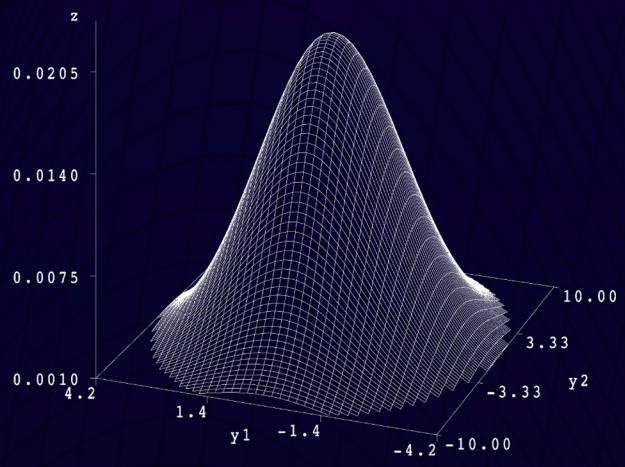
STATISTICS:
a series of TEXTBOOKS and MONOGRAPHS

Univariate and Multivariate General Linear Models

Theory and Applications with SAS

Second Edition



Kevin Kim
Neil Timm

Chapman & Hall/CRC
Taylor & Francis Group

Univariate and Multivariate General Linear Models

Theory and Applications with SAS

Second Edition

STATISTICS: Textbooks and Monographs

Recent Titles

Handbook of Stochastic Analysis and Applications, *edited by D. Kannan and V. Lakshmikantham*

Testing for Normality, *Henry C. Thode, Jr.*

Handbook of Applied Econometrics and Statistical Inference, *edited by Aman Ullah, Alan T. K. Wan, and Anoop Chaturvedi*

Visualizing Statistical Models and Concepts, *R. W. Farebrother and Michaël Schyns*

Financial and Actuarial Statistics: An Introduction, *Dale S. Borowiak*

Nonparametric Statistical Inference, Fourth Edition, Revised and Expanded, *Jean Dickinson Gibbons and Subhabrata Chakraborti*

Computer-Aided Econometrics, *edited by David E.A. Giles*

The EM Algorithm and Related Statistical Models, *edited by Michiko Watanabe and Kazunori Yamaguchi*

Multivariate Statistical Analysis, Second Edition, Revised and Expanded, *Narayan C. Giri*

Computational Methods in Statistics and Econometrics, *Hisashi Tanizaki*

Applied Sequential Methodologies: Real-World Examples with Data Analysis, *edited by Nitis Mukhopadhyay, Sujay Datta, and Saibal Chattopadhyay*

Handbook of Beta Distribution and Its Applications, *edited by Arjun K. Gupta and Saralees Nadarajah*

Item Response Theory: Parameter Estimation Techniques, Second Edition, *edited by Frank B. Baker and Seock-Ho Kim*

Statistical Methods in Computer Security, *edited by William W. S. Chen*

Elementary Statistical Quality Control, Second Edition, *John T. Burr*

Data Analysis of Asymmetric Structures, *Takayuki Saito and Hiroshi Yadohisa*

Mathematical Statistics with Applications, *Asha Seth Kapadia, Wenyaw Chan, and Lemuel Moyé*

Advances on Models, Characterizations and Applications, *N. Balakrishnan, I. G. Bairamov, and O. L. Gebizlioglu*

Survey Sampling: Theory and Methods, Second Edition, *Arijit Chaudhuri and Horst Stenger*

Statistical Design of Experiments with Engineering Applications, *Kamel Rekab and Muzaffar Shaikh*

Quality by Experimental Design, Third Edition, *Thomas B. Barker*

Handbook of Parallel Computing and Statistics, *Erricos John Kontoghiorghes*

Statistical Inference Based on Divergence Measures, *Leandro Pardo*

A Kalman Filter Primer, *Randy Eubank*

Introductory Statistical Inference, *Nitis Mukhopadhyay*

Handbook of Statistical Distributions with Applications, *K. Krishnamoorthy*

A Course on Queueing Models, *Joti Lal Jain, Sri Gopal Mohanty, and Walter Böhm*

Univariate and Multivariate General Linear Models: Theory and Applications with SAS, Second Edition, *Kevin Kim and Neil Timm*

Univariate and Multivariate General Linear Models

Theory and Applications with SAS

Second Edition

Kevin Kim
University of Pittsburgh
Pennsylvania, U.S.A.

Neil Timm
University of Pittsburgh
Pennsylvania, U.S.A.

Chapman & Hall/CRC
Taylor & Francis Group
Boca Raton London New York

Chapman & Hall/CRC is an imprint of the
Taylor & Francis Group, an informa business

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 1-58488-634-X (Hardcover)
International Standard Book Number-13: 978-1-58488-634-1 (Hardcover)

Library of Congress Cataloging-in-Publication Data

Kim, Kevin.
Univariate and multivariate general linear models : theory and applications with SAS. -- 2nd ed. / Kevin Kim and Neil Timm.
p. cm. -- (Statistics : textbooks and monographs)
Rev. ed. of: Univariate & multivariate general linear models / Neil H. Timm, Tammy A. Mieczkowski. c1997.
Includes bibliographical references and index.
ISBN-13: 978-1-58488-634-1 (acid-free paper)
ISBN-10: 1-58488-634-X (acid-free paper)
1. Linear models (Statistics)--Textbooks. 2. Linear models (Statistics)--Data processing--Textbooks. 3. SAS (Computer file) I. Timm, Neil H. II. Timm, Neil H. Univariate & multivariate general linear models. III. Title. IV. Series.

QA279.T56 2007
519.5'35--dc22 2006026561

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

List of Tables

Preface

The general linear model is often first introduced to graduate students during a course on multiple linear regression, analysis of variance, or experimental design; however, most students do not fully understand the generality of the model until they have taken several courses in applied statistics. Even students in graduate statistics programs do not fully appreciate the generality of the model until well into their program of study. This is due in part to the fact that theory and applications of the general linear model are discussed in discrete segments throughout the course of study rather than within a more general framework. In this book, we have tried to solve this problem by reviewing the theory of the general linear model using a general framework. Additionally, we use this general framework to present analyses of simple and complex models, both univariate and multivariate, using data sets from the social and behavioral sciences and other disciplines.

AUDIENCE

The book is written for advanced graduate students in the social and behavioral sciences and in applied statistics who are interested in statistical analysis using the general linear model. The book may be used to introduce students to the general linear model; at the University of Pittsburgh it is used in a one-semester course on linear models. The book may also be used as a supplement to courses in applied statistical methods covering the essentials of estimation theory and hypothesis testing, simple linear regression, and analysis of variance. They should also have some familiarity with matrix algebra and with running SAS procedures.

OVERVIEW

Each chapter of this book is divided into two sections: theory and applications. Standard SAS procedures are used to perform most of the analyses. When standard SAS procedures are not available, PROC IML code to perform the analysis is discussed. Because SAS is not widely used in the social and behavioral sciences, SAS code for analyzing general linear model applications is discussed in detail. The code can be used as a template.

Chapter 1 provides an overview of the general linear model using matrix algebra and an introduction to the multivariate normal distribution as well as to the general theory of hypothesis testing. Applications include the use of graphical methods to

evaluate univariate and multivariate normality and the use of transformations to normality. In Chapter 2 the general linear model without restrictions is introduced and used to analyze multiple regression and ANOVA designs. In Chapter 3 the general linear model with restrictions is discussed and used to analyze ANCOVA designs and repeated measurement designs.

Chapter 4 extends the concepts of the first three chapters to general linear models with heteroscedastic errors and illustrates how the model may be used to perform weighted least squares regression and to analyze categorical data. Chapter 5 extends the theory of Chapter 2 to the multivariate case; applications include multivariate regression analysis, MANOVA, MANCOVA, and analyses of repeated measurement data. This chapter also extends "standard" hypothesis testing to extended linear hypotheses. In Chapter 6, the double multivariate linear model is discussed.

Chapter 7 extends the multivariate linear model to include restrictions and considers the growth curve model. In Chapter 8, the seeming unrelated regression (SUR) and the restricted GMANOVA models are analyzed. Many of the applications in this chapter involve PROC IML code. Finally, Chapter 9 includes analyses of hierarchical linear models, and Chapter 10 treats the analysis of incomplete repeated measurement data.

While the coverage given the general linear model is extensive, it is not exhaustive. Excluded from the book are Bayesian methods, nonparametric procedures, nonlinear models, and generalized linear models, among others.

ACKNOWLEDGMENTS

We would like to thank the reviewers at SAS Institute Inc. and the technical reviewer Vernon M. Chinchilli for their helpful comments and suggestions on the book. We thank Hanna Hicks Schoenrock and Caroline Brickley for making the process of producing this book run so smoothly. We also appreciate the helpful suggestions made on an early draft by doctoral students in the linear models course.

We would especially like to extend our gratitude to Roberta S. Allen. Ms. Allen expertly typed every draft of the book from inception through every revision, including all equations. Thank you for your excellent work and patience with us, Roberta. The authors also want to thank the authors and publishers of copyrighted material for permission to reproduce tables and data sets used in the book.

This book was completed while Neil H. Timm was on sabbatical leave during the 1996-1997 academic year. He would like to thank his colleagues for their support and the School of Education for this opportunity. He would also like to thank his wife, Verena, for her support and encouragement.

Tammy A. Mieczkowski was a full time doctoral student in the Research Methodology program in the School of Education when this book was completed. She also works as a Graduate Student Researcher at the University of Pittsburgh School of Medicine, Department of Family Medicine and Clinical Epidemiology. She would like to thank her family, especially her sister and friend, Karen, for their support, encouragement, understanding, and love.

Neil H. Timm
Tammy A. Mieczkowski

SECOND EDITION

For the second edition, we note that the authorship has changed. The former second author has removed herself from the second edition and a new author, Kevin H. Kim, appears. The order of the authors is alphabetical.

The revision includes corrections to the first edition, expanded material, additional examples, and new material. The theory in Chapters 2, 5, and 8 has been expanded to include recent developments; Chapters 9 and 10 have been rewritten and expanded to include recent developments in structural equation modeling (SEM), Growth Mixture Modeling, Longitudinal Data Analysis, and Hierarchical Linear Models (HLM), Chapters 11, 12, and 13. The material in Chapter 12 on missing data has been expanded to include multiple imputation and the EM algorithm. Also included in the second edition is the addition of Chapter 9 and 10 on Finite Intersection Tests and Power Analysis which illustrates the experimental GLMPOWER procedure. All examples have been revised to include options available using SAS version 9.0.

The second addition also includes applications of the MI, MIANALYZE, TRANSREG, and CALIS procedures. In addition, use of ODS capabilities are illustrated in the examples and functions not known to all users of SAS such as the PROBMC distribution are illustrated. While we still include some output from the SAS procedures in the book, the output has been abbreviated. SAS code and data files have been excluded from the book to save space and are available at the Pitt site www.pitt.edu/˜timm.

Kevin H. Kim
Neil H. Timm

CHAPTER 1

Overview of General Linear Model

1.1 INTRODUCTION

In this chapter, we introduce the structure of the general linear model (GLM) 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 of the linear model is reviewed, along with a general approach to hypothesis testing. Graphical methods and tests for assessing univariate and multivariate normality are also reviewed. The generation of multivariate normal data, the construction of Quantile-Quantile (Q-Q) plots, chi-square plots, scatter plots, and data transformation procedures are reviewed and illustrated to evaluate normality.

1.2 GENERAL LINEAR MODEL

Data analysis in the social and behavioral sciences and numerous other disciplines is associated with a model known as the GLM. Employing matrix notation, univariate and multivariate linear models may be represented using the general form

$$\Omega_0 : y = X\beta + e \tag{1.1}$$

where $y_{n\times 1}$ is a vector of n observations, $X_{n\times k}$ is a known design matrix of full column rank k, $\beta_{k\times 1}$ is a vector of k fixed parameters, $e_{n\times 1}$ is a random vector of errors with mean zero, $\mathcal{E}(e) = 0$, and covariance matrix $\Omega = \mathrm{cov}(e)$. 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 X and Ω.

