective, it should be applied to remove the block-to-block variation. Randomization can then be applied to the assignments of treatments to units within the blocks to further reduce the influence of unknown variables. This strategy of **block what you can and randomize what you cannot** is used in randomized block designs, to be discussed in Section 3.2.

These three principles are generally applicable to physical experiments but not to computer experiments because the same input in a computer experiment gives rise to the same output. Computer experiments (see Santner et al., 2003) are not considered in the book, however.

A simple example will be used to explain these principles. Suppose two keyboards denoted by A and B are being compared in terms of typing efficiency. Six different manuscripts denoted by 1–6 are given to the same typist. First the test is arranged in the following sequence:

$$1.\ A, B, \qquad 2.\ A, B, \qquad 3.\ A, B, \qquad 4.\ A, B, \qquad 5.\ A, B, \qquad 6.\ A, B.$$

Because the manuscripts can vary in length and difficulty, each manuscript is treated as a "block" with the two keyboards as two treatments. Therefore, the experiment is replicated six times (with six manuscripts) and blocking is used to compare the two keyboards with the same manuscript. The design has a serious flaw, however. After typing the manuscript on keyboard A, the typist will be familiar with the content of the manuscript when he or she is typing the same manuscript on keyboard B. This "learning effect" will unfairly help the performance of keyboard B. The observed difference between A and B is the combination of the treatment effects (which measures the intrinsic difference between A and B) and the learning effect. For the given test sequence, it is impossible to disentangle the learning effect from the treatment effect. Randomization would help reduce the unwanted influence of the learning effect, which might not have been known to the investigator who planned the study. By randomizing the typing order for each manuscript, the test sequence may appear as follows:

$$1.\ A, B, \qquad 2.\ B, A, \qquad 3.\ A, B, \qquad 4.\ B, A, \qquad 5.\ A, B, \qquad 6.\ A, B.$$

With four *AB*'s and two *BA*'s in the sequence, it is a better design than the first one. A further improvement can be made. The design is not balanced because B benefits from the learning effect in four trials while A only benefits from two trials. There is still a residual learning effect not completely eliminated by the second design. The learning effect can be completely eliminated by requiring that half of the trials have the order *AB* and the other half the order *BA*. The actual assignment of *AB* and *BA* to the six manuscripts should be done by randomization. This method is referred to as *balanced randomization*. Balance is a desirable design property, which will be discussed later.

For simplicity of discussion, we have assumed that only one typist was involved in the experiment. In a practical situation, such an experiment should involve several typists that are representative of the population of typists so that the conclusions made from the study would apply more generally. This and other aspects of the typing experiment will be addressed in the exercises.

With these principles in mind, a useful addition to the cause-and-effect diagram is to indicate how the proposed experimental design addresses each listed factor. The following designations are suggested: **E** for an experimental factor, **B** for a factor handled by blocking, **O** for a factor held constant at one value, and **R** for a factor handled by randomization. This designation clearly indicates how the proposed design deals with each of the potentially important factors. The designation **O**, for "one value," serves to remind the experimenter that the factor is held constant during the current experiment but may be varied in a future experiment. An illustration is given in Figure 1.3 from the injection molding experiment discussed in Section 1.2.

Other designations of factors can be considered. For example, experimental factors can be further divided into two types (control factors and noise factors), as in the discussion on the choice of factors in Section 1.2. For the implementation of experiments, we may also designate an experimental factor as "hard-to-change" or "easy-to-change." These designations will be considered later as they arise.

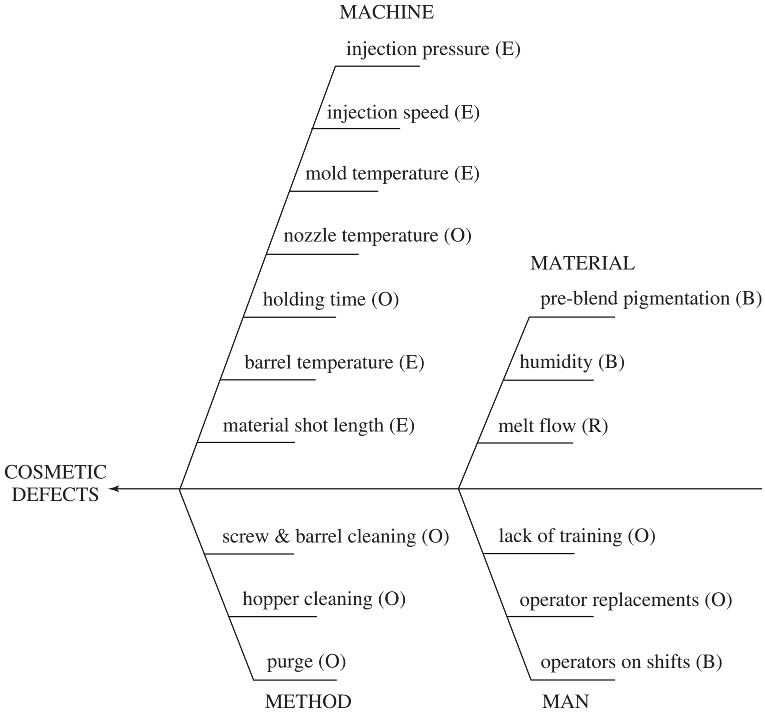

Figure 1.3. Revised cause-and-effect diagram, injection molding experiment.

1.4 SIMPLE LINEAR REGRESSION

Throughout the book, we will often model experimental data by the general linear model (also called the multiple regression model). Before considering the general linear model in Section 1.6, we present here the simplest case known as the simple linear regression model, which consists of a single covariate. We use the following data to illustrate the analysis technique known as *simple linear regression*.

Lea (1965) discussed the relationship between mean annual temperature and a mortality index for a type of breast cancer in women. The data (shown in Table 1.1), taken from certain regions of Great Britain, Norway, and Sweden, consist of the mean annual temperature (in degrees Fahrenheit) and a mortality index for neoplasms of the female breast.

Table 1.1 Breast Cancer Mortality Data

Mortality Index (M):	102.5	104.5	100.4	95.9	87.0	95.0	88.6	89.2
Temperature (T):	51.3	49.9	50.0	49.2	48.5	47.8	47.3	45.1
Mortality Index (M):	78.9	84.6	81.7	72.2	65.1	68.1	67.3	52.5
Temperature (T):	46.3	42.1	44.2	43.5	42.3	40.2	31.8	34.0

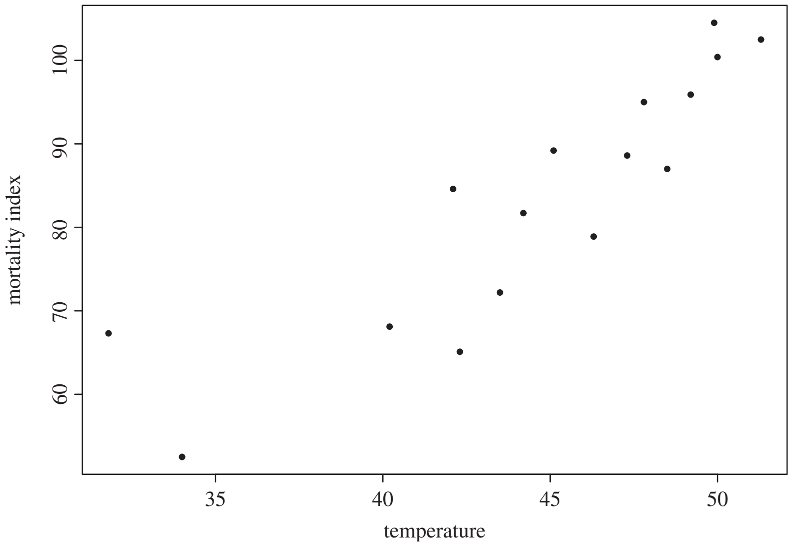

Figure 1.4. Scatter plot of temperature versus mortality index, breast cancer example.

The first step in any regression analysis is to make a scatter plot. A scatter plot of mortality index against temperature (Figure 1.4) reveals an increasing linear relationship between the two variables. Such a linear relationship between a response y and a covariate x can be expressed in terms of the following model:

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where ϵ is the random part of the model which is assumed to be normally distributed with mean 0 and variance σ^2, that is, $\epsilon \sim N(0, \sigma^2)$; because ϵ is normally distributed, so is y with mean $E(y) = \beta_0 + \beta_1 x$ and $\mathrm{Var}(y) = \sigma^2$.

If N observations are collected in an experiment, the model for them takes the form

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, N, \tag{1.1}$$

where y_i is the ith value of the response and x_i is the corresponding value of the covariate.

The unknown parameters in the model are the regression coefficients β_0 and β_1 and the error variance σ^2. Thus, the purpose for collecting the data is to estimate and make inferences about these parameters. For estimating β_0 and β_1, the least squares criterion is used; that is, the *least squares estimators* (LSEs), denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively, minimize the following quantity:

$$L(\beta_0, \beta_1) = \sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 x_i))^2. \tag{1.2}$$

Taking partial derivatives of (1.2) with respect to β_0 and β_1 and equating them to zero yields

$$\begin{aligned} \frac{\partial L}{\partial \beta_0} &= 2\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)(-1) = 0, \\ \frac{\partial L}{\partial \beta_1} &= 2\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0. \end{aligned} \tag{1.3}$$

From (1.3), the following two equations are obtained:

$$\begin{aligned} \sum_{i=1}^{N} y_i &= N\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{N} x_i, \\ \sum_{i=1}^{N} x_i y_i &= \hat{\beta}_0 \sum_{i=1}^{N} x_i + \hat{\beta}_1 \sum_{i=1}^{N} x_i^2. \end{aligned} \tag{1.4}$$

Equations (1.4) are called the *normal equations*. By solving them, the estimators of β_0 and β_1 are obtained as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \tag{1.5}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \text{ where } \bar{x} = \sum_{i=1}^{N} x_i/N \text{ and } \bar{y} = \sum_{i=1}^{N} y_i/N. \tag{1.6}$$

The fitted model is thus

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

The quantities $e_i = y_i - \hat{y}_i$, $i = 1, \ldots, N$, are called *residuals*. Clearly, the ith residual denotes the difference between the observed response y_i and the fitted value $\hat{y}_i$. Residuals are very useful in judging the appropriateness of a given regression model with respect to the available data. Using the fitted model, one can estimate the mean response corresponding to a certain covariate value, say x_0, of x as $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

For the breast cancer data, the response y is the mortality index M and the covariate x is the temperature T. Then we have for $N = 16$,

$$\begin{aligned} &\sum_{i=1}^{16}(T_i - \bar{T})^2 = 467.65, \qquad \sum_{i=1}^{16}(M_i - \bar{M})^2 = 3396.44, \\ &\sum_{i=1}^{16}(T_i - \bar{T})(M_i - \bar{M}) = 1102.57, \qquad \bar{T} = 44.5938, \qquad \bar{M} = 83.34. \end{aligned} \tag{1.7}$$

Using (1.5) and (1.6), the least square estimates of β_0 and β_1 are obtained as follows:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{16}(T_i - \bar{T})(M_i - \bar{M})}{\sum_{i=1}^{16}(T_i - \bar{T})^2} = \frac{1102.57}{467.65} = 2.36,$$

$$\hat{\beta}_0 = \bar{M} - \hat{\beta}_1\bar{T} = -21.79.$$

The fitted regression line is given by $\hat{M} = -21.79 + 2.36T$. This model can now be used to estimate the average mortality index due to breast cancer at a location that has a mean annual temperature of 49°F. Substituting $T = 49$ in the fitted model, we obtain the estimated mean mortality as

$$\hat{M} = -21.79 + 2.36 \times 49 = 93.85.$$

The fitted values and residuals for the 16 data points are shown in Table 1.2, where the residual value e_i is obtained by subtracting $\hat{y}_i$ from y_i.

