# 1. Overview of the General Linear Model

## Theory



## Applications



## 1.1 Introduction

In this chapter we introduce the structure of the general linear model and use the structure to classify the linear models discussed in this book. The multivariate normal distribution which forms the basis for most of the hypothesis testing theory for the linear model is reviewed, along with a general approach to hypothesis testing. Graphical methods for assessing univariate and multivariate normality are also reviewed. The chapter illustrates multivariate normal data generation, Q-Q plots, chi-square plots, scatter plots, and data transformation procedures to evaluate normality.

## 1.2 The General Linear Model

Data analysis in the social and behavioral sciences and numerous other disciplines is associated with the model statisticians call the general linear model (GLM). If we employ matrix notation, univariate and multivariate linear models may be represented using the general structure

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e} \tag{1.1}$$

where $\mathbf{y}_{n\times 1}$ is a vector of $n$ observations, $\mathbf{X}_{n\times k}$ is a known design matrix of full column rank $k$, $\beta_{k\times 1}$ is a vector of $k$ fixed parameters, and $\mathbf{e}_{n\times 1}$ is a random vector of errors with mean zero, $E(\mathbf{e}) = \mathbf{0}$, and covariance matrix $\mathbf{\Omega} = \text{cov}(\mathbf{e})$. If the design matrix is not of full rank, one may reparameterize the model to create an equivalent model of full rank. In this book, we systematically discuss the GLM specified by (1.1) with various structures for $\mathbf{X}$, the design matrix, and $\mathbf{\Omega}$, the covariance matrix of errors.

Depending on the structure of $\mathbf{X}$ and $\mathbf{\Omega}$, the model in (1.1) has many names in the literature. To illustrate, if $\mathbf{\Omega} = \sigma^2\mathbf{I}_n$ in (1.1), the model is called the classical linear regression model or the standard linear regression model. If we partition $\mathbf{X}$ to have the form $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ where $\mathbf{X}_1$ is associated with fixed effects and $\mathbf{X}_2$ is associated with random effects, and if the covariance matrix $\mathbf{\Omega}$ has the form

$$\mathbf{\Omega} = \mathbf{X}_2\mathbf{V}\,\mathbf{X}_2' + \mathbf{\Psi}$$

where $\mathbf{V}$ and $\mathbf{\Psi}$ are covariance matrices, then (1.1) becomes the general linear mixed model (GLMM). If we let $\mathbf{X}$ and $\mathbf{\Omega}$ take the general form

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & \cdots & \cdots & \mathbf{0} \\ \vdots & \mathbf{X}_2 & \cdots & \vdots \\ \vdots & \cdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \cdots & \mathbf{X}_p \end{pmatrix}$$

$$\mathbf{\Omega} = \mathbf{\Sigma} \otimes \mathbf{I}_n$$

where $\mathbf{\Sigma}_{p\times p}$ is a covariance matrix, and $\mathbf{A} \otimes \mathbf{B}$ denotes the Kronecker product of two matrices $\mathbf{A}$ and $\mathbf{B}$ ($\mathbf{A} \otimes \mathbf{B} = a_{ij}\mathbf{B}$), then (1.1) is called Zellner's seemingly unrelated regression (SUR) model or the multiple design multivariate (MDM) model. The SUR model may also be formulated as $p$ separate linear regression models that are not independent:

$$\mathbf{y}_i = \mathbf{X}_i\beta_{ii} + \mathbf{e}_i$$
$$\text{cov}(\mathbf{y}_i, \mathbf{y}_j) = \sigma_{ij}\mathbf{I}_n$$

for $i, j = 1, 2, \ldots, p$ where $\mathbf{y}$, $\beta$ and $\mathbf{e}$ in (1.1) are partitioned:

$$\mathbf{y}' = (\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_p') \quad \text{where} \quad \mathbf{y}_i : n^* \times 1$$
$$\beta' = (\beta_1', \beta_2', \ldots, \beta_p') \quad \text{where} \quad \beta_{ii} : k_i \times 1$$
$$\mathbf{e}' = (\mathbf{e}_1', \mathbf{e}_2', \ldots, \mathbf{e}_p') \quad \text{where} \quad \mathbf{e}_i : n^* \times 1$$

and $\mathbf{\Sigma}_{p\times p} = (\sigma_{ij})$. Alternatively, we may express the SUR model as a restricted multivariate regression model. To do this, we write

$$\mathbf{Y}_{n^*\times p} = \mathbf{X}_{n^*\times k}\tilde{\mathbf{B}}_{k\times p} + \mathbf{E}_{n^*\times p}$$

where $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p)$, $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p)$, $\mathbf{E} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p)$ and

$$\tilde{\mathbf{B}} = \begin{pmatrix} \beta_{11} & \cdots & \cdots & \mathbf{0} \\ \vdots & \beta_{22} & \cdots & \vdots \\ \vdots & \cdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \cdots & \beta_{pp} \end{pmatrix}$$

Letting $\mathbf{X}_1 = \mathbf{X}_2 = \ldots = \mathbf{X}_p = \tilde{\mathbf{X}}_{n^* \times k}$ and $\mathbf{B} = (\beta_{11}, \beta_{22}, \ldots, \beta_{pp})$ in the SUR model, (1.1) becomes the classical multivariate regression model or the MANOVA model. Finally, letting

$$\mathbf{X} = \mathbf{X}_1 \otimes \mathbf{X}_2'$$
$$\mathbf{\Omega} = \mathbf{I}_n \otimes \mathbf{\Sigma}$$

model (1.1) becomes the generalized multivariate analysis of variance (GMANOVA) or the generalized growth curve model. All these models with some further extensions are special forms of the general linear model discussed in this book.

The model in (1.1) is often termed the "classical" model since its orientation is subjects or observations by variables where the number of variables is one. An alternative orientation for the model is to assume that $\mathbf{y}' = (y_1, y_2, \ldots, y_n)$ is a $(1 \times n)$ vector of observations where the number of variables is one. For each observation $y_i$, we may assume that there are $\mathbf{x}_i' = (x_1, x_2, \ldots, x_k)$ or $k$ independent (possibly dummy) variables. With this orientation, (1.1) becomes

$$\mathbf{y} = \mathbf{X}'\beta + \mathbf{e} \tag{1.2}$$

where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$, $\mathbf{e}' = (e_1, e_2, \ldots, e_n)$ and each $\mathbf{x}_i$ contains $k$ independent variables for the $i^{\text{th}}$ observation. Model (1.2) is often called the "responsewise" form. Model (1.1) is clearly equivalent to (1.2) since the design matrix has the same order for either representation; however, in (1.2) $\mathbf{X}$ is of order $k \times n$. Thus, $\mathbf{X}'\mathbf{X}$ using (1.2) becomes $\mathbf{X}\mathbf{X}'$ for the responsewise form of the general linear model.

The simplest example of the GLM is the simple linear regression model:

$$y = \beta_o + \beta_1 x + e \tag{1.3}$$

where x represents the independent variable, $y$ the dependent variable and $e$ a random error. Model (1.3) states that the observed dependent variable for each subject is hypothesized to be a function of a common parameter $\beta_o$ for all subjects and an independent variable x for each subject that is related to the dependent variable by a weighting (i.e., regression) coefficient $\beta_1$ plus a random error $e$. For $k = m + 1$ with $m$ variables, (1.3) becomes (1.4)

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + e \tag{1.4}$$

or using matrix notation, (1.4) is written as

$$y = \mathbf{x}'\beta + e \tag{1.5}$$

where $\mathbf{x}' = (x_1, x_2, \ldots, x_k)$ denotes $k$ independent variables, and $x_1$ is a dummy variable in the vector $\mathbf{x}'$. Then for a sample of $n$ observations, (1.5) has the general form (1.2) where $\mathbf{y}' = (y_1, y_2, \ldots, y_n)$, $\mathbf{e}' = (e_1, e_2, \ldots, e_n)$, and $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ of order $k \times n$ since each column vector $\mathbf{x}_i$ in $\mathbf{X}$ contains $k$ variables. When using the classical form (1.1), $\mathbf{X} \equiv \mathbf{X}'$, a matrix of order $n \times k$. In discussions of the GLM, many authors use either the classical or the responsewise version of the GLM, while we in general prefer (1.1). In some applications (e.g., repeated measurement designs) the form (1.2) is preferred.

## 1.3 The Restricted General Linear Model

In specifying the GLM using (1.1) or (1.2), we have not restricted the $k$-variate parameter vector $\beta$. A linear restriction on the parameter vector $\beta$ will affect the characterization of the model. Sometimes it is necessary to add restrictions to the GLM of the general form

$$\mathbf{R}\beta = \theta \tag{1.6}$$

where $\mathbf{R}_{s\times k}$ is a known matrix with full row rank, rank($\mathbf{R}$) = $s$, and $\theta$ is a known parameter vector, often assumed to be zero. With (1.6) associated with the GLM, the model is commonly called the restricted general linear model (RGLM). Returning to (1.3), we offer an example of this in the simple linear regression model with a restriction

$$\begin{aligned} y &= \beta_o + \beta_1 x + e \\ \beta_o &= 0 \end{aligned} \tag{1.7}$$

so that the regression of $y$ on x is through the origin. For this situation, $\mathbf{R} \equiv (1,0)$ and $\theta = (0, 0)$. Clearly, the estimate of $\beta_1$ using (1.7) will differ from that obtained using the general linear model (1.3) without the restriction.

Assuming (1.1) or (1.4), one first wants to estimate $\beta$ with $\hat{\beta}$ where the estimator $\hat{\beta}$ has some optimal properties like unbiasedness and minimal variance. Adding (1.6) to the GLM, one obtains a restricted estimator of $\beta$, $\hat{\beta}_r$, which in general is not equal to the unrestricted estimator. Having estimated $\beta$, one may next want to test hypotheses regarding the parameter vector $\beta$ and the structure of $\mathbf{\Omega}$. The general form of the null hypothesis regarding $\beta$ is

$$H: \mathbf{C}\beta = \xi \tag{1.8}$$

where $\mathbf{C}_{g\times k}$ is a matrix of full row rank $g$, rank($\mathbf{C}$) = $g$ and $\xi_{g\times 1}$ is a vector of known parameters, usually equal to zero. The hypothesis in (1.8) may be tested using the GLM with or without the restriction given in (1.6). Hypotheses in the form (1.8) are in general testable, provided that $\beta$ is estimable; however, testing (1.8) by assuming (1.6) is more complicated;since the matrix $\mathbf{C}$ may not contain a row identical, inconsistent or dependent on the rows of $\mathbf{R}$ and since the rows of $\mathbf{C}$ must remain independent. Thus, the rank of the augmented matrix must be greater than $s$, $\begin{pmatrix} \mathbf{R} \\ \mathbf{C} \end{pmatrix} = s + g > s$.

Returning to (1.3), we may test the null hypotheses

$$H: \beta = \begin{pmatrix} \beta_o \\ \beta_1 \end{pmatrix} = \xi = \begin{pmatrix} \xi_o \\ \xi_1 \end{pmatrix} \tag{1.9}$$

where $\xi$ is a known parameter vector. The hypothesis in (1.9) may not be inconsistent with the restriction $\beta_o = 0$. Thus, given the restriction, we may test

$$H: \beta_1 = \xi_1 \tag{1.10}$$

so that (1.10) is not inconsistent or dependent on the restriction.

## 1.4 The Multivariate Normal Distribution

To test hypotheses of the form given in (1.8), one must usually make distributional assumptions regarding the observation vector $\mathbf{y}$ or $\mathbf{e}$, namely the assumption of multivariate normality. To define the multivariate normal distribution, recall that the definition of a standard normal random variable $y$ is defined by the density

$$f(y) = (2\pi)^{-1/2} \exp(-y^2/2) \tag{1.11}$$

denoted by $y \sim N(0,1)$. A random variable $y$ has a normal distribution with mean $\mu$ and variance $\sigma^2 > 0$ if $y$ has the same distribution as the random variable

$$\mu + \sigma e \tag{1.12}$$

where $e \sim N(0,1)$. The density for $y$ is given by

$$\begin{aligned} f(y) &= \frac{1}{\sigma\sqrt{2\pi}} \exp[-(y-\mu)^2/2\sigma^2] \\ &= (2\pi\sigma^2)^{-1/2} \exp[-(y-\mu)^2/2\sigma^2] \end{aligned} \tag{1.13}$$

With this as motivation, the definition for a multivariate normal distribution is as follows.

