CHAPTER 13

Discrimination and Classification

13.1 INTRODUCTION

Discrimination or classification methods attempt to use measured characteristics to divide people or objects into prespecified groups. As in regression modeling for prediction in Chapter 11, the criteria for assessing classification models are accuracy of prediction and possibly cost of measuring the relevant characteristics. There need not be any relationship between the model and the actual causal processes involved. The computer science literature refers to classification as *supervised learning*, as distinguished from *cluster analysis* or *unsupervised learning*, in which groups are not prespecified and must be discovered as part of the analysis. We discuss cluster analysis briefly in Note 13.5.

In this chapter we discuss the general problem of classification. We present two simple techniques, logistic and linear discrimination, and discuss how to choose and evaluate classification models. Finally, we describe briefly a number of more modern classification methods and give references for further study.

13.2 CLASSIFICATION PROBLEM

In the classification problem we have a group variable Y for each individual, taking values $1, 2, \ldots, K$, called *classes*, and a set of characteristics $X_1, X_2, \ldots, X_p$. Both X and Y are observed for a *training set* of data, and the goal is to create a rule to predict Y from X for new observations and to estimate the accuracy of these predictions.

The most common examples of classification problems in biostatistics have just two classes: with and without a given disease. In screening and diagnostic testing, the classes are based on whether the disease is currently present; in prognostic models, the classes are those who will and will not develop the disease over some time frame.

For example, the Framingham risk score [Wilson et al., 1998] is used widely to determine the probability of having a heart attack over the next 10 years based on blood pressure, age, gender, cholesterol levels, and smoking. It is a prognostic model used in screening for heart disease risk, to help choose interventions and motivate patients. Various diagnostic classification rules also exist for coronary heart disease. A person presenting at a hospital with chest pain may be having a heart attack, in which case prompt treatment is needed, or may have muscle strain or indigestion-related pain, in which case the clot-dissolving treatments used for heart attacks would be unnecessary and dangerous. The decision can be based on characteristics of the pain,

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2 Copyright © 2004 John Wiley & Sons, Inc.

blood enzyme levels, and electrocardiogram abnormalities. Finally, for research purposes it is often necessary to find cases of heart attack from medical records. This retrospective diagnosis can use the same information as the initial diagnosis and later follow-up information, including the doctors' conclusions at the time of discharge from a hospital.

It is useful to separate the classification problem into two steps:

1. Estimate the probability p_k that $Y = k$.
2. Choose a predicted class based on these probabilities.

It might appear that the second step is simply a matter of choosing the most probable class, but this need not be the case when the consequences of making incorrect decisions depend on the decision. For example, in cancer screening a *false positive*, calling for more investigation of what turns out not to be cancer, is less serious than a *false negative*, missing a real case of cancer. About 10% of women are recalled for further testing after a mammogram [Health Canada, 2001], but the great majority of these are false positives and only 6 to 7% of these women are diagnosed with cancer.

The consequences of misclassification can be summarized by a *loss function* $L(j, k)$, which gives the relative seriousness of choosing class j when in fact class k is the correct one. The loss function is defined to be zero for a correct decision and positive for incorrect decisions. If $L(j, k)$ has the same value for all incorrect decisions, the correct strategy is to choose the most likely class. In some cases these losses might be actual monetary costs; in others the losses might be probabilities of dying as a result of the decision, or something less concrete. What the theory requires is that a loss of 2 is twice as bad as a loss of 1. In Note 13.3 we discuss some of the practical and philosophical issues involved in assigning loss functions.

Finally, the expected proportion in each class may not be the same in actual use as in training data. This imbalance may be deliberate: If some classes are very rare, it will be more efficient if they are overrepresented in the training data. The imbalance may also be due to a variation in frequency of classes between different times or places; for example, the relative frequency of common cold and influenza will depend on the season. We will write π_k for the expected proportion in class k if it is specified separately from the training data. These are called *prior probabilities*.

Given a large enough training set, the classification problem is straightforward (assume initially that we do not have separately specified proportions π_k). For any new observations with characteristics $x_1, \ldots, x_p$, we find all the observations in the training set that have exactly the same characteristics and estimate p_k, the probability of being in class k, as the proportion of these observations that are in class k.

Now that we have probabilities for each class k, we can compute the expected loss for each possible decision. Suppose that there are two classes and we decide on class 1. The probability that we are correct is p_1, in which case there is no loss. The probability that we are incorrect is p_2, in which case the loss is $L(1, 2)$. So the expected loss is $0 \times p_1 + L(1, 2) \times p_2$. Conversely, if we decide on class 2, the expected loss is $L(2, 1) \times p_1 + 0 \times p_2$. We should choose whichever class has the lower expected loss. Even though we are assuming unlimited amounts of training data, the expected loss will typically not be zero. Problems where the loss can be reduced to zero are called *noiseless*. Medical prediction problems are typically very noisy.

Bayes' theorem, discussed in Chapter 6, now tells us how to incorporate separately specified expected proportions (*prior probabilities*) into this calculation: We simply multiply p_1 by π_1, p_2 by π_2, and so on. The expected loss from choosing class 1 is $0 \times p_1 \times \pi_1 + L(1, 2) \times p_2 \times \pi_2$.

Classification is more difficult when we do not have enough training data to use this simple approach to estimation, or when it is not feasible to keep the entire training set available for making predictions. Unfortunately, at least one of these limitations is almost always present. In this chapter we consider only the first problem, the most important in biostatistical applications. It is addressed by building regression models to estimate the probabilities p_k and then following the same strategy as if p_k were known. The accuracy of prediction, and thus the actual average loss, will be greater than in our ideal setting. The error rates in the ideal setting give a lower bound on the error rates attainable by any model; if these are low, improving a model may have a large payoff; if they are high, no model can predict well and improvements in the model may provide little benefit in error rates.

13.3 SIMPLE CLASSIFICATION MODELS

Linear and logistic models for classification have a long history and often perform reasonably well in clinical and epidemiologic classification problems. We describe them for the case of two classes, although versions for more than two classes are available. Linear and logistic discrimination have one important restriction in common: They separate the classes using a linear combination of the characteristics.

13.3.1 Logistic Regression

Example 13.1. Pine et al. [1983] followed patients with intraabdominal sepsis (blood poisoning) severe enough to warrant surgery to determine the incidence of organ failure or death (from sepsis). Those outcomes were correlated with age and preexisting conditions such as alcoholism and malnutrition. Table 13.1 lists the patients with the values of the associated variables. There are 21 deaths in the set of 106 patients. Survival status is indicated by the variable Y. Five potential predictor variables: shock, malnutrition, alcoholism, age, and bowel infarction were labeled X_1, X_2, X_3, X_4, and X_5, respectively. The four variables X_1, X_2, X_3, and X_5 were binary variables, coded 1 if the symptom was present and 0 if absent. The variable $X_4 =$ age in years, was retained as a continuous variable. Consider for now just variables Y and X_1; a 2×2 table could be formed as shown in Table 13.2.

With this single variable we can use the simple approach of matching new observations exactly to the training set. For a patient with shock, we would estimate a probability of death of $7/10 = 0.70$; for a patient without shock, we would estimate a probability of $14/96 = 0.15$.

Once we start to incorporate the other variables, this simple approach will break down. Using all four binary variables would lead to a table with 2^5 cells, and each cell would have too few observations for reliable estimates. The problem would be enormously worse when age is added to the model—there might be no patient in our training set who was an exact match on age.

We clearly need a way to simplify the model. One approach is to assume that to a reasonable approximation, the effect of one variable does not depend on the values of other variables, leading to a linear regression model:

$$P(\text{death}) = \pi = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_5 X_5$$

This model is unlikely to be ideal: If having shock increases the risk of death by 0.55, and the probability can be no larger than 1, the effects of other variables are severely limited. For this reason it is usual to transform the probability to a scale that is not limited by 0 and 1.

The most common reexpression of π leads to the logistic model

$$\log_e \frac{\pi}{1-\pi} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_5 X_5 \qquad (1)$$

commonly written as

$$\text{logit}(\pi) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_5 X_5 \qquad (2)$$

Table 13.1 Survival Status of 106 Patients Following Surgery and Associated Preoperative Variables[a]

ID	Y	X_1	X_2	X_3	X_4	X_5	ID	Y	X_1	X_2	X_3	X_4	X_5
1	0	0	0	0	56	0	301	1	0	1	0	50	1
2	0	0	0	0	80	0	302	0	0	0	0	20	0
3	0	0	0	0	61	0	303	0	0	0	0	74	1
4	0	0	0	0	26	0	304	0	0	0	0	54	0
5	0	0	0	0	53	0	305	1	0	1	0	68	0
6	1	0	1	0	87	0	306	0	0	0	0	25	0
7	0	0	0	0	21	0	307	0	0	0	0	27	0
8	1	0	0	1	69	0	308	0	0	0	0	77	0
9	0	0	0	0	57	0	309	0	0	1	0	54	0
10	0	0	1	0	76	0	401	0	0	0	0	43	0
11	1	0	0	1	66	1	402	0	0	1	0	27	0
12	0	0	0	0	48	0	501	1	0	1	1	66	1
13	0	0	0	0	18	0	502	0	0	1	1	47	0
14	0	0	0	0	46	0	503	0	0	0	1	37	0
15	0	0	1	0	22	0	504	0	0	1	0	36	1
16	0	0	1	0	33	0	505	1	1	1	0	76	0
17	0	0	0	0	38	0	506	0	0	0	0	33	0
19	0	0	0	0	27	0	507	0	0	0	0	40	0
20	1	1	1	0	60	1	508	0	0	1	0	90	0
22	0	0	0	0	31	0	510	0	0	0	1	45	0
102	0	0	0	0	59	1	511	0	0	0	0	75	0
103	0	0	0	0	29	0	512	1	0	0	1	70	1
104	0	1	0	0	60	0	513	0	0	0	0	36	0
105	1	1	0	0	63	1	514	0	0	0	1	57	0
106	0	0	0	0	80	0	515	0	0	1	0	22	0
107	0	0	0	0	23	0	516	0	0	0	0	33	0
108	0	0	0	0	71	0	518	0	0	1	0	75	0
110	0	0	0	0	87	0	519	0	0	0	0	22	0
111	1	1	1	0	70	0	520	0	0	1	0	80	0
112	0	0	0	0	22	0	521	1	0	1	0	85	0
113	0	0	0	0	17	0	523	0	0	1	0	90	0
114	1	0	0	1	49	0	524	1	0	0	1	71	0
115	0	1	0	0	50	0	525	0	0	0	1	51	0
116	0	0	0	0	51	0	526	1	0	1	1	67	0
117	0	0	1	1	37	0	527	0	0	1	0	77	0
118	0	0	0	0	76	0	529	0	0	0	0	20	0
119	0	0	0	1	60	0	531	0	0	0	0	52	1
120	1	1	0	0	78	1	532	1	1	0	1	60	0
122	0	0	1	1	60	0	534	0	0	0	0	29	0
123	1	1	1	0	57	0	535	0	0	0	0	30	1
202	0	0	0	0	28	1	536	0	0	0	0	20	0
203	0	0	0	0	94	0	537	0	0	0	0	36	0
204	0	0	0	0	43	0	538	0	0	1	1	54	0
205	0	0	0	0	70	0	539	0	0	0	0	65	0
206	0	0	0	0	70	0	540	1	0	0	0	47	0
207	0	0	0	0	26	0	541	0	0	0	0	22	0
208	0	0	0	0	19	0	542	1	0	0	1	69	0
209	0	0	0	0	80	0	543	1	0	1	1	68	0
210	0	0	1	0	66	0	544	0	0	1	1	49	0
211	0	0	1	0	55	0	545	0	0	0	0	25	0
214	0	0	0	0	36	0	546	0	1	1	0	44	0
215	0	0	0	0	28	0	549	0	0	0	1	56	0
217	0	0	0	0	59	1	550	0	0	1	1	42	0

Source: Data from Pine et al. [1983].

[a] See the text for labels.

Table 13.2 2 x 2 Table for Survival by Shock Status

		Y		
		Death	Survive	
X_1		1	0	
Shock	1	7	3	10
No Shock	0	14	82	96
		21	85	106

Four comments are in order:

1. The logit of p has range $(-\infty, \infty)$. The following values can easily be calculated:

$$\text{logit}(1) = +\infty$$

$$\text{logit}(0) = -\infty$$

$$\text{logit}(0.5) = 0$$

2. If we solve for π, the expression that results is

$$\pi = \frac{e^{\alpha+\beta_1X_1+\cdots+\beta_5X_5}}{1+e^{\alpha+\beta_1X_1+\cdots+\beta_5X_5}} = \frac{1}{1+e^{-(\alpha+\beta_1X_1+\cdots+\beta_5X_5)}} \tag{3}$$

3. We will write a for the estimate of α, b_1 for the estimate of β_1, and so on. Our estimated probability of death is obtained by inserting these values into equation (3) to get

$$\widehat{P}(\text{death}) = a + b_1X_1 + b_2X_2 + \cdots + b_5X_5$$

4. The estimates are obtained by *maximum likelihood*. That is, we choose the values of a, b_1, b_2, ..., b_5 that maximize the probability of getting the death and survival values that we observed. In the simple situation where we can estimate a probability for each possible combination of characteristics, maximum likelihood gives the same answer as our rule of using the observed proportions. Note 13.1 gives the mathematical details. Any general-purpose statistical program will perform logistic regression.

We can check that with a single variable, logistic regression gives the same results as our previous analysis. In the previous analysis we used only the variable X_1, the presence of shock. If we fit this model to the data, we get

$$\text{logit}(\widehat{\pi}) = -1.768 + 2.615X_1$$

If $X_1 = 0$ (i.e., there is no shock),

$$\text{logit}(\widehat{\pi}) = -1.768$$

or

$$\widehat{\pi} = \frac{1}{1+e^{-(-1.768)}} = 0.146$$

If $X_1 = 1$ (i.e., there is shock),

$$\text{logit}(\widehat{\pi}) = -1.768 + 2.615 = 0.847$$
$$\widehat{\pi} = \frac{1}{1 + e^{-0.847}} = 0.700$$

This is precisely the probability of death given no preoperative shock. The coefficient of X_1, 2.615, also has a special interpretation: It is the logarithm of the odds ratio and the quantity $e^{b_1} = e^{2.615} = 13.7$ is the odds ratio associated with shock (as compared to no shock). This can be shown algebraically to be the case (see Problem 13.1).

Example 13.1. (*continued*) We now continue the analysis of the data of Pine et al. listed in Table 13.1. The output and calculations shown in Table 13.3 can be generated for all the variables. We would interpret these results as showing that in the presence of the remaining variables, malnutrition, is not an important predictor of survival status. All the other variables are significant predictors of survival status. All but variable X_4 are discrete binary variables. If malnutrition is dropped from the analysis, the estimates and standard errors are as given in Table 13.4.

If $\widehat{\pi}$ is the predicted probability of death, the equation is

$$\text{logit}(\widehat{\pi}) = -8.895 + 3.701X_1 + 3.186X_3 + 0.08983X_4 + 2.386X_5$$

For each of the values of X_1, X_3, X_5 (a total of eight possible combinations), a regression curve can be drawn for logit($\widehat{\pi}$) vs. age. In Figure 13.1 the lines are drawn for each of the eight combinations. For example, corresponding to $X_1 = 1$ (shock present), $X_3 = 0$ (no alcoholism), and $X_5 = 0$ (no infarction), the line

Table 13.3 Logistic Regression for Example 13.1

Variable	Regression Coefficient	Standard Error	Z-Value	p-Value
Intercept	−9.754	2.534	—	—
X_1 (shock)	3.674	1.162	3.16	0.0016
X_2 (malnutrition)	1.217	0.7274	1.67	0.095
X_3 (alcoholism)	3.355	0.9797	3.43	0.0006
X_4 (age)	0.09215	0.03025	3.04	0.0023
X_5 (infarction)	2.798	1.161	2.41	0.016

Table 13.4 Estimates and Standard Errors for Example 13.1

Variable	Regression Coefficient	Standard Error
Intercept	−8.895	2.314
X_1 (shock)	3.701	1.103
X_3 (alcoholism)	3.186	0.9163
X_4 (age)	0.08983	0.02918
X_5 (infarction)	2.386	1.071

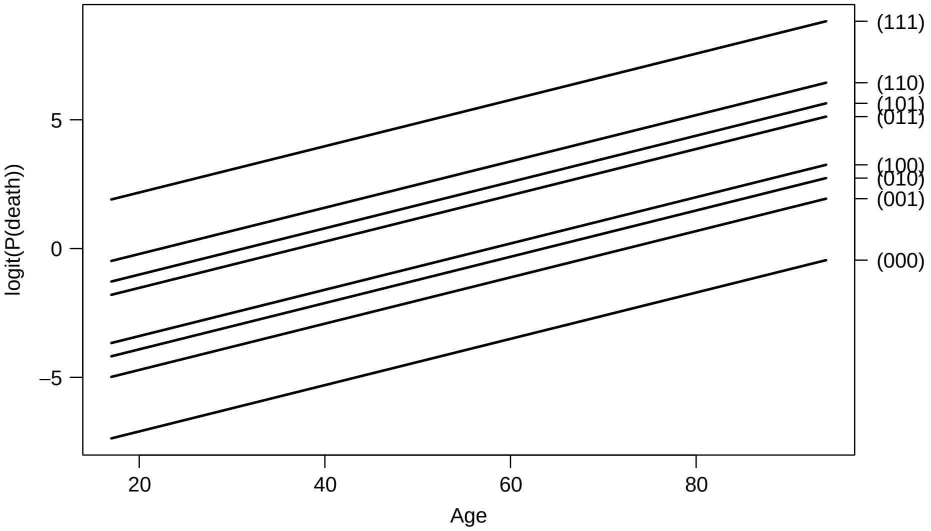


Figure 13.1 Logit of estimated probability of death as a function of age in years and category of status of (X_1, X_3, X_5). (Data from Pine et al. [1983].)

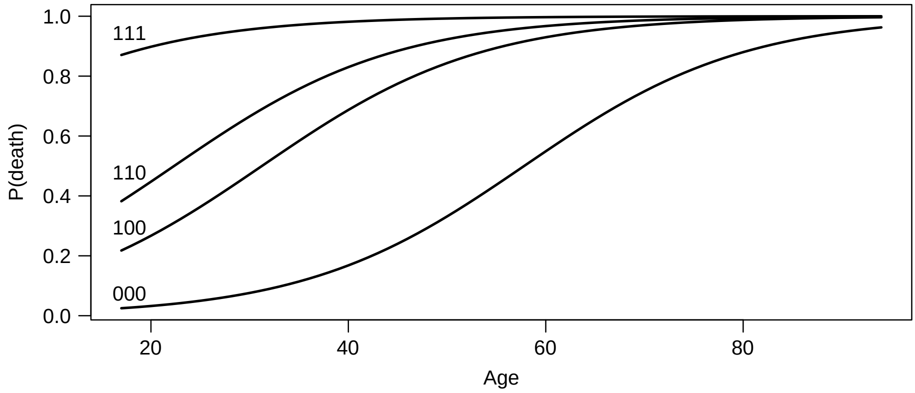


Figure 13.2 Estimated probability of death as a function of age in years and selected values of (X_1, X_3, X_5). (Data from Pine et al. [1983].)

$$\begin{aligned}\text{logit}(\widehat{\pi}) &= -8.895 + 3.701 + 0.08983X_4\\ &= -5.194 + 0.08983X_4\end{aligned}$$

is drawn.

This line is indicated by "(100)" as a shorthand way of writing $(X_1 = 1, X_3 = 0, X_5 = 0)$. The eight lines seem to group themselves into four groups: the top line representing all three symptoms present; the next three lines, groups with two symptoms present; the next three lines, groups with one symptom present; and finally, the group at lowest risk with no symptoms present. In Figure 13.2 the probability of death is plotted on the original probability scale; only four of the eight groups have been graphed. The group at highest risk is the one with all three binary risk factors present. One of the advantages of the model is that we can draw a curve for the situation with all three risk factors present even though there are no patients in that category; but the estimate depends on the model. The curve is drawn on the assumption that the risks are additive in the logistic scale (that is what we *mean* by a linear model). This assumption can be partially tested by including interaction terms involving these three covariates in the model and testing their significance. When this was done, none of the interaction terms were significant, suggesting that the additive model is a reasonable one. Of course, as there are no patients with all three risk factors present, there is no way to perform a complete test of the model.

13.3.2 Linear Discrimination

The first statistical approach to classification, as with so many other problems, was invented by R. A. Fisher. Fisher's linear discriminant analysis is designed for continuous characteristics that have a normal distribution (in fact, a multivariate normal distribution; any sums or differences of multiples of the variables should be normally distributed).

Definition 13.1. A set of random variables $X_1, \ldots, X_k$ is *multivariate normal* if every linear combination of $X_1, \ldots, X_k$ has a normal distribution.

In addition, we assume that the variances and covariances of the characteristics are the same in the two groups. Under these assumptions, Fisher's method finds a combination of variables (a *discriminant function*) for distinguishing the classes:

$$\Delta = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Assuming equal losses for different errors, an observation is assigned to class 1 if $\Delta > 0$ and class 2 if $\Delta < 0$. Estimation of the parameters β again uses maximum likelihood. It is also possible to compute probabilities p_k for membership of each class using the normal cumulative distribution function: $p_1 = \Phi(\Delta)$, $p_2 = 1 - \Phi(\Delta)$, where Φ is the symbol for the cumulative normal distribution.

Because linear discrimination makes more assumptions about the structure of the X's than logistic regression does, it gives more precise estimates of its parameters and more precise predictions [Efron, 1975]. However, in most medical examples the uncertainty in the parameters is a relatively small component of the overall prediction error, compared to model uncertainty and to the inherent unpredictability of human disease. In addition to requiring extra assumptions to hold, linear discrimination is likely to give substantial improvements only when the characteristics determine the classes very accurately so that the main limitation is the accuracy of statistical estimation of the parameters (i.e., a nearly "noiseless" problem).

The robustness can be explained by considering another equivalent way to define Δ. Let D_1 and D_2 be the mean of Δ in groups 1 and 2, respectively, and V be the variance of Δ within each group (assumed to be the same). Δ is the linear combination that maximizes

$$\frac{(D_1 - D_2)^2}{V}$$

the ratio of the between-group and within-group variances.

Truett et al. [1967] applied discriminant analysis to the data of the Framingham study. This was a longitudinal study of the incidence of coronary heart disease in Framingham, Massachusetts. In their prediction model the authors used continuous variables such as age (years) and serum cholesterol (mg/100 mL) as well as discrete or categorical variables such as cigarettes per day (0 = never smoked, 1 = less than one pack a day, 2 = one pack a day, 3 = more than a pack a day) and ECG (0 = normal, 1 = certain kinds of abnormality). It was found that the linear discriminant model gave reasonable predictions. Halperin [1971] came to five conclusions, which have stood the test of time. If the logistic model holds but the normality assumptions for the predictor variables are violated, they concluded that:

1. β_i that are zero will tend to be estimated as zero for large samples by the method of maximum likelihood but not necessarily by the discrimination function method.
2. If any β_i are nonzero, they will tend to be estimated as nonzero by either method, but the discriminant function approach will give asymptotically biased estimates for those β_i and for α.
3. Empirically, the assessment of significance for a variable, as measured by the ratio of the estimated coefficient to its estimated standard error, is apt to be about the same whichever method is used.
4. Empirically, the maximum likelihood method usually gives slightly better fits to the model as evaluated from observed and expected numbers of cases per decile of risk.
5. There is a theoretical basis for the possibility that the discriminant function will give a very poor fit even if the logistic model holds.

Some of these empirical conclusions are supported theoretically by Li and Duan [1989] and Hall and Li [1993], who considered situations similar to this one, where a linear combination

$$\Delta = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

is to be estimated under either of two models. They showed that under some assumptions about the distribution of variables X, using the wrong model would typically lead to estimating

$$\Delta = c\beta_1 X_1 + c\beta_2 X_2 + \cdots + c\beta_p X_p$$

for some constant c. When these conditions apply, using linear discrimination would tend to lead to a similar discriminant function Δ but to poor estimation of the actual class probabilities. See also Knoke [1982]. Problems 13.4, 13.6, and 13.7 address some of these issues.

In the absence of software specifically designed for this method, linear discrimination can be performed with software for linear regression. The details, which are of largely historical interest, are given in Note 13.4.

13.4 ESTIMATING AND SUMMARIZING ACCURACY

When choosing between classification models or describing the performance of a model, it is necessary to have some convenient summaries of the error rates. It is usually important to distinguish between different kinds of errors, although occasionally a simple estimate of the expected loss will suffice.

Statistical methodology is most developed for the case of two classes. In biostatistics, these are typically presence and absence of disease.

13.4.1 Sensitivity and Specificity

In assigning people to two classes (disease and no disease) we can make two different types of error:

1. Detecting disease when none is present
2. Missing disease when it is there

As in Chapter 6, we define the *sensitivity* as the probability of detecting disease given that disease is present (avoiding an error of the first kind) and *specificity* as the probability of not detecting disease given that no disease is present (avoiding an error of the second kind).

The sensitivity and specificity are useful because they can be estimated from separate samples of persons with and without disease, and because they often generalize well between populations. However, in actual use of a classification rule, we care about the probability that a person has disease given that disease was detected (the *positive predictive value*) and the probability that a person is free of disease given that no disease was detected (the *negative predictive value*).

It is a common and serious error to confuse the sensitivity and the positive predictive value. In fact, for a reasonably good test and a rare disease, the positive predictive value depends almost entirely on the disease prevalence and on the specificity. Consider the mammography example mentioned in Section 13.2. Of 1000 women who have a mammogram, about 100 will be recalled for further testing and 7 of those will have cancer. The positive predictive value is 7%, which is quite low, not because the sensitivity of the mammogram is poor but because 93 of those 1000 women are falsely testing positive. Because breast cancer is rare, false positives greatly outnumber true positives, regardless of how sensitive the test is.

When a single binary characteristic is all that is available, the sensitivity and specificity describe the properties of the classification rule completely. When classification is based on a summary criterion such as the linear discriminant function, it is useful to consider the sensitivity and specificity based on a range of possible thresholds.

Example 13.2. Tuberculosis testing is important in attempts to control the disease, which can be quite contagious but in most countries is still readily treatable with a long course of antibiotics. Tests for tuberculosis involve injecting a small amount of antigen under the skin and looking for an inflamed red area that appears a few days later, representing an active T-cell response to the antigen. The size of this indurated area varies from person to person both because of variations in disease severity and because of other individual factors. Some people with HIV infection have no reaction even with active tuberculosis (a state called *anergy*). At the other extreme, migrants from countries where the BCG vaccine is used will have a large response irrespective of their actual disease status (and since the vaccine is incompletely effective, they may or may not have disease).

The diameter of the indurated area is used to classify people as disease-free or possibly infected. It is important to detect most cases of TB (high sensitivity) without too many false positives being subjected to further investigation and unnecessary treatment (high positive predictive value). The diameter used to make the classification varies depending on characteristics of the patient. A 5-mm induration is regarded as positive for close contacts of people with active TB infection or those with chest x-rays suggestive of infection because the prior probability of risk is high. A 5-mm induration is also regarded as positive for people with compromised immune systems due to HIV infection or organ transplant, partly because they are likely to have weaker T-cell responses (so a lower threshold is needed to maintain sensitivity) and partly because TB is much more serious in these people (so the loss for a false negative is higher).

For people at moderately high risk because they are occupationally at higher risk or because they come from countries where TB is common, a 10-mm induration is regarded as positive (their prior probability is moderately elevated). The 10-mm rule is also used for people with poor access to health care or those with diseases that make TB more likely to become active (again, the loss for a false negative is higher in these groups).

Finally, for everyone else, a 15-mm threshold is used. In fact, the recommendation is that they typically not even be screened, implicitly classifying everyone as negative.

Given a continuous variable predicting disease (whether an observed characteristic or a summary produced by logistic or linear discrimination), we would like to display the sensitivity and specificity not just for one threshold but for all possible thresholds. The *receiver operating characteristic* (ROC) *curve* is such a display. It is a graph with "sensitivity" on the y-axis and "$1 -$ specificity" on the x-axis, evaluated for each possible threshold.

If the variable is completely independent of disease, the probability of detecting disease will be the same for people with and without disease, so "sensitivity" and "$1 -$ specificity"

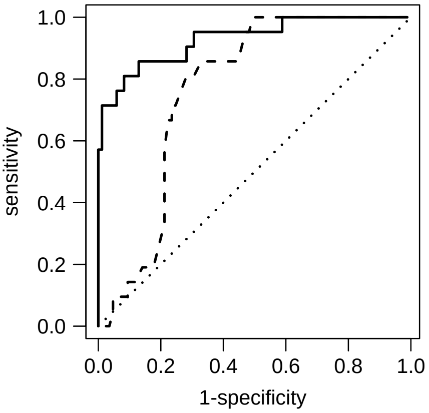


Figure 13.3 Receiver operating characteristic curve for data of Pine et al. [1983]. The solid line is the prediction from all five variables; the dashed line is the prediction from age alone.

will be the same. This is indicated by a diagonal line in Figure 13.3. If higher values of the variable are associated with higher risks of disease, the curve will lie above the diagonal line. By convention, if lower values of the variable are associated with higher risks of disease, the variable is transformed to reverse this, so ROC curves should always lie above the diagonal line.

The area under the ROC curve is a measure of how well the variable discriminates a disease state: If you are given one randomly chosen person with disease and one randomly chosen person without disease, the area under the ROC curve is the probability that the person with disease has the higher value of the variable. The area under the ROC curve is a good analog for binary data of the r^2 value for linear models.

Drawing the ROC curve for two classification rules allows you to compare their accuracy at a range of different thresholds. It might be, for example, that two rules have very different sensitivity when their specificity is low but very similar sensitivity when their specificity is high. In that case, the rules would be equivalently useful in screening low-risk populations, where specificity must be high, but might be very different in clinical diagnostic use.

13.4.2 Internal and External Error Rates

The *internal* or *apparent* or *training* or *in-sample error rates* are those obtained on the same data as those used to fit the model. These always underestimate the true error rate, sometimes very severely. The underestimation becomes more severe when many characteristics are available for modeling, when the model is very flexible in form, and when the data are relatively sparse.

An extreme case is given by a result from computer science called the *perceptron capacity bound* [Cover, 1965]. Suppose that there are d continuous characteristics and n observations from two classes in the training set, and suppose that the characteristics are purely random, having no real association whatsoever with the classes. The probability of obtaining an in-sample error rate of zero for some classification rule based on a single linear combination of characteristics is then approximately

$$1 - \Phi\left(\frac{n-2d}{\sqrt{n}}\right)$$

If d is large and $n/d < 2$, this probability will be close to 1. Even without considering nonlinear models and interactions between characteristics, it is quite possible to obtain an apparent error rate of zero for a model containing no information whatsoever. Note that $n/d > 2$ does not guarantee a good in-sample estimate of the error rate; it merely rules out this worst possible case.

Estimates of error rates are needed for model selection and in guiding the use of classification models, so this is a serious problem. The only completely reliable solution is to compute the error rate on a completely new sample of data, which is often not feasible.

When no separate set of data will be available, there are two options:

1. Use only part of the data for building the model, saving out some data for testing.
2. Use all the data for model building and attempt to estimate the true error rate statistically.

Experts differ on which of these is the best strategy, although the majority probably leans toward the second strategy. The first strategy has the merit of simplicity and requires less programming expertise. We discuss one way to estimate the true error rate, cross-validation, and one way to choose between models without a direct error estimate, the Akaike information criterion.

