

2

Multivariate Data and the Forward Search

Unlike the other chapters in the book, this chapter contains little data analysis. The emphasis is on theory and on the description of the search. In the first half of the chapter we provide distributional results on estimation, testing and on the distribution of quantities such as squared Mahalanobis distances from samples of size n . The second half of the chapter focuses on the forward search

We start in §2.1 by recalling the univariate normal distribution. Sections 2.2 and 2.3 outline estimation and hypothesis testing for the multivariate normal distribution. As we indicated in Chapter 1, forward plots of Mahalanobis distances are one of our major tools. Since the distribution theory for these distances seems, to us, not to be clear in the literature, we devote §2.4 to §2.6 to deriving the distribution using results on the deletion of observations. As a pendant, in §§2.7 and 2.8, we derive this distribution first using an often quoted result of Wilks and then for regression with a multivariate response. The subject of the following three sections is also regression. In §2.9 we introduce added variables which provide useful results for tests for transformations. These results are applied in §2.10 to the mean shift outlier model to provide an alternative derivation of deletion results which is useful in the analysis of spatial data, Chapter 8. This part of the chapter closes in §2.11 where we outline seemingly unrelated regression, a simplification of the results for multivariate regression when each model contains the same explanatory variables.

A general discussion of the forward search is in §2.12. Three aspects of the search require special attention: how to start, how to progress and what to monitor. These three are treated in detail in §§2.13 and 2.14. The

final theoretical section is on the modifications necessary, particularly to the starting procedure, when we have multivariate regression data. The chapter concludes with suggestions for background reading. We do not discuss in detail alternatives to the forward search. In particular §§4.6 and 4.7 of Atkinson and Riani (2000) contain examples in which the forward search breaks the masking which defeats backwards deletion methods. In this regard, we rest our case.

We close our introduction to this chapter, by recalling our advice in the Preface: “If you feel you know enough statistical theory for your present purposes, continue to Chapter 3.”

2.1 The Univariate Normal Distribution

2.1.1 Estimation

Let $y = (y_1, \dots, y_n)^T$ be a random sample of n observations from a univariate normal distribution with mean μ and variance σ^2 . Then the density of the i th observation is

$$f(y_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-(y_i - \mu)^2 / (2\sigma^2)\}.$$

Sometimes we write this distribution as $y_i \sim N(\mu, \sigma^2)$. The loglikelihood of the n observations is

$$L(\mu, \sigma^2; y) = - \sum_{i=1}^n (y_i - \mu)^2 / (2\sigma^2) - (n/2) \log(2\pi\sigma^2).$$

The maximum likelihood estimator of μ is

$$\hat{\mu} = \bar{y} = \sum_{i=1}^n y_i / n,$$

the sample mean. The sum of squares about the sample mean

$$S(\hat{\mu}) = \sum_{i=1}^n (y_i - \hat{\mu})^2 = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2,$$

leads to estimation of σ^2 . The residual mean square estimator is

$$s^2 = S(\hat{\mu}) / (n - 1),$$

which is unbiased. Maximum likelihood produces the biased estimator

$$\hat{\sigma}^2 = S(\hat{\mu}) / n.$$

Obviously

$$s^2 = \frac{n}{n-1} \hat{\sigma}^2.$$

An alternative way of writing the distribution of the y_i is

$$y_i = \mu + \epsilon_i,$$

where now the independent errors $\epsilon_i \sim N(0, \sigma^2)$. These errors are estimated by the least squares residuals

$$e_i = \hat{\epsilon}_i = y_i - \bar{y}. \quad (2.1)$$

2.1.2 Distribution of Estimators

The sample mean from a normal distribution is itself normally distributed

$$\bar{y} \sim N(\mu, \sigma^2/n)$$

and the residual sum of squares has a scaled chi-squared distribution, so

$$S(\hat{\mu}) = n\hat{\sigma}^2 = (n-1)s^2 \sim \sigma^2 \chi_{n-1}^2. \quad (2.2)$$

The least squares residual e_i (2.1) is a linear combination of normally distributed random variables, so is itself normally distributed, with mean zero. The variance can easily be shown (Exercise 2.2) to be $\sigma^2(1-1/n)$. Therefore

$$\frac{n}{n-1} e_i^2 \sim \sigma^2 \chi_1^2. \quad (2.3)$$

For multivariate data we are interested in the distribution of squared Mahalanobis distances which, in the univariate case, reduce to the squared scaled residual

$$d_i^2 = e_i^2/s^2 = (y_i - \bar{y})^2/s^2. \quad (2.4)$$

For a sample from a normal population, \bar{y} and s^2 are independently distributed, leading for example to the t distribution for the statistic for testing hypotheses about the value of μ . However y_i and s^2 are not independent of each other and the distribution of d_i^2 requires some derivation. Since

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \bar{y})^2 = (n-1)s^2$$

it follows that

$$\sum_{i=1}^n d_i^2 = \sum_{i=1}^n e_i^2/s^2 = n-1,$$

so the d_i^2 must have a distribution with a limited range. The results of Cook and Weisberg (1982, p. 19) show that this distribution is, in fact,

a scaled beta. In §2.6 we obtain the related result for the multivariate Mahalanobis distance. But a couple of preliminary distributional results for the univariate case are helpful.

If both μ and σ^2 are known

$$(y_i - \mu)^2 / \sigma^2 \sim \chi_1^2. \quad (2.5)$$

If now μ is still assumed known, but σ^2 is estimated by s_ν^2 , an estimate on ν degrees of freedom which is independent of y_i ,

$$(y_i - \mu)^2 / s_\nu^2 \sim F_{1,\nu} = t_\nu^2, \quad (2.6)$$

given the identity between the square of a t random variable and an F variable on 1 and ν degrees of freedom. Both the chi-squared and the F distributions are often used as asymptotic approximations to the distribution of the squared Mahalanobis distances, for example in probability plotting. One source for s_ν^2 would be the results of a set of readings different from those for which the Mahalanobis distances were being calculated, but taken from the same population. A second source is to use the deletion estimate $s_{(i)}^2$ in which the i th observation is excluded from the data. As a result, the estimate of σ^2 is independent of y_i . This is the path we follow to find the distribution of the Mahalanobis distance for multivariate data. The deletion results that we need are gathered in §2.5.

2.2 Estimation and the Multivariate Normal Distribution

2.2.1 The Multivariate Normal Distribution

For multivariate data let y_i be the $v \times 1$ vector of responses forming the observation from unit i , with y_{ij} the observation on response j . There are n observations, so the data form a matrix Y of dimension $n \times v$, with i th row y_i^T . The mean of $y_i, i = 1, \dots, n$ is the $v \times 1$ vector μ and the $v \times v$ covariance matrix of the data is Σ . If y_i has a v -variate normal distribution, the density is

$$f(y_i; \mu, \Sigma) = |2\pi\Sigma|^{-1/2} \exp\{-(y_i - \mu)^T \Sigma^{-1} (y_i - \mu)/2\}. \quad (2.7)$$

The multiplicative constant before the exponent may also be written as $(2\pi)^{-v/2} |\Sigma|^{-1/2}$. Sometimes we write this distribution as $y_i \sim N_v(\mu, \Sigma)$, omitting the v if the dimension is obvious.

The loglikelihood of the n observations is

$$L(\mu, \Sigma; y) = -(n/2) \log |2\pi\Sigma| - \sum_{i=1}^n (y_i - \mu)^T \Sigma^{-1} (y_i - \mu)/2. \quad (2.8)$$

The maximum likelihood estimator of the vector μ is now

$$\hat{\mu} = \bar{y} = \left(\sum_{i=1}^n y_{i1}/n, \dots, \sum_{i=1}^n y_{iv}/n \right)^T.$$

Alternatively, if J is an $n \times 1$ vector of ones with Y , as before, $n \times v$,

$$\hat{\mu}^T = J^T Y / n. \quad (2.9)$$

This vector of sample means has a normal distribution

$$\hat{\mu} \sim N_v(\mu, \Sigma/n).$$

2.2.2 The Wishart Distribution

The matrix of sums of squares and products about the sample means is the $v \times v$ matrix $S(\hat{\mu})$ with elements

$$S_{jk}(\hat{\mu}) = \sum_{i=1}^n (y_{ij} - \hat{\mu}_j)(y_{ik} - \hat{\mu}_k). \quad (2.10)$$

In (2.2) the residual sum of squares had a scaled chi-squared distribution. The multivariate generalization is that

$$S(\hat{\mu}) \sim W_v(\Sigma, n-1), \quad (2.11)$$

the v -dimensional Wishart distribution on $n-1$ degrees of freedom. Just as the chi-squared distribution can be defined in terms of sums of squares of independent normal random variables, so the Wishart distribution is defined in terms of sums of squares and products of independent multivariate normal random variables.

If the rows y_i^T of the $n \times v$ matrix Y are distributed as $N_v(0, \Sigma)$, $M = Y^T C Y \sim W_v(\Sigma, n)$. We need two results to extend this definition to the sample sum of squares and products matrix (2.10). The first is that $M = Y^T C Y \sim W_v(\Sigma, r)$ if and only if C is symmetric and idempotent, where $r = \text{tr } C = \text{rank } C$.

The second extension is to variables with non-zero mean. We now let y_i be distributed $N_v(\mu_i, \Sigma)$. Then, in addition to the idempotency condition, $M = Y^T C Y \sim W_v(\Sigma, r)$ if and only if $E(CY) = 0$.

Some derivations are in Mardia, Kent, and Bibby (1979), particularly §3.4 and Exercise 3.4.20.

To verify that this condition is satisfied for $S(\hat{\mu})$ (2.10) it is convenient to use the matrix notation introduced in (2.9). We write

$$S(\hat{\mu}) = \sum_{i=1}^n (y_i - \hat{\mu})(y_i - \hat{\mu})^T = E^T E, \quad (2.12)$$

where E is the $n \times v$ matrix of residuals. Then with \hat{Y} the $n \times v$ matrix of fitted values,

$$\begin{aligned} E &= Y - \hat{Y} \\ &= Y - J\hat{\mu}^T \\ &= Y - JJ^TY/n \\ &= (I - H)Y = CY, \end{aligned}$$

where $H = JJ^T/n$. As a result

$$C = I - JJ^T/n. \quad (2.13)$$

Since C is symmetric and idempotent (Exercise 2.3)

$$S(\hat{\mu}) = E^TE = Y^TC^TCY = Y^TCY,$$

a quadratic form in Y , similar to those following the Wishart distribution. But we also require that $E(CY) = 0$. Since

$$E(Y) = J\mu^T, \quad (2.14)$$

and

$$CJ = J - JJ^TJ/n = J - J = 0,$$

it follows straightforwardly that

$$E(CY) = CE(Y) = CJ\mu^T = 0.$$

Further, since $\text{tr } C = n - 1$, the distributional result stated in (2.11), that $S(\hat{\mu}) \sim W_v(\Sigma, n - 1)$, holds.

2.2.3 Estimation of Σ

The maximum likelihood estimator of Σ is (Exercise 2.7)

$$\hat{\Sigma} = S(\hat{\mu})/n, \quad (2.15)$$

which is biased. The unbiased, method of moments, estimator we denote

$$\hat{\Sigma}_u = S(\hat{\mu})/(n - 1). \quad (2.16)$$

From (2.11) we have the distributional results that

$$n\hat{\Sigma} = (n - 1)\hat{\Sigma}_u = S(\hat{\mu}) \sim W_v(\Sigma, n - 1). \quad (2.17)$$

2.3 Hypothesis Testing

2.3.1 Hypotheses About the Mean

The maximum likelihood estimator of the means $\hat{\mu}$ was defined in (2.9) as the vector of sample means of each response. We derive the maximum likelihood test of the hypothesis

$$D\mu = c, \quad (2.18)$$

where D is an $s \times v$ matrix of full row rank s and c an $s \times 1$ vector of constants, both specified by the null hypothesis. One hypothesis sometimes of interest is that all means have the same unspecified value, when $s = v - 1$ (Exercise 2.8).

The maximum likelihood estimator of Σ was defined in (2.15) as $\hat{\Sigma} = S(\hat{\mu})/n = E^T E/n$. Substitution of this estimator, together with $\hat{\mu}$, into (2.8) yields the maximised loglikelihood

$$\begin{aligned} L(\hat{\mu}, \hat{\Sigma}; y) &= -(n/2) \log |2\pi\hat{\Sigma}| - \sum_{i=1}^n (y_i - \hat{\mu})^T \hat{\Sigma}^{-1} (y_i - \hat{\mu})/2 \\ &= -(n/2) \log |2\pi\hat{\Sigma}| - n \operatorname{tr} E(E^T E)^{-1} E^T/2 \\ &= -(n/2) \log |2\pi\hat{\Sigma}| - nv/2. \end{aligned} \quad (2.19)$$

Let the null hypothesis (2.18) be that $\mu = \mu_0$. The residuals under this hypothesis are E_0 yielding via (2.12) a maximum likelihood estimator $\hat{\Sigma}_0$ of Σ . The maximised loglikelihood (2.19) becomes

$$L(\hat{\mu}_0, \hat{\Sigma}_0; y) = -(n/2) \log |2\pi\hat{\Sigma}_0| - nv/2.$$

Then the differences of maximised loglikelihoods

$$T_{LR} = 2\{L(\hat{\mu}, \hat{\Sigma}; y) - L(\hat{\mu}_0, \hat{\Sigma}_0; y)\} = n \log(|\hat{\Sigma}_0|/|\hat{\Sigma}|), \quad (2.20)$$

has asymptotically a chi-squared distribution on v degrees of freedom when the null hypothesis is true. This statistic is the likelihood ratio test for the hypothesis $\mu = \mu_0$. It is sometimes, here and elsewhere, referred to as the likelihood ratio test and is particularly used in Chapter 4 where we are testing hypotheses about transformations of the data.

2.3.2 Hypotheses About the Variance

Most of the examples in the first five chapters of the book are for data in which there is one multivariate normal population, although the data on Swiss bank notes appears to consist of two, or perhaps three, populations. In Chapter 6 on discriminant analysis we have at least two populations, the analysis of which is simplified if all populations have the same covariance

matrix. One of the aims of data transformations in discriminant analysis is to achieve such equality of covariance matrices. We now present a test of this hypothesis.

Suppose there are g groups of v dimensional observations, with n_l observations in the l th group. The maximum likelihood estimator of Σ in group l is denoted $\hat{\Sigma}_l$ and the pooled estimator over all groups is

$$\sum_{l=1}^g n_l \hat{\Sigma}_l / n, \quad \text{where} \quad n = \sum_{l=1}^g n_l. \quad (2.21)$$

The likelihood ratio test for the hypothesis $\Sigma_1 = \dots = \Sigma_g = \Sigma$ is (Exercise 2.10)

$$T_{LR} = n \log \left| \frac{\sum_{l=1}^g n_l \hat{\Sigma}_l}{n} \right| - \sum_{l=1}^g n_l \log |\hat{\Sigma}_l|. \quad (2.22)$$

Asymptotically (2.22) will have a null chi-squared distribution on $(g-1)v(v+1)/2$ degrees of freedom. An asymptotically equivalent statistic with improved distributional properties for small samples was found by Box (1949) who scaled (2.22) with the numbers of observations replaced by the degrees of freedom, giving the statistic

$$T_{LR}(r) = r \left(\nu \log |\hat{\Sigma}_W| - \sum_{l=1}^g \nu_l \log |\hat{\Sigma}_{ul}| \right), \quad (2.23)$$

where

$$\begin{aligned} \hat{\Sigma}_W &= \frac{1}{\nu} \sum_{l=1}^g \nu_l \hat{\Sigma}_{ul} \\ &= \frac{\sum_{l=1}^g \sum_{i_l=1}^{n_l} (y_{i_l} - \bar{y}_l)(y_{i_l} - \bar{y}_l)^T}{\sum_{l=1}^g \nu_l} \end{aligned}$$

is the within groups unbiased estimator of the covariance matrix. Strictly we should write $\hat{\Sigma}_{Wu}$, but we always divide this sum of squares by the degrees of freedom. The notation y_{i_l} shows that observation i belongs to group l . The factor r in equation (2.23), calculated to improve the chi-squared approximation, is given by

$$r = 1 - \frac{2v^2 + 3v - 1}{6(v+1)(g-1)} \left(\sum_{l=1}^g \nu_l^{-1} - \nu^{-1} \right). \quad (2.24)$$

In (2.24) the degrees of freedom are

$$\nu_l = n_l - 1 \text{ and } \nu = \sum_{l=1}^g \nu_l = n - g$$

for a model in which only a constant μ is fitted to each mean. Further degrees of freedom are lost if the covariance matrices are calculated from residuals from regression (§2.8).