1.5 TESTING OF HYPOTHESIS AND INTERVAL ESTIMATION

It is important to realize that the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables as they are functions of the data y_i. From (1.5),

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})y_i}{\sum_{i=1}^{N}(x_i - \bar{x})^2} - \frac{\bar{y}\sum_{i=1}^{N}(x_i - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}.$$

Table 1.2 Observed Responses, Fitted Values, and Residuals, Breast Cancer Example

y_i	$\hat{y}_i$	e_i
102.5	99.16	3.34
104.5	95.85	8.65
100.4	96.09	4.31
95.9	94.20	1.70
87.0	92.55	−5.55
95.0	90.90	4.10
88.6	89.72	−1.12
89.2	84.54	4.66
78.9	87.37	−8.47
84.6	77.46	7.14
81.7	82.42	−0.72
72.2	80.77	−8.57
65.1	77.94	−12.84
68.1	72.98	−4.88
67.3	53.18	14.12
52.5	58.37	−5.87

Since $\sum_{i=1}^{N}(x_i - \bar{x}) = 0$, the second term vanishes and it follows that

$$\hat{\beta}_1 = \sum_{i=1}^{N} w_i y_i, \tag{1.8}$$

where

$$w_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{N}(x_i - \bar{x})^2}. \tag{1.9}$$

Similarly, it can be shown that

$$\hat{\beta}_0 = \sum_{i=1}^{N} v_i y_i, \tag{1.10}$$

where

$$v_i = \frac{1}{N} - \bar{x} w_i. \tag{1.11}$$

Using (1.8)–(1.11), the following expressions can be obtained for the mean, variance, and covariance of $\hat{\beta}_1$ and $\hat{\beta}_0$:

$$E(\hat{\beta}_1) = \beta_1, \tag{1.12}$$

$$E(\hat{\beta}_0) = \beta_0, \tag{1.13}$$

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \tag{1.14}$$

$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \right), \tag{1.15}$$

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sigma^2. \tag{1.16}$$

Formulas (1.12)–(1.16) are special cases of general mean and variance–covariance formulas for the least square estimators in multiple linear regression. See (1.32) and (1.33).

From (1.12) and (1.13), we observe that $\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased estimators of β_1 and β_0, respectively. Clearly, to estimate $\mathrm{Var}(\hat{\beta}_0)$, $\mathrm{Var}(\hat{\beta}_1)$, and $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1)$, it is necessary to obtain an estimate of σ^2. This estimate can be obtained from the residuals e_i. The sum of squares of the residuals, also called the *residual sum of squares* (RSS), is given by

$$\mathrm{RSS} = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2,$$

which, after some straightforward algebraic manipulation, reduces to

$$\text{RSS} = \sum_{i=1}^{N}(y_i - \bar{y})^2 - \hat{\beta}_1^2 \sum_{i=1}^{N}(x_i - \bar{x})^2. \tag{1.17}$$

The degrees of freedom associated with RSS is $N - 2$. Thus, the mean square error is given by $\text{MSE} = \text{RSS}/(N - 2)$. It can be shown that $E(\text{RSS}) = (N - 2)\sigma^2$. Consequently, $E(\text{MSE}) = \sigma^2$, and MSE is an unbiased estimator of σ^2.

In order to know whether the covariate x has explanatory power, it is necessary to test the null hypothesis H_0: $\beta_1 = 0$. Under the assumption of normality of the error term ϵ_i in (1.1), each y_i is a normally distributed random variable. Since $\hat{\beta}_1$ can be expressed as a linear combination of y_i's and its mean and variance are given by (1.12) and (1.14), it follows that

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2 \Big/ \sum_{i=1}^{N}(x_i - \bar{x})^2\right).$$

The estimated standard deviation (i.e., standard error) of $\hat{\beta}_1$ is thus given by

$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{\text{MSE}}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}. \tag{1.18}$$

For testing H_0, the following statistic should be used

$$t = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}.$$

The above statistic follows a t distribution with $N - 2$ degrees of freedom. The higher the value of t, the more significant is the coefficient β_1. For the two-sided alternative H_1: $\beta_1 \neq 0$, p value $= \text{Prob}(|t_{N-2}| > |t_{\text{obs}}|)$, where $\text{Prob}(\cdot)$ denotes the probability of an event, t_ν is a random variable that has a t distribution with ν degrees of freedom, and t_{obs} denotes the observed or computed value of the t statistic. The t critical values can be found in Appendix B. A very small p value indicates that we have observed something which rarely happens under H_0, suggesting that H_0 is not true. In practice, H_0 is rejected at level of significance α if the p value is less than α. Common values of α are 0.1, 0.05, and 0.01.

Generally the p **value** gives the probability under the null hypothesis that the t statistic for an experiment conducted in comparable conditions will exceed the observed value $|t_{\text{obs}}|$. The p value has the following interpretation: The smaller the p value, the larger the evidence against the null hypothesis. Therefore, it provides a quantitative measure of the significance of effects in the problem under study. The same interpretation can be applied when other test statistics and null hypotheses are considered.

The $100(1-\alpha)\%$ confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm t_{N-2,\alpha/2}\ \mathrm{se}(\hat{\beta}_1),$$

where $t_{N-2,\alpha/2}$ is the upper $\alpha/2$ point of the t distribution with $N-2$ degrees of freedom. If the confidence interval does not contain 0, H_0 is rejected at level α.

Another way of judging the explanatory power of the covariate is by splitting up the total variation associated with the response data into two components. The quantity $\sum_{i=1}^{N}(y_i-\bar{y})^2$ measures the total variation in the data and is called the corrected total sum of squares (CTSS). From (1.17), we observe that

$$\mathrm{CTSS} = \mathrm{RegrSS} + \mathrm{RSS}, \tag{1.19}$$

where $\mathrm{RegrSS} = \hat{\beta}_1^2 \sum_{i=1}^{N}(x_i-\bar{x})^2$ is called the *corrected regression sum of squares*. Thus, the total variation in the data is split into the variation explained by the regression model plus the residual variation. This relationship is given in a table called the ANalysis Of VAriance or ANOVA table displayed in Table 1.3.

Based on (1.19), we can define

$$R^2 = \frac{\mathrm{RegrSS}}{\mathrm{CTSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{CTSS}}. \tag{1.20}$$

Because the R^2 value measures the "proportion of total variation explained by the fitted regression model $\hat{\beta}_0 + \hat{\beta}_1 x$," a higher R^2 value indicates a better fit of the regression model. It can be shown that R^2 is the square of the product-moment correlation r between $\mathbf{y} = (y_i)$ and $\mathbf{x} = (x_i)$, $i = 1, \ldots, N$, which is given by

$$r = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}}.$$

The mean square is the sum of squares divided by the corresponding degrees of freedom, where the degrees of freedom are those associated with each sum of squares. As explained earlier, the mean square error, or the residual mean square, is an unbiased estimator of σ^2.

Table 1.3 ANOVA Table for Simple Linear Regression

Source	Degrees of Freedom	Sum of Squares	Mean Squares
Regression	1	$\hat{\beta}_1^2 \sum_{i=1}^{N}(x_i-\bar{x})^2$	$\hat{\beta}_1^2 \sum_{i=1}^{N}(x_i-\bar{x})^2$
Residual	$N-2$	$\sum_{i=1}^{N}(y_i-\hat{y}_i)^2$	$\sum_{i=1}^{N}(y_i-\hat{y}_i)^2/(N-2)$
Total (corrected)	$N-1$	$\sum_{i=1}^{N}(y_i-\bar{y})^2$	

If the null hypothesis H_0: $\beta_1 = 0$ holds, the F statistic

$$\frac{\hat{\beta}_1^2 \sum_{i=1}^{N} (x_i - \bar{x})^2}{\sum_{i=1}^{N} e_i^2/(N-2)}$$

(the regression mean square divided by the residual mean square) has an F distribution with parameters 1 and $N - 2$, which are the degrees of freedom of its numerator and denominator, respectively. The p value is calculated by evaluating

$$\text{Prob}(F_{1,N-2} > F_{\text{obs}}), \tag{1.21}$$

where $F_{1,N-2}$ has an F distribution with parameters 1 and $N - 2$, and F_{obs} is the observed value of the F statistic. The F critical values can be found in Appendix D. The p value in (1.21) can be obtained from certain pocket calculators or by interpolating the values given in Appendix D. An example of an F distribution is given in Figure 2.1 (in Chapter 2) along with its critical values.

Let us now complete the analysis of the breast cancer mortality data. From Table 1.2, we obtain RSS $= \sum_{i=1}^{16} e_i^2 = 796.91$. Consequently,

$$\hat{\sigma}^2 = \text{MSE} = \frac{\text{RSS}}{14} = 56.92.$$

From the computations in Section 1.4, CTSS $= 3396.44$, and RegrSS $=$ CTSS $-$ RSS $= 2599.53$. Table 1.4 shows the resulting ANOVA table.

The R^2 is obtained as $2599.53/3396.44 = 0.7654$, which means 76.54% of the variation in the mortality indices is explained by the fitted model.