13.4.3 Cross-Validation

Statistical methods to estimate true error rate are generally based on the idea of refitting a model to part of the data and using the refitted model to estimate the error rate on the rest of the data. Refitting the model is critical so that the data left out are genuinely independent of the model fit. It is important to note that refitting ideally means redoing the entire model selection process, although this is feasible only when the process was automated in some way.

In *10-fold cross-validation*, the most commonly used variant, the data are randomly divided into 10 equal pieces. The model is then refitted 10 times, each time with one of the 10 pieces left out and the other nine used to fit the model. The classification errors (either the expected loss or the false positive and false negative rates) are estimated for the left-out data from the refitted model. The result is an estimate of the true error rate, since each observation has been classified using a model fitted to data not including that observation. Clearly, 10-fold cross-validation takes 10 times as much computer time as a single model selection, but with modern computers this is usually negligible. Cross-validation gives an approximately unbiased estimate of the true error rate, but a relatively noisy one.

13.4.4 Akaike's Information Criterion

Akaike's information criterion (AIC) [Akaike, 1973] is an asymptotic estimate of expected loss for a particular loss function, one that is proportional to the logarithm of the likelihood. It is extremely simple to compute but can only be used for models fitted by maximum likelihood and requires great caution when used to compare models fitted by different modeling techniques. In the case of linear regression, model selection with AIC is equivalent to model selection with Mallow's C_p, discussed in Chapter 11, so it can be seen as a generalization of Mallow's C_p to nonlinear models.

The primary difficulty in model selection is that increasing the number of variables always decreases the apparent error rate even if the variables contain no useful information. The AIC is based on the observation that for one particular loss function, the log likelihood, the decrease depends only on the number of variables added to the model. If a variable is uninformative, it will on average increase the log likelihood by 1 unit. When comparing model A to model B, we can compute

$$\begin{gathered}\log(\text{likelihood of A}) - \log(\text{likelihood of B}) \\ -(\text{no. parameters in A} - \text{no. parameters in B}) \qquad (4)\end{gathered}$$

If this is positive, we choose model A, if it is negative we choose model B. The AIC is most often defined as

$$\text{AIC} = -2\ \log(\text{likelihood of model}) + 2(\text{no. parameters in model}) \qquad (5)$$

so that choosing the model with the lower AIC is equivalent to our strategy based on equation (4). Sometimes the AIC is defined without the factor of -2, in which case the largest value indicates the best model: It is important to check which definition is being used.

Akaike showed that given two fixed models and increasing amounts of data, this criterion would eventually pick the best model. When the number of candidate models is very large, like the 2^p models in logistic regression with p characteristics, AIC still tends to overfit to some extent. That is, the model chosen by the AIC tends to have more variables than the best model.

In principle, the AIC can be used to compare models fitted by different techniques, but caution is needed. The log likelihood is only defined up to adding or subtracting an arbitrary constant, and different programs or different procedures within the same program may use different constants for computational convenience. When comparing models fitted by the same procedure, the choice of constant is unimportant, as it cancels out of the comparison. When comparing models fitted by different procedures, the constant does matter, and it may be difficult to find out what constant has been used.

13.4.5 Automated Stepwise Model Selection

Automated stepwise model selection has a deservedly poor reputation when the purpose of a model is causal inference, as model choice should then be based on a consideration of the probable cause-and-effect relationships between variables. When modeling for prediction, however, this is unimportant: We do not need to know *why* a variable is predictive to know that it *is* predictive.

Most statistical packages provide tools that will automatically consider a set of variables and attempt to find the model that gives the best prediction. Some of these use AIC, but more commonly they use significance testing of predictors. Stepwise model selection based on AIC can be approximated by significance-testing selection using a critical p-value of 0.15.

Example 13.3. We return to the data of Pine et al. [1983] and fit a logistic model by stepwise search, optimizing the AIC. We begin with a model using none of the characteristics and giving the same classification for everyone. Each of the five characteristics is considered for adding to the model, and the one optimizing the AIC is chosen. At subsequent steps, every variable is considered either for adding to the model or for removal from the model. The procedure stops when no change improves the AIC.

This procedure is not guaranteed to find the best possible model but can be carried out much more quickly than an exhaustive search of all possible models. It is at least as good as, and often better than, forward or backward stepwise procedures that only add or only remove variables.

Starting with an empty model the possible changes were as follows:

	d.f.	Deviance	AIC		d.f.	Deviance	AIC
+ X4	1	90.341	94.341	+ X5	1	97.877	101.877
+ X1	1	91.977	95.977	+ X2	1	99.796	103.796
+ X3	1	95.533	99.533	<none>		105.528	107.528

The d.f. column counts the number of degrees of freedom for each variable (in this case, one for each variable, but more than one if a variable had multiple categories). The deviance is -2 log likelihood. The best (lowest AIC) choice was to add X4 (age). In the second step, X1 (shock) was added, and then X3 (alcoholism). The possible changes in the fourth step were:

	d.f.	Deviance	AIC		d.f.	Deviance	AIC
+ X5	1	56.073	66.073	− X4	1	76.970	82.970
<none>		61.907	69.907	− X3	1	79.088	85.088
+ X2	1	60.304	70.304	− X1	1	79.925	85.925

Table 13.5 Step 1 Using Linear Discrimination

	d.f.	SS	RSS	AIC
+ X1	1	2.781	14.058	−210.144
+ X4	1	2.244	14.596	−206.165
+ X3	1	1.826	15.014	−203.172
+ X5	1	1.470	15.370	−200.691
+ X2	1	0.972	15.867	−197.312
<none>			16.840	−193.009

Table 13.6 Subsequent Steps Using Linear Discrimination

	d.f.	SS	RSS	AIC
<none>			10.031	−239.922
+ X2	1	0.164	9.867	−239.673
− X5	1	0.733	10.764	−234.447
− X4	1	0.919	10.950	−232.627
− X3	1	1.733	11.764	−225.029
− X1	1	2.063	12.094	−222.095

and the lowest AIC came with adding X5 (infarction) to the model. Finally, adding X2 also reduced the AIC, and no improvement could be obtained by deleting a variable, so the procedure terminated. The model minimizing AIC uses all five characteristics.

We can perform the same classification using linear discrimination. The characteristics clearly do not have a multivariate normal distribution, but it will be interesting to see how well the robustness of the methods stands up in this example.

At the first step we have the data shown in Table 13.5.

For this linear model the residual sum of squares and the change in residual sum of squares are given and used to compute the AIC. The first variable added is X1. In subsequent steps X3, X4, and X5 are added, and then we have the data shown in Table 13.6.

The procedure ends with a model using the four variables X1, X3, X4, and X5. The fifth variable (malnutrition) is not used. We can now compare the fitted values from the two models shown in Figure 13.4. It is clear that both discriminant functions separate the surviving and dying patients very well and that the two functions classify primarily the same people as being at high risk. Looking at the ROC curves suggests that the logistic discriminant function is very slightly better, but this conclusion could not be made reliably without independent data.

13.5 MODERN CLASSIFICATION TECHNIQUES

Most modern classification techniques are similar in spirit to automated stepwise logistic regression. A computer search is made through a very large number of possible models for p_k, and a criterion similar to AIC or an error estimate similar to cross-validation is used to choose a model. All these techniques are capable of approximating any relationship between p_k and X arbitrarily well, and as a consequence will give very good prediction if n is large enough in relation to p.

Modern classification techniques often produce “black-box” classifiers whose internal structure can be difficult to understand. This need not be a drawback: As the models are designed for prediction rather than inference about associations, the opaqueness of the model reduces the

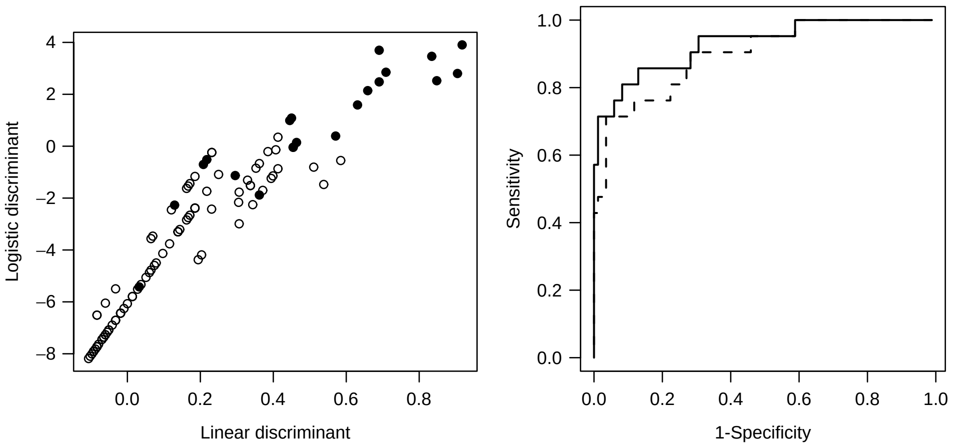


Figure 13.4 Comparison of discriminant functions and ROC curves from logistic and linear models for data of Pine et al. [1983]. Solid circles are deaths; open circles are survival. The solid line is the logistic model; the dashed line is the linear model.

temptation to leap to unjustified causal conclusions. On the other hand, it can be difficult to decide which variables are important in the classification and how strongly the predictions have been affected by outliers. There is some current statistical research into ways of opening up the black box, and techniques may become available over the next few years.

At the time of writing, general-purpose statistical packages often have little classification functionality beyond logistic and linear discrimination. It is still useful for the nonspecialist to understand the concepts behind some of these techniques; we describe two samples.

13.5.1 Recursive Partitioning

Recursive partitioning is based on the idea of classifying by making repeated binary decisions. A *classification tree* such as the left side of Figure 13.5 is constructed step by step:

1. Search every value c of every variable X for the best possible prediction by $X > c$ vs. $X \leq c$.
2. For each of the two resulting subsets of the data, repeat step 1.

In the tree displayed, each split is represented by a logical expression, with cases where the expression is true going left and others going right, so in the first split in Figure 13.5 the cases with white blood cell counts below 391.5 mL^{-1} go to the left.

An exhaustive search procedure such as this is sure to lead to overfitting, so the tree is then *pruned* by snipping off branches. The pruning is done to minimize a criterion similar to AIC:

$$\text{loss} + \text{CP} \times \text{number of splits}$$

The value of CP, called the *cost-complexity penalty*, is most often chosen by 10-fold cross-validation (Section 13.4.3). Leaving out 10% of the data, a tree is grown and pruned with many different values of CP. For each tree pruned, the error rate is computed on the 10% of data left out. This is repeated for each of the ten 10% subsets of the data. The result is a cross-validation estimate of the loss (error rate) for each value of CP, as in the right-hand side of Figure 13.5.

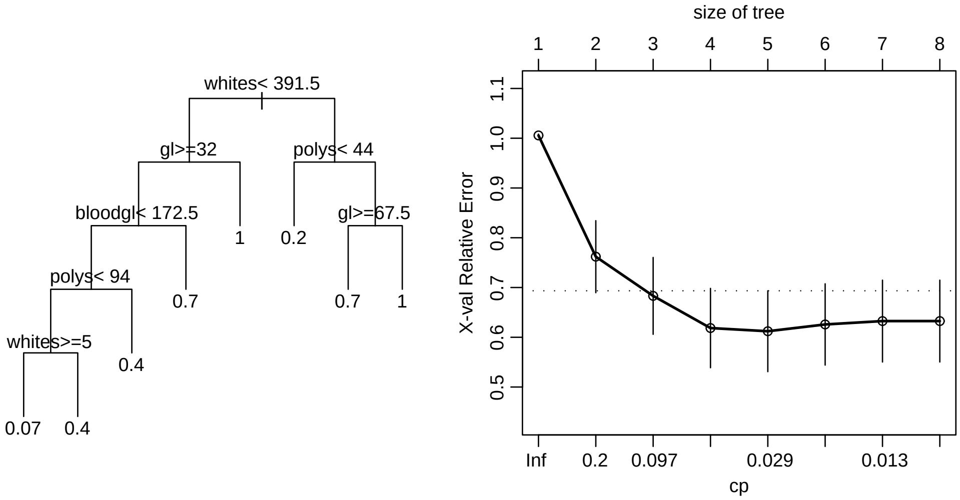


Figure 13.5 Classification tree and cross-validated error rates for differential diagnosis of acute meningitis.

Because cross-validation is relatively noisy (see the standard error bars on the graph), we choose the largest CP (smallest tree) that gives an error estimate within one standard error of the minimum, represented by the horizontal dotted line on the graph.

Example 13.4. In examining these methods we use data from Spanos et al. [1989], made available by Frank Harrell at a site linked from the Web appendix to the chapter. The classification problem is to distinguish viral from bacterial meningitis, based on a series of 581 patients treated at Duke University Medical Center. As immediate antibiotic treatment for acute bacterial meningitis is often life-saving, it is important to have a rapid and accurate initial classification. The definitive classification based on culturing bacteria from cerebrospinal fluid samples will take a few days to arrive. In some cases bacteria can be seen in the cerebrospinal fluid, providing an easy decision in favor of bacterial meningitis with good specificity but inadequate sensitivity.

The initial analysis used logistic regression together with transformations of the variables, but we will explore other possibilities. We will use the following variables:

- *AGE*: in years
- *SEX*
- *BLOODGL*: glucose concentration in blood
- *GL*: glucose concentration in cerebrospinal fluid
- *PR*: protein concentration in cerebrospinal fluid
- *WHITES*: white blood cells per milliliter of cerebrospinal fluid
- *POLYS*: % of white blood cells that are polymorphonuclear leukocytes
- *GRAM*: result of Gram smear (bacteria seen under microscope): 0 negative, > 0 positive
- *ABM*: 1 for bacterial, 0 for viral meningitis

The original analysis left GRAM out of the model and used it only to override the predicted classification if GRAM > 0. This is helpful because the variable is missing in many cases, and because the decision to take a Gram smear appears to be related to suspicion of bacterial meningitis.

In the resulting tree, each *leaf* is labeled with the probability of bacterial meningitis for cases ending up in that leaf. Note that they range from 1 down to 0.07, so that in some cases bacterial meningitis is almost certain, but it is harder to be certain of viral meningitis.

It is interesting to note what happens when Gram smear status is added to the variable list for growing a tree. It is by far the most important variable, and prediction error is distinctly reduced. On the other hand, bacterial meningitis is predicted not only in those whose Gram smear is positive, but also in those whose Gram smear is negative. Viral meningitis is predicted only in a subset of those whose Gram smear is missing. If the goal of the model were to classify the cases retrospectively from hospital records, this would not be a problem. However, the original goal was to construct a diagnostic tool, where it is undesirable to have the prediction strongly dependent on another physician choice. Presumably, the Gram smear was being ordered based on other information available to the physician but not to the investigators.

Classification trees are particularly useful where there are strong interactions between characteristics. Completely different variables can be used to split each subset of the data. In our example tree, blood glucose is used only for those with high white cell counts and high glucose in the cerebrospinal fluid. This ability is particularly useful when there are missing data.

On the other hand, classification trees do not perform particularly well when there are smooth gradients in risk with a few characteristics. For example, the prediction of acute bacterial meningitis can be improved by adding a new variable with the ratio of blood glucose to cerebrospinal fluid glucose.

The best known version of recursive partitioning, and arguably the first to handle overfitting carefully, is the CART algorithm of Breiman et al. [1984]. Our analysis used the free "rpart" package [Therneau, 2002], which automates both fitting and the cross-validation analysis. It follows the prescriptions of Breiman et al. [1984] quite closely.

A relatively nontechnical overview of recursive partitioning in biostatistics is given by Zhang and Singer [1999]. More recently, techniques using multiple classification trees (*bagging*, *boosting*, and *random forests*) have become popular and appear to work better with very large numbers of characteristics than do other methods.

13.5.2 Neural Networks

The terminology *neural network* and the original motivation were based on a model for the behavior of biological neurons in the brain. It is now clear that real neurons are much more complicated, and that the fitting algorithms for neural networks bear no detailed relationship to anything happening in the brain. Neural networks are still very useful black-box classification tools, although they lack the miraculous powers sometimes attributed to them.

A computational neuron in a neural net is very similar to a logistic discrimination function. It takes a list of inputs $Z_1, Z_2, \dots, Z_m$ and computes an output that is a function of a weighted combination of the inputs, such as

$$\operatorname{logit}(\alpha + \beta_1 Z_1 + \cdots + \beta_m Z_m) \tag{6}$$

There are many variations on the exact form of the output function, but this is one widely used variation. It is clear from equation (6) that even a single neuron can reproduce any classification from logistic regression.

The real power of neural network models comes from connecting multiple neurons together in at least two layers, as shown in Figure 13.6. In the first layer the inputs are the characteristics $X_1, \dots, X_p$. The outputs of these neurons form a "hidden layer" and are used as inputs to the second layer, which actually produces the classification probability p_k.

Example 13.5. A neural net fitted to the acute meningitis data has problems because of missing observations. Some form of imputation or variable selection would be necessary for a

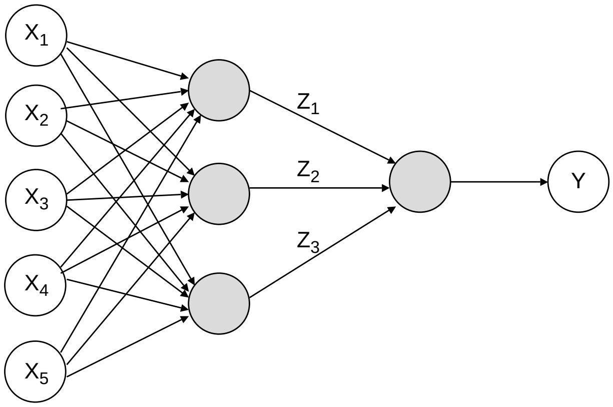


Figure 13.6 Simple neural network with three hidden nodes.

serious analysis of these data. We used the neural network package that accompanies Venables and Ripley [2002], choosing a logistic output function and two hidden nodes (Z_1 and Z_2). That is, the model was

$$\begin{aligned}\text{logit}(p) &= -0.52 + 2.46Z_1 - 2.31Z_2\\ \text{logit}(Z_1) &= 0.35 + 0.11\text{POLYS} + 0.58\text{WHITES} - 0.31\text{SEX} + 0.39\text{AGE}\\ &\quad - 0.47\text{GL} - 2.02\text{BLOODGL} - 2.31\text{PR}\\ \text{logit}(Z_2) &= 0.22 + 0.66\text{POLYS} + 0.25\text{WHITES} - 0.06\text{SEX} + 0.31\text{AGE}\\ &\quad + 0.03\text{GL} + 0.33\text{BLOODGL} - 0.02\text{PR}\end{aligned}$$

The sensitivity of the classification was approximately 50% and the specificity nearly 90%.

Two hidden nodes is the minimum interesting number (one hidden node just provides a transformation of a logistic regression model), and we did not want to use more than this because of the relatively small size of the data set.

NOTES

13.1 Maximum Likelihood for Logistic Regression

The regression coefficients in the logistic regression model are estimated using the maximum likelihood criterion. A full discussion of this topic is beyond the scope of this book, but in this note we outline the procedure for the situation involving one covariate. Suppose first that we have a Bernoulli random variable, Y, with probability function

$$\begin{aligned}P[Y = 1] &= p\\ P[Y = 0] &= 1 - p\end{aligned}$$

A mathematical trick allows us to combine these into one expression:

$$P[Y = y] = p^y(1 - p)^{(1-y)}$$

using the fact that any number to the zero power is 1. We observe n values of Y, $y_1, y_2, \ldots, y_n$ (a sequence of zeros and ones). The probability of observing this sequence is proportional to

$$\prod_{j=1}^{n} p^{y_j}(1-p)^{1-y_j} = p^{\Sigma y_j}(1-p)^{n-\Sigma y_j} \tag{7}$$

This quantity is now considered as a function of p and defined to be the likelihood. To emphasize the dependence on p, we write

$$L\left(p|\sum y_j, n\right) = p^{\Sigma y_j}(1-p)^{n-\Sigma y_j} \tag{8}$$

Given the value of $\sum y_j$, what is the "best" choice for a value for p? The maximum likelihood principle states that the value of p that maximizes $L(p|\sum y_j, n)$ should be chosen. It can be shown by elementary calculus that the value of p that maximizes $L(p|\sum y_j, n)$ is equal to $\sum y_j/n$. You will recognize this as the proportion of the n values of Y that have the value 1. This can also be shown graphically; Figure 13.7 is a graph of $L(p|\sum y_j, n)$ as a function of p for the situation $\sum y = 6$ and $n = 10$. Note that the graph has one maximum and that it is not quite symmetrical.

In the logistic regression model the probability p is assumed to be a function of an underlying covariate, X; that is, we model

$$\text{logit}(p) = \alpha + \beta X$$

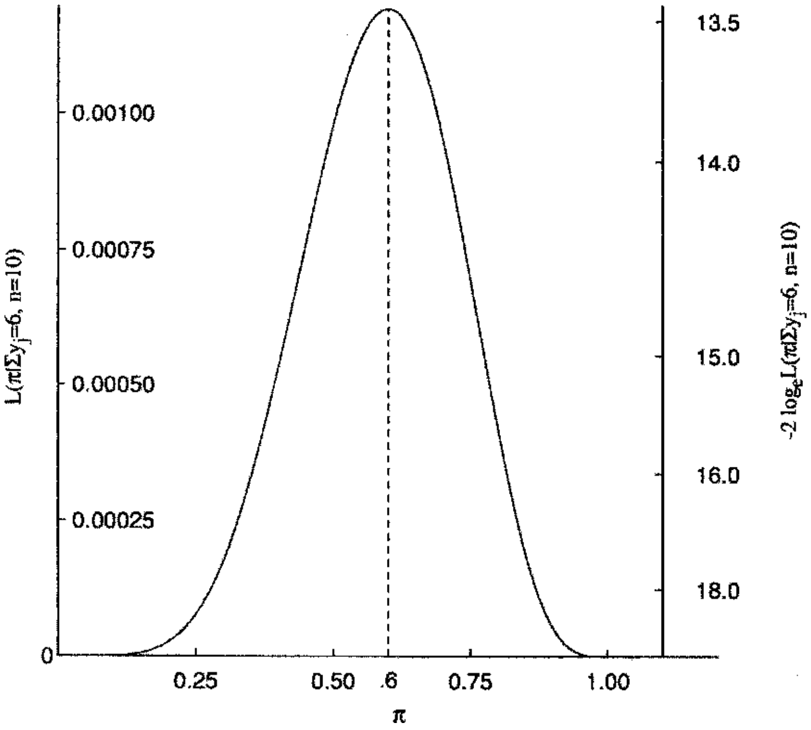


Figure 13.7 Likelihood function, $L(\pi|6, 10)$.

where α and β are constants. Conversely,

$$p = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}} = \frac{1}{1+e^{-(\alpha+\beta X)}} \tag{9}$$

For fixed values of X the probability p is determined (since α and β are parameters to be estimated from the data). A set of data now consists of *pairs* of observations: (y_j, x_j), $j = 1, \ldots, n$, where y_j is again a zero–one variable and x_j is an observed value of X for set j. For each outcome, indexed by set j, there is now a probability $p(j)$ determined by the value of x_j. The likelihood function is written

$$L(p(1), \ldots, p(n)|y_1, \ldots, y_n, x_1, \ldots, x_n, n) = \prod_{j=1}^{n} p(j)^{y_j}[1-p(j)]^{1-y_j} \tag{10}$$

but $p(j)$ can be expressed as

$$p(j) = \frac{e^{\alpha+\beta X_j}}{1+e^{\alpha+\beta X_j}} \tag{11}$$

where x_j is the value of the covariate for subject j. The likelihood function can then be written and expressed as a function of α and β as follows:

$$\begin{aligned} L(\alpha, \beta|y_1, \ldots, y_n; x_1, \ldots, x_n; n) &= \prod_{j=1}^{n} \left(\frac{e^{\alpha+\beta x_j}}{1+e^{\alpha+\beta x_j}}\right)^{y_j} \left(\frac{1}{1+e^{\alpha+\beta x_j}}\right)^{1-y_j} \\ &= \prod_{j=1}^{n} \frac{(e^{\alpha+\beta x_j})^{y_j}}{1+e^{\alpha+\beta x_j}} \\ &= \frac{e^{\Sigma_{j=1}^{n} y_j(\alpha+\beta x_j)}}{\prod_{j=1}^{n}(1+e^{\alpha+\beta x_j})} \end{aligned} \tag{12}$$

The maximum likelihood criterion then requires values for α and β to be chosen so that the likelihood function above is maximized. For more than one covariate, the likelihood function can be deduced similarly.

13.2 Logistic Discrimination with More Than Two Groups

Anderson [1972] and Jones [1975], among others, have considered the case of logistic discrimination with more than two groups. Following Anderson [1972], let for two groups

$$P(G_1|X) = \frac{\exp(\alpha_0 + \alpha_1 X_1 + \cdots + \alpha_p X_p)}{1+\exp(\alpha_0 + \alpha_1 X_1 + \cdots + \alpha_p X_p)}$$

Then

$$P(G_2|X) = \frac{1}{1+\exp(\alpha_0 + \alpha_1 X_1 + \cdots + \alpha_p X_p)}$$

This must be so because $P(G_1|X) + P(G_2|X) = 1$; that is, the observation X belongs to either the G_1 or G_2. For k groups, define

$$P(G_s|X) = \frac{\exp(\alpha_{s0} + \alpha_{s1} X_1 + \cdots + \alpha_{sp} X_p)}{1+\sum_{j=1}^{k-1}\exp(\alpha_{j0} + \alpha_{j1} X_1 + \cdots + \alpha_{jp} X_p)}$$

for groups $s = 1, \ldots, k-1$, and for group G_k, let

$$P(G_k|X) = \frac{1}{1 + \sum_{j=1}^{k-1} \exp(\alpha_{j0} + \alpha_{j1}X_1 + \cdots + \alpha_{jp}X_p)} \tag{13}$$

Most statistical packages provide this analysis, which is often called *polytomous logistic regression* (or occasionally and incorrectly, "polychotomous" logistic regression).

13.3 Defining Losses

In order to say that one prediction is better than another, we need some way to compare the relative importance of false positive and false negative errors. Even looking at total error rate implicitly assigns a relative importance. When the main adverse or beneficial effects are directly comparable, this is straightforward. We can compare the monetary costs of false negatives and false positives, or the probability of death caused by a false positive or false negative. In most cases, however, there will not be direct comparability. When evaluating a cancer screening program, the cost of false negatives is an increase in the risk of death, due to untreated cancer. The cost of a false positive includes the emotional effects and health risks of further testing needed to rule out disease. Even without weighing monetary costs against health costs we can see that it is not clear how many false negatives are worth one false positive. The problem is much more controversial, although perhaps no more difficult when monetary costs are important, as they usually are.

It can be shown [Savage, 1954] that the ability to make consistent choices between courses of action whose outcome is uncertain implies the ability to rate all the possible outcomes on the same scale, so this problem cannot be avoided. Perhaps the most important general guidance we can give is that it is important to recognize that different people will assign different losses and so prefer different classification rules.

13.4 Linear Discrimination Using Linear Regression Software

Given two groups of size n_1 and n_2, it has been shown by Fisher [1936] that the discriminant analysis is equivalent to a multiple regression on the dummy variable Y defined as follows:

$$\begin{aligned} Y &= \frac{n_2}{n_1 + n_2} \text{members of group 1} \\ &= \frac{-n_1}{n_1 + n_2} \text{members of group 2} \end{aligned} \tag{14}$$

We can now treat this as a regression analysis problem. The multiple regression equation obtained will define the regions in the sample space identical to these defined by the discriminant analysis model.

13.5 Cluster Analysis

Cluster analysis is a set of techniques for dividing observations into classes based on a set of characteristics, without the classes being specified in advance. Cluster analysis may be carried out in an attempt to discover classes that are hypothesized to exist but whose structure is unknown, but may also be used simply to create relatively homogeneous subsets of the data.

One application of cluster analysis to clinical epidemiology is in refining the definition of a new syndrome. The controversial *Gulf War syndrome* has been analyzed this way by various authors. Everitt et al. [2002] found five clusters: one *healthy* cluster and four with different distributions of symptoms. On the other hand, Hallman et al. [2003] found only two clusters: healthy and not. Cherry et al. [2001] found six clusters, three of which were relatively healthy and three representing distinct clusters of symptoms. This lack of agreement suggests that there is little evidence for genuine, strongly differentiated clusters.

Cluster analysis has become more visible in biostatistics in recent years with the rise of genomic data. A popular analysis for RNA expression data is to cluster genes based on their patterns of expression across tissue samples or experimental conditions, following Eisen et al. [1998]. The goal of these analyses is intermediate: The clusters are definitely not biologically meaningful in themselves, but are likely to contain higher concentrations of related genes, thus providing a useful starting point for further searches.

Another very visible example of cluster analysis is given by the Google News service (*http://news.google.com*). Google News extracts news stories from a very large number of traditional newspapers and other sources on the Web and finds clusters that indicate popular topics. The most prominent clusters are then displayed on the Web page.

Cluster analysis has a number of similarities to both factor analysis and principal components analysis, discussed in Chapter 14.

13.6 Predicting Categories of a Continuous Variable

In some cases the categorical outcome being predicted is defined in terms of a continuous variable. For example, low birthweight is defined as birthweight below 2500 g, diabetes may be diagnosed by a fasting blood glucose concentration over 140 mg/dL on two separate occasions, hypertension is defined as blood pressure greater than 140/90 mmHg. An obvious question is whether it is better to predict the categorical variable directly or to predict the continuous variable and then divide into categories.

In contrast to the question of whether a predictor should be dichotomized, to which we can give a clear "no!," categorizing an outcome variable may be helpful or harmful. Using the continuous variable has the advantage of making more information available, but the disadvantage of requiring the model to fit well over the entire range of the response. For example, when fitting a model to (continuous) birthweight, the parameter values are chosen by giving equal weight to a 100-g error at a weight of 4000 g as at 2450 g. When fitting a model to (binary) low birthweight, more weight is placed on errors near 2500 g, where they are more important. See also Problem 13.5.

13.7 Further Reading

Harrell [2001] discusses regression modeling for prediction, including binary outcomes, in a medical context. This is a good reference for semiautomatic modeling that uses the available features of statistical software and incorporates background knowledge about the scientific problem. Lachenbruch [1977] covers discriminant analysis, and Hosmer and Lemeshow [2000] discuss logistic regression for prediction (as well as for inference). Excellent but very technical summaries of modern classification methods are given by Ripley [1996] and Hastie et al. [2001]. Venables and Ripley [2002] describe how to use many of these methods in widely available software. As already mentioned, Zhang and Singer [1999] describe recursive partitioning and its use in health sciences. Two excellent texts on screening are Pepe [2003] and Zhou et al. [2002].

PROBLEMS

13.1 For the logistic regression model $\text{logit}(\pi) = \alpha + \beta X$, where X is a dichotomous 0–1 variable, show that e^{β} is the odds ratio associated with the exposure to X.

