ABSTRACT

Title of dissertation: Fixed versus Mixed Parameterization in Logistic Regression Models: Application to Meta-Analysis

Chin-Fang Weng, Master of Arts, 2008

Dissertation directed by: Professor Eric V. Slud
Statistics Program
Department of Mathematics

Three methods: fixed intercept generalized model (GLM), random intercept generalized mixed model (GLMM), and conditional logistic regression (clogit) are compared in a meta-analysis of 43 studies assessing the effect of diet on cancer incidence in rats. We also perform simulation studies to assess distributional behavior of regression estimates and tests of fit. Other simulations assess the effects of model misspecification, and increasing the sample size, either by adding additional studies or by increasing the sizes of a fixed number of studies. Estimates of fixed effects seem insensitive to increasing the sample sizes, but the deviance test of fit is biased. Conditional logistic regression avoids the possibility of bias when the number of studies is very large in a GLM analysis and also avoids effects of misspecification of the random effect distribution in a GLMM analysis, but at the cost of some information loss.

Fixed versus Mixed Parameterization in Logistic Regression Models:Application to Meta-Analysis

by

Chin-Fang Weng

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Master of Arts
2008

Advisory Committee:
Professor Eric V. Slud, Chair/Advisor
Professor Benjamin Kedem
Professor Grace L. Yang

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

The problems of statistics fall into two main classes: statistical modeling and inference. The validity of most statistical analysis is conditional on the model being correct. We need some model to proceed with an analysis, but a wrong model can lead to a wrong conclusion. In this paper, our central question in formulating a statistical model is whether certain sources of variation should be analyzed as systematic (nonrandom) effects, analyzed as random variables, or eliminated from the analysis via conditioning.

In meta-analysis this question arises when one attempts to combine data from many nonidentical studies having a common parameter of research interest, among which study to study differences must be taken into account. A real world example, which we analyze in detail later in this thesis, is described below.

Freedman (1994) [8] performed a meta-analysis of a collection of studies, each of which was meant to relate tumor incidence in rats to dietary fat and/or total caloric intake. The studies were similar in design. Female Sprague-Dawley rats were randomly assigned to various diet groups differing in fat content and in some studies also differing in total caloric intake. At a certain point all rats were dosed with a known carcinogen. At the termination of the studies the rats were sacrificed

and the number of rats which developed breast tumors was determined. There were variations in the conduct of the studies, however, so it was felt that the data could not be combined without accounting for study to study differences that might be statistically significant.

The data are hierarchically structured: there were 43 studies with 2 to 4 groups in each study and 20 to 40 rats for each group. Most of the original studies were intended to study only dietary fat, and total caloric intake was not always varied and in fact not always recorded. Freedman used a generalized linear model (GLM) with 43 nuisance parameters to describe the data. Maximum likelihood in models with many nuisance parameters may lead to inconsistent estimators, meaning that inferences about diet may be invalid under this data structure.

Fay et al. (1998)[10] used conditional logistic regression to analyze the same data and developed a statistic to adjust for small sample size problems and possible model misspecification. This raises other questions: How much information is lost by using conditional logistic regression? What if a generalized mixed model (GLMM) is used? How do the three methods of analysis compare to each other? Although our ultimate goal is the efficiency and consistency of models, there are some technical difficulties to overcome, such as missing values in the data set, limitations of available software, and combining data to make inferences which the original studies were never intended to address.

We reanalyzed Freedman's data using fixed effect logistic regression, random intercept logistic regression and conditional logistic regression. We found that Freedman's model could be improved by adding a quadratic term and that the three meth-

ods produced nearly identical estimates of the regression coefficients. The study to study differences can be modeled as either fixed effects or random effects without affecting the estimates of the regression coefficients, which are the only parameters of interest.

The data structure has many nuisance parameters (the study effects) and small to moderate sample sizes in each of the 104 treatment groups. It is known that with many nuisance parameters and small group sizes maximum likelihood estimators may be inconsistent. On the other hand, the data involve observations on 2844 individual rats. Can one use large sample theory in this situation? We investigate the accuracy of various large sample approximations using simulation studies, varying sample sizes and numbers of studies to assess the performance of various estimation schemes and tests of goodness of fit and to compare the results of using random effects instead of the fixed effects modeling employed by Freedman. We also use simulation to examine robustness of distributional assumptions for random effects.

The major conclusion which we draw from these investigations is that for data structured like Freedman's (many studies and small to moderate sample sizes in each study) conditional regression seems to be the analysis of choice. It avoids the need to impose assumptions about the nuisance parameters and produces consistent estimators when the number of nuisance parameters grows large. However, it involves a loss of some information (5–10% in our simulations) so caution is needed.

1.2 Organization of the Thesis

Chapter 2 reviews selected literature on meta-analysis, generalized linear models, generalized linear mixed models, and conditional logistic regression. In Chapter 3 we analyze the rat data using three methods: GLM, GLMM, and conditional logistic regression. We formulate research questions, some of which come from the literature review and some from the rat data analyses. Chapter 4 presents simulation results which address the research questions of Chapter 3. In Chapter 5, we summarize the results qualitatively and suggest topics for future research.

Chapter 2

Literature Review

2.1 Meta-Analysis

The term *meta-analysis* was first coined in 1976 by the psychologist Glass [11], and since then there has been a surge of interest in quantitative methods for summarizing results from a series of related studies in fields such as education, psychology and the biomedical sciences.

In many different data analytic contexts, when all the studies are designed in a similar way and the measure of outcome is similar, a combined estimation can be conducted. Although the studies may have some minor differences, one may prefer to ignore those differences and combine the results anyway. When the combination is valid, the result of combining many studies usually provide more accurate estimators than does any individual study. However, when study to study differences are large, ignoring these differences and pooling the data can lead to biased estimates and invalid inferences. Meta-analysis permits one to combine data from many studies, taking account of study differences while producing valid inferences on parameters of interest which affect all studies. Meta-analysis is used in general data structures: linear models, analysis of variance models and generalized linear models for categorical or count data.

As in the analysis of variance, in meta-analysis there are three related con-

ceptualizations of statistical models: fixed, random, and mixed effects models [12]. In a fixed-effect model, we assume no "between-study" heterogeneity of treatment effects. The true values of treatment effects are fixed but unknown constants. One tries to parameterize essentially all of the variation in the study results, including study to study differences. Only the "within-study" sampling variability is assumed to be random.

In random-effect models, we rely on the assumption that the study effects and treatment effects are normally distributed or follow some other specified distribution across studies. This assumption is typically unverifiable except in certain large sample situations. Between-study variation in treatment effects is described by estimating the variance component for a treatment by study interaction term. Finally, the mixed-effect model involves the combination of fixed and random effect models. Treatment effects are regarded as fixed and study to study variation is regarded as random.

An example of meta-analysis for binary outcome with no covariate is the heart attack data of Yusuf et al. [21]. The data structure is a $2 \times 2 \times K$ table in which the third dimension consists of K levels of a confounding factor "site" in a multisite study. In site k there are n_{0k} subjects in a control group and n_{1k} in a treatment group, with y_{0k} and y_{1k} observed deaths in the control and treatment groups, respectively. The responses y_{0k} and y_{1k} are assumed to follow binomial distributions with probabilities of deaths p_{0k} and p_{1k} and numbers of trials n_{0k} and n_{1k}, respectively. The estimand of common interest across all K tables is, for example, the *risk ratio*, p_{1k}/p_{0k}, which may be assumed the same for all k. One way to analyze this kind

of stratified table is to compute a Mantel-Haenszel chi-square. This technique is sometimes referred to as meta-analysis.

Because of the way meta-analysis is conducted, often only summary data are available and individual unit (e.g. patient) level data are missing. These summary data come from many studies, in which each study forms a cluster. The groups within each cluster are likely to be more similar to one another, due to shared environmental conditions, than they are to groups from other clusters. Here, the effects describing individual studies (clusters) are nuisance parameters, that is, parameters that are present in a model but are not of inferential interest. Consider the following analysis of covariance example.

The ANCOVA model is

$$E[y_{ij}] = \mu + \alpha_i + \beta x_{ij}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, n_i.$$

In ANCOVA, usually we are interested in the α_i, $i = 1, \ldots, I$, which are interpreted as group or treatment effects, while β is a nuisance parameter. However, in meta-analysis the study effect α_i is a nuisance parameter and β is the parameter of interest. What if we ignore the study effect α_i and directly combine the data to perform meta-analysis? Such an approach leads to a misspecified model:

$$E[y_{ij}] = \mu + \beta x_{ij}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, n_i.$$

In this misspecified model the least squares estimate of β is

$$\hat{\beta} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}) y_{ij}}{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2},$$

where $\bar{x} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} x_{ij} / \sum_{i=1}^{I} n_i$. This estimator is biased. Under the correct

model, the slope β measures the relationship between y and x after adjusting for study to study differences. The correct estimator is

$$\hat{\beta} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i) y_{ij}}{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2},$$

where $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$. The difference between the estimators is illustrated in Figure 2.1.

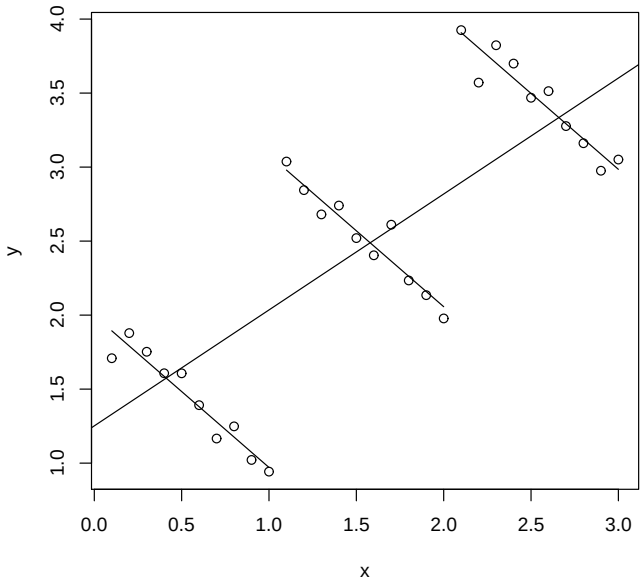

Figure 2.1: *Effect of misspecifying ANCOVA model. Lines with negative slope indicate correct least squares estimate of β, while the line with positive slope is based on misspecified estimate.*

Note that in meta-analysis the sample size, $N = \sum_{i=1}^{I} n_i$, increases as well as the number of studies (nuisance parameters). The large number of nuisance parameters can cause difficulties with the fitting process and with the properties of ordinary ML estimators [4]. Is there a better way to overcome this problem?