With this result on the test of equality of covariance matrices we have the results we need on estimation of μ and Σ and for testing hypotheses about their values. All are based on aggregate statistics summed over the data. One use of the forward search is to see how these quantities vary as we increase the number of observations in the subset. We shall look at forward plots of several test statistics, particularly, in Chapter 4 on transformations. But now we consider some statistical properties of the Mahalanobis distances for individual observations.

2.4 The Mahalanobis Distance

This book contains many forward plots of Mahalanobis distances. As we shall see, these can be highly informative about the structure of the data. In this and the succeeding four sections, we derive a series of results about the distribution of the squared distances which, unlike the distances themselves, have a tractable distribution. If we require numerical values for the distribution of the distances themselves, we can proceed as we shall do in the construction of the boundaries for the forward plot of distances in Figure 3.6, using the square root of the values from the distribution of the squared distances. In Figure 3.6 these are taken from the asymptotic chi-squared distribution of the squared distances shown in Figure 3.7. Here we find the exact distribution of the squared distances.

The squared population Mahalanobis distance of the i th observation, that is the distance when μ and Σ are both known, is

$$d_i^2(\mu, \Sigma) = (y_i - \mu)^T \Sigma^{-1} (y_i - \mu) \sim \chi_v^2, \quad (2.25)$$

the generalization of the univariate result in (2.5). If Σ in this expression is replaced by $\hat{\Sigma}_\nu$, an unbiased estimator of Σ on ν degrees of freedom which is independent of y_i , the distance

$$d_i^2(\mu, \hat{\Sigma}_\nu) = (y_i - \mu)^T \hat{\Sigma}_\nu^{-1} (y_i - \mu) \sim T^2(v, \nu), \quad (2.26)$$

where $T^2(v, \nu)$ is Hotelling's T^2 with parameters v and ν . This is the generalization of the result for the univariate squared scaled residual in (2.6) which followed a squared t , or F , distribution. Here the F distribution arises since

$$T^2(v, \nu) = \frac{\nu v}{\nu - v + 1} F_{v, \nu - v + 1}. \quad (2.27)$$

Hotelling's T^2 distribution is described in §3.5 of Mardia, Kent, and Bibby (1979).

In the analysis of examples in this book we use the squared Mahalanobis distance

$$d_i^2 = (y_i - \hat{\mu})^T \hat{\Sigma}_u^{-1} (y_i - \hat{\mu}) = e_i^T \hat{\Sigma}_u e_i, \quad (2.28)$$

or its square root d_i , in which both the mean and variance are estimated. As was argued above for the squared scaled residual (2.4), the distribution of this squared distance is affected by the lack of independence between y_i and the estimators of μ and Σ .

We obtain the distribution of the squared Mahalanobis distance in two steps using the deletion of observations. If μ and Σ are estimated with observation i deleted, the results of (2.26) and (2.27) indicate that the deletion distance will follow an F distribution. We find an expression for the squared distance (2.28) as a function of this deletion distance and then rewrite the F distribution as a scaled beta to obtain the required distribution. We start with deletion results.

2.5 Some Deletion Results

2.5.1 The Deletion Mahalanobis Distance

Let $\hat{\mu}_{(i)}$, often read as “mu hat sub i ”, be the estimator of μ when observation i is deleted and, likewise, let $\hat{\Sigma}_{u(i)}$ be the unbiased estimator of Σ based on the $n - 1$ observations when y_i is deleted. The squared deletion Mahalanobis distance is

$$d_{(i)}^2 = (y_i - \hat{\mu}_{(i)})^T \hat{\Sigma}_{u(i)}^{-1} (y_i - \hat{\mu}_{(i)}) = e_{(i)}^T \hat{\Sigma}_{u(i)}^{-1} e_{(i)}. \quad (2.29)$$

We first find an expression for the residual $e_{(i)}$.

Consider just the j th response. Then the residual for y_{ij} , when y_i is excluded from estimation, is

$$\begin{aligned} e_{ij(i)} &= y_{ij} - \bar{y}_{\cdot j(i)} \\ &= y_{ij} - \left(\sum_{l=1}^n y_{lj} - y_{ij} \right) / (n - 1) \\ &= (ny_{ij} - n\bar{y}_{\cdot j}) / (n - 1). \end{aligned} \quad (2.30)$$

So, for the vector of residuals in (2.29)

$$e_{(i)} = e_i + \frac{e_i}{n - 1} = \frac{n}{n - 1} e_i. \quad (2.31)$$

We also need the residual for any other observation y_l , when y_i is excluded. In a similar manner to (2.30)

$$\begin{aligned}
e_{lj(i)} &= y_{lj} - \bar{y}_{\cdot j(i)} \\
&= y_{lj} - \left(\sum_{l=1}^n y_{lj} - y_{ij} \right) / (n-1) \\
&= y_{lj} - \bar{y}_{\cdot j} + (y_{ij} - \bar{y}_{\cdot j}) / (n-1) \\
&= e_{lj} + e_{ij} / (n-1).
\end{aligned} \tag{2.32}$$

Now, for the vector of responses

$$e_{l(i)} = e_l + e_i / (n-1).$$

These results yield expressions for the change in the sum of products $S(\hat{\mu})$ and so in $\hat{\Sigma}_{u(i)}$ on the deletion of y_i . For all observations, an element of $S(\hat{\mu})$ can, from (2.12), be written as

$$S(\hat{\mu})_{jk} = \sum_{l=1}^n e_{lj} e_{lk}.$$

When observation i is deleted, the residuals change and one term is lost from the sum. Then

$$\begin{aligned}
S(\hat{\mu})_{jk(i)} &= \sum_{l \neq i=1}^n e_{lj(i)} e_{lk(i)} \\
&= \sum_{l \neq i=1}^n \{e_{lj} + e_{ij} / (n-1)\} \{e_{lk} + e_{ik} / (n-1)\} \\
&= \sum_{l=1}^n e_{lj} e_{lk} - n e_{ij} e_{ik} / (n-1),
\end{aligned}$$

since the residuals sum to zero over l . Therefore

$$S(\hat{\mu})_{(i)} = S(\hat{\mu}) - n e_i e_i^T / (n-1) = E^T E - n e_i e_i^T / (n-1). \tag{2.35}$$

The unbiased deletion estimator of Σ is

$$\hat{\Sigma}_{u(i)} = S_{(i)}(\hat{\mu}) / (n-2). \tag{2.36}$$

To find $\hat{\Sigma}_{u(i)}^{-1}$ we need a general preliminary result.

2.5.2 The (Bartlett)-Sherman-Morrison-Woodbury Formula

The estimator of Σ is a function of the matrix of sums of squares and products $S(\hat{\mu})_{(i)}$ defined in (2.35). Since we require the inverse of this matrix,

we need an explicit expression for $(E^T E - \alpha e_i e_i^T)^{-1}$. In the development of deletion methods for diagnostics in regression similar results are provided by the Sherman-Morrison-Woodbury formula, sometimes with the name of Bartlett added. It is customary to state the result and then to confirm it gives the correct answer. See, for example Atkinson and Riani (2000, p. 23 and Exercise 2.11). Here we give a constructive derivation due to M. Knott.

Let X be $n \times v$, with i th row x^T , and let $C = X^T X$. A simplified version of the required inverse is

$$(X^T X - x x^T)^{-1} = (C - x x^T)^{-1} = B, \quad (2.37)$$

so that

$$BC - B x x^T = I. \quad (2.38)$$

Postmultiplication of (2.38) by C^{-1} and x , followed by rearrangement leads to

$$Bx = C^{-1}x / (1 - x^T C^{-1}x).$$

Substitution for Bx in (2.38) together with postmultiplication of both sides by C^{-1} , leads to the desired result

$$(C - x x^T)^{-1} = C^{-1} + C^{-1} x x^T C^{-1} / (1 - x^T C^{-1}x), \quad (2.40)$$

on the assumption that all necessary inverses exist. A similar derivation can be used to obtain the more general result in which x^T is replaced by a matrix of dimension $m \times v$, resulting from the deletion of m rows of X .

2.5.3 Deletion Relationships Among Distances

Let $C = E^T E$. Then

$$\hat{\Sigma}_u = C / (n - 1)$$

and, from (2.28), the Mahalanobis distance

$$d_i^2 = (n - 1) e_i^T C^{-1} e_i. \quad (2.42)$$

It is convenient to write

$$e_i^T C^{-1} e_i = d_i^2 / (n - 1) \quad \text{as} \quad g_i \quad \text{and} \quad \alpha = n / (n - 1).$$

From (2.35) and (2.40)

$$\begin{aligned} S^{-1}(\hat{\mu})_{(i)} &= (C - \alpha e_i e_i^T)^{-1} \\ &= C^{-1} + \alpha C^{-1} e_i e_i^T C^{-1} / (1 - \alpha e_i^T C^{-1} e_i). \end{aligned} \quad (2.44)$$

Then

$$e_i^T S^{-1}(\hat{\mu})_{(i)} e_i = g_i + \alpha g_i^2 / (1 - \alpha g_i) = g_i / (1 - \alpha g_i).$$

Finally we combine the definition of $d_{(i)}^2$ (2.29) with that of $\hat{\Sigma}_{u(i)}$, the unbiased deletion estimator of Σ to obtain

$$\begin{aligned} d_{(i)}^2 &= \frac{(n-2)n^2}{(n-1)^2} e_i^T S^{-1}(\hat{\mu})_{(i)} e_i \\ &= \frac{(n-2)n^2}{(n-1)^3} \frac{d_i^2}{1 - n d_i^2 / (n-1)^2}. \end{aligned} \quad (2.46)$$

The inversion of this relationship provides an expression for d_i^2 as a function of the squared deletion distance

$$d_i^2 = \frac{(n-1)^3 d_{(i)}^2}{n^2(n-2) + n(n-1)d_{(i)}^2}. \quad (2.47)$$

2.6 Distribution of the Squared Mahalanobis Distance

In the deletion Mahalanobis distance Σ is estimated on $\nu = n - 2$ degrees of freedom. So, from (2.27), the distance with known mean $d_i^2(\mu, \hat{\Sigma}_\nu)$ in (2.26) has the scaled F distribution

$$d_i^2(\mu, \hat{\Sigma}_\nu) \sim \frac{v(n-2)}{n-v-1} F_{v, n-v-1}.$$

But the squared deletion distance is a quadratic form in $y_i - \bar{y}_{(i)}$ whereas in (2.26) we have a quadratic in $y_i - \mu$. But, as in (2.3), the variance of $y_i - \bar{y}_{(i)}$ is $n/(n-1)$ times that of $y_i - \mu$. The distribution of the deletion Mahalanobis distance is then given by

$$d_{(i)}^2 \sim \frac{n}{(n-1)} \frac{v(n-2)}{(n-v-1)} F_{v, n-v-1}. \quad (2.49)$$

To find a compact expression for the distribution of the squared Mahalanobis distance it is helpful to write the F distribution as a scaled ratio of two independent chi-squared random variables

$$F_{v, n-v-1} = \frac{\chi_v^2/v}{\chi_{n-v-1}^2/(n-v-1)}.$$

Then, from (2.49)

$$d_{(i)}^2 \sim \frac{n(n-2)}{n-1} \frac{\chi_v^2}{\chi_{n-v-1}^2},$$

where, again, the two chi-squared variables are independent. It then follows from (2.47) that the distribution of the Mahalanobis distance is given by

$$d_i^2 \sim \frac{(n-1)^2}{n} \frac{\chi_v^2}{\chi_v^2 + \chi_{n-v-1}^2}.$$

We now apply two standard distributional results. The first is that

$$\chi_\nu^2 = \Gamma\left(\frac{\nu}{2}, \frac{1}{2}\right).$$

The second is that if X_1 and X_2 are independently $\text{Gamma}(p, \lambda)$ and $\text{Gamma}(q, \lambda)$, then $X_1/(X_1 + X_2) \sim \text{Beta}(p, q)$. We finally obtain

$$d_i^2 \sim \frac{(n-1)^2}{n} \text{Beta}\left(\frac{v}{2}, \frac{n-v-1}{2}\right). \quad (2.52)$$

For moderate n the range of this distribution is approximately $(0, n)$ rather than the unbounded range for the F distribution of the deletion distances.

This result derives the exact distribution of the individual squared Mahalanobis distances. That the distribution is bounded follows from the relationship

$$\sum_{i=1}^n d_i^2 = v(n-1), \quad (2.53)$$

the proof of which is left to the exercises. A summary of the distributional results for various Mahalanobis distances is in Table 2.1

2.7 Determinants of Dispersion Matrices and the Squared Mahalanobis Distance

We now outline an alternative derivation of the distribution of the Mahalanobis distance due to Wilks (1963). As a measure of outlyingness of a multivariate observation he proposed the scatter ratio

$$R_i = |S_{(i)}(\hat{\mu})|/|S(\hat{\mu})|$$

and showed that

$$R_i \sim \text{Beta}\left(\frac{n-v-1}{2}, \frac{v}{2}\right).$$

To relate this ratio of determinants to the Mahalanobis distance we need the standard result for the $n \times p$ matrix X

$$|X^T X - x x^T| = |X^T X| \{1 - x^T (X^T X)^{-1} x\}, \quad (2.56)$$

TABLE 2.1. Summary of distributional results for the squared Mahalanobis distances used in this book; y_i is a $v \times 1$ vector of responses from $N_v(\mu, \Sigma)$, x_i is a $p \times 1$ vector of regression variables and B is a $p \times v$ matrix of regression parameters

Reference	μ	Σ	Distribution
(2.25)	known	known	χ_v^2
(2.26) and (2.27)	known	estimated independently of y_i on ν degrees of freedom	$T^2(v, \nu) = \frac{\nu v}{\nu - v + 1} F_{v, \nu - v + 1}$
Exercise 2.4	known	unknown ($v = 1$) σ^2 estimated by s^2	$(n - 1) \text{Beta} \left(\frac{1}{2}, \frac{n-1}{2} \right)$
(2.49)	unknown	unknown	$\frac{n}{(n-1)} \frac{v(n-2)}{n-v-1} F_{v, n-v-1}$
(deletion distance)	estimated by $\hat{y}_{(i)}$	estimated by $\hat{\Sigma}_{u(i)}$	
(2.52) and Exercise 2.5	unknown estimated by $\hat{\mu}$	unknown estimated by $\hat{\Sigma}_u$	$\frac{(n-1)^2}{n} \text{Beta} \left(\frac{v}{2}, \frac{n-v-1}{2} \right)$ for $v = 1$ the distribution of the squared scaled residual (2.4)
(2.79)	$\mu = B^T x_i$ unknown estimated by $\hat{B}^T x_i$	unknown estimated by $\hat{\Sigma}_u$	$(n - p)(1 - h_i) \text{Beta} \left(\frac{v}{2}, \frac{n-v-p}{2} \right)$

The limiting distribution is χ_v^2 .

where here x^T is one of the rows of X and it is assumed that $(X^T X)^{-1}$ exists (Rao 1973, p. 32). We now apply this relationship to the matrix of residuals.

We recall (2.35)

$$S(\hat{\mu})_{(i)} = S(\hat{\mu}) - ne_i e_i^T / (n-1) = E^T E - ne_i e_i^T / (n-1) = C - ne_i e_i^T / (n-1).$$

Then, in (2.56)

$$|S(\hat{\mu})_{(i)}| = |S(\hat{\mu})| \{1 - ne_i^T C^{-1} e_i / (n-1)\},$$

so that Wilks' result becomes

$$1 - \frac{n}{n-1} e_i^T C^{-1} e_i \sim \text{Beta} \left(\frac{n-v-1}{2}, \frac{v}{2} \right).$$

Next we recall that if the random variable $X \sim \text{Beta}(\alpha, \beta)$, $1-X \sim \text{Beta}(\beta, \alpha)$, when, invoking the notation of (2.42), we obtain

$$\frac{n}{(n-1)^2} d_i^2 \sim \text{Beta} \left(\frac{v}{2}, \frac{n-v-1}{2} \right),$$

which is (2.52).

2.8 Regression

In many of the examples in this book the data, perhaps after transformation and the removal of outliers, follow the multivariate normal distribution (2.7) in which each observation y_{ij} on the j th response has mean μ_j . However, in some examples, the mean has a regression structure. The simplest, considered in this section, is when the regressors for each of the v responses are the same. Then (2.14) becomes

$$E(Y) = XB, \tag{2.57}$$

where Y is $n \times v$, X is the $n \times p$ matrix of regression variables and B is a $p \times v$ matrix of parameters. The $v \times v$ covariance matrix of the data remains Σ . Then, for an individual observation

$$E(y_{ij}) = \mu_{ij} = x_i^T \beta_j, \tag{2.58}$$

where x_i^T is the i th row of X and β_j the j th column of B .