13.2 For the data of Table 13.7, the logistic regression model using only the variable X_1, malnutrition, is

$$\text{logit}(\widehat{\pi}) = -0.646 + 1.210X_1$$

Table 13.7 Comparison of Logistic Regression and Linear Regression (One Predictor Variable)

	Logistic Regression	Normal Regression
Dependent variable	Y discrete (binary)	Y continuous
Covariates	X categorical or continuous	X categorical or continuous
Distribution of Y (given X)	Binomial$(n\pi)$	Normal(μ, σ^2)
Model	$E(Y) = \pi$	$E(Y) = \mu$
Link to X	$\text{logit}(\pi_j) = \alpha + \beta X_j$	$\mu_j = \alpha + \beta X_j$
Data	$y_1, y_2, \ldots, y_n; x_1, x_2, \ldots, x_n$	$y_1, y_2, \ldots, y_n; x_1, x_2, \ldots, x_n$
Likelihood function (LF)	$\prod_{j=1}^{n} \pi_j^{y_j}(1-\pi_j)^{1-y_j}$ $= \prod_{j=1}^{n} \left(\frac{e^{\alpha+\beta x_j}}{1+e^{\alpha+\beta x_j}}\right)^{y_j} \left(\frac{1}{1+e^{\alpha+\beta x_j}}\right)^{1-y_j}$	$\prod_{j=1}^{n} \left(\frac{1}{\sqrt{2\sigma\pi}}\right)^{n} \exp\left(-1/2 \sum \left(\frac{y_j - \mu_j}{\sigma}\right)\right)$ $= \prod_{j=1}^{n} \left(\frac{1}{\sqrt{2\sigma\pi}}\right)^{n} \exp\left(-1/2 \sum \left(\frac{y_j - \alpha - \beta x_j}{\sigma}\right)^2\right)$
Fitting criterion (for choosing estimates of α, β)	Maximize LF	Maximize LF
-2 log LF (is proportional to)	$-2\sum y_j(\alpha + \beta X_j) + 2\sum \ln(1 + e^{\alpha+\beta X_j})$	$\frac{1}{\sigma^2} \sum (y_j - \alpha - \beta X_j)^2$
Equivalent fitting criterion	Minimize -2 log LF (*not* least squares)	Minimize -2 log LF (least squares)
Notation	$D(X) = \underset{\text{over } \alpha, \beta}{\text{minimum}} (-2 \log LF) = \text{deviance}$	$D(X) = \underset{\text{over } \alpha, \beta}{\text{minimum}} (-2 \log LF) = \text{deviance}$
Testing: $H_0 : \beta = 0$ in model	$D - D(X)$ is approximately chi-square	$D - D(X)$ is chi-square
Alternative test $H_0 : \beta = 0$ in model	$\frac{D - D(X)}{D(X)/(n-2)}$ is approximately $F_{1,n-2}$	$\frac{D - D(X)}{D(X)/(n-2)} = F_{1,n-2}$

Table 13.8 2 x 2 Table for Vital Status vs. Nutritional Status

		Y		
		Death	Survive	
X_1		1	0	
Malnutrition	1	11	21	32
No malnutrition	0	10	64	74
		21	85	106

The 2 × 2 table associated with these data is shown in Table 13.8.

(a) Verify that the coefficient of X_1 is equal to the logarithm of the odds ratio for malnutrition.

(b) Calculate the probability of death given malnutrition using the model above and compare it with the probability observed.

(c) The standard error of the regression coefficient is 0.5035; test the significance of the observed value, 1.210. Set up 95% confidence limits on the population value and translate these limits into limits for the population odds ratio.

(d) Calculate the standard error of the logarithm of the odds ratio from the 2 × 2 table and compare it with the value in part (c).

13.3 The full model for the data of Table 13.2 is given in Section 13.2.

(a) Calculate the logit line for $X_2 = 0$, $X_3 = 1$, and $X_5 = 1$. Plot logit($\widehat{\pi}$) vs. age in years.

(b) Plot $\widehat{\pi}$ vs. age in years for part (a).

(c) What is the probability of death for a 60-year-old patient with no evidence of shock, but with symptoms of alcoholism and prior bowel infarction?

13.4 One of the problems in the treatment of acute appendicitis is that perforation of the appendix cannot be predicted accurately. Since the consequences of perforation are serious, surgeons tend to be conservative by removing the appendix. Koepsell et al. [1981] attempted to relate the occurrence (or absence) of perforation to a variety of risk factors to enable better assessment of the risk of perforation. A consecutive series of 281 surgery patients was selected initially; of these, 192 were appropriate for analysis, 41 of whom had demonstrable perforated appendices according to the pathology report. The data are listed in Table 13.9. Of the 12 covariates studied, six are listed here, with the group indicator Y.

Y = perforation status (1 = yes; 0 = no)

X_1 = gender (1 = male; 0 = female)

X_2 = age (in years)

X_3 = duration of symptoms in hours prior to physician contact

X_4 = time from physician contact to operation (in hours)

X_5 = white blood count (in thousands)

X_6 = gangrene (1 = yes; 0 = no)

Table 13.9 Data for Problem 13.4

	Y	X_1	X_2	X_3	X_4	X_5	X_6		Y	X_1	X_2	X_3	X_4	X_5	X_6
1	0	0	41	19	1	16	0	49	0	1	15	6	6	19	0
2	1	1	42	48	0	24	1	50	0	0	17	10	4	9	0
3	0	0	11	24	5	14	0	51	0	0	10	72	6	17	0
4	0	1	17	12	2	9	0	52	0	1	9	8	999	15	0
5	1	1	45	36	3	99	1	53	1	1	3	4	2	18	1
6	0	0	15	24	5	14	0	54	0	0	7	16	1	24	0
7	0	1	17	11	24	8	0	55	0	1	60	14	2	11	0
8	0	1	52	30	1	13	0	56	0	1	11	48	3	8	0
9	0	1	15	26	6	13	0	57	0	1	8	48	24	14	0
10	1	1	18	48	2	20	1	58	0	1	9	12	1	12	0
11	0	0	23	48	5	14	0	59	0	1	19	36	1	99	0
12	1	1	9	336	11	13	1	60	1	0	44	24	1	11	1
13	0	0	18	24	3	13	0	61	0	0	46	9	4	12	0
14	0	0	30	8	15	11	0	62	0	1	11	36	2	13	0
15	0	0	16	19	9	10	0	63	0	1	18	8	2	19	0
16	0	1	9	8	2	15	0	64	0	0	21	24	5	12	0
17	0	1	15	48	4	12	0	65	0	0	31	24	8	16	0
18	1	1	25	120	4	8	1	66	0	0	14	7	4	12	0
19	0	0	17	7	17	14	0	67	0	1	17	6	6	19	0
20	0	1	17	12	2	14	0	68	0	0	15	24	1	9	0
21	1	0	63	72	7	11	1	69	0	0	18	24	4	9	0
22	0	0	19	8	1	15	0	70	0	0	38	48	2	99	0
23	0	1	9	48	24	9	0	71	0	1	13	18	4	18	0
24	1	0	9	48	12	14	1	72	1	0	23	168	4	18	0
25	0	0	17	5	1	14	0	73	0	0	15	3	2	14	0
26	0	0	12	48	3	15	0	74	1	0	34	48	3	16	1
27	0	1	6	48	1	26	0	75	0	1	21	24	47	8	1
28	0	0	8	48	3	99	0	76	0	1	50	8	4	12	0
29	1	1	17	30	6	12	1	77	0	0	10	23	6	16	1
30	0	0	11	8	7	15	0	78	0	0	14	48	12	15	0
31	0	1	16	48	2	11	0	79	0	1	26	48	12	13	0
32	0	1	15	10	12	12	0	80	1	0	16	22	1	14	1
33	0	1	13	24	11	15	1	81	1	0	9	24	12	16	1
34	1	1	26	48	4	11	1	82	0	1	26	5	1	16	0
35	0	1	14	7	4	16	0	83	0	1	29	24	1	30	0
36	0	0	44	20	2	13	0	84	0	1	35	408	72	6	0
37	1	1	13	168	999	10	1	85	0	0	18	168	16	12	0
38	0	0	13	14	22	13	0	86	0	1	12	18	4	12	0
39	0	1	24	10	2	19	0	87	0	1	14	7	3	21	0
40	1	0	12	72	2	16	1	88	1	1	45	24	3	18	1
41	0	1	18	15	1	16	0	89	0	1	16	5	21	12	0
42	0	0	19	15	0	9	0	90	0	0	19	240	163	6	0
43	0	0	11	336	20	8	0	91	1	1	9	48	7	23	1
44	0	1	13	14	1	99	0	92	1	1	50	30	5	15	1
45	0	1	25	10	10	11	0	93	0	0	18	2	10	15	0
46	0	1	16	72	5	7	0	94	0	0	27	2	24	17	1
47	0	1	25	72	45	7	0	95	0	1	48	27	5	16	0
48	0	1	42	12	33	19	1	96	0	1	7	18	5	14	0
97	0	1	16	13	1	11	0	145	0	1	41	24	4	14	0
98	0	1	29	5	24	19	1	146	0	0	28	6	1	15	0
99	0	1	18	48	3	11	0	147	1	0	13	48	9	15	1

Table 13.9 (*continued*)

	Y	X_1	X_2	X_3	X_4	X_5	X_6		Y	X_1	X_2	X_3	X_4	X_5	X_6
100	0	1	18	9	2	14	0	148	0	1	10	15	1	99	0
101	1	1	14	14	1	15	1	149	0	1	16	18	4	14	0
102	0	1	32	240	24	7	0	150	0	1	17	18	10	17	0
103	0	1	23	18	2	17	1	151	0	1	38	9	7	11	0
104	0	1	26	16	2	13	0	152	0	1	12	18	2	13	0
105	0	0	30	24	4	20	0	153	0	0	12	72	3	15	0
106	0	1	44	39	15	11	0	154	0	0	27	16	0	14	1
107	1	1	17	24	4	16	1	155	0	1	31	7	8	14	0
108	0	1	30	36	3	15	1	156	0	0	45	20	4	27	0
109	0	1	18	24	2	11	1	157	1	1	52	48	3	15	1
110	0	1	34	96	1	10	0	158	1	1	26	48	13	16	1
111	0	1	15	12	2	10	0	159	0	0	38	15	1	16	0
112	0	1	10	24	4	99	0	160	0	0	19	24	5	99	0
113	0	1	12	14	13	5	0	161	0	1	14	20	2	15	0
114	0	1	10	12	17	17	0	162	0	0	27	22	8	18	0
115	0	1	28	24	2	15	0	163	0	1	20	21	1	99	0
116	0	1	10	96	8	8	0	164	1	1	11	24	8	10	1
117	0	0	22	12	2	12	0	165	0	1	17	72	20	10	0
118	0	0	30	15	5	12	0	166	0	0	27	24	3	9	0
119	0	1	16	36	3	12	0	167	1	0	52	16	4	13	1
120	0	0	16	30	4	15	0	168	1	1	38	48	2	13	1
121	0	1	9	12	12	15	0	169	0	1	16	19	3	12	0
122	1	1	16	144	4	15	1	170	0	1	19	9	4	17	0
123	0	1	17	36	13	6	0	171	0	0	24	24	2	11	0
124	1	1	12	120	2	11	1	172	0	1	12	17	20	6	1
125	0	1	28	17	26	10	0	173	1	1	51	72	2	16	1
126	1	0	13	48	3	21	1	174	1	1	50	72	6	11	1
127	0	0	23	72	3	13	0	175	0	0	28	12	3	13	0
128	1	0	62	72	2	12	1	176	0	0	19	48	8	14	1
129	0	1	17	24	4	14	0	177	0	1	9	24	999	99	0
130	0	0	12	24	12	15	0	178	0	0	40	48	7	14	0
131	0	1	10	12	10	11	0	179	0	0	17	504	7	99	0
132	0	1	47	48	8	9	0	180	0	1	51	24	1	9	1
133	0	1	43	11	8	13	0	181	0	1	31	24	2	10	0
134	1	1	18	36	2	15	1	182	0	0	25	8	9	8	0
135	0	0	6	24	1	9	0	183	0	0	14	24	8	10	0
136	0	0	24	2	22	10	0	184	0	1	7	24	4	15	0
137	0	0	22	11	24	7	0	185	0	1	27	7	2	14	0
138	1	1	39	36	3	15	1	186	0	1	35	72	3	19	1
139	1	1	43	48	2	11	1	187	0	0	11	12	9	11	0
140	0	1	12	7	1	14	0	188	0	1	20	8	6	12	0
141	0	1	14	48	6	16	0	189	0	1	50	48	27	19	0
142	0	1	21	24	1	17	0	190	0	1	16	6	7	7	0
143	1	1	34	48	12	9	1	191	0	1	45	24	4	20	0
144	1	0	60	24	3	14	1	192	1	1	47	336	4	9	1

For X_4 the code 999 is for unknown; for X_5 the code 99 is an unknown code.

(a) Compare the means of the continuous variables (X_2, X_3, X_4, X_5) in the two outcome groups $(Y = 0, 1)$ by some appropriate test. Make an appropriate comparison of the association of X_5 and Y. State your conclusion at this point.

(b) Carry out a stepwise discriminant analysis. Which variables are useful predictors? How much improvement in prediction is there in using the discriminant procedure? How appropriate is the procedure?

(c) Carry out a stepwise logistic regression and compare your results with those of part (b).

(d) The authors introduced two additional variables in their analysis: $X_7 = \log(X_2)$ and $X_8 = \log(X_3)$. Test whether these variables improve the prediction scheme. Interpret your findings.

(e) Plot the probability of perforation as a function of the duration of symptoms; using the logistic model, generate a separate curve for subjects aged 10, 20, 30, 40, and 50 years. Interpret your findings.

13.5 The Web appendix to this chapter has a data set with daily concentrations of particulate air pollution in Seattle, Washington. The air quality index for fine particulate pollution below 2.5 μm in diameter (PM2.5) will be "unhealthy for sensitive groups" at 40 μg/m^3 and "moderate" at 20 μg/m^3. The Puget Sound Clean Air Agency is interested in predicting high air pollution days so that it can issue burn bans to reduce fireplace use. Using information on weather and pollution from previous days and the time of year, build logistic models to predict when PM2.5 will exceed 20 or 40 μg/m^3. Also build a linear regression model for predicting PM2.5 or log(PM2.5). Summarize the predictive accuracy of these models. Do you get more accurate prediction using the logistic model or categorizing the prediction from the linear model? Does the answer depend on what losses you assign to false positive and false negative predictions?

13.6 A classic in the use of discriminant analysis is the paper by Truett et al. [1967], in which the authors attempted to predict the risk of coronary heart disease using data from the Framingham study, a longitudinal study of the incidence of coronary heart disease in Framingham, Massachusetts. The two groups under consideration were those who did and did not develop coronary heart disease (CHD) in a 12-year follow-up period. There were 2669 women and 2187 men, aged 30 to 62, involved in the study and free from CHD at their first examination. The variables considered were:

- Age (years)
- Serum cholesterol (mg/100 mL)
- Systolic blood pressure (mmHg)
- Relative weight (100 × actual weight ÷ median for sex–height group)
- Hemoglobin (g/100 mL)
- Cigarettes per day, coded as 0 = never smoked, 1 = less than a pack a day, 2 = one pack a day, and 3 = more than a pack a day
- ECG, coded as 0 = for normal, and 1 = for definite or possible left ventricular hypertrophy, definite nonspecific abnormality, and intraventricular block

Note that the variables "cigarettes" and "ECG" cannot be distributed normally, as they are discrete variables. Nevertheless, the linear discriminant function model was tried. It was found that the predictions (in terms of the risk or estimated probability of being in the coronary heart disease groups) fitted the data well. The coefficients of the linear discriminant functions for men and women, including the standard errors, are shown in Table 13.10.

Table 13.10 Coefficients and Standard Errors for Predicting Coronary Heart Disease.

Risk Factors	Women	Men	Standard Errors of Estimated Coefficients	
Constant ($\widehat{\alpha}$)	−12.5933	−10.8986		
Age (years)	0.0765	0.0708	0.0133	0.0083
Cholesterol (mg %)	0.0061	0.0105	0.0021	0.0016
Systolic blood pressure (mmHg)	0.0221	0.0166	0.0043	0.0036
Relative weight	0.0053	0.0138	0.0054	0.0051
Hemoglobin (g %)	0.0355	−0.0837	0.0844	0.0542
Cigarettes smoked (*see code*)	0.0766	0.3610	0.1158	0.0587
ECG abnormality (*see code*)	1.4338	1.0459	0.4342	0.2706

(a) Determine for both women and men in terms of the p-value the most significant risk factor for CHD in terms of the p-value.

(b) Calculate the probability of CHD for a male with the following characteristics: age = 35 years; cholesterol = 220 mg %; systolic blood pressure = 110 mmHg; relative weight = 110; hemoglobin = 130 g%; cigarette code = 3; and ECG code = 0.

(c) Calculate the probability of CHD for a female with the foregoing characteristics.

(d) How much is the probability in part (b) changed for a male with all the characteristics above except that he does not smoke (i.e., cigarette code = 0)?

(e) Calculate and plot the probability of CHD for the woman in part (c) as a function of age.

13.7 In a paper that appeared four years later, Halperin et al. [1971] reexamined the Framingham data analysis (see Problem 13.6) by Truett et al. [1967] using a logistic model. Halperin et al. analyzed several subsets of the data; for this problem we abstract the data for men aged 29 to 39 years, and three variables: cholesterol, systolic blood pressure, and cigarette smoking (0 = never smoked; 1 = smoker); cholesterol and systolic blood pressure are measured as in Problem 13.6. The following coefficients for the logistic and discriminant models (with standard errors in parentheses) were obtained:

	Intercept	**Cholesterol (mg/100 mL)**	**Systolic Blood Pressure**	**Cigarettes**
Logistic	−11.6246	0.0179(0.0036)	0.0277(0.0085)	1.7346(0.6236)
Discriminant	−13.5300	0.0236(0.0039)	0.0302(0.0100)	1.1191(0.3549)

(a) Calculate the probability of CHD for a male with relevant characteristics defined in Problem 13.6, part (b), for both the logistic and discriminant models.

(b) Interpret the regression coefficients of the logistic model.

(c) In comparing the two methods, the authors state: "Empirically, the assessment of significance of a variable, as measured by the ratio of the estimated coefficient to its estimated standard error, is apt to be about the same whichever method is used." Verify that this is so for this problem. (However, see also the discussion in Section 13.3.2.)

13.8 In a paper in *American Statistician*, Hauck [1983] derived confidence bands for the logistic response curve. He illustrated the method with data from the Ontario Exercise Heart Collaborative Study. The logistic model dealt with the risk of myocardial infarction (MI) during a study period of four years. A logistic model based on the two most important variables, smoking (X_1) and serum triglyceride level (X_2), was calculated to be

$$\text{logit}(P) = -2.2791 + 0.7682X_1 + 0.001952(X_2 - 100)$$

where P is the probability of an MI during the four-year observation period. The variable X_1 had values $X_1 = 0$ (nonsmoker) and $X_1 = 1$ (smoker). As in ordinary regression, the confidence band for the entire line is narrowest at the means of X_1 and $(X_2 - 100)$ and spreads out the farther you go from the means. (See the paper for more details.)

(a) The range of values of triglyceride levels is assumed to be from 0 to 550. Graph the probability of MI for smokers and nonsmokers separately.

(b) The standard errors of regression coefficients for smoking and serum triglyceride are 0.3137 and 0.001608, respectively. Test their significance.

13.9 One of the earliest applications of the logistic model to medical screening by Anderson et al. [1972] involved the diagnosis of keratoconjunctivitis sicca (KCS), also known as "dry eyes." It is known that rheumatoid arthritic patients are at greater risk, but the definitive diagnosis requires an ophthalmologist; hence it would be advantageous to be able to predict the presence of KCS on the basis of symptoms such as a burning sensation in the eye. In this study, 40 rheumatoid patients with KCS and 37 patients without KCS were assessed with respect to the presence (scored as 1) or absence (scored as 0) of each of the following symptoms: (1) foreign body sensation; (2) burning; (3) tiredness; (4) dry feeling; (5) redness; (6) difficulty in seeing; (7) itchiness; (8) aches; (9) soreness or pain; and (10) photosensitivity and excess of secretion. The data are reproduced in Table 13.11.

(a) Fit a stepwise logistic model to the data. Test the significance of the coefficients.

(b) On the basis of the proportions of positive symptoms displayed at the bottom of the table, select that variable that should enter the regression model first.

(c) Estimate the probability of misclassification.

(d) It is known that approximately 12% of patients suffering from rheumatoid arthritis have KCS. On the basis of this information, calculate the appropriate logistic scoring function.

(e) Define X = number of symptoms reported (out of 10). Do a logistic regression using this variable. Test the significance of the regression coefficient. Now do a t-test on the X variable comparing the two groups. Discuss and compare your results.

13.10 This problem deals with the data of Pine et al. [1983]. Calculate the posterior probabilities of survival for a patient in the fourth decade arriving at the hospital in shock and history of myocardial infarction and without other risk factors:

(a) Using the logistic model.

(b) Using the discriminant model.

Table 13.11 Data for Problem 13.8

	KCS Patients											Patients Without KCS									
Patient	1	2	3	4	5	6	7	8	9	10	Patient	1	2	3	4	5	6	7	8	9	10
1	1	1	1		1		1			1	1	1						1			
2	1	1	1	1	1			1			2										
3	1	1		1	1	1			1		3		1	1				1			
4	1	1		1	1			1	1		4										
5	1	1	1	1			1			1	5								1		
6	1	1			1					1	6										
7	1	1	1			1		1			7			1				1			
8	1	1		1		1			1		8										
9	1	1	1	1	1		1	1			9										
10	1										10									1	
11	1	1	1	1		1		1			11		1					1			
12	1	1					1	1	1		12										
13	1	1	1	1	1	1	1			1	13							1			
14	1	1	1	1	1		1	1		1	14										
15			1	1			1	1			15							1			
16	1	1		1					1		16										
17		1	1			1		1		1	17							1			
18	1		1	1	1			1			18										
19	1		1	1	1		1				19					1					
20	1	1	1	1			1			1	20										
21					1						21										
22	1	1	1	1	1	1	1				22						1				

continued overleaf

Table 13.11 (*continued*)

	KCS Patients											Patients Without KCS									
Patient	1	2	3	4	5	6	7	8	9	10	Patient	1	2	3	4	5	6	7	8	9	10
23	1	1	1		1					1	23										
24	1	1		1			1	1	1		24					1					
25	1	1	1	1				1			25	1								1	
26				1			1				26										
27	1	1		1	1			1	1	1	27										
28	1	1	1	1						1	28										
29	1		1		1				1		29							1			
30	1	1		1		1				1	30										
31		1	1	1						1	31				1						
32	1	1	1	1	1	1	1			1	32										
33			1	1	1		1		1		33							1			
34	1	1	1	1		1	1			1	34										
35	1		1		1			1			35										1
36	1	1	1	1				1			36										
37	1	1	1	1	1						37							1			1
38	1	1																			
39																					
40		1	1	1			1			1											
Proportion position	$\frac{32}{40}$	$\frac{30}{40}$	$\frac{26}{40}$	$\frac{28}{40}$	$\frac{19}{40}$	$\frac{10}{40}$	$\frac{16}{40}$	$\frac{15}{40}$	$\frac{9}{40}$	$\frac{15}{40}$	Proportion position	$\frac{2}{37}$	$\frac{2}{37}$	$\frac{2}{37}$	$\frac{1}{37}$	$\frac{2}{37}$	$\frac{1}{37}$	$\frac{10}{37}$	$\frac{1}{37}$	$\frac{2}{37}$	$\frac{2}{37}$

(c) Graph the two survival curves as a function of age. Use the values 5, 15, 25, ... for the ages in the discriminant model.

(d) Assume that the prior probabilities are $\pi_1 = P[\text{survival}] = 0.60$ and $\pi_2 = 1 - 0.60 = 0.40$. Recalculate the probabilities in parts (a) and (b).

(e) Define a new variable for the data of Table 13.2 as follows: $X_6 = X_1 + X_2 + X_3 + X_5$. Interpret this variable.

(f) Do a logistic regression and discriminant analysis using variables X_4 and X_6 (defined above). Interpret your results.

(g) Is any information "lost" using the approach of parts (e) and (f)? If so, what is lost? When is this likely to be important?

13.11 This problem requires some programming. Create 100 observations of 20 independent random characteristics (e.g., from a uniform distribution) and one random 0–1 variable. Fit a logistic discrimination model using 1, 2, 5, 10, 15, or 20 of your characteristics, and 20, 40, 60, 80, and 100 of the observations. Compute the in-sample error rate and compare it to the true error rate (1/2).

13.12 This problem deals with the data of Problem 5.14, comparing the effect of the drug nifedipine on vasospasm attacks in patients suffering from Raynaud's phenomenon. We want to make a multivariate comparison of the seven patients with a history of digital ulcers ("yes" in column 4) with the eight patients without a history of digital ulcers ("no" in column 4). Variables to be used are age, gender, duration of phenomenon, total number of attacks on placebo, and total number of attacks on nifedipine.

(a) Carry out a stepwise logistic regression on these data.

(b) Which variable entered first?

(c) State your conclusion.

(d) Make a scatter plot of the logistic scores and indicate the dividing point.

***13.13** This problem deals with the data of Problem 10.10, comparing metabolic clearance rates in three groups of subjects.

(a) Use a discriminant analysis on the *three* groups.

(b) Interpret your results.

(c) Graph the data using different symbols to denote the three groups.

(d) Suppose you "create" a third variable: concentration at 90 minutes minus concentration at 45 minutes. Will this improve the discrimination? Why or why not?

***13.14** Consider two groups, G_1 and G_2 (e.g., "death," "survive"; "disease," "no disease"), and a binary covariate, X, with values 0 or 1 (e.g., "don't smoke," "smoke"; "symptom absent," "symptom present"). The data can be arranged in a 2 × 2 table:

	Group	
X	G_1	G_2
1		
0		
	π_1	π_2

Here π_1 is the prior probability of group G_1 membership; $P(X = i|G_1)$ the likelihood of $X = i$ given G_1 membership, $i = 0, 1$; and $P(G_1|X = i)$ the posterior probability of G_1 membership given that $X = i, i = 0, 1$.

(a) Show that

$$\frac{P(G_1|X = i)}{P(G_2|X = i)} = \frac{\pi_1}{\pi_2} \frac{P(X = i|G_1)}{P(X = i|G_2)}$$

Hint: Use Bayes' theorem.

(b) The expression in part (a) can be written as

$$\frac{P(G_1|X = i)}{1 - P(G_1|X = i)} = \frac{\pi_1}{1 - \pi_1} \frac{P(X = i|G_1)}{P(X = i|G_2)}$$

In words:

posterior odds of group 1 membership = prior odds of group 1 membership × ratio of likelihoods of observed values of X.

Relate the ratio of likelihoods to the sensitivity and specificity of the procedure.

(c) Take logarithms of both sides of the equation in part (b). Relate your result to Note 6.7.

(d) The result in part (b) can be shown to hold for X continuous or multivariate. What are the assumptions [go back to the simple set-up of part (a)].

REFERENCES

Akaike, H. [1973]. Information theory and an extension of the maximum likelihood principle. *Proceedings of the 2nd International Symposium on Information Theory.*

Anderson, J. A. [1972]. Separate sample logistic regression. *Biometrika*, **59**: 19–35.

Anderson, J. A., Whaley, K., Williamson, J., and Buchanan, W. W. [1972]. A statistical aid to the diagnosis of keratoconjunctivitis sicca. *Quarterly Journal of Medicine, New Series*, **41**: 175–189. Used with permission from Oxford University Press.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. [1984]. *Classification and Regression Trees.* Wadsworth Press, Belmont, CA.

Cherry, N., Creed, F., Silman, A., Dunn, G., Baxter, D., Smedley, J., Taylor, S., and Macfarlane, G. J. [2001]. Health and exposure of United Kingdom Gulf War veterans: I. The pattern and extent of ill health. *Occupational and Environmental Medicine*, **58**: 291–298.

Cover, T. M. [1965]. Geometrical and statistical properties of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computing*, **14**: 326–334.

Efron, B. [1975]. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, **70**: 892–898.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. [1998]. Cluster analysis and display of genome-wise expression patterns. *Proceedings of the National Academy of Sciences*, **95**(25): 14863–14868.

Everitt, B., Ismail, K., David, A. S., and Wessely, S. [2002]. Searching for a Gulf War syndrome using cluster analysis. *Psychological Medicine*, **32**(8): 1335–1337.

Fisher, R. A. [1936]. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*. **7**: 179–188.

Hall, P. and Li, K-C. [1993]. On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, **21**: 867–889.

Hallman, W. K., Kipen, H. M., Diefenbach, M., Boyd, K., Kang, H., Leventhal, H., and Wartenberg, D. [2003]. Symptom patterns among Gulf War registry veterans. *American Journal of Public Health*, **93**(4): 624–630.

Halperin, M. and Gurian, J. [1971]. A note on estimation in straight line regression when both variables are subject to error. *Journal of the American Statistical Association*, **66**: 587–589.

Halperin, M., Blockwelder, W. C., and Verter, J. I. [1971]. *Estimation* of the multivariate logistic risk function: a comparison of the discriminant function and maximum likelihood approaches. *Journal of Chronic Diseases*, **24**: 125–158.

Harrell, F. E. [2001]. *Regression Modeling Strategies*. SpringerVerlag, New York.

Hastie, T., Tibshirani, R., and Friedman, J. H. [2001]. *The Elements of Statistical Learning*. SpringerVerlag, New York.

Hauck, W. W. [1983]. A note on confidence bands for the logistic response curve. *American Statistician*, **37**: 158–160.

Health Canada [2001]. *Organized Breast Cancer Screening Programs in Canada: 1997 and 1998 report*. Downloaded from *http://www.hc-sc.gc.ca/pphb-dgspsp/publications_e.html.*

Hosmer, D. W., and Lemeshow, S. [2000]. *Applied Logistic Regression*, 2nd ed. Wiley, New York.

Jones, R. H. [1975]. Probability estimation using a multinomial logistic function. *Journal of Statistical Computation and Simulation*, **3**: 315–329.

Knoke, J. D. [1982]. Discriminant analysis with discrete and continuous variables. *Biometrics*, **38**: 191–200. See also correction in *Biometrics*, **38**: 1143.

Koepsell, T. D., Inui, T. S., and Farewell, V. T. [1981]. Factors affecting perforation in acute appendicitis. *Surgery, Gynecology and Obstetrics*, **153**: 508–510. Used with permission.

Lachenbruch, P. A. [1977]. *Discriminant Analysis*. Hafner Press, New York.

Li, K-C. and Duan, N. [1989]. Regression analysis under link violation, *The Annals of Statistics*, **17**: 1009–1052.

Pepe, M. S. [2003]. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, Oxford.

Pine, R. W. Wertz, M. J., Lennard, E. S., Dellinger, E. P., Carrico, C. J., and Minshew, H. [1983]. Determinants of organ malfunction or death in patients with intra-abdominal sepsis. *Archives of Surgery*, **118**: 242–249. Copyright © 1983 by the American Medical Association.

Ripley, B. D. [1996]. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Savage, L. J. [1954]. *The Foundations of Statistics*. Wiley, New York.

Spanos, A., Harrell, F. E. and Durack, F. T. [1989]. Differential diagnosis of acute meningitis. An analysis of the predictive value of initial observations. *Journal of the American Medical Association*, **262**(19): 2700–2707.

Therneau, T. M. [2002]. *Rpart Software*. Mayo Foundation for Medical Research, Rochester, MN.

Truett, J., Cornfield, J., and Kannel, W. [1967]. A multivariate analysis of the risk of coronary heart disease in Framingham. *Journal of Chronic Diseases*, **20**: 511–524.