Depending on the structure of X and Ω, the model in (1.1) has many names in the literature. To illustrate, if $\Omega = \sigma^2 I_n$ in (1.1), the model is called the classical linear regression model or the standard linear regression model. If we partition X to have the form $X = (X_1, X_2)$ where X_1 is associated with fixed effects and X_2 is associated with random effects, and if covariance matrix Ω has the form

$$\Omega = X_2 V X_2' + \Psi \tag{1.2}$$

where V and Ψ are covariance matrices, then (1.1) becomes the general linear mixed model (GLMM). If we let X and Ω take the general form

$$X = \begin{pmatrix} X_1 & 0 & \dots & 1 \\ 0 & X_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & X_p \end{pmatrix} = \oplus_{i=1}^{p} X_i \tag{1.3}$$

$$\Omega = \Sigma \otimes I_n \tag{1.4}$$

where $\Sigma_{p\times p}$ is a covariance matrix, $A \otimes B$ denotes the Kronecker product of two matrices A and B ($A\otimes = a_{ij}B$), and $\oplus_{i=1}^{p}$ represents the direct sum of the matrices X_i, then (1.1) is Zellner's seemingly unrelated regression (SUR) model or a multiple design multivariate (MDM) model. The SUR model may also be formulated as p separate linear regression models that are not independent

$$y_i = X_i\beta_{ii} + e_i \tag{1.5}$$

$$\operatorname{cov}(y_i, y_j) = \sigma_{ij} I_n \tag{1.6}$$

for $i, j = 1, 2, \dots, p$ where y, β and e in (1.1) are partitioned

$$y' = \begin{pmatrix} y_1' & y_2' & \dots & y_p' \end{pmatrix} \qquad \text{where } y_i : n^* \times 1 \tag{1.7}$$

$$\beta' = \begin{pmatrix} \beta_1' & \beta_2' & \dots & \beta_p' \end{pmatrix} \qquad \text{where } \beta_{ii} : k_i \times 1 \tag{1.8}$$

$$e' = \begin{pmatrix} e_1' & e_2' & \dots & e_p' \end{pmatrix} \qquad \text{where } e_i : n^* \times 1 \tag{1.9}$$

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

$$\Omega_0 : Y_{n^*\times p} = X_{n^*\times k}\tilde{\beta}_{k\times p} + U_{n^*\times p} \tag{1.10}$$

where $Y = (y_1, y_2, \dots, y_p)$, $X = (X_1, X_2, \dots, X_p)$, $U = (e_1, e_2, \dots, e_p)$ and

$$\tilde{\beta} = \begin{pmatrix} \beta_{11} & 0 & \dots & 0 \\ 0 & \beta_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \beta_{pp} \end{pmatrix}. \tag{1.11}$$

Letting $X_1 = X_2 = \cdots = X_p = \widetilde{X}_{n^*\times k}$ and $\tilde{\beta} = (\beta_{11}, \beta_{22}, \dots, \beta_{pp})$ in the SUR model, (1.1) becomes the classical multivariate regression model or the multivariate analysis of variance (MANOVA) model. Finally, letting

$$X = X_1 \otimes X_2' \tag{1.12}$$

$$\Omega = I_n \otimes \Sigma \tag{1.13}$$

model (1.1) becomes the generalized MANOVA (GMANOVA) or the generalized growth curve model. All these models with some further extensions are special forms of the GLM discussed in this book.

The model in (1.1) is 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 $y' = (y_1, y_2, \ldots, y_n)$ 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 $x_i' = (x_1, x_2, \ldots, x_k)$ or k independent (possible dummy) variables. With this orientation (1.1) becomes

$$y = X'\beta + e \tag{1.14}$$

where $X = (x_1, x_2, \ldots, x_k)$, $e' = (e_1, e_2, \ldots, e_n)$ and each x_i contains k independent variables for the i^{th} observation. Model (1.14) is often called the "responsewise" form. Model (1.1) is clearly equivalent to (1.14) since the design matrix has the same order for either representation; however, in (1.14) X is of order $k \times n$. Thus, $X'X$ using (1.14) becomes XX' for the responsewise form of the classical model.

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

$$y = \beta_0 + \beta_1 x + e \tag{1.15}$$

where x represents the independent variable, y the dependent variable and e a random error. Model (1.15) states that the observed dependent variable for each subject is hypothesized to be a function of a common parameter β_0 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 β_1 plus a random error e. For $k = p + 1$ with p variables, (1.15) becomes (1.16)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e \tag{1.16}$$

or using matrix notation, (1.16) is written as

$$y = x'\beta + e \tag{1.17}$$

where $x' = (x_0, x_1, \ldots, x_p)$ denotes k independent variables, and x_0 is a dummy variable in the vector x'. Then for a sample of n observations, (1.17) has the general form (1.14) where $y' = (y_1, y_2, \ldots, y_n)$, $e' = (e_1, e_2, \ldots, e_n)$ and $X = (x_1, x_2, \ldots, x_n)$ of order $k \times n$ since each column vector x_i in X contains k variables. When using the classical form (1.1), $X \equiv X'$, a matrix of order $n \times k$. In discussions of the GLM, many authors will use either the classical or the responsewise version of the GLM, while we will in general prefer (1.1). In some applications (e.g., repeated measurement designs) form (1.14) is preferred.

1.3 Restricted General Linear Model

In specifying the GLM using (1.1) or (1.14), we have not restricted the k-variate parameter vector β. A linear restriction on the parameter vector β will affect the characterization of the model. Sometimes it is necessary to add restrictions to the GLM of the general form

$$R\beta = \theta \tag{1.18}$$

where $R_{s\times k}$ is a known matrix with full row rank, $\mathrm{rank}(R) = s$, and θ is a known parameter vector, often assumed to be zero. With (1.18) associated with the GLM, the model is commonly called the restricted general linear model (RGLM). Returning to (1.15), we offer an example of this in the simple linear regression model with a restriction

$$y = \beta_0 + \beta_1 x + e \qquad \beta_0 = 0 \tag{1.19}$$

so that the regression of y on x is through the origin. For this situation, $R \equiv (1, 0)$ and $\theta = (0, 0)$. Clearly, the estimate of β_1 using (1.19) will differ from that obtained using the general linear model (1.15) without the restriction.

Assuming (1.1) or (1.16), one first wants to estimate β with $\hat{\beta}$ where the estimator $\hat{\beta}$ has some optimal properties like unbiasness and minimal variance. Adding (1.18) to the GLM, one obtains a restricted estimator of β, $\hat{\beta}_r$, which in general is not equal to the unrestricted estimator. Having estimated β, one may next want to test hypotheses regarding the parameter vector β and the structure of Ω. The general form of the null hypothesis regarding β is

$$H : C\beta = \xi \tag{1.20}$$

where $C_{g\times k}$ is a matrix of full row rank g, $\mathrm{rank}(C) = g$ and $\xi_{g\times 1}$ is a vector of known parameters, usually equal to zero. The hypothesis in (1.20) may be tested using the GLM with or without the restriction given in (1.18). Hypotheses in the form (1.20) are in general testable, provided β is estimable; however, testing (1.20) assuming (1.18) is more complicated since the matrix C may not contain a row identical, inconsistent or dependent on the rows of R and the rows of C must remain independent. Thus, the rank of the augmented matrix must be greater than s, $\mathrm{rank}\binom{R}{C} = s + g > s$.

Returning to (1.15), we may test the null hypotheses

$$H : \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \xi = \begin{pmatrix} \xi_0 \\ \xi_1 \end{pmatrix} \tag{1.21}$$

where ξ is a known parameter vector. The hypothesis in (1.21) may not be inconsistent with the restriction $\beta_0 = 0$. Thus, given the restriction, we may test

$$H : \beta_1 = \xi_1 \tag{1.22}$$

so that (1.22) is not inconsistent or dependent on the restriction.

1.4 Multivariate Normal Distribution

To test hypotheses of the form given in (1.20), one usually makes distributional assumptions regarding the observation vector y or 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

$$f(y) = (2\pi)^{-1/2} \exp\left(-\frac{1}{2}y^2\right) \tag{1.23}$$

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

$$\mu + \sigma e \tag{1.24}$$

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

$$\begin{aligned} f(y) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right] \\ &= \left(2\pi\sigma^2\right)^{-1/2} \exp\left[-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right] \end{aligned} \tag{1.25}$$

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

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

$$f(y) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right]. \tag{1.26}$$

Letting $(\mathrm{diag}\Sigma)^{-1/2}$ represent the diagonal matrix with diagonal elements equal to the square root of the diagonal elements of Σ, the population correlation matrix for the elements of the vector y is

$$P = (\mathrm{diag}\Sigma)^{-1/2}\Sigma(\mathrm{diag}\Sigma)^{-1/2} = \left(\frac{\sigma_{ij}}{\sigma_{ii}^{1/2}\sigma_{jj}^{1/2}}\right) = (\rho_{ij}). \tag{1.27}$$

If $y \sim N_p(\mu, \Sigma)$ and $w = F^{-1}(y-\mu)$, then the quadratic form $(y-\mu)'\Sigma^{-1}(y-\mu)$ has a chi-square distribution with p degrees of freedom, written as

$$w'w = (y-\mu)'\Sigma^{-1}(y-\mu) \sim \chi_p^2. \tag{1.28}$$

The quantity $\left[(y-\mu)'\Sigma^{-1}(y-\mu)\right]^{1/2}$ is called the Mahalanobis distance between y and μ.

For a random sample of n independent p-vectors $(y_1, y_2, \ldots, y_n)$ from a multivariate normal distribution, $y_i \sim IN_p(\mu, \Sigma)$, we shall in general write the data matrix $Y_{n\times p}$ in the classical form

$$Y_{n\times p} = \begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_n' \end{pmatrix} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix}. \tag{1.29}$$

The corresponding responsewise representation for Y is

$$Y_{p\times n} = \begin{pmatrix} y_1 & y_2 & \dots & y_n \end{pmatrix} = \begin{pmatrix} y_{11} & y_{12} & \dots & y_{1n} \\ y_{21} & y_{22} & \dots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{p1} & y_{p2} & \dots & y_{pn} \end{pmatrix}. \tag{1.30}$$

The joint probability density function (pdf) for $(y_1, y_2, \dots, y_n)$ or the likelihood function is

$$L = L(\mu, \Sigma|y) = \prod_{i=1}^{n} f(y_i). \tag{1.31}$$

Substituting $f(y)$ in (1.26) for each $f(y_i)$, the pdf for the multivariate normal distribution is

$$[(2\pi)^p|\Sigma|]^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)'\Sigma^{-1}(y_i-\mu)\right]. \tag{1.32}$$

Using the property of the trace of a matrix, $\mathrm{tr}(x'Ax) = \mathrm{tr}(Axx')$, (1.32) may be written as

$$[(2\pi)^p|\Sigma|]^{-n/2} \,\mathrm{etr}\left\{-\frac{1}{2}\Sigma^{-1}\left[\sum_{i=1}^{n}(y_i-\mu)(y_i-\mu)'\right]\right\} \tag{1.33}$$

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

If we let the sample mean be represented by

$$\bar{y} = n^{-1}\sum_{i=1}^{n} y_i \tag{1.34}$$

and the sum of squares and products (SSP) matrix, using the classical form (1.29), is

$$E = \sum_{i=1}^{n}(y_i-\bar{y})(y_i-\bar{y})' = Y'Y - n\bar{y}\bar{y}' \tag{1.35}$$

or using the responsewise form (1.30), the SSP matrix is

$$E = YY' - n\bar{y}\bar{y}'. \tag{1.36}$$

In either case, we may write (1.33) as

$$[(2\pi)^p|\Sigma|]^{-n/2} \,\mathrm{etr}\left\{-\frac{1}{2}\Sigma^{-1}\left[E + n(y_i-\mu)(y_i-\mu)'\right]\right\} \tag{1.37}$$

so that by Neyman's factorization criterion $(E, \bar{y})$ are sufficient statistics for estimating (Σ, μ), (Lehmann, 1991, p. 16).

Theorem 1.1. *Let $y_i \sim IN_p(\mu, \Sigma)$ be a sample of size n, then $\bar{y}$ and E are sufficient statistics for μ and Σ.*

It can also be shown that $\bar{y}$ and E are independently distributed. The distribution of 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

$$c|\Sigma|^{-\nu/2}|E|^{(\nu-p-1)/2}\text{etr}\left(-\frac{1}{2}\Sigma^{-1}E\right) \tag{1.38}$$

where c is an appropriately chosen constant so that the total probability is equal to one. We write that $E \sim W_P(\nu, \Sigma)$. The expectation of E is $\nu\Sigma$.

Given a random sample of observations from a multivariate normal distribution, we usually estimate the parameters μ and Σ.

Theorem 1.2. *Let $y_i \sim IN_p(\mu, \Sigma)$, then the maximum likelihood estimators (MLEs) of μ and Σ are $\bar{y}$ and $E/n = \widehat{\Sigma}$.*

Furthermore, $\bar{y}$ and

$$S = \left(\frac{n}{n-1}\right)\widehat{\Sigma} = \frac{E}{n-1} \tag{1.39}$$

are unbiased estimators of μ and Σ, so that $\mathcal{E}(\bar{y}) = \mu$ and $\mathcal{E}(S) = \Sigma$. Hence, the sample distributions of S is Wishart, $(n-1)S \sim W_p(\nu = n-1, \Sigma)$ or $S = (s_{ij}) \sim W_p[1, \Sigma/(n-1)]$. Since S is proportional to the MLE $\widehat{\Sigma}$ of Σ, the MLE of the population correlation coefficient matrix is

$$R = (\text{diag}S)^{-1/2}S(\text{diag}S)^{-1/2} = \left(\frac{s_{ij}}{s_{ii}^{1/2}s_{jj}^{1/2}}\right) = (r_{ij}), \tag{1.40}$$

where r_{ij} is the sample correlation coefficient.