We now consider estimation of B and Σ . If there are different sets of explanatory variables for the different responses, so that we write X_j , rather than the common X , estimation of B requires knowledge, or an estimate,

of Σ . This is the subject of the next section. Here, with a common X , each estimate $\hat{\beta}_j$ is found from univariate regression of y_{C_j} on X , that is

$$\hat{\beta}_j = (X^T X)^{-1} X^T y_{C_j}. \quad (2.59)$$

The matrix of residuals E has elements

$$e_{ij} = y_{ij} - x_i^T \hat{\beta}_j, \quad (2.60)$$

and the sum of squares and products matrix is

$$S(\hat{\beta}) = E^T E. \quad (2.61)$$

In an extension of (2.11) this matrix again has a Wishart distribution

$$S(\hat{\beta}) \sim W_v(\Sigma, n - p). \quad (2.62)$$

To prove this result in a manner analogous to that of §2.2.2 requires the use of some standard results in regression. Here we use notation similar to that of Atkinson and Riani (2000, Chapter 2).

The hat matrix H is defined as

$$H = X(X^T X)^{-1} X^T, \quad (2.63)$$

so called because the matrix of fitted values $\hat{Y} = HY$. The i th diagonal element of H is

$$h_i = x_i^T (X^T X)^{-1} x_i. \quad (2.64)$$

The theorems relating to the Wishart distribution of the matrix of the sum of squares and products are analogous to those in §2.2.2 with the matrix C (2.13) replaced by the symmetric idempotent matrix $I - H$.

The maximum likelihood estimator of Σ is basically unchanged,

$$\hat{\Sigma} = S(\hat{\beta})/n,$$

which is biased. The unbiased, method of moments, estimator becomes

$$\hat{\Sigma}_u = S(\hat{\beta})/(n - p). \quad (2.66)$$

To find the distribution of the Mahalanobis distance, we again use deletion methods. A standard result in regression diagnostics for the change in the residual sum of squares on deletion of an observation (Atkinson and Riani 2000, p. 26, eq. 2.48) can be written in our notation for the j th response as

$$\sum_{l \neq i=1}^n e_{lj(i)}^2 = \sum_{l=1}^n e_{lj}^2 - e_{ij}^2/(1 - h_i),$$

so that (2.35) becomes

$$S(\hat{\beta}) = S(\hat{\beta})_{(i)} - e_i e_i^T / (1 - h_i) = E^T E - e_i e_i^T / (1 - h_i). \quad (2.67)$$

For a model in which only the mean is fitted, $h_i = 1/n$ and (2.67) reduces to (2.35). The unbiased deletion estimator of Σ for the regression model is

$$\hat{\Sigma}_{u(i)} = S(\hat{\beta})_{(i)} / (n - p - 1). \quad (2.68)$$

The deletion Mahalanobis distance (2.29) is a function of this matrix and of the residuals

$$y_i - \hat{\mu}_{(i)} = y_i - \hat{\beta}_{(i)}^T x_i.$$

The standard result for the deletion estimator $\hat{\beta}_{(i)}$ in regression (for example (2.94) or Atkinson and Riani 2000, p. 23) shows that

$$y_i - \hat{\beta}_{(i)}^T x_i = e_i / (1 - h_i). \quad (2.69)$$

In the deletion distance Σ is estimated with $n - p - 1$ degrees of freedom. Then the distance with known mean $d_i^2(\mu, \hat{\Sigma}_{n-p-1})$ in (2.26) has a scaled F distribution

$$d_i^2(\mu, \hat{\Sigma}_{n-p-1}) \sim \frac{v(n-p-1)}{n-v-p} F_{v, n-v-p}.$$

But the squared deletion distance is a quadratic form in $e_i / (1 - h_i)$. The variance of each element of e_i is $(1 - h_i)$ times that of the corresponding element of y_i , so the vector has variance $1/(1 - h_i)$ times that of $y_i - \mu$ in (2.26). The distribution of the deletion Mahalanobis distance is therefore now given by

$$d_{(i)}^2 \sim \frac{1}{(1 - h_i)} \frac{v(n-p-1)}{(n-v-p)} F_{v, n-v-p}. \quad (2.71)$$

The next stage in the argument is to find the relationship between the distance d_i^2 and the deletion distance. As before, let $C = E^T E$. Then

$$\hat{\Sigma}_u = C / (n - p)$$

and the squared Mahalanobis distance is

$$d_i^2 = (n - p) e_i^T C^{-1} e_i. \quad (2.72)$$

If now we write

$$e_i^T C^{-1} e_i = d_i^2 / (n - p) \quad \text{as} \quad g_i \quad \text{and} \quad \alpha = 1 / (1 - h_i),$$

application of (2.44) leads to

$$e_i^T S_{(i)}^{-1}(\hat{\beta}) e_i = \frac{d_i^2}{n - p - d_i^2 / (1 - h_i)}.$$

The combination of this result with the definition of the unbiased deletion estimator $\hat{\Sigma}_{u(i)}$ in (2.36) together with the residuals $e_i / (1 - h_i)$ (2.69) yields the required relationship

$$d_{(i)}^2 = \frac{(n - p - 1)}{(1 - h_i)^2 (n - p)} \frac{d_i^2}{1 - d_i^2 / \{(n - p)(1 - h_i)\}}, \quad (2.75)$$

which reduces to (2.46) when the linear model contains just a mean, that is when $p = 1$ and $h_i = 1/n$. The inversion of this relationship again provides an expression for d_i^2 as a function of the squared deletion distance

$$d_i^2 = \frac{(1 - h_i)^2(n - p)d_{(i)}^2}{(n - p - 1) + (1 - h_i)d_{(i)}^2}. \quad (2.76)$$

To find the distribution of this squared Mahalanobis distance we start from (2.71) proceeding as in §2.6. The distribution of the deletion distance (2.71) is again written as the distribution of the ratio of two independent chi-squared random variables

$$d_{(i)}^2 \sim \frac{n - p - 1}{1 - h_i} \frac{\chi_v^2}{\chi_{n-v-p}^2}.$$

It now follows from (2.76) that the distribution of the Mahalanobis distance is given by

$$d_i^2 \sim (n - p)(1 - h_i) \frac{\chi_v^2}{\chi_v^2 + \chi_{n-v-p}^2}.$$

Finally, we again use the relationship between beta and gamma random variables employed in §2.6 to obtain

$$d_i^2 \sim (n - p)(1 - h_i) \text{Beta} \left(\frac{v}{2}, \frac{n - v - p}{2} \right). \quad (2.79)$$

so that the range of support of the distribution of d_i^2 depends upon h_i . For some balanced experimental designs, such as two-level factorials, all h_i are equal when all n observations are included in the subset and so $\sum h_i = p$, when each $h_i = p/n$. Then the distribution (2.79) reduces to

$$d_i^2 \sim \frac{(n - p)^2}{n} \text{Beta} \left(\frac{v}{2}, \frac{n - v - p}{2} \right).$$

However, the distances will not all have the same distribution in the rest of the search. A more complicated instance of unequal leverages is when we include in the model a constructed variable for transformation of the response §4.4, the value for which depends on each observed y_{ij} . But then fitting the regression model requires estimation of Σ . We discuss the resulting regression procedure in §2.11.

2.9 Added Variables in Regression

The previous section concludes our work on the distribution theory of squared Mahalanobis distances. In this and the next section we describe

some more general properties of regression models. For the moment we continue with multiple regression when the matrix of explanatory variables is the same for all responses, that is when (2.57) holds. Each response is then analysed separately using univariate regression. Our purpose is to provide some background to the development of approximate score tests for transformation of the data developed in §4.4. Even if the means of the data do not contain any regression structure, which is the usual situation, the algebra of added variables described in this section provides a convenient way of using the forward search to assess transformations.

The underlying idea is that multiple regression can be performed as a series of simple linear regressions on single explanatory variables, although both the response and the regression variable have to be adjusted for the variables already in the model. We then perform a regression of residuals on residuals. To begin the derivation of the results we extend the univariate regression model to include an extra explanatory variable, the added variable w , so that the model is

$$E(y) = X\beta + w\gamma, \quad (2.81)$$

where y is $n \times 1$, β is $p \times 1$ and γ is a scalar. We find explicit expressions for the least squares estimate of γ and the statistic for testing its value. Added variables are important in the development of regression diagnostics, where they are used to provide graphical representations (added variable plots) for the importance of individual observations to evidence for regression on w . They are also important in providing tests for transformations and in the development of regression diagnostics using the mean shift outlier model, which is briefly introduced in the next section. These uses are described in Atkinson and Riani (2000, §2.2), where full details are given. In this book, since few of our examples include regression, we give a brief summary of the method, which is used in §4.13 where we select a regression model.

An expression for the estimate $\hat{\gamma}$ of γ in (2.81) can be found explicitly from the normal equations for this partitioned model

$$X^T X \hat{\beta} + X^T w \hat{\gamma} = X^T y \quad (2.82)$$

and

$$w^T X \hat{\beta} + w^T w \hat{\gamma} = w^T y. \quad (2.83)$$

If the model without γ can be fitted, $(X^T X)^{-1}$ exists and (2.82) yields

$$\hat{\beta} = (X^T X)^{-1} X^T y - (X^T X)^{-1} X^T w \hat{\gamma}. \quad (2.84)$$

Substitution of this value into (2.83) leads, after rearrangement, to

$$\hat{\gamma} = \frac{w^T (I - H) y}{w^T (I - H) w} = \frac{w^T A y}{w^T A w}. \quad (2.85)$$

Since $A = (I - H)$ is idempotent, $\hat{\gamma}$ can be expressed in terms of the two sets of residuals

$$e = {}^*y = (I - H)y = Ay$$

and

$$w = {}^*w = (I - H)w = Aw \quad (2.86)$$

as

$$\hat{\gamma} = {}^*w^T e / ({}^*w^T {}^*w). \quad (2.87)$$

Thus $\hat{\gamma}$ is the coefficient of linear regression through the origin of the residuals e on the residuals of the new variable w , both after regression on the variables in X .

To calculate the t statistic requires the variance of $\hat{\gamma}$. Since, like any least squares estimate in a linear model, $\hat{\gamma}$ is a linear combination of the observations,

$$\text{var } \hat{\gamma} = \sigma^2 \frac{w^T A^T A w}{(w^T A w)^2} = \frac{\sigma^2}{w^T A w} = \sigma^2 / ({}^*w^T {}^*w). \quad (2.88)$$

Calculation of the test statistic also requires s_w^2 , the residual mean square estimate of σ^2 from regression on X and w , given by (Atkinson and Riani 2000, eq. 2.28)

$$\begin{aligned} (n - p - 1)s_w^2 &= y^T y - \hat{\beta}^T X^T y - \hat{\gamma} w^T y \\ &= y^T A y - (y^T A w)^2 / (w^T A w). \end{aligned} \quad (2.89)$$

The t statistic for testing that $\gamma = 0$ is then

$$t_w = \frac{\hat{\gamma}}{\sqrt{\{s_w^2 / (w^T A w)\}}}. \quad (2.90)$$

If w is the explanatory variable x_k , (2.90) is an alternative way of writing the usual t test for x_k in multiple regression. But the usual regression t tests are hard to interpret in the forward search, decreasing markedly as the search progresses; Figure 3.4 of Atkinson and Riani (2000) is one example. The problem arises in multiple regression because the search orders the observations using all variables including x_k . We obtain t tests for x_k with the correct distributional properties by taking x_k as w in (2.81) with X all the other explanatory variables. The forward search orders the data by all variables except x_k . Because of the orthogonal projection in (2.86), the t test (2.90) is unaffected by the ordering of the observations in the search. A fuller discussion is in Atkinson and Riani (2002a). Our example is in §4.7.

2.10 The Mean Shift Outlier Model

In §2.5 we obtained some deletion results for Mahalanobis distances using results derived from the (Bartlett)-Sherman-Morrison-Woodbury formula

(2.40). In this section we sketch how the mean shift outlier model can be used to obtain deletion results for the more general case of regression, using the relationships for added variables derived in the previous section. The standard results for deletion in regression are summarized, for example, by Atkinson and Riani (2000, §2.3).

Formally the model is similar to that of (2.81). We write

$$E(y) = X\beta + q(i)\phi, \quad (2.91)$$

where the $n \times 1$ vector $q(i)$ is all zeroes apart from a single one in the i th position and ϕ is a scalar parameter. Observation i therefore has its own parameter and, when the model is fitted, the residual for observation i will be zero; fitting (2.91) thus yields the same residual sum of squares as deleting observation i and refitting.

To show this equivalence requires some properties of $q(i)$. Since it is a vector with one nonzero element equal to one, it extracts elements from vectors and matrices, for example:

$$q(i)^T e = e_i, \quad X^T q(i) = x_i \quad \text{and} \quad q(i)^T H q(i) = h_i. \quad (2.92)$$

Then, from (2.85),

$$\hat{\phi} = \frac{q(i)^T A y}{q(i)^T A q(i)} = \frac{e_i}{1 - h_i}. \quad (2.93)$$

If the parameter estimate in the mean shift outlier model is denoted $\hat{\beta}_q$, it follows from (2.84) that

$$\hat{\beta}_q = (X^T X)^{-1} X^T y - (X^T X)^{-1} X^T q(i) \hat{\phi},$$

so that, from (2.93)

$$\hat{\beta}_q - \hat{\beta} = -(X^T X)^{-1} x_i e_i / (1 - h_i). \quad (2.94)$$

Comparison of (2.94) with standard deletion results shows that $\hat{\beta}_q = \hat{\beta}_{(i)}$, confirming the equivalence of deletion and a single mean shift outlier.

The expression for the change in residual sum of squares comes from (2.89). If the new estimate of σ^2 is s_q^2 we have immediately that

$$\begin{aligned} (n - p - 1)s_q^2 &= y^T A y - \{y^T A q(i)\}^2 / \{q(i)^T A q(i)\} \\ &= (n - p)s^2 - e_i^2 / (1 - h_i), \end{aligned} \quad (2.95)$$

where $s_q^2 = s_{(i)}^2$, the deletion estimate.

The mean shift outlier model likewise provides a simple method of finding the effect of multiple deletion. We first need to extend the results on added variables in §2.9 to the addition of m variables, so that Q is an $n \times m$

matrix and γ an $m \times 1$ vector of parameters. We then apply these results to the mean shift outlier model

$$E(y) = X\beta + Q\phi,$$

with Q a matrix that has a single one in each of its columns, which are otherwise zero, and m rows with one nonzero element. These m entries specify the observations that are to have individual parameters or, equivalently, are to be deleted.

2.11 Seemingly Unrelated Regression

When there are different linear models for the v responses, the regression model for the j th response can be written

$$E(y_{C_j}) = X_j\beta_j, \quad (2.96)$$

where y_{C_j} is the $n \times 1$ vector of responses (j th column of matrix Y). Here X_j is an $n \times p$ matrix of regression variables, as was X in (2.57), but now those specifically for the j th response, and β_j is a $p \times 1$ vector of parameters. In our applications we do not need the more general theory in which the number of parameters p_j in the model depends upon the particular response. The extension of the theory to this case is straightforward, but is not considered here.

Because the explanatory variables are no longer the same for all responses the simplification of the regression in §2.8 no longer holds: the covariance Σ between the v responses has to be allowed for in estimation and independent least squares is replaced by generalized least squares. The model for all n observations can be written in the standard form of (2.57) by stacking the equations under each other. In this form the model is that for a vector of nv observations on a heteroscedastic univariate response variable and the vector of parameters β is of dimension $pv \times 1$. If we let Ψ be the $nv \times nv$ covariance matrix of the observations, generalized least squares yields the parameter estimator

$$\hat{\beta} = (X^T \Psi^{-1} X)^{-1} X^T \Psi^{-1} Y, \quad (2.97)$$

with covariance matrix

$$\text{var } \hat{\beta} = (X^T \Psi^{-1} X)^{-1}, \quad (2.98)$$

where X is $nv \times pv$. In the particular form of generalized least squares resulting from stacking equations of the form of (2.96) the parameters for each response are different, the estimates being related only through covariances of the y_{C_j} . This special structure is known as seemingly unrelated regression (Zellner 1962).

In all there are nv observations. When the data are stacked the covariance matrix Ψ is block diagonal with n blocks of the $v \times v$ matrix Σ . As a result of the block diagonal structure the calculation of the parameters β can be achieved without inversion of an $nv \times nv$ matrix.

There are $p^* = vp$ parameters to be estimated. Let X^* be the $n \times p^*$ matrix of explanatory variables formed by copying each column of the X_j in order - first all the elements of the first column of each X_j , then all the second columns and so on up to the last columns of each X_j . The elements of X^* are x_{ij}^* . If β^* is the $p^* \times 1$ vector of parameters, calculation of the least squares estimates requires the $p^* \times p^*$ covariance matrix Ψ . Let J be a vector of ones of dimension $n \times 1$. Then

$$\Psi^{-1} = JJ^T \otimes \Sigma^{-1}, \quad (2.99)$$

a matrix containing $n \times n$ copies of Σ^{-1} . In (2.99) \otimes denotes the Kronecker product. The vector of parameter estimates can then be written in the seemingly standard least squares form $\hat{\beta}^* = A^{-1}B$ where

$$\begin{aligned} A_{jk} &= \sum_{i=1}^n x_{ij}^* \Psi_{jk}^{-1} x_{ik}^* \\ B_j &= \sum_{i=1}^n \sum_{k=1}^v x_{ij}^* \Psi_{jk}^{-1} y_{ik}. \end{aligned} \quad (2.100)$$

Although the pattern is clear, the matrices do not combine according to dimensions and the summations are over the n observations rather than the nv of the stacked data. Discussion of seemingly unrelated regression is to be found in many textbooks on econometrics, for example §2.9 of Harvey (1990).