Using (1.18), $\text{se}(\hat{\beta}_1) = \sqrt{\text{MSE}/\sum_{i=1}^{16}(T_i - \bar{T})^2}$. Substituting $\sum_{i=1}^{16}(T_i - \bar{T})^2 = 467.65$ from (1.7),

$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{56.92}{467.65}} = 0.349.$$

Thus, the computed value of the t statistic for testing H_0: $\beta_1 = 0$ is

$$\frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} = \frac{2.358}{0.349} = 6.756.$$

Table 1.4 ANOVA Table for Breast Cancer Example

Source	Degrees of Freedom	Sum of Squares	Mean Squares
Regression	1	2599.53	2599.53
Residual	14	796.91	56.92
Total (corrected)	15	3396.44	

The corresponding p value is $\text{Prob}(t_{14} > 6.756) = 4.6 \times 10^{-6}$. Since this p value is negligibly small, one can safely reject the null hypothesis at much lower levels of significance than 0.05 or 0.01 and conclude that temperature indeed has a significant effect on mortality due to breast cancer.

From Table 1.4, the computed value of the F statistic is obtained as $2599.53/56.92 = 45.67$ and has a p value of 4.6×10^{-6} associated with it. We thus arrive at the same conclusion regarding the effect of temperature on mortality.

The 95% confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm t_{0.025,14}\text{se}(\hat{\beta}_1) = 2.36 \pm 2.145 \times 0.349 = (1.61, 3.11).$$

Recall that in Section 1.4 we obtained an estimate for the mean mortality corresponding to a temperature of 49°F as $\hat{M} = 93.85$. We are now interested in obtaining a standard error of this estimate and using it to obtain a confidence interval for the mean mortality at $T = 49$.

Let $\hat{y}_{x_0}$ denote the estimated average response from the fitted model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ corresponding to a given value $x = x_0$, that is,

$$\hat{y}_{x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0. \tag{1.22}$$

Then,

$$\begin{aligned}\text{Var}(\hat{y}_{x_0}) &= \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0)\\ &= \text{Var}(\hat{\beta}_0) + x_0^2\text{Var}(\hat{\beta}_1) + 2x_0\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)\\ &= \sigma^2\left(\frac{1}{N} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right).\end{aligned} \tag{1.23}$$

The last step follows from (1.14)–(1.16).

A $100(1 - \alpha)\%$ confidence interval for the estimated mean response corresponding to $x = x_0$ is thus given by

$$\hat{y}_{x_0} \pm t_{N-2,\alpha/2}\sqrt{\text{MSE}}\sqrt{\frac{1}{N} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}. \tag{1.24}$$

Next, consider the prediction of a new or future observation y corresponding to $x = x_0$. We assume that the simple linear regression model developed from the sampled data will be appropriate for the new observation. Therefore, the predicted value $\hat{y}_{\text{pred},x_0}$ is still obtained by substituting $x = x_0$ in the fitted regression model and is the same as the estimated mean response $\hat{y}_{x_0}$ in (1.22).

This prediction, however, differs from estimation of the mean response corresponding to $x = x_0$ in the sense that here we predict *an individual outcome* observed in the future. Since the future value actually observed will fluctuate around the mean response value, the variance σ^2 of an individual observation should be added to the variance of $\hat{y}_{x_0}$. By adding σ^2 to the variance in (1.23), we have

$$\mathrm{Var}(\hat{y}_{\text{pred},x_0}) = \sigma^2 \left(1 + \frac{1}{N} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right).$$

Consequently, a $100(1-\alpha)\%$ confidence interval for a predicted individual response corresponding to $x = x_0$, also called a $100(1-\alpha)\%$ *prediction interval*, is given by

$$\hat{y}_{\text{pred},x_0} \pm t_{N-2,\alpha/2}\sqrt{\text{MSE}}\sqrt{1 + \frac{1}{N} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}. \tag{1.25}$$

Using (1.24), a 95% confidence interval for the estimated mean mortality index corresponding to a temperature of 49°F is obtained as

$$\begin{aligned}
&\hat{M}_{49} \pm t_{14,0.025}\sqrt{\text{MSE}}\sqrt{\frac{1}{16} + \frac{(\bar{T} - 49)^2}{\sum_{i=1}^{16}(T_i - \bar{T})^2}}\\
&= 93.85 \pm t_{14,0.025}\sqrt{56.92}\sqrt{\frac{1}{16} + \frac{(44.59 - 49)^2}{467.65}}\\
&= 93.85 \pm 2.145 \times 7.54 \times 0.323 = (88.63, 99.07).
\end{aligned}$$

Similarly, using (1.25), a 95% confidence interval for the predicted mortality index of an individual corresponding to the temperature of 49°F is obtained as

$$\begin{aligned}
&\hat{M}_{\text{pred},49} \pm t_{14,0.025}\sqrt{\text{MSE}}\sqrt{1 + \frac{1}{16} + \frac{(\bar{T} - 49)^2}{\sum_{i=1}^{16}(T_i - \bar{T})^2}}\\
&= 93.85 \pm t_{14,0.025}\sqrt{56.92}\sqrt{1 + \frac{1}{16} + \frac{(44.59 - 49)^2}{467.65}}\\
&= 93.85 \pm 2.145 \times 7.54 \times 1.051 = (76.85, 110.85).
\end{aligned}$$

1.6 MULTIPLE LINEAR REGRESSION

Experimental data can often be modeled by the general linear model (also called the multiple regression model). Suppose that the response y is related to k covariates (also called explanatory variables, regressors, predictors) $x_1, x_2, \ldots, x_k$ as follows:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon, \tag{1.26}$$

where ϵ is the random part of the model which is assumed to be normally distributed with mean 0 and variance σ^2, i.e., $\epsilon \sim N(0, \sigma^2)$; because ϵ is normally distributed, so is y and $\mathrm{Var}(y) = \sigma^2$. The structural part of the model is

$$\begin{aligned} E(y) &= \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + E(\epsilon) \\ &= \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k. \end{aligned}$$

Here, $E(y)$ is linear in the β's, the regression coefficients, which explains the term **linear model**.

If N observations are collected in an experiment, the model for them takes the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \qquad i = 1, \ldots, N, \tag{1.27}$$

where y_i is the ith value of the response and $x_{i1}, \ldots, x_{ik}$ are the corresponding values of the k covariates.

These N equations can be written in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.28}$$

where $\mathbf{y} = (y_1, \ldots, y_N)^T$ is the $N \times 1$ vector of responses, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$ is the $(k+1) \times 1$ vector of regression coefficients, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)^T$ is the $N \times 1$ vector of errors, and $\mathbf{X}$, the $N \times (k+1)$ **model matrix**, is given as

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nk} \end{pmatrix}. \tag{1.29}$$

The unknown parameters in the model are the regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$ and the error variance σ^2. As in Section 1.4, the least squares criterion is used; that is, the least squares estimator (LSE), denoted by $\hat{\boldsymbol{\beta}}$, minimizes the following quantity:

$$\sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}))^2,$$

which in matrix notation is

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

In other words, the squared distance between the response vector $\mathbf{y}$ and the vector of fitted values $\mathbf{X}\hat{\boldsymbol{\beta}}$ is minimized. In order to minimize the sum of squared residuals, the vector of **residuals**

$$\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

needs to be perpendicular to the vector of *fitted values*

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}};$$

that is, the cross product between these two vectors should be zero:

$$\mathbf{r}^T\hat{\mathbf{y}} = \mathbf{r}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}.$$

An equivalent way of stating this is that the columns of the model matrix **X** need to be perpendicular to **r**, the vector of residuals, and thus satisfy

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}. \tag{1.30}$$

The solution to this equation is the **least squares estimate**, which is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{1.31}$$

From (1.31),

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\mathbf{y}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \qquad (\text{since} \quad E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}) \\ &= \boldsymbol{\beta}. \end{aligned} \tag{1.32}$$

The variance of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned} \text{Var}(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\text{Var}(\mathbf{y})((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2\mathbf{I} \qquad (\text{since} \quad \text{Var}(\mathbf{y}) = \sigma^2\mathbf{I}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2. \end{aligned} \tag{1.33}$$

In fitting the model, one wants to know if any of the covariates (regressors, predictors, explanatory variables) have explanatory power. None of them has explanatory power if the null hypothesis

$$H_0\colon \beta_1 = \cdots = \beta_k = 0 \tag{1.34}$$

holds. In order to test this null hypothesis, one needs to assess how much of the total variation in the response data can be explained by the model relative to the remaining variation after fitting the model, which is contained in the residuals.

Recall how the model was fitted: the residuals are perpendicular to the fitted values so that we have a right triangle. This brings to mind the Pythagorean theorem: The squared length of the hypotenuse is equal to the sum of the squared lengths of its opposite sides. In vector notation, the squared distance of a vector $\mathbf{a}$ is simply $\mathbf{a}^T\mathbf{a} = \sum a_i^2$. Thus, from the least squares fit, we obtain

$$\begin{aligned}\mathbf{y}^T\mathbf{y} &= (\mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{X}\hat{\boldsymbol{\beta}}) + (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} + (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}),\end{aligned}$$

where $\mathbf{y}^T\mathbf{y}$ is the *total sum of squares (uncorrected)*, $\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}$ is the *regression sum of squares (uncorrected)*, and

$$\text{RSS} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

is the *residual* (or *error*) *sum of squares*. In order to test the null hypothesis (1.34), the contribution from estimating the intercept β_0 needs to be removed. Subtracting off its contribution $N\bar{y}^2$, where $\bar{y}$ is the average of the N observations, yields

$$\begin{aligned}\text{CTSS} &= \mathbf{y}^T\mathbf{y} - N\bar{y}^2 = \text{RegrSS} + \text{RSS} \\ &= (\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} - N\bar{y}^2) + (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}),\end{aligned} \tag{1.35}$$

where CTSS is called the *corrected total sum of squares* and is equal to $\sum_{i=1}^N (y_i - \bar{y})^2$, which measures the variation in the data, and RegrSS is called the *corrected regression sum of squares*. Note that if there is a single covariate, the expressions of RegrSS and RSS in (1.35) reduce to the corresponding expressions in Section 1.5. In the remainder of this book, "corrected" will be dropped in reference to various sums of squares but will be implied. As in Section 1.5, the splitting up of the total variation in data into two components is summarized in the ANOVA table displayed in Table 1.5, and a measure of the proportion of the total variation explained by the fitted regression model $\mathbf{X}\hat{\boldsymbol{\beta}}$ is given by

$$R^2 = \frac{\text{RegrSS}}{\text{CTSS}} = 1 - \frac{\text{RSS}}{\text{CTSS}}. \tag{1.36}$$

It can be shown that R is the product-moment correlation between $\mathbf{y} = (y_i)_{i=1}^N$ and $\hat{\mathbf{y}} = (\hat{y}_i)_{i=1}^N$ and is called the *multiple correlation coefficient*.