Venables, W. N., and Ripley, B. D. [2002]. *Modern Applied Statistics with S*, 4th ed. SpringerVerlag, New York.

Wilson P. W. F., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., and Kannel, W. B. [1998]. Prediction of coronary heart disease using risk factor categories. *Circulation*, **97**: 1837–1847.

Zhang, H., and Singer, B. [1999]. *Recursive Partitioning in the Health Sciences*. SpringerVerlag, New York.

Zhou, X.-H., McClish, D. K., and Obuchowski, A. [2002]. *Statistical Methods in Diagnostic Medicine*. Wiley, New York.

CHAPTER 14

Principal Component Analysis and Factor Analysis

14.1 INTRODUCTION

In Chapters 10 and 11 we considered the dependence of a specified response variable on other variables. The response variable identified played a special role among the variables being considered. This is appropriate in many situations because of the scientific question and/or experimental design. What do you do, however, if you have a variety of variables and desire to examine the relationships between them without identifying a specific response variable?

In this chapter we present two methods of examining the relationships among a set of variables without identifying a specific response variable. For these methods, no single variable has a more distinguished role or importance than any other variable. The first technique we examine, principal component analysis, explains as much variability as possible in terms of a few linear combinations of the variables. The second technique, factor analysis, explains the relationships between variables by a few unobserved factors. Both methods depend on the covariances, or correlations, between variables.

14.2 VARIABILITY IN A GIVEN DIRECTION

Consider the 20 observations on two variables X and Y listed in Table 14.1. These data are such that the original observations had their means subtracted, so that the means of the points are zero. Figure 14.1 plots these points, that is, plots the data points about their common mean.

Rather than thinking of the data points as X and Y values, think of the data points as a point in a plane. Consider Figure 14.2(a); when an origin is identified, each point in the plane is identified with a pair of numbers x and y. The x value is found by dropping a line perpendicular to the horizontal axis; the y value is found by dropping a line perpendicular to the vertical axis. These axes are shown in Figure 14.2(b). It is not necessary, however, to use the horizontal and vertical directions to locate our points, although this is traditional. Lines at any angle θ from the horizontal and vertical, as shown in Figure 14.2(c), might be used. In terms of these two lines, the data point has values found by dropping perpendicular lines to these two directions; Figure 14.2(d) shows the two values. We will call the new values x' and y' and the old values x and y. It can be shown that x' and y' are linear combinations of x and y. This idea of lines in different directions with perpendiculars to describe the position of points is used in principal component analysis.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2 Copyright © 2004 John Wiley & Sons, Inc.

Table 14.1 Twenty Biometric Observations

Observation	X	Y	Observation	X	Y
1	−0.52	0.60	11	0.08	0.23
2	0.04	−0.51	12	−0.06	−0.59
3	1.29	−1.19	13	1.25	−1.25
4	−1.12	1.90	14	0.53	−0.45
5	−1.02	0.31	15	0.14	0.47
6	0.10	−1.15	16	0.48	−0.11
7	−0.32	−0.13	17	−0.61	1.04
8	0.08	−0.17	18	−0.47	0.34
9	0.49	0.18	19	0.41	0.29
10	−0.54	0.20	20	−0.22	0.00

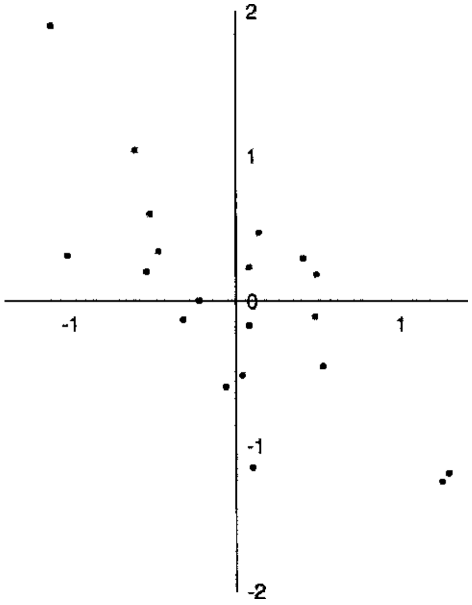


Figure 14.1 Plot of the 20 data points of Table 14.1.

For our data set, the variability in x and y may be summarized by the standard deviation of the x and y values, respectively, as well as the covariance, or equivalently, the correlation between them. Consider now the data of Figure 14.1 and Table 14.1. Suppose that we draw a line in a direction of 30° to the horizontal. The 20 observations give 20 x' values in the X' direction when the perpendicular lines are dropped. Figure 14.3 shows the values in the x' direction. Consider now the points along the line in the x' direction corresponding to the feet of the perpendicular lines. We may summarize the variability among these points by our usual measure of variability, the standard deviation. This would be computed in our usual manner from the 20 values x'. The variability of the data may be summarized by plotting the standard deviation, say $s(\theta)$, in each direction θ at a distance s from the origin. When we look at the standard deviation in all directions, this results in an egg-shaped curve with dents in the side; or a symmetric curve in the shape of a violin or cello body. For the data at hand, this curve is shown in Figure 14.4; the curve is identified as the standard deviation curve. Note that the standard deviation is not the same in all directions. For our data set, the data are spread out more

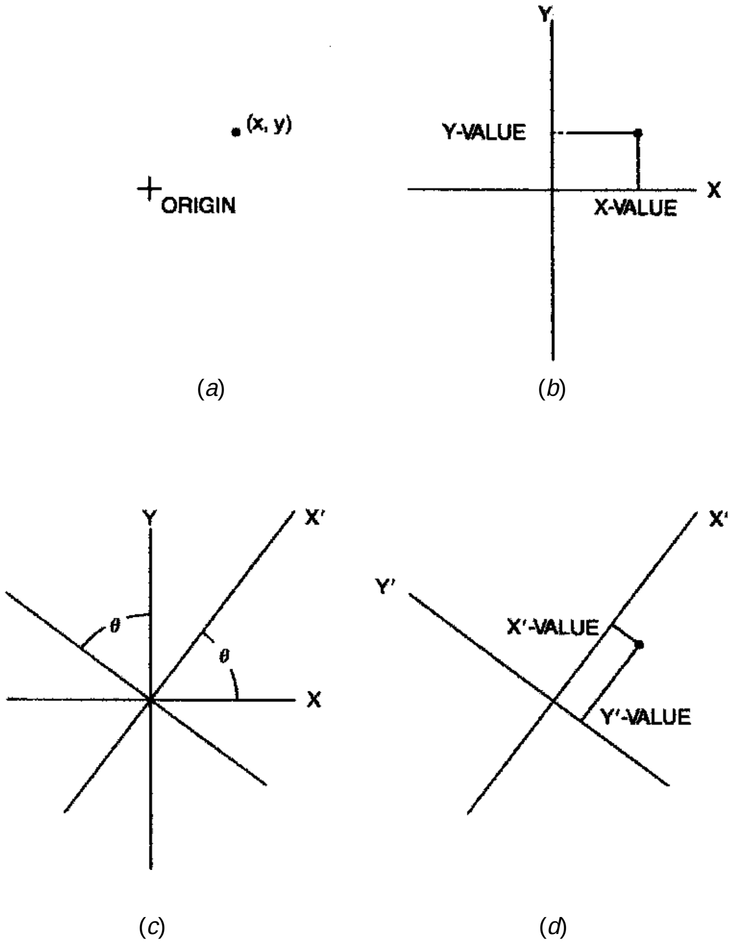


Figure 14.2 Points in the plane, coordinates, and rotation of axes.

along a northwest–southeast direction than in the southwest–northeast direction. The standard deviation curve has a minimum distance at about 38°. The standard deviation increases steadily to a maximum; the maximum is positioned along the line in Figure 14.4, running from the upper left to the lower right. These two directions are labeled directions 1 and 2. If we want to pick one direction that contains as much variability as possible, we would choose direction 1, because the standard deviation is largest in that direction. If all the data points lie on a line, the variability will be a maximum in the direction of the line that contains all the data.

There is some terminology used in finding the value of a data point in a particular direction. The process of dropping a line perpendicular to a direction is called *projecting* the point onto the direction. The value in the particular direction [x' in Figure 14.2(*d*) or Figure 14.3] is called the *projection of the point*. If we know the values x and y, or if we know the values x' and y', we know where the point is in the plane. Two such variables x and y, or equivalently, x' and y', which allow us to find the values of the data, are called a *basis for the variables*.

These concepts may be generalized when there are more than two variables. If we observe three variables x, y, and z, the points may be thought of as points in three dimensions. Suppose that we subtract the means from all the data so that the data are centered about the origin of a three-dimensional plot. As you sit reading this material, picture the points suspended about the

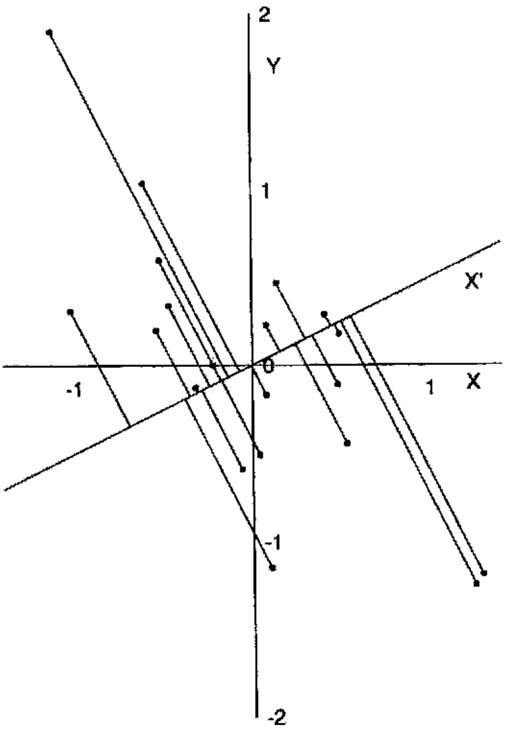


Figure 14.3 Values in the X-direction. X' axis at 30° to the x-axis.

room. Pick an origin. You may draw a line through the origin in any direction. For any point that you have picked in the room, you may drop a perpendicular to the line. Given a line, the point on the line where the perpendicular meets the line is the projection of the point onto the line. We may then calculate the standard deviation for this direction. If the standard deviations are plotted in all directions, a dented egg-shaped surface results. There will be one direction with the greatest variability. When more than three variables are observed, although we cannot picture the situation mentally, mathematically the ideas may be extended; the concept of a direction may be extended in a natural manner. In fact, mathematical statistics is one part of mathematics that heavily uses the geometry of n-dimensional space when there are n variables observed. Fortunately, to understand the statistical methods, we do not need to understand the mathematics!

Let us turn our attention again to Figure 14.4. Rather than plotting the standard deviation curve, it is traditional to summarize the variability in the data by an ellipse. The two perpendicular axes of the ellipse lie along the directions of the greatest variability and the least variability. The ellipse, called the *ellipsoid of concentration*, meets the standard deviation curve along its axes at the points of greatest and least variation. In other directions the standard deviation curve will be larger, that is, farther removed from the origin. In three dimensions, rather than plotting an ellipse we plot an egg-shaped surface, the ellipsoid. (One reason the ellipsoid is used: If you have a bivariate normal distribution in the plane, take a very large sample, divide the plane up into small squares as on graph paper, and place columns whose height is proportional to the number of points; the columns of constant height would lie on an ellipsoid.)

Out of the technical discussion above, we want to remember the following ideas:

1. If we observe a set of variables, we may think of each data point as a point in a space. In this space, when the points are centered about their mean, there is variability in each direction.

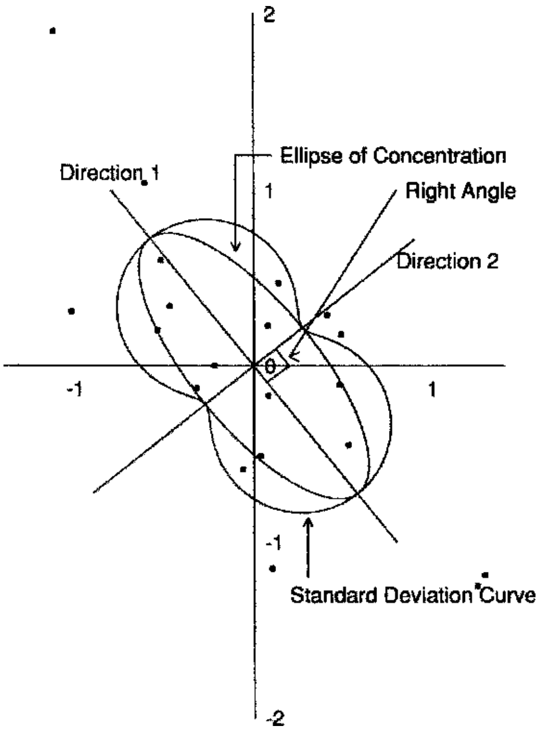


Figure 14.4 Standard deviation in each direction and the ellipse of concentration.

2. The variability is a maximum in one direction. In two dimensions (or more) the minimum lies in a perpendicular direction.
3. The variability is symmetric about each of the particular directions identified.

It is possible to identify the various directions with linear combinations of the variables or coordinates. Each direction for $X_1, \ldots, X_p$ is associated with a sum

$$Y = a_1X_1 + a_2X_2 + \cdots + a_pX_p \tag{1}$$

where

$$a_1^2 + a_2^2 + \cdots + a_p^2 = 1$$

The constants $a_1, a_2, \ldots, a_p$ are uniquely associated with the direction, except that we may multiply each a by -1. The sum that is given in equation (1) is the value of the projection of the points x_1 to x_p corresponding to the given direction.

14.3 PRINCIPAL COMPONENTS

The motivation behind principal component analysis is to find a direction, or a few directions, that explain as much of the variability as possible. Since each direction is associated with a linear sum of the variables, we may say that we want to find a few new variables, which are linear sums of the old variables, which explain as much of the variability as possible. Thus, the first principal component is the linear sum corresponding to the direction of greatest variability:

Definition 14.1. The *first principal component* is the sum

$$Y = a_1X_1 + \cdots + a_pX_p, \qquad a_1^2 + \cdots + a_p^2 = 1 \tag{2}$$

corresponding to the direction of greatest variability when variables $X_1, \ldots, X_p$ are under consideration.

Usually, the first principal component will leave much of the variability unexplained. (In the next section, we discuss a method of quantifying the amount of variability explained.) For this reason we wish to search for a second principal component that explains much of the remaining variability. You might think we would take the next linear combination of variables that explains as much of the variability as possible. But when you examine Figure 14.4, you see that the closer the direction gets to the first principal component (which would be direction 1 in Figure 14.4), the more variability one would have. Thus, essentially, we would be driven to the same variable. Therefore, the search for the second principal component is restricted to variables that are uncorrelated with the first principal component. Geometrically, it can be shown that this is equivalent to considering directions that are perpendicular to the direction of the first principal component. In two dimensions such as Figure 14.4, direction 2 would be the direction of the second principal component. However, in three dimensions, when we have the line corresponding to the direction of the first principal component, the set of all directions perpendicular to it correspond to a plane, and there are a variety of possible directions in which to search for the second principal component. This leads to the following definition:

Definition 14.2. Suppose that we have the first $k - 1$ principal components for variables $X_1, \ldots, X_p$. The kth *principal component* corresponds to the variable or direction that is uncorrelated with the first $k - 1$ principal components and has the largest possible variance.

As a summary of these difficult ideas, you should remember the following:

1. Each principal component is chosen to explain as much of the remaining variability as possible after the preceding principal components have been chosen.
2. Each principal component is uncorrelated to the other principal components. In the case of a multivariate normal distribution, the principal components are statistically independent.
3. Although it is not clear from the above, the following is true: For each k, the first k principal components explain as much of the variability in a sample as may be explained by any k directions, or equivalently, k variables.

14.4 AMOUNT OF VARIABILITY EXPLAINED BY THE PRINCIPAL COMPONENTS

Suppose that we want to perform a principal component analysis upon variables $X_1, \ldots, X_p$. If we were dealing with only one variable, say variable X_j, we summarize its variability by the variance. Suppose that there are a total of n observations, so that for each of the p variables, we have n values. Let X_{ij} be the ith observation on the jth variable. Let $\overline{X}_j$ be the mean of the n observations on the jth variable. Then we estimate the variability, that is, the variance, of the variable X_j by

$$\widehat{\text{var}}(X_j) = \sum_{i=1}^{n} \frac{(X_{ij} - \overline{X}_j)^2}{n-1} \tag{3}$$

A reasonable summary of the variability in the p variables is the sum of the individual variances. This leads us to the next definition.

Definition 14.3. The *total variance, denoted by V*, for variables $X_1, \ldots, X_p$ is the sum of the individual variances. That is,

$$\text{total variance} = V = \sum_{j=1}^{p} \text{var}(X_j) \tag{4}$$

The sample total variance, which we will also denote by V since that is the only type of total variance used in this section, is

$$\text{sample total variance} = V = \sum_{j=1}^{p}\sum_{i=1}^{n} \frac{(X_{ij} - \overline{X}_j)^2}{n-1}$$

We now characterize the amount of variability explained by the principal components. Recall that the principal components are themselves variables; they are linear combinations of the X_j variables. Each principal component has a variance itself. It is natural, therefore, to compare the variance of the principal components with the variance of the X_j's. This leads us to the following definitions.

Definition 14.4. Let $Y_1, Y_2, \ldots$ be the first, second, and subsequent principal components for the variables $X_1, \ldots, X_p$. In a sample the variance of each Y_k is estimated by

$$\text{var}(Y_k) = \sum_{i=1}^{n} \frac{(Y_{ik} - \overline{Y}_k)^2}{n-1} = V_k \tag{5}$$

where Y_{ik} is the value of the kth principal component for the ith observation. That is, we first estimate the coefficients for the kth principal component. The value for the ith observation uses those coefficients and the observed values of the X_j's to compute the value of Y_{ik}. The variance for the kth principal component in the sample is then given by the sample variance for Y_{ik}, $i = 1, 2, \ldots, n$. We denote this variance as seen above by V_k. Using this notation, we have the following two definitions:

1. *The percent of variability explained by the* kth *principal component is*

$$\frac{100V_k}{V}$$

2. *The percent of the variability explained by the first m principal components is*

$$100\sum_{k=1}^{m} \frac{V_k}{V} \tag{6}$$

The following facts about the principal components can be stated:

1. There are exactly p principal components, where p is the number of X variables considered. This is because with p uncorrelated variables, there is a one-to-one correspondence between the values of the principal components and the values of the original data; that is, we can go back and forth so that all of the variability is accounted for; the percent of variability explained by the p principal components is 100%.
2. Because we chose the principal components successively to explain more and more of the variance, we have

$$V_1 \geq V_2 \geq \cdots \geq V_p \geq 0$$

3. The first m principal components explain as much of the total variability as it is possible to explain by m linear functions of the X_j variables.

We now proceed to a geometric interpretation of the principal components. Consider the case where $p = 2$. That is, we observe two variables X_1 and X_2. Plot, as previously in this chapter, the ith data point in the coordinate system that is centered about the means for the X_1 and X_2 variables. Draw a line in the direction of the first principal component and project the data point onto the line. This is done in Figure 14.5.

The square of the distance of the data point from the new origin, which is the sample mean, is given by the following equation, using the Pythagorean theorem:

$$d_i^2 = (X_{i1} - \overline{X}_1)^2 + (X_{i2} - \overline{X}_2)^2 = \sum_{j=1}^{2}(X_{ij} - \overline{X}_j)^2$$

The square of the distance f_i of the projection turns out to be the difference between the value of the first principal component for the ith observation and the mean of the first principal component squared. That is,

$$f_i^2 = (Y_{i1} - \overline{Y}_1)^2$$

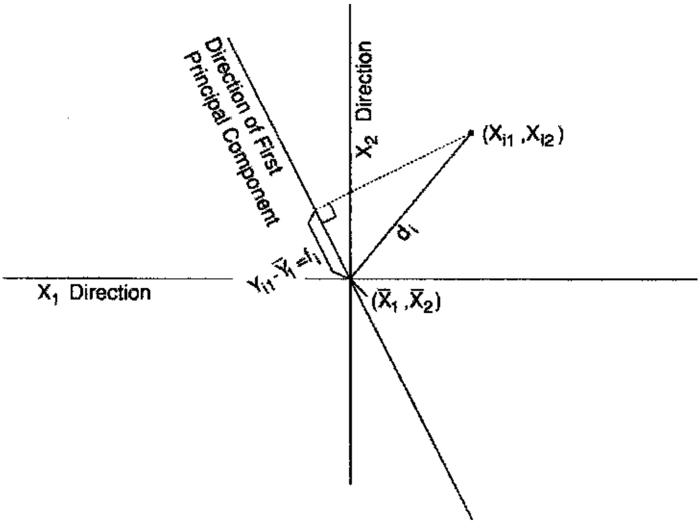

Direction of First Principal Component
X_2 Direction
(X_{i1}, X_{i2})
d_i
$Y_{i1} - \overline{Y}_1 = f_i$
X_1 Direction
$(\overline{X}_1, \overline{X}_2)$

Figure 14.5 Projection of a data point onto the first principal component direction.

It is geometrically clear that the distance d_i is larger than f_i. The ith data point will be better represented by its position along the line if it lies closer to the line, that is, if f_i is close to d_i. One way we might judge the adequacy of the variability explained by the first principal component would be to take the ratio of the sum of the lengths of the f_i's squared to the sum of the lengths of the d_i's squared. If we do this, we have

$$\frac{\sum_{i=1}^{n} f_i^2}{\sum_{i=1}^{n} d_i^2} = \frac{\sum_{i=1}^{n}(Y_{i1} - \overline{Y}_1)^2}{\sum_{i=1}^{n}\sum_{j=1}^{2}(X_{ij} - \overline{X}_j)^2} = \frac{V_1}{V} \tag{7}$$

That is, we have the proportion of the variability explained. If we multiplied the equation throughout by 100, we would have the percent of the variability explained by the first principal component. This gives us an alternative way of characterizing the first principal component. The direction of the first principal component is the line for which the following holds: When the data are projected onto this line, the sum of the squares of the projections is as large as possible; equivalently, the sum of squares is as close as possible to the sum of squares of the lengths of the lines to the original data points from the origin (which is also the mean). From this we see that the percent of variability explained by the first principal component will be 100 if and only if the lengths d_i and f_i are all the same; that is, the first principal component will explain all the variability if and only if all of the data points lie on a single line. The closer all the data points come to lie on a single line, the larger the percent of variability explained by the first principal component.

We now proceed to examine the geometric interpretation in three dimensions. In this case we consider a data point plotted not in terms of the original axes X_1, X_2, and X_3 but rather, in terms of the coordinate system given by the principal components Y_1, Y_2, and Y_3. Figure 14.6 presents such a plot for a particular data point. The figure is a two-dimensional representation of a three-dimensional situation; two of the axes are vertical and horizontal on the paper. The third axis recedes into the plane formed by the page in this book. Consider the ith data point,

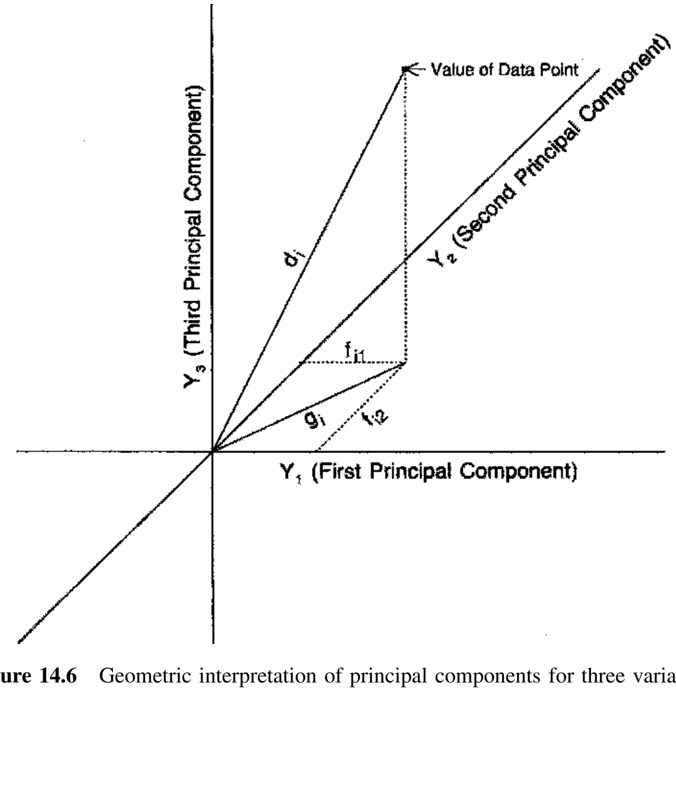

Value of Data Point
Y₂ (Second Principal Component)
Y₃ (Third Principal Component)
dᵢ
fᵢ1
gᵢ
fᵢ2
Y₁ (First Principal Component)

Figure 14.6 Geometric interpretation of principal components for three variables.

which lies at a distance d_i from the origin that is at the mean of the data points. This point also turns out to be the mean of the principal component values. Suppose, now, that we drop a line down into the plane that contains the axes corresponding to the first two principal components. This is indicated by the vertical dotted line in the figure. This point in the plane we could now project onto the value for the first and second principal components. These values, with lengths f_{i1} and f_{i2}, are the same as we would get by dropping perpendiculars directly from the point to those two axes. Again, we might assess the adequacy of the characterization of the data point by the first two principal components by comparing the length of its projection in the plane, g_i, with the length of the line from the origin to the original data point, d_i. If we compare the squares of these two lengths, each summed over all of the data points, and use the Pythagorean theorem again, the following results hold:

$$\begin{aligned}\frac{\sum_{i=1}^{n} g_i^2}{\sum_{i=1}^{n} d_i^2} &= \frac{\sum_{i=1}^{n} f_{i1}^2 + \sum_{i=1}^{n} f_{i2}^2}{\sum_{i=1}^{n} d_i^2} \\ &= \frac{\sum_{i=1}^{n}[(Y_{i1} - \overline{Y}_1)^2/(n-1)] + \sum_{i=1}^{n}[(Y_{i1} - \overline{Y}_2)^2/(n-1)]}{\sum_{i=1}^{n} d_i^2/(n-1)} \\ &= \frac{V_1 + V_2}{V}\end{aligned}$$

Using this equation, we see that the percent of the variability explained by the first two principal components is the ratio of the squared lengths of the projections onto the plane of the first two principal components divided by the squared lengths of the original data points about their mean. This also gives us a geometric interpretation of the total variance. It is the sum for all the data points of the squares of the distance between the point corresponding to the mean of the sample and the original data points. In other words, the first two principal components may be characterized as giving a plane for which the projected points onto the plane contain as high a proportion as possible of the squared lengths associated with the original data points. From this we see that the percent of variability explained by the first two principal components will be 100 if and only if all of the data points lie in some plane through the origin, which is the mean of the data.

The coefficients associated with the principal components are usually calculated by computer; in general, there is no easy formula to obtain them. Thus, the examples in this chapter will begin with the coefficients for the principal components and their variance. (There is an explicit solution when there are only two variables, and this is given in Problem 14.9.)

Example 14.1. We turn to the data of Table 14.1. Equations for the principal components are

$$Y_1 = -0.6245X + 0.7809Y$$

$$Y_2 = 0.7809X + 0.6245Y$$

For the first data point, $(X, Y) = (-0.52, 0.60)$, the values are

$$Y_1 = -0.6245 \times (-0.52) + 0.7809 \times 0.60 = 0.79$$

$$Y_2 = 0.7809 \times (-0.52) + 0.6245 \times 0.60 = -0.03$$

If we compute all of the numbers, we find that the values for each of the 20 data points on the principal components are as given in Table 14.2.

Table 14.2 Data Point Values

Data		Principal Component Values		Data		Principal Component Values	
X	Y	Y_1	Y_2	X	Y	Y_1	Y_2
−0.52	0.60	0.79	−0.03	0.08	0.23	0.13	0.21
0.04	−0.51	−0.42	−0.28	−0.06	−0.59	−0.42	−0.42
1.29	−1.19	−1.74	0.26	1.25	−1.25	−1.76	0.19
−1.12	1.90	2.19	0.31	0.53	−0.45	−0.68	0.13
−1.02	0.31	0.88	−0.60	0.14	0.47	0.28	0.40
0.10	−1.15	−0.96	−0.64	0.48	−0.11	−0.39	0.31
−0.32	−0.13	0.10	−0.33	−0.61	1.04	1.20	0.17
0.08	−0.17	−0.18	−0.04	−0.47	0.34	0.56	−0.16
0.49	0.18	−0.16	0.50	0.41	0.29	−0.02	0.50
0.54	0.20	0.49	−0.29	−0.22	−0.00	0.13	−0.18

From these data we may compute the sample variance of Y_1 and Y_2 as well as the variance of X and Y. We find the following values:

$$V_1 = 0.861, \qquad V_2 = 0.123, \qquad \text{var}(X) = 0.411, \qquad \text{var}(Y) = 0.573$$

From these data we may compute the percent of variability explained by the two principal components, individually and together.

1. Percent of variability explained by the first principal component $= 100 \times 0.861/(0.411 + 0.573) = 87.5\%$.
2. Percent of variability explained by the second principal component $= 100 \times 0.123/(0.411 + 0.573) = 12.5\%$.
3. Percent of variability explained by the first two principal components $= 100 \times (0.861 + 0.123)/(0.411 + 0.573) = 100\%$.

We see that the first principal component of the data in Figure 14.4 contains a high proportion of the variability. This may also be seen visually by examining the plot while orienting your eyes so that the horizontal line is the direction of the first principal component. Certainly, there is much more variability in that direction than in direction 2, the direction of the second principal component.

14.5 USE OF THE COVARIANCE, OR CORRELATION, VALUES AND PRINCIPAL COMPONENT ANALYSIS

The coefficients of the principal components and their variances can be computed by knowing the covariances between the X_j's. One might think that as a general search for relationships among X_j's, the principal component will be appropriate as an exploratory tool. Sometimes, this is true. However, consider what happens when we have different scales of measurement. Suppose, for example, that among our units, one unit is height in inches and another is systolic blood pressure in mmHg. In principal component analysis we are adding the variability in the two variables. Suppose now that we change our measurement of height from inches to feet. Then the standard deviation of the height variable will be divided by 12 and the variance will be divided by 144. In the total variance the contribution of height will have dropped greatly.

Equivalently, the blood pressure contribution (and any other variables) will become much more important. Recomputing the principal components will produce a different answer. In other words, the measurement units are important in finding the principal component because the variance of any individual variable is compared directly to the variance of another variable without regard to whether or not the units are appropriate for the comparison. We reiterate: *The importance of a variable in principal component analysis changes with a change of scale of one or more of the variables*. For this reason, principal component analysis is most appropriate and probably has its best applications when all the variables are measured in the same units; for example, the X_j variables may be measurements of length in inches, with the variables being measurements of different parts of the body, and the covariances between variables such as arm length, leg length, and body length.

In some situations with differing units, one still wants to try principal component analyses. In this case, standardized variables are often used; that is, we divide each variable by its standard deviation. Each rescaled variable then has a variance of 1 and the covariance matrix of the new standardized variables is the correlation matrix of the original variables. The interpretation of the principal components is now less clear. If many of the variables are highly correlated, the first principal component will tend to pick up this fact; for example, with two variables, a high correlation means the variables lie along a line. The ellipse of concentration has one axis along the line; that direction gives us the direction of the first principal component. When standardized variables are used, since each variable has a variance of 1, the sum of the variances is p. In looking at the percent of variability explained, there is no need to compute the total variance separately; it is p, the number of variables. We emphasize that when the correlations are used, there should be some reason for doing this beside the fact that the variables do not have measurements in comparable units.