**Definition 1.1** A $p$-dimensional random vector $\mathbf{y}$ is said to have a multivariate normal distribution with mean $\mu$ and covariance matrix $\mathbf{\Sigma}$ [$\mathbf{y} \sim N_p(\mu, \mathbf{\Sigma})$] if $\mathbf{y}$ has the same distribution as $\mu + \mathbf{F}\mathbf{e}$ where $\mathbf{F}_{p\times p}$ is a matrix of rank $r$, $\mathbf{\Sigma} = \mathbf{F}\mathbf{F}'$ and each element of $\mathbf{e}$ is distributed: $e_i \sim N(0,1)$. The density of $\mathbf{y}$ is given by

$$f(\mathbf{y}) = (2\pi)^{-p/2} |\mathbf{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{y}-\mu)\right] \tag{1.14}$$

For a random sample of $n$ independent $p$-vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ from a multivariate normal distribution, $\mathbf{y}_i \sim \mathrm{IN}_p(\mu, \mathbf{\Sigma})$, we shall in general write the data matrix $\mathbf{Y}_{n\times p}$ in the classical form

$$\mathbf{Y}_{n\times p} \equiv \begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix} \tag{1.15}$$

The corresponding responsewise representation for $\mathbf{Y}$ is

$$\mathbf{Y}_{p\times n} \equiv (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n) = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ y_{p1} & y_{p2} & \cdots & y_{pn} \end{pmatrix} \tag{1.16}$$

The joint probability density function ($pdf$) for $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ or the likelihood function is

$$L = L(\mu, \mathbf{\Sigma} | \mathbf{y}) = \prod_{i=1}^{n} f(\mathbf{y}_i) \tag{1.17}$$

Substituting $f(\mathbf{y})$ in (1.14) for each $f(\mathbf{y}_i)$, the *pdf* is

$$[(2\pi)^p|\boldsymbol{\Sigma}|]^{-n/2}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i-\mu)'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\mu)\right] \tag{1.18}$$

Using the property of the trace of a matrix, $Tr(\mathbf{x}'\mathbf{A}\mathbf{x}) = Tr(\mathbf{A}\mathbf{x}\mathbf{x}')$, (1.18) becomes

$$[(2\pi)^p|\boldsymbol{\Sigma}|]^{-n/2}\,\text{etr}\left\{-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\left[\sum_{i=1}^{n}(\mathbf{y}_i-\mu)(\mathbf{y}_i-\mu)'\right]\right\} \tag{1.19}$$

where etr stands for the exponential of the trace of a matrix.

If we let the sample mean be represented by

$$\bar{\mathbf{y}} = n^{-1}\sum_{i=1}^{n}\mathbf{y}_i \tag{1.20}$$

then the sum of squares and products (*SSP*) matrix, using the classical form (1.15), is

$$\mathbf{E} = \sum_{i=1}^{n}(\mathbf{y}_i-\bar{\mathbf{y}})(\mathbf{y}_i-\bar{\mathbf{y}})' = \mathbf{Y}'\mathbf{Y} - n\bar{\mathbf{y}}\bar{\mathbf{y}}' \tag{1.21}$$

or if we use the responsewise form (1.16), then the SSP matrix is

$$\mathbf{E} = \mathbf{Y}\mathbf{Y}' - n\bar{\mathbf{y}}\,\bar{\mathbf{y}}' \tag{1.22}$$

In either case, we may write (1.19)

$$\left[(2\pi)^p|\boldsymbol{\Sigma}|\right]^{-n/2}\text{etr}\left\{-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\left[\mathbf{E}+n(\bar{\mathbf{y}}-\mu)(\bar{\mathbf{y}}-\mu)'\right]\right\} \tag{1.23}$$

so that by Neyman's factorization criterion $(\mathbf{E}, \bar{\mathbf{y}})$ are sufficient statistics for estimating $(\boldsymbol{\Sigma}, \mu)$, (see, e.g., Lehmann, 1994, p. 16).

**Theorem 1.1** Let $\mathbf{y}_i \sim IN_p(\mu, \boldsymbol{\Sigma})$ of size $n$. Then $\bar{\mathbf{y}}$ and $\mathbf{E}$ are sufficient statistics for $\mu$ and $\boldsymbol{\Sigma}$.

It can also be shown that $\bar{\mathbf{y}}$ and $\mathbf{E}$ are independently distributed. The distribution of $\mathbf{E}$ is known as the Wishart distribution, a multivariate generalization of the chi-square distribution, with $\nu = n-1$ degrees of freedom. The density of the Wishart distribution is

$$c\,|\boldsymbol{\Sigma}|^{-\nu/2}|\mathbf{E}|^{(\nu-p-1)/2}\,\text{etr}\left(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{E}\right) \tag{1.24}$$

where $c$ is an appropriately chosen constant so that the probability over the entire sample space is equal to one. We write this as $\mathbf{E} \sim W_p(\nu, \boldsymbol{\Sigma})$. The expectation of $\mathbf{E}$ is $\nu\boldsymbol{\Sigma}$.

Given a random sample of observations from a multivariate normal distribution, we usually estimate the parameters $\mu$ and $\boldsymbol{\Sigma}$.

**Theorem 1.2** Let $\mathbf{y}_i \sim IN_p(\mu, \boldsymbol{\Sigma})$, then the maximum likelihood estimators of $\mu$ and $\boldsymbol{\Sigma}$ are $\bar{\mathbf{y}}$ and $\mathbf{E}/n = \hat{\boldsymbol{\Sigma}}$.

Furthermore, $\bar{\mathbf{y}}$ and

$$\mathbf{S} = \left(\frac{n}{n-1}\right)\hat{\boldsymbol{\Sigma}} = \mathbf{E}/(n-1) \tag{1.25}$$

are unbiased estimators of $\mu$ and $\boldsymbol{\Sigma}$, respectively, so that $E(\bar{\mathbf{y}}) = \mu$ and $E(\mathbf{S}) = \boldsymbol{\Sigma}$. Hence, the sample distribution of $\mathbf{S}$ is Wishart. $(n-1)\mathbf{S} \sim W_p(v = n-1, \boldsymbol{\Sigma})$ or $\mathbf{S} \sim W_p[n-1, \boldsymbol{\Sigma}/(n-1)]$.

Linear combinations of multivariate normal random variables are again normally distributed. If $\hat{\beta}$ is an unbiased estimate of $\beta$ in (1.5) such that $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ then

$$\hat{\beta} \sim N_k[\beta, \operatorname{cov}\hat{\beta}] \tag{1.26}$$

so that $\hat{\beta}$ has a multivariate normal distribution with covariance structure $\boldsymbol{\Psi} = \operatorname{cov}\hat{\beta}$ when one assumes that $\mathbf{y}$ is multivariate normal, $\mathbf{y} \sim N_n(\mathbf{X}\beta = \mu, \boldsymbol{\Omega} = \boldsymbol{\Sigma})$.

## 1.5 Hypothesis Testing

Having assumed a linear model for a random sample of observations, used the observations to obtain estimates of the population parameters, and decided upon the structure of the restriction $\mathbf{R}$ (if any) and the hypothesis test matrix $\mathbf{C}$, one next tests hypotheses. Two commonly used procedures for testing hypotheses are the likelihood ratio (LR) and union-intersection (UI) tests. To construct an LR test, two likelihood functions are compared for a random sample of observations, $L(\hat{\omega})$, the likelihood function maximized under the hypothesis $H$ in (1.8), and the likelihood $L(\hat{\boldsymbol{\Omega}}_o)$, the likelihood function maximized over the entire parameter space $\boldsymbol{\Omega}_o$ unconstrained by the hypothesis. Defining $\lambda$ as the ratio

$$\lambda = \frac{L(\hat{\omega})}{L(\hat{\boldsymbol{\Omega}}_o)} \tag{1.27}$$

the hypothesis is rejected for small values of $\lambda$ since $L(\hat{\omega}) < L(\hat{\boldsymbol{\Omega}}_o)$ does not favor the hypothesis. The test is said to be of size $\alpha$ if, for a constant $\lambda_o$, the

$$P(\lambda < \lambda_o \mid H) = \alpha \tag{1.28}$$

where $\alpha$ is the size of the type I error rate, the probability of rejecting $H$ given $H$ is true. For large sample sizes and under very general conditions, Wald (1943) showed that $-2 \ln \lambda$ converges in distribution to a chi-square distribution as $n \to \infty$, where the degrees of freedom $v$ is equal to the number of independent parameters estimated under $\boldsymbol{\Omega}_o$ minus the number of independent parameters estimated under $\omega$.

To construct a UI test according to Roy (1953), we write the null hypothesis $H$ as an intersection of an infinite number of elementary tests

$$H: \bigcap_i H_i \tag{1.29}$$

and each $H_i$ is associated with an alternative $A_i$ such that

$$A: \bigcup_i A_i \tag{1.30}$$

The null hypothesis $H$ is rejected if any elementary test of size $\alpha$ is rejected. The overall rejection region is defined as the union of all the rejection regions of the elementary tests of $H_i$ vs. $A_i$. Similarly, the region of acceptance for $H$ is the intersection of the acceptance regions. If $T_i$ is a test statistic for testing $H_i$ vs. $A_i$, the null hypothesis $H$ is accepted or rejected if the $T_i \lesseqgtr T_\alpha$ where the

$$P(\sup_i T_i \leq T_\alpha | H) = 1 - \alpha \tag{1.31}$$

and $T_\alpha$ is chosen such that the type I error is $\alpha$.

## 1.6 Generating Multivariate Normal Data

In the hypothesis testing of both univariate and multivariate linear models, the assumption of normally distributed errors is made. The multivariate normal distribution of $\mathbf{y}$ has density given by (1.14), written $\mathbf{y} \sim N_p(\mu, \Sigma)$. When p = 1, (1.14) reduces to a univariate normal distribution (1.13).

Some important properties of the multivariate normal distribution include

(1) If $\mathbf{y}$ is distributed multivariate normal, then each subset of the components of $\mathbf{y}$ has a normal distribution. Thus, each $y_i$ has a univariate normal distribution. However, the converse is not true. If each $y_i$ has a normal distribution, this does not imply that $\mathbf{y}$ has a multivariate normal distribution.

(2) All linear combinations of the $y_i$ are normally distributed.

(3) The conditional distributions of the components of $\mathbf{y}$ are normally distributed, Timm (1975, p. 114-123).

To generate data having a multivariate distribution with mean and covariance matrix

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \sigma_{p1} & \cdots & & \sigma_{pp} \end{bmatrix} \tag{1.32}$$

we use Definition 1.1. Program 1_6.sas uses the IML procedure to generate 50 observations from a multivariate normal distribution with

$$\mu = \begin{bmatrix} 10 \\ 20 \\ 30 \\ 40 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} 3 & 1 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 4 & 20 \end{bmatrix}$$

### Program 1_6.sas

```
/* Program 1_6.sas */
/* Program to create a multivariate normal data set */

options ls=80 ps=60 nodate nonumber;
filename appl 'c:\1_6.dat';
title1 'Output 1.6: Generating a Multivariate Normal Data Set';
```

```
proc iml;
   seed=30195;
   z=normal(repeat(seed,50,4));
   u={10,20,30,40};
   s={3 1 0  0,
      1 4 0  0,
      0 0 1  4,
      0 0 4 20};
   a=root(s);
   uu=repeat(u`,50,1);
   y=(z*a) + uu;
   print y;

file appl;
   do i=1 to nrow(y);
      do j=1 to ncol(y);
        put (y[i,j]) 10.2 +2 @;
      end;
     put;
   end;
closefile appl;
quit;
```

In Program 1_6.sas, PROC IML is used to produce a matrix $\mathbf{Z}$ that contains $n = 50$ row vectors each with $p = 4$ observations. Each is generated from a standard normal distribution $N(0,1)$. Next, new variables $\mathbf{y}_i = \mathbf{z}_i\mathbf{A} + \mu$, must be created where $\mathbf{A}$ is such that $\text{cov}(\mathbf{y}) = \mathbf{A}'\,\text{cov}(\mathbf{z})\mathbf{A} = \mathbf{A}'\,\mathbf{I}\mathbf{A} = \mathbf{A}'\,\mathbf{A} = \mathbf{\Sigma}$ and $E(\mathbf{y}) = \mu$. The Cholesky factorization procedure can be used to obtain $\mathbf{A}$ from $\mathbf{\Sigma}$; the ROOT function of PROC IML is used to perform the Cholesky decomposition and produce the resultant matrix we named $a$. Next, the matrix $uu$ is created by repeating the row vector $u$ 50 times to produce a $50 \times 4$ matrix. Finally, the multivariate normal random variables are created in the statement $y = (z * a) + uu$. The observations are printed and then output to the file named 1_6.dat.