Table 1.5 ANOVA Table for General Linear Model

Source	Degrees of Freedom	Sum of Squares	Mean Squares
Regression	k	$\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} - N\bar{y}^2$	$(\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} - N\bar{y}^2)/k$
Residual	$N - k - 1$	$(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$	$(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(N - k - 1)$
Total (corrected)	$N - 1$	$\mathbf{y}^T\mathbf{y} - N\bar{y}^2$	

The mean square is the sum of squares divided by the corresponding degrees of freedom, where the degrees of freedom are those associated with each sum of squares. The residual mean square is commonly referred to as the *mean square error* (MSE) and is used as an estimate for σ^2. Thus we can denote the MSE as

$$\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(N - k - 1). \tag{1.37}$$

If the null hypothesis (1.34) holds, the F statistic

$$\frac{(\hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{X}\hat{\boldsymbol{\beta}} - N\overline{y}^2)/k}{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(N - k - 1)}$$

(the regression mean square divided by the residual mean square) has an F distribution with parameters k and $N - k - 1$, which are the degrees of freedom of its numerator and denominator, respectively. The p value is calculated by evaluating

$$\text{Prob}(F_{k,N-k-1} > F_{\text{obs}}),$$

where Prob$(\cdot)$ denotes the probability of an event, $F_{k,N-k-1}$ has an F distribution with parameters k and $N - k - 1$, and F_{obs} is the observed value of the F statistic.

In model (1.28), the vector $\boldsymbol{\epsilon}$ follows a multivariate normal distribution with mean vector 0 and covariance matrix $\sigma^2\mathbf{I}$. Using this with (1.32) and (1.33), it can be shown that the least squares estimate $\hat{\boldsymbol{\beta}}$ has a multivariate normal distribution with mean vector $\boldsymbol{\beta}$ and variance–covariance matrix $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, that is,

$$\hat{\boldsymbol{\beta}} \sim MN(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}),$$

where MN stands for the multivariate normal distribution. The (i, j)th entry of the variance–covariance matrix is Cov$(\hat{\beta}_i, \hat{\beta}_j)$ and the jth diagonal element is Cov$(\hat{\beta}_j, \hat{\beta}_j) = $ Var$(\hat{\beta}_j)$. Therefore, the distribution for the individual $\hat{\beta}_j$ is $N(\beta_j, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj})$, which suggests that for testing the null hypothesis

$$H_0\colon \beta_j = 0,$$

the following t statistic be used:

$$\frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}}. \tag{1.38}$$

Under H_0, the t statistic in (1.38) has a t distribution with $N - k - 1$ degrees of freedom. This can also be used to construct confidence intervals since the denominator of the t statistic is the standard error of its numerator $\hat{\beta}_j$:

$$\hat{\beta}_j \pm t_{N-k-1,\alpha/2}\sqrt{\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}, \tag{1.39}$$

where $t_{N-k-1,\alpha/2}$ is the upper $\alpha/2$ quantile of the t distribution with $N-k-1$ degrees of freedom. See Appendix B for t critical values.

Besides testing the individual β_j's, testing linear combinations of the β_j's can be useful. As an example, for testing $\mathbf{a}^T\boldsymbol{\beta} = \sum_{j=0}^{k} a_j\beta_j$, where $\mathbf{a}$ is a $(k+1)\times 1$ vector, it can be shown that

$$\mathbf{a}^T\hat{\boldsymbol{\beta}} \sim N(\mathbf{a}^T\boldsymbol{\beta}, \sigma^2\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}).$$

This suggests using the test statistic

$$\frac{\mathbf{a}^T\hat{\boldsymbol{\beta}}}{\sqrt{\hat{\sigma}^2\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}}, \tag{1.40}$$

which has a t distribution with $N-k-1$ degrees of freedom.

More generally, for any setting of $(x_1,\ldots,x_k)$, the $E(y)$ value, $\beta_0 + \sum_{j=1}^{k}\beta_j x_j$, can be rewritten in vector notation as $\mathbf{x}^T\boldsymbol{\beta}$, where $\mathbf{x}^T = (1, x_1, \ldots, x_k)$. It can be estimated by its sample counterpart $\mathbf{x}^T\hat{\boldsymbol{\beta}}$. The corresponding test statistic is the same as in (1.40) with $\mathbf{a}$ replaced by $\mathbf{x}$. The $100(1-\alpha)\%$ confidence interval for $\mathbf{x}^T\boldsymbol{\beta}$ is given by

$$\mathbf{x}^T\hat{\boldsymbol{\beta}} \pm t_{N-k-1,\alpha/2}\sqrt{\hat{\sigma}^2\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}}. \tag{1.41}$$

This is an extension of formula (1.24) for simple linear regression. If $\mathbf{x}^T\hat{\boldsymbol{\beta}}$ is used to *predict* a future y value at $(x_1,\ldots,x_k)$, the $100(1-\alpha)\%$ prediction interval is obtained from (1.41) with $\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}$ replaced by $(1+\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x})$. This prediction interval is an extension of formula (1.25) for simple linear regression.

Extra Sum of Squares Principle

The *extra sum of squares principle* will be useful later for developing test statistics in a number of situations. Suppose that there are two models, say Model I and Model II. Model I is a special case of Model II, denoted by Model I $\subset$ Model II. Let

$$\text{Model I:}\quad y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$$

and

$$\text{Model II:}\quad y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \beta_{k+1}x_{i,k+1} + \cdots + \beta_q x_{iq} + \epsilon_i'.$$

Model I $\subset$ Model II since $\beta_{k+1} = \cdots = \beta_q = 0$ in Model I. Then, for testing the null hypothesis that Model I is adequate, that is,

$$H_0:\quad \beta_{k+1} = \cdots = \beta_q = 0 \tag{1.42}$$

holds, the extra sum of squares principle employs the F statistic:

$$\frac{(\text{RSS}(\text{Model I}) - \text{RSS}(\text{Model II}))/(q-k)}{\text{RSS}(\text{Model II})/(N-q-1)}, \tag{1.43}$$

where RSS stands for the residual sum of squares. It follows that

$$\begin{aligned}&\text{RSS}(\text{Model I}) - \text{RSS}(\text{Model II})\\ &\quad = \text{RegrSS}(\text{Model II}) - \text{RegrSS}(\text{Model I}),\end{aligned}$$

where RegrSS denotes the regression sum of squares; thus, the numerator of the F statistic in (1.43) is the gain in the regression sum of squares for fitting the more general Model II relative to Model I, that is, the *extra sum of squares*. When (1.42) holds, the F statistic has an F distribution with parameters $q - k$ (the difference in the number of estimated parameters between Models I and II) and $N - q - 1$. The extra sum of squares technique can be implemented by fitting Models I and II separately, obtaining their respective residual sums of squares, calculating the F statistic above, and then computing its p value ($=$ Prob($F > F_{\text{obs}}$)).

1.7 VARIABLE SELECTION IN REGRESSION ANALYSIS

In the fitting of the general linear model (1.27), those covariates whose regression coefficients are not significant may be removed from the full model. A more parsimonious model (i.e., one with fewer covariates) is preferred as long as it can explain the data well. This follows from the **principle of parsimony** (or Occam's razor), a principle attributed to the fourteenth-century English philosopher, William of Occam, which states "entities should not be multiplied beyond necessity." It is also known that a model that fits the data too well may give poor predictions, a phenomenon that can be justified by the following result in regression analysis (Draper and Smith, 1998): The average variance of $\hat{y}_i$ $(= \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})$ over $i = 1, \ldots, N$ is proportional to $k + 1$, the number of parameters in the regression model in (1.26). As k increases, a large model can better fit the data but the prediction variance also increases.

The goal of variable selection in regression analysis is to identify the smallest subset of the covariates that explains the data well; one hopes to capture the true model or at least the covariates of the true model with the most significant regression coefficients. One class of strategies is to use a model selection criterion to evaluate all possible subsets of the covariates and select the subset (which corresponds to a model) with the best value of the criterion. This is referred to as **best subset regression**. To maintain a balance between data fitting and prediction, a good model selection criterion should reward good model fitting as well as penalize model complexity. The R^2 in (1.36) is not a suitable criterion because it increases as the number of covariates increases. That is, it does not penalize excessively large models. An alternative criterion is the adjusted R^2 (Wherry, 1931), which takes into consideration the reduction in degrees of freedom for estimating the residual variance with inclusion of covariates in the model. For a model containing k covariates, the *adjusted* R^2 is given by

$$R_a^2 = 1 - \frac{\text{RSS}/(N-k-1)}{\text{CTSS}/(N-1)}. \tag{1.44}$$

Note that the difference between R_a^2 and the expression for R^2 in (1.36) is in the degrees of freedom in the denominator and the numerator of (1.44). If an insignificant variable is added to a model, the R^2 will increase, but the adjusted R_a^2 may decrease.

Another commonly used criterion is the **C_p statistic** (Mallows, 1973). Suppose there are a total of q covariates. For a model that contains p regression coefficients, which consist of those associated with $p-1$ covariates $(p-1<q)$ and an intercept term β_0, define its C_p value as

$$C_p = \frac{\text{RSS}}{s^2} - (N-2p), \tag{1.45}$$

where RSS is the residual sum of squares for the model, s^2 is the mean square error (see (1.37)) for the model containing all q covariates and β_0, and N is the total number of observations. As the model gets more complicated, the RSS term in (1.45) decreases while the value p in the second term increases. The counteracting effect of these two terms prevents the selection of extremely large or small models. If the model is true, $E(\text{RSS}) = (N-p)\sigma^2$. Assuming that $E(s^2) = \sigma^2$, it is then approximately true that

$$E(C_p) \approx \frac{(N-p)\sigma^2}{\sigma^2} - (N-2p) = p.$$

Thus one should expect the best fitting models to be those with $C_p \approx p$. Further theoretical and empirical studies suggest that models whose C_p values are low and are close to p should be chosen.

Alternatively, two commonly used criteria are the *Akaike information criterion* (AIC) and the *Bayesian information criterion* (BIC), which are defined as

$$\text{AIC} = N \ln\left(\frac{\text{RSS}}{N}\right) + 2p, \tag{1.46}$$

and

$$\text{BIC} = N \ln\left(\frac{\text{RSS}}{N}\right) + p\ln(N), \tag{1.47}$$

where p is the number of parameters (i.e., regression coefficients) in the model. Unlike the C_p statistic, AIC and BIC can be defined for general statistical problems beyond regression (see McQuarrie and Tsai, 1998). Another advantage is that, in software like R, both AIC and BIC are available in the stepwise regression routine.

