Previously known as *Statistical Methods for Health Sciences*, this bestselling resource is one of the first books to discuss the methodologies used for the analysis of clustered and correlated data. While the fundamental objectives of its predecessors remain the same, *Analysis of Correlated Data with SAS and R, Third Edition* incorporates several additions that take into account recent developments in the field.

**New to the Third Edition**
- The introduction of R codes for almost all of the numerous examples solved with SAS
- A chapter devoted to the modeling and analyzing of normally distributed variables under clustered sampling designs
- A chapter on the analysis of correlated count data that focuses on over-dispersion
- Expansion of the analysis of repeated measures and longitudinal data when the response variables are normally distributed
- Sample size requirements relevant to the topic being discussed, such as when the data are correlated because the sampling units are physically clustered or because subjects are observed over time
- Exercises at the end of each chapter to enhance the understanding of the material covered
- An accompanying CD-ROM that contains all the data sets in the book along with the SAS and R codes

Assuming a working knowledge of SAS and R, this text provides the necessary concepts and applications for analyzing clustered and correlated data from epidemiologic and medical investigations.
Analysis of Correlated Data with SAS and R

Third Edition
Analysis of Correlated Data with SAS and R

THIRD EDITION

Mohamed M. Shoukri
Mohammad A. Chaudhary
To the memory of my mother. To my wife, Sue, and my children, Nader, Nadene, and Tamer

— Mohamed M. Shoukri

To my mother, Nazir Begum, and my father, Siddiq Ahmed

— Mohammad A. Chaudhary
Contents

Preface to the First Edition ................................................................. ix
Preface to the Second Edition ............................................................ xi
Preface to the Third Edition ................................................................. xiii
Authors .............................................................................................. xv

1. Analyzing Clustered Data ............................................................... 1
2. Analysis of Cross-Classified Data ................................................... 33
3. Modeling Binary Outcome Data ..................................................... 83
4. Analysis of Clustered Count Data ................................................ 133
5. Analysis of Time Series ................................................................. 159
6. Repeated Measures and Longitudinal Data Analysis ................... 207
7. Survival Data Analysis ................................................................. 243

References ......................................................................................... 283
Index ..................................................................................................... 291
Preface to the First Edition

A substantial portion of epidemiologic studies, particularly in community medicine, veterinary herd health, field trials and repeated measures from clinical investigations, produce data that are clustered and quite heterogeneous. Such clustering will inevitably produce highly correlated observations; thus, standard statistical techniques in non-specialized biostatistics textbooks are no longer appropriate in the analysis of such data. For this reason it was our mandate to introduce to our audience the recent advances in statistical modeling of clustered or correlated data that exhibit extra variation or heterogeneity.

This book reflects our teaching experiences of a biostatistics course in the University of Guelph’s Department of Population Medicine. The course is attended predominantly by epidemiology graduate students, as well as, students from Animal Science and researchers from disciplines which involve the collection of clustered and over-time data. The material in this text assumes that the reader is familiar with basic applied statistics, principles of linear regression and experimental design, but stops short of requiring a cognizance of the details of the likelihood theory and asymptotic inference. We emphasize the “how to” rather than the theoretical aspect; however, on several occasions the theory behind certain topics could not be omitted, but is presented in a simplified form.

The book is structured as follows: Chapter 1 serves as an introduction in which the reader is familiarized with the effect of violating the assumptions of homogeneity and independence on the ANOVA problem. Chapter 2 discusses the problem of assessing measurement reliability. The computation of the intraclass correlation as a measure of reliability allowed us to introduce this measure as an index of clustering in subsequent chapters. The analysis of binary data summarized in $2 \times 2$ tables is taken up in Chapter 3. This chapter deals with several topics including, for instance, measures of association between binary variables, measures of agreement and statistical analysis of medical screening tests. Methods of cluster adjustment proposed by Donald and Donner (1987), Rao and Scott (1992) are explained. Chapter 4 concerns the use of logistic regression models in studying the effects of covariates on the risk of disease. In addition to the methods of Donald and Donner, and Rao and Scott to adjust for clustering, we explain the Generalized Estimating Equations (GEE) approach proposed by Liang and Zeger (1986). A general background on time series models is introduced in Chapter 5. Finally, in Chapter 6 we show how repeated measures data are analyzed under the linear additive model for continuously distributed data and also for other types of data using the GEE.
We wish to thank Dr. A. Meek, the Dean of the Ontario Veterinary College, for his encouragement in writing this book; Dr. S. W. Martin, Chair of the Department of Population medicine, for facilitating the use of the departmental resources; the graduate students who took the course “Statistics for the Health Sciences”; special thanks to Dr. J. Sargeant for being so generous with her data and to Mr. P. Page for his invaluable computing expertise. Finally, we would like to thank J. Tremblay for her patience and enthusiasm in the production of this manuscript.

M.M. Shoukri
V.L. Edge
Guelph, Ontario
July 1995
Preface to the Second Edition

The main structure of the book has been kept similar to the first edition. To keep pace with the recent advances in the science of statistics, more topics have been covered. In Chapter 2 we introduce the coefficient of variation as a measure of reproducibility, and comparing two dependent reliability coefficients. Testing for trend using Cochran-Armitage chi-square, under cluster randomization has been introduced in Chapter 4. In this chapter we discussed the application of the PROC GENMOD in SAS, which implements the GEE approach, and “Multi-level analysis” of clustered binary data under the “Generalized Linear Mixed Effect Models,” using Schall’s algorithm, and GLIMMIX SAS macro. In Chapter 5 we added two new sections on modeling seasonal time series; one uses combination of polynomials to describe the trend component and trigonometric functions to describe seasonality, while the other is devoted to modeling seasonality using the more sophisticated ARIMA models. Chapter 6 has been expanded to include analysis of repeated measures experiment under the “Linear Mixed Effects Models,” using PROC MIXED in SAS. We added Chapter 7 to cover the topic of survival analysis. We included a brief discussion on the analysis of correlated survival data in this chapter.

An important feature of the second edition is that all the examples are solved using the SAS package. We also provided all the SAS programs that are needed to understand the material in each chapter.

M.M. Shoukri, Guelph, Ontario
C.A. Pause, London, Ontario
July 1998
Preface to the Third Edition

It was brought to our attention by many of our colleagues that the title of the previous edition did not reflect the focus of the book, which was the analysis of correlated data. We therefore decided to change the title of the third edition to *Analysis of Correlated Data with SAS and R*. We believe that the change in the title is appropriate and reflects the main focus of the book.

The fundamental objectives of the new edition have been kept similar to those of the previous two editions. However, this edition contains major structural changes. The first chapter in the previous editions has been deleted and is replaced with a new chapter devoted to the issue of modeling and analyzing normally distributed variables under clustered sampling designs. A separate chapter is devoted to the analysis of correlated count data with extensive discussion on the issue of overdispersion. Multilevel analyses of clustered data using the “Generalized Linear Mixed Effects Models” fitted by the PROC GLIMMIX in SAS are emphasized. Chapter 6 has been expanded to include the analysis of repeated measures and longitudinal data when the response variables are normally distributed, binary and count. The “Linear Mixed Effects Models” are fitted using PROC MIXED and PROC GLIMMIX in SAS. An important feature of the third edition is the introduction of R codes for almost all the examples solved with SAS. The freeware R package can be downloaded from The Comprehensive R Archive Network (CRAN) at http://cran.r-project.org/ or at any of its mirrors. The reader of this book is expected to have prior working knowledge of SAS and R packages. Readers who have experience with S-PLUS will have no problem working with R. Readers completely new to R will benefit from the many tutorials available on the R web site.

The important features of the third edition are

1. We provide a large number of examples to cover the material, together with their SAS and the R codes.
2. In each chapter, we provide the reader with sample size requirements relevant to the topic discussed, with special emphasis on situations when the data are correlated either because the sampling units are physically clustered or because subjects are observed over time.
3. At the end of each chapter, a set of exercises is provided to enhance the understanding of the material covered.
4. We provide a CD that contains all the data along with the SAS and R codes.
Dr. Chaudhary’s work was supported in part by the Bill and Melinda Gates Foundation as part of the Consortium to Respond Effectively to the AIDS-TB Epidemic (CREATE) project (19790.01).

M.M. Shoukri, London, Ontario, Canada
M.A. Chaudhary, Baltimore, Maryland
Authors

Mohamed M. Shoukri received his MSc and PhD degrees from the Department of Mathematics and Statistics, University of Calgary, Alberta, Canada. He held several faculty positions at various Canadian universities, and taught applied statistics at Simon Fraser University, the University of British Columbia, and the University of Windsor, and was a full professor with tenure at the University of Guelph, Ontario, Canada. His papers have been published in the *Journal of the Royal Statistical Society* (series C), *Biometrics, Journal of Statistical Planning and Inference, The Canadian Journal of Statistics, Statistics in Medicine, Statistical Methods in Medical Research*, and many other journals. He is a fellow of the Royal Statistical Society of London and an elected member of the International Statistical Institute. He is now principal scientist and the acting chairman of the Department of Biostatistics and Epidemiology at the Research Center of King Faisal Specialist Hospital.

Mohammad A. Chaudhary is an associate scientist (biostatistics) at the Department of International Health, Johns Hopkins Bloomberg School of Public Health. He received his PhD degree in biostatistics from the University of North Carolina at Chapel Hill and his master degrees from Islamia University and the University of Southampton. He has taught at Islamia University, Punjab University, and at the University of Memphis while he was a postdoctoral fellow at St. Jude Children Research Hospital. Before joining his current position, he served as a scientist (biostatistics) at King Faisal Specialist Hospital and Research Center, Riyadh. His papers have been published in *Statistics in Medicine, Journal of Statistical Planning and Inference, Computer Methods and Programs in Medicine, Biometrical Journal, Journal of Statistical Research, Cancer Research, Lancet*, and other biomedical research journals. His current research interests include statistical computing and the design and analysis of randomized clinical trials evaluating novel interventions for the prevention of TB and HIV-AIDS.
1

Analyzing Clustered Data

CONTENTS
1.1 Introduction ................................................................................................. 1
   1.1.1 The Basic Feature of Cluster Data ................................................ 2
   1.1.2 Sample and Design Issues ............................................................. 7
1.2 Regression Analysis for Clustered Data .................................................. 11
1.3 Generalized Linear Models ....................................................................... 15
   1.3.1 Marginal Models (Population Average Models) ........................ 16
   1.3.2 Random Effects Models ................................................................. 17
   1.3.3 Generalized Estimating Equation (GEE) ..................................... 17
1.4 Fitting Alternative Models for Clustered Data ....................................... 19

1.1 Introduction

Clusters are aggregates of individuals or items that are the subject of investigation. A cluster may be a family, school, herd of animals, flock of birds, hospital, medical practice, or an entire community. Data obtained from clusters may be the result of an intervention in a randomized clinical or a field trial. Sometimes interventions in randomized clinical trials are allocated to groups of patients rather than to individual patients. This is called cluster randomization or cluster allocation, and is particularly common in human and animal health research. There are several reasons why investigators wish to randomize clusters rather than the individual study subjects. The first being the intervention may naturally be applicable to clusters. For example, Murray et al. (1992) evaluated the effect of school-based interventions in reducing adolescent tobacco use. A total of 24 schools (of various sizes) were randomized to an intervention condition (SFG = smoke-free generation) or to a control condition (EC = existing curriculum). The number (and proportion) of children in each school who continued to use smokeless tobacco after 2 years of follow-up is given in Table 1.1.