2.1.1 Approaches to Meta-Analysis

In many models $Y_{ij},\ j = 1, ..., n_i$, the outcomes in cluster i are assumed to be independent. In meta-analysis, without fixed effects parameters for cluster differences, this assumption is implausible. The use of generalized linear mixed models for non-normal data can incorporate the correlation via random effects. There are two approaches to modeling such data.

One approach is through *generalized estimating equations* (GEE) [7]. It involves dropping the usual assumption of independence between the outcomes Y_{ij}, $j = 1, ..., n_i$, and modeling the correlation structure explicitly. Usually the correlation parameters are not of particular interest (i.e., they are nuisance parameters) but they need to be included in the model in order to obtain consistent estimates of those parameters that are of interest and to correctly calculate the standard errors of these estimates. The correlation specification does not have to be completely correct, but efficiency is better if it is [13].

The alternative approach to modeling clustered data is based on considering the hierarchical structure of the study. It is called multilevel modeling. On each branch, outcomes at the same level are assumed to be conditionally independent given cluster effect and the correlation is a result of shared branch effects within the multilevel structure. For multilevel models, the effects of levels may be described by fixed parameters or random effects or both. In general, the linear mixed model for normally distributed responses can be written in the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{2.1}$$

where $\boldsymbol{\beta}$ are the fixed effects, $\mathbf{u}$ is a vector of random effects and $\mathbf{e}$ $N(\mathbf{0}, \sigma^2\mathbf{I})$). The matrices $\mathbf{X}$ and $\mathbf{Z}$ are design matrices. In (2.1) the random effect enters the model as a predictor and reflects heterogeneity caused by omitting certain explanatory variables or random measurement error in the explanatory variables.

For multilevel data, if the primary goal is to estimate the random effects, the generalized linear model could be conceived as a Bayesian model with a parameter vector $\mathbf{u}$ having some prior distribution, for example $N(\mathbf{0}, \mathbf{D})$. Then Bayesian methods using Markov Chain Monte Carlo or Gibbs Sampling, as implemented in the software BUGS, might be more appropriate to estimate u than the frequentist approach adopted here [5].

2.2 General Form of Model

First, we start with a more general form of the generalized linear mixed model (GLMM):

$$f_{Y_{ij}|\mathbf{u}}(y_{ij}|\mathbf{u}) = \exp\left[(y_{ij}\eta_{ij} - b(\eta_{ij}))/\tau^2 - c(y_{ij}, \tau)\right] \tag{2.2}$$

for $i = 1, \ldots, I$, $j = 1, \ldots, n_i$ and $\mathbf{u}' = [\mathbf{u}_1' \ldots \mathbf{u}_I']$. Assume

$$\begin{aligned} Y_{ij}|\mathbf{u} &\sim f_{y_{ij}|u_i}(y_{ij}|\mathbf{u}_i) \text{ independently,} \\ E[Y_{ij}|\mathbf{u}] &= \mu_{ij}, \\ g(\mu_{ij}) &= \eta_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{u}_i, \\ \mathbf{u} &\sim f_{\mathbf{u}}(\mathbf{u}). \end{aligned}$$

The so-called dispersion parameter τ is often known to be 1, as in binomial or Poisson models. We can see that the GLMM model has the following features:

1. The distribution of Y_{ij} is from an exponential family, conditionally given the random effect $\mathbf{u}$.

2. The link function, $g(.)$, is applied to the conditional mean of Y_{ij} given $\mathbf{u}$ to obtain the conditional linear predictors.

3. The linear predictor η_{ij} is assumed to consist of two components, the fixed effect portion, $\mathbf{x_{ij}}'\boldsymbol{\beta}$, and the random effect portion, $\mathbf{z_{ij}}'\mathbf{u}_i$, and a distribution is assigned to $\mathbf{u}$.

We next describe the generalized linear mixed effect model for our particular count data: Y_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J_i$. Suppose that given a vector of random effects $\mathbf{u}$, the $Y_{ij}|\mathbf{u}$ are conditionally independent and have binomial distributions with parameters (n_{ij}, π_{ij}). Assume

$$\operatorname{logit} \pi_{ij} = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \eta_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{u} \tag{2.3}$$

where

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_{11} \\ \vdots \\ \mathbf{x}_{I,J_I} \end{bmatrix}, \qquad \mathbf{Z} = \begin{bmatrix} \mathbf{z}_{11} \\ \vdots \\ \mathbf{z}_{I,J_I} \end{bmatrix},$$

$\mathbf{X}$ is the model matrix for the fixed effects, $\boldsymbol{\beta}$ is the vector of fixed effect parameters, the parameters of interest, $\mathbf{Z}$ is the model matrix for the random effects, and $\mathbf{u}$ is the vector of random effects. It is assumed that $\mathbf{u}$ has zero mean vector and

covariance matrix $\mathbf{D}$. One might specify in addition that $\mathbf{u}$ has a multivariate normal distribution. Because the Y_{ij} are binomially distributed, this model is a mixed effect logistic regression model. In the following subsections we state specific models which are special cases of the general model above.

2.3 Fixed Intercept GLM

This model has no random effects $\mathbf{u}$, no model matrix for random effects, and in fact is an ordinary GLM. The linear predictor for this GLM has the form

$$g(\mu_{ij}) = \mathbf{x_{ij}}'\boldsymbol{\beta}. \tag{2.4}$$

or

$$\text{logit}(\pi_{ij}) = \alpha_i + \beta_1 TFA_{ij} + \beta_2 (TFA_{ij})^2 + \beta_3 KCA_{ij}. \tag{2.5}$$

where

n_{ij} is the number of rats in SET i, group j, $i = 1, 2, \ldots, 43$,

Y_{ij} is the number of tumors in SET i, group j, $j = 1, \ldots, J_i$, $i = 1, 2, \ldots, 43$,

$Y_{ij} \sim Binomial(n_{ij}, \pi_{ij})$,

$\boldsymbol{\beta} = [\alpha_1, \ldots, \alpha_{43}, \beta_1, \beta_2, \beta_3]'$.

Assume the SET effects, α_i, are unknown but fixed. The SET parameters, α_i, $i = 1, 2, \ldots, 43$, are treated as fixed intercepts. Also, the SET parameters are nuisance parameters, while $\beta_1, \beta_2, \beta_3$ are fixed parameters of interest.

The likelihood for this fixed effect GLM is

$$L = \prod_{i=1}^{43} \prod_{j=1}^{J_i} \binom{n_{ij}}{y_{ij}} \frac{\exp[(\alpha_i + \beta_1 TFA_{ij} + \beta_2 (TFA_{ij})^2 + \beta_3 KCA_{ij}) y_{ij}]}{[1 + \exp(\alpha_i + \beta_1 TFA_{ij} + \beta_2 (TFA_{ij})^2 + \beta_3 KCA_{ij})]^{n_{ij}}}.$$

Write

$$\eta_{ij} = \alpha_i + \beta_1 TFA_{ij} + \beta_2 (TFA_{ij})^2 + \beta_3 KCA_{ij}.$$

The log-likelihood is

$$\log L = \log(const) + \sum\sum y_{ij}\eta_{ij} - \sum\sum n_{ij} \log[1 + \exp(\eta_{ij})]. \tag{2.6}$$

The goodness of fit of a GLM is often tested using the *deviance* statistic. If $L(\boldsymbol{\mu})$ represents the likelihood for a GLM parameterized in terms of its mean vector $\boldsymbol{\mu}$, the deviance is defined as

$$-2[\log L(\boldsymbol{\mu}) - \log L(\mathbf{Y})]$$

Here $L(\mathbf{Y})$ is the likelihood function of the *saturated model*, under which the MLE of μ_{ij} is Y_{ij}. If the number of groups is fixed and the sample size in each group goes to infinity, the deviance has a limiting χ^2 distribution [1] with degrees of freedom equal to the difference in the number of parameters in the saturated model and model of interest. Large deviance values indicate that the model does not fit the data. These distributional results do not hold if the number of groups increases with the sample size. This topic is discussed further in Section 4.2.1.

2.3.1 Many Nuisance Parameters and Inconsistent MLE

In this model, the SET parameters, α_i, $i = 1, 2, \ldots, 43$, are treated as fixed intercepts. If the total sample size $N = \sum_i \sum_j n_{ij}$ is increased by adding more SETs, the number of SET parameters α_i, $i = 1, \ldots, I$, also increases. Even though the number of regression coefficients remain the same, the ordinary ML estimators of $\beta_1, \beta_2, \beta_3$

may not be consistent. Asymptotic optimality properties of ML estimators, such as consistency, require the number of parameters to be fixed or to increase slowly as N increases.

In classical estimation problems, bias is small relative to standard error, and the bias goes away as the sample gets large. That is typically the case when the number of nuisance parameters is small relative to the sample size. There is a genuine concern, however, when bias does not disappear as the sample size gets large, or when bias is large relative to standard error, resulting in inconsistent estimation.

This happens in many models when the number of parameters has an order similar to that of the number of clusters. For example, consider the logistic matched-pairs model. The data form two dependent binomial samples. Cluster i consists of the responses (y_{i1}, y_{i2}) for matched pair i. Observation t in Cluster i has $y_{it} = 1$ (a success) or 0 (a failure), $t = 1, 2$. These data form a 2x2 table. Now an extension of the logit matched-pairs model allows $T > 2$ observations in each cluster. ML estimators of β_t have an approximate asymptotic bias of order $T/(T-1)$ (Andersen 1980, pp. 244-245) [4]. Neyman and Scott (1948) demonstrated that maximum likelihood estimators can be severely biased even as the sample size gets large. This is a common occurrence if there are "infinitely many" nuisance parameters.