For a random sample of n independent identically distributed (iid) p-vectors $(y_1, y_2, \ldots, y_n)$ from any distribution with mean μ and covariance matrix Σ, by the central limit theorem (CLT), the pdf for the random variable $z = \sqrt{n}(\bar{y} - \mu)$ converges in distribution to a multivariate normal distribution with mean 0 and covariance matrix Σ,

$$z = \sqrt{n}(\bar{y} - \mu) \xrightarrow{d} N_p(0, \Sigma). \tag{1.41}$$

And, the quadratic form,

$$n(\bar{y} - \mu)'\Sigma^{-1}(\bar{y} - \mu) = z'z \xrightarrow{d} \chi^2_p, \tag{1.42}$$

converges in distribution to a chi-square distribution with p degrees of freedom. The quantity $\left[(\bar{y} - \mu)'\Sigma^{-1}(\bar{y} - \mu)\right]^{1/2}$ is the Mahalanobis distance from $\bar{y}$ to μ.

The mean μ and covariance matrix Σ are the first two moments of a random vector y. We now extend the classical measures of skewness and kurtosis, $\mathcal{E}(y-\mu)^3/\sigma^3$ and the $\mathcal{E}(y-\mu)^4/\sigma^4$, to the multivariate case following Mardia (1970). Letting $y_i \sim (\mu, \Sigma)$, Mardia's sample measures of multivariate skewness $(\beta_{1,p})$ and kurtosis $(\beta_{2,p})$ are based upon the scaled random variables, $z_i = S^{-1/2}(y_i - \bar{y})$, $i = 1, \ldots, n$. Mardia's measures of the population values are respectively,

$$b_{1,p} = \frac{1}{n^2} \sum_{i,j=1}^{n} \left[(y_i - \bar{y})' S^{-1} (y_i - \bar{y})\right]^3 \tag{1.43}$$

$$b_{2,p} = \frac{1}{n} \sum_{i,j=1}^{n} \left[(y_i - \bar{y})' S^{-1} (y_i - \bar{y})\right]^2 . \tag{1.44}$$

When $y_i \sim IN_p(\mu, \Sigma)$, the population values of the moments are $\beta_{1,p} = 0$ and $\beta_{2,p} = p(p+2)$. Under normality, Mardia showed that the statistic $X^2 = nb_{1,p}/6$ converges to a chi-square distribution with $\nu = p(p+1)(p+2)/6$ degrees of freedom. And, that the multivariate kurtosis statistic converges to a normal distribution with mean $\mu = p(p+2)$ and variance $\sigma^2 = 8p(p+2)/n$. When $n > 50$, one may use the test statistics to evaluate multivariate normality. Rejection of normality indicates either the presence of outliers or that the distribution is significantly different from a multivariate normal distribution. Romeu and Ozturk (1993) performed a comparative study of goodness-of-fit tests for multivariate normality and showed that Mardia's tests are most stable and reliable. They also calculated small sample empirical critical values for the tests.

When one finds that a distribution is not multivariate normal, one usually replaces the original observations with some linear combination of variables which may be more nearly normal. Alternatively, one may transform each variable in the vector using a Box-Cox power transformation, as outlined for example by Bilodeau and Brenner (1999, p. 95). However, because marginal normality does not ensure ensure multivariate normality a joint transformation may be desired. Shapiro and Wilk (1965) W statistic or Royston (1982, 1992) approximation may be used to test for univariate normality one variable at a time. Since marginal normality does not ensure multivariate normality, a multivariate test must be used to evaluate joint normality for a set of p variables.

1.5 Elementary Properties of Normal Random Variables

Theorem 1.3. *Let $y \sim IN_p(\mu, \Sigma)$ and $w = A_{m\times p} y$, then $w \sim IN_m(A\mu, A\Sigma A')$.*

Thus, linear combinations of multivariate normal random variables are again normally distributed. If one assumes that the random variable y is multivariate normal, $y \sim N_n(X\beta = \mu, \Omega = \sigma^2 I)$, and $\hat{\beta}$ is an unbiased estimate of β such that $\hat{\beta} = (X'X)^{-1}X'y$ then by Theorem 1.3,

$$\hat{\beta} \sim N_k \left[\beta, \operatorname{cov}\left(\hat{\beta}\right)\right]. \tag{1.45}$$

Given that a random vector y is multivariate normal, one may partition the p elements of a random vector y into two subvectors y_1 and y_2 where the number of elements $p = p_1 + p_2$. The joint distribution of the partitioned vector is multivariate normal and written as

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N_p \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right]. \tag{1.46}$$

Given the partition of the vector y, one is often interested in the conditional distribution of y_1 given the subset y_2, $y_1|y_2$. Provided Σ_{22} is nonsingular, we have the following general result.

Theorem 1.4. *$y_1|y_2 = z \sim N_{p_1}(\mu_{1\cdot 2}, \Sigma_{11\cdot 2})$, where $\mu_{1\cdot 2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2)$ and $\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.*

Letting the subset y_1 contain a single element, then

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N_p \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12}' \\ \sigma_{21} & \Sigma_{22} \end{pmatrix} \right]. \tag{1.47}$$

The multiple correlation coefficient R is the maximum correlation possible between y_1 and the linear combination of the random vector y_2, $a'y_2$. Using the Cauchy-Schwarz inequality, one can show that the multiple correlation coefficient $R = [\sigma_{12}'\Sigma_{22}^{-1}\sigma_{21}/\sigma_{11}]^{1/2} \geq 0$, which is seen to be the correlation between y_1 and $z = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2)$. The sample, biased overestimate of R^2, the squared multiple correlation coefficient, is $\hat{R}^2 = s_{12}'S_{22}^{-1}s_{21}/s_{11}$. Even if $R^2 = 0$, $\mathcal{E}(\hat{R}^2) = (p-1)/(n-1)$. Thus, if the sample size n is small relative to p, the bias can be large.

The partial correlation coefficient between variables y_i and y_j is the ordinary simple correlation ρ between y_i and y_j with the variables in the subset y_2 held fixed and represented by $\rho_{ij|y_2}$. Letting $\Sigma_{11\cdot 2} = (\sigma_{ij|y_2})$, the matrix of partial variances and covariances, the partial correlation between the (i, j) element is

$$\rho_{ij|y_2} = \frac{\sigma_{ij|y_2}}{\sigma_{ii|y_2}^{1/2}\sigma_{jj|y_2}^{1/2}}. \tag{1.48}$$

Replacing $\Sigma_{11\cdot 2}$ with the sample estimate $S_{11\cdot 2} = (s_{ij|y_2})$, the MLE of $\rho_{ij|y_2}$ is

$$r_{ij|y_2} = \frac{s_{ij|y_2}}{s_{ii|y_2}^{1/2}s_{jj|y_2}^{1/2}}. \tag{1.49}$$

1.6 Hypothesis Testing

Having assumed a linear model for a random sample of observations, used the observations to obtain an estimate of the population parameters, and decided upon the structure of the restriction R (if any) and the hypothesis test matrix C, one next test hypotheses. Two commonly used procedures for testing hypotheses are

the likelihood ratio (LR) and union-intersection (UI) test. To construct a 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 (1.20), and the likelihood $L(\hat{\Omega}_0)$, the likelihood function maximized over the entire parameter space Ω_0 unconstrained by the hypothesis. Defining λ as the ratio

$$\lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega}_0)} \tag{1.50}$$

the hypothesis is rejected for small values of λ since $L(\hat{\omega}) < L(\hat{\Omega}_0)$ does not favor the hypothesis. The test is said to be of size α if for a constant λ_0, the

$$P(\lambda < \lambda_0 | H) = \alpha \tag{1.51}$$

where α 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 ν is equal to the number of independent parameters estimated under Ω_0 minus the number of independent parameters estimated under ω.

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

$$H : \bigcap_i H_i \tag{1.52}$$

and each H_i is associated with an alternative A_i such that

$$A : \bigcup_i A_i. \tag{1.53}$$

The null hypothesis H is rejected if any elementary test of size α is rejected. The overall rejection region being 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 $T_i \lesseqgtr c_\alpha$ where the

$$P\left(\sup_i T_i \leq c_\alpha | H\right) = 1 - \alpha \tag{1.54}$$

and c_α is chosen such that the Type I error is α.

1.7 Generating Multivariate Normal Data

In hypothesis testing of both univariate and multivariate linear models, the assumption of multivariate normality is made. The multivariate normal distribution of a random vector y with p variables has the density function given in (1.26), written as $y \sim N_p(\mu, \Sigma)$. For $p = 1$, the density function reduces to the univariate normal

distribution. Some important properties of normally distributed random variables were reviewed in Section 1.5. To generate data having a multivariate normal distribution with mean $\mu' = (\mu_1, \mu_2, \ldots, \mu_p)$ and covariance matrix $\Sigma = (\sigma_{ij})$, we use Definition 1.1. Program 1_7.sas uses the IML procedure to generate 50 observations from a multivariate normal distribution with structure

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

In Program 1_7.sas, PROC IML is used to produce a matrix Z that contains $n = 50$ observation vectors with $p = 4$ variables. Each vector is generated by using the standard normal distribution, $N(0, 1)$. Using Theorem 1.3, new variables $y_i = z_i A + \mu$ are created where the matrix A is such that the $\text{cov}(y) = A'\text{cov}(z)A = A'IA = A'A = \Sigma$ and $E(y) = \mu$. The Cholesky factorization procedure is used to obtain A from Σ: the ROOT function in PROC IML performs the Cholesky decomposition and stores the result in the matrix named `a` in the program. Next, the matrix `u` is created by repeating the mean row vector `u` 50 times to produce a 50×4 data matrix. The multivariate normal random variables are created using the statement: `y=(z*a)+uu`. The observations are printed and output to the file named 1_7.dat. The seed in the program allows one to always create the same data set.

1.8 Assessing Univariate Normality

Before one tests hypotheses, it is important that one examines the distributional assumptions for the sample data under review. While the level of the test (Type I error) for means is reasonably robust to nonnormality, this is not the case when investigating the covariance structure. However, very skewed data and extreme outliers may result in errors in statistical inference of the population means. Thus, one usually wants to verify normality and investigate data for outliers.

If a random vector y is distributed multivariate normally, then its components y_i are distributed univariate normal. Thus, one step in evaluating multivariate normality of a random vector is to evaluate the univariate normality of it components. One can construct and examine histograms, stem-and-leaf plots, box plots, and Quantile-Quantile (Q-Q) probability plots for the components of a random p-vector.

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).

One may also evaluate normality by performing the Shapiro and Wilk (1965) W test when sample sizes are less than or equal to 50. The test is known to show a reasonable sensitivity to nonnormality (Shapiro, Wilk, & Chen, 1968). For $50 \leq n \leq$

2000, Royston (1982, 1992) approximation is recommended and is implemented in the SAS procedure UNIVARIATE.

When individual variables are found to be nonnormal, one can often find a Box and Cox (1964) power transformation that may be applied to the data to achieve normality. The Box-Cox power transformation has the general structure for $y > 0$:

$$x = \begin{cases} \frac{(y^{\lambda}-1)}{\lambda} & : \lambda \neq 0 \\ \log(y) & : \lambda = 0 \end{cases}.$$

Note that the random dependent random variable y must be positive for all values. Thus, one may have to add a constant to the sample data before applying the transformation. After applying a Box-Cox type transformation, one should again check whether transformed data are more nearly normal.

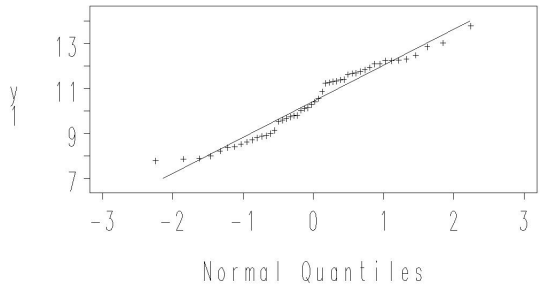
1.8.1 Normally and Nonnormally Distributed Data

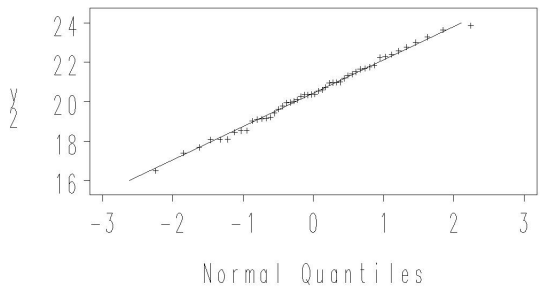
Program 1_8_1.sas produces Q-Q plots for the univariate normally distributed random variables generated by Program 1_7.sas. The Q-Q plots for the normal data show that the observations lie close to a line, but not exactly; the tail, especially, falls off from the line (Graph 1.8.1). Recall that we know that these data are normally distributed. Thus, when using Q-Q plots for diagnostic purposes, we cannot expect that even normally distributed data will lie exactly on a straight line. Note also that some of the Shapiro-Wilk test statistics are significant at the nominal $\alpha = 0.05$ level even for normal data. When performing tests for sample sizes less than or equal to 50, it is best to reduce the level of the normality test to the nominal $\alpha = 0.01$ level.