It would be impossible to assign students to intervention and control groups because the intervention is through the educational program that is received by all students in a school.
TABLE 1.1
Smokeless Tobacco Use among Schoolchildren

<table>
<thead>
<tr>
<th>Control (EC)</th>
<th>Intervention (SFG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5/103</td>
<td>0/42</td>
</tr>
<tr>
<td>3/174</td>
<td>1/84</td>
</tr>
<tr>
<td>6/83</td>
<td>9/149</td>
</tr>
<tr>
<td>6/75</td>
<td>11/136</td>
</tr>
<tr>
<td>2/152</td>
<td>4/58</td>
</tr>
<tr>
<td>7/102</td>
<td>1/55</td>
</tr>
<tr>
<td>7/104</td>
<td>10/219</td>
</tr>
<tr>
<td>3/74</td>
<td>4/160</td>
</tr>
<tr>
<td>1/55</td>
<td>2/63</td>
</tr>
<tr>
<td>23/225</td>
<td>5/85</td>
</tr>
<tr>
<td>16/125</td>
<td>1/96</td>
</tr>
<tr>
<td>12/207</td>
<td>10/194</td>
</tr>
</tbody>
</table>

Second, even if individual allocation is possible, there is the possibility of contamination. For example, in a randomized controlled intervention trial the purpose was to reduce the rate of cesarean section. The intervention was that each physician should consult with his/her fellow physician before making the decision to operate, and the control was to allow the physician to make the decision without consulting his/her colleague. Ten hospitals were randomized to receive the intervention, while ten other hospitals were kept as controls. In this example, cluster randomization is desired even if the randomization by the physician is possible, because of the possibility of significant crossover contamination. Because the physicians work together, it is likely that a physician in the control group, who did not receive the intervention, might still be affected by it via interactions with colleagues in the intervention group.

Third, cluster allocation is sometimes cheaper or more practical than individual allocation. Many public health interventions are relatively less costly when implemented at an organizational level (e.g., community) than at an individual level.

Similar to cluster randomization, cluster allocation is common in observational studies. In this case, it is sometimes more efficient to gather data from organizational units such as farms, census tracts, or villages rather than from individuals (see Murray et al., 1992).

1.1.1 The Basic Feature of Cluster Data

When subjects are sampled, randomized, or allocated by clusters, several statistical problems arise. Observations within a cluster tend to be more alike than observations selected at random. If observations within a cluster are correlated, one of the assumptions of estimation and hypothesis testing is violated. Because of this correlation, the analyses must be modified to take into account the cluster design effect. When cluster designs are used, there
are two sources of variations in the observations. The first is the one between subjects within a cluster, and the second is the variability among clusters. These two sources of variation cause the variance to inflate and must be taken into account in the analysis.

The effect of the increased variability due to clustering is to increase the standard error of the effect measure, and thus widen the confidence interval and inflate the type I error rate, compared to a study of the same size using individual randomization. In other words, the effective sample size and consequently the power are reduced. Conversely, failing to account for clustering in the analysis will result in confidence intervals that are falsely narrow and the \( p \)-values falsely small. Randomization by cluster accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception (Cornfield, 1978).

Failing to account for clustering in the analysis is similar to another error that relates to the definition of the unit of analysis. The unit of analysis error occurs when several observations are taken on the same subject. For example, a patient may have multiple observations (e.g., systolic blood pressure) repeated over time. In this case, the repeated data points cannot be regarded as independent, since measurements are taken on the same subject. For example, if five readings of systolic blood pressure are taken from each of the 15 patients, assuming 75 observations to be independent is wrong. Here, the patient is the appropriate unit of analysis and is considered as a cluster.

To recognize the problem associated with the unit of analysis, let us assume that we have \( k \) clusters each of size \( n \) units (the assumption of equal cluster size will be relaxed later on). The data layout (Table 1.2) may take the form:

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Units</th>
<th>( 1 )</th>
<th>( 2 )</th>
<th>( \ldots )</th>
<th>( i )</th>
<th>( \ldots )</th>
<th>( k )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>( Y_{11} )</td>
<td>( y_{21} )</td>
<td>( \ldots )</td>
<td>( y_{i1} )</td>
<td>( \ldots )</td>
<td>( y_{1k} )</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>( y_{12} )</td>
<td>( Y_{22} )</td>
<td>( \ldots )</td>
<td>( y_{22} )</td>
<td>( \ldots )</td>
<td>( y_{2k} )</td>
</tr>
<tr>
<td></td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \ldots )</td>
<td>( \vdots )</td>
<td>( \ldots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td></td>
<td>( j )</td>
<td>( Y_{1j} )</td>
<td>( y_{2j} )</td>
<td>( \ldots )</td>
<td>( y_{ij} )</td>
<td>( \ldots )</td>
<td>( y_{jk} )</td>
</tr>
<tr>
<td></td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \ldots )</td>
<td>( \vdots )</td>
<td>( \ldots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td></td>
<td>( n )</td>
<td>( y_{1n} )</td>
<td>( y_{2n} )</td>
<td>( \ldots )</td>
<td>( y_{in} )</td>
<td>( \ldots )</td>
<td>( y_{kn} )</td>
</tr>
<tr>
<td>Total</td>
<td>( y_1 )</td>
<td>( y_2 )</td>
<td>( \ldots )</td>
<td>( y_i )</td>
<td>( \ldots )</td>
<td>( y_k )</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>( \bar{y}_1 )</td>
<td>( \bar{y}_2 )</td>
<td>( \ldots )</td>
<td>( \bar{y}_i )</td>
<td>( \ldots )</td>
<td>( \bar{y}_k )</td>
<td></td>
</tr>
</tbody>
</table>

The grand sample mean is denoted by \( \bar{y} = \frac{1}{nk} \sum_{i=1}^{k} \sum_{j=1}^{n} \bar{y}_{ij} \).

If the observations within a cluster are independent, then the variance of the \( i \)th cluster mean is

\[
V(\bar{y}_i) = \frac{\sigma_y^2}{n}
\]  

(1.1)
where \( \sigma_y^2 = E(y_{ij} - \mu)^2 \) and \( \mu \) is the population mean. Assuming that the variance is constant across clusters, the variance of the grand mean is

\[
V(\bar{y}) = \frac{\sigma_y^2}{nk} \tag{1.2}
\]

Now, if \( k \) clusters are sampled from a population of clusters, and because members of the same cluster are similar, the variance within the cluster would be smaller than would be expected if members were assigned at random. To articulate this concept, we first assume that the \( j \)th measurement within the \( i \)th cluster \( y_{ij} \) is such that

\[
y_{ij} = \mu + b_i + e_{ij},
\]

so that \( E(b_i) = 0, V(b_i) = \sigma_b^2, E(e_{ij}) = 0, V(e_{ij}) = \sigma_e^2, \) and \( b_i \) and \( e_{ij} \) are independent of each other. Under this setup, we can show that

\[
V(y_{ij}) = \sigma_b^2 + \sigma_e^2 \equiv \sigma_y^2 \tag{1.3}
\]

\[
\text{Cov}(y_{ij}, y_{il}) = \sigma_b^2 \text{ if } j \neq l \tag{1.4}
\]

Therefore, the correlation between any pair of observations within a cluster is

\[
\text{Corr}(y_{ij}, y_{il}) = \rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2} \tag{1.5}
\]

This correlation is known as the intracluster correlation (ICC). Therefore,

\[
\sigma_e^2 = \sigma_y^2 (1 - \rho) \tag{1.6}
\]

Equation 1.4 shows that if the observations within a cluster are not correlated (\( \rho = 0 \)), then the within-cluster variance is identical to the variance of randomly selected individuals. Since \( 0 \leq \rho < 1 \), the within-cluster variance \( \sigma_e^2 \) is always less than \( \sigma_y^2 \).

Simple algebra shows that

\[
V(\bar{Y}_i) = \frac{\sigma_y^2}{n}[1 + (n - 1)\rho]\tag{1.7}
\]

and

\[
V(\bar{Y}) = \frac{\sigma_y^2}{nk}[1 + (n - 1)\rho]\tag{1.8}
\]
Note that for binary response, $\sigma^2$ is replaced with $\pi(1 - \pi)$. The quantity $[1 + (n - 1)\rho]$ is called the variance inflation factor (VIF) or the “design effect” (DEFF) by Kerry and Bland (1998). It is also interpreted as the relative efficiency of the cluster design relative to the random sampling of subjects and is the ratio of Equation 1.8 to Equation 1.2:

$$\text{DEFF} = [1 + (n - 1)\rho] \tag{1.9}$$

The DEFF represents the factor by which the total sample size must be increased if a cluster design is to have the same statistical power as a design in which individuals are sampled or randomized. If the cluster sizes are not equal, which is commonly the case, the cluster size $n$ should be replaced with $\bar{n} = \frac{\sum n_i^2}{N}$, where $N = \sum n_i$.

Since $\rho$ is unknown, we estimate its value from the one-way ANOVA layout (Table 1.3):

**TABLE 1.3**

<table>
<thead>
<tr>
<th>SOV</th>
<th>DF</th>
<th>Sum of square</th>
<th>Mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between clusters</td>
<td>$k - 1$</td>
<td>BSS</td>
<td>$\frac{\text{BSS}}{k - 1}$</td>
</tr>
<tr>
<td>Within clusters</td>
<td>$N - k$</td>
<td>WSS</td>
<td>$\frac{\text{WSS}}{N - k}$</td>
</tr>
<tr>
<td>Total</td>
<td>$N - 1$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

where $\text{BSS} = n \sum_i (\bar{y}_i - \bar{\mu})^2$, $\text{WSS} = \sum_{ij}^k (y_{ij} - \bar{y}_i)^2$, and the ICC is estimated by

$$\hat{\rho} = \frac{\text{BMS} - \text{WMS}}{\text{BMS} + (n_0 - 1)\text{WMS}} = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \hat{\sigma}_e^2}$$

where $\hat{\sigma}_b^2 = (\text{BMS} - \text{WMS})/n_0$ and $\hat{\sigma}_e^2 = \text{WMS}$ are the sample estimates of $\sigma_b^2$ and $\sigma_e^2$, respectively.

$$n_0 = \frac{N}{k} - \frac{\sum_{i=1}^k (n_i - \overline{n})^2}{k(k - 1)\overline{n}} \quad \overline{n} = \frac{N}{k} \tag{1.10}$$

Note that when $n_1 = n_2 = \cdots = n_k$, then $n_0 = \overline{n} = n$.