For moderate to large q, fitting all subsets is computationally infeasible. An alternative strategy is based on adding or dropping one covariate at a time from a given model, which requires fewer model fits but can still identify good fitting models. It need not identify the best-fitting models as in any optimization that optimizes sequentially (and locally) rather than globally. The main idea is to compare the current model with a new model obtained by adding or deleting a covariate from the current model. Call the smaller and larger models Model I and Model II, respectively. Based on the extra sum of squares principle in Section 1.6, one can compute the F statistic in (1.43), also known as a *partial* F statistic, to determine if the covariate should be added or deleted. The partial F statistic takes the form

$$\frac{\text{RSS(Model I)} - \text{RSS(Model II)}}{\text{RSS(Model II)}/\nu}, \tag{1.48}$$

where ν is the degrees of freedom of the RSS (residual sum of squares) for Model II. Three versions of this strategy are considered next.

One version is known as **backward elimination**. It starts with the full model containing all q covariates and computes partial F's for all models with $q - 1$ covariates. At the mth step, Model II has $q - m + 1$ covariates and Model I has $q - m$ covariates, so that $\nu = N - (q - m + 1) - 1 = N - q + m - 2$ in the partial F in (1.48). At each step, compute the partial F value for each covariate being considered for removal. The one with the lowest partial F, provided it is smaller than a preselected value, is dropped. The procedure continues until no more covariates can be dropped. The preselected value is often chosen to be $F_{1,\nu,\alpha}$, the upper α critical value of the F distribution with 1 and ν degrees of freedom. Choice of the α level determines the stringency level for eliminating covariates. Typical α's range from $\alpha = 0.1$ to 0.2. A conservative approach would be to choose a smaller F (i.e., a large α) value so that important covariates are not eliminated. The statistic in (1.48) does not have a proper F distribution because the RSS term in its denominator has a noncentral χ^2 distribution (unless Model II is the true full model). Therefore the F critical values in the selection procedure serve only as guidelines. The literature often refers to them as *F-to-remove* values to make this distinction.

Another version is known as **forward selection**, which starts with the model containing an intercept and then adds one covariate at a time. The covariate with the largest partial F (as computed by (1.48)) is added, provided that it is larger than a preselected F critical value, which is referred to as an *F-to-enter* value. The forward selection procedure is not recommended as it often misses important covariates. It is combined with backward elimination to form the following stepwise selection procedure.

The **stepwise selection** procedure starts with two steps of the forward selection and then alternates between one step of backward elimination and one step of forward selection. The F-to-remove and F-to-enter values is usually chosen to be the same. A typical choice is $F_{1,\nu,\alpha}$ with $\alpha = 0.05$, 0.1, 0.15. The choice varies from data to data and can be changed as experience dictates. Among the three selection procedures, stepwise selection is known to be the most effective and is therefore recommended for general use.

For a comprehensive discussion on variable selection, see Draper and Smith (1998).

1.8 ANALYSIS OF AIR POLLUTION DATA

In this section, we consider an application of multiple linear regression and variable selection. Researchers at General Motors collected data on 60 U.S. Standard Metropolitan Statistical Areas (SMSAs) in a study of whether air pollution contributes to mortality (Gibbons and McDonald, 1980). The response variable for analysis is age adjusted mortality (denoted by *MORTALITY*). The predictors include variables measuring demographic characteristics of the cities, variables measuring climate characteristics, and variables recording the pollution potential of three different air pollutants. The goal of this analysis is to find out if air pollution affects mortality after *adjusting* for the effects of other significant variables. The data are available at

http://lib.stat.cmu.edu/DASL/Stories/AirPollutionandMortality.html.

The complete list of the 14 predictor variables and their codes are given below:

1. *JanTemp*: mean January temperature (degrees Fahrenheit).
2. *JulyTemp*: mean July temperature (degrees Fahrenheit).
3. *RelHum*: relative humidity.
4. *Rain*: annual rainfall (inches).
5. *Education*: median education.
6. *PopDensity*: population density.
7. *%NonWhite*: percentage of non-whites.
8. *%WC*: percentage of white collar workers.
9. *pop*: population.
10. *pop/house*: population per household.
11. *income*: median income.
12. *HCPot*: HC pollution potential.
13. *NOxPot*: nitrous oxide pollution potential.
14. *SO2Pot*: sulphur dioxide pollution potential.

Note that these data are *observational* in which mortality and the 14 variables are observed together rather than them being *experimental* in which the 14 variables are set to specified values and then the resulting mortality is observed.

Consequently, any relationship found between mortality and the 14 variables in the analysis of these data need not imply causality. For information on observational studies and causality, see Rosenbaum (2002).

Among the 60 data points, the 21st (Fort Worth, TX) has two missing values and is discarded from the analysis. Thus the total sample size N is 59. The three pollution variables HCPot, NOxPot, and SO2Pot are highly skewed. A log transformation makes them nearly symmetric. Therefore these three variables are replaced by log(HCPot), log(NOxPot), and log(SO2Pot), respectively, and are called logHC, logNOx, and logSO2, respectively. Figure 1.5 shows the

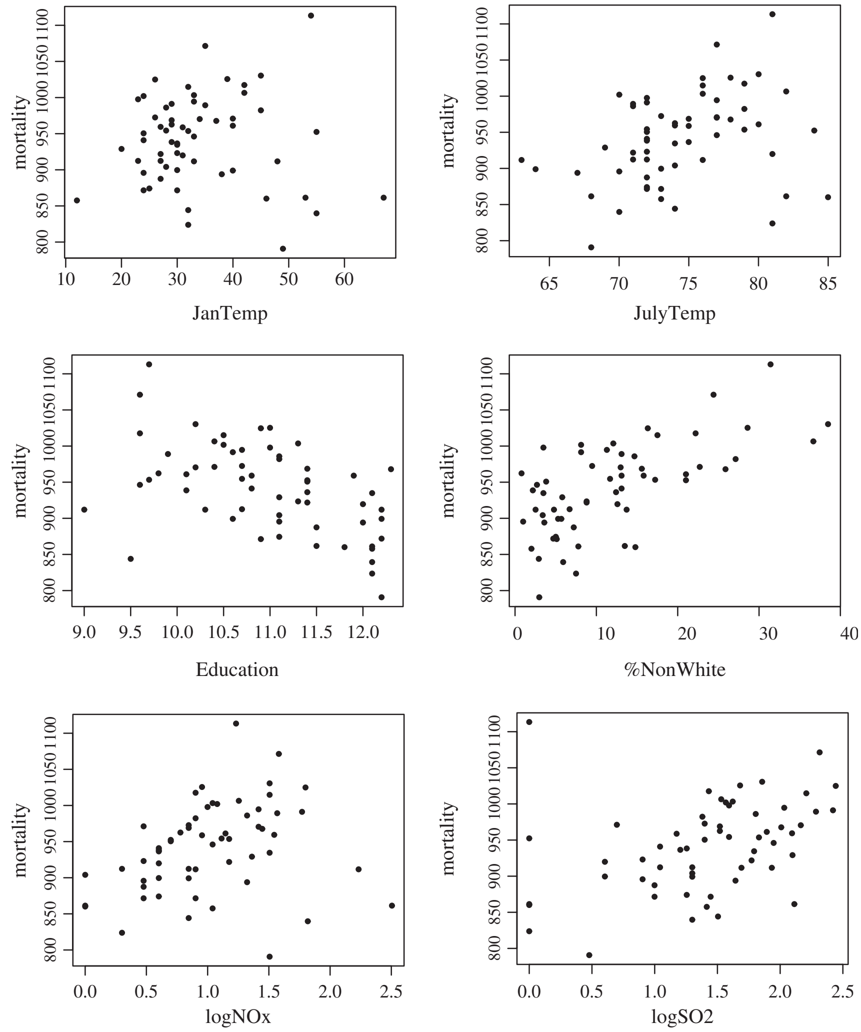

Figure 1.5. Scatter plots of MORTALITY versus JanTemp, JulyTemp, Education, %NonWhite, logNOx, and logSO2, air pollution example.

scatter plots of mortality against six selected predictors. They clearly show an increasing trend of mortality as a function of %NonWhite, logSO2, and logNOx and a decreasing trend as a function of Education, which seem to support our intuition about these variables. There is also a discernable increasing trend in JanTemp and JulyTemp. While the information gleaned from the plots can be useful and suggestive, they should *not* be taken as conclusive evidence because of the potentially high correlations between the predictor variables. Such correlations can complicate the relationship between the response and the predictors. Consequently, a trend in the scatter plots may not hold up in the final analysis. We will return to this point after presenting the fitted model in (1.49).

To incorporate the joint effects of the predictors, a regression model should be fitted to the data. The output for the multiple regression analysis is shown in Tables 1.6 and 1.7.

From Table 1.7, R^2 and R_a^2 can easily be computed as follows:

$$R^2 = 1 - \frac{52610}{225993} = 0.761,$$

$$R_a^2 = 1 - \frac{52610/44}{225993/58} = 0.693.$$

Table 1.6 Multiple Regression Output, Air Pollution Example

Predictor	Coefficient	Standard Error	t	p Value
Constant	1332.7	291.7	4.57	0.000
JanTemp	−2.3052	0.8795	−2.62	0.012
JulyTemp	−1.657	2.051	−0.81	0.424
RelHum	0.407	1.070	0.38	0.706
Rain	1.4436	0.5847	2.47	0.018
Education	−9.458	9.080	−1.04	0.303
PopDensi	0.004509	0.004311	1.05	0.301
%NonWhite	5.194	1.005	5.17	0.000
%WC	−1.852	1.210	−1.53	0.133
pop	0.00000109	0.00000401	0.27	0.788
pop/hous	−45.95	39.78	−1.16	0.254
income	−0.000549	0.001309	−0.42	0.677
logHC	−53.47	35.39	−1.51	0.138
logNOx	80.22	32.66	2.46	0.018
logSO2	−6.91	16.72	−0.41	0.681

Table 1.7 ANOVA Table, Air Pollution Example

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F	p Value
Regression	14	173383	12384	10.36	0.000
Residual	44	52610	1196		
Total	58	225993			

Any standard software will provide these values. The estimate s^2 of the error variance σ^2 is given by the mean square error, which, from Table 1.7, is 1196.