Because (2.100) contains Ψ , estimation of Ψ , or equivalently Σ , is required for the procedure to be operational. The estimation proceeds in two or more steps:

1. Obtain $\hat{\Sigma}_0$, an estimate of Σ , from the independent regressions as in (2.61), but with y_{C_j} regressed on X_j .
2. Seemingly unrelated regression using (2.100) with $\hat{\Psi}^{-1}$ in (2.99) calculated using $\hat{\Sigma}_0^{-1}$.
3. Iteration in the estimation of Ψ is possible, starting with the estimate of Σ obtained from Step 2 and repeating the seemingly unrelated regression calculations until there is no significant change in the estimates of the covariance matrices.

Much of the emphasis so far in this chapter has been on the distribution of the statistics we have calculated, particularly the Mahalanobis distances. However, such results are not available for the seemingly unrelated regression procedure of this section. Under the assumption of normally distributed errors, the estimate in β from generalized least squares in (2.97)

has the normal distribution. But with Ψ estimated from the data, the distribution is not readily determined. If the exact distribution is important, recourse may be had to simulation. But, we use the asymptotic results which apply when Ψ is known.

2.12 The Forward Search

Examples of the forward search were given in Chapter 1. During these we monitored the behaviour of the minimum Mahalanobis distance for units not in the subset as the data were fitted to increasingly large subsets. In this chapter we have introduced further quantities that can be monitored during the forward search. We now briefly describe the search, which is made up of three steps: the first is the choice of an initial subset, the second the way in which we progress in the forward search and the third is the monitoring of the statistics during the progress of the search. In subsequent sections we discuss in some detail how to start the search and what quantities it is interesting to monitor. Here we discuss its properties.

The purpose of the forward search is to identify observations which are different from the majority of the data and to determine the effect of these observations on inferences made about the correct model. There may be a few outliers or it may be, as in the data on Swiss bank notes, that the observations can be divided into groups, so that it is appropriate to fit a different model to each group. Although it is convenient to refer to such observations as “outliers”, they may well form a large part of the data and indicate unsuspected structure. Such structure is often impossible to detect from a model fitted to all the data. The effect of the outliers is masked and backwards methods using the deletion of observations fail to show any important features.

If the values of the parameters of the model were known, there would be no difficulty in detecting the outliers, which would have large Mahalanobis distances. The difficulty arises because the outliers are included in the data used for fitting the model, leading to parameter estimates which can be badly biased. In particular, the estimates of the elements of the covariance matrix can be seriously inflated, so masking the existence of outlying observations. Many methods for outlier detection therefore seek to divide the data into two parts, a larger “clean” part and the outliers. The clean data are then used for parameter estimation. The forward search provides subsets of increasingly large size which are designed to exclude the outliers until there are no clean data remaining outside the subset. At this point outliers start to be used in estimation, when test statistics and Mahalanobis distances may change appreciably.

Some methods for the detection of multiple outliers therefore use very robust methods to sort the data into a clean part and potential outliers.

An example is the use of the resampling algorithm of Rousseeuw and van Zomeren (1990) for the detection of multivariate outliers using the minimum volume ellipsoid. The algorithm selects random samples of size $v + 1$ from which the vector of means μ and the covariance matrix Σ are estimated. The process is repeated many times, perhaps one thousand, and the estimates chosen which give the smallest ellipsoid containing approximately half the data. The resulting parameter estimates are very robust. However Woodruff and Rocke (1994) show that such estimators, although very robust, have higher variance than those based on larger subsets. Such larger subsets are therefore more reliable when used in outlier detection procedures, provided they are outlier free. See also Hawkins and Olive (2002).

In the forward search for multivariate data we find such larger initial subsets of outlier free observations by starting from m_0 observations which are not outlying at a specified level in any univariate or bivariate boxplot. The properties of robust bivariate boxplots are described in the next section. We then increment this set starting subset by selecting observations that have small Mahalanobis distances and so are unlikely to be outliers. In some versions of the forward search, for example Hadi (1992) and Hadi and Simonoff (1993), the emphasis is on using the forward search to find a single set of parameter estimates and of outliers. These are determined by the point at which the algorithm stops, which may be either deterministic or data dependent. The emphasis in this book is very different: at each stage of the forward search we use information such as parameter estimates and plots of Mahalanobis distances to guide us to a suitable model.

At some stage in the forward search let the set of m observations used in fitting be $S_*^{(m)}$. The mean and estimated covariance matrix of this subset are $\hat{\mu}_m^*$ and $\hat{\Sigma}_{um}^*$. From these parameter estimates we can calculate a set of n squared Mahalanobis distances d_{im}^{*2} . Suppose that the subset $S_*^{(m)}$ is clear of outliers. There will then be $n - m$ observations not used in fitting that may contain outliers. We do not seek to identify these outliers by a formal test. Our interest is in the evolution, as m goes from m_0 to n , of quantities such as Mahalanobis distances, test statistics and other diagnostic quantities. We also look at the sequence of parameter estimates and related quantities such as the eigenvectors of $\hat{\Sigma}_{um}^*$. We monitor changes that occur, which can always be associated with the introduction of a particular group of observations, in practice usually one observation, into the subset of size m used for fitting. Interpretation of these changes is complemented by examination of changes in the forward plot of Mahalanobis distances.

Given that we have fitted the model to a subset of dimension $m \geq m_0$, the forward search moves to dimension $m + 1$ by selecting the $m + 1$ units with the smallest squared Mahalanobis distances, the units being chosen by ordering all squared distances d_{im}^{*2} , $i = 1, \dots, n$. In most moves from m to $m + 1$ just one new unit joins the subset. It may also happen that two or more units join $S_*^{(m)}$ as one or more leave. However our experience is

that such an event is unusual, only occurring when the search includes one unit that belongs to a cluster of outliers. At the next step the remaining outliers in the cluster seem less outlying and so several may be included at once. Of course, several other units then have to leave the subset.

Remark 1: The search starts with a robustified estimator of μ and Σ found by use of a bivariate boxplot. Let this estimator of μ be $\hat{\mu}_0^*$ and let the estimator at the end of the search be $\hat{\mu}_n^* = \hat{\mu}$. In the absence of outliers and systematic departures from the model

$$E(\hat{\mu}_0^*) = E(\hat{\mu}) = \mu;$$

that is, both parameter estimates are unbiased estimators of the same quantity. The same property holds for the sequence of estimates $\hat{\mu}_m^*$ produced in the forward search. Therefore, in the absence of outliers, we expect estimates of the mean to remain sensibly constant during the forward search. However, because of the way in which we select the observations for inclusion in the subset, those with smaller Mahalanobis distances will be selected first. As a result the estimate of Σ , unlike that of μ , will increase during the forward search. Therefore, unless outliers are present, the distances d_{im}^{*2} will trend steadily downwards during the search. The use of the scaled distances defined in (2.104) overcomes this tendency. A comparison of plots of scaled and unscaled distances is in Figure 2.5.

Remark 2: Now suppose there are k outliers. Starting from a clean subset, the forward procedure will include these towards the end of the search, usually in the last k steps. Until these outliers are included, we expect that the conditions of Remark 1 will hold and that plots of Mahalanobis distances will remain reasonably smooth until the outliers are incorporated in the subset used for fitting. The forward plot of scaled distances for the data on municipalities in Emilia-Romagna, Figure 3.24, is a dramatic example in which the pattern is initially stable, but changes appreciably at the end of the search when the two gross outliers enter.

Remark 3: If there are indications that the data should be transformed, it is important to remember that outliers in one transformed scale may not be outliers in another scale. If the data are analyzed using the wrong transformation, the k outliers may enter the search well before the end.

The search avoids the initial inclusion of outliers and provides a natural ordering of the data according to the specified null model. In our approach we use a robust starting point combined with unbiased estimators during the search that are multiples of the maximum likelihood estimators. The estimators are therefore fully efficient for the multivariate normal model. The zero breakdown point of these estimators is an advantage for the forward search. The introduction of atypical influential observations is signalled by sharp changes in the curves that monitor Mahalanobis distances and test statistics at every step. In this context, the robustness of the method does not derive from the choice of a particular estimator with a high breakdown point, but from the progressive inclusion of units into a subset which, in the

first steps, is outlier free. As a result of the forward search, the observations are ordered according to the specified null model and it becomes clear how many of them are compatible with a particular specification. Our approach enables us to analyze the inferential effect of the atypical units (“outliers”) on the results of statistical analyses.

Remark 4: The procedure is not sensitive to the method used to select an initial subset; even if outliers are included at the start they are often removed in the first few steps. For example, two forms of robust bivariate boxplot are described in the next section, either of which can be used to provide an initial subset. For speed of calculation we use the less robust. Although the first steps of the search may depend on which of the two methods is used to find the initial subset, the later stages are independent of it. What is important in the procedure is that the initial subset is either free of outliers or breaks the masking of outliers which are masked in the complete set of n observations. The removal of outliers is visible in some searches where there are sometimes numerous interchanges in the first few steps. Examples in which the search recovers from a start that contains outliers include Exercise 3.4 and Figure 7.20. An example for spatial data in which the search recovers from a start that is not very robust is given by Cerioli and Riani (1999).

2.13 Starting the Search

We now describe two methods for finding a “central” part of the data by looking at two dimensional projections. Both methods fit curves to bivariate scatter plots. These fitted curves can provide useful extra information when they are included in scatterplot matrices. In the first section we describe the babyfood data which are used here to illustrate the construction of the boxplots and, in §§4.2 and 4.7, to illustrate the transformation of multivariate data. The two methods of construction are described in §§2.13.2 and 2.13.3. Finally, in §2.13.4, we discuss the use of the intersection of these bivariate central parts in starting the forward search.

2.13.1 The Babyfood Data

Box and Draper (1987, p. 265) present part of a larger set of data on the storage of a babyfood. The data are in Table A.5. Unlike other data that we have so far seen, these include five explanatory variables. There are 27 readings on four responses which are the initial viscosity of the babyfood and its viscosity after three, six and nine months storage. The distribution of viscosity, a non-negative property, is highly skewed and we can expect that the data will require transformation. The ratios of the maximum to the minimum of each response in Table 4.1 reinforce this expectation. We

discuss the transformation of these data in some detail in Chapter 4. Here we only look at a scatterplot of the first two responses, both transformed and untransformed, to show the effect of data skewness on the construction of the two kinds of robust contour.

2.13.2 Robust Bivariate Boxplots from Peeling

A method for using the peeling of points from convex hulls to find a central part of the data is described by Zani, Riani, and Corbellini (1998). Once the central part of the data has been found, smooth contours are found by the use of B -splines (Micula 1998, de Boor 2002). The method is virtually non-parametric, in that almost no distributional assumptions are made in deriving the fitted B -spline. The method can be described in three steps.

Step 1 The Inner Region. The inner region is the two dimensional extension of the interquartile range of the univariate boxplot, where it is often called a “hinge”. In one dimension we take the length of the box between the first and third quartiles, which therefore contains 50% of the values. In two dimensions we look for a similar region centred on a robust estimator of location, containing a fixed percentage of the data. A natural, nonparametric way of finding a central region in two-dimensions is to use convex hull peeling. The most extreme group of observations in a multivariate sample can be thought of as those lying on the convex hull, with those on the convex hull of the remaining sample, the second most extreme group and so on. The output of the peeling is a series of nested convex polygons (hulls). We call the $(1 - \alpha)\%$ -hull the biggest hull containing not more than $(1 - \alpha)\%$ of the data (the points on the boundary belong to the hull). Usually, even if the outermost hulls assume very different shapes and are influenced by outliers, the 50%-hull seems to capture the underlying correlation of the two variables.

Since each convex hull contains several observations, the nominal 50%-hull found by peeling may contain less than 50% of the data, the effect being greater if the sample size is small. It also might not be smooth. To overcome this problem we fit a B -spline curve to the 50%-hull found by peeling. The inner region is therefore formed by those units which lie inside or on the boundary of the B -spline curve superimposed on the 50%-hull.

As an illustration of our method we use the scatterplot of the logged values of the first two variables in the babyfood data. There are 27 observations. Panel (a) of Figure 2.1 shows the outermost hull, which passes through seven points. Panel (b) shows this hull together with the second hull, also passing through seven points, so that 13 remain. The third hull, of five points in Panel (c), is the 50% hull, since it is the largest containing not more than 50% of the data. Inside it are eight data points. In Panel (d) a B -spline curve is fitted to the 50% hull. This inner region seems free from outliers and is robust, while keeping the correlation in the data and allowing for different spreads in the various directions. It is worth noting

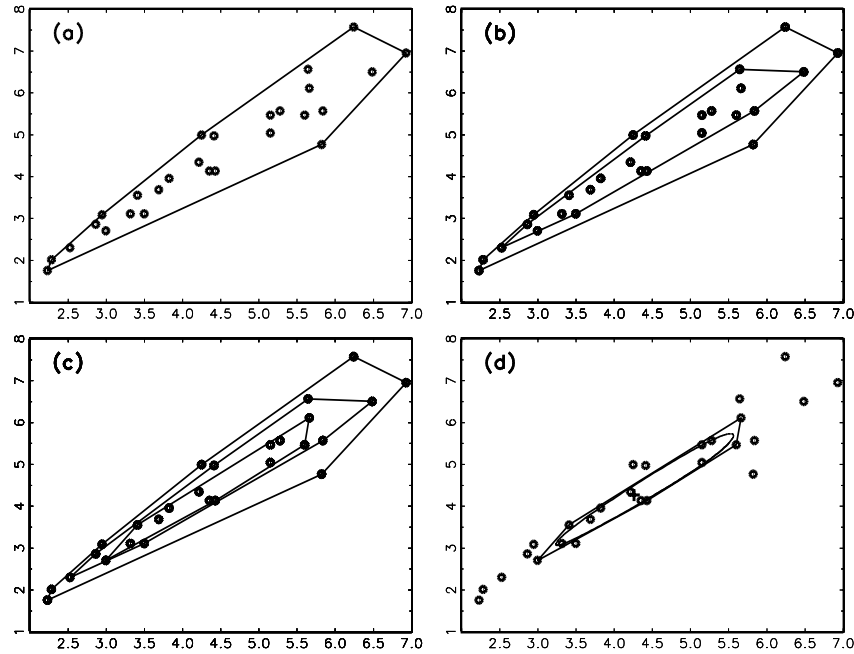


FIGURE 2.1. Logged babyfood data, y_1 and y_2 : the first three convex hulls containing respectively 7, 7 and 5 points. Panel (d) shows the B -spline fitted to the 50% hull of five points and the robust centre, marked $+$ almost coincident with an observation

that the fitted spline contains only seven data points, since one observation, with coordinates 5.15 and 5.47, lies inside the 50% hull, but outside the spline.

Step 2 The Robust Centroid. We find a robust bivariate centroid using the componentwise arithmetic means of the observations inside the inner region defined by the fitted spline. In this way we exploit both the efficiency properties of the arithmetic mean and the natural trimming offered by the hulls. This mean of the values of logged y_1 and logged y_2 is marked with a cross in Panel (d) of Figure 2.1. This cross gives the appearance of being near the centre of the nearly elliptical spline.

A useful requirement of estimators of location is affine invariance (for example Woodruff and Rocke 1994) ensuring that different rescalings of the individual variables leave the estimator of location unchanged. If we require such a property of our estimator we need to take the mean of the observations over the convex hull, rather than over the fitted B-spline. References to other ways of finding robust bivariate centres are given at the end of the chapter.