14.6 STATISTICAL RESULTS FOR PRINCIPAL COMPONENT ANALYSIS

Suppose that we have a sample of size n from a multivariate normal distribution with unknown covariances. Let $V_i(\text{pop})$ be the true (unknown) population value for the variance of the ith principal component when computed from the (unknown) true variances; let V_i be the variance of the principal components computed from the sample covariances. Then the following are true:

1.
$$\frac{V_i - V_i(\text{pop})}{V_i(\text{pop})\sqrt{2/(n-1)}}, \qquad i = 1, \ldots, p \tag{8}$$
for large n is approximately a standard normal, $N(0, 1)$, random variable. These variables are approximately statistically independent.

2. $100(1-\alpha)\%$ confidence intervals for $V_i(\text{pop})$ for large n are given by
$$\left(\frac{V_i}{1 + z_{1-\alpha/2}\sqrt{2/(n-1)}}, \frac{V_i}{1 - z_{1-\alpha/2}\sqrt{2/(n-1)}}\right) \tag{9}$$
where $z_{1-\alpha/2}$ is the $1-\alpha/2$ percentile value of the $N(0, 1)$ distribution.

Further statistical results on principal component analysis are given in Morrison [1976] and Timm [1975].

Principal component analysis is a least squares technique, as were analysis of variance and multiple linear regression. Outliers in the data can have a large effect on the results (as in other cases where least squares techniques are used).

14.7 PRESENTING THE RESULTS OF A PRINCIPAL COMPONENT ANALYSIS

We have seen that principal component analysis is designed to explain the variability in data. Thus, any presentation should include:

1. The variance of the principal components
2. The percent of the total variance explained by each individual principal component
3. The percent of the total variance explained cumulatively by the first m terms (for each m)

It is also useful to know how closely each variable X_j is related to the values of the principal components Y_i; this is usually done by presenting the correlations between each variable and each of the principal components. Let

$$Y_i = a_{i1}X_1 + \cdots + a_{ip}X_p$$

The correlation between one of the original variables X_j and the kth principal component Y_i is given by

$$r_{jk} = \text{correlation of } X_j \text{ and } Y_k = \frac{a_{kj}\sqrt{V_k}}{s_j} \tag{10}$$

In this equation, V_i is the variance of the ith principal component, while s_j is the standard deviation of X_j. These results are summarized in Table 14.3.

By examining the variables that are highly correlated with a principal component, we can see which variables contribute most to the principal component. Alternatively, glancing across the rows for each variable X_j we may see which principal component has the highest correlation with the variable. An X_i that has the highest correlations with the first few principal components is contributing more to the overall variability than variables with small correlations with the first few principal components. In Section 14.9, several examples of principal component analysis are given, including an example of the use of such a summary table (Table 14.4).

Table 14.3 Summary of a Principal Component Analysis Using Covariances

	Correlation of the Principal Components and the X_j's				
Variables	1	2	$\cdots$	p	Standard Deviations of the X_j
X_1	$\frac{a_{11}\sqrt{V_1}}{s_1}$	$\cdots$	$\cdots$	$\frac{a_{p1}\sqrt{V_p}}{s_1}$	s_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_p	$\frac{a_{1p}\sqrt{V_1}}{s_p}$	$\cdots$	$\cdots$	$\frac{a_{pp}\sqrt{V_p}}{s_p}$	s_p
Variance of principal component	V_1	V_2	$\cdots$	V_p	
% of total variance	$\frac{100V_1}{V}$	$\cdots$	$\cdots$	$\frac{100V_p}{V}$	
Cumulative % of total variance	$\frac{100V_1}{V}$	$\frac{100(V_1+V_2)}{V}$	$\cdots$	1	

Table 14.4 Data for Example 14.2

Principal Component	Variance Explained	Percent of Total Variance	Cumulative Percent of Total Variance
1	7.82	41.1	41.1
2	4.46	23.5	64.6
3	1.91	10.1	74.7
4	0.88	4.6	79.4
5	0.76	4.0	83.3
6	0.56	2.9	86.3
7	0.45	2.4	88.6
8	0.38	2.0	90.7
9	0.35	1.9	92.5
10	0.31	1.6	94.1
11	0.19	1.0	95.1
12	0.18	0.9	96.1
13	0.16	0.8	96.9
14	0.14	0.7	97.7
15	0.13	0.7	98.3
16	0.10	0.5	98.9
17	0.10	0.5	99.4
18	0.06	0.3	99.7
19	0.05	0.3	100.0

14.8 USES AND INTERPRETATION OF PRINCIPAL COMPONENT ANALYSIS

Principal component analysis is a technique for explaining variability. Following are some of the uses of principal components:

1. Principal component analysis is a search for linear relationships for explaining variability in a multivariate sample. The first few principal components are important because they may summarize a large proportion of the variability. However, the understanding of which variables contribute to the variability is important only if most of the variance comes about because of important relationships among the variables. After all, we can increase the variance of a variable, say X_1, by increasing the error of measurement. If we have a phenomenally large error of measurement, the variance of X_1 will be much larger than the variances of the rest of the variables. In this case, the first principal component will be approximately equal to X_1, and the amount of variability explained will be close to 1. However, such knowledge is not particularly useful, since the variability in X_1 does not make X_1 the most important variable, but in this case, reflects a very poorly measured quantity. Thus, to decide that the first few principal components are important summary variables, you must feel that the relationships among them come from linear relationships which may shed some light on the data being studied.

2. In some cases the first principal component is relatively uninteresting, with more informative relationships being found in the next few components. One simple case comes from analyzing physical measurements of plants or animals to display species differences: the first principal component may simply reflect differences in size, and the next few components give the more interesting differences in shape.

3. We may take the first two principal components and plot the values for the first two principal components of the data points. We know that among all possible plots in only two dimensions, this one gives the best fit in one precise mathematical sense. However, it should be noted that other techniques of multivariate analysis give two-dimensional plots that are the best fit or most interesting in other precise mathematical senses (see Note 14.1).

4. In some situations we have many measurements of somewhat related variables. For example, we might have a large number of size measurements on different portions of the human body. It may be that we want to perform a statistical inference, but the large number of variables for the relatively small number of cases involved makes such statistical analysis inappropriate. We may summarize the data by using the values on the first few principal components. *If the variability is important* (!), we have then reduced the number of variables without getting involved in multiple comparison problems. We may proceed to statistical analysis. For example, suppose that we are trying to perform a discriminant analysis and want to use size as one of the discriminating variables. However, for each of a relatively small number of cases we may have many anthropometric measurements. We might take the first principal component as a variable to summarize all the size relationships. One of the examples of principal component analysis below gives a principal component analysis of physical size data.

14.9 PRINCIPAL COMPONENT ANALYSIS EXAMPLES

Example 14.2. Stoudt et al. [1970] consider measurements taken on a sample of adult females from the United States. The correlations among these measurements (as well as weight and age) are given in Table 11.21. The variance explained for each principal component is presented in Table 14.4.

These data are very highly structured. Only three (of 19) principal components explain over 70% of the variance. Table 14.5 summarizes the first three principal components. The

Table 14.5 Example 14.2: First Three Principal Components

	Correlation of the Principal Components and the Variables		
Variables	1	2	3
SITHTER	0.252	0.772	0.485
SITHTNORM	0.235	0.748	0.470
KNEEHT	0.385	0.722	−0.392
POPHT	0.005	0.759	−0.444
ELBOWHT	0.276	0.243	0.783
THIGHHT	0.737	−0.007	0.204
BUTTKN	0.677	0.476	−0.348
BUTTPOP	0.559	0.411	−0.444
ELBOWBR	0.864	−0.325	−0.033
SEATBR	0.832	−0.050	0.096
BIACROM	0.504	0.350	−0.053
CHEST	0.890	−0.228	−0.018
WAIST	0.839	−0.343	−0.106
ARMGTH	0.893	−0.267	0.068
ARMSKIN	0.733	−0.231	0.124
INFRASCA	0.778	−0.371	0.056
HT	0.251	0.923	−0.051
WT	0.957	−0.057	0.001
AGE	0.222	−0.488	−0.289
Variance of principal components	7.82	4.46	1.91
Percent of total variance	41.1	23.5	10.1
Cumulative percent of total variance	41.1	64.6	74.7

first component, in the direction of greatest variation, is associated heavily with the weight variables. The highest correlation is with weight, 0.957. Other variables associated with size—such as chest and waist measurements, arm girth, and skinfolds—also are highly correlated with the first principal component. The second component is most closely associated with physical length measurements. Height is the most highly correlated variable. Other variables with correlations above 0.7 are the sitting heights (normal and erect), knee height, and popliteal height.

Since we are working with a correlation matrix, the total variance is 19, the number of variables. The average variance, in fact the exact variance, per variable is 1. Only these first three principal components have variance greater than 1. The other 16 directions correspond to a variance of less than 1.

Example 14.3. Reeck and Fisher [1973] performed a statistical analysis of the amino acid composition of protein. The mole percent of the 18 amino acids in a sample of 207 proteins was examined. The covariances and correlations are given in Table 14.6. The diagonal entries and numbers above them give the variances and covariances; the lower numbers are the correlations. The mnemonics are:

Asp	Aspartic acid	Met	Methionine
Thr	Threonine	Ile	Isoleucine
Ser	Serine	Leu	Leucine
Glu	Glutamic acid	Tyr	Tyrosine
Pro	Proline	Phe	Phenylalanine
Gly	Glycine	Trp	Tryptophan
Ala	Alanine	Lys	Lysine
Cys/2	Half-cystine	His	Histidine
Val	Valine	Arg	Arginine

The principal component analysis applied to the data produced Table 14.7, where k is the dimension of the subspace used to represent the data and C is the proportion of the total variance accounted for in the best k-dimensional representation.

In contrast to Example 14.2, eight principal components are needed to account for 70% of the variance. In this example there are no simple linear relationships (or directions) that account for most of the variability. In this case the principal component correlations are not presented, as the results are not very useful.

14.10 FACTOR ANALYSIS

As in principal component analysis, factor analysis looks at the relationships among variables as expressed by their correlations or covariances. While principal component analysis is designed to model and explain as much of the variability as possible, factor analysis seeks to explain the relationships among the variables. The assumption of the model is that the relationships may be explained by a few unobserved variables, which will be called *factors*. It is hoped that fewer factors than the original number of variables will be needed to explain the relationships among the variables. Thus, conceptually, one may simplify the understanding of the correlations between the variables.

It is difficult to present the technique without having the model and many of the related issues discussed first. However, it is also difficult to understand the related issues without examples. Thus, it is suggested that you read through the material about the mathematical model, go through the examples, and then with this understanding, reread the material about the mathematical model.

Table 14.6 Example 14.3: Reeck and Fisher [1973] Covariance/Correlation Matrix[a]

	Asp	Thr	Ser	Glu	Pro	Gly	Ala	Cys/2	Val	Met	Ile	Leu	Tyr	Phe	Trp	Lys	His	Arg
Asp	6.5649	0.2449	0.7879	−1.5329	−1.9141	−1.8328	−1.7003	−0.4974	−0.1374	0.0810	0.6332	−1.0855	0.6413	0.1879	0.3873	0.7336	0.0041	−1.5633
Thr	0.0517	3.4209	1.3998	−1.3341	−0.3531	−0.7752	−0.6428	0.4468	0.3603	−0.3502	0.1620	−1.2836	0.1804	−0.0978	0.1114	−0.3348	−0.2594	−0.8938
Ser	0.1219	0.2999	6.3687	−1.6465	0.1876	−0.8922	−1.3593	−0.3123	0.6659	−0.6488	−0.3738	−1.1125	0.4403	0.0432	0.2552	−1.6972	−0.3025	−1.4289
Glu	−0.1789	−0.2157	−0.1951	11.1880	−0.5866	−2.1665	−0.7732	−0.1443	−1.5346	0.0002	−0.3804	1.6210	−1.1824	−0.6684	−0.6778	0.0192	−0.3154	0.1169
Pro	−0.3566	−0.0911	−0.0355	−0.0837	4.3891	1.4958	−0.4259	1.0159	−0.7017	−0.4171	−0.8453	−0.9980	−0.0868	−0.1187	0.1163	−0.7021	−0.1612	0.4801
Gly	−0.2324	−0.1362	−0.1149	−0.2105	0.2320	9.4723	1.2857	0.1737	−0.3883	−0.4226	−0.2812	−2.3936	−0.8971	−0.7784	−0.2637	−1.0861	−0.2526	−0.0037
Ala	−0.2417	−0.1266	−0.1962	−0.0842	−0.0741	0.1522	7.5371	−2.1250	0.8498	0.1810	−0.4183	1.2480	−1.3374	−0.4320	−0.5219	−1.1641	−0.2730	0.0701
Cys/2	−0.0717	0.0892	−0.0457	−0.0159	0.1790	0.0208	−0.2857	7.3393	−1.3667	−0.4788	−1.3959	−2.3443	0.5408	−0.6282	0.1136	0.2727	−0.7482	0.1447
Val	−0.0275	0.1001	0.1356	−0.2357	−0.1721	−0.0648	0.1590	−0.2592	3.7885	−0.0632	0.5700	0.2767	−0.1348	−0.2303	−0.2792	−0.7921	−0.0632	−0.8223
Met	0.0294	−0.1759	−0.2388	0.0001	−0.1849	−0.1275	0.0612	−0.1642	−0.0302	1.1589	0.2493	0.2438	−0.1397	0.2060	−0.0159	0.1715	0.1457	0.0945
Ile	0.1426	0.0505	−0.0855	−0.0656	−0.2328	−0.0527	−0.0879	−0.2974	0.1690	0.1337	3.0023	−0.1857	−0.2785	−0.0870	−0.1296	0.2361	−0.0829	−0.3956
Leu	−0.1701	−0.2786	−0.1770	0.1946	−0.1912	−0.3122	0.1825	−0.3474	0.0571	0.0928	−0.0430	6.2047	−1.0362	0.2515	−0.2332	−0.6337	0.3951	1.0593
Tyr	0.1605	0.0625	0.1119	−0.2267	−0.0266	−0.1869	−0.3123	0.1280	−0.0444	−0.0832	−0.1031	−0.2667	0.1823	0.1659	0.9201	−0.5061	0.0855	0.1436
Phe	0.0525	−0.0379	0.0123	−0.1431	−0.0406	−0.1811	−0.1126	−0.1660	−0.0847	0.1370	−0.0360	0.0723	0.2262	1.9512	0.2223	−0.8382	0.3434	0.1796
Trp	0.1576	0.0628	0.1054	−0.2113	0.0579	−0.0893	−0.1982	0.0437	−0.1495	−0.0154	−0.0780	−0.0976	0.1823	0.1659	0.9201	−0.5061	0.0855	0.1436
Lys	0.1061	−0.0670	−0.2491	0.0021	−0.1241	−0.1307	−0.1571	0.0373	−0.1507	0.0590	0.0505	−0.0942	0.0733	−0.2223	−0.1954	7.2884	−0.1830	−1.0898
His	0.0014	−0.1194	−0.1020	−0.0803	−0.0655	−0.0699	−0.0847	−0.2351	−0.0276	0.1152	−0.0408	0.1350	0.0314	0.2093	0.0759	0.0577	1.3795	0.2280
Arg	−0.3068	−0.2430	−0.2847	0.0176	0.1152	−0.0006	0.0128	0.0269	−0.2124	0.0441	−0.1148	0.2138	−0.0882	0.0646	0.0753	−0.2030	0.0976	3.9550

[a] Diagonal and upper entries are variances and covariances. Below the diagonal are the correlations.

Table 14.7 Principal Component Analysis Data

k	C	k	C	k	C
1	0.13	7	0.66	13	0.90
2	0.26	8	0.70	14	0.93
3	0.37	9	0.75	15	0.95
4	0.46	10	0.79	16	0.98
5	0.55	11	0.83	17	1.00
6	0.61	12	0.86	18	1.00

We now turn to the model. We observe jointly distributed random variable $X_1, \ldots, X_p$. The assumption is that each X is a linear sum of the factors plus some remaining residual variability. That is, the model is the following:

$$\begin{array}{ccccccccccccc} X_1 & = & E(X_1) & + & \lambda_{11}F_1 & + & \lambda_{12}F_2 & + & \cdots & + & \lambda_{1k}F_k & + & e_1 \\ \vdots & & \vdots & & \vdots & & \vdots & & & & \vdots & & \vdots \\ X_p & = & E(X_p) & + & \lambda_{p1}F_1 & + & \lambda_{p2}F_2 & + & \cdots & + & \lambda_{pk}F_k & + & e_p \end{array} \qquad (11)$$

In this model, each X_i is equal to its expected value, plus a linear sum of k factors and a term for residual variability. This looks like a series of multiple regression equations; each of the variables X_i is regressed on the variables $F_1, \ldots, F_k$. There are, however, major differences between this model and the multiple regression model of Chapter 11. Observations and assumptions about this model are the following:

1. The factors F_j are *not* observed; only the $X_1, \ldots, X_p$ are observed, although the X_i variables are expressed in terms of these smaller number of factors F_j.
2. The e_i (which are also unobserved) represent variability in the X_i not explained by the factors. We do *not* assume that these residual variability terms have the same distribution.
3. Usually, the number of factors k is unknown and must be determined from the data. We shall first consider the model and the analysis where the number of factors is known; later, we consider how one might search for the appropriate number of factors.

Assumptions made in the model, in addition to the linear equations given above, are the following:

1. The factors F_j are standardized; that is, they have mean zero and variance 1.
2. The factors F_j are uncorrelated with each other, and they are uncorrelated with the e_i terms. See Section 14.12 for a relaxation of this requirement.
3. The e_i's have mean zero and are uncorrelated with each other as well as with the F_j's. They may have different variances.

It is a fact that if p factors F are allowed, there is no need for the residual variability terms e_i. One can reproduce any pattern of covariances or correlations using p factors when p variables X_i are observed. This, however, is not very useful because we have summarized the p variables which were observed with p unknown variables. Thus, in general, we will be interested in k factors, where k is less than p.

Let ψ_i be the variance of e_i. With the assumptions of the model above, the variance of each X_i can be expressed in terms of the coefficients λ_{ij} of the factors and the residual variance ψ_i.

The equation giving the relationship for k factors is

$$\operatorname{var}(X_i) = \lambda_{i1}^2 + \cdots + \lambda_{ik}^2 + \psi_i \tag{12}$$

In words, the variance of each X_i is the sum of the squares of the coefficients of the factors, plus the variance of e_i. The variance of X_i has two parts. The sum of the coefficients λ_{ij} squared depends on the factors; the factors contribute in common to all of the X_i's. The e_i's correlate only with their own variable X_i and not with other variables in the model. In particular, they are uncorrelated with all of the X_i's except for the one corresponding to their index. Thus, we have broken down the variance into a part related to the factors that each variable has in common, and the unique part related to the residual variability term. This leads to the following definition.

Definition 14.5. $c_i = \sum_{j=1}^{k} \lambda_{ij}^2$ is called the *common part of the variance* of X_i, c_i is also called the *communality* of X_i, ψ_i is called the *unique* or *specific part of the variance* of X_i, and ψ_i is also called the *uniqueness* or *specificity*.

Although factor analysis is designed to explain the relationships between the variables and not the variance of the individual variables, if the communalities are large compared to the specificities of the variables, the model has also succeeded in explaining not only the relationships among the variables but the variability in terms of the common factors.

Not only may the variance be expressed in terms of the coefficients of the factors, but the covariance between any two variables may also be expressed by

$$\operatorname{cov}(X_i, X_j) = \lambda_{i1}\lambda_{j1} + \cdots + \lambda_{ik}\lambda_{jk} \qquad \text{for } i \neq j \tag{13}$$

These equations explain the relationships among the variables. If both X_i and X_j have variances equal to 1, this expression gives the correlation between the two variables. There is a standard name for the coefficients of the common factors.

Definition 14.6. The coefficients λ_{ij} are called the *factor loadings* or *loadings*. λ_{ij} represents the loading of variable X_i and factor F_j.

In general, $\operatorname{cov}(X_i, F_j) = \lambda_{ij}$. That is, λ_{ij} is the covariance between X_i and F_j. If X_i has variance 1, for example if it is standardized, then since F_j has variance 1, the factor loading is the correlation coefficient between the variable and the factor.

We illustrate the method by two examples.

Example 14.4. We continue with the measurement data of U.S. females of Example 14.2. A factor analysis with three underlying factors was performed on these data. Since we are trying to explain the correlations between the variables, it is useful to examine the fit of the model by comparing the observed and modeled correlations. We do this by examining the residual correlation.

Definition 14.7. The *residual correlation* is the observed correlation minus the fitted correlation from the factor analysis model.

Table 14.8 gives the residual correlations below the diagonal; on the diagonal are the estimated uniquenesses, the part of the (standardized) variance not explained by the three factors.

A rule of thumb is that the correlation has been explained reasonably when the residual is less than 0.1 in absolute value. This is convenient because it is easy to scan the residual matrix for a zero after a decimal point. Of course, depending on the purpose, more stringent requirements may be considered.

Table 14.8 Residual Correlations: Example 14.4

		STHTER	STHTNORM	KNEEHT	POPHT	ELBOWHT
		1	2	3	4	5
STHTER	1	0.034				
STHTNORM	2	0.002	0.151			
KNEEHT	3	−0.001	0.001	0.191		
POPHT	4	0.001	0.002	0.048	0.276	
ELBOWHT	5	−0.001	−0.011	0.011	−0.004	0.474
THIGHHT	6	−0.009	0.004	0.003	−0.076	0.035
BUTTKN	7	−0.002	0.000	−0.016	−0.056	−0.021
BUTTPOP	8	−0.002	0.011	−0.042	−0.064	−0.035
ELBOWBR	9	0.000	0.013	−0.004	0.014	−0.010
SEATBR	10	−0.002	0.013	0.016	−0.041	0.020
BIACROM	11	0.004	−0.005	−0.000	0.014	−0.089
CHEST	12	0.003	0.004	0.003	0.030	−0.015
WAIST	13	0.005	−0.004	0.002	0.032	0.006
ARMGTH	14	−0.001	−0.004	0.004	−0.009	0.003
ARMSKIN	15	−0.005	0.016	0.025	−0.012	−0.004
INFRASCA	16	−0.002	0.006	0.020	0.016	0.004
HT	17	0.000	−0.001	−0.000	0.003	0.008
WT	18	−0.000	−0.009	−0.004	−0.005	0.008
AGE	19	0.002	0.024	0.003	0.024	−0.042

		THIGHHT	BUTTKN	BUTTPOP	ELBOWBR	SEATBR
		6	7	8	9	10
THIGHHT	6	0.499				
BUTTKN	7	0.062	0.251			
BUTTPOP	8	0.040	**0.136**	0.425		
ELBOWBR	9	−0.012	−0.017	−0.016	0.158	
SEATBR	10	0.035	0.070	0.010	−0.016	0.338
BIACROM	11	0.049	−0.035	−0.039	0.012	−0.042
CHEST	12	−0.038	−0.044	−0.017	0.036	−0.056
WAIST	13	−0.067	−0.023	−0.021	0.037	−0.029
ARMGTH	14	0.005	0.005	0.007	−0.014	0.008
ARMSKIN	15	0.048	0.019	0.021	−0.030	0.047
INFRASCA	16	0.004	−0.025	−0.007	−0.003	−0.030
HT	17	−0.003	−0.001	0.001	0.004	−0.014
WT	18	0.017	0.009	−0.004	−0.011	0.019
AGE	19	−**0.172**	−0.056	−0.034	0.078	0.002

		BIACROM	CHESTGRH	WSTGRTH	RTARMGRH	RTARMSKN
		11	12	13	14	15
BIACROM	11	0.679				
CHEST	12	0.072	0.148			
WAIST	13	−0.008	0.032	0.172		
ARMGTH	14	−0.014	−0.014	−0.031	0.134	
ARMSKIN	15	−0.053	−0.041	−0.046	0.075	0.487
INFRASCA	16	−0.010	0.013	0.003	0.013	**0.171**
HT	17	0.002	−0.000	−0.002	−0.001	0.003
WT	18	−0.003	0.000	0.004	0.009	−0.030
AGE	19	−**0.106**	0.033	0.105	−0.017	−0.012

		INFRASCA	HT	WT	AGE
		16	17	18	19
INFRASCA	16	0.317			
HT	17	0.002	0.056		
WT	18	−0.018	0.001	0.057	
AGE	19	−0.017	0.016	−0.034	0.770

Table 14.9 Factor Loadings for a Three-Factor Model: Example 14.4

		Factor Loadings (Pattern)[a]		
Variable	Number	Factor 1	Factor 2	Factor 3
SITHTER	1		0.346	0.920
SITHTNORM	2		0.332	0.859
KNEEHT	3		0.884	0.146
POPHT	4	−0.271	0.801	
ELBOWHT	5	0.222	−0.120	0.680
THIGHHT	6	0.672	0.125	0.181
BUTTKN	7	0.436	0.741	
BUTTPOP	8	0.339	0.679	
ELBOWBR	9	0.914		
SEATBR	10	0.781	0.171	0.150
BIACROM	11	0.344	0.390	0.225
CHEST	12	0.916	0.114	
WAIST	13	0.898		−0.126
ARMGTH	14	0.929		
ARMSKIN	15	0.714		
INFRASCA	16	0.823		
HT	17		0.804	0.538
WT	18	0.929	0.265	0.103
AGE	19	0.328	−0.124	−0.328
VP		7.123	3.632	2.628
Proportion var.		0.375	0.191	0.138
Cumulative var.		0.375	0.566	0.704

[a] Loadings less than 0.1 have been omitted.

In this example there are four large absolute values of residuals (−0.172, 0.171, 0.136, and −0.106). This suggests that more factors are needed. (In Problem 14.10 we consider analysis of these data with more factors.) The factor loadings are presented in Table 14.9. Loadings below 0.1 in absolute value are omitted, making it easier to see which variables are related to which factors. In this example the first factor has high loadings on weight and bulk measurements (variables 14, 18, 12, 9, 13, 16, 10, 15, and 6) and might be called a *weight* factor. The second factor has high loadings on length or height measurements (variables 3, 17, 4, 7, and 8) and might be considered a *height* factor. The third factor seems to be a *sitting height* factor.

The variables have been reordered so that variables loading on the same factor appear together. When this is done, clusters of correlated variables often appear, which may be appreciated visually by replacing correlations by symbols or colors. Figure 14.7 is a graph of the correlation data from Table 11.21 using circles whose radius is proportional to the correlation, shaded light gray for positive correlations and dark gray for negative correlations.

The sum of the squares of loadings for a factor (VP) is the portion of the sum of the X_i variances (the total variance) that is explained by the factor. The table also gives this as a proportion of the total and as a cumulative proportion of the total. In all, these factors explain 70% of the variability in the measurements.

Example 14.5. As a second example, consider coronary artery disease patients with left main coronary artery disease. This patient group was discussed in Chaitman et al. [1981]. In this factor analysis, 12 variables were considered and four factors were used with 357 cases. The factor analysis was based on the correlation matrix. The variables and their mnemonics (names) are:

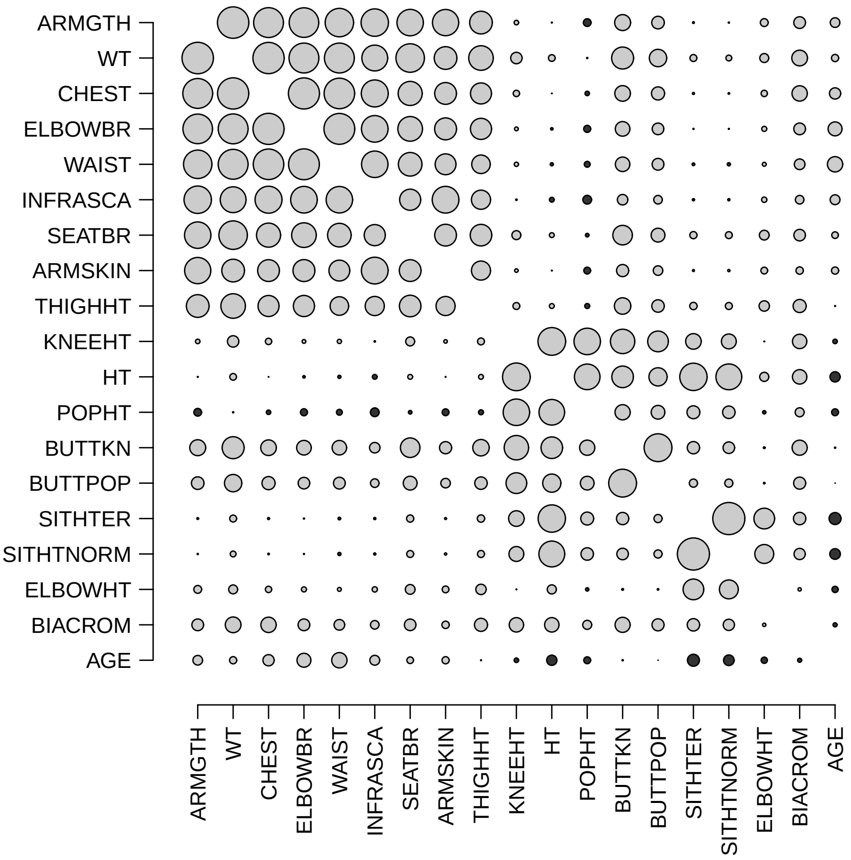


Figure 14.7 Correlations for Example 14.4. The radius of the circle is proportional to the absolute value of the correlation. Light gray circles indicate positive correlations; dark gray circles, negative. (Data from Stoudt et al. [1970].)

- *SEX*: 0 = male, 1 = female.
- *PREVMI*: 0 = history of prior myocardial infarction, 1 = no such history.
- *FEPCHEP*: time in weeks since the first episode of anginal chest pain; this analysis was restricted to patients with anginal chest pain.
- *CHCLASS*: severity of impairment due to angina (chest pain); ranging from I (mildly impaired) to IV (any activity is limited; almost totally bedridden).
- *LMCA*: the percent diameter narrowing of the left main coronary artery; this analysis was restricted to 50% or more narrowing.
- *AGE*: in years.
- *SCORE*: the amount of impairment of the pumping chamber (left ventricle) of the heart; score ranges from 5 (normal) to 30 (not attained).
- *PS70*: the number of proximal (near the beginning of the blood supply) segments of the coronary arteries with 70% or more diameter narrowing.
- *LEFT*: this variable (and RIGHT) tells if the right artery of the heart carries as much blood as normal. LEFT (dominant) implies that the right coronary artery carries little blood; 8.8% of these cases fell into this category. Code: LEFT = 1 (left dominant); LEFT = 0 otherwise.