## Result and Interpretation 1_6.sas

This is the output of Program 1_6.sas. The data are saved in the file 1_6.dat.

**Output 1.6:** *Generating a Multivariate Normal Data Set*

Output 1.6: Generating a Multivariate Normal Data Set

| Y | | | |
|---|---|---|---|
| 10.277241 | 18.515521 | 29.670682 | 35.682188 |
| 11.929652 | 20.275932 | 28.911167 | 36.047376 |
| 8.2156513 | 17.403286 | 28.721552 | 35.264579 |
| 8.4058699 | 21.693912 | 32.140953 | 49.366754 |
| 10.034067 | 19.206552 | 31.13305 | 41.664014 |
| 12.852388 | 20.338035 | 29.292826 | 41.662329 |
| 7.7777858 | 19.422314 | 30.000313 | 34.771545 |
| 10.551284 | 17.678773 | 31.285126 | 44.347948 |
| 10.105934 | 21.405632 | 29.357324 | 35.774756 |
| 12.249358 | 21.003614 | 30.638821 | 44.324676 |
| 8.7131087 | 23.276061 | 30.084822 | 40.906227 |
| 12.086015 | 18.060828 | 30.285533 | 41.081205 |
| 9.7895428 | 19.771547 | 32.692761 | 49.95753 |
| 11.620772 | 19.016928 | 28.224879 | 32.90378 |
| 8.9938978 | 20.103495 | 29.681585 | 39.858339 |
| 10.863083 | 20.553824 | 31.148947 | 45.86955 |

**Output 1.6** (*continued*)

```
12.484503 19.623805  29.90136 38.509333
11.402297 22.778862 28.018182 31.069522
11.821743 22.568622 29.524165 36.428279
11.318856 20.383911 29.374577 35.644656
11.381783 21.328723 31.799191 44.314454
13.771969 20.979002 30.335263 38.938178
9.5884659 22.248491 29.881163 41.837656
8.6288368 16.496882 29.921442 40.953565
11.304779  20.34674 29.286882  34.35696
12.304391 18.544842 29.169465 36.569666
13.019728 22.288911 28.364377 35.729171
 8.877829 19.951443 30.726531 44.400961
12.226032 20.373853 30.003996 43.396151
 8.532775 18.091436 32.765839 51.091007
8.8049882 20.945813 29.453855 34.251643
8.3840127 19.969099  29.94458 41.713592
9.5339415 20.027807 29.632425 39.880545
11.233662 19.164775 29.175422 38.209971
10.407781 21.653868 31.753313 49.197686
12.102968  21.75833 29.168331  38.05888
12.228035  19.13339 29.878038 36.567964
9.6612809  21.85244  29.61101 37.617794
9.8002753 23.859038  29.01271 38.490044
 9.138417 18.084905 28.812975 36.371415
11.259012 22.395983 29.811016 39.231856
7.8569112 21.512452 30.526549 40.452245
8.8971645 18.458986 30.576723 43.875282
7.8857517 21.186746 28.030963 34.526299
11.684868 23.639538 30.187498 40.505556
9.7260651 20.968905 28.914758 38.067146
 11.66679 23.005702 28.672645 30.684908
7.9976197   20.6001 29.810047  39.05618
10.134748 19.093801 29.053383 34.423978
 11.74687   20.7355 30.442122 43.669143
```

## 1.7 Assessing Univariate Normality with Q-Q Plots

Before we perform hypothesis tests on data, we should examine the distributional assumptions. Hypothesis testing in the presence of assumption violations may result in errors in statistical inference.

By the properties of the multivariate normal distribution of $\mathbf{y}$, the components $y_i$ are univariate normally distributed. Thus, one step in evaluating the multivariate normality assumption of $\mathbf{y}$ is to evaluate the univariate normality of each $y_i$. One can construct and examine histograms, stem-and-leaf plots, box plots, and Quantile-Quantile (Q-Q) probability plots.

Q-Q plots are plots of the observed, ordered quantile versus the quantile values expected if the observed data are normally distributed. Departures from a straight line are evidence against the assumption that the population from which the observations are drawn is normally distributed. Outliers may be detected from these plots as points well separated from the other observations. The behavior at the ends of the plots can provide information about the length of the tails of the distribution, and the symmetry or asymmetry of the distribution (Singh, 1993).

### 1.7.1 Normally Distributed Data

Program 1_7_1.sas produces a Q-Q plot for the univariate normally distributed random variable $y1$ from the data set generated by Program 1_6.sas.

#### Program 1_7_1.sas

```
/* Program 1_7_1.sas */
/* Program to create Q-Q plot of data */

options ls=80 ps=60 nodate nonumber;
filename appl 'c:\ 1_6.dat';
title1 'Output 1.7.1: Q-Q plot of 1.6 Data (y1)';

data ex171;
   infile appl;
   input y1-y4;
proc sort;
   by y1;
proc univariate noprint;
   var y1;
   output out=stats n=nobs mean=mean std=std;
data quantile;
   set ex171;
   if _n_=1 then set stats;
   i=1;
   p=(i - .5) / nobs;
   z=probit(p);
   normal = mean + z*std;
proc print;
proc corr;
   var y1 z;
run;

filename out 'c:\ 1_7_1.cgm';
goptions device=cgmmwwc gsfname=out gsfmode=replace
   colors=(black) hsize=5in vsize=4in;
proc gplot data=quantile;
   plot y1*z normal*z /overlay frame;
   symbol1 v=;
   symbol2 i=join v=none l=1;
run;
```

First the DATA step reads in the data and then PROC SORT sorts the values of $y1$ in ascending order. PROC UNIVARIATE outputs the number of observations (nobs), the mean of $y1$ (MEAN), and the standard deviation of $y1$ (STD). In the next DATA step the observed ordered quantiles are defined (P), and then the PROBIT function defines the quantiles expected if $z$ is a normally distributed random variable with mean 0 and standard deviation of 1. Next the variable NORMAL is created so it can be compared to $y1$ in the plots. A filename for the plot is specified as well as options for the graphics plot. The option DEVICE = CGMMWWC specifies a cgm plot that can be accessed with Microsoft Word software. PROC GPLOT overlays the plots of $y1$ versus $z$ and normal versus $z$ where the plot of normal versus $z$ is represented as the straight line (since we specified $I$ = JOIN). The plotting symbol is specified by the v = option; when no value is specified, the default symbol is used.

## Result and Interpretation 1_7_2.sas

Output 1.7.1 contains results of Program 1_7_1.sas and the resulting Q-Q plot.

**Output 1.7.1:** *Q-Q plot of 1.6 Data (y1)*

Output 1.7.1: Q-Q plot of 1.6 Data (y1)

| OBS | Y1 | Y2 | Y3 | Y4 | NOBS | MEAN | STD | I | P | Z | NORMAL |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 7.78 | 19.42 | 30.00 | 34.77 | 50 | 10.4254 | 1.60353 | 1 | 0.01 | -2.32635 | 6.6950 |
| 2 | 7.86 | 21.51 | 30.53 | 40.45 | 50 | 10.4254 | 1.60353 | 2 | 0.03 | -1.88079 | 7.4095 |
| 3 | 7.89 | 21.19 | 28.03 | 34.53 | 50 | 10.4254 | 1.60353 | 3 | 0.05 | -1.64485 | 7.7878 |
| 4 | 8.00 | 20.60 | 29.81 | 39.06 | 50 | 10.4254 | 1.60353 | 4 | 0.07 | -1.47579 | 8.0589 |
| 5 | 8.22 | 17.40 | 28.72 | 35.26 | 50 | 10.4254 | 1.60353 | 5 | 0.09 | -1.34076 | 8.2755 |
| 6 | 8.38 | 19.97 | 29.94 | 41.71 | 50 | 10.4254 | 1.60353 | 6 | 0.11 | -1.22653 | 8.4586 |
| 7 | 8.41 | 21.69 | 32.14 | 49.37 | 50 | 10.4254 | 1.60353 | 7 | 0.13 | -1.12639 | 8.6192 |
| 8 | 8.53 | 18.09 | 32.77 | 51.09 | 50 | 10.4254 | 1.60353 | 8 | 0.15 | -1.03643 | 8.7634 |
| 9 | 8.63 | 16.50 | 29.92 | 40.95 | 50 | 10.4254 | 1.60353 | 9 | 0.17 | -0.95417 | 8.8954 |
| 10 | 8.71 | 23.28 | 30.08 | 40.91 | 50 | 10.4254 | 1.60353 | 10 | 0.19 | -0.87790 | 9.0177 |
| 11 | 8.80 | 20.95 | 29.45 | 34.25 | 50 | 10.4254 | 1.60353 | 11 | 0.21 | -0.80642 | 9.1323 |
| 12 | 8.88 | 19.95 | 30.73 | 44.40 | 50 | 10.4254 | 1.60353 | 12 | 0.23 | -0.73885 | 9.2406 |
| 13 | 8.90 | 18.45 | 30.58 | 43.88 | 50 | 10.4254 | 1.60353 | 13 | 0.25 | -0.67449 | 9.3438 |
| 14 | 8.99 | 20.10 | 29.68 | 39.86 | 50 | 10.4254 | 1.60353 | 14 | 0.27 | -0.61281 | 9.4427 |
| 15 | 9.14 | 18.05 | 28.81 | 36.37 | 50 | 10.4254 | 1.60353 | 15 | 0.29 | -0.55338 | 9.5380 |
| 16 | 9.53 | 20.03 | 29.63 | 39.88 | 50 | 10.4254 | 1.60353 | 16 | 0.31 | -0.49585 | 9.6303 |
| 17 | 9.59 | 22.25 | 29.88 | 41.84 | 50 | 10.4254 | 1.60353 | 17 | 0.33 | -0.43991 | 9.7200 |
| 18 | 9.66 | 21.85 | 29.61 | 37.62 | 50 | 10.4254 | 1.60353 | 18 | 0.35 | -0.38532 | 9.8075 |
| 19 | 9.73 | 20.97 | 28.91 | 38.07 | 50 | 10.4254 | 1.60353 | 19 | 0.37 | -0.33185 | 9.8933 |
| 20 | 9.79 | 19.77 | 32.69 | 49.96 | 50 | 10.4254 | 1.60353 | 20 | 0.39 | -0.27932 | 9.9775 |
| 21 | 9.80 | 23.86 | 29.01 | 38.49 | 50 | 10.4254 | 1.60353 | 21 | 0.41 | -0.22754 | 10.0605 |
| 22 | 10.03 | 19.21 | 31.13 | 41.66 | 50 | 10.4254 | 1.60353 | 22 | 0.43 | -0.17637 | 10.1426 |
| 23 | 10.11 | 21.41 | 29.36 | 35.77 | 50 | 10.4254 | 1.60353 | 23 | 0.45 | -0.12566 | 10.2239 |
| 24 | 10.13 | 19.09 | 29.05 | 34.42 | 50 | 10.4254 | 1.60353 | 24 | 0.47 | -0.07527 | 10.3047 |
| 25 | 10.28 | 18.52 | 29.67 | 35.68 | 50 | 10.4254 | 1.60353 | 25 | 0.49 | -0.02507 | 10.3852 |
| 26 | 10.41 | 21.65 | 31.75 | 49.20 | 50 | 10.4254 | 1.60353 | 26 | 0.51 | 0.02507 | 10.4656 |
| 27 | 10.55 | 17.68 | 31.29 | 44.35 | 50 | 10.4254 | 1.60353 | 27 | 0.53 | 0.07527 | 10.5461 |
| 28 | 10.86 | 20.55 | 31.15 | 45.87 | 50 | 10.4254 | 1.60353 | 28 | 0.55 | 0.12566 | 10.6269 |
| 29 | 11.23 | 19.26 | 29.18 | 38.21 | 50 | 10.4254 | 1.60353 | 29 | 0.57 | 0.17637 | 10.7082 |
| 30 | 11.26 | 22.40 | 29.81 | 39.23 | 50 | 10.4254 | 1.60353 | 30 | 0.59 | 0.22754 | 10.7903 |
| 31 | 11.30 | 20.35 | 29.29 | 34.36 | 50 | 10.4254 | 1.60353 | 31 | 0.61 | 0.27932 | 10.8733 |
| 32 | 11.32 | 20.35 | 29.37 | 35.64 | 50 | 10.4254 | 1.60353 | 32 | 0.63 | 0.33185 | 10.9575 |
| 33 | 11.38 | 21.33 | 31.80 | 44.31 | 50 | 10.4254 | 1.60353 | 33 | 0.65 | 0.38532 | 11.0433 |
| 34 | 11.40 | 22.78 | 28.02 | 31.07 | 50 | 10.4254 | 1.60353 | 34 | 0.67 | 0.43991 | 11.1308 |
| 35 | 11.62 | 19.02 | 28.22 | 32.90 | 50 | 10.4254 | 1.60353 | 35 | 0.69 | 0.49585 | 11.2205 |
| 36 | 11.67 | 23.01 | 28.67 | 30.68 | 50 | 10.4254 | 1.60353 | 36 | 0.71 | 0.55338 | 11.3128 |
| 37 | 11.68 | 23.64 | 30.19 | 40.51 | 50 | 10.4254 | 1.60353 | 37 | 0.73 | 0.61281 | 11.4081 |
| 38 | 11.75 | 20.74 | 30.44 | 43.67 | 50 | 10.4254 | 1.60353 | 38 | 0.75 | 0.67449 | 11.5070 |
| 39 | 11.82 | 22.57 | 29.52 | 36.43 | 50 | 10.4254 | 1.60353 | 39 | 0.77 | 0.73885 | 11.6102 |
| 40 | 11.93 | 20.28 | 28.91 | 36.05 | 50 | 10.4254 | 1.60353 | 40 | 0.79 | 0.80642 | 11.7185 |
| 41 | 12.09 | 18.06 | 30.29 | 41.08 | 50 | 10.4254 | 1.60353 | 41 | 0.81 | 0.87790 | 11.8331 |
| 42 | 12.10 | 21.76 | 29.17 | 38.06 | 50 | 10.4254 | 1.60353 | 42 | 0.83 | 0.95417 | 11.9554 |
| 43 | 12.23 | 20.37 | 30.00 | 43.40 | 50 | 10.4254 | 1.60353 | 43 | 0.85 | 1.03643 | 12.0874 |
| 44 | 12.23 | 19.13 | 29.88 | 36.57 | 50 | 10.4254 | 1.60353 | 44 | 0.87 | 1.12639 | 12.2316 |
| 45 | 12.25 | 21.00 | 30.64 | 44.32 | 50 | 10.4254 | 1.60353 | 45 | 0.89 | 1.22653 | 12.3922 |
| 46 | 12.30 | 18.54 | 29.17 | 36.57 | 50 | 10.4254 | 1.60353 | 46 | 0.91 | 1.34076 | 12.5753 |
| 47 | 12.48 | 19.62 | 29.90 | 38.51 | 50 | 10.4254 | 1.60353 | 47 | 0.93 | 1.47579 | 12.7919 |
| 48 | 12.85 | 20.34 | 29.29 | 41.66 | 50 | 10.4254 | 1.60353 | 48 | 0.95 | 1.64485 | 13.0630 |
| 49 | 13.02 | 22.29 | 28.36 | 35.73 | 50 | 10.4254 | 1.60353 | 49 | 0.97 | 1.88079 | 13.4413 |
| 50 | 13.77 | 20.93 | 30.34 | 38.94 | 50 | 10.4254 | 1.60353 | 50 | 0.99 | 2.32635 | 14.1558 |