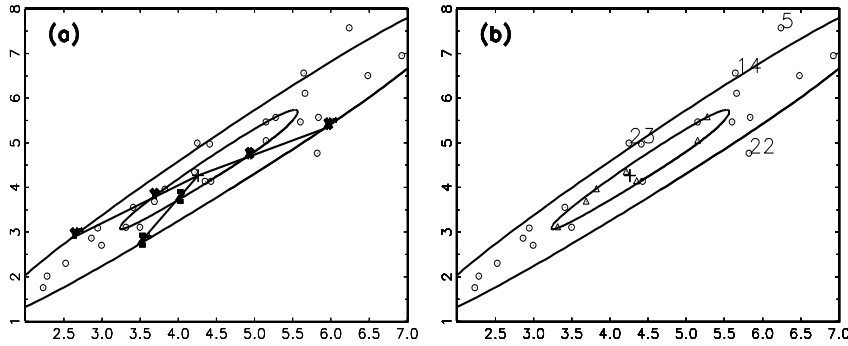


FIGURE 2.2. Logged babyfood data, y_1 and y_2 : scaling the convex hull. The resulting 99% hull indicates four outliers

Step 3 The Outer Region. Once we have found a robust bivariate centre and a curve asymptotically containing half the data (a bivariate “hinge”) we require a method for constructing contours at other levels. To find a small subset for starting the search we may need contours with a nominal content of less than 50%. But, if interest is in a contour which discriminates between “good” and “bad” observations, a much higher nominal content will be needed. If we were using Mahalanobis distances to measure the remoteness of the observations, these contours would be ellipses which would be found analytically. However, with the metric provided by the bivariate hinge, we have to proceed numerically. We find the contours by scaling the hinge, following the procedure suggested by Goldberg and Iglewicz (1992) as modified by Zani et al. (1998).

The left-hand panel of Figure 2.2 shows the method of scaling. Let O be the centre of the data found in Step 2 and let X , Y and Z be three points on the 50% contour. Then, if X' , Y' and Z' are three points on rays from O on the scaled contour, we require that the ratios

$$\frac{OX'}{OX} = \frac{OY'}{OY} = \frac{OZ'}{OZ} = c, \quad (2.101)$$

say. For outlier detection c will be appreciably greater than one.

To find suitable values of c , Goldberg and Iglewicz (1992) use an approximate F distribution for the Mahalanobis distance

$$d_i^2 \sim \{2(n-1)/(n-2)\}F_{2,n-2}, \quad (2.102)$$

which is close to the distribution of the deletion Mahalanobis distance (2.49). The theoretical 50% contour for the babyfood data in Figure 2.2 corresponds to an F value, on 2 and 25 degrees of freedom, of 0.7127. The value for the 99% contour is 5.568. The ratio of these two is 7.813. But

we are concerned with distance, not squared distances, so the value of c in (2.101) is $\sqrt{7.813} = 2.795$.

In a conservative approach to outlier detection we might seek to declare any ambiguous points as outliers, thereby obtaining a central part of the data which has a reduced probability of being contaminated. A possibility here is to replace the F distribution with the asymptotic chi-squared distribution of the distances. In this case the value of 2.795 becomes 2.58. Zani et al. (1998) use this approximation to the distribution of squared Mahalanobis distances combined with simulation to allow for the effect of peeling on the content of the hinge. The value of 2.58 increases slightly to 2.68. This final value is a compromise between the 2.795 based on the F distribution and the 2.58 from the chi-squared approximation. (In Table 2 of Zani et al. (1998) the values are of $c - 1$, the extension of the ray beyond the 50% fitted spline). This approximate 99% contour is plotted in both panels of Figure 2.2. In the right-hand panel some possible outliers in this two-dimensional projection of the data are identified.

Use of the exact beta distribution (2.52) for scaling is recommended if it really is desired to use the bivariate boxplot for bivariate outlier detection. However, we use the forward search for outlier detection for any dimension v , at the same time obtaining information on the inferential importance of each observation. Our interest in the boxplots is to select an initial subset, when several values of c may be tried, until a satisfactory value is found for m_0 .

The convex hulls in Figure 2.1 and the nearly elliptical contours in Figure 2.2 suggest that the logged data are approximately normally distributed. It is interesting to see what happens when we repeat the procedure using untransformed data.

Figure 2.3 shows the four convex hulls fitted to the data in the peeling process to find the 50% hull. These hulls are quite different in shape from the hulls for the transformed data, several being quadrilateral. Since they mostly only contain four observations, four hulls have to be peeled to obtain the 50% hull, rather than three for the transformed data.

Figure 2.4 shows the fitted B -spline and the nominal 99% contour found by scaling up. Because the data are concentrated in the lower-left hand corner of the figure, the 50% curve is relatively small. As a result several observations lie outside the 99% curve found by scaling up. The skew distribution of the data leads to the detection of many apparent outliers.

2.13.3 Bivariate Boxplots from Ellipses

The bivariate boxplots calculated from B-splines provide a useful tool for a preliminary examination of the data. They are however over elaborate as a means of finding a central part of the data which can serve as a starting point for the forward search. In this section we present a computationally simpler method in which ellipses with a robust centroid are fitted to the

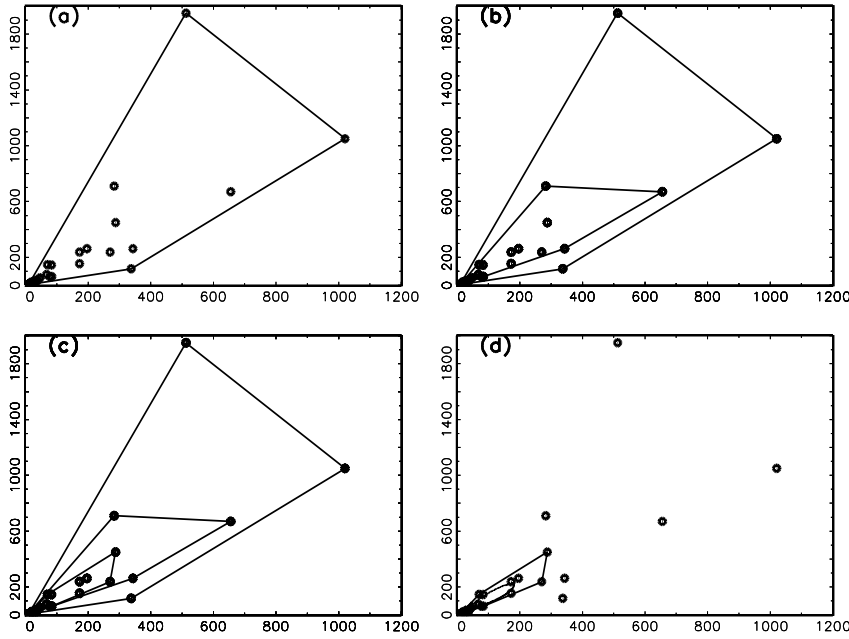


FIGURE 2.3. Untransformed babyfood data, y_1 and y_2 : four convex hulls have to be peeled to obtain the 50% hull as opposed to three for the transformed data in Figure 2.1

data. Our description follows Riani and Zani (1997) who use a version of the “quelplot” of Goldberg and Iglewicz (1992).

The robust centroid of the ellipse is found as the componentwise median of the two variables in the scatterplot. Let this be $\tilde{\mu}$. The shape of the contours is based on a covariance matrix in which the univariate medians are used, but which is otherwise calculated in the usual way. That is, the mean in (2.10) is replaced by $\tilde{\mu}$ to give a 2×2 matrix with elements proportional to

$$S_{jk}(\tilde{\mu}) = \sum_{i=1}^n (y_{ij} - \tilde{\mu}_j)(y_{ik} - \tilde{\mu}_k).$$

The combination of centroid and covariance estimate gives Mahalanobis distances for each observation and a family of ellipses which need to be scaled. The 50% ellipse is that which passes through the point with the median Mahalanobis distance, and so contains exactly 50% of the data. As a matter of minor detail, we use the F distribution for scaling this ellipse. As was stated above, the theoretical value of c for this 50% contour for the babyfood data is 0.7127. Contours for other levels are then found by scaling this ellipse.

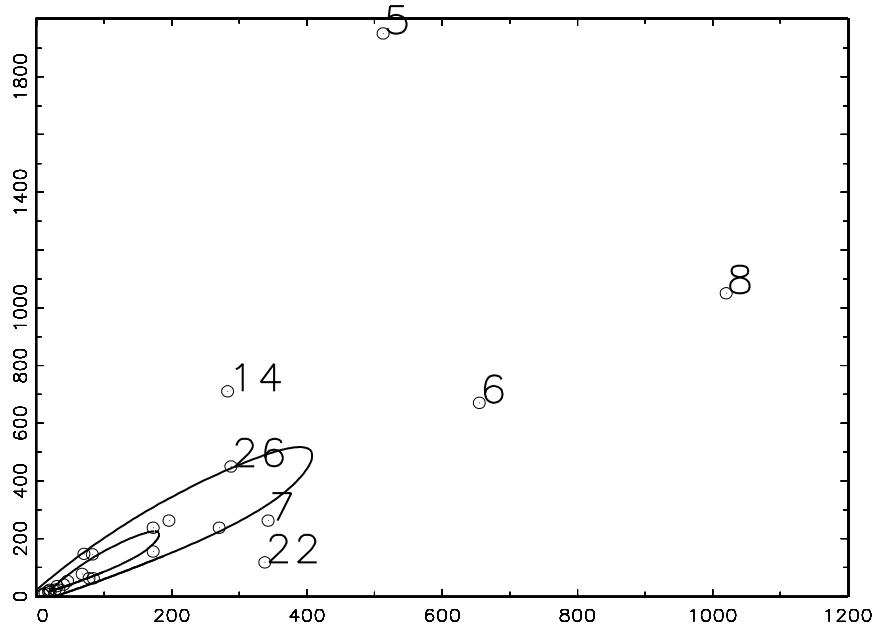


FIGURE 2.4. Untransformed babyfood data, y_1 and y_2 : the scaled 99% convex hull indicates seven outliers

A scatterplot matrix of these ellipses for the original babyfood data is in Figure 4.6 and for the log-transformed data in Figure 4.5. The interpretation of the plots is similar to that for Figures 2.4 and 2.2, with the untransformed data exhibiting many more outliers.

The method of constructing boxplots based on peeling was described above in Section 2.13.2 as virtually non-parametric. It is not completely non-parametric since the F or χ^2 distribution used to find the scaling c is based on the assumption of bivariate normality. The method based on ellipses in this section is hardly non-parametric at all; apart from the use of the median as a measure of location, the theory is based entirely on the normal distribution. Even so, this boxplot is both a useful tool for the examination of data and a method for finding an initial subset, even in data with many outliers.

2.13.4 The Initial Subset

We find an initial subset of m_0 observations from the intersection of units inside a contour of specified content, where we have to adjust the content to yield, at least roughly, the required number of observations. We also eliminate any univariate outliers. The magnitude of the contour depends on the parameter θ giving the scaling

$$d_i^2(\theta) = \{2(n-1)/(n-2)\}\theta. \quad (2.103)$$

The relationship between θ and the scaling c follows from the approximate F distribution for Mahalanobis distances defined in (2.102). For example, a value of $\theta = 1$ corresponds to the 61.8% point of the F distribution on 2 and 25 degrees of freedom. Usually we have to use smaller values to obtain a sufficiently small value for m_0 . An example showing the variation of m_0 with θ is in §4.2.

The value of m_0 is not critical. It should be small enough so that the initial subset contains no masked outliers, but large enough that the initial stages of the search are fairly stable, apart from any initial interchanges. For examples in which we are fitting multivariate models without any structure in the means, a value around $2v$ is often suitable. The procedure is generally robust to the choice of the value of m_0 and allows us to start with a somewhat larger subset if the percentage of contamination permits. Since the method does not involve complicated iterative procedures, there is no computational burden in finding the starting point. As the size of the initial subset can easily be increased or decreased by changing the value of θ , we usually try several values and check whether the last third or so of each search from the various starting points is the same. As we have seen, it is often towards the end of the search that we obtain information about unsuspected structure and outliers for observations basically from a single normal population. However, if there are several populations, as in the Swiss bank note data, the earlier parts of the search are also informative. For example, Figure 3.30 will show the effect of two populations around $m = 100$ when $n = 200$. Larger initial subsets than $2v$ are required for models in which there are more than v parameters to be estimated, for example when we are determining transformations.

We find the initial subset from the intersection in all $v(v-1)/2$ bivariate scatterplots and v univariate boxplots of units within the contour specified by θ . This subset will exclude any observations which are outlying in one or two dimensions. However it will not exclude observations that are not outlying in one or two dimensions but are outlying in three or more. Although it is not difficult to construct such observations, they seem to be rare in practice. Any problem they might cause can be simply reduced by decreasing the value of θ . However, in general, even if one or two have been included in the initial subset, they are detected in the early stages of the search, their large Mahalanobis distance causing them to leave the subset. We do indeed sometimes observe several interchanges in the first two or three steps of the search. All that we require is that the construction of the initial subset reveals outliers which are masked in the whole data set. They do not need to be excluded from the initial subset, merely to be unmasked in it.

2.14 Monitoring the Search

At each step in the forward search we calculate all squared distances d_{im}^{*2} , $i = 1, \dots, n$ for $m_0 \leq m \leq n$. Many of our most informative plots are based on the Mahalanobis distances d_{im}^* , rather than on the squared distances. We plot these as m increase from m_0 to n . Such plots are called “forward” plots.

Mahalanobis Distances. We plot all n distances d_{im}^* for each value of m . This plot is informative about the behaviour of individual units, the distances for which can be followed throughout the search.

Scaled Mahalanobis Distances. At the beginning of the search the few central units may give a very small estimate of the covariance matrix. Consequently, units not in the subset may have very large distances, which decrease as the search proceeds. The result is that the eye tends to focus on the early part of the search, whereas important information is usually in the last third or so, where the outliers, if any, enter and cause changes in inferences.

Virtually constant residual plots in regression were obtained by Atkinson and Riani (2000), for example Figure 1.4, by scaling the least squares residuals e_{im} at subset size m by s_n , the error mean square estimate of σ^2 at the end of the search. These scaled residuals can be written as

$$\frac{e_{im}}{s_n} = \frac{e_{im}}{s_m} \frac{s_m}{s_n}.$$

The Mahalanobis distances

$$d_{im}^* = (e_{im}^T \hat{\Sigma}_{um}^{-1} e_{im})^{0.5}$$

are scaled by the square root of the estimated covariance matrix. If we had independent observations with constant variance σ^2 , Σ would be a diagonal matrix and

$$|\Sigma| = \sigma^{2v}.$$

So the generalization of the scaled residuals is the **scaled** distance

$$d_{im}^* \times \left(\frac{|\hat{\Sigma}_{um}|}{|\hat{\Sigma}_{un}|} \right)^{1/2v}, \quad (2.104)$$

where we rename $\hat{\Sigma}_u$ as $\hat{\Sigma}_{un}$ to stress that the estimator is calculated at the end of the search from all n observations.

As Figure 2.5 shows, for the Swiss bank note data, this rescaling increases emphasis on the right-hand end of the forward plot. The upper panel is the forward plot of the scaled distances: in the lower panel the distances are not scaled. In this example the plot of scaled distances seems superior in all parts of the search. Although quite stable, these scaled distances

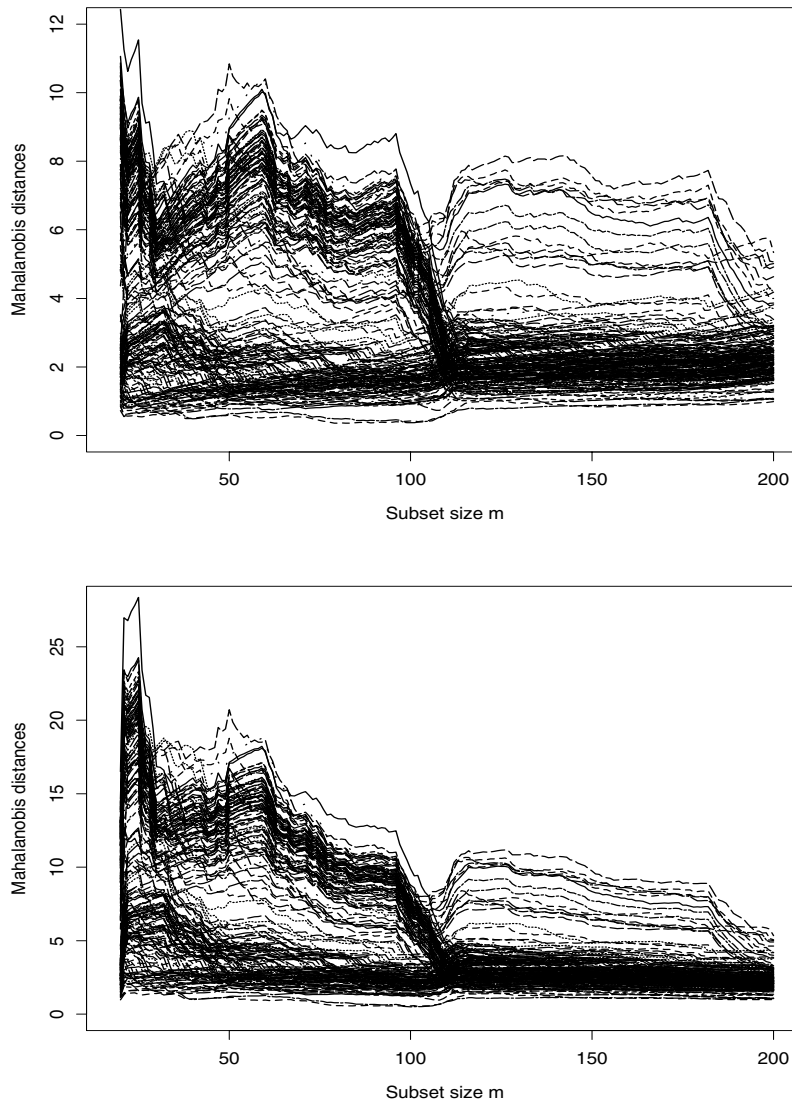


FIGURE 2.5. Swiss bank notes, starting with the first 20 observations on genuine notes: forward plots of Mahalanobis distances - upper panel scaled and, lower panel, unscaled

are somewhat less stable than scaled residuals in regression. This is not surprising since the regression structure means that the residuals fluctuate much less than those here from a structureless sample. We discuss the upper panel in some detail in Chapter 3 as Figure 3.30.