If $\hat{\rho} > 0$, then the variance in a cluster design will always be greater than a design in which subjects are randomly assigned so that, conditional on the cluster size, the confidence interval will be wider and the $p$-values larger. We note further that, if the cluster size ($n$) is large, the DEFF would be large even for small values of $\hat{\rho}$. For example, an average cluster size of 40 and ICC = 0.02 would give DEFF of 1.8, implying that the sample size of a cluster-randomized design should be 180% as large as the estimated sample size of an individually randomized design to achieve the same statistical power.
We have demonstrated that applying conventional statistical methods to cluster data assuming independence between the study subjects is wrong, and one has to employ appropriate approach that duly accounts for the correlated nature of cluster data. The complexity of the approach depends on the complexity of design. For example, the simplest form of cluster data is the one-way layout, where subjects are nested within clusters. Conventionally, this design has two levels of hierarchy: subjects at the first level and clusters at the second level. For example, the sib-ship data (will be shown below) have two levels: the first is observations from sibs and the second is formed by the family identifiers. Data with multiple levels of hierarchy, such as animals within the farms, farms nested within regions, may also be available, and one must account for the variability at each level of hierarchy.

We shall now review some of the studies reported in the medical literature that must be included under “cluster randomization trials” where we clearly identify what is meant by “cluster.”

Russell et al. (1983) investigated the effect of nicotine chewing gum as a supplement to the general practitioners’ advice against smoking. Subjects were assigned by week of attendance (in a balanced design) to one of three groups: (a) nonintervention controls, (b) advice and booklet, and (c) advice and booklet plus the offer of nicotine gum. There were six practices, with recruitment over 3 weeks, 1 week to each regime. There were 18 clusters (practices) and 1938 subjects. The unit of analysis will be subject nested within practice.

1. In a trial of clinical guidelines to improve general-practice management and referral of infertile couples, Emslie et al. (1993) randomized 82 general practices in Grompian region and studied 100 couples in each group. The outcome measure was whether the general practitioner had taken a full sexual history and examined and investigated both partners appropriately, so that the general practitioner would be the unit of analysis.

2. A third example is the Swedish two-county trial of mammographic screening for breast cancer. In this study, clusters (geographical areas) were randomized within strata, comprising 12 pairs of geographical clusters in Ostergotland county and 7 triplets in Kopperberg county. The strata were designed so that clusters within a stratum were similar in socioeconomic terms. It should be noted that for randomization or allocation by cluster, there is a price to be paid at the analysis stage. We can no longer think of our trial subjects as independent individuals, but must do the analysis at the level of the sampling unit. This is because we have two sources of variation, one between subjects within a cluster and the other between clusters; and the variability between clusters must be taken into account. Clustering leads to a loss of power and a need to increase the sample size to cope up with the loss. The excess variation resulting from randomization being at the cluster level rather than the individual level was neglected by Tabar...
et al. (1985). They claimed (without supporting evidence) that such excess variation was negligible. This study was later analyzed by Duffy et al. (2003) who used hierarchical modeling to take clustering into account and found an evidence for an effect. Taking account of the cluster randomization, there was a significant 30% reduction in breast cancer mortality. They concluded that mammographic screening does indeed reduce mortality from breast cancer and that the criticisms of the Swedish two-county trial were unfounded. The fact is that the criticism was founded, because it was wrong to ignore clustering in such a study. Getting the same answer when we do it correctly is irrelevant.

There are several approaches that can be used to allow for clustering ranging from simple to quite sophisticated:

1. Whether the outcome is normally distributed or not, group comparisons may be achieved by adjusting the standard error of the “effect measure” using the DEFF. These are generally approximate methods, but more realistic than ignoring the within-cluster correlation.
2. Multilevel or hierarchical or random effects modeling.
3. When covariates adjustment are required within the regression analyses, the “generalized estimating equation” (GEE) approach is used.
4. Bayesian hierarchical models.

The focus in this book will be on the statistical analysis of correlated data using the first three approaches.

1.1.2 Sample and Design Issues

As we have seen, the main reasons for adopting trials with clusters as the sampling unit are:

- Evaluation of interventions, which by their nature have to be implemented at community level, e.g., water and sanitation schemes, and some educational interventions, e.g., smoking cessation project.
- Logistical convenience, or to avoid the resentment or contamination that might occur if unblended interventions were provided for some individuals, but not others in each cluster.

It might be desirable to capture the mass effect on disease of applying an intervention to a large proportion of community or cluster members, for example, due to “an overall reduction in the transmission of an infectious agent” (Hayes and Bennett, 1999).

The within-cluster correlation can come about by any of several mechanisms, including shared exposure to the same physical or social environment,
self-selection in belonging to the cluster or the group, or sharing ideas or diseases among members of the group.

As we see from Equation 1.9, the DEFF is a function of the cluster size and the ICC. The values of $\rho$ tend to be larger in small groups such as a family, and smaller in large groups such as a county or a village because the degree of clustering often depends on the interaction of group members; family members are usually more alike than individuals in different areas of a large geographic region. Unfortunately, the influence of $\rho$ on study power is directly related to cluster size. Studies with a few large clusters are often very inefficient.

**Example 1.1** Comparison of Means

We assume in this example that $k$ clusters of $n$ individuals are assigned to each of an experimental group $E$ and a control group $C$. The aim is to test the hypothesis $H_0: \mu_E = \mu_C$, where $\mu_E$ and $\mu_C$ are the means of the two groups, respectively, of a normally distributed response variable $Y$ having common but unknown variance $\sigma^2$.

Unbiased estimates of $\mu_E$ and $\mu_C$ are given by the usual sample means $\bar{y}_E$ and $\bar{y}_C$, where

$$\bar{y}_E = \frac{1}{nk} \sum_{i=1}^{k} \sum_{j=1}^{n} y_{ij}$$

and

$$V(\bar{y}_E) = \frac{\sigma^2}{nk} [1 + (n - 1)\rho]$$

(1.11)

with similar expression for $V(\bar{y}_C)$.

For sample size determination, expression 1.11 implies that the usual estimate of the required number of individuals in each group should be multiplied by the inflation factor or DEFF $= [1 + (n - 1)\rho]$ to provide the same statistical power as would be obtained by randomizing $nk$ subjects to each group when there is no clustering effect. More formally, let $z_{\alpha/2}$ denote the value of a standardized score cutting of 100$\alpha/2$% of each tail of a standard normal distribution and $z_\beta$ denote the value of a standardized score cutting of the upper 100$\beta$%. Then the test $H_0: \mu_E = \mu_C$ versus $H_1: \mu_E \neq \mu_C$ has a power of 100(1$-\beta$)% when performed at the 100$\alpha$% level of significance; the number of individuals $n$ required in each group is given by

$$n' = 2(z_{\alpha/2} + z_\beta)^2 \sigma^2 [1 + (n - 1)\rho] / \delta^2$$

(1.12)

where $\delta$ is a “meaningful difference” specified in advance by the investigator. Alternatively, Equation 1.12 may be written as

$$n' = (z_{\alpha/2} + z_\beta)^2 [1 + (n - 1)\rho] / \Delta^2$$

(1.13)

where $\Delta = \frac{\mu_E - \mu_C}{\sigma \sqrt{2}} = \frac{\delta}{\sigma \sqrt{2}}$. 
At $\rho = 0$, formula 1.12 reduces to the usual sample size specification given by Snedecor and Cochran (1981). When $\rho = 1$, there is variability within the cluster, and the usual formula applies with $n$ as the number of clusters that must be sampled from each group. In most epidemiologic applications, however, values of $\rho$ tend to be no greater than 0.6, and advance estimates may also be available from previous data. Obtaining an estimate of $\rho$ is no different in principle from the usual requirement imposed on the investigator to supply an advance estimate of the population variance $\sigma^2$. In the case of unequal cluster sizes, we may replace $n$ in the right-hand side of Equation 1.12 or 1.13 by the average cluster size, $\bar{n}$ (or by $n_0$). A conservative approach would be to replace $n$ by $n_{\text{max}}$, the largest expected cluster size in the sample.

We now consider a test of significance on $\bar{y}_E - \bar{y}_C$. Note that in most applications, the clusters are likely to be of unequal size $n_i, i = 1, 2, \ldots, k$. In this case, formula 1.11 generalizes to

$$V(\bar{y}_E) = \frac{\sigma^2}{N} \left[ 1 + \left( \frac{\sum n_i^2}{N} \right) \frac{\sigma^2}{\sigma^2_{\epsilon}} \right]$$

An estimate $\hat{V}(\bar{y}_E)$ of $V(\bar{y}_E)$ may be calculated by substituting $\hat{\sigma}^2_{\epsilon}$ and $\hat{\sigma}^2_{\epsilon}$ (defined earlier), the sample estimates of $\sigma^2_{\epsilon}$ and $\sigma^2_{\epsilon}$, respectively, in Equation 1.14.

A large sample test on $H_0: \mu_E = \mu_C$ may be obtained by calculating

$$Z = \frac{\bar{y}_E - \bar{y}_C}{\left[ \hat{V}(\bar{y}_E) + \hat{V}(\bar{y}_C) \right]^{1/2}}$$

(See Donner et al., 1981.)

**Example 1.2**

The milk yield data from 10 Ontario farms, 5 large and 5 small farms, are analyzed. Each farm provided 12 observations representing the average milk yield per cow per day for each month. For the purpose of this example, we shall ignore the sampling time as a factor and consider each farm as a cluster size of 12. The following SAS code shows how to read in the data and run the general linear model.

```sas
data milk;
input farm milk size $;
cards;
1 32.33 L
1 29.47 L
1 30.19 L
···
10 24.12 S
;
proc sort data=milk; by size;
```
proc glm data = milk;
class farm;
model milk = farm;
run;

The ANOVA results from the SAS output are given below:

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>9</td>
<td>905.122484</td>
<td>100.569165</td>
</tr>
<tr>
<td>Error</td>
<td>110</td>
<td>734.014075</td>
<td>6.672855</td>
</tr>
</tbody>
</table>

\[
\hat{\sigma}_e^2 = 6.67 \\
\hat{\sigma}_b^2 = \frac{100.57 - 6.67}{12} = 7.83
\]

Therefore, the estimated ICC is \(\hat{\rho} = \frac{7.83}{7.83 + 6.67} = 0.54\).

An important objective of this study was to test whether average milk yield in large farms differs significantly from the average milk yield of small farms. That is to test \(H_0: \mu_s = \mu_l\) versus \(H_1: \mu_s \neq \mu_l\).