Output 1.7.1: Q-Q plot of 1.6 Data (y1)

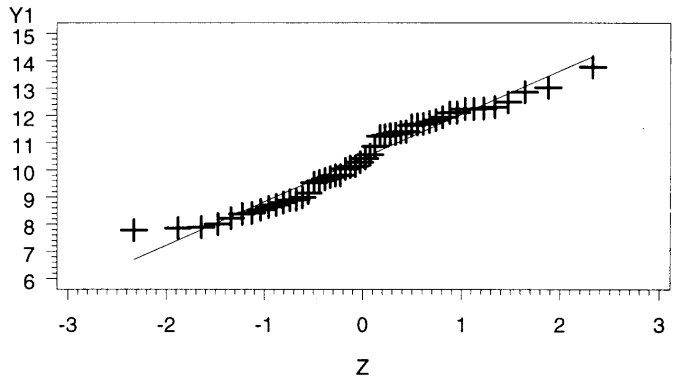
Correlation Analysis

2 'VAR' Variables: Y1 Z

**Output 1.7.1** (*continued*)

```
                              Simple Statistics

Variable           N        Mean     Std Dev         Sum     Minimum     Maximum

Y1                50    10.42540     1.60353   521.27000     7.78000    13.77000
Z                 50           0     0.99740           0    -2.32635     2.32635


    Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 50

                                        Y1                Z

                      Y1           1.00000           0.98033
                                    0.0               0.0001

                      Z            0.98033           1.00000
                                    0.0001            0.0
```

# Output 1.7.1: Q–Q plot of 1.6 Data (y1)




The plot shows that the observations lie close to the line, but not exactly; the tails, especially, fall off from the line. Recall that we know that these data are a random sample from a normal distribution. Thus, when using these plots for diagnostic purposes, we cannot expect that even normally distributed data will lie exactly on a straight line.

## 1.7.2 Nonnormally Distributed Data

The following example presents a Q-Q plot in which the data are not normally distributed: $y1$ is transformed by $1/y^2$ and plotted. Program 1_7_2.sas contains the SAS code.

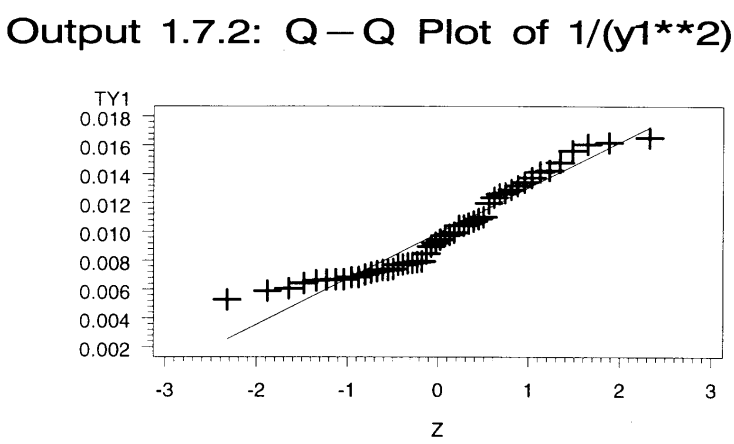
### Program 1_7_2.sas

```
/* Program 1_7_2.sas */
/* Program to create Q-Q plot of 1/(y1**2) Data */
```

```
options ls=80 ps=60 nodate nonumber;
filename appl 'c:\1_6.dat';
title1 'Output 1.7.2: Q-Q Plot of 1/(y1**2)';

data ex172;
   infile appl;
   input y1-y4;
   ty1=1/(y1**2);
proc sort;
   by ty1;
proc univariate noprint;
   var ty1;
   output out=stats n=nobs mean=mean std=std;
data quantile;
   set ex172;
   if _n_=1 then set stats;
   i+1;
   p=(i - .5) / nobs;
   z=probit(p);
   normal = mean + z*std;
proc print;
proc corr;
   var ty1 z;
run;

filename out 'c:\1_7_2.cgm';
goptions device=cgmmwwc gsfname=out gsfmode=replace
     colors=(black) hsize=5in vsize=4in;

proc gplot data=quantile;
   plot ty1*z normal*z /overlay frame;
   symbol1 v=;
   symbol2 i=join v=none l=1;
run;
```

The program is the same as Program 1_7_1.sas with the exception that the variable *ty*1 is created. This transformed variable is then used for plotting.

### Result and Interpretation 1_7_2.sas




Observe that the observations show a marked curvilinear pattern and are not close to the line.

## 1.7.3 Outliers

The following example presents a Q-Q plot with an outlier. Program 1_7_3.sas is used to create an outlier in $y_1$ and to plot the data.

### Program 1_7_3.sas

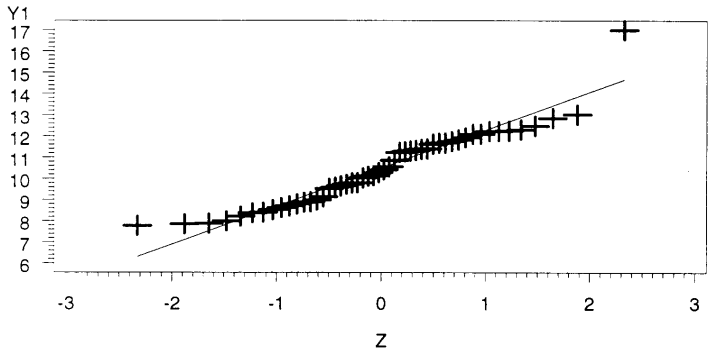
```
/* Program 1_7_3.sas */
/* Program to create Q-Q plot of y1 data with Outlier */

options ls=80 ps=60 nodate nonumber;
filename appl 'c:\ 1_6.dat';
titlel 'Output 1.7.3: Q-Q plot of y1 data with an Outlier';

data ex173;
   infile appl;
   input y1-y4;
   if y1 ge 13.7 then y1 = 17;
proc sort;
   by y1;
proc univariate noprint;
   var y1;
   output out=stats n=nobs mean=mean std=std;
data quantile;
   set ex173;
   if _n_=1 then set stats;
   i+1;
   p=(i - .5) / nobs;
   z=probit(p);
   normal = mean + z*std;
proc print;
proc corr;
   var y1 z;
run;
```

```
filename out 'c:\ 1_7_3.cgm';
goptions device=cgmmwwc gsfname=out gsfmode=replace
   colors=(black) hsize=5in vsize=4in;

proc gplot data=quantile;
   plot y1*z normal*z /overlay frame;
   symbol1 v=;
   symbol2 i=join v=none l=1;
run;
```

Again, the program is the same as Program 1_7_1.sas with exception that the $y1$ value of 13.77 is changed to the value of 17 by using an IF-THEN statement in the DATA step.

### Result and Interpretation 1_7_3.sas

**Output 1.7.3: Q−Q plot of y1 data with an Outlier**




The extreme observation lying far from the line at the top of the plot indicates an outlier.

## 1.7.4 Real Data Example

To illustrate the application of Q-Q plots utilizing real data, we use the data from Rhower given in Timm (1975, p. 281 and p. 345). Rhower was interested in predicting the performance on three standardized tests (Peabody Picture Vocabulary (Y1), Student Achievement Test (Y2), and the Raven Progressive Matrices Test (Y3)) given five paired-associate, learning-proficiency tasks (named (X1), still (X2), named still (X3), named action (X4) and sentence still (X5)) for 32 randomly selected school children in an upper-class, white residential school.

A data set of the residuals resulting after the three dependent variables were regressed on the five independent variables is contained in the file named ycondx.dat. Program 1_7_4.sas produces three Q-Q plots for the residuals of the three dependent variables and then to produce five Q-Q plots of the independent variables.