Keeping in mind the principle of parsimony explained in Section 1.7, the objective is now to fit a reduced model containing fewer number of variables. Table 1.6 shows that four variables, JanTemp, Rain, %NonWhite, and logNOx, are significant at the 0.05 level. However, simply retaining the significant predictors in the fitted multiple regression model may not work well if some variables are strongly correlated. We therefore employ the two variable selection techniques described in Section 1.7. Table 1.8 shows the output for the best subsets regression using BIC and C_p (only 5 out of 14 rows are shown). Each row of the table corresponds to the "best" (the one with minimum BIC or C_p) model for a fixed number of predictor variables. For example, among the $\binom{14}{4}$ subsets of predictors containing 4 variables, the one containing, coincidentally, the predictors JanTemp, Rain, %NonWhite, and logNOx has the lowest C_p value of 8.3 and the lowest BIC value of 608.29 and is considered the best.

Based on the BIC, the best model has five predictors and the lowest BIC value of 605.83. The five selected variables are: JanTemp, Rain, %NonWhite, Education, and logNOx. Use of the C_p criterion gives less conclusive choices. Although the best model with six predictors has the lowest C_p value of 3.7, it does not satisfy the criterion $C_p \approx p$ (note that p is one more than the number of predictors because it also includes the intercept term). Observe that the best model containing five variables ($p = 6$) has its C_p value (4.3) closer to p. Therefore, on the basis of the C_p criterion and the principle of parsimony, one would be inclined to choose the same five variables as the BIC. (Use of the AIC criterion for these data will be left as an exercise.)

Let us now use a stepwise regression approach to find the best model. Table 1.9 summarizes the stepwise regression output. The α values corresponding to F-to-remove and F-to-enter were both taken as 0.15 (the default value in standard statistical software). The output does not show the F-to-enter or F-to-remove value; rather, it shows the t statistic corresponding to each coefficient in the multiple regression model after inclusion or exclusion of a variable at each step. After seven steps, the stepwise method chooses a model with five predictors JanTemp, Rain, %NonWhite, Education, and logNOx. Observe that although, at the third step, logSO2 entered the model, it was dropped at the seventh step. This means that, at the third step, when the model

Table 1.8 Best Subsets Regression Using BIC and C_p, Air Pollution Example

Subset Size	R^2	R_a^2	C_p	BIC	s	Variables
4	69.7	67.4	8.3	608.29	35.62	1,4,7,13
5	72.9	70.3	4.3	605.83	34.02	1,4,5,7,13
6	74.2	71.3	3.7	606.81	33.46	1,4,6,7,8,13
7	75.0	71.6	4.3	609.16	33.29	1,4,6,7,8,12,13
8	75.4	71.5	5.4	612.18	33.32	1,2,4,6,7,8,12,13

Note: The s value refers to the square root of the mean square error [see (1.37)] of the model.

Table 1.9 Stepwise Regression Output, Air Pollution Example

Step	1	2	3	4	5	6	7
Constant	887.9	1208.5	1112.7	1135.4	1008.7	1029.5	1028.7
%NonWhite	4.49	3.92	3.92	4.73	4.36	4.15	4.15
t value	6.40	6.26	6.81	7.32	6.73	6.60	6.66
p value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Education		−28.6	−23.5	−21.1	−14.1	−15.6	−15.5
t value		−4.32	−3.74	−3.47	−2.10	−2.40	−2.49
p value		0.000	0.000	0.001	0.041	0.020	0.016
logSO2			28.0	21.0	26.8	−0.4	
t value			3.37	2.48	3.11	−0.02	
p value			0.001	0.016	0.003	0.980	
JanTemp				−1.42	−1.29	−2.15	−2.14
t value				−2.41	−2.26	−3.25	−4.17
p value				0.019	0.028	0.002	0.000
Rain					1.08	1.66	1.65
t value					2.15	3.07	3.16
p value					0.036	0.003	0.003
logNOx						42	42
t value						2.35	4.04
p value						0.023	0.000
s	48.0	42.0	38.5	37.0	35.8	34.3	34.0
R^2	41.80	56.35	63.84	67.36	69.99	72.86	72.86
R^2(adj)	40.78	54.80	61.86	64.94	67.16	69.73	70.30
C_p	55.0	29.5	17.4	12.7	9.7	6.3	4.3
BIC	634.52	621.62	614.60	612.63	611.76	609.90	605.83

consisted of two variables (%NonWhite and Education), among the remaining 12 predictor variables, inclusion of logSO2 increased the partial F by the maximum amount. Obviously, this partial F (not shown in the table) was more than the cut-off value. At the sixth step, logNOx was included in the model following exactly the same procedure. However, after running a multiple regression with six variables following the inclusion of logNOx, the t value for logSO2 drops drastically with the corresponding p value of 0.98 (see Table 1.9). This is due to a strong positive correlation between logNOx and logSO2, referred to as multicollinearity in the regression literature. Consequently, at the seventh step, the F-to-remove value for logSO2 becomes very small (again, this is not shown in the output) and results in dropping this variable. Eventually, the final model selected by stepwise regression is exactly the same as the one selected by using the C_p statistic as discussed in the preceding paragraphs.

As expected, the R^2 and R_a^2 values in Tables 1.8 and 1.9 increase as the model size increases, although R_a^2 increases more slowly. With the exception of the last column (seventh step) of Table 1.9, the values of R^2, R_a^2, BIC, C_p, and s in Table 1.8 are different from the corresponding values in Table 1.9 because the variables selected by stepwise regression are not necessarily the best (for a given model size) according to the BIC or C_p criterion. Recall that, unlike in Table 1.8, not all subsets of a given size p are searched in stepwise regression.

The final model can be obtained either directly from the last stage of the stepwise regression output, or by running a multiple regression of mortality on the five significant variables. The coefficients of these five variables in the final model will be different from the corresponding coefficients in Table 1.6, as the model now has fewer variables. The fitted model is

$$\begin{aligned}\text{MORTALITY} = 1028.67 - 2.14\ \text{JanTemp} + 1.65\ \text{Rain} - 15.54\ \text{Education}\\ + 4.15\ \%\text{NonWhite} + 41.67\ \text{logNOx},\end{aligned} \tag{1.49}$$

with an R^2 of 0.729 and an adjusted R_a^2 of 0.703.

Note that the signs of coefficients of the predictors of JanTemp, Education, %NonWhite, and logNOx in (1.49) confirm the trends in Figure 1.5. On the other hand, the apparent trend in Figure 1.5 for JulyTemp and logSO2 is not confirmed in the more rigorous regression analysis. This discrepancy can be explained by the high correlations between these two variables and the other significant variables. The observed trend in these two variables is incorporated when the other significant variables are included in the model. In the case of logSO2, it was later dropped when a highly correlated variable logNOx was included in the model (see Steps 6 and 7 of Table 1.9). This analysis demonstrates the *danger of making decisions based on simple graphical displays*.

From (1.49) one can conclude that, after adjusting for the effects of JanTemp, Rain, Education, and %NonWhite, the pollutant NOx still has a significant effect on mortality, while the other two pollutants HC and SO2 do not.

1.9 PRACTICAL SUMMARY

1. Experimental problems can be divided into five broad categories:

   (i) Treatment comparisons,
   (ii) Variable screening,
   (iii) Response surface exploration,
   (iv) System optimization,
   (v) System robustness.

2. Statistical process control tools such as control charts are often used to monitor and improve a process. If a process is stable but needs to be further improved, more active intervention like experimentation should be employed.

3. There are seven steps in the planning and implementation of experiments:

   (i) State objective,
   (ii) Choose response,

(iii) Choose factors and levels,
(iv) Choose experimental plan,
(v) Perform the experiment,
(vi) Analyze the data,
(vii) Draw conclusions and make recommendations.

4. Guidelines for choosing the response:

(i) It should help understand the mechanisms and physical laws involved in the problem.
(ii) A continuous response is preferred to a discrete response.
(iii) A good measurement system should be in place to measure the response.

5. For response optimization, there are three types of responses: nominal-the-best, larger-the-better, and smaller-the-better.

6. A cause-and-effect diagram or a flow chart should be used to facilitate the identification of potentially important factors and to provide a system view of the problem.

7. Three fundamental principles need to be considered in experimental design: replication, randomization, and blocking. Blocking is effective if the within-block variation is much smaller than the between-block variation.

8. Factors can be designated as **E** (experimental), **B** (blocking), **O** (constant level), and **R** (randomization).

9. A step-by-step introduction to simple linear regression, including estimation and hypothesis testing, is given in Sections 1.4 and 1.5. Elementary derivations are presented without the use of matrix algebra.

10. Multiple linear regression, which extends simple linear regression to any number of predictor variables, is covered in Section 1.6. General linear models and least squares estimator are presented using matrix algebra. Analysis of variance, R^2 (multiple correlation coefficient), and the extra sum of squares principle are included.

11. Variable selection in regression analysis is considered in Section 1.7. Criteria for selection include the principle of parsimony, the adjusted R_a^2, the AIC, BIC, and C_p statistics. Variable selection strategies include backward elimination, forward selection, and stepwise selection, the last one being preferred. Illustration of multiple regression and variable selection with air pollution data is given in Section 1.8.

EXERCISES

1. Use a real example to illustrate the seven-step procedure in Section 1.2.

2. Use two examples, one from manufacturing and another from the service sector, to illustrate the construction of the cause-and-effect diagram. Designate each factor on the diagram as **E, B, O**, or **R**.

3. Give examples of hard-to-change factors. How do you reconcile the hard-to-change nature of the factor with the need for randomization?

4. Identify a real-life situation (industrial, social, or even something associated with your daily life), where you may need to conduct an experiment to satisfy any of the five objectives stated in Section 1.1. Among the seven steps necessary to conduct an experiment systematically (as described in Section 1.2), illustrate as many as you can with the help of your example.

5. The prices of gasoline are on the rise once again and everyone's concern is to get as much mileage as possible from his/her automobile. Prepare a cost-and-effect diagram to list all the factors that may be responsible for high fuel consumption in your vehicle. Classify the causes under the heads MAN, MACHINE, METHOD, MATERIAL as in Figure 1.2. Which of the factors in the diagram can be classified as experimental (E) factors with respect to your routine driving process?

***6.** **(a)** For the typing experiment considered in Section 1.3, use a statistical model to quantify the gains from using randomization (as illustrated in the second sequence) and from using balance in addition to randomization.