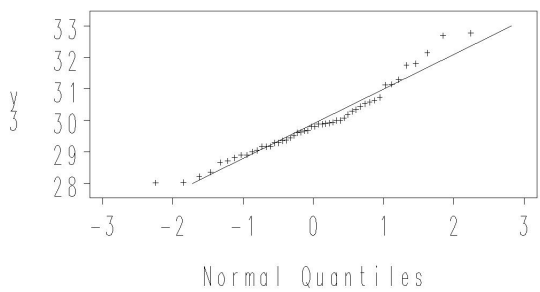
In Program 1_8_1.sas we transform the normal data using the transformations: $ty1 = 1/y^2$, $ty2 = e^y$, $ty3 = \log(y)$, and $ty4 = y^2$ for the four normal variables and generating Q-Q plots for the transformed nonnormal data (Graph 1.8.2). Inspection of the plots clearly show marked curvilinear patterns. The plots are not linear. Next we illustrate how one may find a Box-Cox power transformation using the transformed variable $x = ty4$. To achieve normality the value of λ should be near one-half, the back transformation for the variable. Using the macro `%adxgen` and `%adxtran` in SAS, the value $\lambda = -0.4$ (found in the log file) is obtained for the data; the transformed data are stored in the data set named result. The λ value is near the correct value of -0.5 (or a square root transformation) for the variable. The λ plot in the output indicates that the value of λ should be within the interval: $-0.2 \leq \lambda \leq -0.6$. While the macro uses the minimal value, one often trys other values within the interval to attain near normality.

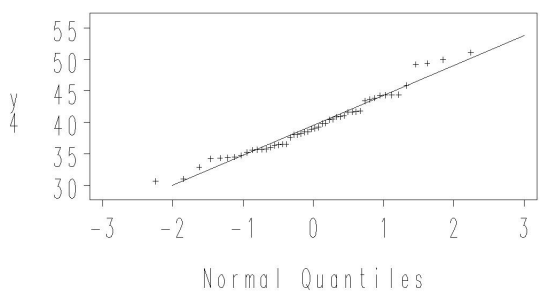
Finally, we introduce an outlier into the normally distributed data, and again generate Q-Q plots (Graph 1.8.3). Inspection of the Q-Q plot clearly shows the extreme observation. The data point is far from the line.

Graph 1.8.1: Q-Q plot of y1-y4

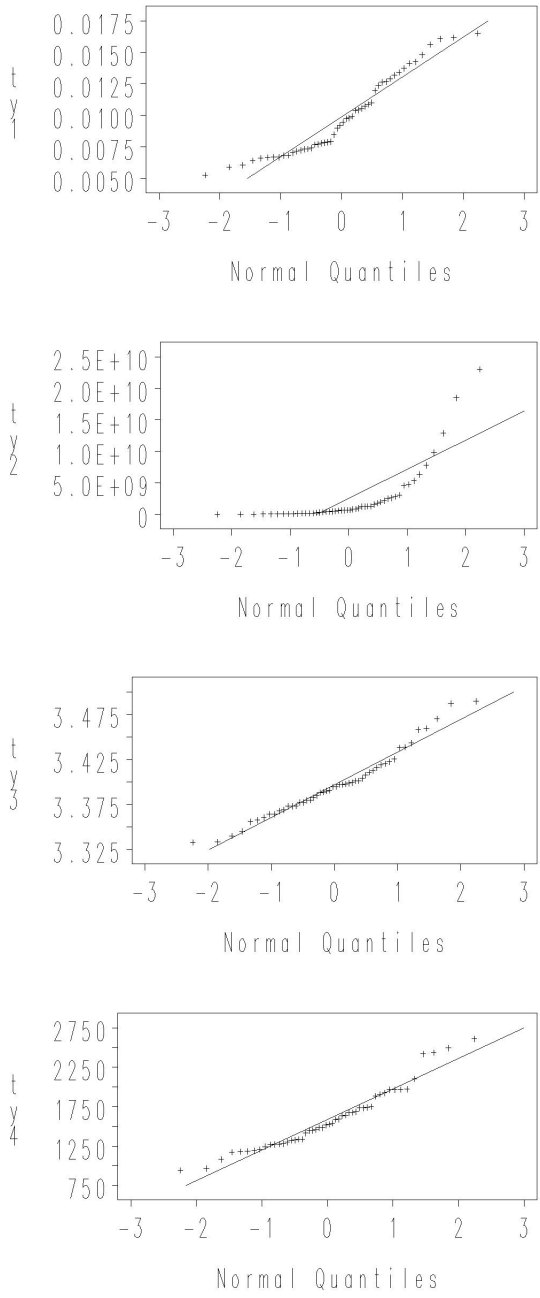




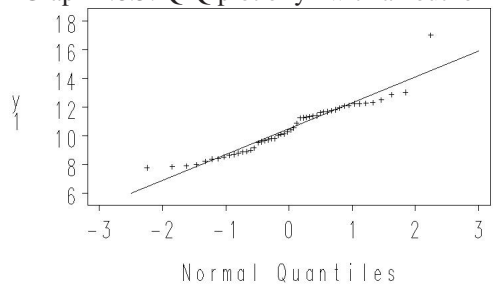




Graph 1.8.2: Q-Q plot of ty1-ty4



Graph 1.8.3: Q-Q plot of y1 with an outlier



1.8.2 Real Data Example

To illustrate the application of plots and tests to evaluate normality, data from Rohwer given in (Timm, 2002, p. 213) are used. The data are in the data set Rohwer.dat.

The data are for 32 selected school children in an upper-class, white residential school and contain three standardized tests: Peabody Picture Vocabulary (`y1`), Student Achievement (`y2`), and the Raven Progressive Matrices test (`y3`) and five paired-associate, learning-proficiency tasks: Named (`x1`), Still (`x2`), Named Still (`x3`), Named Action (`x4`), and sentence still (`x5`). While we will use the data to evaluate multivariate prediction in Chapter 5, we use the raw data to investigate the normality of the three standardized test variables using univariate Q-Q plots and tests for univariate normality. Also illustrated is the use of the Box-Cox transformation. The code for the analysis is contained in Program 1_8_2.sas.

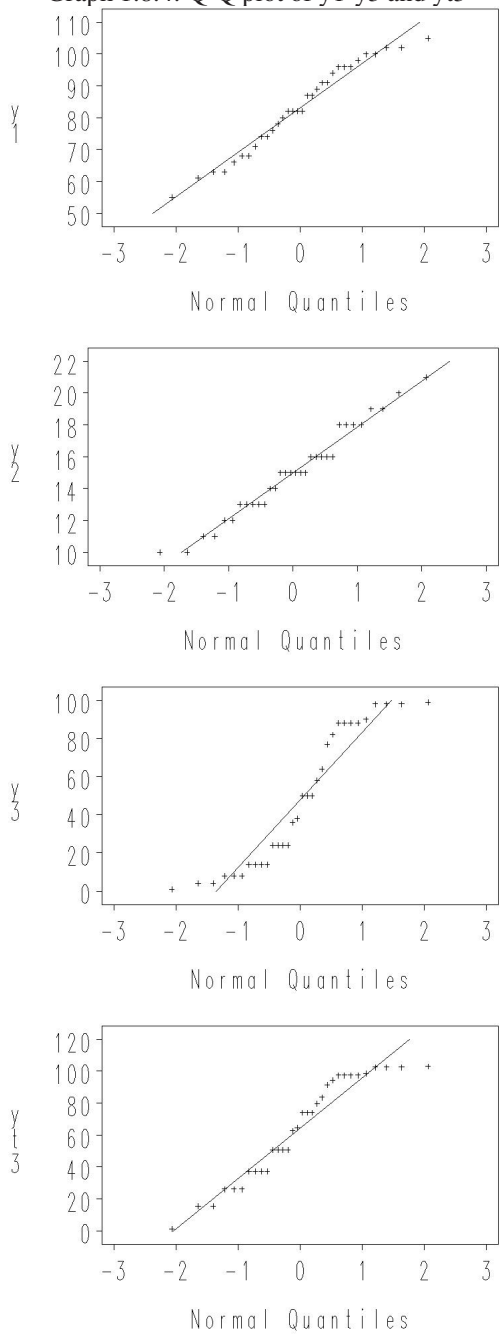
The program produces Q-Q plots for each dependent variable (Graph 1.8.4). Review of the plots indicates that the variables `y1` and `y2` appear normal. However, this is not the case for the variable `y3`. Using the Box-Cox power transformation, a value of $\lambda = 0.4$ is used to transform the data to near normality, `yt3`. The plot is more nearly linear and the Shapiro-Wilk test appears to marginally support normality at the nominal level $\alpha = 0.01$ for the transformed data.

1.9 Assessing Multivariate Normality with Chi-Square Plots

Even though each variable in a vector of variables is normally distributed, marginally normality does not ensure multivariate normality. However, multivariate normality does ensure marginal normality. Thus, one often wants to evaluate whether or not a vector of random variables follows a multivariate normal distribution. To evaluate multivariate normality, one may compute the Mahalanobis distance for the i^{th} observation:

$$D_i^2 = (y_i - \bar{y})' S^{-1} (y_i - \bar{y}) \tag{1.55}$$

Graph 1.8.4: Q-Q plot of y1-y3 and yt3



and plot these distances against the ordered chi-square percentile values $q_i = \chi^2_p \cdot [(i - 1/2)/n]$ where q_i $(i = 1, 2, \ldots, n)$ is the $100\,(i - 1/2)\,/n$ sample quantile of the chi-square distribution.

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) constructs multivariate probability plots with the ordered Mahalanobis distances versus quantiles from a beta distribution, rather than a chi-square distribution. The exact distribution of $b_i = nD_i^2/(n-1)^2$ follows a beta $[a = p/2,$ $b = (n-p-1)/2]$ distribution (Gnanadesikan & Kettenring, 1972). Small (1978) found that as p gets large ($p > 5\%$ of n) relative to n that the chi-square approximation may not be adequate unless $n \geq 25$ and in these cases recommends a beta plot.

When evaluating multivariate normality, one should also compute measures of multivariate skewness and kurtosis. If data follow a multivariate normal distribution, 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) defined the measures of multivariate skewness and kurtosis:

$$\beta_{1,p} = \mathcal{E}[(x-\mu)'\Sigma^{-1}(y-\mu)]^3 \tag{1.56}$$

$$\beta_{2,p} = \mathcal{E}[(y-\mu)'\Sigma^{-1}(y-\mu)]^2 \tag{1.57}$$

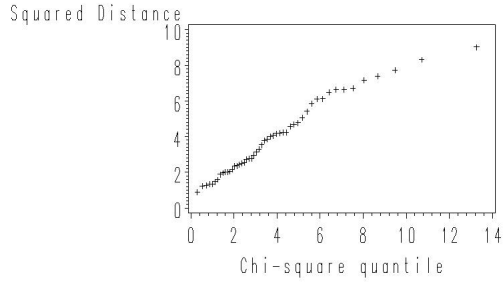
where x and y are identically and independently distributed. Sample estimates of these quantities are:

$$\hat{\beta}_{1,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[(y_i-\bar{y})'S^{-1}(y_j-\bar{y})\right]^3 \tag{1.58}$$

$$\hat{\beta}_{2,p} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[(y_i-\bar{y})'S^{-1}(y_j-\bar{y})\right]^2. \tag{1.59}$$

If $y \sim N_p(\mu, \Sigma)$, then $\beta_{1,p} = 0$ and $\beta_{2,p} = p(p+2)$. Mardia showed that the sample estimate of multivariate kurtosis $X^2 = n\hat{\beta}_{1,p}/6$ has an asymptotic chi-square distribution with $\nu = p(p+1)(p+2)/6$ degrees of freedom. And that $Z = [\hat{\beta}_{2,p} - p(p+2)]/[8p(p+2)/n]^{1/2}$ converges in distribution to a standard normal distribution. Provided the sample size $n \geq 50$ one may develop tests of multivariate normality. Mardia (1974) developed tables of approximate percentiles for $p = 2$ and $n \geq 10$ and alternative large sample approximations. Romeu and Ozturk (1993) investigated ten tests of goodness-of-fit for multivariate normality. They show that the multivariate tests of Mardia are most stable and reliable for assessing multivariate normality. In general, tests of hypotheses regarding means are sensitive to high values of skewness and kurtosis for multivariate data.

Graph 1.9.1: Chi-square Q-Q plot



Output 1.9.1: MULTNORM Macro Univariate and Multivariate Normality Tests for y1-y4.