The ANOVA results separately for each farm size (small and large) are now produced. The SAS code is

proc glm data = milk;
class farm;
model milk = farm;
by size;
run; quit;

### Large farm size

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>283.4424600</td>
<td>70.8606150</td>
</tr>
<tr>
<td>Error</td>
<td>55</td>
<td>407.2490000</td>
<td>7.4045273</td>
</tr>
</tbody>
</table>

### Small farm size

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>379.2289833</td>
<td>94.8072458</td>
</tr>
<tr>
<td>Error</td>
<td>55</td>
<td>326.7650750</td>
<td>5.9411832</td>
</tr>
</tbody>
</table>

For large farms, since \(\bar{y}_l = 28.00, k = 5, n = 12, N = 60\)

\[
\hat{\sigma}_{\alpha}^2 = 7.40, \quad \hat{\sigma}_{\beta l}^2 = \frac{70.86 - 7.4}{12} = 5.29
\]

\[
\hat{\sigma}_{\beta l}^2 / \hat{\sigma}_{\alpha}^2 = 0.71
\]

then \(\hat{V}(\bar{y}_l) = \frac{7.40}{60}[1 + (12)(0.71)] = 1.17\).
For small farms

\[
\bar{y}_s = 25.32, \quad k = 5, \quad n = 12, \quad N = 60
\]

\[
\hat{\sigma}_{es}^2 = 5.94, \quad \hat{\sigma}_{bs}^2 = \frac{94.81 - 5.94}{12} = 7.4
\]

\[
\hat{\sigma}_{es}^2 = 1.24
\]

\[
\hat{V}(\bar{y}_s) = \frac{5.94}{60} [1 + (12)(1.24)] = 1.57
\]

\[
Z = \frac{28 - 25.32}{(1.17 + 1.57)^{1/2}} = 1.61
\]

\[
p\text{-value} = 0.10\text{, and there is no sufficient evidence in the data to support } H_0.\]

The R code to read in the milk data and produce the ANOVA results:

```r
milk <- read.table("x:/xxx/milk.txt",header=T)

# ANOVA results overall
anova(lm(milk ~ factor(farm), data=milk))

# ANOVA results for large farms
anova(lm(milk ~ factor(farm), data=milk[milk$size=='L',]))

# ANOVA results for small farms
anova(lm(milk ~ factor(farm), data=milk[milk$size=='S',]))
```

### 1.2 Regression Analysis for Clustered Data

A fundamental question in many scientific investigations is concerned with how and to what extent a response variable is related to a set of independent variables. For example, a health economist may be interested in the relationship between the effect of intervention and the cost of its administration, a clinical nutritionist in the relationship between obesity and hypertension, or a radiologist in the relationship between the ultrasound diagnosis of cancer and the tumor size. The list of situations of this kind in biomedical research is endless, let alone other areas of applications.

Suppose for a given situation, the actual mathematical relationship between the response variable "Y" and a set of independent variables is known. The investigator is then in a position to understand the factors that control the direction of the response. Unfortunately, there are few situations in practice in which the true mathematical model connecting the response to the independent variables is known. Consequently, one is forced to combine practical experience and mathematical theory to develop an approximate
model that characterizes the main features of the behavior of the response variable.

Regression analysis is among the most commonly used methods of statistical analysis to model the relationship between variables. Its objective is to describe the relationship of response with independent or explanatory variables. In its very general form, a regression model is written as

$$Y = X\beta + e$$

(1.16)

where $Y$ is the ($n \times 1$) vector of dependent variable values, $X$ an ($n \times (p + 1)$) matrix containing the values of the independent variables, $\beta$ the ($p + 1 \times 1$) vector of parameters, and $e$ the ($n \times 1$) vector of error components. It is well known that the method of least squares is the most preferred method of estimation of the parameter of the regression model. There are fundamental assumptions that should be satisfied to use this method to estimate the parameter vector $\beta$:

(a) The components of $Y$ are uncorrelated with each other.

(b) The error components $e$ are assumed to be uncorrelated with mean 0 and common variance $\sigma^2$.

(c) The vector of error components $e$ is uncorrelated with the matrix $X$.

Under the above conditions and provided that $(X^TX)^{-1}$ exists, the least squares estimate of $\beta$ is

$$\hat{\beta} = (X^TX)^{-1}X^TY$$

(1.17)

Equation 1.17 is important in regression analysis since it provides the estimates of $\beta$ once we are sure that conditions (a, b, c) are satisfied and the matrix $X$ is specified.

In addition to the linear models (Equation 1.16), regression models include logistic models for binary responses, log-linear model for counts, and survival analysis for time to events. In this chapter, we discuss the linear-normal model for continuous responses when the basic assumption that all the observations are independent, or at least uncorrelated, is violated. Recall that the assumption of zero correlation would mean that knowing one subject’s response provides no information regarding the status of another subject in the same study. However, the assumption of independence may not hold if the subjects belong to the same cluster as has been already demonstrated. As an example of a regression problem when clusters of subjects are sampled together is Miall and Oldham’s arterial blood pressure levels family study. Owing to their common household environment and their shared genetic makeup, we would expect a family member to have a greater chance of having elevated blood pressure levels if his/her sibling had the same. Data from this study can be usefully thought of as being “clustered” into families. Blood pressure levels from different families are likely to be independent; those from the same
family are not. This dependence among observations from the same cluster must be accounted for in assessing the relationships between risk factors and health outcomes.

Another example cited by Liang and Zeger (1993) is the growth study of Hmong refugee children. In this example, 1000 Hmong refugee children receiving health care at two Minnesota clinics between 1976 and 1985 were examined for their growth patterns. The objective was to study the patterns of growth and its association with age at entry into the United States. It is believed that stature is influenced by both genetic and environmental factors. When the offending environmental factors are removed, the growth process progresses at a faster rate. To study the growth, repeated measurements of height of each child were recorded. The number of visits per child ranged from 1 to 15 and averaged 5. The correlation between repeated observations on height for each child may be a nuisance but cannot be ignored in regression analysis.

The above two examples have common features, although they address questions with different scientific objectives. Data in the above two studies are organized in clusters. For family study, the clusters are formed by families, and in the second example a cluster comprises the repeated observations for a child. Another aspect of similarity between the two studies is that one can safely assume that the response variables (blood pressure in the family study and height in the growth study) are normally distributed. The two studies also differ in the structure of the within-cluster correlation. For example, in the family study one may assume that the correlation between the pairs of sibs within the family is equal, that is, we may assume a constant within-cluster correlation. For repeated measures longitudinal study, the situation is different. Although the repeated observations are correlated, this correlation may not be constant across time (cluster units). It is common sense to assume that observations taken at adjacent time points are more correlated than observations that are taken at separated time points.

In the remainder of this chapter, we shall focus on regression analysis of clustered data assuming common or fixed within-cluster correlation and the response variable is normally distributed. Other types of response variables and different correlation structures will be discussed in detail in subsequent chapters.

Within the framework of linear regression, we illustrate an answer to the question: what happens when the conventional linear regression is used to analyze clustered data?

Let $Y_{ij}$ denote the score on the $j$th offspring in the $i$th family; $X_i$ the score of the $i$th parent, where $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, k$; $n_i$ the number of offspring in the $i$th family; and $k$ the total number of families. We assume that the regression of $Y$ on $X$ is given by

$$Y_{ij} = \mu_y + \beta(X_i - \mu_x) + E_{ij}$$

(1.18)
where $\mu_y = E(Y_{ij})$, $\mu_x = E(X_i)$, $\beta$ is the regression coefficient of $Y$ on $X$, and $E_{ij}$ is the deviation of the $j$th offspring of the $i$th parent. We further assume that

$$\text{Cov}(Y_{ij}, Y_{il}) = \begin{cases} \rho \sigma^2 & j \neq l \\ \sigma^2 & j = l \end{cases}$$

Under this model, Kempthorne and Tandon (1953) showed that the minimum variance unbiased estimator of $\beta$ is given by

$$b_1 = \frac{\sum_{i=1}^k w_i (x_i - \bar{x}) y_i}{\sum w_i (x_i - \bar{x})^2}$$

(1.19)

and

$$V(b_1) = \frac{\sigma^2 (1 - \rho)}{\sum w_i (x_i - \bar{x})^2}$$

where $w_i = n_i / (1 + n_i \rho_e)$, $\rho_e = (\rho - \beta^2) / (1 - \beta^2)$, $\bar{y}_i = \sum_j y_{ij} / n_i$, and $\bar{x} = \sum w_i x_i / \sum w_i$.

The most widely used estimator for $\beta$ ignores the ICC $\rho$ and is given by the usual estimator:

$$b = \frac{\sum_{i=1}^k n_i (\bar{y}_i - \bar{y}) (x_i - \bar{x})}{\sum_{i=1}^k n_i (x_i - \bar{x})^2}$$

(1.20)

where $\bar{y} = \sum n_i \bar{y}_i / N$, $\bar{x} = \sum n_i x_i / N$, and $N = n_1 + n_2 + \cdots + n_k$.

$$V(b) = \frac{\sigma^2 (1 - \rho) \sum n_i (1 + n_i \rho_e) (x_i - \bar{x})^2}{\left[ \sum n_i (x_i - \bar{x})^2 \right]^2}$$

(1.21)

If $\rho_e = 0$, then $w_i = n_i$ and $\text{var}(b) = \text{var}(b_1) = \sigma^2 (1 - \rho) / \sum n_i (x_i - \bar{x})^2$, which means that $b_1$ is fully efficient.

Therefore,

$$\frac{V(b)}{V(b_1)} = \left[ \frac{\sum n_i (1 + n_i \rho_e) (x_i - \bar{x})^2}{\sum n_i (x_i - \bar{x})^2} \right]$$

or equivalently

$$\frac{V(b)}{V(b_1)} = 1 + \frac{\rho_e \sum n_i^2 (x_i - \bar{x})^2}{\sum n_i (x_i - \bar{x})^2}$$

(1.22)

Assuming $\rho_e > 0$, $b$ is always less efficient.

The most important message of Equation 1.22 is that ignoring within-cluster correlation can lead to a loss of power when both within-cluster and cluster-level covariate information are being used to estimate the regression coefficient.

To analyze clustered data, one must therefore model both the regression of $Y$ on $X$ and the within-cluster dependence. If the responses are independent of each other, then ordinary least squares can be used, which produces
regression estimators that are identical to the maximum likelihood in the case of normally distributed responses. In this chapter we consider two different modeling approaches: marginal and random effects. In marginal models, the regression of $Y$ on $X$ and the within-cluster dependence are modeled separately. The random effects models attempt to address both issues simultaneously through a single model. We shall explore both modeling strategies for a much wider class of distributions named “generalized linear models” (GLM) that includes the normal distribution as a special case.

### 1.3 Generalized Linear Models

GLM are a unified class of regression methods for continuous and discrete response variables. There are two components in a GLM, the systematic component and the random component. For the systematic component, one relates $Y$ to $X$ assuming that the mean response $\mu = E(Y)$ satisfies 

$$g(\mu) = X_1\beta_1 + X_2\beta_2 + \cdots + X_p\beta_p$$

which may conveniently be written as

$$g(\mu) = X^T \beta$$ \hspace{1cm} (1.23)

Here, $g(.)$ is a specified function known as the “link function.” The normal regression model for continuous data is a special case of Equation 1.23, where $g(.)$ is the identity link. For binary response variable $Y$, the logistic regression is a special case of Equation 1.23 with the logit transformation as the link. That is,

$$g(\mu) = \log \frac{\mu}{1 - \mu} = \log \frac{P(Y = 1)}{P(Y = 0)} = X^T \beta$$

When the response variable is count, we assume that

$$\log E(Y) = X^T \beta$$

To account for the variability in the response variable that is not accounted for by the systematic component, GLM assume that $Y$ has the probability density function given by

$$f(y) = \exp[(y\theta - b(\theta))/\phi + C(y, \phi)] \hspace{1cm} (1.24)$$

which is a general form of the exponential family of distributions. This includes among others, the normal, binomial, and Poisson as special cases. It can be easily shown that

$$E(Y) = b'(\theta)$$

and

$$V(Y) = \phi b''(\theta)$$
For the normal distribution:

\[ \theta = \mu \text{ (identity link)} \]
\[ b(\theta) = \mu^2/2, \quad \phi = \sigma^2 \]

Hence, \( b'(\theta) = \mu \) and \( b''(\theta) = 1 \), indicating that \( \phi \) is the variance.