For meta-analysis, the many nuisance parameters problem seems inevitable because of the way data are collected. Nevertheless, the GLM is maximized numerically by the function `glm( )` in the software package R [16].

2.4 Random Intercept GLMM

2.4.1 Randomness versus Determinism

The question of formulating a model for patterns of variation is not central in some contexts. For example, in normal theory balanced randomized block designs, it is unimportant whether or not block effects are regarded as random. In other contexts, effects of some direct interest should be represented as random variables only as a "last resort;" for example, an interaction between treatment effects and intrinsic factors of interest (e.g.,"centers") should be taken as random only if they cannot be "explained" in some way. When there is a large number of parameters of secondary interest representing similar effects in an unbalanced design, it will often be good to consider representing them by random variables. This is partly because the occurrence of a large number of nuisance parameters means that unmodified maximum likelihood methods may be inappropriate and partly because higher precision may be achieved by a represention in terms of random variable with a well-behaved distribution [6]. In order to reduce the many nuisance parameters in fixed intercept GLM, a random intercept Mixed GLMM is used.

2.4.2 Random Intercept Mixed Model

Conditional on $\mathbf{u}$, a GLMM resembles an ordinary GLM. Let $E[y_{ij}|\mathbf{u}] = \mu_{ij}$. The *conditional linear predictor* for a GLMM has the form

$$g(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_i. \tag{2.7}$$

for link function $g(\cdot)$. The random effect vector $\mathbf{u}$ is often assumed to have a multivariate normal distribution $N(\mathbf{0}, \mathbf{D})$. In our case, $\mathbf{z}_{ij} = 1$, for $j = 1, \ldots, n_i$, so this simplified $\mathbf{u}_i$ has a univariate normal distribution. The model is called a *random intercept model*. The random intercept GLMM can be also written

$$logit(\pi_{ij}) = \mu + a_i + \beta_1 TFA_{ij} + \beta_2 (TFA_{ij})^2 + \beta_3 KCA_{ij} \tag{2.8}$$

where $\mu + a_i$, $i = 1, 2, ..., 43$, are treated as random intercepts. Here μ is an unknown parameter and the a_i are i.i.d. $N(0, \sigma_a^2)$. Groups in the same set share the same value of a_i but different sets have different values of a_i. These values are unobserved and treated as independent random variables.

This is an alternative way of handling the large number of nuisance parameters in the logistic model for the rat data. It eliminates a_i by averaging with respect to their distribution, yielding a marginal distribution. This mixed effect model only involves 5 parameters. It is more manageable, in comparison to the previous fixed effect model, which had 46 parameters.

Let $\eta_{ij} = \beta_1 TFA_{ij} + \beta_2 (TFA_{ij})^2 + \beta_3 KCA_{ij}$. Then the likelihood for the mixed GLM is

$$L = \prod_{i=1}^{43} \int_{-\infty}^{\infty} \prod_{j=1}^{J_i} \binom{n_{ij}}{y_{ij}} \left(\frac{\exp[(\mu + a_i + \eta_{ij}) y_{ij}]}{[1 + \exp(\mu + a_i + \eta_{ij})]^{n_{ij}}} \right) \frac{1}{\sigma_a \sqrt{2\pi}} \exp\left(-\frac{a_i^2}{2\sigma_a^2} \right) da_i \tag{2.9}$$

Usually, the main focus in using a GLMM is inference about the fixed effects. The random effects part of the model is a mechanism for representing how positive correlation occurs between observation within a cluster. Parameters pertaining to the random effects may themselves be of interest, however. For instance, the estimate $\hat{\sigma}$ of the standard deviation of a random intercept may be a useful summary

of the degree of heterogeneity of the population. When $\sigma = 0$, the GLMM simplifies to the ordinary model treating all observations as independent.

2.4.3 Inference in GLMM

How might one go about fitting a model like (2.7)? In general, evaluation of the likelihood can be quite difficult, because the integration is of a dimension equal to the dimension of $\mathbf{u}_i$. In our case, the random intercepts reduce to a product of one-dimensional integrals and hence can be evaluated numerically. Inference using ML would proceed using the usual asymptotic approximations:

- ML estimates are asymptotically normal, with standard error estimates coming from second derivatives of the log likelihood.
- Tests for fixed effect parameters would be based on the likelihood ratio test, comparing twice the negative of the log likelihood for nested models. Alternatively, Wald tests could be formed.
- Tests of whether variances of random effects are zero can be based on the likelihood ratio statistic.

This likelihood is maximized numerically over $\vartheta = (\mu, \beta_1, \beta_2, \beta_3, \sigma_a^2)$ by D. Bates' R function `glmer()`. These numerical calculations were also checked using Slud's software [18].

2.5 Conditional Logistic Regression

ML estimators of logistic model parameters work best when the sample size N is large compared to the number of parameters. If the number of parameters increases in proportion to the data accumulated as in the GLM described above, consistency, normality, or optimality might not hold. It is our objective to seek a modified likelihood function that depends on as few of the nuisance parameters as possible while sacrificing as little information as possible. Conditional likelihood is one of the approaches in solving nuisance parameter problems like this, when sufficient statistics for cluster-level parameter are available.

As in the fixed GLM model, the model for the conditional approach is:

$$\text{logit}(\pi_{ij}) = \alpha_i + \beta_1 TFA_{ij} + \beta_2 (TFA_{ij})^2 + \beta_3 KCA_{ij} \tag{2.10}$$

Again, there are n_{ij} rats in SET i, group j, $i = 1, 2, \ldots, 43$, $j = 1, \ldots, J_i$ and Y_{ij} observed tumors in SET i, group$j$, with assumed distribution $Y_{ij} \sim Binomial(n_{ij}, \pi_{ij})$, and fixed but unknown SET effects,α_i.

The conditional likelihood approach eliminates nuisance parameters from the likelihood by conditioning on their sufficient statistics. In the present problem, the nuisance parameters are the intercepts in each set, $\mu + \alpha_i$. The sufficient statistics are $S_i = S_i(Y_{i1}, \ldots, Y_{iJ_i}) = \sum_j Y_{ij}$, the total number of rats with tumors in set i. We also can start from the GLMM and then go on to the conditional approach. In GLMM, we assume that the SET effects follow a specified distribution, while in the conditional approach, we treat the SET effects as fixed unknown parameters. The

conditional likelihood for conditional logistic regression is:

$$L_c = \prod_{i=1}^{43} \prod_{j=1}^{J_i} \left[\frac{\binom{n_{ij}}{Y_{ij}} \exp[(\beta_1 TFA_{ij} + \beta_2 (TFA_{ij})^2 + \beta_3 KCA_{ij}) Y_{ij}]}{\sum_{\mathbf{z}_i} \binom{n_{ij}}{z_{ij}} \exp[(\beta_1 TFA_{ij} + \beta_2 (TFA_{ij})^2 + \beta_3 KCA_{ij}) z_{ij}]} \right] \tag{2.11}$$

where the sum in the denominator ranges over all vectors $\mathbf{z}_i = [z_{i1}, \ldots, z_{i,J_i}]'$ such that $\sum_j z_{ij} = S_i = \sum_j Y_{ij}$.

A conditional likelihood is used just like an ordinary likelihood. The conditional ML estimates of parameters contained in L_c are the parameter values maximizing L_c. Calculating using iterative methods, the estimators are asymptotically normal with covariance matrix equal to the negative inverse of the matrix of second partial derivatives of the conditional log likelihood. For the rat data we estimate $\beta_1, \beta_2, \beta_3$ by maximizing L_c numerically by means of Fay's software [10] in R.

2.5.1 Random Effect versus Conditional ML Approaches

Compared with the random effects approach, the conditional ML approach has certain advantages. As $I \to \infty$, the conditonal MLE's are consistent [3]. One does not have to assume a parametric distribution for u_i. It is difficult to check this assumption in the random effect approach. Conditional ML is also appropriate with retrospective sampling. In that case, bias can occur with a random effect approach because clusters are not randomly sampled [14].

However, the conditional ML approach has several disadvantages. It is restricted to the canonical link (the logit), for which reduced sufficient statistics exist for u_i. More important, it is restricted to inference about within-cluster fixed effects. The conditioning removes the source of variability needed for estimating between-

cluster effects. Also, this approach does not provide information about u_i. Finally, in more general models with covariates, conditional ML can be less efficient that the random effects approach for estimating the fixed effects. Neuhaus and Esperance (1996) [15] note that conditional ML may lose considerable efficiency when sample sizes are small and covariates have strong within cluster correlation. As this correlation approaches +1, the covariate effect resembles a between cluster effect, which conditional MLE can not estimate.

Chapter 3

Analysis of the Rat Data

3.1 Data example - The rat data

The data used in this paper are taken from a large data set, which was originally collected from many separate studies and analyzed by Freedman et al. (1990) [8], and then reanalyzed by Freedman (1994) [9] and by Fay et al. (1998) [10]. In this data example of meta-analysis, statistical models are used to answer the biological question: What are the effects on mammary tumor development of increasing energy intake (KCA) and fat intake (TFA), respectively?

The experimental data were drawn from public articles using MEDLINE, covering the years 1966–1987. There were 104 groups consisting of 2844 Sprague-Dawley rats altogether in the 43 sets (clusters/experiments/studies). There were typically 20–40 rats in each group and 2–4 groups per set. Experiments received essentially the same treatment and shared similar environments except for diets "controlled" to be different.