Variable	n	Test	Multivariate Skewness & Kurtosis	Test Statistic Value	p-value
y1	50	Shapiro-Wilk	.	0.95519	0.0560
y2	50	Shapiro-Wilk	.	0.99170	0.9775
y3	50	Shapiro-Wilk	.	0.95347	0.0475
y4	50	Shapiro-Wilk	.	0.96571	0.1540
	50	Mardia Skewness	1.0846	9.81142	0.9715
	50	Mardia Kurtosis	20.5357	-1.76789	0.0771

While Andrews, Gnanadesikan, and Warner (1971) have developed a multivariate extension of the Box-Cox power transformation for multivariate data, determination of the appropriate transformation is complicated (see, Chambers, 1977; Velilla & Barrio, 1994). In general, one applies the Box-Cox transformation a variable at a time or uses some linear combination of the variables in the analysis when multivariate normality is not satisfied.

1.9.1 Multivariate Normal Data

To illustrate the construction of a chi-square plot, the data in the multivariate data set 1_7.dat are used. Program 1_9_1.sas contains the code for the chi-square plots. The program uses the SAS macro `%multnorm` which calculates Mardia's test statistics for multivariate skewness and kurtosis and also the Shapiro-Wilk W statistics for each variable. Inspection of the plot and the multivariate statistics indicate that the data are clearly multivariate normal (Graph 1.9.1, Output 1.9.1).

1.9.2 Real Data Example

Using the Rohwer data set described in Section 1.8.2, we developed chi-square plots for the raw data and the transformed data. Program 1_9_2.sas contains the code for the example. The p-values for Mardia Skewness and Kurtosis for the raw data are: 0.93807 and 0.05000, respectively. Upon transformation of the third variable, the corresponding values become: 0.96408 and 0.05286. While the data are nearly normal, the transformation does not show a significant improvement in joint normality.

1.10 USING SAS INSIGHT

Outliers in univariate data only occur in the tail of the Q-Q plot since the plots are based upon ordered variables. However, for multivariate data this is not the case since multivariate vector observations cannot be ordered. Instead, ordered squared distances are used so that the location of an outlier within the distribution is uncertain. It may involve any distance in the multivariate chi-square Q-Q plot. To evaluate the data for potential outliers, one may use the tool SAS INSIGHT interactively.

When SAS is executed, it creates temporary data sets in the Library `WORK`. To access the Library interactively, click on `Solution` → `Analysis` → `Interactive Data Analysis`. This executes the SAS INSIGHT software. Using SAS INSIGHT, click on the data set called `WORK`. The data sets used and created by the SAS program are displayed. For the multivariate Q-Q plot, select the data set `_CHIPLOT`. Displayed will be the coordinates of the multivariate Q-Q plot. From the tool bar select `Analyze` → `Fit(Y X)`. This will invoke a `Fit(Y X)` software window; next, move the variables `MAHDIST` to the window labeled `Y` and the variable `CHISQ` to the window labeled `X`. Then, select `Apply` from the menu. This will produce the multivariate Q-Q plot generated by macro `%multnorm` (Graph 1.10.1). The observation number will appear by clicking on a data point. By double clicking on a value, the window `Examine Observations` appears which display the residual and predicted squared distances (Figure 1.10.1). Also contained in the output is a plot of these values. By clicking on data points, extreme observations are easily located. To illustrate the use of SAS INSIGHT two data sets using real data are investigated.

1.10.1 Ramus Bone Data

To illustrate the use of SAS INSIGHT for the location of outliers Ramus bone data from Elston and Grizzle (1962) are used. The data are in the data set Ramus.dat. Using Program 1_10_1.sas, the data are investigated for normality. One observes that while all variables are univariate normal, the test of multivariate normality is rejected (Output 1.10.1). This is due in part to the small sample size. Following the procedure discussed above, we observe that observation 9 appears extreme (Graph 1.10.1). Removing this observation from the data set using Program 1_10_1a.sas, the data become more normal, but remain skewed (Output 1.10.2). For a multivariate

Graph 1.10.1: Chi-square Q-Q plot generated by SAS INSIGHT

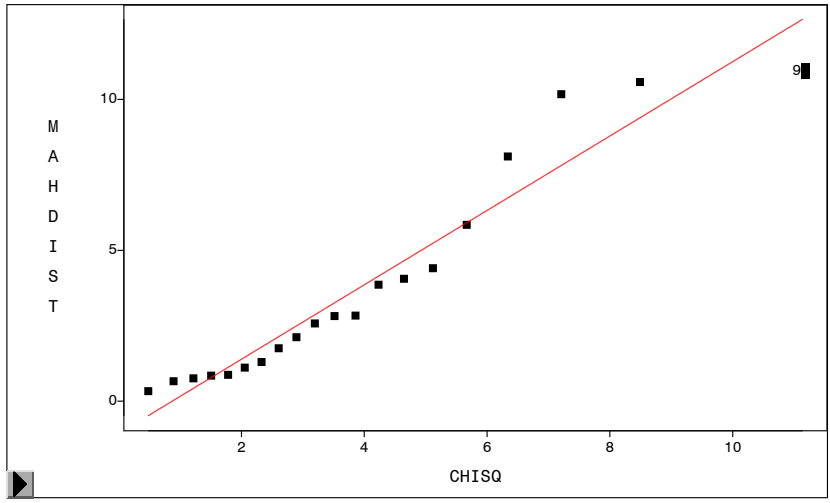
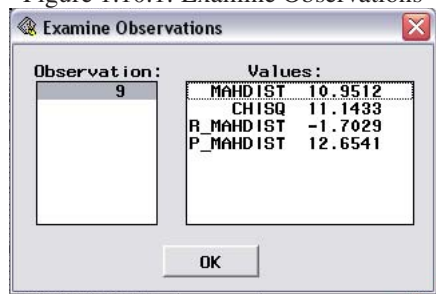


Figure 1.10.1: Examine Observations



Output 1.10.1: MULTNORM Macro Univariate and Multivariate Normality Tests for Ramus Bone Data.

Variable	n	Test	Multivariate Skewness & Kurtosis	Test Statistic Value	p-value
y1	20	Shapiro-Wilk	.	0.9479	0.3360
y2	20	Shapiro-Wilk	.	0.9628	0.6020
y3	20	Shapiro-Wilk	.	0.9578	0.5016
y4	20	Shapiro-Wilk	.	0.9180	0.0905
	20	Mardia Skewness	11.3431	46.1170	0.0008
	20	Mardia Kurtosis	28.9174	1.5871	0.1125

Output 1.10.2: MULTNORM Macro Univariate and Multivariate Normality Tests for Ramus Bone Data without Observation 9.

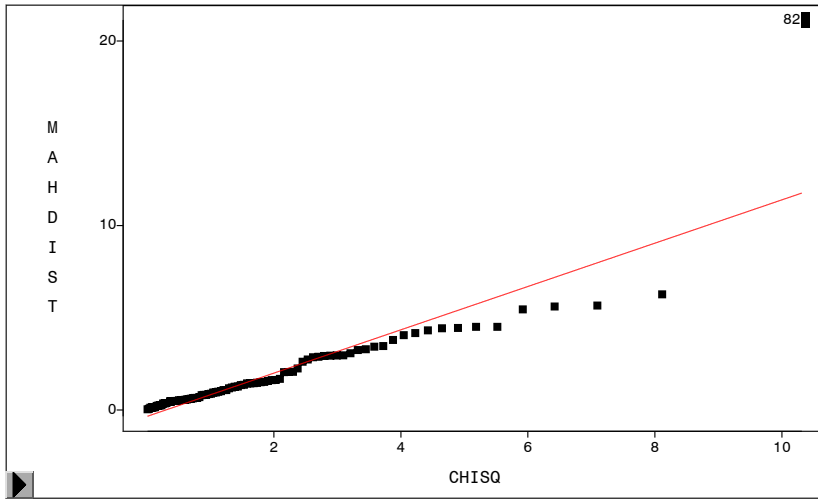
Variable	n	Test	Multivariate Skewness & Kurtosis	Test Statistic Value	p-value
y1	19	Shapiro-Wilk	.	0.9436	0.3064
y2	19	Shapiro-Wilk	.	0.9519	0.4249
y3	19	Shapiro-Wilk	.	0.9533	0.4490
y4	19	Shapiro-Wilk	.	0.9210	0.1180
	19	Mardia Skewness	11.0359	43.0477	0.0020
	19	Mardia Kurtosis	29.0259	1.5810	0.1139

analysis of this data set, one should consider linear combination of the Ramus data over the years of growth since the skewness is not easily removed from the data.

1.10.2 Risk-Taking Behavior Data

For our second example, data from a large study by Dr. Stanley Jacobs and Mr. Ronald Hritz at the University of Pittsburgh are used. Students were assigned to three experimental conditions and administered two parallel forms of a test given under high and low penalty. The data set is in the file Stan_Hz.dat. Using Program 1_10_2.sas, the data are investigated for multivariate normality. The test of multivariate normality is clearly rejected (Output 1.10.3). Using SAS INSIGHT, observation number 82 is clearly an outlier (Graph 1.10.2). Removing the observation (Program 1_10_2a.sas), the data are restored to multivariate normality (Output 1.10.4). These examples clearly indicate the importance of removing outliers from multivariate data.

Graph 1.10.2: Chi-square Q-Q plot of risk-taking behavior data



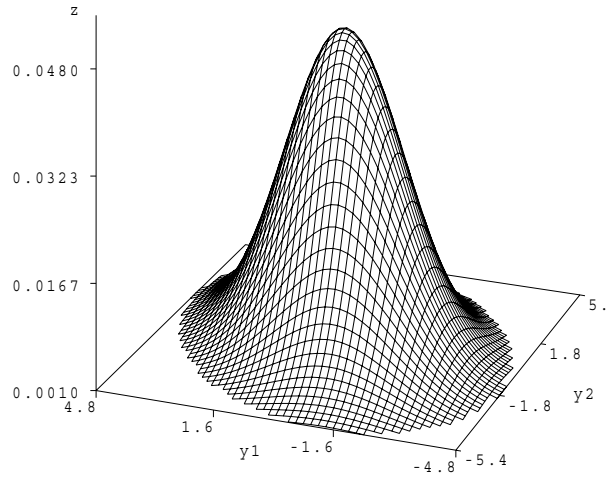
Output 1.10.3: MULTNORM Macro Univariate and Multivariate Normality Tests of Risk-taking Behavior Data.

Variable	n	Test	Multivariate Skewness & Kurtosis	Test Statistic Value	p-value
resL	87	Shapiro-Wilk	.	0.9888	0.6674
resH	87	Shapiro-Wilk	.	0.9520	0.0027
	87	Mardia Skewness	0.7450	11.4348	0.0221
	87	Mardia Kurtosis	10.7652	3.2240	0.0013

Output 1.10.4: MULTNORM Macro Univariate and Multivariate Normality Tests of Risk-taking Behavior Data with Observation 82.

Variable	n	Test	Multivariate Skewness & Kurtosis	Test Statistic Value	p-value
resL	86	Shapiro-Wilk	.	0.98969	0.7354
resH	86	Shapiro-Wilk	.	0.97449	0.0863
	86	Mardia Skewness	0.13474	2.04568	0.7274
	86	Mardia Kurtosis	7.04198	-1.11054	0.2668

Graph 1.11.1: Bivariate Normal Distribution with u=(0, 0), var(y1)=3, var(y2)=4, cov(y1,y2)=1, r=.289



1.11 THREE-DIMENSIONAL PLOTS

Three dimensional scatter plots of multivariate data often help with the visualization of data. They are generated using the G3D procedure in SAS. The first part of Program 1_11.sas (adapted from Khattree & Naik, 1995, p. 65) produces a plot of a bivariate normal distribution with mean and covariance matrix:

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 3 & 1 \\ 1 & 4 \end{pmatrix}.$$

This is the covariance matrix of variables `y1` and `y2` from the simulated multivariate normal data generated by Program 1_7.sas. The three-dimensional plot is given in Graph 1.11.1.

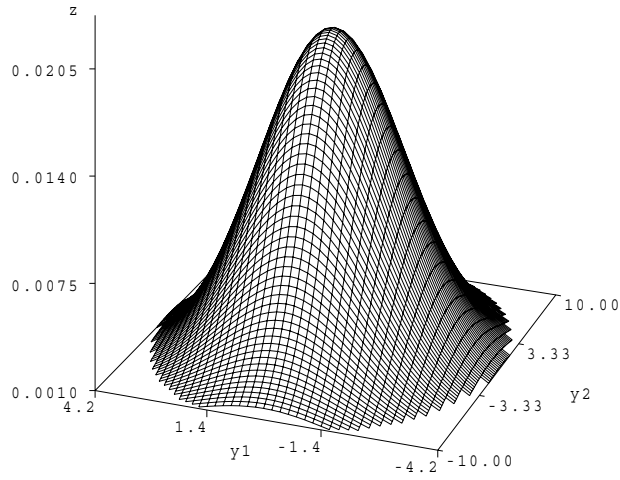
To see how plots vary, a second plot is generated in Program 1_11.sas using variables `y1` and `y4` from the simulated data in data set 1_7.dat. The covariance matrix for population parameters for the plot are:

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 3 & 0 \\ 0 & 20 \end{pmatrix}.$$

The plot is displayed in Graph 1.11.2.