For the Poisson distribution:

\[ \theta = \ln \mu \text{ (log-link)} \]
\[ b(\theta) = \mu = e^\theta, \quad \phi = 1 \]
\[ b'(\theta) = e^\theta = \mu \]
\[ b''(\theta) = e^\theta = \mu \]

Hence, \( E(Y) = V(Y) = \phi \mu \).

The scale parameter \( \phi \) in Equation 1.24 is called the over-dispersion parameter. If \( \phi > 1 \), then the variance of the counts is larger than the mean.

When the link function in Equation 1.23 and the random component are specified by the GLM, we can estimate the regression parameters \( \beta \) by solving the estimating equation:

\[
U = \sum_{i=1}^{p} \left( \frac{\partial \mu_i(\beta)}{\partial \beta} \right)^T V^{-1}(Y_i)[Y_i - \mu_i(\beta)] = 0 \quad (1.25)
\]

The above equation provides valid estimates when the responses are independently distributed. For clustered data, the GLM and hence Equation 1.25 are not sufficient, since the issue of within-cluster correlation is not addressed. We now discuss the two modeling approaches commonly used to analyze clustered data.

1.3.1 Marginal Models (Population Average Models)

As already mentioned, in a marginal model, the regression of \( Y \) on \( X \) and the within-cluster dependence are modeled separately. We assume:

1. The marginal mean or “population average” of the response, \( \mu_{ij} = E(Y_{ij}) \), depends on the explanatory variables \( X_{ij} \) through a link function \( g(\mu_{ij}) = X_{ij}^T \beta \), where \( g \) is a specified link function.

2. The marginal variance depends on the marginal mean through \( V(Y_{ij}) = \phi V(\mu_{ij}) \), where \( V(\cdot) \) is a known variance function, such as \( V(\mu_{ij}) = \phi \) for normally distributed response and \( V(\mu_{ij}) = \mu_{ij} \) for count data similar to the GLM setup.
3. Cov($Y_{ij}, Y_{il}$), the covariance between pairs within clusters, is a function of the marginal means and another additional parameter $\alpha$, i.e.,

$$\text{Cov}(Y_{ij}, Y_{il}) = \gamma(\mu_{ij}, \mu_{il}, \alpha)$$

where $\gamma(.)$ is a known function.

### 1.3.2 Random Effects Models

There are several names given to these types of models: multilevel, hierarchical, random coefficients, or mixed effects models. The fundamental feature of these models is the assumption that parameters vary from cluster to cluster, reflecting natural heterogeneity due to unmeasured cluster-level covariates.

The general setup for the random effects GLM was described by Zeger and Karim (1991) as follows:

1. Conditional on random effects $b_i$, specific to the $i$th cluster, the response variable $Y_{ij}$ follows a GLM with

$$g[E(Y_{ij}|b_i)] = X_{ij}^T \beta + Z_{ij}^T b_i$$  \hspace{1cm} (1.26)

where $Z_{ij}$, a $q \times 1$ vector of covariates, is a subset of $X_{ij}$.

2. Conditional on $b_i$, $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})^T$ are statistically independent.

3. The $b_i$’s are independent observations from a distribution $F(.)$, indexed by some unknown parameter $\alpha$. The term “random effect” was assigned to the variable $b_i$, because we treat $b_i$ as a random sample from $F$. The random effects $b_i$ are not related to $X_{ij}$.

### 1.3.3 Generalized Estimating Equation (GEE)

The GEE provides a tool for practical statistical inference on the $\beta$ coefficient under the marginal model when the data are clustered. The regression estimates are obtained by solving the equation:

$$U_1(\beta, \alpha) = \sum_{i=1}^{k} \left( \frac{\partial \mu_i}{\partial \beta} \right)^T \text{Cov}^{-1}(Y_i, \beta, \alpha)[y_i - \mu_i(\beta)] = 0$$  \hspace{1cm} (1.27)

where $\mu_i(\beta) = E(Y_i)$, the marginal mean of $Y_i$. One should note that $U_1(.)$ has the same form of $U(.)$ in Equation 1.25, except that $Y_i$ is now $n_i \times 1$ vector, which consists of the $n_i$ observations of the $i$th cluster, and the covariance matrix Cov($Y_i$), which depends on $\beta$ and $\alpha$, a parameter that characterizes the within-cluster dependence.

For a given $\alpha$, the solution $\hat{\beta}$ to Equation 1.27 can be obtained by an iteratively reweighted least squares calculations. The solution to these equations...
is a consistent estimate of $\beta$, provided that $\mu_i(\beta)$ is correctly specified. The consistency property follows because $(\frac{\partial \mu_i}{\partial \beta})^T \text{Cov}^{-1}(Y_i)$ does not depend on the $Y_i$’s, so Equation 1.27 converges to 0 and has consistent roots as long as $E(Y_i - \mu_i(\beta)) = 0$ (see Liang and Zeger (1986) and Zeger and Liang (1986)).

If a $\sqrt{k}$ consistent estimate of $\alpha$ is available, $\hat{\beta}$ are asymptotically normal, even if the correlation structure is misspecified. Liang and Zeger (1986) proposed a “robust” estimate of the covariance matrix of $\hat{\beta}$ as

$$V(\hat{\beta}) = A^{-1}MA^{-1}$$

where

$$A = \sum_{i=1}^{k} \tilde{D}_i^T \tilde{V}_i^{-1} \tilde{D}_i$$

$$M = \sum_{i=1}^{k} \tilde{D}_i^T \tilde{V}_i^{-1} \text{Cov}(Y_i) \tilde{V}_i^{-1} \tilde{D}_i$$

$$\text{Cov}(Y_i) = (Y_i - \mu_i(\widetilde{\beta}))(Y_i - \mu_i(\widetilde{\beta}))^T$$

where $\sim$ denotes evaluation at $\tilde{\beta}$ and $\tilde{\alpha}(\tilde{\beta})$.

Liang and Zeger (1986) proposed a simple estimator for $\alpha$ based on Pearson’s residuals

$$\hat{r}_{ij} = \frac{y_{ij} - \mu_{ij}(\beta)}{\sqrt{\text{var}(y_{ij})}}$$

For example, under common correlation structure that $\text{Cor}(y_{ij}, y_{il}) = \alpha$ for all $i, j$, and $l$, an estimate of $\alpha$ is

$$\hat{\alpha} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \sum_{l=1}^{n_i-1} \hat{r}_{ij} \hat{r}_{il}}{\left[ \sum_{i=1}^{k} \left( \frac{n_i}{2} \right) - p \right]}$$

(1.28)

where $p$ in the denominator of Equation 1.28 is the number of regression parameters.

One of the limitations of the above approach is that estimation of $\beta$ and $\alpha$ from Equations 1.27 and 1.28 are done as if $(\beta, \alpha)$ are independent of each other. Consequently, very little information from $\beta$ is used when estimating $\alpha$. This can lead to a significant loss of information on $\alpha$. To improve on the efficiency of estimating $\beta$, Prentice (1988) and Liang et al. (1992) discussed estimating $\theta = (\beta, \alpha)$ jointly by solving

$$U_2(\beta, \alpha) = \sum_{i=1}^{k} \left( \frac{\partial \mu_i^*}{\partial \beta} \right) [\text{Cov}(Z_i, \delta)]^{-1}(Z_i - \mu_i^* \theta) = 0$$
where

\[ Z_i = (Y_{i1}, \ldots, Y_{in_i}, Y_{i1}^2, \ldots, Y_{in_i}^2, Y_{i1}Y_{i2}, Y_{i1}Y_{i3}, \ldots, Y_{i1}Y_{i(n_i-1)}, Y_{i1}Y_{in_i})^T \]

where \( \mu^*_i = E(Z_i, \theta) \), which is completely specified by the modeling assumptions of the GLM. They called this expanded procedure that uses both the \( Y_{ij} \)'s and \( Y_{ij}Y_{il} \) the GEE2.

GEE2 seems to improve the efficiency of both \( \beta \) and \( \alpha \). On the other hand, the robustness property for \( \beta \) of GEE is no longer true. Hence, correct inferences about \( \beta \) require correct specification of the within-cluster dependence structure. The same authors suggest using a sensitivity analysis when making inference on \( \beta \). That is, one may repeat the procedure with different models for the within-cluster dependence structure to examine the sensitivity of \( \tilde{\beta} \) to choose the dependence structure.

### 1.4 Fitting Alternative Models for Clustered Data

#### Example 1.3
We will use a subset of the data from a survey conducted by Miall and Oldham (1955) to assess the correlations in systolic and diastolic blood pressures among family members living within 25 miles of Rhonda Fach Valley in South Wales. The purpose of the following analysis is to illustrate the effect of the within-cluster correlation in the case of the normal linear regression model. The part of the data that we use for this illustration consists of the observations made on siblings and their parents. Each observation consists of systolic and diastolic blood pressures to the nearest 5 or 10 mmHg. In this analysis, we will not distinguish among male and female siblings. The following variables will be used to run a few models in this section.

- **familyid**: Family ID
- **subjid**: Sibling ID
- **sbp**: Sibling systolic blood pressure
- **msbp**: Mother systolic blood pressure
- **age**: Sibling age
- **armgirth**: Sibling arm girth
- **cenmsbp**: Mother systolic blood pressure centered
- **cenage**: Sibling age centered within the family
- **cengirth**: Sibling arm girth centered within the family

The records with missing values of sibling age, mother systolic blood pressure, sibling arm girth, and sibling systolic blood pressure are deleted. The dataset
Analysis of Correlated Data with SAS and R

consists of 488 observations on 154 families. The family size ranges from 1 to 10. We begin by first analyzing the data using the GEE approach. This is followed by a series of models using the multilevels modeling approach. All models are fitted using the SAS procedures GENMOD for the GEE and the MIXED for the multilevels approach. The equivalent R code (R 2.3.1—A Language and Environment Copyright 2006, R Development Core Team) is also provided.

The SAS code to read in the data and fit this model is

```sas
data fam;
  input familyid subjid sbp age armgirth msbp;
datalines;
  1 1 85 5 5.75 120
  1 2 105 15 8.50 120
  .......
  200 5 135 40 12.50 255
213 1 120 64 11.00 110
;

* Computing the overall mean for msbp;
proc means data=fam noprint; var msbp; output out=msbp mean=mmsbp; run;

* Computing cluster-specific means for age and armgirth;
proc means data=fam noprint; class familyid; var age armgirth;
  output out=fmeans mean=mage marmgirth; run;
data fmeans; set fmeans; if familyid= . then delete; drop _TYPE_ _FREQ_; run;

* Centering msbp at overall mean and age and armgirth at cluster-specific means;
data fam; merge fam fmeans; by familyid; if _n_=1 then set msbp(drop=_TYPE_ _FREQ_);
cenage=age-mage; cengirth=armgirth-marmgirth; cenmsbp=msbp-mmsbp; keep familyid subjid sbp age armgirth msbp cenage cengirth cenmsbp;
proc genmod data=fam;
class familyid;
model sbp = msbp cenage cengirth/dist = n lnk=id;
  repeated subject = familyid / type = cs corrw;
run;
```
The following is the partial output showing the analysis of the GEE parameter estimates:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>95% Confidence Limits</th>
<th>Z</th>
<th>Pr &gt;</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>119.0985</td>
<td>0.9198</td>
<td>117.2957</td>
<td>120.9013</td>
<td>129.48</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>cenmsbp</td>
<td>0.2024</td>
<td>0.0349</td>
<td>0.1340</td>
<td>0.2707</td>
<td>5.80</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>cenage</td>
<td>0.1802</td>
<td>0.1984</td>
<td>-0.2086</td>
<td>0.5690</td>
<td>0.91</td>
<td>0.3636</td>
</tr>
<tr>
<td>cengirth</td>
<td>3.5445</td>
<td>0.8263</td>
<td>1.9249</td>
<td>5.1641</td>
<td>4.29</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

This model treats the within-cluster correlation as nuisance. It is assumed that the within-subject correlation structure is exchangeable or compound symmetry. The estimated working correlation under the common (compound symmetry) structure is 0.328. The analysis indicates that the arm girth and the mother systolic blood pressure are significant predictors of the sibling systolic blood pressure levels.