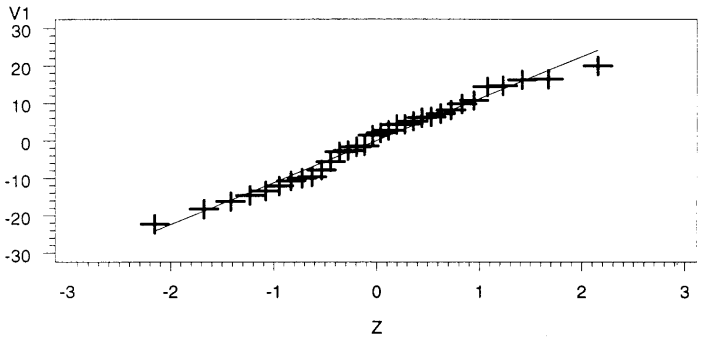
## Program 1_7_4.sas

```
/* Program 1_7_4.sas                                    */
/* Program to create Q-Q plot of a dataset              */
/* To run this program on your own dataset change       */
/* the name of the file in the file=_____statement      */
/* and the number of columns in the p=___ statement */

options ls=80 ps=60 nodate nonumber;

%let file = ycondx.dat;
%let p = 3;

  /* macro to expand the string of variables that are processed */
%macro expand(cols);
   %do j=1 %to &cols;
      v&j
   %end;
%mend expand;

  /* macro to perform Q-Q plotting of the variables    */
%macro qq(cols);
   %do i=1 %to &cols;
proc sort data=ex174;
   by v&i;
proc univariate noprint data=ex174;
   var v&i;
   output out=stats n=nobs mean=mean std=std;
data quantile;
   set ex174;
   if _n_=1 then set stats;
   k+1;
   pr=(k - .5) / nobs;
   z=probit(pr);
   normal = mean + z*std;
proc print data=quantile;
   title "Output 1.7.4: Q-Q plot, variable V&i, &file";
proc corr data=quantile;
   var v&i z;

filename out "1_7_4_&i'.cgm";
goptions device=cgmmwwc gsfname=out gsfmode=replace
   colors=(black) hsize=5in vsize=4in;

proc gplot data=quantile;
   title "Output 1.7.4: Q-Q plot, variable V&i, &file";
   plot v&i*z normal*z /overlay frame;
   symbol1 v=;
   symbol2 i=join v=none l=1;
run;
%end;
%mend qq;

data ex174;
   infile "&file";
   input %expand(&p);
proc print data=ex174;
   title "Output 1.7.4: &file";
%qq(&p);
```

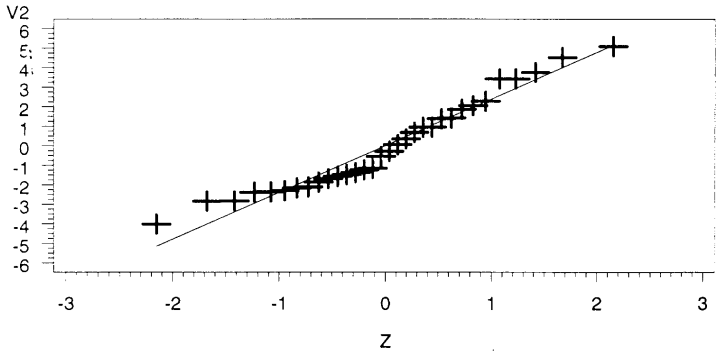
This program produces one Q-Q plot for each column of data in the data file. The name of the data file is contained in the statement `%let file =` and the number of columns must be specified in the statement `%let p=`. Here, to produce the output that follows we first ran the program with `file = ycondx.dat` and $p$=3 and then we ran it with `file=rhowerx.dat` and $p$=5. To run this program on your own data set, simply change the name of the data file and the number of columns to correspond to your data set. Q-Q plots are output to files named 1_7_4&*i*.cgm where the &*i* is equal to 1 for the first column of data, 2 for the second column, and so on.

### Result and Interpretation 1_7_4.sas

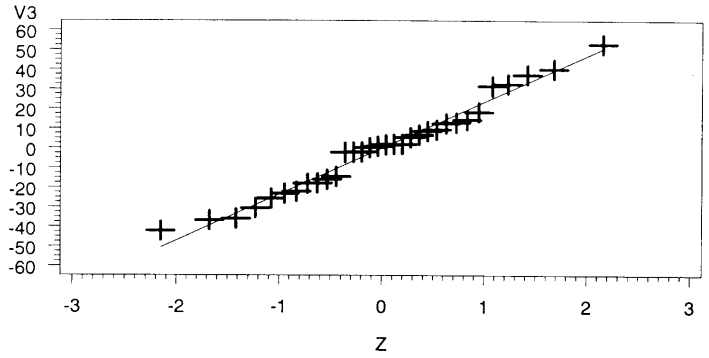
Output 1.7.4: Q – Q plot, variable V1, ycondx.dat



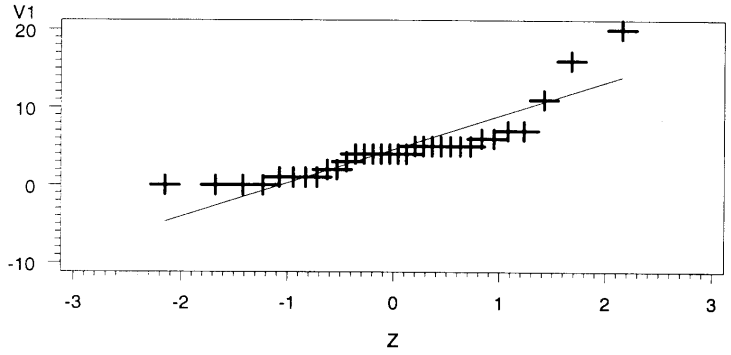

Output 1.7.4: Q – Q plot, variable V2, ycondx.dat



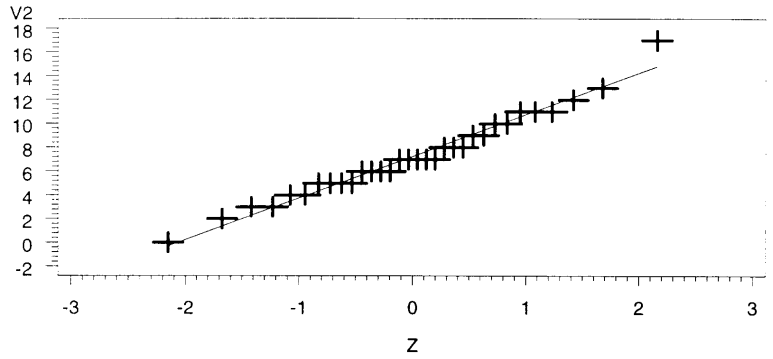
Output 1.7.4: Q – Q plot, variable V3, ycondx.dat



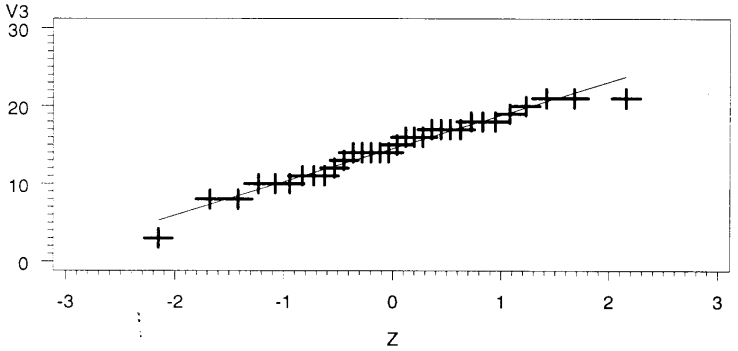

Output 1.7.4: Q – Q plot, variable V1, rhowerx.dat



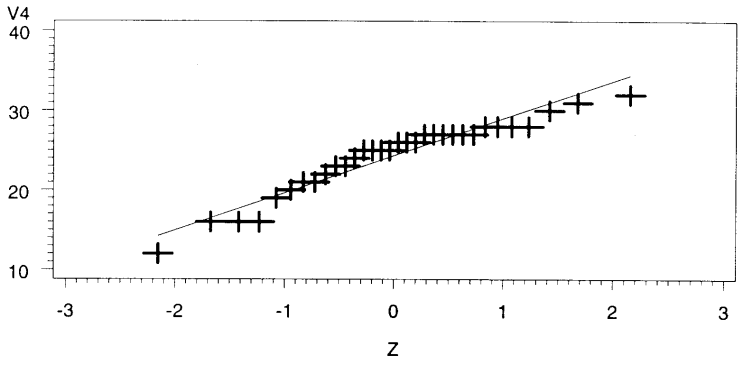
Output 1.7.4: Q – Q plot, variable V2, rhowerx.dat



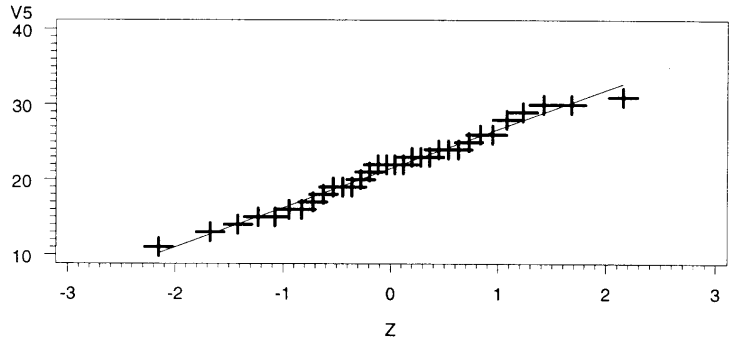

Output 1.7.4: Q – Q plot, variable V3, rhowerx.dat



**Output 1.7.4: Q – Q plot, variable V4, rhowerx.dat**




**Output 1.7.4: Q – Q plot, variable V5, rhowerx.dat**




From the plots generated using the ycondx.dat file we see that the residuals for variables Y1 and Y3 appear to be univariate normally distributed, but that Y2 does not appear to satisfy the assumption of normality. With a sample size as small as 32, however, it is difficult to be certain. In general, for Q-Q plots a sample size of at least 50 observations is desirable.

The plots of the independent variables using the file rhowerx.dat appear to reveal a possibility of several outliers for variable 1, a possible outlier for variables 2 and 3, and a curved pattern for variable 4. Variable 5 appears to lie close to the line. With a sample size of only 32 it is difficult to be certain if the plots show departures from normality.

## 1.8 Assessing Multivariate Normality with Chi-Square Plots

Recall that marginal normality does not ensure multivariate normality. To evaluate multivariate normality, one may compute the Mahalanobis distance for the $i^{th}$ observation

$$D_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \tag{1.33}$$

and plot these distances against the ordered chi-square percentile value $\chi_p^2[(i-(1/2)/n)]$. If the data are multivariate normal, the plotted pairs should be close to a straight line. As with Q-Q plots, points far from the line may be multivariate outliers.

Singh (1993) constructed probability plots resembling Shewart-type control charts, where warning points were placed at the $\alpha$ 100% critical value of the distribution of Mahalanobis distances, and a maximum point limit was also defined. Thus, any observation falling beyond the maximum limit was considered an outlier, and any point between the warning limit and the maximum limit required further investigation.

Singh (1993) constructed multivariate probability plots with the ordered Mahalanobis distances versus quantiles from a beta distribution, rather than the chi-square distribution. The exact distribution of $nD_i^2/(n-1)^2$ is in fact a beta distribution $B(p/2,(n-p-1)/2)$, Gnandesikan and Kettenring (1972). The chi-square distribution is only an approximation, which may not be good enough as $p$ gets large, Small (1978).

### 1.8.1 Normally Distributed Data

For an example of a chi-square plot of multivariate normally distributed data, see the chi-square plot of the multivariate normally distributed data, **y**, produced by Program 1_8_1.sas.