(b) Suppose that the following sequence is obtained from using balanced randomization:

$$1.\ A, B,\quad 2.\ A, B,\quad 3.\ A, B,\quad 4.\ B, A,\quad 5.\ B, A,\quad 6.\ B, A.$$

Would you use it for the study? If not, what would you do? What aspect of the sequence makes you uneasy? Can you relate it to the possibility that the advantage of the learning effect may diminish over time and express it in more rigorous terms? (*Hint*: The terms in the model should represent the effects you identified as potentially influencing the comparison.)

7. The typing experiment can be further improved by employing more typists that are representative of the population of typists. Suppose three typists are chosen for the study. Devise an experimental plan and discuss its pros and cons. (Some of the more elaborate plans may involve strategies that will be introduced in the next chapter.)

***8.** Give an elementary proof of the variance and covariance formulas (1.14)–(1.16) by using equations (1.8)–(1.11).

9. In the winter, a plastic rain gauge cannot be used to collect precipitation data because it will freeze and crack. As a way to record snowfall, weather observers were instructed to collect the snow in a metal standard 2.5 can, allow the snow to melt indoors, pour it into a plastic rain gauge, and then record the measurement. An estimate of the snowfall is then obtained by multiplying the measurement by 0.44. (The factor 0.44 was theoretically derived as the ratio of the surface area of the rectangular opening of the rain gauge and of the circular metal can.) One observer questioned the validity of the 0.44 factor for estimating snowfall. Over one summer, the observer recorded the following rainfall data collected in the rain gauge and in the standard 2.5 can, both of which were mounted next to each other at the same height. The data (courtesy of Masaru Hamada) appear in Table 1.10, where the first column is the amount of rain (in inches) collected in the standard

Table 1.10 Rainfall Data

x	y	x	y	x	y
0.11	0.05	2.15	0.96	1.25	0.62
1.08	0.50	0.53	0.32	0.46	0.23
1.16	0.54	5.20	2.25	0.31	0.17
2.75	1.31	0.00	0.06	0.75	0.33
0.12	0.07	1.17	0.60	2.55	1.17
0.60	0.28	6.67	3.10	1.00	0.43
1.55	0.73	0.04	0.04	3.98	1.77
1.00	0.46	2.22	1.00	1.26	0.58
0.61	0.35	0.05	0.05	5.40	2.34
3.18	1.40	0.15	0.09	1.02	0.50
2.16	0.91	0.41	0.25	3.75	1.62
1.82	0.86	1.45	0.70	3.70	1.70
4.75	2.05	0.22	0.12	0.30	0.14
1.05	0.58	2.22	1.00	0.07	0.06
0.92	0.41	0.70	0.38	0.58	0.31
0.86	0.40	2.73	1.63	0.72	0.35
0.24	0.14	0.02	0.02	0.63	0.29
0.01	0.03	0.18	0.09	1.55	0.73
0.51	0.25	0.27	0.14	2.47	1.23

Note: x = amount of rain (in inches) collected in metal can, y = amount of rain (in inches) collected in plastic gauge.

2.5 can (x) and the second column is the amount of rain (in inches) collected in the rain gauge (y).

(a) Plot the residuals $y_i - 0.44x_i$ for the data. Do you observe any systematic pattern to question the validity of the formula $y = 0.44x$?

(b) Use regression analysis to analyze the data in Table 1.10 by assuming a general β_0 (i.e., an intercept term) and $\beta_0 = 0$ (i.e., regression line through the origin). How well do the two models fit the data? Is the intercept term significant?

***(c)** Because of evaporation during the summer and the can being made of metal, the formula $y = 0.44x$ may not fit the rainfall data collected in the summer. An argument can be made that supports the model with an intercept. Is this supported by your analyses in (a) and (b)?

10. The data in Table 1.11 (Weisberg, 1980, pp. 128–129) records the average weight of the brain weight (g) and body weight (kg) for 62 mammal species. The objective is to model brain weight as a function of body weight.

Table 1.11 Brain and Body Weight Data

Body Wt	Brain Wt	Body Wt	Brain Wt	Body Wt	Brain Wt
3.385	44.5	521.000	655.0	2.500	12.10
0.480	15.5	0.785	3.5	55.500	175.00
1.350	8.1	10.000	115.0	100.000	157.00
465.000	423.0	3.300	25.6	52.160	440.00
36.330	119.5	0.200	5.0	10.550	179.50
27.660	115.0	1.410	17.5	0.550	2.40
14.830	98.2	529.000	680.0	60.000	81.00
1.040	5.5	207.000	406.0	3.600	21.00
4.190	58.0	85.000	325.0	4.288	39.20
0.425	6.4	0.750	12.3	0.280	1.90
0.101	4.0	62.000	1320.0	0.075	1.20
0.920	5.7	6654.000	5712.0	0.122	3.00
1.000	6.6	3.500	3.9	0.048	0.33
0.005	0.1	6.800	179.0	192.000	180.00
0.060	1.0	35.000	56.0	3.000	25.00
3.500	10.8	4.050	17.0	160.000	169.00
2.000	12.3	0.120	1.0	0.900	2.60
1.700	6.3	0.023	0.4	1.620	11.40
2547.000	4603.0	0.010	0.3	0.104	2.50
0.023	0.3	1.400	12.5	4.235	50.40
187.100	419.0	250.000	490.0		

(a) Obtain a scatter plot of brain weight versus body weight and observe whether there is any indication of a relationship between the two. Comment on the plot.

(b) Now take a log transformation of both the variables and plot log(brain weight) against log(body weight). Does this transformation improve the relationship?

(c) Fit an appropriate regression model to express brain weight as a function of body weight. What percentage of variation in the response is explained by the model?

(d) Estimate the expected average brain weight of mammals that have an average body weight of 250 kg. Obtain a 95% confidence interval for this estimate.

11. The modern Olympic Games are an international athletic competition that has been held at a different city every four years since their inauguration in 1896, with some interruptions due to wars. The data for the gold medal performances in the men's long jump (distance in inches) is given in Table 1.12. The first column Year is coded to be zero in 1900. The performance of long jump is expected to be improved over the years.

(a) Draw a scatter plot of the response "Long Jump" against the covariate "Year." Comment on any striking features of the data by relating them to the world events of specific years.

(b) Run simple linear regression on the data. Does the linear regression model fit the data well?

(c) Conduct an F test at 0.05 level to decide if there is a linear relationship between the performance of long jump and the year of the game.

(d) Get an estimate of the mean long jump performance in year 1896, and obtain a 95% confidence interval for the estimate.

(e) Analyze the regression output. Are there any outliers in the data? If so, remove the outliers and reanalyze the data. Obtain the residual plots and take a careful look. Do they still reveal any special pattern pertaining to the record setting nature of the data?

Table 1.12 Long Jump Data

Year	Long Jump	Year	Long Jump	Year	Long Jump	Year	Long Jump
−4	249.75	24	293.13	56	308.25	80	336.25
0	282.88	28	304.75	60	319.75	84	336.25
4	289.00	32	300.75	64	317.75	88	343.25
8	294.50	36	317.31	68	350.50	92	342.50
12	299.25	48	308.00	72	324.50		
20	281.50	52	298.00	76	328.50		

12. **(a)** For the air pollution data in Section 1.8, use the AIC criterion to find the best model for subset sizes four to eight.

***(b)** Explain why the best model according to AIC is different from the best model chosen by BIC in Table 1.8. (*Hint*: Compare the second terms in (1.46) and (1.47).)

13. The data in Table 1.13 is from 1980 U.S. Census Undercount (Ericksen et al., 1989). There are 66 rows and 10 columns. The first column is the place where the data is collected. There are eight predictors:

1. *Minority*: minority percentage.
2. *Crime*: rate of serious crimes per 1000 population.
3. *Poverty*: percentage poor.
4. *Language*: percentage having difficulty speaking or writing English.
5. *Highschool*: percentage age 25 or older who had not finished high school.
6. *Housing*: percentage of housing in small, multi-unit buildings.
7. *City*: a factor with two levels: "city" (major city), "state" (state remainder).
8. *Conventional*: percentage of households counted by conventional personal enumeration.

The response is undercount (in terms of percentage). Use regression to investigate the relationship between undercount and the eight predictors.

(a) Perform regression analysis using all the predictors except city. Show the regression residuals. Which predictors seem to be important? Draw the residual plot against the fitted value. What can you conclude from this plot?

***(b)** Explain how the variable "*City*" differs from the others.

(c) Use both best subset regression and stepwise regression to select variables from all the predictors (excluding the variable "*City*"). Compare your final models obtained by the two methods.

14. For 2005, the consumption of gasoline was measured in the 50 states and the District of Columbia in the United States by the Federal Highway Administration. The response y is the consumption of gallons of gasoline per population of 16 years olds and older. Consider the following four predictor variables—x_1: state gasoline tax (cents per gallon), x_2: per capita income (1000s of dollars), x_3: paved highways (1000s of miles), and x_4: licensed drivers per 1000 persons in population of 16 years olds or older (exceeds 1000 in some states, because learner's permits are counted as licenses and learner's permit holders can be under 16 years of age). The data in Table 1.14 are derived from the tables at

http://www.fhwa.dot.gov/policy/ohim/hs05/index.htm.