We now illustrate the application of the random effects model to analyze clustered data. We followed an informative strategy given by Singer (1998) for fitting multilevel data. Therefore, we shall present three nested random effects models and discuss the relative advantages of each model.

**PROC MIXED for Clustered Data**

Here we illustrate how two levels of clustered data are analyzed using PROC MIXED in SAS (SAS Institute 1995, and 1996). By two levels we mean a situation where subjects are nested within organizational units. The subjects in the dataset are the siblings and the clusters are the families. We are interested in examining the behavior of a level 1 outcome (siblings outcome) as function of level 1 and level 2 (family) covariates. The siblings-level outcomes are the systolic blood pressures (sbp), and the covariates measured at the siblings level are age (cenage) and arm girth (cengirth). There are several family-level outcomes, but we shall restrict to the systolic blood pressure of the mother (cenmsbp).

The first baseline model is called an “unconditional means” model, which examines the variability in sbp across families.

**Model 1: Unconditional Mean Model**

Under this model, we express sbp \((y_{ij})\) as a one-way random effects model

\[
y_{ij} = \mu + b_i + e_{ij}
\]

\(b_i \approx N(0, \sigma_b^2)\) and \(e_{ij} \approx N(0, \sigma_e^2)\)
The model has one fixed effect (\(\mu\)) and two variance components—one representing the variation between family means (\(\sigma_b^2\)) and the other variation among siblings within families (\(\sigma_e^2\)).

The SAS code to fit this model is

```sas
proc mixed data=family noclprint noitprint;
class familyid;
model sbp=/bw;
random familyid;
run;
```

The selected output is shown below.

### Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>familyid</td>
<td>106.98</td>
</tr>
<tr>
<td>Residual</td>
<td>166.22</td>
</tr>
</tbody>
</table>

### Solution for Fixed Effects

| Effect | Estimate | Standard Error | DF  | T-Value | Pr > |t|
|--------|----------|----------------|-----|---------|------|
| Intercept | 118.67  | 1.0554         | 487 | 112.44  | <0.0001 |

The mixed procedure produces a set of information: the “familyid” estimate is an estimate of the parameter (\(\sigma_b^2\)), while the “residual” estimate is an estimate of the parameter (\(\sigma_e^2\)). The maximum likelihood estimate of the ICC is

\[
\hat{\rho} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2} = \frac{106.98}{106.98 + 166.22} = 0.39
\]

This tells us that there is a great deal of clustering of systolic blood pressure levels of siblings within families.

There is another approach that generalizes more easily to data with multiple levels. This approach expresses the subject-level outcome \(y_{ij}\) using a pair of linked models: one at the subject level (level 1) and another at the family level (level 2). At level 1, we express \(y_{ij}\) as the sum of an intercept for the subject’s family (\(\beta_{i0}\)) and random error (\(e_{ij}\)):

\[
y_{ij} = \beta_{i0} + e_{ij} \quad \text{where} \quad e_{ij} \approx N(0,\sigma_e^2).
\]
Analyzing Clustered Data

At the higher level (family level), we express the family-level intercept as the sum of an overall mean ($\beta$) and random deviation ($u_{i0}$) so that

$$\beta_{i0} = \beta + u_{i0} \quad \text{where} \quad u_{i0} \approx N(0, \tau_0)$$

Therefore,

$$y_{ij} = \beta + u_{i0} + e_{ij}$$

The SAS code for this model is

```sas
proc mixed data=family noclprint noitprint covtest;
    class familyid;
    model sbp = /s ddfm=bw;
    random intercept/sub = familyid;
run;
```

The parameter estimates under this model are the same as in the previous model. The purpose of the `covtest` option in the “proc mixed” statement is to test the hypothesis for the variance components.

**Model 2: Including a Family-Level Covariate**

In this model, we include the mother’s systolic blood pressure score (msbp) as a predictor of the siblings score. Following Singer (1998), the msbp is centered at the overall mean so that it has mean 0 and allows a meaningful interpretation of the intercept. For this model we write

$$y_{ij} = \beta_{i0} + e_{ij}$$

$$\beta_{i0} = \gamma_{00} + \gamma_{01} x_i + u_{i0}$$

where

$$e_{ij} \approx N(0, \sigma^2_e), \quad u_{i0} \approx N(0, \sigma^2_{0})$$

Therefore, $y_{ij} = (\gamma_{00} + \gamma_{01} x_i) + (u_{i0} + e_{ij})$, where $x_i = \text{msbp} - \text{mean(msbp)} = \text{cenmsbp}$.

The above model has two components, a fixed part enclosed in the first bracket and a random part enclosed in the second bracket. The SAS code to fit this model is

```sas
proc mixed data=family noclprint noitprint;
    class familyid;
    model sbp = cenmsbp /s ddfm=bw notest;
    random intercept/subject = familyid;
run;
```

Note that there is another option in the “model” statement, which is `ddfm = bw`. This allows SAS to use the “between/within” method for computing the denominator degrees of freedom for tests of fixed effects. See SAS documentation or Littell et al. (1996) for details.
The SAS output is

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>familyid</td>
<td>67.4679</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>163.89</td>
</tr>
</tbody>
</table>

Solution for Fixed Effects

| Effect  | Estimate | Standard Error | DF  | t-Value | Pr > |t| |
|---------|----------|----------------|-----|---------|-----|----|
| Intercept | 119.11   | 0.9161         | 152 | 130.01  | <0.0001 |
| cenmsbp  | 0.2005   | 0.02748        | 152 | 7.29    | <0.0001 |

Note that there are two fixed effects to be estimated: the intercept and the covariate (MSBP). The null hypothesis, which states that there is no relationship between mother’s systolic blood pressure levels and the siblings, is not supported by the data. Also note that the variance components estimates are 67.47 and 163.89 for $\tau_0$ and $\sigma^2_e$, respectively. These estimates under the present model have different interpretations. In model 1, there were no covariates, so these were unconditional components. After adding the mother’s blood pressure as a covariate, these are now conditional components. Note that the conditional within-family component is slightly reduced (from 166.22 to 163.89). The variance component representing variation between families $\tau_0$ or $\sigma^2$ has diminished markedly (from 106.98 to 67.47). This tells us that the cluster- or family-level covariate (mother’s systolic blood pressure) explains a large percentage of the family-to-family variation. One way of measuring how much variation exists in family mean blood pressures as explained by the mother’s blood pressure levels is to compute how much the variance component for this term $\tau_0$ has diminished between the two models. Following Bryk and Raudenbush (1992), we compute this as $(106.98 - 67.47)/106.98 = 36.9\%$. This is interpreted by saying that about 36% of the explainable variation in the sibling’s mean systolic blood pressure levels is explained by the mother’s systolic blood pressure levels.

**Model 3: Including Sib-Level Covariate**

The simplest model may be written as

$$y_{ij} = \beta_0 + \beta_1 Z_{ij} + e_{ij}$$

Here, $Z_{ij}$ is the age of the $j$th subject within the $i$th family centered at its mean value. The other terms are defined as before. Let

$$\beta_0 = \beta_0 + u_{i0}$$

$$\beta_1 = \beta_{11} + u_{i1}$$
Hence,

\[ y_{ij} = \beta_{00} + u_{i0} + (\beta_{11} + u_{i1})Z_{ij} + e_{ij} \]

\[ = (\beta_{00} + \beta_{11}Z_{ij}) + (u_{i0} + u_{i1}Z_{ij} + e_{ij}) \]

where \( e_{ij} \approx N(0, \sigma^2_e) \), \( u_i = (u_{i0}, u_{i1}) \approx \text{BIVN}(0, \Sigma) \), and \( e_{ij} \) is independent of the bivariate normal random vector \( u_i \). The \( \Sigma \) is a 2 \times 2 symmetric matrix whose elements are \( \delta_{00} = V(u_{i0}), \delta_{11} = V(u_{i1}) \), and \( \delta_{01} = \text{Cov}(u_{i0}, u_{i1}) \). The SAS code to fit this model is

```sas
proc mixed data=family noclprint noitprint;
class familyid;
model sbp=cenmsbp/s ddfm=bw notest;
random intercept cenage/ subject=familyid type=un;
run;
```

Notice that the random statement has two random effects—one for intercept and one for the \( Z \)-slope. There is also an additional option in the random statement, “type=un,” indicating that an unstructured specification for the covariance of \( u_i \) is assumed. Partial SAS output is shown below.

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>UN(1,1)</td>
<td>familyid</td>
<td>78.9632</td>
</tr>
<tr>
<td>UN(2,1)</td>
<td>familyid</td>
<td>4.3682</td>
</tr>
<tr>
<td>UN(2,2)</td>
<td>familyid</td>
<td>1.0199</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>133.16</td>
</tr>
</tbody>
</table>

| Effect    | Estimate | Standard Error | DF  | t-Value | Pr > |t|   |
|-----------|----------|----------------|-----|---------|------|-----|
| Intercept | 118.43   | 0.8977         | 152 | 131.92  | <0.0001 |
| cenmsbp   | 0.1973   | 0.02687        | 152 | 7.34    | <0.0001 |

We shall start by first interpreting the fixed effects. The estimate of the intercept 118.43 indicates the estimated average sibling systolic blood pressure levels after controlling for the mother’s systolic blood pressure. The estimate of the cenmsbp indicates that the average slope representing the relationship between siblings’ blood pressure and the mother’s systolic blood pressure is 0.20. The standard errors of these estimates are very small, resulting in small, \( p \)-values. We conclude that, on average, there is a statistically significant
relationship between siblings’ systolic blood pressures and the mother’s systolic blood pressure.

The covariance parameter estimates tell us how much these intercepts and slopes vary across families. The $\hat{\delta}_{00} = 78.96$ represents the variability in the intercepts, $\hat{\delta}_{11} = 4.37$ the variability in the slopes, and $\hat{\delta}_{01} = 1.02$ the covariance between intercepts and slopes. We can say that the intercepts vary considerably; in other words, families do differ in average systolic blood pressure levels after controlling for the effects of the mother’s blood pressure levels. We also note that the slopes do not considerably vary between the families, and there is no evidence that the effects of mother’s blood pressure on sibling systolic blood pressure differ between the families.