For the first plot, a cross-wise plot would result in an oval shape, whereas in the second plot, a circular shape results. This is due to the structure of the covariance matrix. Using SAS INSIGHT for the data set, one may generate contour plots for

Graph 1.11.2: Bivariate Normal Distribution with u=(0, 0), var(y1)=3, var(y2)=20, cov(y1,y2)=0, r=0



the data. See Khattree and Naik (1995) for more graphical displays of multivariate data using SAS.

CHAPTER 2

Unrestricted General Linear Models

2.1 INTRODUCTION

Unrestricted (univariate) linear models are linear models that specify a relationship between a set of random, independent, identically distributed (iid) dependent variables $y' = (y_1, y_2, \ldots, y_n)$ and a matrix of fixed, nonrandom, independent variables $X = (x_{ik})$ such that

$$\mathcal{E}(y_i) = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ik}\beta_k \qquad i = 1, 2, \ldots, n. \tag{2.1}$$

The variance of each y_i is constant (σ^2) or homogeneous, and the relationship is linear in the unknown, nonrandom parameters $\beta' = (\beta_1, \beta_2, \ldots, \beta_k)$. Special classes of such models are called multiple linear regression models, analysis of variance (ANOVA) models, and intraclass covariance models. In this chapter, we review both ordinary least squares (OLS) and maximum likelihood (ML) estimation of the model parameters, hypothesis testing, model selection and prediction in multiple linear regression model, and the general linear mixed model (GLMM) is introduced. Estimation theory and hypothesis testing for the GLMM are not discussed until Chapter 11. Applications discussed include multiple linear regression analyses and the analysis of variance for several experimental designs.

2.2 LINEAR MODELS WITHOUT RESTRICTIONS

For multiple regression, ANOVA, and intraclass covariance models, we assume that the covariance matrix for the vector y has the structure

$$\Omega = \sigma^2 I_n \tag{2.2}$$

where I_n is an $n \times n$ identity matrix. The error structure for the observation is said to be homogeneous or spherical. Models of the form (1.1) with covariance structure (2.2) are called unrestricted (univariate) linear models.

To estimate β, the vector of unknown, nonrandom, fixed effects regression coefficients, the method of OLS is commonly utilized. The least squares criterion requires minimizing the error sum of squares, $\sum_{i=1}^{n} e_i^2 = \mathrm{tr}(ee')$, where $\mathrm{tr}(\cdot)$ is the trace

operator. Minimizing the error sum of squares leads to the normal equations

$$(X'X)\hat{\beta} = X'y. \tag{2.3}$$

Because X has full column rank, $\mathrm{rank}(X) = k$, the ordinary least squares estimator (OLSE) of β is the unique solution to the normal equation (2.3),

$$\hat{\beta} = (X'X)^{-1}X'y. \tag{2.4}$$

This is also called the best linear unbiased estimator (BLUE) of β since among all parametric functions $\psi = c'\beta$, $\hat{\psi} = c'\hat{\beta}$ is unbiased for ψ and has smallest variance. The mean and variance of the parametric functions are

$$\mathcal{E}(\hat{\psi}) = \psi \tag{2.5}$$

$$\mathrm{var}(\hat{\psi}) = \sigma^2 c'(X'X)^{-1}c. \tag{2.6}$$

If the matrix X is not of full rank k, one may either reparameterize the model to full rank or use a generalized inverse of $X'X$ in (2.4) to solve the normal equations.

Definition 2.1. A generalized inverse of a real matrix A is any matrix G that satisfies the condition $AGA = A$. The generalized inverse of A is written as $G = A^-$.

Because A^- is not unique, (2.3) has no unique solution if $X'X$ is not full rank k; however, linear combinations of β, $\psi = c'\beta$, may be found that are unique even though a unique estimate of β is not available. Several SAS procedures use a full rank design matrix while others do not; more will be said about this when the applications are discussed. For a thorough discussion of the analysis of univariate linear models see Searle (1971) and Milliken and Johnson (1984). Following Searle (1987), and Timm and Carlson (1975), we will usually assume in our discussion in this chapter that the design matrix X is of full rank.

2.3 HYPOTHESIS TESTING

Once the parameter β has been estimated, the next step is usually to test hypotheses about β. For hypothesis testing, we assume that the vector e follows a spherical multivariate normal distribution,

$$y \sim N_n(\mu = X\beta, \Omega = \sigma^2 I_n). \tag{2.7}$$

The MLE of the population parameters β and σ^2 assuming normality are

$$\hat{\beta}_{MLE} = (X'X)^{-1}X'y \tag{2.8}$$

$$\begin{aligned}\hat{\sigma}^2_{MLE} &= \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{n} \\ &= \frac{(y'y - n\bar{y}^2)}{n}\end{aligned} \tag{2.9}$$

where the likelihood function using (1.31) has the form

$$L(\sigma^2, \beta|y) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{(y - X\beta)'(y - X\beta)}{2\sigma^2}\right]. \tag{2.10}$$

We see that the OLSE (2.4) and the MLE (2.8) estimate of β are identical.

To test the hypothesis $H : C\beta = \xi$ (1.20), we may create a likelihood ratio test which requires maximizing (2.10) with respect to β and σ^2 under the hypothesis $L(\hat{\omega})$ and over the entire parameter space $L(\widehat{\Omega}_0)$. Over the entire parameter space, Ω_0, the MLE of β and σ^2 are given in (2.8). The corresponding estimates under the hypothesis are

$$\hat{\beta}_\omega = \hat{\beta} - (X'X)^{-1}C'[C(X'X)^{-1}C']^{-1}(C\hat{\beta} - \xi) \tag{2.11}$$

$$\hat{\sigma}^2_\omega = \frac{(y - X\hat{\beta}_\omega)'(y - X\hat{\beta}_\omega)}{n} \tag{2.12}$$

(see Timm, 1975, p. 178). Substituting the estimates under ω and Ω_0 into the likelihood function (2.10), the likelihood ratio defined in (1.50) becomes

$$\begin{aligned} \lambda = \frac{L(\hat{\omega})}{L(\widehat{\Omega}_0)} &= \frac{(2\pi\hat{\sigma}^2_\omega)^{-n/2}}{(2\pi\hat{\sigma}^2)^{-n/2}} \\ &= \left[\frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{(y - X\hat{\beta}_\omega)'(y - X\hat{\beta}_\omega)}\right]^{n/2} \end{aligned} \tag{2.13}$$

so that

$$\Lambda = \lambda^{2/n} = \frac{E}{E + H} \tag{2.14}$$

where

$$E = y'y - n\bar{y}^2 = y'(I - X(X'X)^{-1}X')y \tag{2.15}$$

$$H = (C\hat{\beta} - \xi)'[C(X'X)^{-1}C']^{-1}(C\hat{\beta} - \xi). \tag{2.16}$$

Details are provided in Timm (1975, 1993b, Chapter 3) and Searle (1987, Chapter 8).

The likelihood ratio test is to reject

$$H : C\beta = \xi \qquad \text{if } \Lambda < c \tag{2.17}$$

where c is determined such that the $P(\Lambda < c|H) = \alpha$. The statistic Λ is related to a beta distribution represented generally as U_{p,ν_h,ν_e} where p is the number of variables, ν_h is the degrees of freedom for the hypothesis, $\nu_h = \text{rank}(C) = g$, and ν_e is the degrees of freedom for error, $\nu_e = n - \text{rank}(X)$.

Theorem 2.1. *When* $p = 1$,

$$\left[\frac{\nu_e(1 - U_{1,\nu_h,\nu_e})}{\nu_h U_{1,\nu_h,\nu_e}}\right] = F_{\nu_h,\nu_e}. \tag{2.18}$$

From Theorem 2.1, we see that rejecting H for small values of U is equivalent to rejecting H for large values F where

$$F = \frac{H/\nu_h}{E/\nu_e} = \frac{MS_h}{MS_e} \tag{2.19}$$

is the F statistic, and MS_h refers to the mean square for hypothesis and MS_e refers to the mean square for error.

2.4 Simultaneous Inference

While the parametric function that led to the rejection of H may not be of interest to the researcher, one can easily find the combination of the parameters β that led to rejection. To find the function $\psi_{\max} = c'\beta$, observe that by (2.19) and (2.15)-(2.16) with $\hat{\psi} = C\hat{\beta}$ that

$$(\hat{\psi} - \psi)'(C(X'X)^{-1}C')^{-1}(\hat{\psi} - \psi) > gMS_eF^{1-\alpha} \tag{2.20}$$

where $F^{1-\alpha}$ is the upper $1-\alpha$ percentage value of the F distribution. using the Cauchy-Schwarz (C-S) inequality, $(X'y)^2 \leq (X'x)(y'y)$, with $x = Fa$ and $y = F^{-1}b$ and $G = F'F$, we have that $(a'b)^2 \leq (a'Ga)(b'G^{-1}b)$. Letting $b = (\hat{\psi} - \psi)$ and $G = C(X'X)^{-1}C'$, the

$$\sup_a \frac{[a'(\hat{\psi} - \psi)]^2}{a'C(X'X)^{-1}C'a} \leq (\hat{\psi} - \psi)'(C(X'X)^{-1}C')^{-1}(\hat{\psi} - \psi) \tag{2.21}$$

or for (2.20), the

$$\sup_a \frac{[a'(\hat{\psi} - \psi)]^2}{a'C(X'X)^{-1}C'a} \geq (gMS_eF^{1-\alpha})^{1/2}. \tag{2.22}$$

By again applying the C-S inequality

$$(a'a)^2 \leq (a'a)(a'a) \text{ or } (a'a) \leq [(a'a)(a'a)]^{1/2},$$

we have that

$$[a'(\hat{\psi} - \psi)]^2 \leq a'(\hat{\psi} - \psi)(\hat{\psi} - \psi)'a. \tag{2.23}$$

Hence, for (2.22) the

$$\sup_a \frac{a'G_1a}{a'G_2a} \geq (gMS_eF^{1-\alpha}) \tag{2.24}$$

where $G_1 = (\hat{\psi} - \psi)(\hat{\psi} - \psi)'$ and $G_2 = C(X'X)^{-1}C'$. Recall, however, that the supremum of the ratio of two quadratic forms is the largest characteristic root of the determinantal equation $|G_1 - \lambda G_2| = 0$ with associated eigenvector a_*. Solving $|G_2^{-1}G_1 - \lambda I| = 0$, we find that there exists a matrix P say such that $G_2^{-1}G_1P =$

ΛP where Λ are the roots and P is the matrix of associated eigenvectors of the determinantal equation. For (2.24), we have that

$$\begin{aligned} G_2^{-1}G_1|G_2^{-1}(\hat{\psi}-\psi)| &= G_2^{-1}(\hat{\psi}-\psi)(\hat{\psi}-\psi)'G_2^{1-}(\hat{\psi}-\psi) \\ &= [(\hat{\psi}-\psi)'G_2^{-1}(\hat{\psi}-\psi)]G_2^{-1}(\hat{\psi}-\psi) \\ &= \lambda G_2^{-1}(\hat{\psi}-\psi) \end{aligned} \tag{2.25}$$

so that $a_* = G_2^{-1}(\hat{\psi}-\psi)$ is the eigenvector of $G_2^{-1}G_1$ for the maximum root of $|G_2^{-1}G_1 - \lambda I| = 0$. Thus, the eigenvector a_* may be used to find the linear parametric function of β that is most significantly different from ξ. The function is

$$\psi_{\max} = (a_*'C)\beta = c'\beta. \tag{2.26}$$

The Scheffé-type simultaneous confidence interval for $\psi = c'\beta$ for all nonnull vectors c' such that the $\sum_i c_i = 0$ and the c_i are elements of the vector c is given by

$$\hat{\psi} - c_0\hat{\sigma}_{\hat{\psi}} \leq \psi \leq \hat{\psi} + c_0\hat{\psi}_{\hat{\psi}} \tag{2.27}$$

where $\psi = c'\beta$, $\hat{\psi} = c'\hat{\beta}$, $\hat{\sigma}_{\hat{\psi}}$ is an estimate of the standard error of $\hat{\psi}$ given by (2.6) and

$$c_0^2 = gF_{g,\nu_e}^{1-\alpha} \tag{2.28}$$

where $g = \nu_h$ (see Scheffé, 1959, p. 69).

With the rejection of the test of size α for the null overall hypothesis $H : C\beta = 0$, one may invoke Scheffé's S_2-method to investigate the infinite, nonhierarchical family of contrasts orthogonal to the significant contrast c found using the S-method. Any contrast $\hat{\psi}$ is significantly different from zero if

$$|\hat{\psi}| > [(\nu_h - 1)F_{\nu_h-1,\nu_e}^{1-\alpha}]^{1/2} = S_2 \tag{2.29}$$

where ν_h is the degrees of freedom of the null hypothesis, and $F_{\nu_h-1,\nu_e}^{1-\alpha}$ is the upper $(1-\alpha)$ 100% critical value of the F distribution with degrees of freedom $\nu_h - 1$ and ν_e. Scheffé (1970) showed that the experimentwise Type I error rate for the procedure is controlled at the nominal level α, the level of the overall test. However, the Per-Family Error Rate (PFE), the expected number of Type I errors within the family will increase, Klockars, Hancock, and Krishnaiah (2000). Because these tests are guaranteed to control the overall experimentwise error rate at the level α, they are superior to Fisher's protected t-tests which only weakly control the experimentwise Type I error rate, due to the protection of the significant overall F test, Rencher and Scott (1990).