Table 1.13 Ericksen Data

Place	Minority	Crime	Poverty	Language	Highschool	Housing	City	Conventional	Undercount
Alabama	26.1	49	18.9	0.2	43.5	7.6	state	0	−0.04
Alaska	5.7	62	10.7	1.7	17.5	23.6	state	100	3.35
Arizona	18.9	81	13.2	3.2	27.6	8.1	state	18	2.48
Arkansas	16.9	38	19.0	0.2	44.5	7.0	state	0	−0.74
California	24.3	73	10.4	5.0	26.0	11.8	state	4	3.60
Colorado	15.2	73	10.1	1.2	21.4	9.2	state	19	1.34
Connecticut	10.8	58	8.0	2.4	29.7	21.0	state	0	−0.26
Delaware	17.5	68	11.8	0.7	31.4	8.9	state	0	−0.16
Florida	22.3	81	13.4	3.6	33.3	10.1	state	0	2.20
Georgia	27.6	55	16.6	0.3	43.6	10.2	state	0	0.37
Hawaii	9.1	75	9.9	5.7	26.2	17.0	state	29	1.46
Idaho	4.2	48	12.6	1.0	26.3	9.1	state	56	1.53
Illinois	8.1	48	7.7	1.0	29.8	13.5	state	0	1.69
Indiana	7.1	48	9.4	0.5	33.6	9.9	state	0	−0.68
Iowa	2.3	47	10.1	0.3	28.5	10.4	state	0	−0.59
Kansas	7.9	54	10.1	0.5	26.7	8.5	state	14	0.94
Kentucky	7.7	34	17.6	0.2	46.9	10.6	state	0	−1.41
Louisiana	31.4	54	18.6	1.1	42.3	9.7	state	0	2.46
Maine	0.7	44	13.0	1.0	31.3	19.5	state	40	2.06
Maryland	16.7	58	6.8	0.8	28.2	10.5	state	0	2.03
Massachusetts	3.8	53	8.5	2.1	27.4	26.9	state	4	−0.57
Michigan	7.0	61	8.7	0.7	29.9	9.4	state	8	0.89
Minnesota	2.1	48	9.5	0.5	26.9	10.7	state	11	1.57
Mississippi	35.8	34	23.9	0.2	45.2	7.2	state	0	1.52
Missouri	7.8	45	11.2	0.3	34.9	9.1	state	0	0.81
Montana	1.5	50	12.3	0.4	25.6	12.8	state	75	1.81
Nebraska	4.8	43	10.7	0.5	26.6	9.7	state	33	0.36
Nevada	13.0	88	8.7	1.6	24.5	11.7	state	10	5.08
New Hampshire	1.0	47	8.5	0.8	27.7	20.3	state	0	−1.49
New Jersey	19.0	64	9.5	3.6	32.6	23.7	state	0	1.44
New Mexico	38.4	59	17.6	4.6	31.1	10.7	state	58	2.69
New York	8.0	48	8.9	1.3	29.3	21.6	state	0	−1.48
North Carolina	23.1	46	14.8	0.2	45.2	8.2	state	0	1.36
North Dakota	1.0	30	12.6	0.5	33.6	15.1	state	70	0.35
Ohio	8.9	52	9.6	0.5	32.1	11.3	state	0	0.97
Oklahoma	8.6	50	13.4	0.5	34.0	8.0	state	0	−0.12
Oregon	3.9	60	10.7	0.8	24.4	7.9	state	13	0.93
Pennsylvania	4.8	33	8.8	0.6	33.6	13.3	state	0	−0.78
Rhode Island	4.9	59	10.3	3.2	38.9	29.6	state	0	0.74
South Carolina	31.0	53	16.6	0.2	46.3	7.9	state	0	6.19
South Dakota	0.9	32	16.9	0.4	32.1	12.0	state	84	0.42
Tennessee	16.4	44	16.4	0.2	43.8	9.4	state	0	−2.31
Texas	30.6	55	15.0	4.7	38.7	7.7	state	1	0.27
Utah	4.7	58	10.3	0.9	20.0	11.3	state	14	1.14
Vermont	0.9	50	12.1	0.5	29.0	20.8	state	0	−1.12
Virginia	20.0	46	11.8	0.5	37.6	10.3	state	0	1.11

(*continued*)

Table 1.13 (*Continued*)

Place	Minority	Crime	Poverty	Language	Highschool	Housing	City	Conventional	Undercount
Washington	5.4	69	9.8	1.0	22.4	9.4	state	4	1.48
West Virginia	3.9	25	15.0	0.2	44.0	9.0	state	0	−0.69
Wisconsin	1.7	45	7.9	0.4	29.5	12.8	state	9	1.45
Wyoming	5.9	49	7.9	0.7	22.1	13.2	state	100	4.01
Baltimore	55.5	100	22.9	0.7	51.6	23.3	city	0	6.15
Boston	28.4	135	20.2	4.4	31.6	52.1	city	0	2.27
Chicago	53.7	66	20.3	6.7	43.8	51.4	city	0	5.42
Cleveland	46.7	101	22.1	1.6	49.1	36.4	city	0	5.01
Dallas	41.6	118	14.2	3.1	31.5	12.9	city	0	8.18
Detroit	65.4	106	21.9	1.6	45.8	18.6	city	0	4.33
Houston	45.1	80	12.7	5.1	31.6	8.9	city	0	5.79
Indianapolis	22.5	53	11.5	0.3	33.3	13.6	city	0	0.31
Los Angeles	44.4	100	16.4	12.7	31.4	15.0	city	0	7.52
Milwaukee	27.2	65	13.8	1.6	46.4	27.2	city	0	3.17
New York City	44.0	101	20.0	8.9	39.8	32.2	city	0	7.39
Philadelphia	41.3	60	20.6	2.2	45.7	21.7	city	0	6.41
Saint Louis	46.7	143	21.8	0.5	51.8	40.9	city	0	3.60
San Diego	23.6	81	12.4	4.2	21.1	11.2	city	0	0.47
San Francisco	24.8	107	13.7	9.2	26.0	20.3	city	0	5.18
Washington DC	72.6	102	18.6	1.1	32.9	21.0	city	0	5.93

Analyze the data and answer the following questions:

(a) What impression can you draw from the scatter plots regarding the variables affecting consumption of gasoline?

(b) After performing a multiple regression analysis, which predictor variables are seen to significantly affect the response?

***(c)** Provide intuitive explanations for the impact of each of the four predictor variables on the response as suggested by the multiple regression output. Try to account for any inconsistency that may be apparent.

(d) Use the following three methods to obtain the best regression model:

i. Use best subset selection using BIC and C_p.
ii. Conduct regression analysis using backward elimination.
iii. Perform regression analysis using stepwise regression.

Are the results consistent?

***(e)** Obtain the residual plots and take a careful look at them. Do they suggest any specific pattern that may provide any clue regarding the improvement of the model?

Table 1.14 Gasoline Consumption Data

x_1	x_2	x_3	x_4	y	x_1	x_2	x_3	x_4	y
18.00	27.8	96.0	1011.7	708.1	27.75	26.9	69.3	944.0	608.2
8.00	34.5	14.4	976.9	537.7	25.30	31.3	93.3	958.7	568.6
18.00	28.4	59.8	871.4	609.3	24.80	33.4	34.6	859.2	593.0
21.70	25.7	98.7	927.9	626.7	19.50	37.0	15.6	954.5	651.5
18.00	35.0	169.9	832.8	568.0	10.50	41.3	38.6	863.2	620.0
22.00	36.1	87.6	925.0	574.9	18.88	26.2	63.8	871.2	618.8
25.00	45.4	21.2	987.5	564.6	23.25	38.2	113.3	726.7	365.4
23.00	35.9	6.1	910.7	636.5	27.10	29.2	103.1	919.0	623.3
20.00	51.8	1.5	737.2	279.6	23.00	31.4	86.8	901.0	645.2
14.50	31.5	120.6	942.5	589.5	28.00	31.3	124.8	853.4	559.5
7.50	30.1	117.6	852.5	712.7	17.00	28.1	112.9	799.8	641.1
16.00	32.2	4.3	849.0	446.4	24.00	30.0	64.5	931.8	522.7
25.00	27.1	47.1	890.8	530.0	30.00	33.3	120.7	849.2	508.0
19.00	34.4	138.8	796.3	511.3	30.00	33.7	6.5	868.0	437.3
18.00	30.1	95.6	875.6	645.8	16.00	27.2	66.2	892.4	715.3
20.70	30.6	114.0	855.1	642.8	22.00	30.9	83.9	926.8	641.7
24.00	30.8	135.5	918.5	520.0	21.40	30.0	90.5	919.5	639.5
18.50	27.7	78.0	865.5	648.0	20.00	30.2	304.2	851.8	658.1
20.00	27.6	60.9	878.5	637.3	24.50	26.6	43.6	887.0	552.8
26.00	30.6	22.8	928.2	652.8	20.00	32.8	14.4	1106.9	666.2
23.50	39.2	31.0	850.8	603.3	17.50	35.5	72.0	870.2	652.0
21.00	41.8	35.9	902.5	544.4	31.00	35.3	83.4	940.2	529.9
19.00	32.0	121.5	900.0	609.9	27.00	25.9	37.0	896.8	555.0
20.00	35.9	132.0	761.0	640.8	29.90	32.2	114.1	907.4	558.1
18.40	24.7	74.2	870.3	698.0	14.00	34.3	27.7	934.0	741.1
17.00	30.6	125.8	901.5	672.3					

REFERENCES

AIAG (Automotive Industry Action Group) (1990), *Measurement Systems Analysis Reference Manual*, Troy, MI: AIAG.

Barton, R. R. (1997), Pre-experiment planning for designed experiments: graphical methods, *Journal of Quality Technology* **29**, 307–316.

Coleman, D. E., and Montgomery, D. C. (1993), A systematic approach to planning for a designed industrial experiment [with discussion], *Technometrics* **35**, 1–27.

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

Ericksen, E. P., Kadane, J. B., and Tukey, J. W. (1989), Adjusting the 1980 census of population and housing, *Journal of the American Statistical Association* **84**, 927–944.

Gibbons, D. I., and McDonald, G. C. (1980), Examining regression relationships between air pollution and mortality, General Motors Research Publication, GMR-3278, General Motors Research Laboratories, Warren, Michigan.

Hinkelmann, K., and Kempthorne, O. (1994), *Design and Analysis of Experiments*, Vol. 1, New York: John Wiley & Sons.

Knowlton, J., and Keppinger, R. (1993), The experimentation process, *Quality Progress* February, 43–47.

Lea, A. J. (1965), New observations on distribution of neoplasms of female breast in certain countries, *British Medical Journal* **1**, 488–490.

León, R. V., Shoemaker, A. C., and Tsui, K. L. (1993), Discussion of "A systematic approach to planning for a designed industrial experiment" by Coleman, D. E., and Montgomery, D. C., *Technometrics* **35**, 21–24.

Mallows, C. L. (1973), Some comments on C_p, *Technometrics* **15**, 661–676.

McQuarrie, A. D. R., and Tsai, C.-L. (1998), *Regression and Time Series Model Selection*, River Edge, NJ: World Scientific.

Phadke, M. S. (1989), *Quality Engineering Using Robust Design*, Englewood Cliffs, NJ: Prentice-Hall.

Rosenbaum, P. R. (2002), *Observational Studies*, 2nd ed., New York: Springer-Verlag.

Rosenberger W. F., and Lachin, J. M. (2002), *Randomization in Clinical Trials: Theory and Practice*, New York: John Wiley & Sons.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*, New York: Springer.

Weisberg, S. (1980), *Applied Linear Regression*, New York: John Wiley & Sons.

Wherry, R. J. (1931), A new formula for predicting the shrinkage of the coefficient of multiple correlation, *The Annals of Mathematical Statistics* **2**, 440–457.