Freedman et al. recorded the study indicator (SET), the number of animals in the group (N), the number of animals developing mammary tumors (NTUM), the percent of calories from fat (PCF) and the total calories consumed (KCA) for each of the 104 groups of animals in the data. The variable KCA was often not reported in the original articles. In such a case Freedman et al. imputed the missing KCA

variable as 52.5 kCal. The justification was that rats typically self-regulate their diets and eat essentially a constant amount of total calories.

3.1.1 Data Structure

Let Z_{ijk}, $i = 1, \ldots, 43$, $j = 1, \ldots, J_i$, $k = 1, \ldots, n_{ij}$, be the response indicating whether rat k in group j of set i does or does not develop tumors. Here J_i is the number of groups in set i, and n_{ij} is the number of rats in group j of set i. So, Z_{ijk} is 1 if a rat develops tumor and 0 otherwise. For group j in set i, the Z_{ijk} $k = 1, \ldots, n_{ij}$ are assumed to be independent Bernoulli variables with parameter π_{ij}. Let $Y_{ij} = \sum_{k=1}^{n_{ij}} Z_k \sim$ Binomial(n_{ij}, π_{ij}), $j = 1, \ldots, J_i$, $i = 1, \ldots, 43$. Then Y_{ij} is the outcome variable, the number of rats developing tumors in each group. So, the effective sample size for these data is $N = 104$. We assume Y_{ij}, $j = 1, \ldots, J_i$, are independent and binomially distributed, given set i, with success probability π_{ij}. Thus, the set identifier is treated as a grouping variable and also as a factor covariate. The variable "SET" is really a multicolumn dummy covariate. Besides "SET", there are two other covariates: proportion of dietary fat (PCF) and total calories (KCA).

3.1.2 Data issues

The formula TFA = KCA x PCF/100 shows the relation between fat and calories (where PCF is the average percent calories from fat). The KCA values in 79 of the groups were imputed by Freedman et al. [8]. Based on the belief that most

rats will self-regulate their intake to eat a constant amount of the diet, the missing KCA values were imputed as the constant 52.5 kCal. Freedman's paper [9] shows that there is some discrepancy between analysis with and without inclusion of the experiments for which KCA was missing. I created an indicator variable (MISSING-KCA) to investigate the effect of the missing variable. This MISSING-KCA turned out to be not significant in the fixed effect model.

In an attempt to reduce the total number of SETs, a cluster analysis was used to combine 43 SETs into 7 clusters. However, adding this cluster indicator to the fixed effect model had no significant effect. Similarly, using Group ID and Cluster indicator in place of SET had no significant effect on the model fit.

3.2 Fixed effect models

Notation:

TFA2 = calories from fat,

KCA2 = total calories,

SET = factor indicating set or experiment.

Three models for logit(π_{ij})

Model A : SET + TFA2 + KCA2

Model B : SET + TFA2 + KCA2 + $(\text{TFA2})^2$

Model C : SET + TFA2 + KCA2 + TFA2 * SET

Model A is simplest and was originally used by Freedman [8]. Model B fits better than Model A, but is less intuitive because of the square term of TFA2. Model

C says that the slope of TFA2 depends on SET, so no meta-analysis is possible. This could be due to inadequate specification of model (e.g., relevant explanatory variables have been omitted or the link function is incorrect) or to a more complex structure. (A more complicated model, SET + TFA2 + KCA2 + TFA2*SET + KCA2*SET, was examined, but this model is overparameterized because KCA2 does not vary in most SETs, leading to confounding of SET and KCA2*SET. The GLM fit of this model leads to singularities.)

From Table 3.1, comparing the deviance statistics to the nominal χ^2 critical values would suggest that Model A and Model B do not fit the data, but Model C does fit, contradicting scientific belief. However, the theory behind the deviance test of fit is based on asymptotics for fixed I and all $n_{ij} \to \infty$. The applicability of these asymptotics is questionable here.

Table 3.1: *Comparison for models with different explanatory variables*

Model	Residual Deviance	Degrees of Freedom	Nominal p-value
Model A	91.924	59	0.003
Model B	81.241	58	0.024
Model C	22.621	16	0.125

We also tried a more general family of link functions, which is

$$g(\pi, \alpha) = \log\left[\frac{(1-\pi)^{-\alpha} - 1}{\alpha}\right]. \tag{3.1}$$

If $\alpha = 1$ then $g(\pi) = \log[\pi/(1-\pi)]$, the logit link. As $\alpha \to 0$, then $g(\pi) \to$

$\log[-\log(1-\pi)]$, the complementary log-log link. Unlike the logistic transformation, the response function of the complementary log-log is not symmetric about $\pi = 0.5$. Deviances based on the complementary log-log link function were larger than those for the logit link and the same predictors.

Further, we look at the regression coefficients and their standard errors. From Table 3.2, the regression coefficients including the SETs coefficients in Model A and Model B are significant, but a lot of large estimated standard errors appeared in Model C. This may indicate near singularity of Model C and raises doubts about the validity of the computation.

Also, we look at the dispersion parameter, ϕ, so that $\text{Var}(Y_{ij}) = n_{ij}\pi_{ij}(1-\pi_{ij})\phi$. The default for the dispersion parameter is 1 for the model fitting in software R procedure `glm()`. By using "quasibinomial", allowing a dispersion parameter to be fitted from the data, the fitted dispersion parameters are 1.577 and 1.300 respectively for Models A and B, which indicates there is dependence among observations. Combining the information above, we choose Model B for further investigation.

Table 3.2: *Comparison for regression coefficients and standard errors under three different models with dispersion fixed at 1*

	Model A (s.e.)	Model B(s.e.)	Model C (s.e.)
TFA2	0.081 (0.008)	0.172 (0.028)	0.889 (3.28e+02)
KCA2	0.125 (0.023)	0.126 (0.018)	0.105 (3.09e-02)
$(\text{TFA2})^2$	– (–)	-0.003 (0.001)	– (–)

3.3 Deviance issues

Let Y_{ij} be the outcome variable, the number of rats developing tumors in group j within set i and let n_{ij} be the number of rats in the group. In the rat data, there are 4 groups having 100 % cancer rate. We refit Model B on the data obtained after removing the 4 SETs (out of 43) containing those groups, which left 96 groups (out of the original 104), and compare the data analysis to the previous based on the full data set. From Table 3.3, we see that the parameter estimates are almost equal, but the residual deviance of the reduced data is considerably smaller. The square term, $(\text{TFA2})^2$, is still needed. This analysis raises a question of whether the $y_{ij} = n_{ij}$ is a serious issue for the model fit.

Table 3.3: *Fixed model analysis of full data (43 SETs) and reduced data (39 SETs)*

	Reduced	Full
TFA2	0.171 (0.029)	0.172 (0.028)
KCA2	0.125 (0.018)	0.126 (0.018)
$(\text{TFA2})^2$	-0.004 (0.001)	-0.003 (0.001)
Residual Deviance	63.991	81.241
d.f.	54	58
p-value	0.166	0.024

3.4 Mixed Effect and Conditional Analyses

We fit Model B with random SET effect, using Bates' glmer() function in R software. In order to justify the use of random effect model, the SET effects in the fixed effect model and the random effect model were tested by using the Shapiro-Wilk normality test, and both appear normal. The mixed effect model point estimates of regression coefficients are very close to those from the fixed effect model, as shown in Table 3.4. Bates' `glmer()` function does not provide a test for significance of the random SET effect, so we will conduct a simulation to test for the random SET effect in the simulation study described in the next Chapter. Neither Bates' `glmer()` function nor the conditional logistic regression software of Fay provides a test of model fit.

Table 3.4: *Mixed model analysis of full data (43 SETs) and reduced data (39 SETs)*

	Reduced	Full
Intercept	-6.840 (0.920)	-6.897 (0.928)
TFA2	0.168 (0.027)	0.169 (0.027)
KCA2	0.122 (0.017)	0.125 (0.018)
$(\text{TFA2})^2$	-0.004 (0.001)	-0.004 (0.001)
Var(SET)	0.829 (—-)	0.912(—-)
Residual Deviance	197.2	228.1
Log likelihood (from Eq. 2.9)	-98.59	-114.1
d.f.	91	99

Finally, we use Fay's software for conditional logistic regression. Results are presented in Table 3.5. The coefficient estimates agree well with those of the full likelihood methods assuming either fixed or random SET effects.

This analysis is equally valid for either fixed effect or random effect models, but it may lose information about the regression coefficients as a result of looking only at the conditional distribution of the data. The efficiency of the conditional logistic analysis will be examined in Chapter 4.

Table 3.5: *Conditional logistic regression analysis of full data (43 SETs) and reduced data (39 SETs)*

	Reduced	Full
TFA2	0.168 (0.026)	0.172 (0.028)
KCA2	0.123 (0.018)	0.126 (0.018)
$(\text{TFA2})^2$	-0.004 (0.001)	-0.003 (0.001)
Log likelihood (Eq. 2.11)	-1386.364	-1434.096

3.5 Research Questions

1. How well does asymptotic theory describe the behavior of estimates with the rat data structure and low dimensional parameters?

2. What is the relative efficiency of conditional logistic regression compared to maximum likelihood in the generalized linear mixed model? How does the

relative efficiency depend on the group sizes n_{ij}, the number of SETs, or the true tumor probabilities π_{ij}?

3. How well does asymptotic theory describe the behavior of estimates in the fixed effect model with many nuisance parameters? Does the classical theory apply as the number of SETs gets large? Does the validity of the theory depend on the true tumor probabilities π_{ij}?

4. What is the relative efficiency of conditional logistic regression compared to maximum likelihood in the fixed effect model with many nuisance parameters? How does the relative efficiency depend on the group sizes n_{ij}, the number of SETs, or the true tumor probabilities π_{ij}?

5. Is the analysis of the mixed model robust against departures of the random effect distribution from normality?

We address these questions by simulation studies in Chapter 4.

Chapter 4

Simulation Studies

Each simulation study in this chapter varies one or several of the following factors.

Group size n_{ij}: The original n_{ij}, satisfying $20 \leq n_{ij} \leq 40$ rats for each group, or $n_{ij} \times 10$.