#### Program 1_8_1.sas

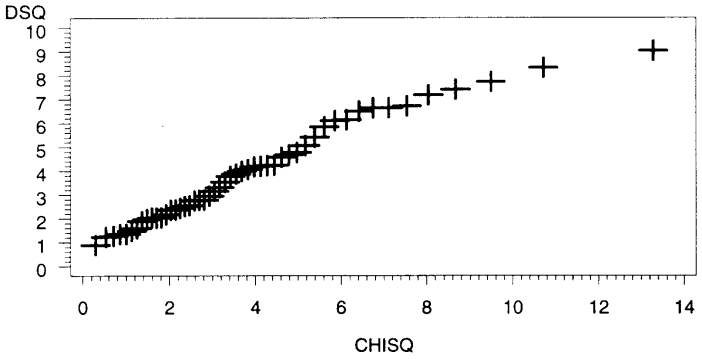
```
/* Program 1_8_1.sas                              */
/* Program to create Chi-Square Plot of Y Data    */
/* Data set from 1_6.sas is used, with a column   */
/* of observation numbers added to the file       */

options ls=80 ps=61 nodate nonumber;
filename app1 'c:  1_6.da2';
title1 'Output 1.8.1: Chi-Square Plot of the 1.6 Dataset';

data ex181;
   infile app1;
   input tag $ y1 - y4;
   label tag = 'id
      y1 = 'var1
      y2 = 'var2
      y3 = 'var3
      y4 = 'var4
   %let id=tag;
   %let var=y1 y2 y3 y4;
```

```
proc iml;
   reset;
   start dsquare;
      use _last_;
      read all var {&var} into y [colname=vars rowname=&id];
      n=nrow(y);
      p=ncol(y);
      r1=&id;
      print y;
      m=y[ :,];
      d=y - j(n,1) * m;
      s=d` * d / (n-1);
      dsq=vecdiag(d * inv(s) * d`);
      r=rank(dsq);
      val=dsq; dsq [r, ] = val;
      val=r1; &id [r] = val;
      z=((1:n)` - .5) / n;
      chisq = 2 * gaminv(z,p/2);
      result = dsq||chisq;
      cl={'dsq' 'chisq'};
      create dsquare from result [colname=cl rowname=&id];
      append from result [rowname=&id];
   finish;
run dsquare;
quit;
proc print data=dsquare;
   var tag dsq chisq;
run;

filename out 'c:\ 1_8_1.cgm';
goptions device=cgmmwwc gsfname=out gsfmode=replace
   colors=(black) hsize=5in vsize=4in;

proc gplot data=dsquare;
   plot dsq*chisq /frame;
   symbol1 v=;
run;
```

The data set used is the file output by Program 1_6.sas but with a column of observation numbers added as the first column; the new file is named 1_6.da2. Program 1_8_1.sas first reads the data with the DATA step and labels the variables. PROC IML then computes a vector of Mahalanobis distances, named dsq. Next, the ordered chi-square percentile values are computed, named chisq. The values are printed and then plotted with PROC GPLOT. The graphics file is output to the file 1_8_1.cgm.

### Result and Interpretation 1_8_1.sas

Output 1.8.1: Chi – Square Plot of the 1.6 Dataset




Upon examining the plot one can see that the observations lie somewhat close to a straight line, but not exactly, even for data known to be from a multivariate normal distribution. For multivariate chi-square plots, one would like to have approximately 100 observations.

## 1.8.2 Real Data Example

For an example of a chi-square plot using real data, see Program 1_8_2.sas, which produces a multivariate chi-square plot of the Rhower data (from Section 1.7.4). Namely, this is a plot of the residuals resulting after the three dependent variables were regressed on the five independent variables.

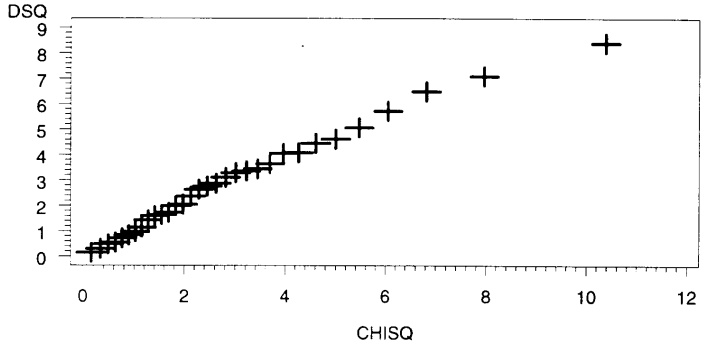
### Program 1_8_2.sas

```
/* Program 1_8_2.sas                    */
/* Program to create Chi-Square Plot */
/* of Residuals from Rhower data     */

options ls=80 ps=60 nodate nonumber;
filename rhower 'c:\ycondx.da2';
title1 'Output 1.8.2: Chi-Square Plot of Residuals';

data ex182;
   infile rhower;
   input tag $ yc1-yc3;
   label tag = 'id'
      yc1 = 'var1'
      yc2 = 'var2'
      yc3 = 'var3';
   %let id=tag;
   %let var=yc1 yc2 yc3;
proc iml;
   reset;
   start dsquare;
      use _last_;
      read all var {&var} into y {colname=vars rowname=&id};
      n=nrow(y);
```

```
      p=ncol(y);
      r1=&id;
      print y;
      m=y[ :,];
      d=y - j(n,1) * m;
      s=d` * d / (n-1);
      dsq=vecdiag(d * inv(s) * d`);
      r=rank(dsq);
      val=dsq; dsq [r, ] = val;
      val=r1; &id [r] = val;
      z=((1:n)` - .5) / n;
      chisq = 2 * gaminv(z,p/2);
      result = dsq||chisq;
      cl={'dsq' 'chisq'};
      create dsquare from result [colname=cl rowname=&id];
      append from result [rowname=&id];
   finish;
   run dsquare;
   quit;
proc print data=dsquare;
   var tag dsq chisq;
run;

filename out 'c:\ 1_8_2.cgm';
goptions device=cgmmwwc gsfname=out gsfmode=replace
   colors=(black) hsize=5in vsize=4in;

proc gplot data=dsquare;
   plot dsq*chisq /frame;
   symbol1 v=;
run;
```

### Result and Interpretation 1_8_2.sas

**Output 1.8.2: Chi – Square Plot of Residuals**




Recall that variable 2 appeared nonnormal from the univariate Q-Q plot. According to the multivariate chi-square plot shown in Output 1.8.2, the points appear to lie close to a straight line, thus indicating that the data are normally distributed.

## 1.9 Scatter Plots

Two-dimensional and three-dimensional scatter plots of variables may be used to detect possible skewness, kurtosis, and outlying observations.

### 1.9.1 Two-Dimensional Plots

PROC PLOT may be used to produce simple bivariate scatter plots. Program 1_9_1.sas produces scatter plots of the Peabody Picture Vocabulary variable (y1) with each of the independent variables in the Rhower data set.

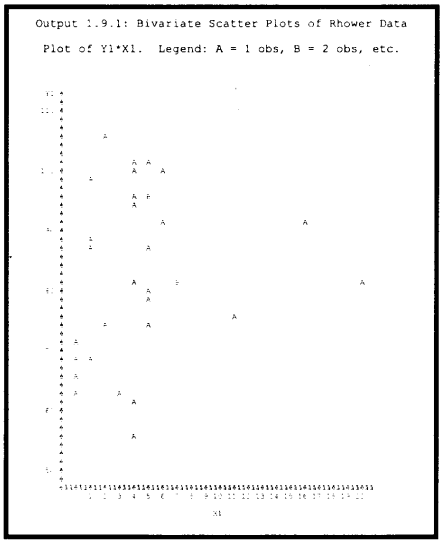
#### Program 1_9_1.sas

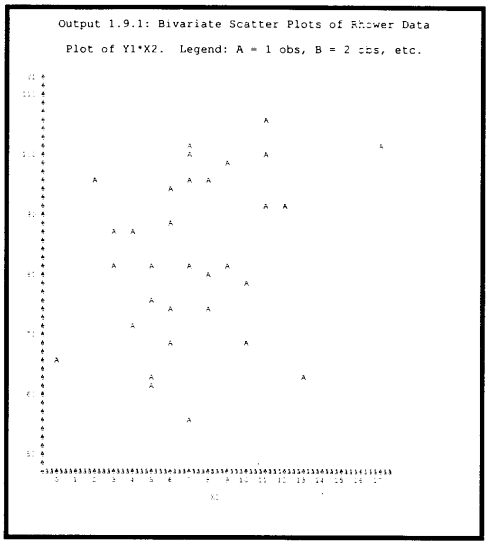
```
/* Program 1_9_1.sas */
/* Program to produce bivariate scatter plot of Rhower Data */

options ls=80 ps=60 nodate nonumber;
filename rhower 'c:\5_6.dat';
title 'Output 1.9.1: Bivariate Scatter Plots of Rhower Data';

data ex191;
   infile rhower;
   input y1-y3 x0-x5;
proc plot;
   plot y1*x1;
   plot y1*x2;
   plot y1*x3;
   plot y1*x4;
   plot y1*x5;
run;
```

### Result and Interpretation 1_9_1.sas

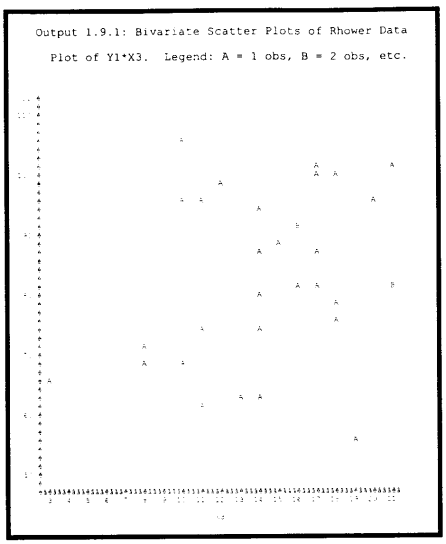
The output is given in Output 1.9.1.

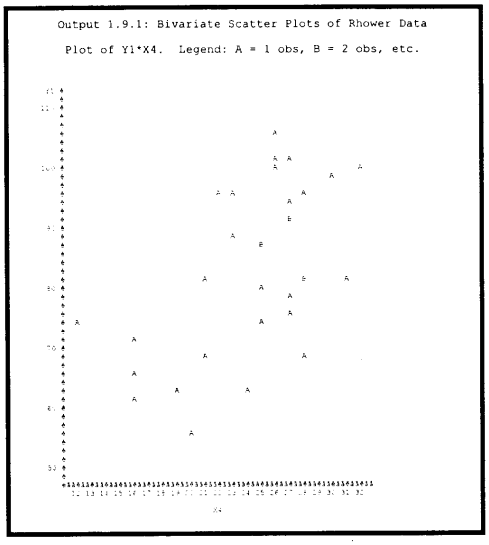
**Output 1.9.1** *Bivariate Scatter Plots of Rhower Data*

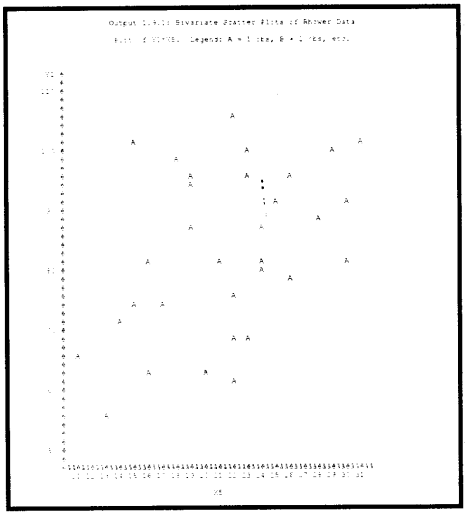


**Output 1.9.1** (*continued*)



From the plot of Y1 versus X1, three of the observations appeared to be outliers. The plot of Y1 versus X3 reveals a possible outlier as well. The other plots show no obvious outliers.

### 1.9.2 Three-Dimensional Plots

Three-dimensional scatter plots may be generated using the G3D procedure. The first part of Program 1_9_2.sas (adapted from Khatree and Naik, 1995, p. 65) produces a plot of a bivariate normal distribution with a mean and covariance matrix

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 3 & 1 \\ 1 & 4 \end{pmatrix}$$

This is the covariance matrix of variables $y1$ and $y2$ from the simulated multivariate normal data generated by Program 1_6.sas.