Finally, we compare the residual error variance of the unconditional model to that of the present model. Recall that the estimated variance of the unconditional model was 166.22. Here we have the conditional estimate of 133.16. Inclusion of the sibling’s age is therefore explained as $(166.22 - 133.16)/166.22 = 20.0\%$ of the explainable variation within families.

**Model 4: Including One Family-Level Covariate and Two Subject-Level Covariates**

Following Singer (1998), it is always helpful to write the outcome variable as a function of the covariates measured at the lowest (subject) level. Thereafter, we write the slopes and the intercepts as functions of the higher level (in our example, a family).

$$y_{ij} = B_{i0} + B_{i1}Z_{ij} + B_{i2}a_{ij} + e_{ij}$$

$$B_{i0} = \gamma_{00} + \gamma_{01}x_i + u_i$$

$$B_{i1} = \gamma_{10} + \gamma_{11}x_i + u_1$$

$$B_{i2} = \gamma_{20} + \gamma_{21}x_i + u_2$$

where $Z_{ij}$ is the centered arm girth of the $j$th subject within the $i$th family, $a_{ij}$ is the centered age of the $j$th subject within the $i$th family, and $x_i$ is the centered mother’s systolic blood pressure in the $i$th family.

Hence,

$$y_{ij} = \gamma_{00} + \gamma_{01}x_i + u_i + Z_{ij}(\gamma_{10} + \gamma_{11}x_i + u_1) + a_{ij}(\gamma_{20} + \gamma_{21}x_i + u_2) + e_{ij}$$

Simplifying, we get

$$y_{ij} = [\gamma_{00} + \gamma_{01}x_i + \gamma_{10}Z_{ij} + \gamma_{20}a_{ij} + \gamma_{11}x_iZ_{ij} + \gamma_{21}x_ia_{ij}] + [u_i + Z_{ij}u_1 + a_{ij}u_2 + e_{ij}]$$

Terms in the first bracket should appear in the model statement, while those in the second bracket should appear in the random statement.

The SAS code to fit the model is

```sas
proc mixed data=family covtest noclprint noitprint;
class familyid;
```

model sbp=acenmsbp cenceage cencegirth acenmsbp*cenceage acenmsbp*cengirth/ s ddfm=bw notest;
  random cenceage cencegirth/subject=familyid type=un;
  run;

The variable cengirth has a zero variance component; we fit the model after removing cengirth from the random statement. The results are given below.

Covariance Parameter Estimates

| Cov Parm | Subject  | Estimate | Standard Error | Z-Value | Pr > |Z| |
|----------|----------|----------|----------------|---------|-------|---|
| UN(1,1)  | familyid| 81.8458  | 15.4989        | 5.28    | <0.0001 |
| UN(2,1)  | familyid| 4.7531   | 1.7062         | 2.79    | 0.0053 |
| UN(2,2)  | familyid| 0.6185   | 0.3115         | 1.99    | 0.0235 |
| Residual |          | 124.29   | 11.2232        | 11.07   | <0.0001 |

Solution for Fixed Effects

| Effect          | Estimate | Standard Error | DF | t-Value | Pr > |t| |
|-----------------|----------|----------------|----|---------|-------|---|
| Intercept       | 119.07   | 0.9208         | 152| 129.32  | <0.0001 |
| acenmsbp        | 0.2043   | 0.02777        | 152| 7.36    | <0.0001 |
| cenceage        | 0.02669  | 0.1624         | 330| 0.16    | 0.8695 |
| cencegirth      | 4.2129   | 0.7010         | 330| 6.01    | <0.0001 |
| acenmsbp*cenceage| 0.008845 | 0.004130       | 330| 2.14    | 0.0330 |
| acenmsbp*cengirth| -0.02641| 0.01736        | 330| -1.52   | 0.1292 |

Interpretation of the above output has been left as an exercise to the reader.

The R code reads the data; computes the centered variables cenceage, cencegirth, and acenmsbp; and fits the alternative models discussed in this example. Note that the packages “nlme” and “gee” should be installed and loaded for functions “lme” and “gee”, respectively, to run.

fam <- read.table("x:/xxx/familydata.txt",header=T)
cenceage <- unlist(tapply(fam[,4], fam[,1], scale, scale=FALSE))
cencegirth <- unlist(tapply(fam[,5], fam[,1], scale, scale=FALSE))
acenmsbp <- scale(fam[,6], center = TRUE, scale = FALSE)
family <- data.frame(fam, cenceage, cencegirth, acenmsbp)

# Generalized estimating equations (GEE)
fam.gee <- gee(sbpb ~ acenmsbp+cenceage+cencegirth, familyid, data=family, family = gaussian, corstr = "exchangeable")
summary(fam.gee)

# Unconditional mean model—Mixed Model 1
fam.lme1 <- lme(fixed = sbp ~ 1, random=~1 | familyid, data = family)
summary(fam.lme1)
# Mixed model including one cluster-level covariate, cenmsbp—Mixed Model 2
fam.lme2 <- lme(fixed = sbp ~ cenmsbp, random = ~1 | familyid, data = family)
summary(fam.lme2)

# Mixed model including sib-level covariate, cengirth—Mixed Model 3
fam.lme3 <- lme(fixed = sbp ~ cenmsbp, random = ~cenage | familyid, data = family)
summary(fam.lme3)

# Mixed model including one sibling-level covariate, cengurth—*Mixed Model 4
fam.lme4 <- lme(fixed = sbp ~ cenmsbp + cenage + cengirth + cenmsbp * cenage + cenmsbp*cengirth, random = ~cenage -1 | familyid, data = family)
summary(fam.lme4)

---

Appendix

1. Linear combinations of random variables

   Let \( x = (x_1, x_2, \ldots, x_k) \) be a set of random variables such that \( E(x_i) = \mu_i \), \( V(x_i) = \sigma_i^2 \), and \( \text{Cov}(x_i, x_j) = c_{ij} \). We define a linear combination of the random variable \( x \) as

   \[
   y = \sum_{i=1}^{k} w_i x_i
   \]

   where \( w_1, w_2, \ldots, w_k \) are constants.

   \[
   E(y) = \sum_{i=1}^{k} w_i \mu_i \tag{A.1}
   \]

   and

   \[
   V(y) = \sum_{i=1}^{k} w_i^2 \sigma_i^2 + \sum_{i=1}^{k} \sum_{j=1 \atop i \neq j}^{k} w_i w_j c_{ij} \tag{A.2}
   \]

2. Consider two linear combinations \( y_s = \sum_{i=1}^{k} a_{si} x_i \) and \( y_t = \sum_{i=1}^{k} a_{ti} x_i \). Then the covariance between \( y_s \) and \( y_t \) is

   \[
   \text{Cov}(y_s, y_t) = \sum_{i=1}^{k} \sum_{j=1}^{k} a_{si} a_{tj} \text{Cov}(x_i, x_j) \tag{A.3}
   \]
3. The delta method

Let $g(.)$ be a differentiable function of $x$. Then to the first order of approximation

$$E[g(x)] = g(E(x))$$

and

$$V[g(x)] = \left(\frac{\partial g}{\partial x}\right)^2_{x=\mu} V(x) \quad (A.4)$$

In general, if $g$ is a differentiable function of $x_1, x_2, \ldots, x_k$, then to the first approximation

$$V[g(x_1, x_2, \ldots, x_k)] \approx \sum_{i=1}^{k} \left(\frac{\partial g}{\partial x_i}\right)^2_{x=\mu} V(x_i) + \sum_{i=1}^{k} \sum_{j=1}^{k} \left(\frac{\partial g}{\partial x_i}\right) \left(\frac{\partial g}{\partial x_j}\right) \text{cov}(x_i, x_j) \quad (A.5)$$

where the $\cdot$ on top of $\frac{\partial g}{\partial x_i}$ means that they are evaluated at $\mu$.

If we have two differentiable functions $g_1(x_1, \ldots, x_k)$ and $g_2(x_1, \ldots, x_k)$, then to the first order of approximation

$$\text{Cov}(g_1, g_2) = \sum_{i=1}^{k} \sum_{j=1}^{k} \left(\frac{\partial g_1}{\partial x_i}\right) \left(\frac{\partial g_2}{\partial x_j}\right) \text{Cov}(x_i, x_j) \quad (A.6)$$

**Exercises**

1.1 Suppose that we have $k$ clusters, each of size $n$, and that the model generating the data is the one-way random effects:

$$y_{ij} = \mu + b_i + e_{ij}, \quad b_i \approx N(0, \sigma^2_b), \quad e_{ij} \approx N(0, \sigma^2_e)$$

Under the same model assumptions, and given that

$$\text{var}(\hat{\sigma}^2_e) = \frac{2\sigma^4_e}{k(n-1)}$$

$$\text{var}(\hat{\sigma}^2_b) = \frac{2}{n^2} \left[ \frac{(n\sigma^2_b + \sigma^2_e)^2}{k - 1} + \frac{\sigma^4_e}{k(n-1)} \right]$$

and $\text{cov}(\hat{\sigma}^2_e, \hat{\sigma}^2_b) = -2\sigma^4_e/kn(n-1)$,

use the delta method to show that

$$\text{var}(\hat{\rho}) = \frac{2(1 - \rho)^2(1 + (n-1)\rho)^2}{kn(n-1)}$$
1.2 Under the same assumptions of exercise 1.1, it is known that

\[
\text{MSB} \approx (n\sigma^2 + \sigma^2_e) \chi^2_{k-1}
\]

\[
\text{MSW} \approx \sigma^2_e \chi^2_{k(n-1)}
\]

On defining upper and lower points of the \(F\)-distribution as \(F_u\) and \(F_l\) by

\[
\Pr[F_l \leq F^k_{k(n-1)} \leq F_u] = 1 - \alpha
\]

construct an exact \((1 - \alpha)100\%\) confidence interval on \(\rho = \sigma^2_b / (\sigma^2_b + \sigma^2_e)\).

1.3 Suppose that we have a two-arm cluster randomized clinical trial as described in this chapter. Let \(\theta = \mu_E / \mu_C\). Use the delta method to find a first-order approximation to the maximum likelihood of \(\theta\). Hence, find an approximate \((1 - \alpha)100\%\) confidence interval on \(\theta\) assuming that the number of clusters is the same in each arm.

1.4 Suppose that we have \(H\) treatment groups and \(k_h\) clusters are randomized in the \(h\)th group (\(h = 1, 2, \ldots, H\)) with \(n_{hj}\) denoting the \(j\)th cluster size within the \(h\)th group (\(j = 1, 2, \ldots, k_h\)). It is required to test the hypothesis \(H_0: \mu_1 = \mu_2 = \cdots = \mu_H\). Cochran (1937) suggested that for individual randomization, the statistic

\[
G^2_H = \sum_{h=1}^{H} w_i (\bar{y}_{h'} - \bar{y})^2
\]

has approximately \(\chi^2\) distribution with \((H - 1)\) degrees of freedom. Here \(\bar{y}_{h'}\) is the \(h\)th group mean,

\[
\bar{y} = \frac{\sum_{i=1}^{H} w_i \bar{y}_i}{\sum_{i=1}^{H} w_i}
\]

and \(w_i = \text{var}(\bar{y}_i)^{-1}\)

For the case of cluster randomization,

(i) What is \(w_i\) and \(\text{var}(\bar{y})\).