Number of SETs: 43 SETs or 86 SETs.

Tumor rates for each group, π_{ij}: The original π_{ij}, satisfying $0.07 \leq \pi_{ij} \leq 0.98$, or adjusted π_{ij}, satisfying $0.10 \leq \pi_{ij} \leq 0.90$. The adjusted tumor rates are $\exp(\eta_{ij}/2-1)/(1+\exp(\eta_{ij}/2-1))$, where η_{ij} are the original linear predictors.

Method of analysis: Fixed-intercept GLM, random-intercept GLMM, or conditional logistic regression.

Random effect distribution: $N(0, \sigma^2)$, scaled Student t with 4 d.f., shifted and scaled gamma(4,1) and uniform$[-\sigma\sqrt{3}, \sigma\sqrt{3}]$. The shift and scale factors are chosen so that all four distributions have mean zero and variance σ^2.

All the simulations are done by using the R (Version 2.6.2) software package. Each simulation is based on 1000 Monte Carlo replications.

4.1 Bounded π_{ij} and Many Nuisance Parameters in Fixed-effect GLM

4.1.1 Modeling Issues

In the logistic regression model, the responses Y_{ij} are independent and $Y_{ij} \sim binomial(n_{ij}, \pi_{ij})$. The binomial pmf can be written in exponential family form as

$$f(y|\pi) = h(y)c(\pi)\exp[w_1(\pi)t_1(y)],$$

where

$$\begin{aligned} h(y) &= \begin{cases} \binom{n}{y} & \text{for } \; y = 0, 1, \ldots, n \\ 0, & \text{otherwise,} \end{cases} \\ c(\pi) &= (1-\pi)^n, \;\; 0 < \pi < 1, \\ w_1(\pi) &= \log(\pi/(1-\pi), \;\; 0 < \pi < 1, \\ t_1(y) &= y. \end{aligned}$$

Note that the parameter values $\pi = 0$ and $\pi = 1$ are sometimes included in the binomial model, but we should exclude them here because in logistic regression $\pi = \exp(\eta)/(1+\exp(\eta))$, where η is the logit or log odds of success. The logit can be any real number, but π must lie in the open interval $(0, 1)$.

4.1.2 Convergence and Existence of Finite Estimators

The log-likelihood function for the logistic regression model is a strictly concave function of the π_{ij}. Maximum likelihood estimates of the π_{ij} exist and are unique in unrestricted binomial models. However, the existence, finiteness, and uniqueness of

maximum likelihood estimates of the logistic regression parameters depend on the patterns of data points in the observation space [2].

In some simulations performed in this Chapter, there are SETs having π_{ij} very close to 1 for all groups j. Therefore, with a nonnegligible probability, simulated data may have $Y_{ij} = n_{ij}$ for all groups in set i. In such a case the MLE of π_{ij} is 1 and the MLE of α_i is infinity under the fixed effect model.

In practice, the `glm()` function in R fails to recognize that $\hat{\alpha}_i = \infty$. Instead it produces a large but finite estimate of α_i. After a few cycles of iterative fitting, the log likelihood looks flat at the working estimate, and convergence criteria are satisfied. The software then reports unreasonably large positive values of $\hat{\alpha}_i$ and huge standard errors for $\hat{\alpha}_i$ in about 12% of simulated GLM analyses. The large standard errors arise because they are calculated from the inverse matrix of negative second derivatives, which will all be very close to zero [1].

4.1.3 Research Question

As mentioned above, the logistic regression model assumes $0 < \pi_{ij} < 1$. However, in practical situations we sometimes observe $Y_{ij}/n_{ij} = 1$. We would like to know how this affects the estimation. Also, we know ML estimators of logistic model parameters work best when the sample size n is large compared to the number of parameters. If the number of nuisance parameters increases in proportion to the data accumulated, as in the fixed effect model, does the increasing number of nuisance parameters cause bad behavior of the analysis? Or is the bad behavior of

analysis entirely due to cases where $Y_{ij}/n_{ij} = 1$?

4.1.4 Simulation Design

We design two pairs of comparisons:

1. Situation A vs. Situation B

   Situation A: n_{ij} satisfy $20 \leq n_{ij} \leq 40$ rats for each group; number of SETs $= 43$; π_{ij} satisfy $0.07 \leq \pi_{ij} \leq 0.98$.

   Situation B: n_{ij} satisfy $20 \leq n_{ij} \leq 40$ rats for each group; number of SETs $= 43$; π_{ij} satisfy $0.10 \leq \pi_{ij} \leq 0.90$.

2. Situation C vs. Situation D

   Situation C: n_{ij} satisfy $20 \leq n_{ij} \leq 40$ rats for each group; number of SETs $= 86$; π_{ij} satisfy $0.07 \leq \pi_{ij} \leq 0.98$.

   Situation D: n_{ij} satisfy $20 \leq n_{ij} \leq 40$ rats for each group; number of SETs $= 86$; π_{ij} satisfy $0.10 \leq \pi_{ij} \leq 0.90$.

With the different situations in the simulation, the true values for each situation are varied. Instead of looking at actual values of sample moments, we compare their relative values. Let $\hat{\beta}_{\text{Monte Carlo}}$ be the average of $\hat{\beta}$ over 1000 Monte Carlo replications and let $\text{se}_{\text{Monte Carlo}}$ be the sample standard deviation of the Monte Carlo realizations of $\hat{\beta}$. Two statistics are calculated:

1. Relative bias $= [(\hat{\beta}_{\text{Monte Carlo}} - \beta_{\text{True}})/|\beta_{\text{True}}|] \times 100\%$.

2. Relative s.e. = coefficient of variation = $(\text{se}_{\text{Monte Carlo}}/\beta_{\text{True}}) \times 100\%$.

The SET parameters α_i are nuisance parameters, so we do not report results on all of them. Instead we only report results on SET 1, which is typical of the 36 SETs with only a single value of KCA2.

4.1.5 Simulation Results and Discussion

4.1.5.1 Effect of Bounding π_{ij}

In the comparison of Situation A vs. Situation B, we examine how the phenomenon of $Y_{ij} = n_{ij}$ affects the estimation of the regression coefficients and SET effects. From Table 4.1, the estimated regression coefficients of TFA2, KCA2, and $(\text{TFA2})^2$ look good in both situations. Note that the relative s.e. values always are larger in Situation B than in Situation A. This is because we adjusted the π_{ij} away from 0 or 1, and the variance will be the maximum when $p = 0.5$.

However, for the SET 1 effect, the relative bias is -40.59%, and Relative s.e. is 114.91% in SET 1 whenever $Y_{ij} = n_{ij}$ for all groups within SET 1. In Situation A, the estimates of SET 1 fall into two groups: one group is centered around -5 and contains about 88% of 1000 simulations, and the other group is centered around 11 and contains 12% of 1000 simulations. See Figure 4.1. In contrast, in Situation B the estimates of SET 1 are centered around -4 and have a bell shaped distribution. See Figure 4.2. These results indicate that when $y_{ij} = n_{ij}$, estimates produced by `glm()` are unreliable.

Table 4.1: *Effect of Bounding* π_{ij}*, where* $20 \leq n_{ij} \leq 40$ *and number of SETs = 43 for both situations. In situation A,* $0.07 \leq$ *original* $\pi_{ij} \leq 0.98$ *and in situation B,* $0.10 \leq$ *adjusted* $\pi_{ij} \leq 0.90$*.*

	Original π_{ij}		Adjusted π_{ij}	
Coefficient	Rel. Bias	Rel. se	Rel. Bias	Rel. se
TFA2	2.34	16.24	2.42	28.62
KCA2	2.96	15.40	2.56	29.35
$(\text{TFA2})^2$	2.72	29.63	3.89	50.43
SET 1	−40.59	114.91	1.57	29.69

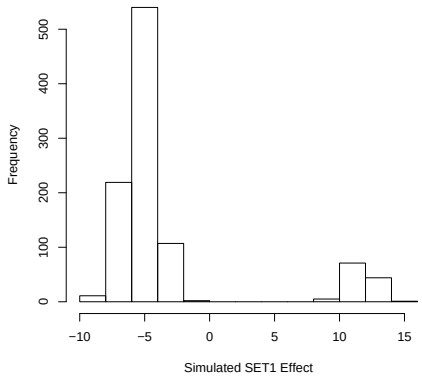

Figure 4.1: *When* $0.07 \leq$ *original* $\pi_{ij} \leq 0.98$*, in 88% of 1000 simulations the SET 1 effect is centered around* -5*, and in 12% of 1000 simulations it is centered around 11.*

We can explain the reason numerically based on the theory presented above.