To construct an UI test of $H : C\beta = \xi$ one writes the null hypothesis as the intersection hypothesis

$$H = \bigcap_a H_a \tag{2.30}$$

where a is a nonnull g-dimensional vector. Hence, H is the intersection of a set of elementary tests H_a. We would reject H if we can reject H_a for any a. By the UI principle, it follows that if we could reject for any a, we could reject for the $a = a_*$ that maximizes (2.22); thus, the UI test for this situation is equivalent to the F test or a likelihood ratio test. For additional details, see for example Casella and Berger (1994, Section 11.2.2).

2.5 MULTIPLE LINEAR REGRESSION

Multiple linear regression procedures are widely applied in the social and physical sciences, in business and industry, and in the health sciences to explain variation in a dependent (criterion) variable by employing a set of independent (predictor) variables using observational (nonexperimental) data. In these studies, the researcher's objective is to establish an optimal model by selecting a subset of available predictors that accounts for the variation in the dependent variable. In such studies, the primary goal is to discover the relationship between the dependent variable and the "best" subset of predictor variables. Multiple linear regression analysis is also used with experimental data. In these situations, the regression coefficients are employed to evaluate the marginal or partial effect of a predictor on the dependent variable given the other predictor variables in the model. In both of these cases, one is usually concerned with estimating model parameters, model specification and variable selection. The primary objective is to develop an "optimal" model using a sampling plan with fixed or random predictors, based upon an established theory. Generally speaking, one is concerned with model calibration using sample data that employs either fixed or random predictors. A second distinct phase of the study may involve model validation. For this phase of the study, one needs to define a measure of predictive precision.

Regression models are also developed to predict some random continuous outcome variable. In these situations, predictor variables are selected to maximize the predictive power of the linear regression model. Studies in this class are not concerned with model calibration, but predictive precision. As a result, regression coefficients are not interpreted as indices of the effects of a predictor variable on the criterion. And, variable selection methods that maximize prediction accuracy and/or minimize the mean squared error of prediction are of primary interest. While the predictors may be fixed, they are usually considered to be random when investigation prediction.

Even though both paradigms are widely used in practice and appear to be interrelated, they are distinct and depend on a set of model assumptions. And, the corresponding model assumptions affect measures of model fit, prediction, model validation, and variable selection which are not always clearly understood when put into practice. Model calibration studies are primarily concerned with understanding the relationship among the predictors to account for variation in the criterion variable and prediction studies are primarily concerned with selecting variables that maximize predictive precision.

Regression models may be applied using a random dependent variable and sev-

eral fixed predictors making specific assumptions regarding the random errors: (i) the classical multiple linear regression (CR) model. Models are also developed with fixed predictors, assuming that the random errors have normal structure: (ii) the classical normal multiple linear regression (CNR) model. The CR model and CNR model are called the general linear model and the normal general linear model, respectively. For these models the set of predictors is not random; it remains fixed under repeated stratified sampling of the dependent variable. For the CR or CNR models, the model calibration phase of the study and the model validation phase of a study are usually separated into two distinct phases called model calibration and model validation. The validation phase of the study may require a second data set to initiate the cross-validation process. The second data set may be obtained from a single sample by splitting the sample or by obtaining an independent set of observations.

An alternative framework for model development occurs when the dependent variable and the set of predictors are obtained from a multivariate population as a random sample of independent and identically distributed observations. If one develops linear models of the joint variation of all variables in the study, we have what is called the (iii) random, classical (distribution-free) multiple linear regression (RCR) model. For the RCR model, the joint distribution of the dependent and independent random variables is unknown. While one may employ robust regression procedures to develop a RCR model, in most applications of multiple linear regression, one assumes a structural linear model where the dependent and independent variables follow a multivariate normal distribution: this is the (iv) jointly normal multiple linear regression (JNR) model. Another model closely related to the JNR model is the (iv) random, classical normal multiple linear regression (RCN) model. For the RCN model, the conditional distribution of the dependent variable is assumed to be normal and the marginal distribution of the independent variables is unknown. For both the JNR and RCN models, model calibration and model validation need not be separated. They may be addressed in a single study without cross-validation.

In the following sections, model assumptions, sampling plans, model calibration and model prediction, goodness of model fit criteria, model selection criteria, predictive precision, mean squared error of prediction in multiple linear regression models, and model validation when the independent variables are consider fixed or random are reviewed. While the concepts of variation free-ness and weak exogeneity may also be of interest, the topic of exogeneity in linear regression models is discussed in detail by Ericsson (1994) and will not be reviewed here.

2.5.1 Classical and Normal Regression Models

Estimation

The CR model is most commonly used in experimental and observational studies to "discover" the relationship between a random dependent variable y and p fixed independent variables x_i. The goal of the study is model specification or model calibration. Assuming the relationship between y and the independent variable is linear

in the elements of the parameter vector β the linear regression model is represented as $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + e$ where the unknown parameter vector $\beta' = (\beta_0, \beta_\ell')$, and the vector of regression coefficients $\beta_\ell' = (\beta_1, \ldots, \beta_p)$ is associated with the p independent variables and β_0 is the model's intercept. The random unknown error e is assumed to have mean zero, $\mathcal{E}(e) = 0$, and common unknown error variance σ^2, $V(e) = \sigma^2$. Organizing the n observations into a vector, the observation vector is y and the n fixed row vectors x_i' in the matrix $X = (1, X_\ell)$ is the $n \times k$ (design) matrix of fixed variables with full rank $k = p+1$ for a model with p parameters. The vector 1 is a column vector of n 1's and the matrix X_ℓ contains the independent variables. The linear model $y = X\beta + e$ has mean $\mathcal{E}(y) = X\beta$ and covariance matrix $\Omega = \sigma^2 I_n$. The primary goal of an analysis is to estimate the unknown parameter vector β and σ^2 using the collected data, the calibration sample.

To estimate the unknown parameter vector β, the n row vectors x_i' of the matrix X are fixed for some set of (optimally) defined values that define the strata for the i^{th} subpopulation. For each subpopulation or strata, a single observation y_i is selected and the n observations are organized to form the elements of the observation vector y. Using the stratified sampling scheme, the elements in y are independent, but not identically distributed since they are obtained from distinct subpopulations. Given this sampling process (for an elementary discussion of this sampling scheme, one may consult Graybill, 1976, p. 154-158), it is not meaningful to estimate the population mean or variance of the row vectors in the matrix X since each row is not sampled from the joint distribution of independent variables. Furthermore, σ^2 is the variance of the random variable y given x_i' and the means $\mathcal{E}(y_i|x_i') = x_i'\beta = \mu_i$ are the conditional means of y_i given x_i' with population mean vector $\mu = X\beta$. For the CR model, we are not directly concerned with estimating the marginal or unconditional mean of y, $\mathcal{E}(y_i) = \mu_y$, or the marginal or unconditional variance of y, $V(y_i) = \sigma_y^2$ for all i. In the CR model, the variance of the unknown random error e is the same as variance of y, namely σ^2. Thus, as correctly noted by Goldberger (1991, p. 179), the sample estimator $\sum_i (y_i - \bar{y})^2/(n-1) = SST/(n-1) = \tilde{\sigma}_y^2$ is not an unbiased estimator of the population variance of y. Using properties of quadratic forms, the expected value of the sample estimator is $\mathcal{E}|\tilde{\sigma}_y^2| = \sigma^2 + \beta_\ell'(X_\ell' P_1 X_\ell)\beta_\ell/(n-1)$ where $P_1 = (I - 1(1'1)^{-1}1')$ is the projection (symmetric and idempotent) matrix for the CR model. Furthermore, because the matrix X is fixed in the CR model the sample covariance matrix associated with the independent variables may not be used as an estimate of the unknown population covariance matrix Σ_{xx} since X is not a random matrix. The row vectors x_i' are not selected from the joint multivariate distribution of the independent variables. Given the CR model, the OLSE for the parameter vector is given in (2.4). Adding the assumption of normality to the CR model yields the CNR model. Then the MLE of the parameters is given in (2.8). The mean squared error "risk" of the estimate $\hat{\beta}$, assuming either a CR or CNR model, is

$$\mathcal{E}[(\beta - \hat{\beta})'(\beta - \hat{\beta})] = \sigma^2 \mathrm{tr}[(X'X)^{-1}]. \tag{2.31}$$

While the OLS estimator is the best linear unbiased estimator, shrinkage estimators due to Stein may have uniformly smaller risk (Dempster, Laird, & Rubin, 1977; Srivastava & Bilodeau, 1989).

Having found the estimate of β, the vector $\hat{y} = X\hat{\beta}$ characterizes the empirical relationship or fit between the random variable y and the vector of fixed independent variables $x' = (x_0, x_1, \ldots, x_p)$ where $x_0 = 1$, based upon the sampling plan. The vector $e = y - \hat{y}$ is the vector of estimated errors or the vector of residuals, where the residuals $e_i = y_i - \hat{y}_i$ have mean zero. Letting $SSE = \sum_i^n (y_i - \hat{y}_i)^2 = \|e\|^2$ represent the error sum of squares, the minimum variance unbiased estimator and maximum likelihood estimators of σ^2 are, respectively,

$$s^2 = \frac{SSE}{(n-k)} = MSE \tag{2.32}$$

and

$$\hat{\sigma}^2 = \frac{SSE}{n}. \tag{2.33}$$

These variance estimates are estimates of the conditional variance of the random variable y given the fixed vector of observations.

Model Fit

Having established the relationship between y and x it is customary to report a measure of the proportion of the variation about the sample mean y, $\bar{y}$, that can be accounted for by the regression function. The measure of sample fit is Fisher's correlation-like ratio, $\tilde{\eta}^2$, defined as

$$\begin{aligned}\tilde{\eta}^2 &= \frac{\|\hat{y} - 1\bar{y}\|^2}{\|y - 1\bar{y}\|^2} = \frac{\sum_i^n (\hat{y}_i - \bar{y})^2}{\sum_i^n (y_i - \bar{y})^2} = \frac{SSB}{SST} \\ &= 1 - \frac{SSE}{SST} = 1 - \frac{\|e\|^2}{(y'y - n\bar{y}^2)} \\ &= 1 - \frac{\sum_i^n e_i^2}{\sum_i^n (y_i - \bar{y})^2}\end{aligned} \tag{2.34}$$

where the vector 1 represents a vector of n 1's and from an analysis of the variation about the mean, the total sum of squares (SST) is equal to the sum of the squares deviations between the fitted values and the mean (SSB) plus the sum of the squares error (SSE). The correlation-like ratio lies between zero and one (provided an intercept is included in the model). If the relationship between y and x is linear, then Fisher's correlation-like ratio becomes the coefficient of multiple determination, $\tilde{R}^2$. And, expression (2.34) becomes

$$\tilde{\eta}^2 = \tilde{R}^2 = \frac{(\hat{\beta}' X y - n\bar{y}^2)}{(y'y - n\bar{y}^2)} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{2.35}$$

where the deviations between the fitted values and the mean are replaced by deviations due to the linear relationship, represented by (SSR), and is called the sum

of square regression. The coefficient of determination $\tilde{R}^2 = 1$ if $e = 0$, and the vector $y = X\beta$ so that y is an exact linear function of the variables in x. The sample fit is exact in that all the variation in the elements of the vector y is accounted for (linearly) or explained by the variation among the elements in the vector x. When the coefficient of determination $\tilde{R}^2 = 0$, each element of the vector $\hat{y}$ is identical to the sample mean of the observations, $\hat{y}_i = \bar{y}$ for all i observations. Then, the best regression equation is the sample mean vector $\bar{y} = 1\bar{y}$ so that none of the variation in y is accounted for (linearly) by the variation in the independent variables. Thus the quantity $\tilde{R}^2$ is often used as a measure of goodness-of-fit for the estimated regression model. However, since $\tilde{R}^2$ tends to increase as additional independent variables are included in the CR or CNR model, many authors suggest the adjusted coefficient of determination as a measure of goodness of model fit which takes into account the size of k relative to n. While one may account for the size of k relative to n in any number of ways, the most popular correction is the adjusted coefficient of determination defined as

$$\begin{aligned}\tilde{R}_a^2 &= 1 - \left(\frac{n-1}{n-k}\right)\left(\frac{SSE}{SST}\right) = 1 - \left(\frac{SSE/(n-k)}{SST/(n-1)}\right) \\ &= 1 - \frac{MSE}{\tilde{\sigma}_y^2} \\ &= 1 - \frac{(n-1)(1-\tilde{R}^2)}{n-k}. \end{aligned} \tag{2.36}$$

While the numerator MSE is an unbiased estimate of σ^2, the denominator of the ratio is not an unbiased estimate of the variance of y for either the CR or CNR models. Although $\tilde{R}_a^2$ may be expressed using a "sample variance-like" formula

$$\tilde{R}_a^2 = 1 - \frac{s_{y\cdot x}^2}{\tilde{\sigma}_y^2}. \tag{2.37}$$

Neither $\tilde{R}^2$ or $\tilde{R}_a^2$ is an estimator of the population coefficient of determination R^2 since we have selected the rows of the matrix X selectively and not at random. Furthermore, with a fixed matrix X, we can always find a design matrix X that makes $\tilde{R}^2 = \tilde{R}_a^2 = 1$. Since the matrix X is not randomly created, but fixed, one can always obtain a set of n-linearly independent $n \times 1$ column vectors to create a basis for the matrix X. Then, y may be represented by $y = X\beta$ exactly making the coefficient of determination or adjusted coefficient of determination unity. Given these limitations, Goldberger (1991, p. 177) concluded that the most important thing about $\tilde{R}^2$ and $\tilde{R}_a^2$ for the CR and CNR models is that they are not very useful.