Table 14.10 Correlations (as the Bottom Entry in Each Cell) and the Residual Correlations (as the Top Entry) in Each Cell[a]

	SEX	PREMI	FEPCHEP	CHCLASS	LMCA	AGE
SEX	0.933					
	1.000					
PREVMI	0.053	0.802				
	0.040	1.000				
FEPCHEP	−0.013	−0.043	0.714			
	−0.002	−0.161	1.000			
CHCLASS	0.056	−0.000	−0.001	0.796		
	0.073	−0.117	0.217	1.000		
LMCA	0.010	0.049	0.005	−0.037	0.989	
	0.012	0.036	0.041	0.004	1.000	
AGE	−0.026	0.019	0.012	−0.001	0.024	0.727
	−0.013	−0.107	0.286	0.227	0.065	1.000
SCORE	0.000	−0.001	−0.000	0.000	0.000	0.000
	0.030	−0.427	0.143	0.185	0.019	0.175
PS70	−0.028	−0.057	−0.027	0.062	−0.016	0.013
	−0.054	−0.188	0.129	0.087	−0.034	0.044
LEFT	0.015	−0.011	−0.015	0.025	0.011	−0.005
	−0.027	−0.022	0.014	0.099	0.063	0.064
RIGHT	0.009	−0.007	−0.009	0.015	0.006	−0.003
	0.054	0.017	−0.033	−0.062	−0.049	−0.077
NOVESLS	0.000	0.000	0.000	−0.000	0.000	0.000
	−0.033	−0.183	0.206	0.014	−0.034	0.130
LVEDP	0.014	0.023	0.001	0.024	0.019	−0.015
	0.020	−0.072	0.119	0.135	0.041	0.109

	SCORE	PS70	LEFT	RIGHT	NOVESLS	LVEDP
SCORE	0.021					
	1.000					
PS70	0.001	0.514				
	0.198	1.000				
LEFT	−0.000	−0.004	0.281			
	0.007	0.004	1.000			
RIGHT	−0.000	−0.004	0.002	0.175		
	−0.041	−0.013	−0.767	1.000		
NOVESLS	0.000	0.000	0.000	0.000	0.000	
	0.284	0.693	−0.071	0.073	1.000	
LVEDP	0.000	−0.025	−0.007	−0.004	0.000	0.930
	0.175	0.029	0.068	−0.086	0.063	1.000

[a] The diagonal entry on top is the estimated uniqueness for each variable. Four factors were used.

- *RIGHT*: there are three types of dominance of the coronary arteries: LEFT above, unbalanced (implicitly coded when LEFT = 0 and RIGHT = 0), and RIGHT. Right dominance is the usual case and occurs when the right coronary artery carries a usual amount of blood. 85.8% of these cases are right dominant: RIGHT = 1; otherwise, RIGHT = 0.
- *NOVESLS*: the number of diseased vessels with $\geq$ 70% stenosis or narrowing of the three major arterial branches above and beyond the left main disease.
- *LVEDP*: the left ventricular end diastolic pressure. This is the pressure in the heart when it is relaxed between beats. A damaged or failing heart has a higher pressure.

Table 14.11 Factor Loadings: Example 14.5

	Factor[a]			
	1	2	3	4
SEX				
PREVMI	−0.103	−0.396	−0.174	
FEPCHEP	0.152		0.535	
CHCLASS			0.125	0.428
LMCA				
AGE				0.502
SCORE		0.108	0.981	0.158
PS70		0.683	0.117	
LEFT	−0.818			0.124
RIGHT	0.917			−0.121
NOVESLS		0.980	0.166	
LVEDP			0.143	0.215
VP[b]	1.525	1.487	1.210	0.872
Proportion var.	0.127	0.124	0.101	0.073
Cumulative var.	0.127	0.251	0.352	0.425

[a] Loadings below 0.100 are omitted.
[b] VP is the portion of sum of squares explained by the factor.

Factor analysis is designed primarily for continuous variables. In this example we have many discrete variables, and even dummy or indicator variables. The analysis is considered more descriptive or explanatory in this case.

Examining the residual values in Table 14.10, we see a fairly satisfactory fit; the maximum absolute value of a residual is 0.062, but most are much smaller. Examination of the uniqueness diagonal column on top shows that the number of vessels diseased, NOVESLS, and SCORE are explained essentially by the factors (uniqueness = 0.000). Some other variables retain almost all of their variability: SEX (uniqueness = 0.993) and LMCA (uniqueness = 0.989). Since we have explained most of the relationships among the variables without using the variability of these factors, SEX and LMCA must be weakly related to the other factors. This is readily verified by looking at the correlation matrix; the maximum absolute correlation involving either of the variables is $r = 0.073$, $r^2 = 0.005$. They explain $\frac{1}{2}$ of 1% or less of the variability in the other variables.

Let us now look at the factor loading (or correlation) values in Table 14.11. The first factor has heavy loadings on the two *dominance* variables. This factor could be labeled a dominance factor. The second factor looks like a *coronary artery disease* (CAD) *factor*. The third is a heart attack, a *ventricular function* factor. The fourth might be labeled a *history* variable.

The first factor exists largely by definition; if LEFT = 1, then RIGHT = 0, and vice versa. The second factor is also expected; if proximal segments are diseased, the arteries are diseased. The third factor makes biological sense. A damaged ventricle often occurs because of a heart attack. The factor with moderate loadings on AGE, FEPCHEP, and CHCLASS is not as clear.

14.11 ESTIMATION

Many methods have been suggested for estimation of the factor loadings and the specificities, that is, the coefficients λ_{ij} and the variance of the residual term e_i. Consider equation (11) and suppose that we change the scale of X_i. Effectively, this is the same as looking at a new variable cX_i; the new value is the old value multiplied by a constant. Multiplying through the equations of equation (11) by the constant, and remembering that we have restricted the factors to have variance 1, we see that factor loading should be multiplied by the same factor as X_i. Only one method of estimation has this property, which also implies that we can use either the covariance matrix or correlation matrix as input to the estimation. This method is the maximum likelihood method; it is our method of choice. The method seems to give the best fit, where fit is examined as described below. There are drawbacks to the method. There can be multiple possible solutions, and software may not converge to the best solution, particularly if the best solution involves a communality of 1.00 for some variable (the "Heywood case"). The examples in this chapter are fairly well behaved, and essentially the same solution was obtained with the programs BMDP and R. For a review of other methods, we recommend the book by Gorsuch [1983]. This book, which is cited extensively below, contains a nice review of many of the issues of factor analysis. Two shorter volumes are those of Kim and Mueller [1983, 1999].

14.12 INDETERMINACY OF THE FACTOR SPACE

There appears to be something magical about factor analysis; we are estimating coefficients of variables that are not even observed. It is difficult to imagine that one can estimate this at all. In point of fact, it is not possible to estimate the F_i uniquely, but one can estimate the F_i up to a certain indeterminacy. It is necessary to describe this indeterminacy in mathematical terms.

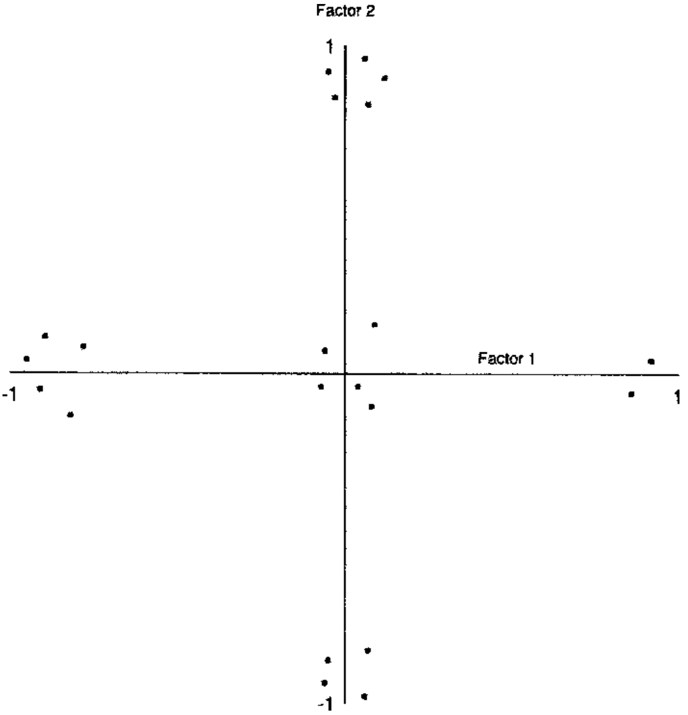
Mathematically, the factors are unique except for possible linear combinations. Geometrically, suppose that we think of the factors (e.g., a model with $k = 2$) as corresponding to values in a plane. Let this plane exist in three-dimensional space. For example, the subspace corresponding to the two factors (i.e., the plane) might be the plane of the paper of this book. Within this three-dimensional space, factor analysis would determine which plane contains the two factors. However, any two perpendicular directions in the factor plane would correspond to factors that equally well fit the data in terms of explaining the covariances or correlations between the variables. Thus, we have the factors identified up to a certain extent, but we are allowed to rotate them within a subspace.

This indeterminacy allows one to "fiddle" with different combinations of factors (i.e., rotations) so that the factors are considered "easy to interpret." As discussed at some length below, one of the strengths and weaknesses of factor analysis is the possibility of finding factors that represent some abstract concept. This task is easiest when the factors are associated with some subset of the variables. That is, one would like factors that have high loadings (in terms of absolute value) on some subset of variables and very low (near zero in absolute value) loadings on the rest of the variables. In this case, the factor is closely associated with the subset of the variables that have large absolute loadings. If these variables have something in common conceptually (e.g., they are all measures of blood pressure) or in a psychological study they all seem to be related to aggressive behavior, one might then identify the specific factor as a blood pressure factor or an aggression factor.

Another complication in the literature of factor analysis is related to the choice of a specific basis in the factor subspace. Suppose for the moment that we are dealing with the correlations among the X_i's. In this case, as we saw before, the loadings on the factors are correlations of the factor with the variable. Thus each loading will be in absolute value less than or equal to 1. It will be easy to interpret our factors if the absolute value is near zero or near 1. Consider Figure 14.8(*a*) and (*b*), plots of the loadings on factors 1 and 2, with a separate point for each of the variables X_i. In Figure 14.8(*a*) there is a very nice pattern. The variables corresponding to points on the factor 1 axis of ±1 or on the factor 2 axis of ±1 are variables associated with each of the factors. The variables plotted near zero on both factors have little relationship to the two factors; in particular, factor 1 would be associated with the variables having points near ±1 along its axis, including variables 1 and 10 as labeled. This would be considered a very nice loading pattern, and easy to interpret, having the simple structure as described above. In Figure 14.8(*b*) we see that if we look at the original factors 1 and 2, it is difficult to interpret the data points, but should we rotate by θ as indicated in the figure, we would have factors easy to interpretation (i.e., each factor associated with a subset of the X_i variables). By looking at such plots and then drawing lines and deciding on the angle θ visually, we have what is called *visual rotation*. When the factor subspace contains a variety of factors (i.e., $k > 2$), the situation is not as simple. If we rotate factors 1 and 2 to find a simple interpretation, we will have altered the relationship between factors 1 and 2 and the other factors, and thus, in improving the relationship between 1 and 2 to have a simple form, we may weaken the relationship between 1 and 5, for example. Visual rotation of factors is an art that can take days or even weeks. The advantage of such rotation is that the mind can weigh the different trade-offs. One drawback of visual rotation is that it may be rotated to give factors that conform to some pet hypothesis. Again, the naming and interpretation of factors are discussed below. Thus, visual rotation can take an enormous amount of time and is subject to the biases of the data analyst (as well as to his or her creativity).

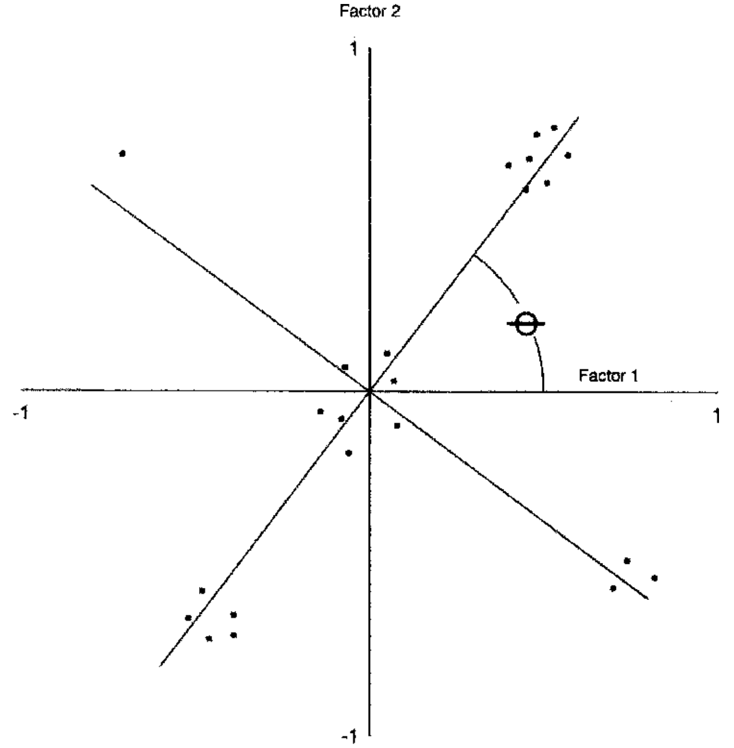
Because of the time constraints for analysis, the complexity of the rotation, and the potential biases, considerable effort has been devoted to developing analytic methods of rotating the factors to get the best rotation. By *analytic* we mean that there is an algorithm describing whether or not a particular rotation for all of the factors is desirable. The computer software, then, finds the best orientation.

Note 14.2 describes two popular criteria, the *varimax method* and the *quartimax method*. A factor analysis is said to have a *general factor* if there is a factor that is associated with all or almost all of the variables. The varimax method can be useful but does not allow general factors and should not be used when such factors may occur. Otherwise, it is considered one of the most



a. Very good loading pattern.
All loadings with absolute value near zero or one.

Figure 14.8 Two-factor loading patterns. (*Continued overleaf*)



b. This pattern suggests rotating the factors by the angle Θ to have a simple structure.

Figure 14.8 *continued*

satisfactory methods. (In fact, factor analysis was developed in conjunction with the study of intelligence. In particular, one of the issues was: Does intelligence consist of one general factor or a variety of uncorrelated factors corresponding to different types of intelligence? Another alternative model for intelligence is a general factor plus other factors associated with some subset of measures of performance thought to be associated with intelligence.)

The second popular method is the quartimax method. This method, in contrast to the varimax method, tends to have one factor with large loadings on all the variables and not many large loadings among the rest of the factors. In the examples of this chapter we have used the varimax method. We do not have the space to get into all the issues involved in the selection of a rotation method.

Returning to visual rotation, suppose that we have the pattern shown in Figure 14.9. We see that there are no perpendicular axes for which the loadings are 1 or -1, but if we took two axes corresponding to the dashed lines, the interpretation might be simplified. Factors corresponding to the two dashed lines are no longer uncorrelated with each other, and one may wonder to what extent they are "separate" factors. Such factors are called *oblique factors,* the word *oblique* coming from the geometric picture and the fact that in geometry, oblique lines are lines that do not intersect at a right angle. There are a number of analytic methods for getting oblique rotations, with snappy names such as *oblimax*, *biquartimin*, *binormamin*, and *maxplane*. References to these may be found in Gorsuch [1983]. If oblique axes or bases are used, the formulas for the variance and covariances of the X_i's as given above no longer hold. Again, see Gorsuch for more in-depth consideration of such issues.

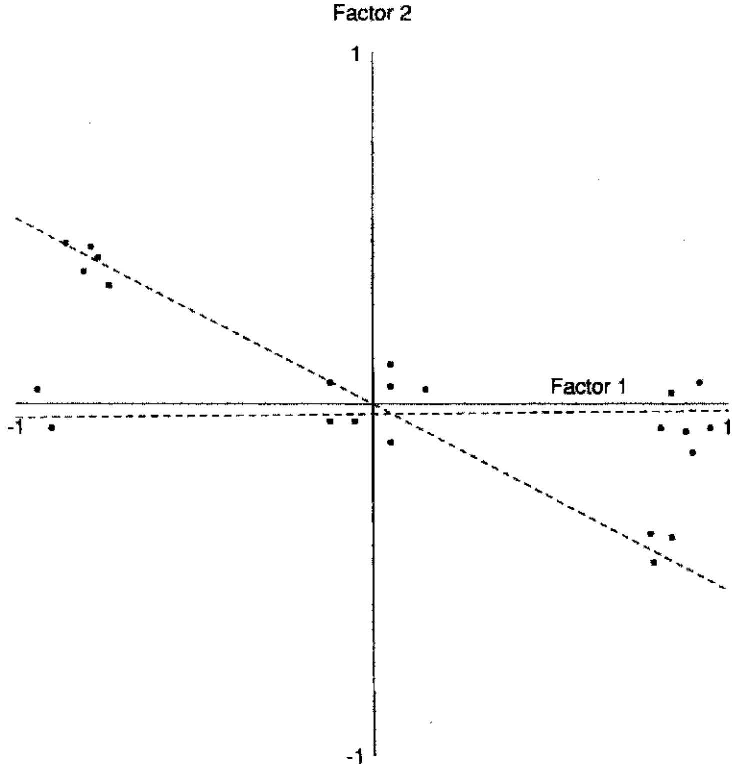


Figure 14.9 Orthogonal and oblique axes for factor loadings.

To try a factor analysis it is not necessary to be expert with every method of estimation and rotation. An exploratory data analysis may be performed to see the extent to which things simplify. We suggest the use of the maximum likelihood estimation method for estimating the coefficients λ_{ij}, where the rotation is performed using the varimax method unless one large general factor is suspected to occur.

Example 14.6. We return to Examples 14.4 and 14.5 and examine plots of the correlations of the variables with the factors. Figure 14.10 shows the plots for Example 14.4, where the numbers on the plot correspond to the variable numbers in Table 14.9.

The plot for factors 2 and 3 looks reasonable (absolute values near 0 or 1). The other two plots have in-between points making interpretation of the factors difficult. This, along with the large residuals mentioned above, suggests trying an analysis with a few more factors.

The plots for Example 14.5 are given in Figure 14.11. These plots suggest factors fairly easy of interpretation, with few, if any, points with moderate loadings on several factors. The interpretation of the factors, discussed in Example 14.5, was fairly straightforward.

14.13 CONSTRAINED FACTOR ANALYSIS

In some situations there are physical constraints on the factors that affect the fitting and interpretation of the factor analysis model. One important application of this sort is in the study of air pollution. Particulate air pollution consists of small particles of smoke, dust, or haze, typically 10 μm in size or smaller. These particles come from a relatively small number of sources, such

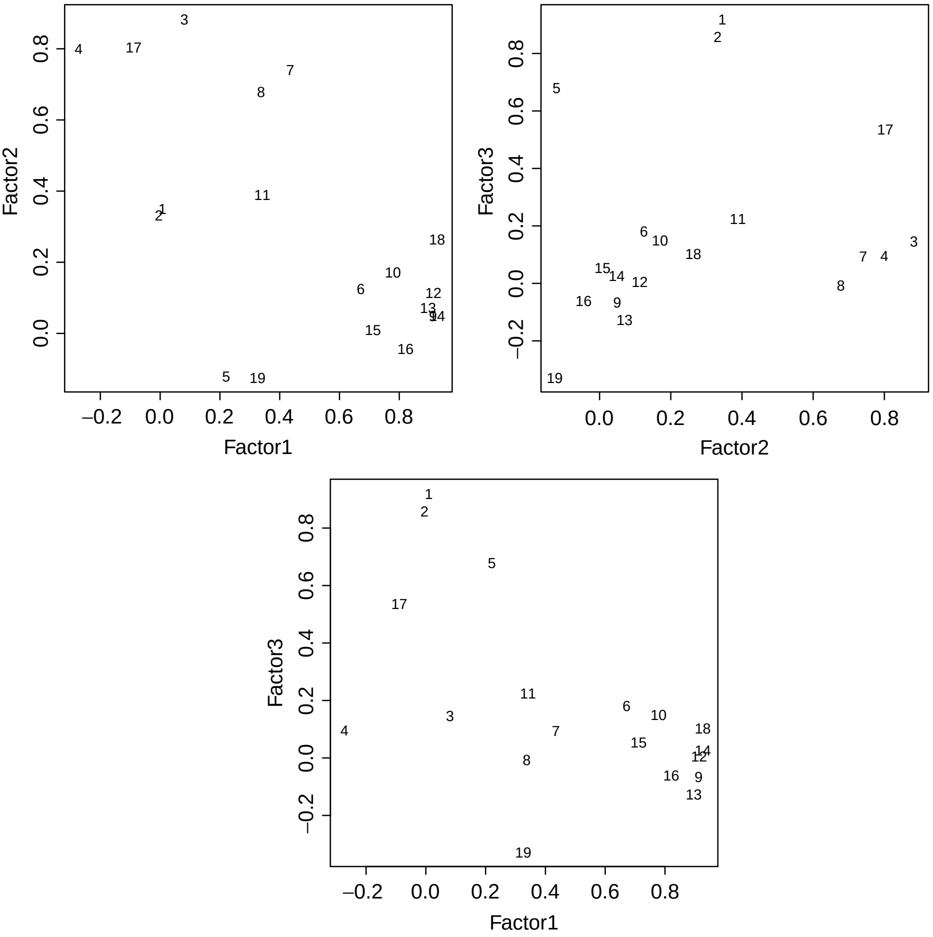


Figure 14.10 Factor loadings for Example 14.4.

as car and truck exhaust, smoke from fireplaces, road dust, and chemical reactions between gases in the air. Particles from different sources have differing distributions of chemical composition, so the chemical composition of particles in the air will be approximately an average of those for each source, weighted according to that source's contribution to overall pollution. That is, we have a factor analysis model in which the factor loadings λ represent the contribution of each source to overall particulate air pollution, the factors F characterize the chemical composition of each source, and the uniquenesses c_i are due largely to measurement error.

In this context the factor analysis model is modified slightly by removing the intercept in each of the regression models of equation (11). Rather than constraining each factor to have zero mean and unit variance, we constrain all the coefficients F and λ to be nonnegative. That is, a source cannot contain a negative amount of some chemical element and cannot contribute a negative concentration of particles. These physical constraints reduce the rotational indeterminacy of the model considerably. On the other hand, it is not reasonable to require that factors are orthogonal to each other, so that oblique rotations must be considered, restoring some of the indeterminacy.

The computation is even more difficult than for ordinary factor analysis, and specialized software is needed [Paatero, 1997, 1999; Henry, 1997]. The full data are needed rather than just a correlation or covariance matrix.

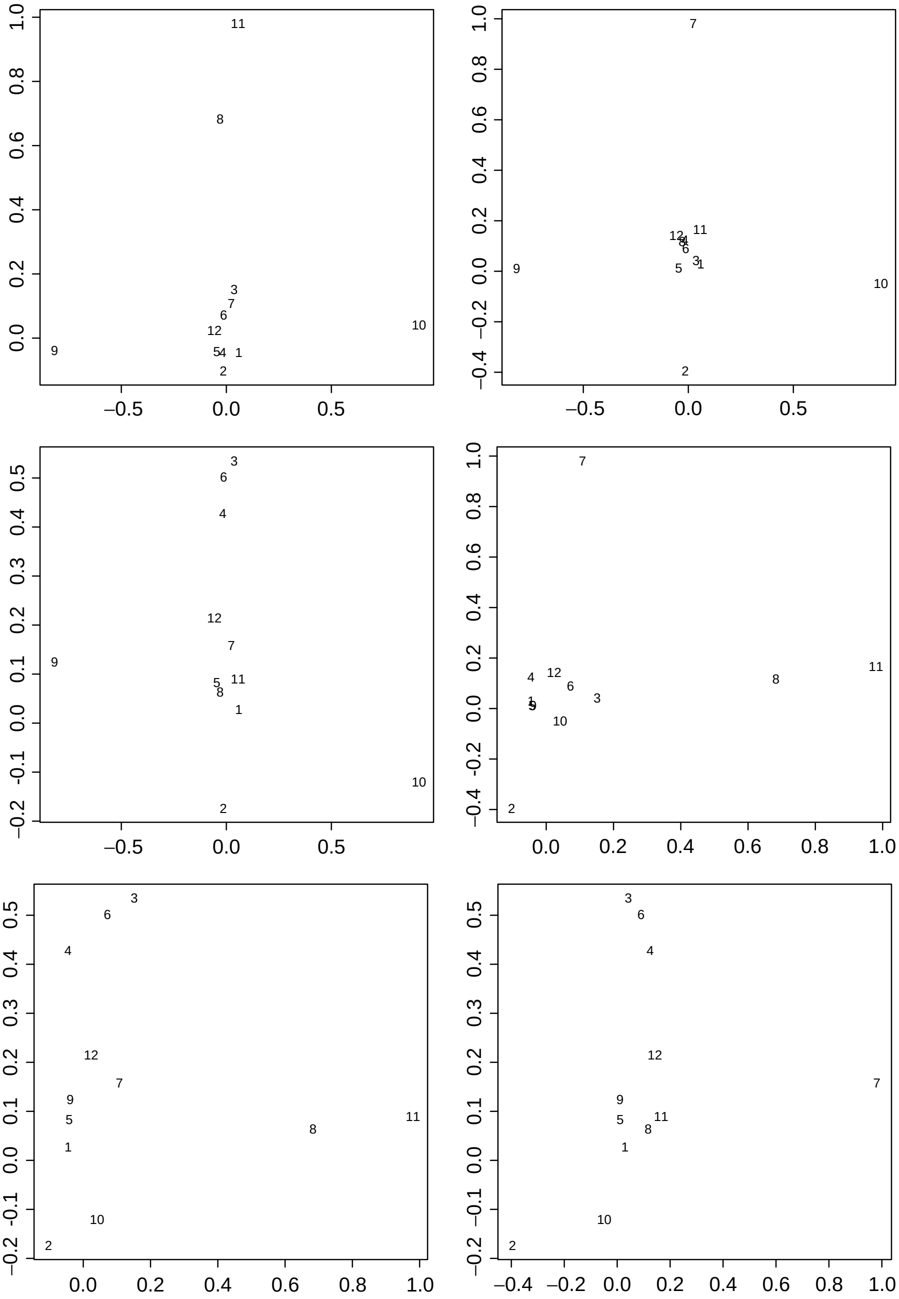


Figure 14.11 Factor loadings for Example 14.5.

Example 14.7. In February 2000, the U.S. Environmental Protection Agency held a workshop on source apportionment for particulate air pollution [U.S. EPA, 2000]. The main part of the workshop was a discussion of two constrained factor analysis methods which were used to investigate fine particulate air pollution from Phoenix, Arizona. Data were available for 981 days, from March 1995 through June 1998, on concentrations of 44 chemical elements and on carbon content, divided into organic carbon and elemental carbon.

The UNMIX method [Henry, 1997] gave a five-factor model:

Source	Concentration (μg/m^3)
Vehicles	4.7
Secondary aerosol	2.6
Soil	1.8
Diesel	1.2
Vegetative burning	0.7
Unidentified	1.6

and the PMF method [Paatero, 1997] gave a six-factor model:

Source	Concentration (μg/m^3)
Motor vehicles	3.5
Coal-fired power	2.1
Soil	1.9
Smelter	0.5
Biomass burning	4.4
Sea salt	0.1

Some of these factors were expected and their likely composition known a priori, such as vehicle exhaust with large amounts of both organic and elemental carbon, and soil with aluminium and silicon. Others were found and interpreted as a result of the analysis; the diesel source had both the elemental carbon characteristic of diesel exhaust and the maganese attributed to fuel additives. The secondary aerosol source in the UNMIX results probably corresponds to the coal-fired power source of PMF and perhaps some of the other burning; it would consist of sulfate and nitrate particles formed by chemical reactions in the atmosphere.

The attributions of fine particles to combustion, soil, and chemical reactions in the atmosphere were reasonably consistent between these methods, but separating different types of combustion proved much more difficult. This is probably a typical case and illustrates that the indeterminacy in the basic factor analysis model can partly, but not entirely, be overcome by substantive knowledge.

14.14 DETERMINING THE NUMBER OF FACTORS

In this section we consider what to do when the number of factors is unknown. Estimation methods of factor analysis begin with knowledge of k, the number of factors. But this number is usually not known or hypothesized. There is no universal agreement on how to select k; below we examine a number of ways of doing this. The first step is always carried out.

1. Examine the values of the residual correlations. In this section we suppose that we are trying to model the correlations between variables rather than their covariances. Recall that with maximum likelihood estimation, fitting one is the same as fitting the other. In looking at the residual correlations, as done in Examples 14.4 and 14.5, we may feel that we have done a good job if all of the correlations have been fit to within a specified difference. If the residual correlations reveal large discrepancies, the model does not fit.

2. There are statistical tests *if* we can assume that multivariate normality holds and we use the maximum likelihood estimation method. In this case, there is an asymptotic chi-square test for any hypothesized fixed number of factors. Computation of the test statistic is complex and given in Note 14.3. However, it is available in many statistical computer programs. One approach is to look at successively more factors until the statistic is not statistically significant; that is, there are enough factors so that one would not reject at a fixed significance level the hypothesis that the number of factors is as given. This is analogous to a stepwise regression procedure. If we do this, we are performing a stepwise procedure, and the true and nominal significance levels differ (as usual in a stepwise analysis).

3. Looking at the roots of the correlation matrix:

a. If the correlations are arranged in a square pattern or matrix, as usually done, this pattern is called a *correlation matrix.* Suppose that we perform a principal component analysis and examine the variances of the principal components $V_1 \geq V_2 \geq \cdots \geq V_p$. These values are called the *eigenvalues* or *roots* of the correlation matrix. If we have the correlation matrix for the entire population, Guttman [1954] showed that the number of factors, k, must be greater than or equal to the number of roots greater than or equal to 1. That is, the number of factors in the factor analytic model must be greater than or equal to the number of principal components whose variance is greater than or equal to 1. Of course, in practice we do not have the population correlation matrix but an estimate. The number of such roots greater than or equal to 1 in a sample may turn out to be smaller or larger. However, because of Guttman's result, a reasonable starting value for k is the number of roots greater than or equal to 1 for the sample correlation matrix. For a thorough factor analysis, values of k above and below this number should be tried and the residual patterns observed. The number of factors in Examples 14.4 and 14.5 was chosen by this method.

b. *Scree* is the name for the rubble at the bottom of a cliff. The scree test plots the variances of the principal components. If the plot looks somewhat like Figure 14.12, one looks to separate the climb of the cliff from the scree at the bottom of the cliff. We are directed to pick the cliff, components 1, 2, 3, and possibly 4, rather than the rubble. A clear plastic ruler is laid across the bottom points, and the number of values above the line is the number of important factors. This advice is reasonable when a sharp demarcation can be seen, but often the pattern has no clear breakpoint.

c. Since we are interested in the correlation structure, we might plot as a function of k (the number of factors) the maximum absolute value of all the residuals of the estimated

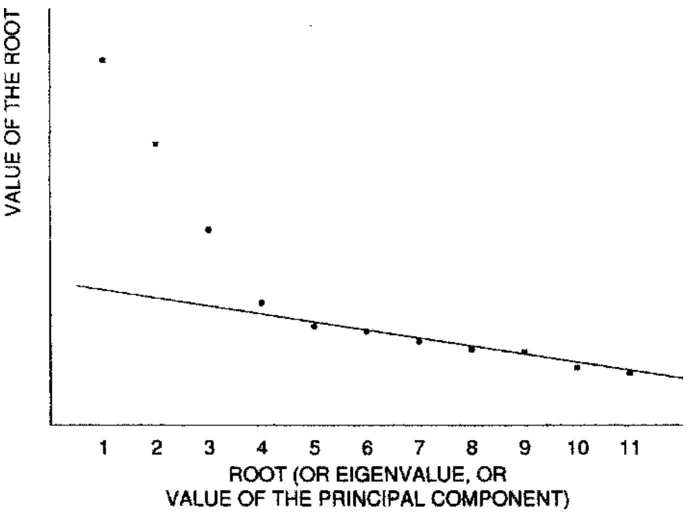


Figure 14.12 Plot for the scree test.