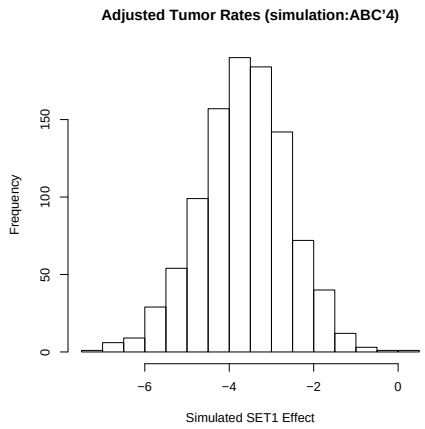

Figure 4.2: *When* $0.10 \leq$ *adjusted* $\pi_{ij} \leq 0.90$*, the estimates of SET 1 are centered around* -4 *and have a bell shaped distribution.*

The likelihood for the fixed effect model is:

$$L = \prod_{i=1}^{43} \prod_{j=1}^{J_i} \binom{n_{ij}}{y_{ij}} \frac{\exp[(\alpha_i + \beta_1 TFA_{ij} + \beta_2 (TFA_{ij})^2 + \beta_3 KCA_{ij}) y_{ij}]}{[1 + \exp(\alpha_i + \beta_1 TFA_{ij} + \beta_2 (TFA_{ij})^2 + \beta_3 KCA_{ij})]^{n_{ij}}}$$

The log-likelihood is:

$$\begin{aligned} \log L &= \log(const) + \sum \sum y_{ij} (\alpha_i + \beta_1 TFA_{ij} + \beta_2 (TFA_{ij})^2 + \beta_3 KCA_{ij}) \\ &\quad - \sum \sum n_{ij} \log[1 + \exp(\alpha_i + \beta_1 TFA_{ij} + \beta_2 (TFA_{ij})^2 + \beta_3 KCA_{ij})]. \end{aligned}$$

We differentiate $\log L$ with respect to α_i and set it equal to zero to obtain:

$$\frac{d \log L}{d \alpha_i} = \sum_{j=1}^{J_i} \left[Y_{ij} - n_{ij} \left(\frac{\exp(\alpha_i + \beta_1 TFA_{ij} + \beta_2 (TFA_{ij})^2 + \beta_3 KCA_{ij})}{1 + \exp(\alpha_i + \beta_1 TFA_{ij} + \beta_2 (TFA_{ij})^2 + \beta_3 KCA_{ij}} \right) \right] = 0$$

Therefore, if $y_{ij} = n_{ij}$ for each j,

$$\frac{\exp(\alpha_i + \beta_1 TFA_{ij} + \beta_2 (TFA_{ij})^2 + \beta_3 KCA_{ij})}{1 + \exp(\alpha_i + \beta_1 TFA_{ij} + \beta_2 (TFA_{ij})^2 + \beta_3 KCA_{ij})}$$

is forced to be 1. If the covariates are not constant, this forces $\hat{\alpha}_i = \infty$. Our average estimated regression coefficients are $\hat{\beta}_{TFA2} = 0.172$, $\hat{\beta}_{KCA2} = 0.126$, and

$\hat{\beta}_{(TFA2)^2} = -0.003$. For Group 2 in SET 1, TFA2 = 20 and KCA2 = 52.5. So

$$\hat{\beta}_{TFA2}(TFA2) + \hat{\beta}_{KCA2}(KCA2) + \hat{\beta}_{(TFA2)^2}(TFA2)^2 = 8.855.$$

If $\hat{\alpha}_i \approx 11$, the typical value of our "bad" estimates, then the logit is $\log(\hat{\pi_{ij}}/(1 - \hat{\pi_{ij}}) = 19.855$. So $\hat{\alpha}_i$ appears to go to infinity, although in fact $\hat{\alpha}_i \approx 11$ is finite.

In the 1000 Monte Carlo replications, there are 12% bad estimates for SET 1. This also can be explained. In SET 1, we have

$$\pi_{11} = 0.894 \qquad n_{11} = 14,$$

$$\pi_{12} = 0.981 \qquad n_{12} = 16.$$

Therefore

$$P[Y_{11} = n_{11}, Y_{12} = n_{12}] = \binom{14}{14}(0.894)^{14}\binom{16}{16}(0.981)^{16} = 0.12256.$$

Note that the bad behavior of estimation on SET 1 happened only when all the groups within that SET are have $Y_{ij} = n_{ij}$. When some but not all groups within one SET have $Y_{ij} = n_{ij}$, the estimate of that SET parameter involves data from other groups in the same SET, and the estimate will be finite.

4.1.5.2 Ratio of (Number of Nuisance Parameters)/(Number of Observations)

In the previous comparison, there were 43 SETs in both Situation A and Situation B. In the following comparison there are 86 SETs in both Situation C and Situation D. We investigate whether or not more nuisance parameters make the estimation

worse. Usually, we want the ratio of (number of nuisance parameters)/(number of observations) to go to zero. But here we force the ratio to be constant by doubling the number of SETs and keeping the n_{ij} the same. From Table 4.2, Figure 4.3 and Figure 4.4 we see the results are not much different from the previous comparison. So for the estimation, the ratio of (number of nuisance parameters)/(number of observations) does matter but not the number of nuisance parameters.

Table 4.2: *Effect of bounding π_{ij} while adding nuisance parameters (Situation C vs. Situation D), where $20 \leq n_{ij} \leq 40$ and number of SETs increase to 86 for both situations. In situation C $0.07 \leq$ original $\pi_{ij} \leq 0.98$, in situation D $0.10 \leq$ Adjusted $\pi_{ij} \leq 0.90$.*

	Original π_{ij}		Adjusted π_{ij}	
Coefficient	Rel. Bias	Rel. se	Rel. Bias	Rel. se
TFA2	1.44	11.61	1.49	20.43
KCA2	2.38	10.62	3.12	20.28
$(\text{TFA2})^2$	1.07	21.09	1.36	36.36
SET 1	−38.37	110.04	2.21	21.86

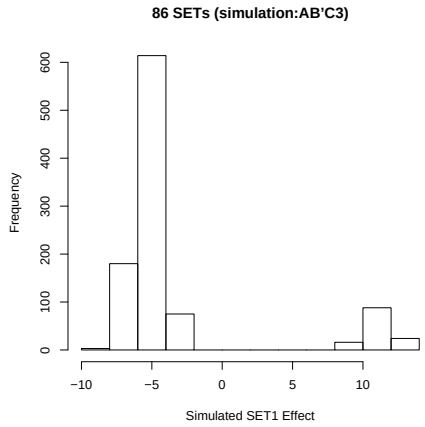

Figure 4.3: *SET1 effect (Situation C) where* $20 \leq n_{ij} \leq 40$ *and* $0.07 \leq$ *original* $\pi_{ij} \leq 0.98$*, but number of SETs increases to 86.*

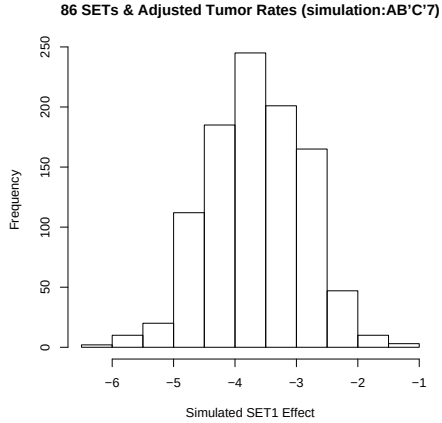

Figure 4.4: *SET 1 effect (Situation D), where* $20 \leq n_{ij} \leq 40$ *and* $0.10 \leq$ *adjusted* $\pi_{ij} \leq 0.90$*, but number of SETs increasing to 86*

4.1.5.3 The Consequence of Imputation for Missing Values and Bounded π_{ij} on SET 1

When there is only one value of KCA2 in a SET, the SET effect and the KCA2 effect are confounded. (One can not decompose the sum $\alpha_i + \beta_2 KCA2_i$ into its

components.) The result is that there is correlation between $\hat{\alpha}_i$ and $\hat{\beta}_2 KCA2_i$, since only their sum can be estimated. All SETs have two or more values of TFA2, so there is no confounding of SET and TFA2 effects.

When some probabilities are close to 1, there is also the problem of "bad" estimates of SET effects. Figure 4.5 and Figure 4.6 illustrate the situation for SET 1 under Situation A. Note that in each scatterplot there are two clouds of points, corresponding to "good" and "bad" estimates. These clouds of points correspond to the bimodal histogram in Figure 4.1. There is no correlation between $\hat{\beta}_{TFA2}$ and the SET 1 effect, but considerable correlation between $\hat{\beta}_{KCA2}$ and the SET 1 effect.

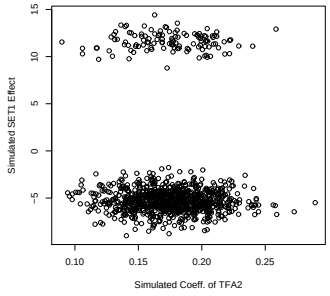

Figure 4.5: *TFA (no missing value) vs. SET 1, where* $20 \leq n_{ij} \leq 40$ *and* $0.07 \leq$ *original* $\pi_{ij} \leq 0.98$.

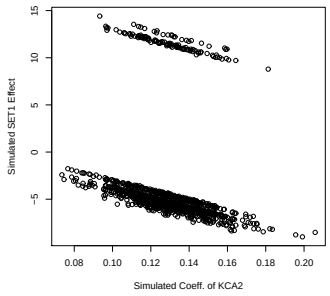

Figure 4.6: *KCA (imputation for missing value) vs. SET 1, where* $20 \leq n_{ij} \leq 40$ *and* $0.07 \leq$ *original* $\pi_{ij} \leq 0.98$.

The pattern is somewhat different under Situation B. Here the SET 1 estimates have a unimodal distribution, so the scatterplots of Figure 4.7 and Figure 4.8 each consist of a single cloud of points. There is no correlation between $\hat{\beta}_{TFA2}$ and the SET 1 effect, but a high correlation between $\hat{\beta}_{KCA2}$ and the SET 1 effect.

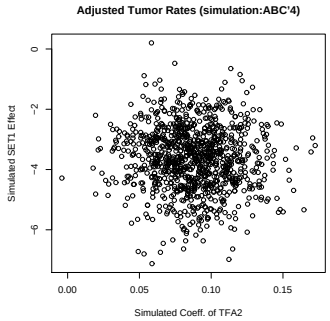

Figure 4.7: *TFA(no missing value) vs. SET 1 (Situation B), where* $20 \leq n_{ij} \leq 40$ *and* $0.10 \leq$ *adjusted* $\pi_{ij} \leq 0.90$.

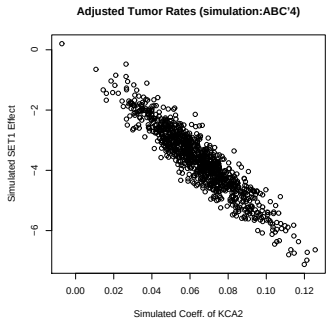

Figure 4.8: *KCA (imputation for missing value) vs. SET 1 (Situation B), where* $20 \leq n_{ij} \leq 40$ *and* $0.10 \leq$ *adjusted* $\pi_{ij} \leq 0.90$.