Ordered Mahalanobis Distances. As we saw, for example in Figure 1.17, we find it informative to plot particular ordered Mahalanobis distances. We discuss three useful plots. But we begin with some results about ordered distances and the forward search.

The progress of the search depends on ordering the squared Mahalanobis distances. From the subset $S_*^{(m-1)}$ we calculate the n squared distances $d_{i,m-1}^{*2}$, $i = 1, \dots, n$, and order them to obtain the n distances $d_{[i],m-1}^{*2}$. The new subset $S_*^{(m)}$ at step m then consists of the units corresponding to the m ordered distances $d_{[1],m-1}^{*2}$ to $d_{[m],m-1}^{*2}$. The units corresponding to the ordered distances are then divisible into two sets. Let $U_{[i],m}$ denote the unit with the i th ordered Mahalanobis distance at step m . Then

$$U_{[1],m-1}, \dots, U_{[m],m-1} \in S_*^{(m)}$$

and

$$U_{[m+1],m-1}, \dots, U_{[n],m-1} \notin S_*^{(m)}.$$

To move to the new subset $S_*^{(m+1)}$ we form the n distances d_{im}^{*2} and order them. It is not certain that all the units which were in $S_*^{(m)}$ will be in $S_*^{(m+1)}$. There are three cases which need to be distinguished:

1. **Normal Progression.** If

$$d_{[m]m}^* = \max d_{i,m}^* \quad i \in S_*^{(m)},$$

the next unit to join will be $U_{[m+1]m}$ which $\notin S_*^{(m)}$ with distance $d_{[m+1]m}^*$;

2. **Inversion.** Now suppose that

$$d_{[m+1]m}^* = \max d_{i,m}^* \quad i \in S_*^{(m)}.$$

Then $U_{[m+1]m}$ will remain in the subset. But there must be a unit, say $U_{NEW} \notin S_*^{(m)}$ for which

$$d_{NEW,m}^* \leq d_{[m]m}^*.$$

This unit will join the subset while $U_{[m+1]m}$ will remain in the subset. The minimum distance among units not in the subset will obviously be $d_{NEW,m}^* \leq d_{[m]m}^*$;

3. **Interchange.** An interchange occurs when two or more new units enter the subset, when one or more must leave. Instead of the one new unit U_{NEW} when inversion occurs we have a set S_{NEW} , containing at least two members, such that

$$i \in S_{NEW} \text{ if } d_{i,m}^* \leq d_{[m+1]m}^* \cap i \notin S_*^{(m)}.$$

Then the minimum distance among units not in the subset can be written as

$$d_{NEW,m}^* = \min d_{i,m}^* \quad i \in S_{NEW}.$$

To obtain an upper bound for this distance let the number of units in S_{NEW} be $n_{NEW}(\geq 2)$. Then

$$d_{NEW,m}^* \leq d_{[m+2-n_{NEW}]m}^*.$$

The smallest distance among units not in the subset is monitored up to step $n - 1$.

Largest Distance among Units in the Subset. Here we monitor

$$\max d_{i,m}^* \quad i \in S_*^{(m)},$$

the largest distance among units in the subset. For normal progression this will be $d_{[m]m}^*$. As we have seen above, for inversion the distance is $d_{[m+1]m}^*$; it will be larger than this when an interchange occurs. The forward plot of this largest distance will show a peak when the first outlier is included. The peak is therefore one step later than it is for the preceding plot of the smallest distance not in the subset. The largest distance is monitored up to step n .

In general, when there is one outlier, the size of the peak in the plot of the largest distance amongst units in the subset is smaller than that in the plot of the smallest distance among units not in the subset. This arises because $d_{i,m}^{*2}$ for units not belonging to the subset has an unbounded distribution, whereas that for the maximum over units in the subset is the maximum of m scaled beta distributions.

“Gap” Plot. The forward plots of the minimum and maximum distances trend upwards, which can sometimes obscure interpretation. In the gap plot we look at the difference of the two preceding quantities, that is

$$\min_{i \notin S_*^{(m)}} d_{i,m}^* - \max_{i \in S_*^{(m)}} d_{i,m}^*. \quad (2.105)$$

For normal progression this difference is

$$d_{[m+1]m}^* - d_{[m]m}^*, \quad (2.106)$$

where both distances are calculated using the same subset of size m . If there is an inversion, an upper bound on the value is

$$d_{[m]m}^* - d_{[m+1]m}^*,$$

the negative of the value for normal progression. The bound is even more negative if there is an interchange, the magnitude depending on the value of n_{NEW} . We plot both the true difference (2.105), which can be negative and the difference in order statistics (2.106), which is always positive.

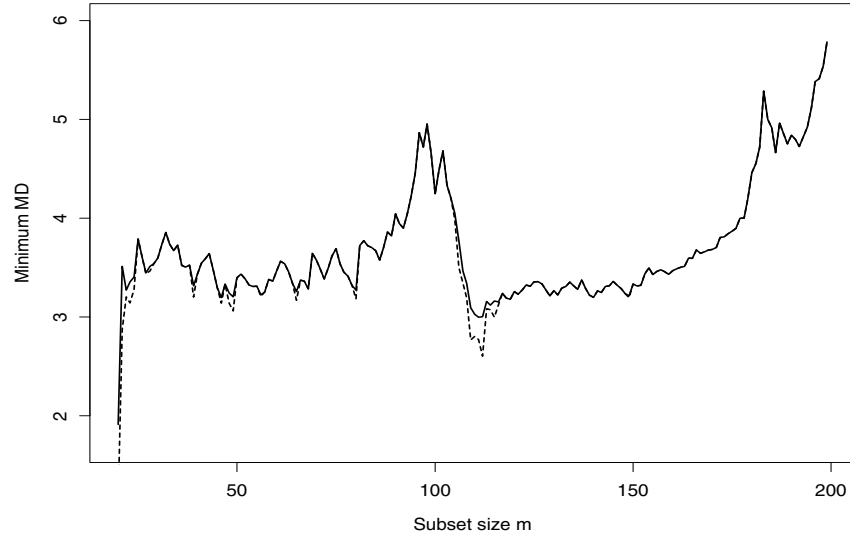


FIGURE 2.6. Swiss bank notes starting with the first 20 observations on genuine bank notes: forward plots of, dotted line, minimum distances of units not in the subset and, solid line, of the ordered distance $d_{[m+1]m}^*$. There is an indication of appreciable interchange around $m = 100$

As an example of these plots, Figure 2.6 shows the forward plot of the minimum distance amongst units not in the subset for the Swiss bank note data, which we have seen in Figure 1.17, together with the forward plot of $d_{[m+1]m}^*$. These two are the same for much of the search, but different at the beginning and around $m = 100$. Both regions in which some interchanges are occurring. The earlier one is associated with instability at the beginning of the search. The later reflects the interchanges which occur as units from the group of forgeries start to enter the subset, with the appreciable change in covariance matrix and distances that we saw in Figures 1.16 and 2.5.

Covariance Matrix. The estimate of the covariance matrix Σ does not remain constant during the forward search as observations are sequentially selected that have small Mahalanobis distances. To see how the variance is increasing we can look at forward plots of the ratios

$$\text{tr } \hat{\Sigma}_{um} / \text{tr } \hat{\Sigma}_{un} \quad \text{or} \quad |\hat{\Sigma}_{um}| / |\hat{\Sigma}_{un}|. \quad (2.107)$$

In the absence of outliers these ratios increase smoothly. Non-monotonic increase of the curve is evidence of the existence of masked outliers which are preventing the unambiguous ordering of the data by its closeness to the model. Large increases at the end of the search are more easily interpreted as being due to isolated outliers.

In addition to the magnitude of the covariance matrices, we also look at the evolution of their structure, through forward plots of the eigenvalues and also of the eigenvectors for two dimensional subsets of the data. The proportion of the total variation in the data explained by each eigenvalue is an important property for principal components analysis in Chapter 5

Parameter Estimates. The means of most of the sets of multivariate data analysed in this book are without structure. However, if there is regression structure, it is interesting to look at forward plots of the parameter estimates of the linear models for the various responses. Forward plots of the estimated transformation parameters are crucial in our strategy for determining the correct transformation of data outlined in §4.6, whatever the structure of the means.

2.15 The Forward Search for Regression Data

If, as in Section 2.8, there is a regression structure in the means of the multivariate observations, we need to allow for this in finding a central set of observations to form the starting point for the forward search. We describe a method for the case when the regressors for all responses are the same.

2.15.1 Univariate Regression

An appealing feature of the multivariate regression model (2.57) was that estimation was by independent least squares on each response. We proceed by using a series of v forward searches, one on each of the responses, to order the univariate observations by their closeness to the regression model. We then take the intersection of the units in these ordered sets to give us approximately the required m_0 observations in the initial subset. To start, we sketch the forward search for univariate regression models. Many of the principles are the same as those in the search for multivariate data. The details are in Chapter 2 of Atkinson and Riani (2000).

For the univariate linear regression model $E(Y) = X\beta$, with X of rank p , let b be any estimate of β . With n observations the residuals from this estimate are $e_i(b) = y_i - x_i^T b$, ($i = 1, \dots, n$). The least median of squares estimate $\hat{\beta}_p^*$ is the value of b minimizing the median of the squared residuals $e_i^2(b)$. Thus $\hat{\beta}_p^*$ minimizes the scale estimate

$$\sigma^2(b) = e_{[\text{med}]}^2(b), \quad (2.108)$$

where $e_{[k]}^2(b)$ is the k th ordered squared residual. In order to allow for estimation of the parameters of the linear model the median is taken as

$$\text{med} = [(n + p + 1)/2], \quad (2.109)$$

the integer part of $(n + p + 1)/2$.

The parameter estimate satisfying (2.108) has, asymptotically, a breakdown point of 50%. Thus, for large n , almost half the data can be outliers, or come from some other model and LMS will still provide an unbiased estimate of the regression line. This is the maximum breakdown that can be tolerated. For a higher proportion of outliers there is no longer a model that fits the majority of the data.

The definition of $\hat{\beta}_p^*$ in (2.108) gives no indication of how to find such a parameter estimate. Since the surface to be minimized has many local minima, approximate methods are used. Rousseeuw (1984) finds an approximation to $\hat{\beta}_p^*$ by searching only over elemental sets, that is, subsets of p observations, taken at random. We follow this procedure. Depending on the dimension of the problem we find the starting point for the forward search either by sampling 1,000 subsets or by exhaustively evaluating all subsets. We take as our initial subset for each response that yielding the minimum value in (2.108), so obtaining an outlier free start for our forward search.

For regression models we have v searches, one for each response. In a generalisation of the previous notation, suppose at some stage in the forward search the set of m observations used in fitting response j is $S_{*j}^{(m)}$. The parameters of the linear model are estimated by least squares yielding the parameter estimates $\hat{\beta}_{jm}^*$. From these parameter estimates we can calculate a set of n residuals e_{ijm}^* . The forward search for the j th response moves to dimension $m+1$ by selecting the $m+1$ units with the smallest squared least squares residuals, the units being chosen by ordering all squared residuals e_{ijm}^{*2} , $i = 1, \dots, n$. As with the search for multivariate data, most moves from m to $m+1$ introduce just one new unit to the subset although it may happen that two or more units join $S_{*j}^{(m)}$ as one or more leave.

The procedure is again not sensitive to the method used to select an initial subset, even if unmasked outliers are included at the start. For example, the least median of squares criterion (2.108) for regression can be replaced by that of least trimmed squares (LTS). This criterion provides estimators with better properties than LMS estimators. They are found by minimizing the sum of the smallest h squared residuals

$$S_h(b) = \sum_{i=1}^h e_{[i]}^2(b), \quad (2.110)$$

for some h with $[(n + p + 1)/2] \leq h < n$. The rate of convergence of LTS estimates is $n^{-1/2}$ as opposed to $n^{-1/3}$ for LMS. But, for the moderate sized datasets of the size considered in Atkinson and Riani (2000), the largest having 200 observations, there seems to be little difference in the abilities of the two methods to detect outliers and so to provide a clean starting point for the forward search.

2.15.2 Multivariate Regression

To adapt the searches for univariate regression to multivariate regression we need to find a starting point and to describe how the search moves forward.

As a result of the univariate forward search on response j we have, for each m , a subset of observations $S_{*j}^{(m)}$ which are used for fitting the j th model. For any particular value of m , say k , the subsets $S_{*j}^{(k)}$, $j = 1, \dots, v$, will contain some observations in common, but they will not in general be identical, except when $k = n$. To find an initial subset of size m_0 we consider the observations in common in these subsets. Let there be $m(k)$ such observations, that is

$$m(k) = S_{*1}^{(k)} \cap \dots \cap S_{*v}^{(k)}. \quad (2.111)$$

We start with $k = m_0$ and increase k until the first time when there are at least m_0 common units in the intersection. These units form the initial subset.

The v forward searches order the observations by their closeness to the fitted univariate models. If there were no interchanges during the search, we would have a single list of the order in which observations on each response enter the subset and $S_{*j}^{(m)}$ would consist of the first m units on this list. However, when there is an interchange, some units leave $S_{*j}^{(m)}$, and it is not true that

$$S_{*j}^{(m)} \subset S_{*j}^{(m+1)}.$$

The lists of units used in (2.111) to calculate $m(k)$ therefore need to include information on units which leave the subsets as the search progresses in addition to those which enter.

Once we have an initial subset of m_0 units, the search progresses much as it did in the absence of regression in §2.12. Given $S_*^{(m)}$ individual regressions are fitted to the v responses. From the parameter estimates we can calculate the $n \times v$ matrix of residuals with elements e_{ijm}^* and so the set of n squared Mahalanobis distances d_{im}^{*2} . The search moves to dimension $m+1$ by selecting the $m+1$ units with the smallest squared Mahalanobis distances, the units being chosen by ordering the squared distances d_{im}^{*2} , $i = 1, \dots, n$.

2.16 Further Reading

There are numerous books on multivariate analysis, many of which provide important background reading for the multivariate normal distribution and associated inferential and data analysis procedures on which our book is

based. Since the analysis of multivariate data requires numerical computing, there is a time trend in the books towards data analysis and also the plotting of data. There is also a trend away from mathematics, which may reflect an increasing cultural impatience with mathematical manipulation for its own sake. Whatever the reasons for the latter trend, our book seems to us extreme in following both these tendencies.

The theory is presented by Morrison (1967). Anderson (1984) gives the matrix algebra in great detail. Muirhead (1982) focuses on distribution theory, without any mention of data. We have already mentioned the mathematically less advanced books of Flury and Riedwyl (1988), Flury (1997) and Johnson and Wichern (1997). A useful reference for mathematical results is Mardia, Kent, and Bibby (1979) as is Seber (1984). Krzanowski (2000) describes both applications and theory. The two parts of Krzanowski and Marriott (1994) and Krzanowski and Marriott (1995) cover respectively distributions, ordination and inference and classification, covariance structures and repeated measurements.

Throughout we deal with data which presumably arise from several multivariate normal distributions, perhaps, of course, with outliers. There may also be explanatory variables, which may be discrete or continuous. However we do not consider data in which some of the multivariate responses have discrete distributions. Much of the literature apparently about discrete multivariate data analysis is concerned with the analysis of contingency tables in which there is a Poisson response and categorical explanatory variables. An example is Agresti (2002). The forward search for Poisson generalized linear models described in Atkinson and Riani (2000, §6.10 - §6.12) extends to such data. Chapter 3 of Fahrmeir and Tutz (2001) describes methods for multicategorical responses, again based on generalized linear models.