### Program 1_9_2.sas

```
/* Program 1_9_2.sas */
/* Program to create 3-D Plots of bivariate normal distributions*/

options ls=80 ps=60 nodate nonumber;
title 'Output 1.9.2: Bivariate Normal Distribution';

title2 'with u=(0, 0), var(y1)=3, var(y2)=4, cov(y1,y2)=1, r=.289';
data bivar;
   vy1=3;
   vy2=4;
   r=.289;
   keep y1 y2 z;
   cons=1/(2*3.14159265*sqrt(vy1*vy2*(1-r*r)));
   do y1=-10 to 10 by .2;
      do y2=-10 to 10 by .2;
         zy1=y1/sqrt(vy1);
         zy2=y2/sqrt(vy2);
         d=((zy1**2)+(zy2**2)-2*r*zy1*zy2)/(1-r**2);
         z=cons*exp(-d/2);
         if z > .001 then output;
     end;
   end;
run;

filename out1 'c:\ 1_9_2_1.cgm';
goptions device=cgmmwwc gsfname=out1 gsfmode=replace
   colors=(black) hsize=6in vsize=5in;

proc g3d data=bivar;
   plot y1*y2=z;
run;

/* A Second Plot*/

title2 'with u=(0, 0), var(y1)=3, var(y2)=20, cov=0, r=0';
data bivar2;
   vy1=3;
   vy2=20;
   r=0;
   keep y1 y2 z;
   cons=1/(2*3.14159265*sqrt(vy1*vy2*(1-r*r)));
```

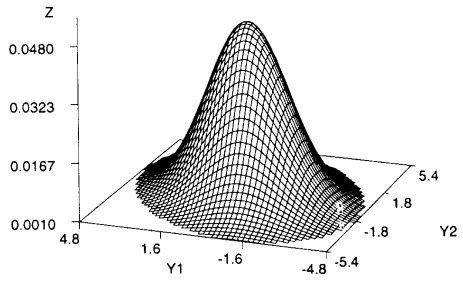
```
    do y1=-10 to 10 by .2;
       do y2=-10 to 10 by .2;
          zy1=y1/sqrt(vy1);
          zy2=y2/sqrt(vy2);
          d=((zy1**2)+(zy2**2)-2*r*zy1*zy2)/(1-r**2);
          z=cons*exp(-d/2);
          if z > .001 then output;
       end;
    end;
 run;


filename out2 'c:\ 1_9_2_2.cgm';
goptions device=cgmmwwc gsfname=out2 gsfmode=replace
    colors=(black) hsize=6in vsize=5in;

proc g3d data=bivar2;
   plot y1*y2=z;
run;
```

### Result and Interpretation 1_9_2.sas

The three-dimensional plot is given in Output 1.9.2.

**Output 1.9.2: Bivariate Normal Distribution**

with u=(0, 0), var(y1)=3, var(y2)=4, cov(y1,y2)=1, r=.289



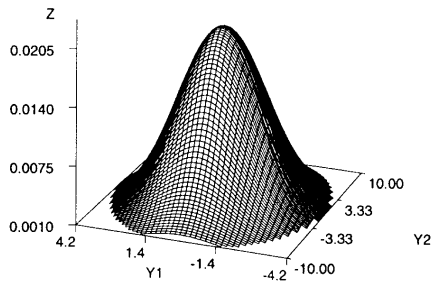

To see how plots vary, see the three-dimensional plot of the bivariate normal distribution having the same covariance matrix as variables $y1$ and $y4$ of 1_6.dat that is also plotted using Program 1_9_2.sas, here

$$\boldsymbol{\Sigma} = \begin{pmatrix} 3 & 0 \\ 0 & 20 \end{pmatrix}$$

The output is given in the continuation of Output 1.9.2 that follows.

## Output 1.9.2: Bivariate Normal Distribution

with u=(0. 0), var(y1)=3, var(y2)=20, cov=0, r=0




Notice that in the first plot, a cross-wise slice would result in an oval shape, whereas in the second plot, a circular shape would result. This is because in the first plot there is a covariance of one, but for the second plot the covariance is zero. See Khattree and Naik for many more graphical representations of multivariate data using SAS software.

## 1.10 Multivariate Skewness and Kurtosis

Another method for examining the assumption of multivariate normality is to compute the measures of skewness and kurtosis. If the data are multivariate normally distributed, these measures should be near zero. If the distribution is leptokurtic (has heavy tails), the measure of kurtosis will be large. If the distribution is platykurtic (has light tails), the kurtosis coefficient will be small.

Mardia (1970) defines the measures of multivariate skewness and kurtosis as

$$\begin{aligned} \beta_{1,p} &= E\left\{(\mathbf{x}-\mu)'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mu)\right\}^3 \\ \beta_{2,p} &= E\left\{(\mathbf{y}-\mu)'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mu)\right\}^2 \end{aligned} \tag{1.34}$$

where $\mathbf{x}$ and $\mathbf{y}$ are identically and independently distributed. Sample estimates of these quantities are

$$\begin{aligned} b_{1,p} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[(\mathbf{y}_i-\bar{\mathbf{y}})'\mathbf{S}^{-1}(\mathbf{y}_j-\bar{\mathbf{y}})\right]^3 \\ b_{2,p} &= \frac{1}{n}\sum_{i=1}^{n}\left[(\mathbf{y}_i-\bar{\mathbf{y}})'\mathbf{S}^{-1}(\mathbf{y}_i-\bar{\mathbf{y}})\right]^2 \end{aligned} \tag{1.35}$$

When $\mathbf{y} \sim N_p(\mu,\boldsymbol{\Sigma})$, then $\beta_{1,p}=0$, $\beta_{2,p}=p(p+2)$, $\kappa_1 = nb_{1,p}/6$ has an asymptotic chi-square distribution with $p(p+1)(p+2)/6$ degrees of freedom, and $\kappa_2=\left[b_{2,p}-p(p+2)\right]/\left[8p(p+2)/n\right]^{1/2}$ has an asymptotic standard normal distribution. Thus, when $n>50$ one may develop large sample tests of multivariate normality. Mardia (1974) developed tables of approximate percentiles for $p=2$ and $n \geq 10$ and alternative large sample

approximations. In general, tests of hypotheses regarding means are sensitive to highly skewed multivariate distributions.

To illustrate the calculations to evaluate multivariate skewness and kurtosis, we again will use the Rhower data; however, given the small sample size of only $n = 32$ the significance level associated with the test may be in error. The SAS code to perform the calculations and to estimate $p$-values is given in Program 1_10.sas.

### Program 1_10.sas

```
/* Program 1_10.sas                                              */
/* Program to calculate Mardia's measure of multivariate         */
/* skewness and kurtosis                                         */

options ls=80 ps=60 nodate nonumber;
title 'Output 1.10: Mardias Multivariate Skewness & Kurtosis';

data Rhower;
   infile 'c:\ 5_1.dat';
   input y1-y3 x0-x5;
proc print data=Rhower;

proc iml;
   use Rhower;
   v={y1 y2 y3};
   w={x0 x1 x2 x3 x4 x5};
   read all var v into y;
   read all var w into x;
   beta=inv(x`*x)*x`*y;
   n=nrow(y);
   p=ncol(y);
   k=ncol(x);
   s=(y`*y-y`*x*beta)/(n-k);
   s_inv=inv(s);
   q=i(n)-x*(inv(x`*x)*x`);
   d=q*y*s_inv*y`*q;

   b1=(sum(d#d))/(n*n);
   b2=trace(d#d)/n;

   kappa1= n*b1/6;
   kappa2=(b2-p*(p+2))/sqrt(8*p*(p+2)/n);

   dfchi=p*(p-1)*(p+2)/6;

   pvalskew=1-probchi(kappa1,dfchi);
   pvalkurt=2*(1-probnorm(abs(kappa2)));

   print s; print s_inv;
   print b1; print kappa1; print pvalskew;
   print b2; print kappa2; print pvalkurt;
quit;
```

First the three dependent and five independent variables of the Rhower data set are read within the DATA step and then printed. Next, using PROC IML, the matrix $\mathbf{Y}$ is defined as the $32 \times 3$ matrix of dependent variables, and the matrix $\mathbf{X}$ is defined as the $32 \times 5$ matrix of independent variables. The variance-covariance matrix is then computed and named $s$. Next, $b1$ and $b2$ of (1.35) are computed and then Kappa1 and Kappa2 and the corresponding $p$-values, pvalskew and pvalkurt, are printed.

## Result and Interpretation 1_10.sas

**Output 1.10:** *Mardias Multivariate Skewness & Kurtosis*

Output 1.10: Mardias Multivariate Skewness & Kurtosis

| OBS | Y1 | Y2 | Y3 | X0 | X1 | X2 | X3 | X4 | X5 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 68 | 15 | 24 | 1 | 0 | 10 | 8 | 21 | 22 |
| 2 | 82 | 11 | 8 | 1 | 7 | 3 | 21 | 28 | 21 |
| 3 | 82 | 13 | 88 | 1 | 7 | 9 | 17 | 31 | 30 |
| 4 | 91 | 18 | 82 | 1 | 6 | 11 | 16 | 27 | 25 |
| 5 | 82 | 13 | 90 | 1 | 20 | 7 | 21 | 28 | 16 |
| 6 | 100 | 15 | 77 | 1 | 4 | 11 | 18 | 32 | 29 |
| 7 | 100 | 13 | 58 | 1 | 6 | 7 | 17 | 26 | 23 |
| 8 | 96 | 12 | 14 | 1 | 5 | 2 | 11 | 22 | 23 |
| 9 | 63 | 10 | 1 | 1 | 3 | 5 | 14 | 24 | 20 |
| 10 | 91 | 18 | 98 | 1 | 16 | 12 | 16 | 27 | 30 |
| 11 | 87 | 10 | 8 | 1 | 5 | 3 | 17 | 25 | 24 |
| 12 | 105 | 21 | 88 | 1 | 2 | 11 | 10 | 26 | 22 |
| 13 | 87 | 14 | 4 | 1 | 1 | 4 | 14 | 25 | 19 |
| 14 | 76 | 16 | 14 | 1 | 11 | 5 | 18 | 27 | 22 |
| 15 | 66 | 14 | 38 | 1 | 0 | 0 | 3 | 16 | 11 |
| 16 | 74 | 15 | 4 | 1 | 5 | 8 | 11 | 12 | 15 |
| 17 | 68 | 13 | 64 | 1 | 1 | 6 | 10 | 28 | 23 |
| 18 | 98 | 16 | 88 | 1 | 1 | 9 | 12 | 30 | 18 |
| 19 | 63 | 15 | 14 | 1 | 0 | 13 | 13 | 19 | 16 |
| 20 | 94 | 16 | 99 | 1 | 4 | 6 | 14 | 27 | 19 |
| 21 | 82 | 18 | 50 | 1 | 4 | 5 | 16 | 21 | 24 |
| 22 | 89 | 15 | 36 | 1 | 1 | 6 | 15 | 23 | 28 |
| 23 | 80 | 19 | 88 | 1 | 5 | 8 | 14 | 25 | 24 |
| 24 | 61 | 11 | 14 | 1 | 4 | 5 | 11 | 16 | 22 |
| 25 | 102 | 20 | 24 | 1 | 5 | 7 | 17 | 26 | 15 |
| 26 | 71 | 12 | 24 | 1 | 0 | 4 | 8 | 16 | 14 |
| 27 | 102 | 16 | 24 | 1 | 4 | 17 | 21 | 27 | 31 |
| 28 | 96 | 13 | 50 | 1 | 5 | 8 | 20 | 28 | 26 |
| 29 | 55 | 16 | 8 | 1 | 4 | 7 | 19 | 20 | 13 |
| 30 | 96 | 18 | 98 | 1 | 4 | 7 | 10 | 23 | 19 |
| 31 | 74 | 15 | 98 | 1 | 2 | 6 | 14 | 25 | 17 |
| 32 | 78 | 19 | 50 | 1 | 5 | 10 | 18 | 27 | 26 |

S

| | | |
|---|---|---|
| 149.9613 | 10.822548 | 49.215424 |
| 10.822548 | 6.8088795 | 23.991269 |
| 49.215424 | 23.991269 | 659.0046 |

S_INV

| | | |
|---|---|---|
| 0.0075446 | -0.011479 | -0.000146 |
| -0.011479 | 0.1859442 | -0.005912 |
| -0.000146 | -0.005912 | 0.0017435 |

B1

1.9804688

KAPPA1

10.5625

PVALSKEW

0.3926023

B2

8.8696232

**Output 1.10** (*continued*)

```
                    KAPPA2
                  -3.165713

                    PVALKURT
                    0.001547
```

From Output 1.10, it appears that for the Rhower data, there is no evidence of multivariate skewness. The distribution may, however, be platykurtic since the Kurtosis coefficient is small (-3.166) and the $p$-value is significant ($p < .01$).