Though these simulation results look surprising and are not seen in the real data, could these situations happen? One of our conjectures is that when $n_{ij} = Y_{ij}$ happens in laboratory studies, the laboratory researchers may put the "bad" results aside and not bother to publish them. Is this a "publication bias" in Meta-analysis?

4.2 Behavior of Deviance in Fixed Effects Analyses

4.2.1 Sampling Distribution of the Deviance

In the following we consider the case where there is no nuisance parameter. One way of assessing the adequacy of a model is to compare it with a more general model with the maximum number of parameters that can be estimated while still remaining within the class of generalized linear models with the same distribution and same link function as the model of interest. If there are N binomial(n_i, π_i) observations Y_i, $i = 1, \ldots, N$, with potentially different π_i, then a saturated model

can be specified with N parameters. Under the saturated model $\hat{\pi}_i = Y_i/n_i$.

Let $\pi_{\max}$ denote the parameter vector for the saturated model and $\mathbf{p}_{\max}$ denote the maximum likelihood estimator of $\pi_{\max}$. The likelihood function, $L(\mathbf{p}_{max}; y)$ for the saturated model will be larger than any other likelihood function for these observations with the same assumed distribution and link function, because it provides the most detailed description of the data. Suppose that the model of interest is a GLM with $\pi_i = \pi_i(\boldsymbol{\beta})$ and let $L(\mathbf{b}; y)$ denote the maximum value of the likelihood function for the model of interest. Then the likelihood ratio

$$\lambda = \frac{L(\mathbf{p}_{\max}; y)}{L(\mathbf{b}; y)}$$

provides a way of assessing the goodness of fit for the model. The deviance, also called the log likelihood ratio statistic, is

$$D = 2[\log L(\mathbf{p}_{\max}; y) - \log L(\mathbf{b}; y)] = 2 \log \lambda.$$

If the fitted model is correct, the number of parameters N is fixed and all $n_i \to \infty$, then the sampling distribution of the deviance is, approximately,

$$D \sim \chi^2(N - p)$$

where

N is the number of parameters in the saturated model,

p is the number of parameters in the model of interest.

If one fits a linear model and the response variables Y_i are Normally distributed with known variance σ^2, then $D = SSE/\sigma^2$, where SSE is the usual residual sum of squares and D has an exact chi-squared distribution. In this case, however, D

depends on σ^2 which, in practice, is usually unknown. This means that D cannot be used directly as a goodness of fit statistic.

For Y_i's with other distributions, the sampling distribution of D may be only approximately chi-squared. If the number of parameters under the saturated model increases as the sample size N increases, the approximation will often not improve and the usual likelihood ratio approximation argument does not apply.

However, for the Binomial distribution with fixed I and large values of n_i, $i = 1, \ldots, I$, D can be calculated and used directly as a goodness of fit statistic [20]. Notice that D does not involve any nuisance parameters (like σ^2 for normal response data). The deviance for a binomial model is

$$\begin{aligned} D &= 2[l(p_{max}; y) - l(\mathbf{b}; y)] \\ &= 2\sum_{i=1}^{I}\left[y_i \log\left(\frac{y_i}{\pi_i(\mathbf{b})}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i(1-\pi_i(\mathbf{b}))}\right)\right]. \end{aligned}$$

This has the form

$$D = 2\sum \text{Obs} \log\left(\frac{\text{Obs}}{\text{Exp}}\right)$$

where Obs denotes the observed frequencies Y_{ij} and $n_{ij} - Y_{ij}$ and Exp denotes the corresponding estimated expected frequencies of successes and failures.

4.2.2 Research Question

In previous chapter, the data analysis shows that the dispersion parameter is 1.3 and the deviance D is much greater than the nominal expected value of $N - p = 58$, assuming the validity of the classical asymptotic theory. How accurate is the nominal

χ^2 approximation to the distribution of the deviance statistic? We want to know whether the behavior of the deviance is related to group sample sizes, n_{ij}, the number of nuisance parameters, or misspecification of the model, such as missing predictors, measurement errors, etc.

4.2.3 Simulation Design

We design two groups of comparisons:

1. Situation A vs. Situation B

   Situation A: n_{ij} satisfy $20 \leq n_{ij} \leq 40$ rats for each group, number of SETs $= 43$, π_{ij} satisfy $0.07 \leq \pi_{ij} \leq 0.98$.

   Situation B: n_{ij} satisfy $20 \leq n_{ij} \leq 40$ rats for each group, number of SETs $= 86$, π_{ij} satisfy $0.07 \leq \pi_{ij} \leq 0.98$.

2. Situation C vs. Situation D vs. Situation E

   Situation C: n_{ij} satisfy $200 \leq n_{ij} \leq 400$ rats for each group, number of SETs $= 43$, π_{ij} satisfy $0.07 \leq \pi_{ij} \leq 0.98$.

   Situation D: n_{ij} satisfy $20 \leq n_{ij} \leq 40$ rats for each group, number of SETs $= 43$, π_{ij} satisfy $0.10 \leq \pi_{ij} \leq 0.90$.

   Situation E: n_{ij} satisfy $200 \leq n_{ij} \leq 400$ rats for each group, number of SETs $= 43$, π_{ij} satisfy $0.10 \leq \pi_{ij} \leq 0.90$.

4.2.4 Simulation results and discussion

In the comparison of Situation A vs. Situation B, we investigate how a large numbers of nuisance parameters may affect the behavior of the deviance. In comparisons among A, C, D and E, we investigate the effect on deviance of group size, n_{ij}, and of bounds imposed on π_{ij}.

From Table 4.3, Figure 4.9 and Figure 4.10 we can see that the distribution of the deviance in both situations is stochastically larger than the nominal χ^2 distribution. The means and variances are both larger than under a true χ^2 distribution and the graphs show that the deviance histograms are to the right of the theoretical density. This means that doubling the sample size does not improve the agreement of the deviance distribution to the theoretical χ^2 distribution if the number of nuisance parameters is also doubled. It appears that the group sizes are too small to justify the large sample theory of the deviance statistic. As a result, the deviance test of fit is biased. For example, in Situation A the nominal $\chi^2_{58,0.05}$ cutoff point is 76.8, but from the simulation we find $P[D > 77.7] = 0.10$. Therefore a nominal 0.05 level deviance test has actual level greater than 0.10. In Situation B the results are similar: $\chi^2_{119,0.05} = 145.5$, but from the simulation $P[D > 147.8] = 0.10$.

The effect of bounding the π_{ij} away from 0 and 1 is shown in the comparison of Situations A and D. Table 4.3 shows that the mean and variance of the deviance under Situation D are closer to the nominal values of 58 and 116 than under Situation A. Figure 4.9 and Figure 4.11 support this finding: under Situation D the deviance histogram agrees more closely with the asymptotic χ^2 distribution. The tail prob-

Table 4.3: *Effect of Nuisance Parameters, Group Size and Bounding π_{ij} on Mean and Variance of Deviance*

	d.f.	mean	variance
Situation A($20 \leq n_{ij} \leq 40$, SETs = 43, $0.07 \leq \pi_{ij} \leq 0.98$)	58	62.30	138.95
Situation B($20 \leq n_{ij} \leq 40$, SETs = 86, $0.07 \leq \pi_{ij} \leq 0.98$)	119	127.40	251.33
Situation C($200 \leq n_{ij} \leq 400$, SETs = 43, $0.07 \leq \pi_{ij} \leq 0.98$)	58	58.26	117.23
Situation D($20 \leq n_{ij} \leq 40$, SETs = 43, $0.10 \leq \pi_{ij} \leq 0.90$)	58	60.42	126.94
Situation E($200 \leq n_{ij} \leq 400$, SETs = 43, $0.10 \leq \pi_{ij} \leq 0.90$)	58	58.19	107.28

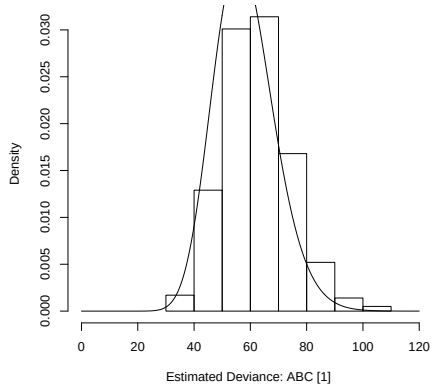

Figure 4.9: *The behavior of the deviance. Situation A (*$20 \leq n_{ij} \leq 40$*, SETs = 43,* $0.07 \leq \pi_{ij} \leq 0.98$*).*

abilities in Situation D agree somewhat more closely with the nominal probability, but the agreement is still not good: the Monte Carlo 90 percentile of the deviance is 75.2, so that $0.05 < P[D > 76.8] < 0.10$, where $\chi^2_{58,0.05} = 76.8$. Apparently the

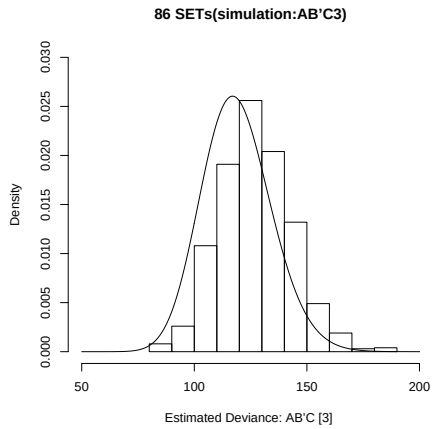

Figure 4.10: *The behavior of the deviance. Situation B (*$20 \leq n_{ij} \leq 40$*, SETs = 86,* $0.07 \leq \pi_{ij} \leq 0.98$*).*

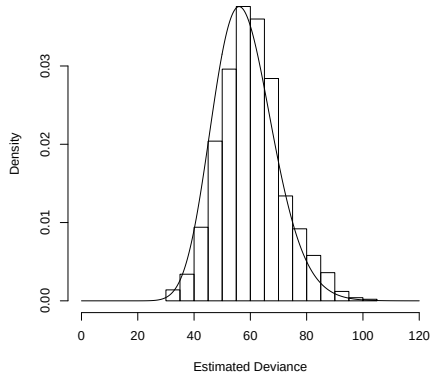

Figure 4.11: *The behavior of the deviance. Situation D (*$20 \leq n_{ij} \leq 40$*, SETs = 43,* $0.10 \leq \pi_{ij} \leq 0.90$*).*

possibility of all rats developing tumors has some effect on the distribution of the deviance. This is mitigated somewhat if the π_{ij} are bounded away from 0 and 1.