(ii) Show that \(G_H\) is equivalent to \(Z\) given in Equation 1.15. State your assumptions.

1.5 Under the one-way random effects, define the within-cluster coefficient of variation as \(\theta = \sigma_e / \mu_s\) and its maximum likelihood estimator as \(\hat{\theta} = (\text{MSW})^{1/2} / \bar{y}\). Use the delta method to derive a first-order approximation to the variance of \(\hat{\theta}\). Assume the data are normally distributed.
1.6 Run a sequence of models using SAS programs and R programs on 13 families, a subset of Miall and Oldham’s data provided below:

<table>
<thead>
<tr>
<th>FID</th>
<th>SID</th>
<th>FBP</th>
<th>MBP</th>
<th>SBP</th>
<th>AGE</th>
<th>SEX</th>
<th>ARMGIRTH</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>104.47</td>
<td>103.70</td>
<td>94.30</td>
<td>3.66</td>
<td>1</td>
<td>5.90</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>104.47</td>
<td>103.70</td>
<td>96.51</td>
<td>15.47</td>
<td>0</td>
<td>11.00</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>104.47</td>
<td>103.70</td>
<td>96.51</td>
<td>15.43</td>
<td>1</td>
<td>9.80</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>104.47</td>
<td>103.70</td>
<td>102.00</td>
<td>17.16</td>
<td>0</td>
<td>10.00</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>104.47</td>
<td>103.70</td>
<td>92.00</td>
<td>2.00</td>
<td>1</td>
<td>3.20</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>104.47</td>
<td>103.70</td>
<td>103.00</td>
<td>12.28</td>
<td>1</td>
<td>8.00</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>104.47</td>
<td>103.70</td>
<td>91.90</td>
<td>2.11</td>
<td>1</td>
<td>3.30</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>109.48</td>
<td>102.55</td>
<td>96.70</td>
<td>4.35</td>
<td>0</td>
<td>5.90</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>109.48</td>
<td>102.55</td>
<td>95.00</td>
<td>8.57</td>
<td>0</td>
<td>8.60</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>109.48</td>
<td>102.55</td>
<td>104.80</td>
<td>9.51</td>
<td>0</td>
<td>6.70</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>109.48</td>
<td>102.55</td>
<td>96.00</td>
<td>7.14</td>
<td>1</td>
<td>6.00</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>109.48</td>
<td>102.55</td>
<td>90.00</td>
<td>1.28</td>
<td>1</td>
<td>2.30</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>107.44</td>
<td>103.02</td>
<td>94.00</td>
<td>5.46</td>
<td>1</td>
<td>5.90</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>107.44</td>
<td>103.02</td>
<td>96.52</td>
<td>15.70</td>
<td>1</td>
<td>10.50</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>107.44</td>
<td>103.02</td>
<td>102.00</td>
<td>9.16</td>
<td>0</td>
<td>8.00</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>107.44</td>
<td>103.02</td>
<td>100.00</td>
<td>20.92</td>
<td>0</td>
<td>11.39</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>107.44</td>
<td>103.02</td>
<td>94.00</td>
<td>6.34</td>
<td>1</td>
<td>8.00</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>111.68</td>
<td>106.51</td>
<td>97.00</td>
<td>4.44</td>
<td>0</td>
<td>4.99</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>111.68</td>
<td>106.51</td>
<td>96.49</td>
<td>14.26</td>
<td>0</td>
<td>7.90</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>111.68</td>
<td>106.51</td>
<td>101.00</td>
<td>17.44</td>
<td>0</td>
<td>9.00</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>111.68</td>
<td>106.51</td>
<td>105.00</td>
<td>9.17</td>
<td>1</td>
<td>7.00</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>111.68</td>
<td>106.51</td>
<td>96.60</td>
<td>4.71</td>
<td>0</td>
<td>6.00</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>111.68</td>
<td>106.51</td>
<td>108.00</td>
<td>12.47</td>
<td>0</td>
<td>7.80</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>110.18</td>
<td>107.44</td>
<td>89.00</td>
<td>1.48</td>
<td>1</td>
<td>3.00</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>110.18</td>
<td>107.44</td>
<td>97.99</td>
<td>11.01</td>
<td>1</td>
<td>4.50</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>110.18</td>
<td>107.44</td>
<td>93.00</td>
<td>3.89</td>
<td>1</td>
<td>5.00</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>110.18</td>
<td>107.44</td>
<td>10.00</td>
<td>12.09</td>
<td>1</td>
<td>7.50</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>110.18</td>
<td>107.44</td>
<td>110.00</td>
<td>9.70</td>
<td>1</td>
<td>9.10</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>110.18</td>
<td>107.44</td>
<td>96.48</td>
<td>15.46</td>
<td>1</td>
<td>9.90</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>110.18</td>
<td>107.44</td>
<td>92.00</td>
<td>8.30</td>
<td>1</td>
<td>7.00</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>110.18</td>
<td>107.44</td>
<td>96.48</td>
<td>14.42</td>
<td>0</td>
<td>9.50</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>109.98</td>
<td>108.09</td>
<td>98.17</td>
<td>11.42</td>
<td>0</td>
<td>7.50</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>109.98</td>
<td>108.09</td>
<td>95.00</td>
<td>5.07</td>
<td>1</td>
<td>5.40</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>109.98</td>
<td>108.09</td>
<td>104.30</td>
<td>12.75</td>
<td>1</td>
<td>9.60</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>109.98</td>
<td>108.09</td>
<td>96.47</td>
<td>13.43</td>
<td>1</td>
<td>8.00</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>109.98</td>
<td>108.09</td>
<td>95.50</td>
<td>8.54</td>
<td>0</td>
<td>6.60</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>109.98</td>
<td>108.09</td>
<td>104.00</td>
<td>9.03</td>
<td>0</td>
<td>5.90</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>110.18</td>
<td>108.15</td>
<td>90.00</td>
<td>6.89</td>
<td>1</td>
<td>7.40</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>110.18</td>
<td>108.15</td>
<td>104.20</td>
<td>16.84</td>
<td>0</td>
<td>10.20</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>110.18</td>
<td>108.15</td>
<td>96.90</td>
<td>6.83</td>
<td>0</td>
<td>6.70</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>110.18</td>
<td>108.15</td>
<td>97.99</td>
<td>11.51</td>
<td>0</td>
<td>7.80</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>110.18</td>
<td>108.15</td>
<td>98.00</td>
<td>6.04</td>
<td>1</td>
<td>7.50</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>110.18</td>
<td>108.15</td>
<td>96.60</td>
<td>7.25</td>
<td>1</td>
<td>6.80</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>110.18</td>
<td>108.15</td>
<td>98.00</td>
<td>5.67</td>
<td>1</td>
<td>5.00</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>110.18</td>
<td>108.15</td>
<td>96.91</td>
<td>18.03</td>
<td>0</td>
<td>10.60</td>
</tr>
</tbody>
</table>

(Continued)
Analysis of Correlated Data with SAS and R

(Continued)

<table>
<thead>
<tr>
<th>FID</th>
<th>SID</th>
<th>FBP</th>
<th>MBP</th>
<th>SBP</th>
<th>AGE</th>
<th>SEX</th>
<th>ARMGIRTH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>108.69</td>
<td>103.60</td>
<td>95.90</td>
<td>7.00</td>
<td>1</td>
<td>6.60</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>108.69</td>
<td>103.60</td>
<td>101.50</td>
<td>17.35</td>
<td>1</td>
<td>11.40</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>108.69</td>
<td>103.60</td>
<td>96.00</td>
<td>8.67</td>
<td>1</td>
<td>5.50</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>108.69</td>
<td>103.60</td>
<td>96.90</td>
<td>5.39</td>
<td>0</td>
<td>5.60</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>108.69</td>
<td>103.60</td>
<td>98.60</td>
<td>8.49</td>
<td>1</td>
<td>7.50</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>108.69</td>
<td>103.60</td>
<td>96.98</td>
<td>20.70</td>
<td>1</td>
<td>11.00</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>108.69</td>
<td>103.60</td>
<td>96.94</td>
<td>11.80</td>
<td>0</td>
<td>7.50</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>113.43</td>
<td>104.33</td>
<td>96.51</td>
<td>14.03</td>
<td>0</td>
<td>9.90</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>113.43</td>
<td>104.33</td>
<td>96.51</td>
<td>15.50</td>
<td>0</td>
<td>8.00</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>113.43</td>
<td>104.33</td>
<td>96.51</td>
<td>13.92</td>
<td>1</td>
<td>7.00</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>113.43</td>
<td>104.33</td>
<td>98.27</td>
<td>11.48</td>
<td>0</td>
<td>8.00</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>113.43</td>
<td>104.33</td>
<td>97.12</td>
<td>9.85</td>
<td>1</td>
<td>7.00</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>111.52</td>
<td>108.42</td>
<td>96.47</td>
<td>15.05</td>
<td>0</td>
<td>10.50</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>111.52</td>
<td>108.42</td>
<td>97.00</td>
<td>5.75</td>
<td>1</td>
<td>7.00</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>111.52</td>
<td>108.42</td>
<td>106.50</td>
<td>9.57</td>
<td>1</td>
<td>8.00</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>111.52</td>
<td>108.42</td>
<td>97.73</td>
<td>12.06</td>
<td>0</td>
<td>9.50</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>112.62</td>
<td>103.85</td>
<td>98.19</td>
<td>18.61</td>
<td>0</td>
<td>10.00</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>112.62</td>
<td>103.85</td>
<td>96.51</td>
<td>14.45</td>
<td>1</td>
<td>9.50</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>112.62</td>
<td>103.85</td>
<td>96.98</td>
<td>10.99</td>
<td>1</td>
<td>8.50</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>107.28</td>
<td>112.91</td>
<td>96.00</td>
<td>4.54</td>
<td>0</td>
<td>5.00</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>107.28</td>
<td>112.91</td>
<td>104.00</td>
<td>12.94</td>
<td>0</td>
<td>7.90</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>111.84</td>
<td>108.56</td>
<td>96.00</td>
<td>4.00</td>
<td>0</td>
<td>5.50</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>111.84</td>
<td>108.56</td>
<td>98.31</td>
<td>10.65</td>
<td>0</td>
<td>7.00</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>111.84</td>
<td>108.56</td>
<td>97.00</td>
<td>8.53</td>
<td>0</td>
<td>7.00</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>111.84</td>
<td>108.56</td>
<td>98.31</td>
<td>10.17</td>
<td>0</td>
<td>8.00</td>
</tr>
<tr>
<td>13</td>
<td>5</td>
<td>111.84</td>
<td>108.56</td>
<td>99.00</td>
<td>5.49</td>
<td>0</td>
<td>4.99</td>
</tr>
</tbody>
</table>
References


R 2.3.1 (2006). A Language and Environment, R Development Core Team, Stanford University, California.


Weil, C.S. (1970). Selection of the valid number of sampling units and a consideration of their combination in toxicological studies involving reproduction, teratogenesis or carcinogenesis. *Food and Cosmetics Toxicology*, 8, 177–182.