In summary, for the CR or CNR models: (i) the predictor variables are nonrandom and fixed, (ii) the sampling plan for the model employs a stratified sampling process where the criterion variable is obtained for fixed values of the predictor so that the dependent variables are independent, but are not identically distributed, (iii) because the predictors are fixed, the values of the predictors may be chosen to create an optimal design to minimize the mean square error of the estimate $\hat{\beta}$, and (iv) neither $\tilde{R}^2$ or $\tilde{R}_a^2$ are necessarily very useful in the evaluation of model fit.

Model Selection

Because X is fixed and not random in the CR and CRN models, the matrix X may represent an over fitted model. Letting X^t represent the true model and X^u an under fitted model, the relationship among the variables are often assumed to be nested in that $X^u \subseteq X^t \subseteq X$ with corresponding parameter vectors β^u, β^t, and β, respectively. Next, suppose a sample of n observations from an over fitted model is used to estimate the unknown parameter vector β so that the estimator has the expression given in (2.4). Then, if we obtain n new observations y where the observations have the linear identical form $y = X\beta + e$, the predicted value of y is $\hat{y} = X\hat{\beta}$. Since the matrix X is fixed, the average mean squared error of prediction for the n new observations is

$$\begin{aligned}\delta_f^2 &= \frac{\mathcal{E}[(y-\hat{y})'(y-\hat{y})]}{n} = \frac{\mathcal{E}[(X\beta - X\hat{\beta})'(X\beta - X\hat{\beta})]}{n} = \frac{\mathcal{E}[e'e]}{n} \\ &= \frac{\mathrm{tr}[(X'X)\sigma^2(X'X)^{-1}]}{n} + \frac{n\sigma^2}{n} \\ &= \sigma^2\left(1 + \frac{k}{n}\right). \end{aligned} \tag{2.38}$$

The quantity δ_f^2 is called the final prediction error (FPE).

An unbiased estimator of δ_f^2 is obtained by substituting for σ^2 in (2.38), the unbiased estimator given in (2.32). An unbiased estimate of the FPE and its associated variance follow

$$\hat{\delta}_u^2 = s^2\left(1 + \frac{k}{n}\right), \tag{2.39}$$

$$\mathrm{var}(\hat{\delta}_u^2) = \left(\frac{2\sigma^4}{n-k}\right)\left(1 + \frac{2k}{n} + \frac{k^2}{n}\right), \tag{2.40}$$

Picard and Berk (1990). For the CNR model, the errors follow a multivariate normal distribution so the MLE of the FPE is

$$\begin{aligned}\hat{\delta}_{MLE}^2 &= \frac{SSE}{(n-k)}\left(\frac{n+k}{n}\right) = \frac{SSE}{n}\left(\frac{n+k}{n-k}\right) \\ &= \hat{\sigma}^2\left(\frac{n+k}{n-k}\right) \\ &= \hat{\sigma}^2\left(1 + \frac{2k}{n-k}\right) \end{aligned} \tag{2.41}$$

where $\hat{\sigma}^2$ is given in (2.33). As the number of parameters vary, the final prediction error balances the variance between the best linear predictor of y and the variance of $X\hat{\beta}$. Models with small final prediction error are examined to select the "best" candidate model. The effect of data splitting on the estimate of FPE is discussed by Picard and Berk (1990).

Mallows (1973) took an alternative approach in developing a model selection criterion, again for fixed X. There are $2^p - 1$ possible submodels. Let subscript j represent different models where $j = 1, \ldots, (2^p - 1)$ and k_j represents number of parameters in the j^{th} model. He considered the relative error of estimating y for a submodel $\hat{y}_j = X_j\hat{\beta}_j$ defined by

$$\begin{aligned} J_j &= \frac{\mathcal{E}[(y - \hat{y}_j)'(y - \hat{y}_j)]}{\sigma^2} \\ &= \frac{\sum_i^n \text{var}(\hat{y}_i)^2 + \sum_i^n (\text{bias in } \hat{y}_i)^2}{\sigma^2} \\ &= k_j + \frac{(\hat{\beta}_j - \beta)'X'X(\hat{\beta}_j - \beta)}{2\sigma^2} = k_j + \hat{\lambda}_j \end{aligned} \tag{2.42}$$

where $\lambda_j = \mathcal{E}(\hat{\lambda}_j)$ is the noncentrality parameter in the CR and CNR models. Letting $SSE_j = \|e_j\|^2$ for k_j parameters, Mallows proposed an estimator $\widehat{J}_j = SSE_j/s^2 - n + 2k_j$ of J_j by obtaining unbiased estimators of the numerator and denominator of the relative error ratio. His familiar C_p criterion for model selection is

$$C_p = \left(\frac{SSE_j}{s^2} - n + 2k_j\right). \tag{2.43}$$

Mallows (1995) suggests that any model in which $C_p < k_j$ may be a potential candidate model. Both the FPE and C_p criteria may be used with either the CR and CNR models.

In 1973, Hirotugu Akaike derived an estimator of the (relative) Kullback-Leibler distance based on Fisher's maximized log-likelihood for the CNR model. His measure for model selection is called Akaike's information criterion (AIC).
Akaike (1973) criterion is defined as

$$AIC = -2\log(\text{likelihood}) + 2(\text{number of parameters estimated}). \tag{2.44}$$

For the CNR model,

$$-2\log(\text{likelihood}) = n\log(2\pi) + n\log(\sigma^2) + \frac{(y - X\beta)'(y - X\beta)}{\sigma^2} \tag{2.45}$$

and since the number of parameters to be estimated is k_j for β and 1 for σ_j^2, the AIC criterion is

$$AIC_j = n\log(2\pi) + n\log(\sigma_j^2) + n + 2(k_j + 1). \tag{2.46}$$

Ignoring the constants, the AIC criterion becomes $AIC_j = n\log(\sigma^2) + 2d$, for $d = k_j + 1$. Substituting a MLE and an unbiased estimator for the unknown parameter σ_j^2, the AIC criteria follow

$$AIC_{MLE} = n\log(\hat{\sigma}_j^2) + 2d \tag{2.47}$$

$$AIC_u = n\log(s_j^2) + 2d. \tag{2.48}$$

The model with the smallest AIC value is said to fit best. McQuarrie and Tsai scale the AIC criterion by dividing it by the sample size n.

When using Akaike's AIC fit criterion, one selects a model with too many variables (an overfit model) when there are too many parameters relative to the sample size n, $[n/d < 40]$. For this situation, Sugiura (1978) proposed a corrected AIC (CAIC) criterion defined as

$$CAIC_j = AIC_j + \frac{2d(d+1)}{n-d-1}. \tag{2.49}$$

A word of caution, most statistical packages do not calculate $CAIC_j$, but calculate AIC_{MLE}. Users must make their own adjustments. One may simply, for example, substitute AIC_{MLE} for the AIC_j. Monte Carlo studies performed by McQuarrie and Tsai (1998) use both the statistic AIC_u and the criterion $CAIC$, where the latter includes the biased MLE for the variance estimate and Suguira's penalty correction. An estimate of AIC that includes an unbiased estimate of the variance and Suguira's penalty correction is represented by $CAIC_u$, a "doubly" corrected criterion.

Schwarz (1978) and Akaike (1978) developed model selection criteria using a Bayesian approach which incorporates a large penalty factor for over fitting. The criteria select models based upon the largest posterior probability of being correct. In large samples, their posterior probabilities are approximated using a Taylor series expansion. Scaling the first two terms in the series by n their criterion is labeled BIC for Bayesian information criterion (or also SIC for Schwarz information criterion or SBC for Schwarz-Bayesian criterion). Hannan and Quinn (1979) developed another criterion when analyzing autoregressive time series models. Applying their criterion to CNR models, the criterion is represented by HQ. Formula for the two criteria are

$$BIC_j = n\log(\sigma_j^2) + d\log(n), \tag{2.50}$$

$$HQ_j = n\log(\sigma_j^2) + 2d\log[\log(n)]. \tag{2.51}$$

One may again substitute either a MLE for σ_j^2, $\hat{\sigma}_j^2$, or the minimum variance unbiased estimator, s_j^2. When an unbiased estimate is substituted, the criteria are represented by BIC_u, and HQ_u, respectively. In either case, one investigates potential candidate models for a subset of variables that have the smallest BIC and HQ values. For very large samples, the HQ criterion behaves very much like the AIC criterion. Using the scaling factor $n/(n-d-1)$ for the HQ criteria, the scaled corrected criteria (CHQ) proposed by McQuarrie and Tsai (1998, p. 35) is defined

$$CHQ_j = \log(\hat{\sigma}_j^2) + \frac{2k_j\log[\log(n)]}{n-d-1}. \tag{2.52}$$

Many authors suggest the investigation of models that minimize s_j^2 or equivalently the sum of squares error criterion SSE_j since $s_j^2 = SSE_j/(n-k_j)$. However, since $1-\tilde{R}_a^2 = s_j^2/(SST/n)$, the denominator is constant as k_j varies, so that minimizing s_j^2 is equivalent to selecting a model that maximizes that statistic $\tilde{R}_a^2$.

One may also relate Mallow's criterion to $\tilde{R}^2$. Letting $\tilde{R}_j^2$ denote the coefficient of determination for the submodel,

$$C_p = \left(\frac{1-\tilde{R}_j^2}{1-\tilde{R}^2}\right)(n-k) - n + 2k_j. \tag{2.53}$$

A common practice in the CR or CNR models with fixed X is to assume that the "true" model is within the class of models under study, a nested set of models. Thus, the best model is defined by a parameter space defined by a subset of the collected variables. Model selection criteria that select the true nested model asymptotically with probability one are said to be consistent. The criteria BIC and HQ are consistent. This is not the case for the selection criteria AIC, C_p, and $\tilde{R}_a^2$. When using these criteria, one tends to select a model with too many independent variables, McQuarrie and Tsai (1998, p. 42 & 370); however, the AIC_u and CAIC criteria tend to overfit least. If a researcher is not sure that the "true" model is among the nested variables, the model is nonnested; however, one may still want to locate a model that is an approximation to the true model. In this case, the approximation is usually evaluated by comparing the average minimum mean square error of prediction for any two models.

In large samples, a model selection criterion that chooses the model with minimum mean squared error is said to be asymptotically efficient. The FPE criterion δ_f^2, and the criteria AIC, C_p, and $\tilde{R}_a^2$ are all asymptotically efficient criteria. But, in small samples, they may lead to overfitting. No selection criterion is both consistent and asymptotically efficient, so there is not a single criterion that is best for all situations. However, based on extensive Monte Carlo studies conducted by McQuarrie and Tsai (1998) using random normal errors, they found that the asymptotically efficient criterion CAIC and the consistent criterion CHQ performed best. The criteria were most likely to find the correct or closest candidate model. They are least likely to under fit and minimize over fitting. For weakly identified models no criterion is best; however, criteria with weak penalty functions tend to overfit excessively.

Finally, the C_p and FPE criteria may only be used to select a submodel from within the class of nested models. The information criteria allow one to rank order potential candidate models whether the models are nested or nonnested. For a comprehensive review of model selection criteria for nonnested models, one may consult McAleer (1995).

Model Selection, Likelihood Ratio Tests

For the CNR model, the likelihood ratio test statistic for testing the null hypothesis that a subset of the regression coefficients β_i associated with any $h = p - m$ variables (excluding the intercept-even thought it is included in the regression model) is zero versus the alternative hypothesis that the coefficients are not zero, one may employ the F statistic

$$F = \frac{(n-k)}{n} \cdot \frac{(\tilde{R}^2 - \tilde{R}_m^2)}{(1-\tilde{R}^2)} \sim F(\nu_h, \nu_e) \tag{2.54}$$