To conclude this section, we provide some references to detailed points. An expression for the effect of deletion of observation k on the i th Mahalanobis distance, which is a generalization of (2.46) and (2.47), is given by Riani and Zani (1997). The result of Wilks (1963) used in §2.7 as the start of an alternative derivation of the distribution of the squared Mahalanobis distance, is described by Barnett and Lewis (1994, p. 288). It is used by Penny (1996a) to derive the distribution of an outlier test. Further discussion of her results is in Fung (1996b) and Penny (1996b). Campbell (1985) presents a succinct summary of these deletion and distributional results. Grubbs (1950) gives the univariate result on Beta distributions for a simple sample and considers the distribution of the order statistics.

The methods of starting the search described in §§2.13.2 and 2.13.3 employ only two of the many methods that have been investigated for finding bivariate centres. Small (1990) provides a survey on multidimensional medians. The lengthy review of Liu, Parelius, and Singh (1999) and associated discussion provides many references on finding robust centres using the idea

of data “depth”. A more recent reference is Van Aelst, Rousseeuw, Hubert, and Struyf (2002) who apply robust regression to this problem.

2.17 Exercises

In all exercises y is a v -variate random variable with $E(y) = \mu$ and $\text{cov}(y) = \Sigma$. Unless otherwise stated, the normality of y may also be assumed.

Exercise 2.1 Show, without assuming normality, that $E\{(y - \mu)^T \Sigma^{-1}(y - \mu)\} = v$.

Exercise 2.2 Show that the variance of the residual e_i (2.1) is as given in §2.1.2. What distributional assumptions did you make?

Exercise 2.3 The distribution of the sum of squares and products matrix $S(\hat{\mu})$ (2.10) depends on the projection matrix C (2.13). Show that C is symmetric and idempotent and prove the result claimed at the end of §2.2.2.

Exercise 2.4 When μ is known, the squared Mahalanobis distance $d_i^2(\mu, \hat{\Sigma}_\nu)$ is defined in (2.26). Derive the distribution of this quantity when $v = 1$ and $\hat{\Sigma}_\nu = s^2$.

Exercise 2.5 Find the distribution of the scaled squared residual about the mean, which is called d_i^2 in (2.4).

Exercise 2.6 An extension of Exercise 2.5. Let $e_i = y_i - x_i^T \hat{\beta}$ be the residual from univariate regression as in §2.9. Find the distribution of the scaled squared residual e_i^2/s^2 . You may find equation (2.95) helpful. Relate your answer to that you found for Exercise 2.5.

Exercise 2.7 Show that:

1) the loglikelihood of the n observations is

$$L(\mu, \Sigma; y) = -(n/2) \log |2\pi\Sigma| - \sum_{i=1}^n (y_i - \mu)^T \Sigma^{-1} (y_i - \mu)/2;$$

2) the maximum likelihood estimators of μ and of the covariance matrix Σ are given by:

$$\hat{\mu} = \bar{y} = \left(\sum_{i=1}^n y_{i1}/n, \dots, \sum_{i=1}^n y_{iv}/n \right)^T$$

and

$$\hat{\Sigma} = S(\hat{\mu})/n;$$

3) the maximised multivariate normal loglikelihood is given by (equation 2.19)

$$-(n/2) \log |2\pi\hat{\Sigma}| - nv/2.$$

Exercise 2.8 Find the form of the matrix D when the test of equality of the v means is formulated as $D\mu = c$ (equation 2.18). What is the row rank of D ?

Exercise 2.9 In order to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ the usual test statistic is

$$T^2 = n(\bar{y} - \mu_0)^T \hat{\Sigma}_u^{-1} (\bar{y} - \mu_0).$$

The quantity T^2 has Hotelling's T^2 distribution with dimension v and degrees of freedom $n - 1$. We reject H_0 if $T^2 \geq T_{\alpha, v, n-1}^2$ and accept H_0 otherwise. Show the connection between T^2 and the corresponding likelihood ratio test (equation 2.20).

Exercise 2.10 Suppose there are g groups of v dimensional normal observations. Find the likelihood ratio test of the equality of the covariance matrices of the g groups. When is this important?

Exercise 2.11 Verify the (Bartlett)-Sherman-Morrison-Woodbury formula (equation 2.40). Show that the inverse of $A - UV^T$ is given by $\{A^{-1} + A^{-1}U(I_m - V^T A^{-1}U)^{-1}V^T A^{-1}\}$, when the dimensions are: A is $p \times p$, with U and V $p \times m$. Apply this formula for the deletion of m rows of X .

Exercise 2.12 Find $\sum_{i=1}^m d_{im}^{*2}$, where the distances are calculated for the subset $S_*^{(m)}$. Give bounds for $\sum_{i=1}^n d_{im}^{*2}$ when there is no inversion or interchange in going from $S_*^{(m)}$ to $S_*^{(m+1)}$. Suggest a data configuration for which your lower bound is achieved. What happens to your lower bound when there is an inversion and when there is an interchange?

Exercise 2.13 The hat matrix H is defined in equation (2.63). Prove it is (a) symmetric and (b) idempotent. (c) Find $\text{tr}(H)$. For what model is C (equation 2.13) the hat matrix?

Exercise 2.14 The explanatory variables in the first 16 rows of the baby-food data have coded levels of 1 and -1 . The experimental design is a 2^{5-1} fractional factorial. If x_5 is omitted, the design is a full 2^4 factorial. Suppose a first-order model, including a constant term, is fitted to the results of a full 2^k factorial experiment. Calculate the values of the leverage measures h_i (equation 2.64) and confirm that the value of the sum of the leverage measures $\sum_{i=1}^n h_i$ agrees with the result you found in Exercise 2.13.

How does your answer change when some interaction terms of the form $x_i x_j$ are included in the model?

Exercise 2.15 Figure 1.16 in Chapter 1 is a forward plot, for the Swiss bank note data, of the elements of the estimated covariance matrix for a search starting from 20 observations on genuine notes. The left panel of Figure 2.7 is a forward plot of the determinant of this matrix. The right panel shows the trace. Relate these two figures to one another and give reasons for the difference between the two panels of Figure 2.7. What different features of the data are revealed by the two panels?

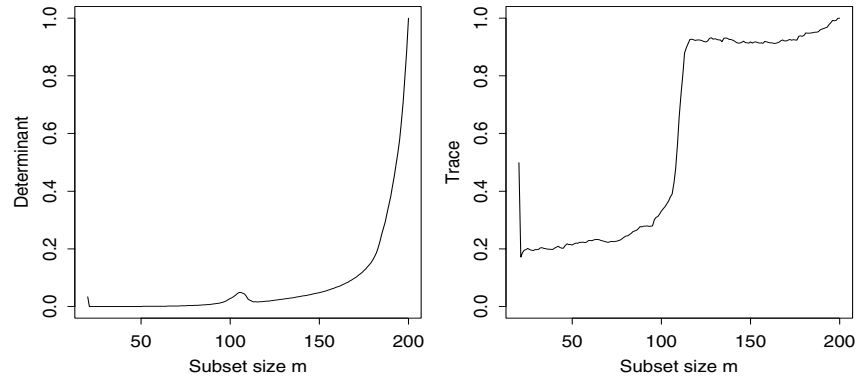


FIGURE 2.7. Swiss bank notes starting with the first 20 observations on genuine bank notes: forward plots of the estimated covariance matrix; left panel, the determinant and, right panel, the trace

2.18 Solutions

Exercise 2.1

$$\begin{aligned}
 \mathbb{E}(y - \mu)^T \Sigma^{-1} (y - \mu) &= \mathbb{E} \operatorname{tr} (y - \mu)^T \Sigma^{-1} (y - \mu) \\
 &= \operatorname{tr} \Sigma^{-1} \mathbb{E} (y - \mu) (y - \mu)^T \\
 &= \operatorname{tr} \Sigma^{-1} \Sigma \\
 &= \operatorname{tr} I_v \\
 &= v.
 \end{aligned}$$

Exercise 2.2

$$\begin{aligned}
 \operatorname{var}(e_i) &= \operatorname{var}(y_i - \bar{y}) \\
 &= \operatorname{var}(y_i) + \operatorname{var}(\bar{y}) - 2\operatorname{cov}(y_i, \bar{y}) \\
 &= \sigma^2 + \frac{\sigma^2}{n} - \frac{2}{n} \operatorname{cov}(y_i, \sum_{i=1}^n y_i) \\
 &= \sigma^2 + \frac{\sigma^2}{n} - \frac{2\sigma^2}{n} \\
 &= \frac{n-1}{n} \sigma^2.
 \end{aligned}$$

Note that no distributional assumptions have been made.

Exercise 2.3

A matrix C is symmetric when $C = C^T$. We have that

$$C^T = (I - JJ^T/n)^T = (I - JJ^T/n) = C$$

A matrix C is idempotent when $CC = C$. We have that

$$\begin{aligned} CC &= (I - JJ^T/n)(I - JJ^T/n) = I - JJ^T/n - JJ^T/n + nJJ^T/n^2 \\ &= I - JJ^T/n = C. \end{aligned}$$

Note that

$$\text{rank}(C) = \text{tr}C = \text{tr}I - \text{tr}J^T J/n = \text{tr}I - \text{tr}(n/n) = n - 1.$$

Given that C is symmetric and idempotent with rank $(n - 1)$ we have that

$$S(\hat{\mu}) = Y^T CY$$

is distributed as $W_v(\Sigma, n - 1)$.

Exercise 2.4

We have to find the distribution of

$$\frac{(y_i - \mu)^2}{s^2},$$

which be rewritten as

$$\frac{(y_i - \mu)^2}{s^2} = (n - 1) \frac{(y_i - \mu)^2 / \sigma^2}{(y_i - \mu)^2 / \sigma^2 + \sum_{j \neq i=1}^n (y_j - \mu)^2 / \sigma^2}. \quad (2.112)$$

It is straightforward to see that

$$\frac{(y_i - \mu)^2}{s^2} \sim (n - 1) \frac{\chi_1^2}{\chi_1^2 + \chi_{n-1}^2}.$$

Given that the χ_1^2 in the denominator is independent of the χ_{n-1}^2 , the resulting distribution is Beta, that is

$$\frac{(y_i - \mu)^2}{s^2} \sim (n - 1) \text{Beta}\left(\frac{1}{2}, \frac{n - 1}{2}\right).$$

Exercise 2.5

Equation (2.4) can be rewritten as

$$\frac{(y_i - \bar{y})^2}{s^2} = \frac{(n - 1)^2}{n} \frac{\frac{(y_i - \bar{y})^2}{((n - 1)/n)\sigma^2}}{\sum_{i=1}^n \frac{(y_i - \bar{y})^2}{\sigma^2}}$$

Now, given that $(n-1)s^2$ can be decomposed as the sum of two quantities,

$$(n-1)s^2 = \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{n}{n-1} (y_i - \bar{y})^2 + \sum_{j \neq i=1}^n (y_j - \bar{y}_{(i)})^2,$$

where $\bar{y}_{(i)} = \sum_{j \neq i=1}^n y_j / (n-1)$, we obtain

$$\frac{(y_i - \bar{y})^2}{s^2} = \frac{(n-1)^2}{n} \frac{\frac{(y_i - \bar{y})^2}{((n-1)/n)\sigma^2}}{\frac{(y_i - \bar{y})^2}{((n-1)/n)\sigma^2} + \frac{\sum_{j \neq i=1}^n (y_j - \bar{y}_{(i)})^2}{\sigma^2}}. \quad (2.113)$$

From equation (2.113) and exercise (2.2) it follows that

$$\frac{(y_i - \bar{y})^2}{s^2} \sim \frac{(n-1)^2}{n} \frac{\chi_1^2}{\chi_1^2 + \chi_{n-2}^2}.$$

We now have to prove that the χ_1^2 which appears both in the numerator and the denominator

$$\frac{n}{n-1} \frac{(y_i - \bar{y})^2}{\sigma^2} \sim \chi_1^2$$

is independent of the χ_{n-2}^2 of the denominator

$$\frac{\sum_{j \neq i=1}^n (y_j - \bar{y}_{(i)})^2}{\sigma^2} \sim \chi_{n-2}^2.$$

The proof we give has two steps. First we write the χ^2 variables as idempotent quadratic forms. Then, we show that the product of the matrices of the two quadratic forms is equal to zero so we conclude that the two random variables are independent. The numerator of equation (2.113) can be rewritten as

$$\frac{n}{n-1} (y_i - \bar{y})^2 = y^T Q_1 y,$$

where $y = (y_1, \dots, y_n)^T$, $Q_1 = \frac{n}{n-1} q(i)q(i)^T (I_n - JJ^T/n)$, and $q(i)$ is a vector which has a 1 in i th position and 0 elsewhere: $q(i) = (0, \dots, 0, 1, 0, \dots, 0)^T$. Q_1 is symmetric and idempotent with trace (rank) equal to 1. On the other hand,

$$\sum_{j \neq i=1}^n (y_j - \bar{y}_{(i)})^2 = y^T Q_2 y,$$

where $Q_2 = I_n - q(i)q(i)^T - \{J - q(i)\}\{J - q(i)\}^T / (n-1)$. Q_2 is symmetric and idempotent with trace (rank) equal to $n-2$. Since

$$\{I_n - q(i)q(i)^T\}q(i) = 0 \quad \text{and} \quad \{J - q(i)\}^T q(i) = 0,$$

it follows that $Q_1 Q_2 = 0$. We thus conclude that the two χ^2 random variables are independent. Using the independence argument between the two χ^2 random variables and the relationship between Gamma and Beta,

$$\frac{(y_i - \bar{y})^2}{s^2} \sim \frac{(n-1)^2}{n} \frac{\chi_1^2}{\chi_1^2 + \chi_{n-2}^2} \sim \frac{(n-1)^2}{n} \text{Beta}\left(\frac{1}{2}, \frac{n-2}{2}\right). \quad (2.114)$$

Note that (2.114) is just the special case of (2.52) for $v = 1$.

Exercise 2.6

We now require the distribution of

$$e_i^2/s^2 = k \frac{e_i^2/\{\sigma^2(1-h_i)\}}{\sum_{j=1}^n e_j^2/\sigma^2}, \quad (2.115)$$

where

$$k = (n-p)(1-h_i). \quad (2.116)$$

Since

$$\text{var } e_i = \sigma^2(1-h_i),$$

$$\frac{e_i^2}{\sigma^2(1-h_i)} \sim \chi_1^2.$$

From (2.95)

$$(n-p)s^2 = (n-p-1)s_{(i)}^2 + e_i^2/(1-h_i).$$

The residual sum of squares $(n-p-1)s_{(i)}^2 \sim \sigma^2 \chi_{n-p-1}^2$ independently of y_i and of $\hat{\beta}_{(i)}$. But, from (2.94),

$$e_i = (1-h_i)(y_i - x_i^T \hat{\beta}_{(i)}).$$

Thus e_i^2 and $s_{(i)}^2$ are independent and

$$e_i^2/s^2 \sim k \frac{\chi_1^2}{\chi_1^2 + \chi_{n-p-1}^2} \sim k \text{Beta}\left(\frac{1}{2}, \frac{n-p-1}{2}\right),$$

since the two χ^2 variables are independent, with k defined in (2.116). This is the result in (2.79) for $v = 1$.

The result of Exercise 2.5 (2.114) is obtained when just the mean is fitted, so that $p = 1$ and $h_i = 1/n$.