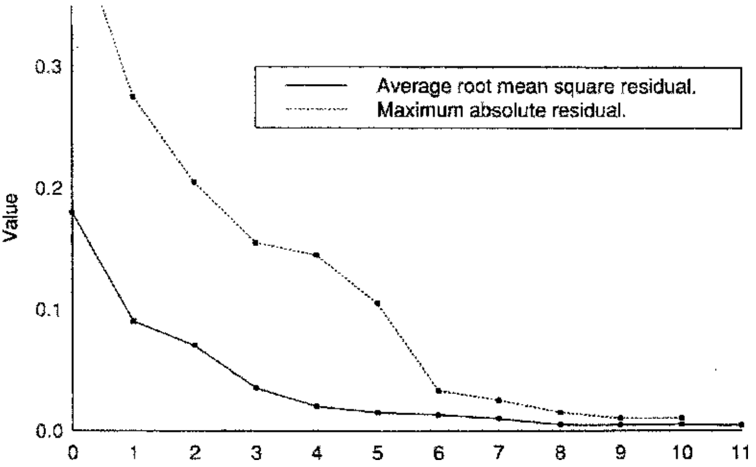


Figure 14.13 Plot of the maximum absolute residual and the average root mean square residual.

correlations. Another useful plot is the square root of the sum of the squares of all of the residual correlations divided by the number of such residual correlations, which is $p(p-1)/2$. If there is a break in the plots of the curves, we would then pick k so that the maximum and average squared residual correlations are small. For example, in Figure 14.13 we might choose three or four factors. Gorsuch suggests: "In the final report, interpretation could be limited to those factors which are well stabilized over the range which the number of factors may reasonably take."

14.15 INTERPRETATION OF FACTORS

Much of the debate about factor analysis stems from the naming and interpretation of factors. Often, after a factor analysis is performed, the factors are identified with concepts or objects. *Is a factor an underlying concept or merely a convenient way of summarizing interrelationships among variables*? A useful word in this context is *reify*, meaning to convert into or to regard something as a concrete thing. Should factors be reified?

As Gorsuch states: "A prime use of factor analysis has been in the development of both the theoretical constructs for an area and the operational representatives for the theoretical constructs." In other words, a prime use of factor analysis requires reifying the factors. Also, "The first task of any research program is to establish empirical referents for the abstract concepts embodied in a particular theory."

In psychology, how would one deal with an abstract concept such as aggression? On a questionnaire a variety of possible "aggression" questions might be used. If most or all of them have high loadings on the same factor, and other questions thought to be unrelated to aggression had low loadings, one might identify that factor with aggression. Further, the highest loadings might identify operationally the questions to be used to examine this abstract concept.

Since our knowledge is of the original observations, without a unique set of variables loading a factor, interpretation is difficult. Note well, however, that there is no law saying that one must interpret and name any or all factors.

Gorsuch makes the following points:

1. "The factor can only be interpreted by an individual with extensive background in the substantive area."

2. "The summary of the interpretation is presented as the factor's name. The name may be only descriptive or it may suggest a causal explanation for the occurrence of the factor. Since the name of the factor is all most readers of the research report will remember, it should be carefully chosen." *Perhaps it should not be chosen at all in many cases.*
3. "The widely followed practice of regarding interpretation of a factor as confirmed solely because the post-hoc analysis 'makes sense' is to be deplored. Factor interpretations can only be considered hypotheses for another study."

Interpretation of factors may be strengthened by using cases from other populations. Also, collecting other variables thought to be associated with the factor and including them in the analysis is useful. They should load on the same factor. Taking "marker" variables from other studies is useful in seeing whether an abstract concept has been embodied in more or less the same way in two different analyses.

For a perceptive and easy-to-understand discussion of factor analysis, see Chapter 6 in Gould [1996], which deals with scientific racism. Gould discusses the reification of intelligence in the Intelligence Quotient (IQ) through the use of factor analysis. Gould traces the history of factor analysis starting with the work of Spearman. Gould's book is a cautionary tale about scientific presuppositions, predilections, and perceptions affecting the interpretation of statistical results (it is not necessary to agree with all his conclusions to benefit from his explanations). A recent book by McDonald [1999] has a more technical discussion of reification and factor analysis. For a semihumorous discussion of reification, see Armstrong [1967].

NOTES

14.1 Graphing Two-Dimensional Projections

As noted in Section 14.8, the first two principal components can be used as plot axes to give a two-dimensional representation of higher-dimensional data. This plot will be best in the sense that it shows the maximum possible variability. Other multivariate graphical techniques give plots that are "the best" in other senses.

Multidimensional scaling gives a two-dimensional plot that reproduces the distances between points as accurately as possible. This view will be similar to the first two principal components when the data form a football (ellipsoid) shape, but may be very different when the data have a more complicated structure. Other *projection pursuit techniques* specifically search for views of the data that reveal holes, clusters, lines, and other departures from an ellipsoidal shape. A relatively nontechnical review of this concept is given by Jones and Sibson [1987].

Rather than relying on a single two-dimensional projection, it is also possible to display animated sequences of projections on a computer screen. The projections can be generated by random rotations of the data or by projection pursuit methods that attempt to show "interesting" projections. The free computer program GGobi (*http://www.ggobi.org*) implements many of these techniques.

Of course, more sophisticated searches performed by computer mean that more caution in interpretation is needed from the analyst. Substantial experience with these techniques is needed to develop a feeling for which graphs indicate real structure as opposed to overinterpreted noise.

14.2 Varimax and Quartimax Methods of Choosing Factors in a Factor Analysis

Many analytic methods of choosing factors have been developed so that the loading matrix is easy to interpret, that is, has a simple structure. These many different methods make the factor analysis literature very complex. We mention two of the methods.

1. *Varimax method.* The varimax method uses the idea of maximizing the sum of the variances of the squares of loadings of the factors. Note that the variances are high when the λ_{ij}^2 are near 1 and 0, some of each in each column. In order that variables with large communalities are not overly emphasized, weighted values are used. Suppose that we have the loadings λ_{ij} for one selection of factors. Let θ_{ij} be the loadings for a different set of factors (the linear combinations of the old factors). Define the weighted quantities

$$\gamma_{ij} = \theta_{ij} \Bigg/ \sqrt{\sum_{j=1}^{m} \lambda_{ij}^2}$$

The method chooses the θ_{ij} to maximize the following:

$$\sum_{j=1}^{k} \left[\frac{1}{p} \sum_{i=1}^{p} \gamma_{ij}^4 - \frac{1}{p^2} \left(\sum_{i=1}^{p} \gamma_{ij}^2 \right)^2 \right]$$

Some problems have a factor where all variables load high (e.g., general IQ). Varimax should not be used if a general factor may occur, as the low variance discourages general factors. Otherwise, it is one of the most satisfactory methods.

2. *Quartimax method.* The quartimax method works with the variance of the square of all p_k loadings. We maximize over all possible loadings θ_{ij}:

$$\max_{\theta_{ij}} \left[\sum_{i=1}^{p} \sum_{j=1}^{k} \theta_{ij}^4 - \frac{1}{pm} \left(\sum_{i=1}^{p} \sum_{j=1}^{k} \theta_{ij}^2 \right) \right]$$

Quartimax is used less often, since it tends to include one factor with all major loadings and no other major loadings in the rest of the matrix.

14.3 Statistical Test for the Number of Factors in a Factor Analysis When $X_1, \ldots, X_p$ Are Multivariate Normal and Maximum Likelihood Estimation Is Used

This note presupposes familiarity with matrix algebra. Let A be a matrix and A' denote the transpose of A; if A is square, let $|A|$ be the determinant of A and $\mathrm{Tr}(A)$ be the trace of A. Consider a factor analysis with k factors and estimated *loading matrix*

$$\Lambda = \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1k} \\ \vdots & \ddots & \vdots \\ \lambda_{n1} & \cdots & \lambda_{nk} \end{pmatrix}$$

The test statistic is

$$X^2 = \left(n - 1 - \frac{2p+5}{6} - \frac{2k}{3} \right) \log_e \left(\frac{|\Lambda\Lambda' + \psi|}{|S|} \right) \mathrm{Tr}(S(\Lambda\Lambda' + \psi)^{-1})p$$

where S is the sample covariance matrix, ψ a diagonal matrix where $\psi_{ii} = s_i - (\Lambda\Lambda')_{ii}$, and s_i the sample variance of X_i. If the true number of factors is less than or equal to k, X^2 has a chi-square distribution with $[(p-k)^2 - (p+k)]/2$ degrees of freedom. The null hypothesis of only k factors is rejected if X^2 is too large.

One could try successively more factors until this is not significant. The true and nominal significance levels differ as usual in a stepwise procedure. (For the test to be appropriate, the degrees of freedom must be > 0.)

PROBLEMS

The first four problems present principal component analyses using correlation matrices. Portions of computer output (BMDP program 4M) are given. The coefficients for principal components that have a variance of 1 or more are presented. Because of the connection of principal component analysis and factor analysis mentioned in the text (when the correlations are used), the principal components are also called *factors* in the output. With a correlation matrix the coefficient values presented are for the standardized variables. You are asked to perform a subset of the following tasks.

(a) Fill in the missing values in the "variance explained" and "cumulative proportion of total variance" table.

(b) For the principal component(s) specified, give the percent of the total variance accounted for by the principal component(s).

(c) How many principal components are needed to explain 70% of the total variance? 90%? Would a plot with two axes contain most (say, $\geq$ 70%) of the variability in the data?

(d) For the case(s) with the value(s) as given, compute the case(s) values on the first two principal components.

14.1 This problem uses the psychosocial Framingham data in Table 11.20. The mnemonics go in the same order as the correlations presented. The results are presented in Tables 14.12 and 14.19. Perform tasks (a) and (b) for principal components 2 and 4, and task (c).

14.2 Measurement data on U.S. females by Stoudt et al. [1970] were discussed in this chapter. The same correlation data for adult males were also given (Table 14.14). The principal

Table 14.12 Problem 14.1: Variance Explained by Principal Components[a]

Factor	Variance Explained	Cumulative Proportion of Total Variance
1	4.279180	0.251716
2	1.633777	0.347821
3	1.360951	?
4	1.227657	0.500092
5	1.166469	0.568708
6	?	0.625013
7	0.877450	0.676627
8	0.869622	0.727782
9	0.724192	0.770381
10	0.700926	0.811612
11	0.608359	?
12	0.568691	0.880850
13	0.490974	0.909731
14	?	0.935451
15	0.386540	0.958189
16	0.363578	0.979576
17	?	?

[a] The variance explained by each factor is the eigenvalue for that factor. Total variance is defined as the sum of the diagonal elements of the correlation (covariance) matrix.

Table 14.13 Problem 14.1: Principal Components

		Unrotated Factor Loadings (Pattern) for Principal Components				
		Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
TYPEA	1	0.633	−0.203	0.436	−0.049	0.003
EMOTLBLE	2	0.758	−0.198	−0.146	0.153	−0.005
AMBITIOS	3	0.132	−0.469	0.468	−0.155	−0.460
NONEASY	4	0.353	0.407	−0.268	0.308	0.342
NOBOSSPT	5	0.173	0.047	0.260	−0.206	0.471
WKOVRLD	6	0.162	−0.111	0.385	−0.246	0.575
MTDISSAG	7	0.499	0.542	0.174	−0.305	−0.133
MGDISSAT	8	0.297	0.534	−0.172	−0.276	−0.265
AGEWORRY	9	0.596	0.202	0.060	−0.085	−0.145
PERSONWY	10	0.618	0.346	0.192	−0.174	−0.206
ANGERIN	11	0.061	−0.430	−0.470	−0.443	−0.186
ANGEROUT	12	0.306	0.178	0.199	0.607	−0.215
ANGRDISC	13	0.147	−0.181	0.231	0.443	−0.108
STRESS	14	0.665	−0.189	0.062	−0.053	0.149
TENSION	15	0.771	−0.226	−0.186	0.039	0.118
ANXSYMPT	16	0.594	−0.141	−0.352	0.022	0.067
ANGSYMPT	17	0.723	−0.242	−0.256	0.086	−0.015
	VP[a]	4.279	1.634	1.361	1.228	1.166

[a] The VP for each factor is the sum of the squares of the elements of the column of the factor loading matrix corresponding to that factor. The VP is the variance explained by the factor.

component analysis gave the results of Table 14.15. Perform tasks (a) and (b) for principal components 2, 3, and 4, and task (c).

14.3 The Bruce et al. [1973] exercise data for 94 sedentary males are used in this problem (see Table 9.16). These data were used in Problems 9.9 to 9.12. The exercise variables used are DURAT (duration of the exercise test in seconds), $\text{VO}_{2\ \text{MAX}}$ [the maximum oxygen consumption (normalized for body weight)], HR [maximum heart rate (beats/min)], AGE (in years), HT (height in centimeters), and WT (weight in kilograms). The correlation values are given in Table 14.17. The principal component analysis is given in Table 14.18. Perform tasks (a) and (b) for principal components 4, 5, and 6, and task (c) (Table 14.19). Perform task (d) for a case with $\text{DURAT} = 600, \text{VO}_{2\ \text{MAX}} = 38, \text{HR} = 185, \text{AGE} = 29, \text{HT} = 165$, and $\text{WT} = 71$. (*N.B.*: Find the value of the *standardized* variables.)

14.4 The variables are the same as in Problem 14.3. In this analysis 43 active females (whose individual data are given in Table 9.14) are studied. The correlations are given in Table 14.21. the principal component analysis in Tables 14.22 and 14.23. Perform tasks (a) and (b) for principal components 1 and 2, and task (c). Do task (d) for the two cases in Table 14.24 (use standard variables). See Table 14.21.

Problems 14.5, 14.7, 14.8, 14.10, 14.11, and 14.12 consider maximum likelihood factor analysis with varimax rotation (from computer program BMDP4M). Except for Problem 14.10, the number of factors is selected by Guttman's root criterion (the number of eigenvalues greater than 1). Perform the following tasks as requested.

Table 14.14 Problem 14.2: Correlations

		STHTER	STHTHL	KNEEHT	POPHT	ELBWHT
		1	2	3	4	5
STHTER	1	1.000				
STHTHL	2	0.873	1.000			
KNEEHT	3	0.446	0.443	1.000		
POPHT	4	0.410	0.382	0.798	1.000	
ELBWHT	5	0.544	0.454	−0.029	−0.062	1.000
THIGHHT	6	0.238	0.284	0.228	−0.029	0.217
BUTTKNHT	7	0.418	0.429	0.743	0.619	0.005
BUTTPOP	8	0.227	0.274	0.626	0.524	−0.145
ELBWELBW	9	0.139	0.212	0.139	−0.114	0.231
SEATBRTH	10	0.365	0.422	0.311	0.050	0.286
BIACROM	11	0.365	0.335	0.352	0.275	0.127
CHESTGRH	12	0.238	0.298	0.229	0.000	0.258
WSTGRTH	13	0.106	0.184	0.138	−0.097	0.191
RTARMGRH	14	0.221	0.265	0.194	−0.059	0.269
RTARMSKN	15	0.133	0.191	0.081	−0.097	0.216
INFRASCP	16	0.096	0.152	0.038	−0.166	0.247
HT	17	0.770	0.717	0.802	0.767	0.212
WT	18	0.403	0.433	0.404	0.153	0.324
AGE	19	−0.272	−0.183	−0.215	−0.215	−0.192

		THIGH-HT	BUTT-KNHT	BUTT-POP	ELBW-ELBW	SEAT-BRTH
		6	7	8	9	10
THIGHHT	6	1.000				
BUTTKNHT	7	0.348	1.000			
BUTTPOP	8	0.237	0.736	1.000		
ELBWELBW	9	0.603	0.299	0.193	1.000	
SEATBRTH	10	0.579	0.449	0.265	0.707	1.000
BIACROM	11	0.303	0.365	0.252	0.311	0.343
CHESTGRH	12	0.605	0.386	0.252	0.833	0.732
WSTGRTH	13	0.537	0.323	0.216	0.820	0.717
RTARMGRH	14	0.663	0.342	0.224	0.755	0.675
RTARMSKN	15	0.480	0.240	0.128	0.524	0.546
INFRASCP	16	0.503	0.212	0.106	0.674	0.610
HT	17	0.210	0.751	0.600	0.069	0.309
WT	18	0.684	0.551	0.379	0.804	0.813
AGE	19	−0.190	−0.151	−0.108	0.156	0.043

		BIACROM	CHESTGRH	WSTGRTH	RTARMGRH	RTARMSKN
		11	12	13	14	15
BIACROM	11	1.000				
CHESTGRH	12	0.418	1.000			
WSTGRTH	13	0.249	0.837	1.000		
RTARMGRH	14	0.379	0.784	0.712	1.000	
RTARMSKN	15	0.183	0.558	0.552	0.570	1.000
INFRASCP	16	0.242	0.710	0.727	0.667	0.697
HT	17	0.381	0.189	0.054	0.139	0.060
WT	18	0.474	0.885	0.821	0.849	0.562
AGE	19	−0.261	0.062	0.299	−0.115	−0.039

		INFRASCP	HT	WT	AGE
		16	17	18	19
INFRASCP	16	1.000			
HT	17	−0.003	1.000		
WT	18	0.709	0.394	1.000	
AGE	19	0.045	−0.270	−0.058	1.000

Table 14.15 Problem 14.2: Variance Explained by the Principal Components[a]

Factor	Variance Explained	Cumulative Proportion of Total Variance
1	7.839282	0.412594
2	4.020110	0.624179
3	1.820741	0.720007
4	1.115168	0.778700
5	0.764398	0.818932
6	?	0.850389
7	0.475083	?
8	0.424948	0.897759
9	0.336247	0.915456
10	?	0.931210
11	0.252205	0.944484
12	?	0.955404
13	0.202398	0.966057
14	0.169678	0.974987
15	0.140613	0.982388
16	0.119548	?
17	0.117741	0.994872
18	0.055062	0.997770
19	0.042365	1.000000

[a] The variance explained by each factor is the eigenvalue for that factor. Total variance is defined as the sum of the diagonal elements of the correlation (covariance) matrix.

Table 14.16 Exercise Data for Problem 14.3

		Univariate Summary Statistics	
	Variable	Mean	Standard Deviation
1	DURAT	577.10638	123.83744
2	VO_2 MAX	35.63298	7.51007
3	HR	175.39362	18.59195
4	AGE	49.78723	11.06955
5	HT	177.39851	6.58285
6	WT	79.00000	8.71286

Table 14.17 Problem 14.3: Correlation Matrix

		DURAT	VO_2 MAX	HR	AGE	HT	WT
DURAT	1	1.000					
VO_2 MAX	2	0.905	1.000				
HR	3	0.678	0.647	1.000			
AGE	4	−0.687	−0.656	−0.630	1.000		
HT	5	0.035	0.050	0.107	−0.161	1.000	
WT	6	−0.134	−0.147	0.015	−0.069	0.536	1.000

Table 14.18 Problem 14.3: Variance Explained by the Principal Components[a]

Factor	Variance Explained	Cumulative Proportion of Total Variance
1	3.124946	0.520824
2	1.570654	?
3	0.483383	0.863164
4	?	0.926062
5	?	0.984563
6	0.092621	1.000000

[a] The variance explained by each factor is the eigenvalue for that factor. Total variance is defined as the sum of the diagonal elements of the correlation (covariance) matrix.

Table 14.19 Problem 14.3: Principal Components

		Unrotated Factor Loadings (Pattern) for Principal Components	
		Factor 1	Factor 2
DURAT	1	0.933	−0.117
VO2 MAX	2	0.917	−0.120
HR	3	0.832	0.057
AGE	4	−0.839	−0.134
HT	5	0.128	0.860
WT	6	−0.057	0.884
	VP[a]	3.125	1.571

[a] The VP for each factor is the sum of the squares of the elements of the column of the factor loading matrix corresponding to that factor. The VP is the variance explained by the factor.

Table 14.20 Exercise Data for Problem 14.4

		Univariate Summary Statistics	
	Variable	Mean	Standard Deviation
1	DURAT	514.88372	77.34592
2	VO2 MAX	29.05349	4.94895
3	HR	180.55814	11.41699
4	AGE	45.13953	10.23435
5	HT	164.69767	6.30017
6	WT	61.32558	7.87921

Table 14.21 Problem 14.4: Correlation Matrix

		DURAT	VO2 MAX	HR	AGE	HT	WT
DURAT	1	1.000					
VO2 MAX	2	0.786	1.000				
HR	3	0.528	0.337	1.000			
AGE	4	−0.689	−0.651	−0.411	1.000		
HT	5	0.369	0.299	0.310	−0.455	1.000	
WT	6	0.094	−0.126	0.232	−0.042	0.483	1.000

Table 14.22 Problem 14.4: Variance Explained by the Principal Components[a]

Factor	Variance Explained	Cumulative Proportion of Total Variance
1	3.027518	?
2	1.371342	0.733143
3	?	?
4	0.416878	0.918943
5	?	0.972750
6	?	1.000000

[a] The variance explained by each factor is the eigenvalue for that factor. Total variance is defined as the sum of the diagonal elements of the correlation (covariance) matrix.

Table 14.23 Problem 14.4: Principal Components

		Unrotated Factor Loadings (Pattern) for Principal Components	
		Factor 1	Factor 2
DURAT	1	0.893	−0.201
VO_2 MAX	2	0.803	−0.425
HR	3	0.658	0.162
AGE	4	−0.840	0.164
HT	5	0.626	0.550
WT	6	0.233	0.891
	VP[a]	3.028	1.371

[a] The VP for each factor is the sum of the squares of the elements of the column of the factor loading matrix corresponding to that factor. The VP is the variance explained by the factor.

Table 14.24 Data for Two Cases, Problem 14.3

	Subject 1	Subject 2
DURAT	660	628
VO_2 MAX	38.1	38.4
HR	184	183
AGE	23	21
HT	177	163
WT	83	52

a. Examine the residual correlation matrix. What is the maximum residual correlation? Is it < 0.1? < 0.5?

b. For the pair(s) of variables, with mnemonics given, find the fitted residual correlation.

c. Consider the plots of the rotated factors. Discuss the extent to which the interpretation will be simple.

d. Discuss the potential for naming and interpreting these factors. Would you be willing to name any? If so, what names?

e. Give the uniqueness and communality for the variables whose numbers are given.

f. Is there any reason that you would like to see an analysis with fewer or more factors? If so, why?

g. If you were willing to associate a factor with variables (or a variable), identify the variables on the shaded form of the correlations. Do the variables cluster (form a dark group), which has little correlation with the other variables?

14.5 A factor analysis is performed upon the Framingham data of Problem 14.1. The results are given in Tables 14.25 to 14.27 and Figures 14.14 and 14.15. Communalities were obtained from five factors after 17 iterations. The communality of a variable is its squared multiple correlation with the factors; they are given in Table 14.26. Perform tasks (a), (b)

Table 14.25 Problem 14.5: Residual Correlations

		TYPEA 1	EMOTLBLE 2	AMBITIOS 3	NONEASY 4	NOBOSSPT 5	WKOVRLD 6
TYPEA	1	0.219					
EMOTLBLE	2	0.001	0.410				
AMBITIOS	3	0.001	0.041	0.683			
NONEASY	4	0.003	0.028	−0.012	0.635		
NOBOSSPT	5	−0.010	−0.008	0.001	−0.013	0.964	
WKOVRLD	6	0.005	−0.041	−0.053	−0.008	0.064	0.917
MTDISSAG	7	0.007	−0.010	−0.062	−0.053	0.033	0.057
MGDISSAT	8	0.000	0.000	0.000	0.000	0.000	0.000
AGEWORRY	9	0.002	0.030	0.015	0.017	0.001	−0.017
PERSONWY	10	−0.002	−0.010	0.007	0.007	−0.007	−0.003
ANGERIN	11	0.007	−0.006	−0.028	0.005	−0.018	0.028
ANGEROUT	12	0.001	0.056	0.053	0.014	−0.070	−0.135
ANGRDISC	13	−0.011	0.008	0.044	−0.019	−0.039	0.006
STRESS	14	0.002	−0.032	−0.003	0.018	0.030	0.034
TENSION	15	−0.004	−0.006	−0.016	−0.017	0.013	0.024
ANXSYMPT	16	0.004	−0.026	−0.028	−0.019	0.009	−0.015
ANGSYMPT	17	−0.000	0.018	−0.008	−0.012	−0.006	0.009

		MTDISSAG 7	MTDISSAT 8	AGEWORRY 9	PERSONWY 10	ANGERIN 11	ANGEROUT 12
MTDISSAG	7	0.574					
MGDISSAT	8	0.000	0.000				
AGEWORRY	9	0.001	−0.000	0.572			
PERSONWY	10	−0.002	0.000	0.001	0.293		
ANGERIN	11	0.010	−0.000	0.015	−0.003	0.794	
ANGEROUT	12	0.006	−0.000	−0.006	−0.001	−0.113	0.891
ANGRDISC	13	−0.029	−0.000	0.000	0.001	−0.086	0.080
STRESS	14	−0.017	−0.000	−0.015	0.013	0.022	−0.050
TENSION	15	0.004	−0.000	−0.020	0.007	−0.014	−0.045
ANXSYMPT	16	0.026	−0.000	0.037	−0.019	0.011	−0.026
ANGSYMPT	17	0.004	−0.000	−0.023	0.006	0.012	0.049

		ANGRDISC 13	STRESS 14	TENSION 15	ANXSYMPT 16	ANGSYMPT 17
ANGRDISC	13	0.975				
STRESS	14	−0.011	0.599			
TENSION	15	−0.005	0.035	0.355		
ANXSYMPT	16	−0.007	0.015	0.020	0.645	
ANGSYMPT	17	0.027	−0.021	−0.004	−0.008	0.398

Table 14.26 Problem 14.5: Communalities

1	TYPEA	0.7811
2	EMOTLBLE	0.5896
3	AMBITIOS	0.3168
4	NONEASY	0.3654
5	NOBOSSPT	0.0358
6	WKOVRLD	0.0828
7	MTDISSAG	0.4263
8	MGDISSAT	1.0000
9	AGEWORRY	0.4277
10	PERSONWY	0.7072
11	ANGERIN	0.2063
12	ANGEROUT	0.1087
13	ANGRDISC	0.0254
14	STRESS	0.4010
15	TENSION	0.6445
16	ANXSYMPT	0.3555
17	ANGSYMPT	0.6019

Table 14.27 Problem 14.5: Factors (Loadings Smaller Than 0.1 Omitted)

		Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
TYPEA	1	0.331	0.185	0.133	0.753	0.229
EMOTLBLE	2	0.707	0.194		0.215	
AMBITIOS	3				0.212	0.515
NONEASY	4	0.215	0.105	0.163	0.123	−0.516
NOBOSSPT	5		0.101		0.142	
WKOVRLD	6				0.281	
MTDISSAG	7		0.474	0.391	0.178	
MGDISSAT	8		0.146	0.971	−0.143	
AGEWORRY	9	0.288	0.576			
PERSONWY	10	0.184	0.799	0.138	0.127	
ANGERIN	11	0.263			−0.238	0.272
ANGEROUT	12	0.128	0.179		0.196	−0.148
ANGRDISC	13	0.117			0.102	
STRESS	14	0.493	0.189		0.337	
TENSION	15	0.753	0.193		0.190	
ANXSYMPT	16	0.571	0.138			
ANGSYMPT	17	0.748	0.191			
	VP[a]	2.594	1.477	1.181	1.112	0.712

[a] The VP for each factor is the sum of the squares of the elements of the column of the factor pattern matrix corresponding to that factor. When the rotation is orthogonal, the VP is the variance explained by the factor.

(TYPEA, EMOTLBLE) and (ANGEROUT, ANGERIN), (c), (d), and (e) for variables 1, 5, and 8, and tasks (f) and (g). In this study, the TYPEA variable was of special interest. Is it associated particularly with one of the factors?

14.6 This question requires you to do the fitting of the factor analysis model. Use the Florida voting data of Problem 9.34 available on the Web appendix to examine the structure of

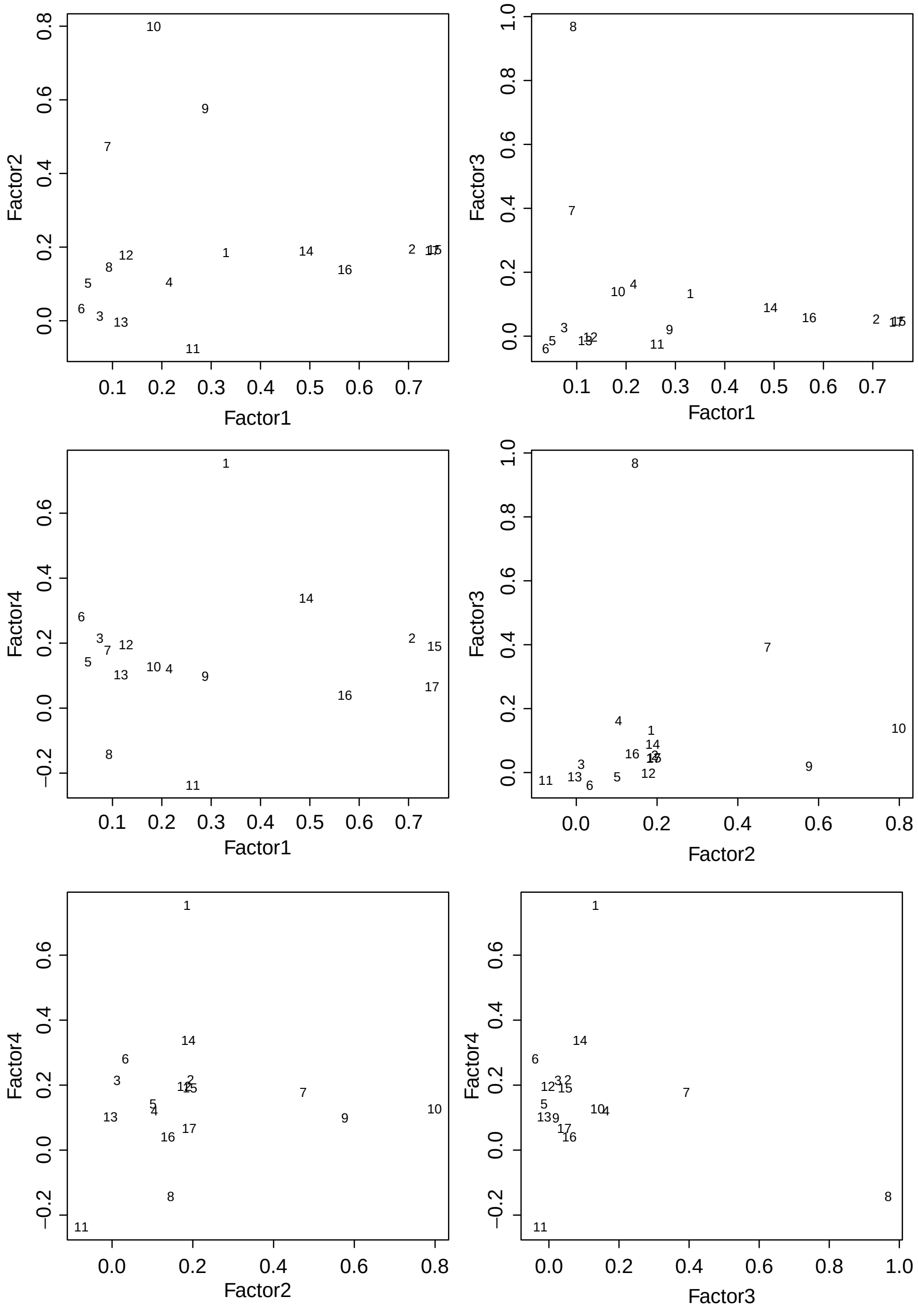


Figure 14.14 Problem 14.5, plots of factor loadings.

voting in the two Florida elections. As the counties are very different sizes, you will need to convert the counts to proportions voting for each candidate, and it may be useful to use the logarithm of this proportion. Fit models with one, two, or three factors and try to interpret them.