## 1.11 Box-Cox Transformations

Data that are not normally distributed can often be transformed to near normality. Count data can be made to be more nearly normal by taking the square root. Proportions are often transformed using the logit, $\text{logit}(p) = \frac{1}{2}\log\left(\frac{p}{1-p}\right)$, and correlations are transformed using Fisher's $r$ to $z$ transformation, $\frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$, or the arc-sine square root. Alternately, the sample data can be used for finding an appropriate power transformation by using the Box-Cox family of power transformations

$$x^{(\lambda)} = \begin{cases} \dfrac{x^{\lambda}-1}{\lambda} & \lambda \neq 0 \quad x > 0 \\ \ln x & \lambda = 0 \end{cases} \tag{1.36}$$

where x is the observed data. The appropriate $\lambda$ to use for the transformation is that value of $\lambda$ that maximizes the expression

$$L(\lambda) = -\frac{n}{2}\ln\left[\frac{1}{n}\sum_{i=1}^{n}(x_i^{(\lambda)} - \overline{x^{(\lambda)}})^2\right] + (\lambda - 1)\sum_{i=1}^{n}\ln x_i \tag{1.37}$$

$$\overline{x^{(\lambda)}} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(\lambda)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(x_i)^{\lambda}-1}{\lambda}\right)$$

Program 1_11.sas computes and graphs the function L ($\lambda$) for values of $\lambda$: -1.0 (.1) 1.3. The appropriate value to use for transforming nonnormally distributed observations is obtained from the plot by locating the $\hat{\lambda}$ that maximizes the likelihood. When the data are normally distributed there will be no maximum value for the graph.

### Program 1_11.sas

```
/* Program 1_11.sas                                          */
/* Program to compute Box Cox Transformations                */
/* To run this program on your own dataset change            */
/* the name of the file in the file=____ statement           */
/* the number of rows in the n=____ statement and            */
/* the number of columns (variables) in the p=___ statement */
```

```
options ls=80 ps=60 nodate nonumber mprint;
%let file=c:\expl_6.dat;
%let n=50;
%let p=1;

 /*macro to expand the string of variables that are processed */
%macro expand(cols);
   %do j=1 %to &cols;
      x&j
   %end;
%mend expand;

 /*macro to perform the Box-Cox transformation on the data matrix */
%macro loop(cols);
   %do i=1 %to &cols;
      proc iml;
         use matrix;
         read all var {x&i};
         in=i(&n);
         allh={-1.0, -0.9, -0.8, -0.7, -0.6, -0.5, -0.4, -0.3, -0.2, -0.1,
           0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3};
         one=j(&n,1,1);
         c=in-(one*(inv(one`*one))*one`);
         do k=1 to 23 by 1;
            h=allh[k,1];
            xh=x&i##h;
            hinv=1/h;
            vhinv=j(&n,1,hinv);
            y=(xh-one)#vhinv;
            my=(one`*y)/&n;
            ycy=y`*c*y;
            lnx=log(x&i);
            slnx=one`*lnx;
            ycyn=ycy/&n;
            if ycyn > 0 then lhp1=-(&n/2)*log(ycyn); else lhp1=.;
            lhp2=(h-1)*slnx;
            if ycyn > 0 then lh=lhp1+lhp2; else lh=.;
            lhs=lhs//lh;
         end;
         Lambda=allh||lhs;
         print, "Lambda and corresponding likelihood for variables x&i",
           lambda;
         call pgraf(lambda, '*', 'lambda', 'likelihood',
         "plot of lambda vs likelihood for variable x&i");
      quit;
   %end;
%mend loop;

 /*input the data and process the macro */
data matrix;
   infile "&file";
   input (%expand(&p)) (&p*:25.) ;
   title "Output 1.11: Box-Cox Transformation plots of &file";
proc print;
%loop(&p)
run;
```

First the name of the data file is specified in the statement `%let file = exp1_6.dat`, and the number of observations and the number of columns are specified in the following two statements, respectively. In the macro named loop, PROC IML computes the values of $L(\lambda)$ of (1.37) for each of the values of $\lambda$ in the vector named **allh**. The likelihood value is named lh and at the end of each iteration through the values of **allh**, the value of lh is appended to the bottom of a vector named **lhs**. The vectors **allh** and **lhs** are then appended side by side and plotted using the PGRAF call.

## Result and Interpretation 1_11.sas

Output 1.11 gives the results of using Program 1_11.sas for variable $y1$ of the data generated by Program 1_6.sas, but transformed by the exponential transformation $y_i^* = \exp(y_i)$.

**Output 1.11:** *Box-Cox Transformation plots of exp1_6.dat*

Output 1.11: Box-Cox Transformation plots of exp1_6.dat

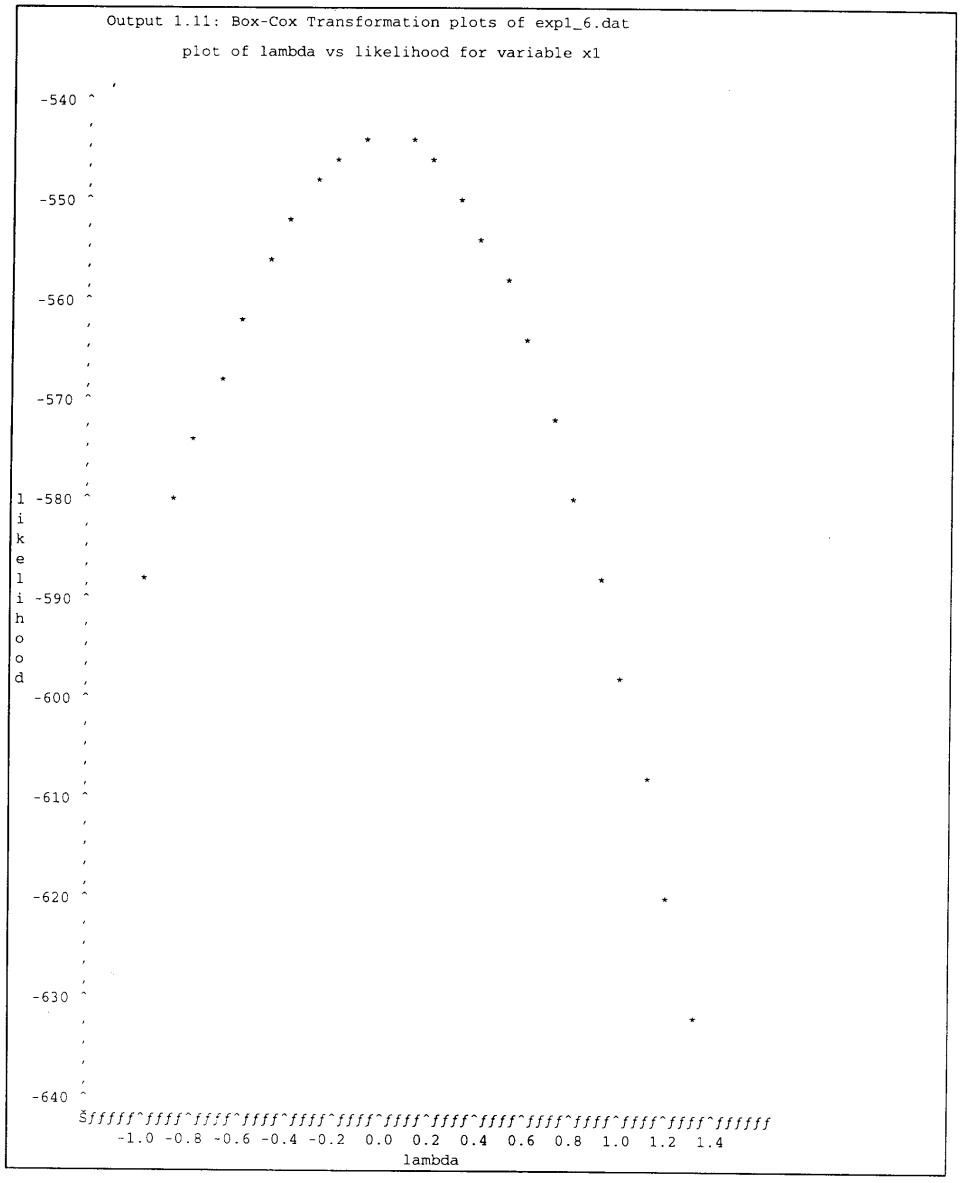
| OBS | X1 |
|---|---|
| 1 | 29063.57 |
| 2 | 151698.83 |
| 3 | 3698.38 |
| 4 | 4473.25 |
| 5 | 22789.77 |
| 6 | 381699.28 |
| 7 | 2386.98 |
| 8 | 38226.50 |
| 9 | 24487.89 |
| 10 | 208847.22 |
| 11 | 6082.12 |
| 12 | 177373.85 |
| 13 | 17846.14 |
| 14 | 111387.69 |
| 15 | 8053.79 |
| 16 | 52212.82 |
| 17 | 264210.87 |
| 18 | 89527.14 |
| 19 | 136181.37 |
| 20 | 82360.05 |
| 21 | 87709.27 |
| 22 | 957392.80 |
| 23 | 14595.46 |
| 24 | 5590.57 |
| 25 | 81208.83 |
| 26 | 220662.85 |
| 27 | 451228.04 |
| 28 | 7171.21 |
| 29 | 204031.97 |
| 30 | 5078.52 |
| 31 | 6667.42 |
| 32 | 4376.54 |
| 33 | 13820.96 |
| 34 | 75634.08 |
| 35 | 33116.29 |
| 36 | 180406.47 |
| 37 | 204440.99 |
| 38 | 15697.88 |

**Output 1.11** (*continued*)

| | |
|---|---|
| 39 | 18038.71 |
| 40 | 9306.02 |
| 41 | 77575.93 |
| 42 | 2583.53 |
| 43 | 7311.21 |
| 44 | 2659.12 |
| 45 | 118760.90 |
| 46 | 16748.52 |
| 47 | 116633.26 |
| 48 | 2973.87 |
| 49 | 25203.74 |
| 50 | 126357.47 |

| LAMBDA | |
|---|---|
| -1 | -587.4075 |
| -0.9 | -580.1185 |
| -0.8 | -573.3246 |
| -0.7 | -567.0842 |
| -0.6 | -561.4603 |
| -0.5 | -556.5192 |
| -0.4 | -552.3292 |
| -0.3 | -548.9588 |
| -0.2 | -546.4744 |
| -0.1 | -544.9384 |
| 0.1 | -544.9296 |
| 0.2 | -546.5448 |
| 0.3 | -549.28 |
| 0.4 | -553.1478 |
| 0.5 | -558.1424 |
| 0.6 | -564.2375 |
| 0.7 | -571.3853 |
| 0.8 | -579.5187 |
| 0.9 | -588.5554 |
| 1 | -598.4042 |
| 1.1 | -608.9711 |
| 1.2 | -620.1652 |
| 1.3 | -631.9018 |

**Output 1.11** (*continued*)



Notice that the value of the likelihood reaches the maximum at lambda equals 0. This suggests that the natural log transformation should be used, as we would have expected, since we transformed the normally distributed variable by exp(y).

When using the Box-Cox transformation, one must transform the data and reevaluate normality, a variable at a time. Alternatively, one may estimate the shape $(\hat{\eta})$ and scale $(\hat{\lambda})$ parameters of a gamma distribution using $D_i^2$ and construct gamma probability plots to assess multivariate normality (Gnandesikan, 1980). One may also construct multivariate Box-Cox transformations to improve joint multivariate normality (Andrews, 1971). However, such procedures are not readily available except through custom software (Gnandesikan, 1997). Only recently have regression graphics reached the desktop for the multiple regression case (Cook and Weisberg, 1994).