Exercise 2.7

Since the y_i 's are independent (because they arise from a random sample)

the likelihood function (joint density) $Lik(\mu, \Sigma; y)$ is the product of the densities of the y_i 's

$$\begin{aligned} Lik(\mu, \Sigma; y) &= \prod_{i=1}^n f(y_i; \mu, \Sigma) \\ &= \prod_{i=1}^n \frac{1}{(\sqrt{2\pi})^v |\Sigma|^{1/2}} \exp\{-(y_i - \mu)^T \Sigma^{-1} (y_i - \mu)/2\} \\ &= \frac{1}{(\sqrt{2\pi})^{nv} |\Sigma|^{n/2}} \exp\{-\sum_{i=1}^n (y_i - \mu)^T \Sigma^{-1} (y_i - \mu)/2\}. \end{aligned}$$

The log likelihood is given by:

$$L(\mu, \Sigma; y) = -(n/2) \log |2\pi\Sigma| - \sum_{i=1}^n (y_i - \mu)^T \Sigma^{-1} (y_i - \mu)/2. \quad (2.117)$$

This solves part 1) of the exercise. As concerns part 2), in order to derive the expressions for the maximum likelihood estimators of μ and Σ , we first write the quadratic form in equation (2.117) in a way that facilitates finding the maximum. Since a scalar quantity is equal to its trace,

$$\begin{aligned} \sum_{i=1}^n (y_i - \mu)^T \Sigma^{-1} (y_i - \mu) &= \sum_{i=1}^n \text{tr}(y_i - \mu)^T \Sigma^{-1} (y_i - \mu) \\ &= \text{tr} \Sigma^{-1} \sum_{i=1}^n (y_i - \mu)(y_i - \mu)^T. \end{aligned} \quad (2.118)$$

Now, by adding and subtracting \bar{y} in the sum in the right hand side of (2.118), we obtain

$$\begin{aligned} \sum_{i=1}^n (y_i - \mu)(y_i - \mu)^T &= \sum_{i=1}^n (y_i - \bar{y} + \bar{y} - \mu)(y_i - \bar{y} + \bar{y} - \mu)^T \\ &= \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})^T + n(\bar{y} - \mu)(\bar{y} - \mu)^T \\ &= S(\hat{\mu}) + n(\bar{y} - \mu)(\bar{y} - \mu)^T. \end{aligned} \quad (2.119)$$

The other two terms in equation (2.119) vanish because $\sum_{i=1}^n (y_i - \bar{y}) = 0$. Using (2.119) and (2.118) in (2.117) we obtain

$$\begin{aligned} L(\mu, \Sigma; y) &= -\frac{nv}{2} \log 2\pi - \frac{n}{2} \log |\Sigma| \\ &\quad - \frac{1}{2} \text{tr} \Sigma^{-1} \{S(\hat{\mu}) + n(\bar{y} - \mu)(\bar{y} - \mu)^T\} \\ &= -\frac{nv}{2} \log 2\pi - \frac{n}{2} \log |\Sigma| \end{aligned} \quad (2.120)$$

$$- \frac{1}{2} \text{tr}(\Sigma^{-1} S(\hat{\mu})) - \frac{n}{2} (\bar{y} - \mu)^T \Sigma^{-1} (\bar{y} - \mu). \quad (2.121)$$

To find the maximum likelihood estimator for μ we differentiate $L(\mu, \Sigma; y)$ in (2.121) with respect to μ and set the resulting expression equal to 0:

$$\frac{\partial L(\mu, \Sigma; y)}{\partial \mu} = -0 - 0 - 0 + n(\Sigma^{-1}\bar{y} - \Sigma^{-1}\mu) = 0$$

which gives

$$\hat{\mu} = \bar{y}.$$

It is clear that $\hat{\mu} = \bar{y}$ maximizes $\log L(\mu, \Sigma; y)$ with respect to μ because the last term in (2.121) is ≤ 0 and the term vanishes for $\hat{\mu} = \bar{y}$. Before differentiating $\log L(\mu, \Sigma; y)$ to find $\hat{\Sigma}$, we substitute $\mu = \bar{y}$ in (2.121) and rewrite $\log |\Sigma|$ in terms of Σ^{-1} to obtain

$$L(\hat{\mu}, \Sigma; y) = -\frac{nv}{2} \log 2\pi + \frac{n}{2} \log |\Sigma^{-1}| - \frac{1}{2} \text{tr}\{\Sigma^{-1}S(\hat{\mu})\}. \quad (2.122)$$

We now differentiate (2.122) with respect to Σ^{-1} , remembering that

$$\frac{\partial \text{tr}(AB)}{\partial A} = B + B^T - \text{diag}(B)$$

and that

$$\frac{\partial \log |A|}{\partial A} = 2A^{-1} - \text{diag}(A^{-1}).$$

We obtain

$$\frac{\partial L(\hat{\mu}, \Sigma; y)}{\partial \Sigma^{-1}} = -0 + n\Sigma - \frac{n}{2} \text{diag}(\Sigma) - S(\hat{\mu}) + \frac{1}{2} \text{diag}S(\hat{\mu}) = 0, \quad (2.123)$$

whence

$$\hat{\Sigma} - \frac{1}{2} \text{diag}(\hat{\Sigma}) = \frac{1}{n} \{S(\hat{\mu}) - \frac{1}{2} \text{diag}S(\hat{\mu})\}$$

or

$$\hat{\Sigma} = \frac{S(\hat{\mu})}{n}.$$

Note that we solved (2.123) for Σ rather than Σ^{-1} , even though we differentiated with respect to Σ^{-1} . Otherwise we would have obtained $\{S(\hat{\mu})/n\}^{-1}$ as the maximum likelihood estimator for Σ^{-1} . We have exploited the property of invariance of maximum likelihood estimators.

For part 3) of the exercise, we have from equation (2.122) that the log-likelihood maximized with respect to $\hat{\mu}$ and $\hat{\Sigma}$ is

$$\begin{aligned} L(\hat{\mu}, \hat{\Sigma}; y) &= -nv \log \sqrt{2\pi} + \frac{n}{2} \log |\hat{\Sigma}^{-1}| - \frac{1}{2} \text{tr}(\hat{\Sigma}^{-1}S(\hat{\mu})) \\ &= -nv \log \sqrt{2\pi} + \frac{n}{2} \log |\hat{\Sigma}^{-1}| - \frac{nv}{2} \\ &= -\frac{n}{2} \log(2\pi)^v - \frac{n}{2} \log |\hat{\Sigma}| - \frac{nv}{2} \\ &= -\frac{n}{2} \log |2\pi \hat{\Sigma}| - \frac{nv}{2}. \end{aligned}$$

Exercise 2.8

In order to test the hypothesis of equality of means, the matrix D and the vectors c and μ are

$$\begin{pmatrix} 1 & -1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & -1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 1 & -1 \end{pmatrix},$$

$c = (0, \dots, 0)^T$ and $\mu = (\mu_1, \dots, \mu_v)^T$. The first row of the matrix D imposes the constraint $\mu_1 - \mu_2 = 0$, the second $\mu_2 - \mu_3 = 0$, ..., the last $\mu_{v-1} - \mu_v = 0$. D in this case has dimension $(v-1) \times v$ and has full row rank.

Exercise 2.9

We start by rewriting the expression which defines the likelihood ratio test (2.20)

$$T_{LR} = n \log(|\hat{\Sigma}_0|/|\hat{\Sigma}|).$$

Given that $n\hat{\Sigma}_0$ can be decomposed as

$$n\hat{\Sigma}_0 = \sum_{i=1}^n (y_i - \mu_0)(y_i - \mu_0)^T = \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})^T + n(\bar{y} - \mu_0)(\bar{y} - \mu_0)^T,$$

$$\begin{aligned} T_{LR} &= n \log \left\{ \left| \sum_{i=1}^n \frac{(y_i - \bar{y})(y_i - \bar{y})^T}{n} + \frac{n(\bar{y} - \mu_0)(\bar{y} - \mu_0)^T}{n} \right| / |\hat{\Sigma}| \right\} \\ &= n \log \left\{ \left| \hat{\Sigma} + (\bar{y} - \mu_0)(\bar{y} - \mu_0)^T \right| / |\hat{\Sigma}| \right\}. \end{aligned}$$

Now, since $|A + bb^T| = |A|(1 + b^T A^{-1}b)$, we can rewrite the former equation as

$$\begin{aligned} T_{LR} &= n \log \frac{|\hat{\Sigma}| \left\{ 1 + (\bar{y} - \mu_0)^T \hat{\Sigma}^{-1} (\bar{y} - \mu_0) \right\}}{|\hat{\Sigma}|} \\ &= n \log \left\{ 1 + n(\bar{y} - \mu_0)^T \hat{\Sigma}_u^{-1} (\bar{y} - \mu_0) / (n-1) \right\} \\ &= n \log \left\{ 1 + T^2 / (n-1) \right\}. \end{aligned}$$

This implies that the likelihood ratio test is a monotone function of Hotelling's T^2 statistic.

Exercise 2.10

We have a random sample of size n_l from each of $N_v(\mu_l, \Sigma_l; y)$, $l = 1, 2, \dots, g$. The likelihood function is

$$\begin{aligned} \text{Lik}(\mu_1, \mu_2, \dots, \mu_g, \Sigma_1, \Sigma_2, \dots, \Sigma_g, y) &= \prod_{l=1}^g \text{Lik}(\mu_l, \Sigma_l; y \in \text{group } l) \\ &= \frac{1}{(\sqrt{2\pi})^{nv}} \prod_{l=1}^g |\Sigma_l|^{n_l/2} \exp \left[-\frac{1}{2} \text{tr} \Sigma_l^{-1} \{ n_l \Sigma_l + n_l (\bar{y}_l - \mu_l)(\bar{y}_l - \mu_l)^T \} \right]. \end{aligned}$$

The maximum likelihood estimate of μ_l ($v \times 1$ vector of means for group l) is \bar{y}_l under both H_0 and H_1 because there is no restriction on the population means. The maximum likelihood estimate of Σ_l is $\sum_{l=1}^g n_l \hat{\Sigma}_l / n$ under H_0 where $n = \sum_{l=1}^g n_l$. Under the alternative H_1 , the maximum likelihood estimate of Σ_l is $\hat{\Sigma}_l$. So the maximized likelihood in the two cases is:

$$\begin{aligned} \max_{H_1} \text{Lik} &= \frac{1}{(\sqrt{2\pi})^{nv}} \prod_{l=1}^g |\hat{\Sigma}_l|^{-n_l/2} \exp(-n_l v/2), \\ \max_{H_0} \text{Lik} &= \frac{1}{(\sqrt{2\pi})^{nv}} \left| \sum_{l=1}^g n_l \hat{\Sigma}_l / n \right|^{-n/2} \exp(-nv/2). \end{aligned}$$

The maximized loglikelihoods are

$$\begin{aligned} L_1 = \max_{H_1} L &= -\frac{nv}{2} \log 2\pi - \frac{1}{2} \sum_{l=1}^g n_l \log |\hat{\Sigma}_l| - nv/2, \\ L_0 = \max_{H_0} L &= -\frac{nv}{2} \log 2\pi - \frac{n}{2} \log \left| \sum_{l=1}^g n_l \hat{\Sigma}_l / n \right| - nv/2. \end{aligned}$$

The likelihood ratio test ($2L_1 - 2L_0$) is equal to

$$n \log \left| \sum_{l=1}^g n_l \hat{\Sigma}_l / n \right| - \sum_{l=1}^g n_l \log |\hat{\Sigma}_l|.$$

Exercise 2.11

We must show that the product of $(C - xx^T)$ with the right hand side of (2.40) gives the identity matrix

$$\begin{aligned} (C - xx^T) \left(C^{-1} + \frac{C^{-1}xx^TC^{-1}}{1 - x^TC^{-1}x} \right) &= \\ I_p - xx^TC^{-1} + \frac{xx^TC^{-1}}{1 - x^TC^{-1}x} - \frac{xx^TC^{-1}xx^TC^{-1}}{1 - x^TC^{-1}x} &= \\ I_p + \frac{-xx^TC^{-1} + xx^TC^{-1}x^TC^{-1}x + xx^TC^{-1} - xx^TC^{-1}xx^TC^{-1}}{1 - x^TC^{-1}x} &= \\ I_p + \frac{xx^TC^{-1}x^TC^{-1}x - xx^TC^{-1}xx^TC^{-1}x}{1 - x^TC^{-1}x} &= I_p. \end{aligned}$$

For the generalization of the Sherman-Morrison-Woodbury formula, we have to show that the product of $(A - UV^T)$ with $\{A^{-1} + A^{-1}U(I_m - V^T A^{-1}U)^{-1}V^T A^{-1}\}$ gives the identity matrix.

$$\begin{aligned} (A - UV^T)\{A^{-1} + A^{-1}U(I_m - V^T A^{-1}U)^{-1}V^T A^{-1}\} &= \\ I_p + U(I_m - V^T A^{-1}U)^{-1}V^T A^{-1} - UV^T A^{-1} &= \\ -UV^T A^{-1}U(I_m - V^T A^{-1}U)^{-1}V^T A^{-1} &= \\ I_p - UV^T A^{-1} + U(I_m - V^T A^{-1}U)(I_m - V^T A^{-1}U)^{-1}V^T A^{-1} &= \\ I_p - UV^T A^{-1} + UV^T A^{-1} = I_p. \end{aligned}$$

This generalization can be applied to the deletion of m rows of matrix X because $X_{(m)}^T X_{(m)}$ can be written as $X_{(m)}^T X_{(m)} = X^T X - X_m X_m^T$.

Exercise 2.12

We start with $m = n$. From (2.28)

$$d_i^2 = (y_i - \hat{\mu})^T \hat{\Sigma}_u^{-1} (y_i - \hat{\mu}),$$

where, from (2.16)

$$\hat{\Sigma}_u = S(\hat{\mu})/(n-1) = \left\{ \sum_{i=1}^n (y_i - \hat{\mu})(y_i - \hat{\mu})^T \right\} / (n-1).$$

Then

$$\begin{aligned} \sum_{i=1}^n d_i^2 &= (n-1) \operatorname{tr} \sum_{i=1}^n (y_i - \hat{\mu})^T S(\hat{\mu})^{-1} (y_i - \hat{\mu}) \\ &= (n-1) \operatorname{tr} \sum_{i=1}^n S(\hat{\mu})^{-1} (y_i - \hat{\mu})(y_i - \hat{\mu})^T \\ &= (n-1) \operatorname{tr} I_v = (n-1)v. \end{aligned}$$

Thus $\sum_{i=1}^m d_{im}^{*2} = (m-1)v$.

Since there is no limit on how remote a unit not included in the subset can be, the upper bound on $\sum_{i=1}^n d_{im}^{*2} = \infty$. If all units sit exactly on an ellipsoid, all d_{im}^{*2} will be equal and the sum $= n(m-1)v/m$. If there is no inversion or interchange and all units in $S_*^{(m)}$ sit on the ellipsoid, units not in $S_*^{(m)}$ must have distances greater than the average value $(m-1)v/m$ and so $\sum_{i=1}^n d_{im}^{*2} \geq n(m-1)v/m$. If all units are on the ellipsoid, the choice of units to include or exclude would be arbitrary, the decision having no effect on Mahalanobis distances. With an inversion, exactly one unit will have a smaller distance than the m comprising $S_*^{(m)}$. The minimum value of this distance is zero, so $\sum_{i=1}^n d_{im}^{*2} \geq (n-1)(m-1)v/m$. With an interchange, if all units in the subset have the same Mahalanobis distance, at least two

units must have values less than the average; the maximum number of units with zero distances is $n - m$. In this case, $\sum_{i=1}^n d_{im}^{*2} \geq (m - 1)v$. The step of the search going from $S_*^{(m)}$ to $S_*^{(m+1)}$ will then destroy this structure and the sum of all the distances will increase.

Exercise 2.13

- (a) $H^T = (X(X^T X)^{-1} X^T)^T = X(X^T X)^{-1} X^T = H$.
 (b) $HH = X(X^T X)^{-1} X^T X(X^T X)^{-1} X^T = X(X^T X)^{-1} X^T$.
 (c) $\text{tr } H = \sum_{i=1}^n h_i = \text{tr } \{X(X^T X)^{-1} X^T\} = \text{tr } (X^T X)(X^T X)^{-1} = \text{tr } I_p = p$.

When X contains only the constant term, that is $X = J$.

$$H = J(J^T J)J^T = \frac{1}{n} J J^T.$$

The vector of residuals e can be written as

$$e = (I - H)y = (I - \frac{1}{n} J J^T)y = Cy.$$

So, C is the hat matrix for the model which contains only the constant term.

Exercise 2.14

There are $p = k + 1$ columns in X , the column for the constant term, which is a vector of ones, and k columns, one for each variable in the model, which contain 2^{k-1} entries of $+1$ and the same number of -1 entries. The columns are mutually orthogonal, so

$$X^T X = \text{diag } (n, \dots, n) \quad \text{and} \quad (X^T X)^{-1} = \text{diag } (1/n, \dots, 1/n),$$

where $n = 2^k$.

The hat matrix $H = X(X^T X)^{-1} X^T$ is $n \times n$. The leverage measures h_i are the diagonal terms of H :

$$h_i = \sum_{j=1}^p x_{ij}^2 / n = p/n = (k + 1)/n.$$

Then $\sum_{i=1}^n h_i = p$, in agreement with the results of Exercise 2.13.

The interaction terms give additional columns of X formed by multiplication of columns i and j . These columns are, as before, orthogonal to all others. So p increases to some larger value p^+ , when all $h_i = p^+/n$.

Exercise 2.15

The trace of the estimated covariance matrix is a function only of the variances of the variables; the determinant also includes the correlations. The

right panel of Figure 2.7 shows that the inclusion of units from the second group causes an appreciable increase in the variances of the variables (signalled by a sudden change of slope in the trace). The left panel of Figure 2.7 shows that the increase of the variances due to the initial inclusion of the units from the group of forgeries is partially counterbalanced by the increase in the covariances. Due to this compensation, the overall effect on the determinant seems to be negligible compared to that on the variances (see the left panel of Figure 2.7). This conclusion is in agreement with what we had already seen in Figure 1.16. This figure showed that around $m = 105$ - $m = 110$ there was not only a big increase in the variances of variables 4, 5, 6, but also an increase in absolute values of the covariances between variables 6 and 4, and between variables 6 and 5.

Exploring Multivariate Data with the Forward Search

Atkinson, A.; Riani, M.; cccc, A.

2004, XXIV, 624 p., Hardcover

ISBN: 978-0-387-40852-1