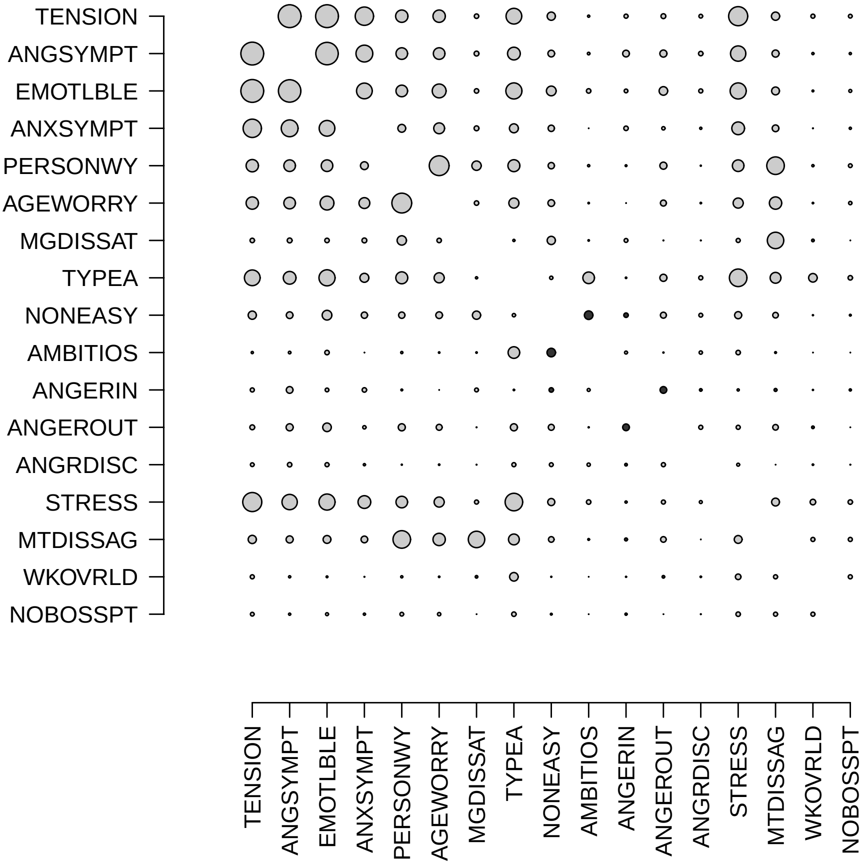


Figure 14.15 Shaded correlation matrix for Problem 14.5.

14.7 Starkweather [1970] performed a study entitled "Hospital Size, Complexity, and Formalization." He states: "Data on 704 United States short-term general hospitals are sorted into a set of dependent variables indicative of organizational formalism and a number of independent variables separately measuring hospital size (number of beds) and various types of complexity commonly associated with size." Here we used his data for a factor analysis of the following variables:

- *SIZE:* number of beds.
- *CONTROL:* a hospital was scored: 1 proprietary control; 2 nonprofit community control; 3 church operated; 4 public district hospital; 5 city or county control; 6 state control.
- *SCOPE* (of patient services): "A count was made of the number of services reported for each sample hospital. Services were weighted 1, 2, or 3 according to their relative impact on hospital operations, as measured by estimated proportion of total operating expenses."
- *TEACHVOL:* "The number of students in each of several types of hospital training programs was weighted and the products summed. The number of paramedical students

Table 14.28 Problem 14.7: Correlation Matrix

		SIZE 1	CONTROL 2	SCOPE 3	TEACHVOL 4	TECHTYPE 5	NONINPRG 6
SIZE	1	1.000					
CONTROL	2	−0.028	1.000				
SCOPE	3	0.743	−0.098	1.000			
TEACHVOL	4	0.717	−0.040	0.643	1.000		
TECHTYPE	5	0.784	−0.034	0.547	0.667	1.000	
NONINPRG	6	0.523	−0.051	0.495	0.580	0.440	1.000

Table 14.29 Problem 14.7: Communalities[a]

1	SIZE	0.8269
2	CONTROL	0.0055
3	SCOPE	0.7271
4	TEACHVOL	0.6443
5	TECHTYPE	1.0000
6	NONINPRG	0.3788

[a] Communalities obtained from two factors after eight iterations. The communality of a variable is its squared multiple correlation with the factors.

Table 14.30 Problem 14.7: Residual Correlations

		SIZE 1	CONTROL 2	SCOPE 3	TEACHVOL 4	TECHTYPE 5	NONINPRG 6
SIZE	1	0.173					
CONTROL	2	0.029	0.995				
SCOPE	3	0.013	−0.036	0.273			
TEACHVOL	4	−0.012	0.012	−0.014	0.356		
TECHTYPE	5	−0.000	0.000	−0.000	−0.000	0.000	
NONINPRG	6	−0.020	−0.008	−0.027	0.094	−0.000	0.621

was weighted by 1.5, the number of RN students by 3, and the number of interns and residents by 5.5. These weights represent the average number of years of training typically involved, which in turn constitute a rough measure of the relative impact of students on hospital operations."

- *TECHTYPE:* types of teaching programs. The following scores were summed: 1 for practical nurse training program; 2 for RN; 3 for medical students; 4 for interns; 5 for residents.
- *NONINPRG:* noninpatient programs. Sum the following scores: 1 for emergency service; 2 for outpatient care; 3 for home care.

The results are given in Tables 14.28 to 14.31, and Figures 14.16 and 14.17. The factor analytic results follow. Perform tasks (a), (c), (d), and (e) for 1, 2, 3, 4, 5, and 6, and tasks (f) and (g).

Table 14.31 Problem 14.7: Factors (Loadings 14.31 Smaller Than 0.1 Omitted)

		Factor 1	Factor 2
SIZE	1	0.636	0.650
CONTROL	2		
SCOPE	3	0.357	0.774
TEACHVOL	4	0.527	0.605
TECHTYPE	5	0.965	0.261
NONINPRG	6	0.312	0.530
	VP[a]	1.840	1.743

[a] The VP for each factor is the sum of the squares of the elements of the column of the factor pattern matrix corresponding to that factor. When the rotation is orthogonal, the VP is the variance explained by the factor.

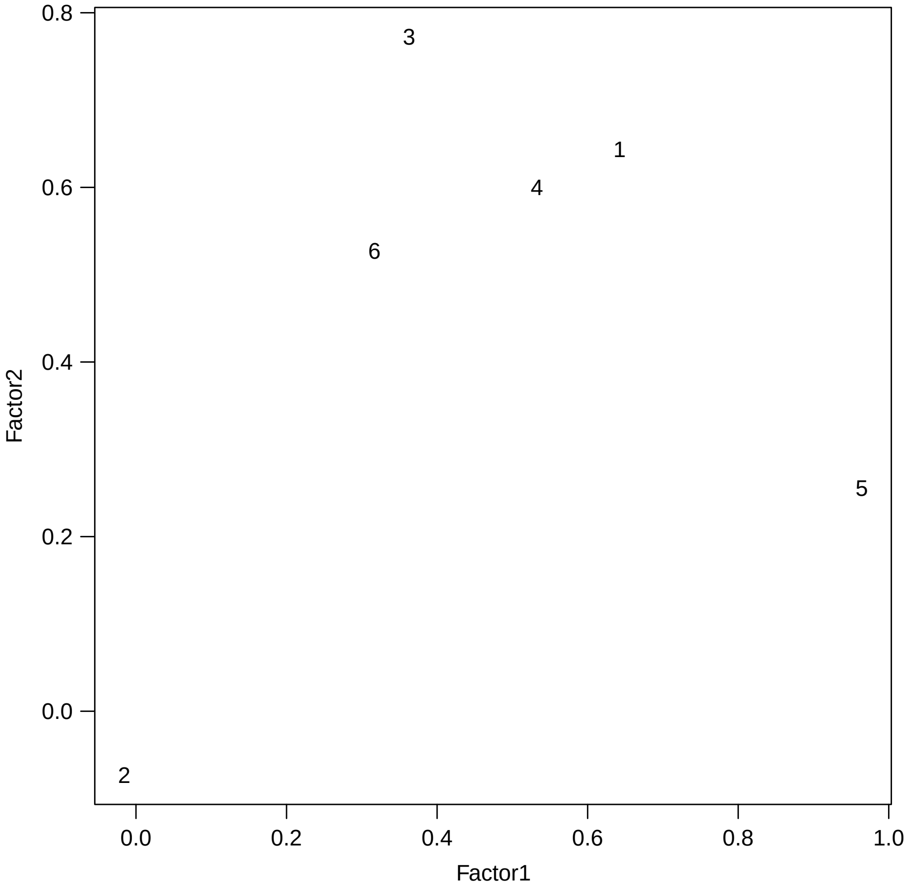


Figure 14.16 Problem 14.7, plot of factor loadings.

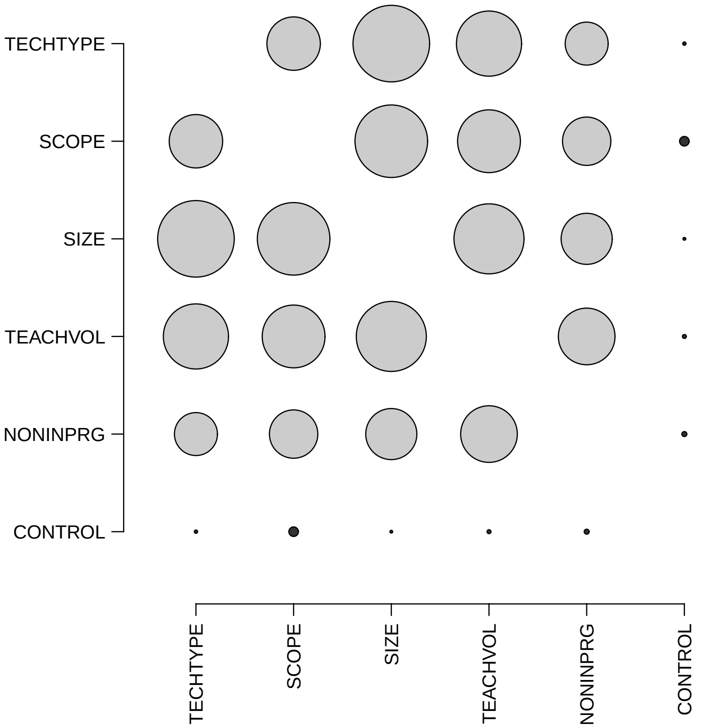


Figure 14.17 Shaded correlation matrix for Problem 14.7.

Table 14.32 Problem 14.8: Residual Correlations

		DURAT	VO_2 MAX	HR	AGE	HT	WT
DURAT	1	0.067					
VO_2 MAX	2	0.002	0.126				
HR	3	−0.005	−0.011	0.678			
AGE	4	0.004	0.011	−0.092	0.441	6	
HT	5	−0.006	0.018	−0.021	0.0106	0.574	
WT	6	0.004	−0.004	−0.008	0.007	0.605	0.301

14.8 This factor analysis examines the data used in Problem 14.3, the maximal exercise test data for sedentary males. The results are given in Tables 14.32 to 14.34 and Figures 14.18 and 14.19. Perform tasks (a), (b) (HR, AGE), (c), (d), and (e) for variables 1 and 5, and tasks (f) and (g).

14.9 Consider two variables, X and Y, with covariances (or correlations) given in the following notation. Prove parts (a) and (b) below.

	Variable	
Variable	**1**	**2**
X	*a*	*c*
Y	*c*	*b*

Table 14.33 Problem 14.8: Communalities[a]

1	DURAT	0.9331
2	VO$_{2\ \text{MAX}}$	0.8740
3	HR	0.5217
4	AGE	0.5591
5	HT	0.4264
6	WT	0.6990

[a] Communalities obtained from two factors after six iterations. The communality of a variable is its squared multiple correlation with the factors.

Table 14.34 Problem 14.8: Factors

		Factor	Factor
		1	2
DURAT	1	0.962	0.646
VO$_{2\ \text{MAX}}$	2	0.930	−0.092
HR	3	0.717	
AGE	4	−0.732	−0.154
HT	5		0.833
WT	6		0.833
	VP[a]	2.856	1.158

[a] The VP for each factor is the sum of the squares of the elements of the column of the factor pattern matrix corresponding to that factor. When the rotation is orthogonal, the VP is the variance explained by the factor.

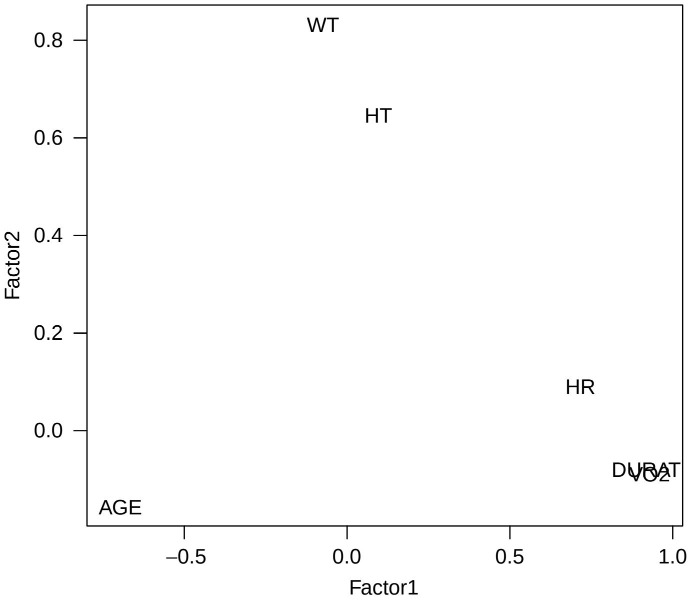


Figure 14.18 Problem 14.8, plot of factor loadings.

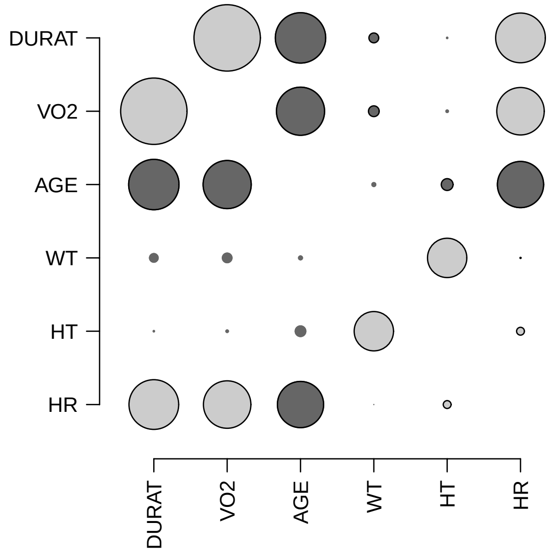


Figure 14.19 Shaded correlation matrix for Problem 14.8.

(a) We suppose that $c \neq 0$. The variance explained by the first principal component is

$$V_1 = \frac{(a+b) + \sqrt{(a-b)^2 + 4c^2}}{2}$$

The first principal component is

$$\sqrt{\frac{c^2}{c^2 + (V_1 - a)^2}}X + \frac{c}{|c|}\sqrt{\frac{(V_1 - a)^2}{c^2 + (V_1 - a)^2}}Y$$

(b) Suppose that $c = 0$. The first principal component is X if $a \geq b$, and is Y if $a < b$.

(c) The introduction to Problems 9.30–9.33 presented data on 20 patients who had their mitral valve replaced. The systolic blood pressure before and after surgery had the following variances and covariance:

	SBP	
	Before	**After**
Before	349.74	21.63
After	21.63	91.94

Find the variance explained by the first and second principal components.

14.10 The exercise data of the 43 active females of Problem 14.4 are used here. The findings are given in Tables 14.35 to 14.37 and Figures 14.20 and 14.21. Perform tasks (a), (c), (d), (f), and (g). Problem 14.8 examined similar exercise data for sedentary males.

Table 14.35 Problem 14.10: Residual Correlations

		DURAT	VO_2 MAX	HR	AGE	HT	WT
DURAT	1	0.151					
VO_2 MAX	2	0.008	0.241				
HR	3	0.039	−0.072	0.687			
AGE	4	0.015	0.001	−0.013	0.416		
HT	5	−0.045	0.013	−0.007	−0.127	0.605	
WT	6	0.000	0.000	0.000	−0.000	0.000	0.000

Table 14.36 Problem 14.10: Communalities[a]

1	DURAT	0.8492
2	VO_2 MAX	0.7586
3	HR	0.3127
4	AGE	0.5844
5	HT	0.3952
6	WT	1.0000

[a] Communalities obtained from two factors after 10 iterations. The communality of a variable is its squared multiple correlation with the factors.

Table 14.37 Problem 14.10: Factors

		Factor 1	Factor 2
DURAT	1	0.907	0.165
VO_2 MAX	2	0.869	
HR	3	0.489	0.271
AGE	4	−0.758	−0.102
HT	5	0.364	0.513
WT	6		0.997
	VP[a]	2.529	1.371

[a] The VP for each factor is the sum of the squares of the elements of the column of the factor pattern matrix corresponding to that factor. When the rotation is orthogonal, the VP is the variance explained by the factor.

Which factor analysis do you feel was more satisfactory in explaining the relationship among variables? Why? Which analysis had the more interpretable factors? Explain your reasoning.

14.11 The data on the correlation among male body measurements (of Problem 14.2) are factor analyzed here. The computer output gave the results given in Tables 14.38 to 14.40 and Figure 14.22. Perform tasks (a), (b) (POPHT, KNEEHT), (STHTER, BUTTKNHT), (RTARMSKN, INFRASCP), and (e) for variables 1 and 11, and tasks (f) and (g). Examine the diagonal of the residual values and the communalities. What values are on the diagonal of the residual correlations? (The diagonals are the 1–1, 2–2, 3–3, etc. entries.)

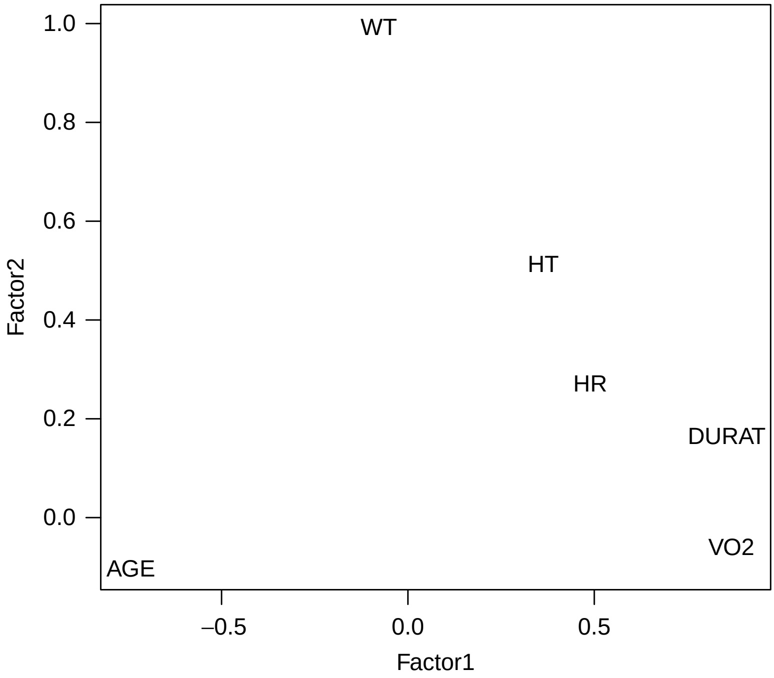


Figure 14.20 Problem 14.10, plot of factor loadings.

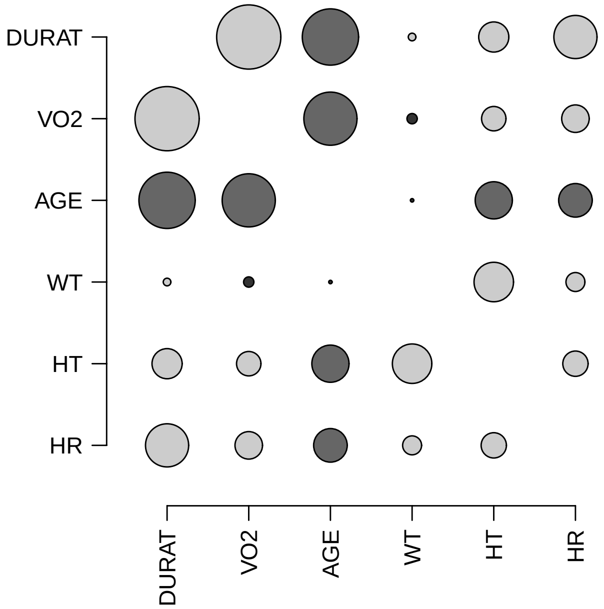


Figure 14.21 Shaded correlation matrix for Problem 14.10.

Table 14.38 Problem 14.11: Residual Correlations

		STHTER	STHTNORM	KNEEHT	POPHT	ELBWHT
		1	2	3	4	5
STHTER	1	0.028				
STHTNORM	2	0.001	0.205			
KNEEHT	3	0.000	−0.001	0.201		
POPHT	4	0.000	−0.006	0.063	0.254	
ELBWHT	5	−0.001	−0.026	−0.012	0.011	0.519
THIGHHT	6	−0.003	0.026	0.009	−0.064	−0.029
BUTTKNHT	7	0.001	−0.004	−0.024	−0.034	−0.014
BUTTPOP	8	−0.001	0.019	−0.038	−0.060	−0.043
ELBWELBW	9	−0.001	0.008	0.007	−0.009	0.004
SEATBRTH	10	−0.002	0.023	0.015	−0.033	−0.013
BIACROM	11	0.006	−0.009	0.009	0.035	−0.077
CHESTGRH	12	−0.001	0.004	−0.004	0.015	−0.007
WSTGRTH	13	0.001	−0.004	−0.002	0.008	0.006
RTARMGRH	14	0.002	0.011	0.012	−0.006	−0.021
RTARMSKN	15	−0.002	0.025	−0.002	−0.012	0.009
INFRASCP	16	−0.002	0.003	−0.009	−0.002	0.020
HT	17	−0.000	0.001	−0.003	−0.003	0.007
WT	18	0.000	−0.007	0.001	0.004	0.007
AGE	19	−0.001	0.006	0.010	−0.014	−0.023

		THIGHHT	BUTTKNHT	BUTTPOP	ELBWELBW	SEATBRTH
		6	7	8	9	10
THIGHHT	6	0.462				
BUTTKNHT	7	0.012	0.222			
BUTTPOP	8	0.016	0.076	0.409		
ELBWELBW	9	0.032	−0.002	0.006	0.215	
SEATBRTH	10	0.023	0.020	−0.017	0.007	0.305
BIACROM	11	−0.052	−0.019	−0.027	0.012	−0.023
CHESTGRH	12	−0.020	−0.013	−0.011	0.025	−0.020
WSTGRTH	13	−0.002	0.006	0.009	−0.006	−0.009
RTARMGRH	14	0.009	0.000	0.013	0.011	−0.017
RTARMSKN	15	0.038	0.039	0.015	−0.019	0.053
INFRASCP	16	−0.025	0.008	−0.000	−0.022	0.001
HT	17	0.005	0.005	0.005	0.000	−0.001
WT	18	−0.004	−0.005	−0.007	−0.006	0.004
AGE	19	−0.012	−0.010	−0.014	0.011	0.007

		BIACROM	CHESTGRH	WSTGRTH	RTARMGRH	RTARMSKN
		11	12	13	14	15
BIACROM	11	0.684				
CHESTGRH	12	0.051	0.150			
WSTGRTH	13	−0.011	0.000	0.095		
RTARMGRH	14	−0.016	−0.011	−0.010	0.186	
RTARMSKN	15	−0.065	−0.011	0.009	0.007	0.601
INFRASCP	16	−0.024	−0.005	0.014	−0.022	0.199
HT	17	−0.008	0.000	−0.003	−0.005	0.004
WT	18	0.006	0.002	0.002	0.006	−0.023
AGE	19	−0.015	−0.006	−0.002	0.014	−0.024

		INFRASCP	HT	WT	AGE
		16	17	18	19
INFRASCP	16	0.365			
HT	17	0.003	0.034		
WT	18	−0.003	0.001	0.033	
AGE	19	−0.022	0.002	0.002	0.311

Table 14.39 Problem 14.11: Communalities[a]

1	STHTER	0.9721
2	STHTNORM	0.7952
3	KNEEHT	0.7991
4	POPHT	0.7458
5	ELBWHT	0.4808
6	THIGHHT	0.5379
7	BUTTKNHT	0.7776
8	BUTTPOP	0.5907
9	ELBWELBW	0.7847
10	SEATBRTH	0.6949
11	BIACROM	0.3157
12	CHESTGRH	0.8498
13	WSTGRTH	0.9054
14	RTARMGRH	0.8144
15	RTARMSKN	0.3991
16	INFRASCP	0.6352
17	HT	0.9658
18	WT	0.9671
19	AGE	0.6891

[a] Communalities obtained from four factors after six iterations. The communality of a variable is its squared multiple correlation with the factors.

Table 14.40 Problem 14.11: Factors (Loadings Smaller Than 0.1 Omitted)

		Factor 1	Factor 2	Factor 3	Factor 4
			Unrotated[a]		
STHTER	1	0.100	0.356	0.908	−0.104
STHTNORM	2	0.168	0.367	0.795	
KNEEHT	3	0.113	0.875	0.128	
POPHT	4	−0.156	0.836	0.133	
ELBWHT	5	0.245	−0.151	0.617	−0.131
THIGHHT	6	0.675	0.131	0.114	−0.230
BUTTKNHT	7	0.308	0.819	0.100	
BUTTPOP	8	0.188	0.742		
ELBWELBW	9	0.873			0.131
SEATBRTH	10	0.765	0.209	0.247	
BIACROM	11	0.351	0.298	0.213	−0.242
CHESTGRH	12	0.902	0.137	0.118	
WSTGRTH	13	0.892			0.323
RTARMGRH	14	0.873			−0.198
RTARMSKN	15	0.625			
INFRASCP	16	0.794			
HT	17		0.836	0.507	−0.098
WT	18	0.907	0.308	0.218	−0.049
AGE	19		−0.135	−0.160	0.801
	VP[a]	6.409	3.964	2.370	0.978

[a] The VP for each factor is the sum of the squares of the elements of the column of the factor pattern matrix corresponding to that factor. When the rotation is orthogonal, the VP is the variance explained by the factor

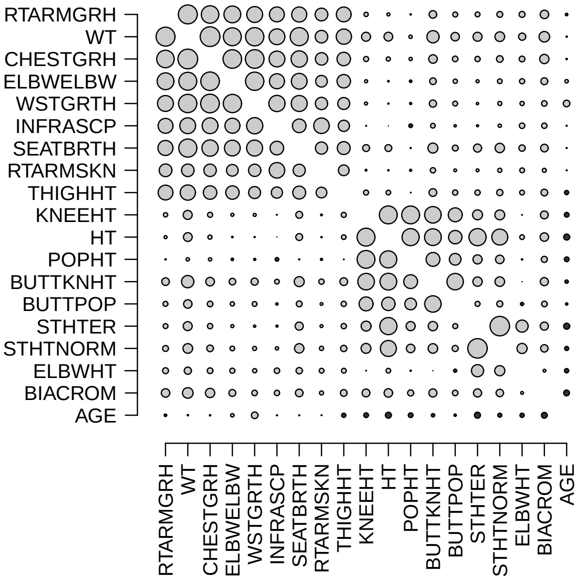


Figure 14.22 Shaded correlation matrix for Problem 14.11.

REFERENCES

Armstrong, J. S. [1967]. Derivation of theory by means of factor analysis, or, Tom Swift and his electric factor analysis machine. *American Statistician* **21**: 17–21.

Bruce, R. A., Kusumi, F., and Hosmer, D. [1973]. Maximal oxygen intake and nomographic assessment of functional aerobic impairment in cardiovascular disease. *American Heart Journal*, **85**: 546–562.

Chaitman, B. R., Fisher, L., Bourassa, M., Davis, K., Rogers, W., Maynard, C., Tyros, D., Berger, R., Judkins, M., Ringqvist, I., Mock, M. B., Killip, T., and participating CASS Medical Centers [1981]. Effects of coronary bypass surgery on survival in subsets of patients with left main coronary artery disease. Report of the Collaborative Study on Coronary Artery Surgery. *American Journal of Cardiology*, **48**: 765–777.

Gorsuch, R. L. [1983]. *Factor Analysis.* 2nd ed. Lawrence Erlbaum Associates, Mahwah, NJ.

Gould, S. J. [1996]. *The Mismeasure of Man.* Revised, Expanded Edition. W.W. Norton, New York.

Guttman, L. [1954]. Some necessary conditions for common factor analysis. *Psychometrika*, **19**(2): 149–161.

Henry, R. C. [1997]. History and fundamentals of multivariate air quality receptor models. *Chemometrics and Intelligent Laboratory Systems* **37**: 525–530.

Jones, M. C., and Sibson, R. [1987]. What is projection pursuit? *Journal of the Royal Statistical Society, Series A*, **150**: 1–36.

Kim, J.-O., and Mueller, C. W. [1999]. *Introduction to Factor Analysis: What It Is and How to Do It.* Sage University Paper 13. Sage Publications, Beverly Hills, CA.

Kim, J.-O., and Mueller, C. W. [1983]. *Factor Analysis: Statistical Methods and Practical Issues.* Sage University Paper 14. Sage Publications, Beverly Hills, CA.

McDonald, R. P. [1999]. *Test Theory: A Unified Treatment. Lawrence* Erlbaum Associates, Mahwah, NJ.

Morrison, D. R. [1990]. *Multivariate Statistical Methods*, 3rd ed. McGraw-Hill, New York.

Paatero, P. [1997]. Least squares formulation of robust, non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems*, **37**: 23–35.

Paatero, P. [1999]. The multilinear engine: a table-driven least squares program for solving multilinear problems, including n-way parallel factor analysis model. *Journal of Computational and Graphical Statistics*, **8**: 854–888.

Reeck, G. R., and Fisher, L. D. [1973]. A statistical analysis of the amino acid composition of proteins. *International Journal of Peptide Protein Research*, **5**: 109–117.

Starkweather, D. B. [1970]. Hospital size, complexity, and formalization. *Health Services Research*, Winter, 330–341. Used with permission from the Hospital and Educational Trust.

Stoudt, H. W., Damon, A., and McFarland, R. A. [1970]. *Skinfolds, Body Girths, Biacromial Diameter, and Selected Anthropometric Indices of Adults: United States, 1960–62*. Vital and Health Statistics. Data from the National Survey. Public Health Service Publication 1000, Series 11, No. 35. U.S. Government Printing Office, Washington, DC.

Timm, N. H. [2001]. *Applied Multivariate Analysis*. Springer-Verlag, New York.

U.S. EPA [2000]. *Workshop on UNMIX and PMF as Applied to* $PM_{2.5}$. National Exposure Research Laboratory, Research Triangle Park, NC. *http://www.epa.gov/ttn/amtic/unmixmtg.html*.