Situations C and E illustrate the effect of increasing sample size while controlling the number of nuisance parameters. Table 4.3 shows that the mean and variance

are close to the nominal values in both situations. Figure 4.12 and Figure 4.13 show good agreement between the actual and theoretical χ^2 distributions. The sample quantiles also agree well with the theoretical χ^2 quantiles. For instance, under Situation C the sample 95 percentile is 76.5, under Situation E the 95 percentile is 75.9, and the theoretical quantile is 76.8. The agreement is comparably good for other quantiles. We conclude that with very large (200–400) samples in each group, the asymptotic theory for the deviance is reliable and the chance that all rats develop tumors in some group is negligible.

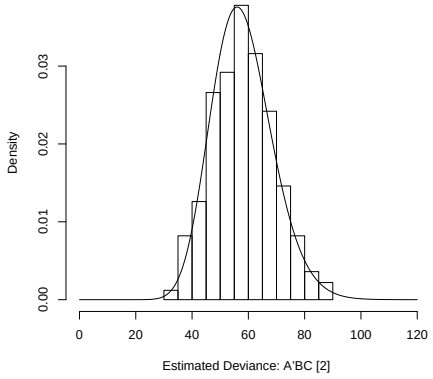

Figure 4.12: *The behavior of the deviance. Situation C(*$200 \leq n_{ij} \leq 400$*, SETs = 43,* $0.07 \leq \pi_{ij} \leq 0.98$*).*

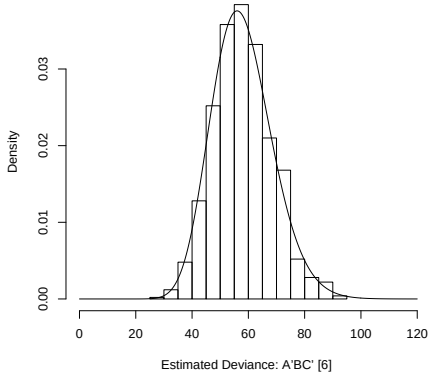

Figure 4.13: *The behavior of the deviance. Situation E(*$200 \leq n_{ij} \leq 400$*, SETs = 43,* $0.10 \leq \pi_{ij} \leq 0.90$*).*

4.3 Tests of Variance of Random Effects in GLMM

When using maximum likelihood analysis, tests are often based on the improvement in the maximized value of the log likelihood. The difference in twice the log likelihood is compared to a chi-squared distribution for statistical significance. For testing whether a single variance component is equal to zero the usual method must be slightly modified [17], [19]. Ordinarily we would take twice the difference in log likelihoods of the model with and without the random effect and compare that directly to a ${\chi_1}^2$ critical value. The modification is either to calculate a p-value and then cut it in half, or to compare to a cutoff point with twice the nominal α level. The test of

$$H_0 : \sigma_u^2 = 0 \quad \text{versus} \quad H_1 : \sigma_u^2 > 0$$

is a one-sided test. The usual likelihood ratio test is inherently two-sided and must be adjusted to reflect this fact. Specifically, consider the ML estimators of σ_u^2 in a balanced, one way random model. Under the normal theory, the ML estimator is

$$\hat{\sigma}_u^2 = [(1 - 1/k)MS(Betw) - MS(Error)]^+/n$$

where $[\cdot]^+ = \max(.\ , 0)$ denotes positive part.

If $\sigma_u^2 = 0$, then the ML estimator is often zero. In that case, the likelihood ratio test (LRT) statistic is given by

$$\begin{aligned} LRT &= -2[\log L(\sigma_u^2 = 0) - \log L(\sigma_u^2 = \hat{\sigma}_u^2)] \\ &= -2[\log L(\sigma_u^2 = 0) - \log L(\sigma_u^2 = 0)] \\ &= 0 \end{aligned}$$

About half the time the estimate would be zero so that the LRT statistic would be zero. With a point mass of approximately 0.5 at 0, the usual asymptotic distribution theory (suggesting a χ_1^2 distribution) clearly breaks down because the estimate gets "stuck" on the boundary. So the actual large-sample distribution under H_0 is a 50:50 mixture of a χ_1^2 and 0.

The software `glmer()` in `R` does not test hypotheses about variances of random effects, nor does it produce standard errors or confidence bounds for variances of random effects. Bates' `glmer()` was used and Slud's GLMM software is used for checking. The two programs agree with each other to six decimal places for regression coefficients and standard error. For the mixed-effect model the log likelihood is -1597.033. For the Fixed-effect model, which we fit as a GLMM with the variance

of random effect set equal to 10^{-7}, the log likelihood is -1736.198. The difference is $-1597.033 - (-1736.198) = 142.165$. The value is compared to a chi-square cutoff of $\chi^2_{1,0.90} = 2.71$ instead of $\chi^2_{1,0.95} = 3.84$. The test is highly significant at α level $= 0.05$. This GLMM likelihood optimization estimated the variance of the random effect as 0.882.

The asymptotic theory for GLMMs applies as the number of clusters increases, rather than as the numbers of observations within the clusters increase. To compute a confidence interval for σ_a^2 we performed a parametric bootstrap analysis. Using the parameter values as estimated by `glmer()`, we generated 1000 bootstrap replications of the original rat data. The random effects a_i were distributed as $N(0, 0.9115)$. This is also the `glmer()` fitted value. Using these simulated random effects, we then generated 104 binomial observations from the conditional logistic model, given the a_i. The bootstrap mean of $\hat{\sigma}_a^2$, the variance of the random effects, was 0.8847 and its standard deviation is 0.2262. The 95% bootstrap confidence interval for the variance of the random effects was [0.4414, 1.3279], which does not contain 0. Note that the bootstrap mean is not identical to the original `glmer()` estimate but is close to the estimate based on the log-likelihood analysis discussed in the previous paragraph.

4.4 Robustness of Mixed Analysis against Non-Normality of Random Effects

The issue of sensitivity to assumptions about the random effects distribution is not completely resolved. Some literature indicates that the choice is not so crucial while other papers indicate that it may be [1]. We explored the topic a little. The following distributions of random effects were simulated:

1. $N(0, \sigma^2)$.

2. scaled Student t with 4 d.f.. This distribution has heavy tails.

3. shifted and scaled gamma(4,1). This distribution is skewed.

4. uniform$[-\sigma\sqrt{3}, \sigma\sqrt{3}]$. This distribution has light tails.

All four distributions have mean zero and are scaled to have variance $\sigma^2 = 0.9115$, the value estimated by `glmer()` on the rat data.

We generated random SET effects from these distributions and then analyzed the data using GLMM, under the working assumption that they were normal random intercepts. Figure 4.14 displays the histogram of estimated SET effect variances for the case where the simulated random effect were $N(0, \sigma^2)$. For simulated uniform$[-\sigma\sqrt{3}, \sigma\sqrt{3}]$ SET effects, the results were very close to those with normal random effects. For Student t with 4 d.f., estimates of fixed effects were similar to normals, but the variance of the random effect is underestimated and its s.e. is higher. For Gamma(4,1), estimates, the standard errors of the fixed effects are like those in the normal case. The random effect variance was underestimated, but

with same standard error as in the normal case. Overall, estimation of fixed effect parameters seems quite robust against misspecification of random effects, but the estimation of random effect variances is sensitive to distributional misspecification.

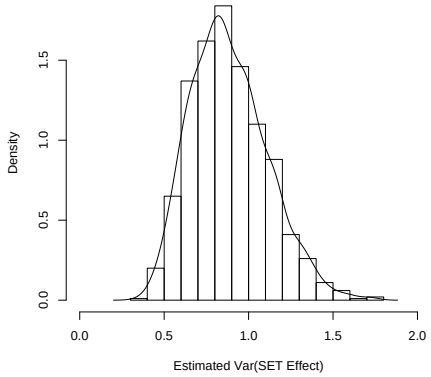

Figure 4.14: *Histogram and kernel estimate of random effect density (normal random effect - Correct model)*

4.5 Efficiency Loss - Comparing Coefficient Estimator Variance for Three Methods

Of the three approaches to meta-analysis considered in this thesis, conditional logistic regression is the simplest and fixed effect analysis is the most complicated. We want to compare their efficiencies. The Monte Carlo variances of regression coefficients from the different approaches were compared. Two sets of comparisons are made:

1. GLM vs. Clogist:

   n_{ij} satisfy $20 \leq n_{ij} \leq 40$ rats for each group; number of SETs = 43.

   n_{ij} satisfy $80 \leq n_{ij} \leq 160$ rats for each group; number of SETs = 43.

2. GLMM vs. Clogist

   n_{ij} satisfy $20 \leq n_{ij} \leq 40$ rats for each group; number of SETs = 43.

   n_{ij} satisfy $20 \leq n_{ij} \leq 40$ rats for each group; number of SETs = 86.

In each case 1000 Monte Carlo samples are generated. Samples with fixed effects use the parameter estimates obtained from the `glm()` analysis of the rat data. Samples with random effects use the parameter estimates obtained from the `glmer()` analysis of the rat data. The relative efficiency for a particular scalar parameter is defined as the ratio of the Monte Carlo variance of the conditional logistic estimator divided by the Monte Carlo variance of the MLE, which should be asymptotically most efficient. Recall that conditional logistic regression involves loss of all information about intercepts and may also lose information about regression coefficients.

The asymptotic theory for fixed effect models assumes $\min\{n_{ij}\} \to \infty$ while the number of groups is fixed. For small n_{ij}, as in the original rat data, Table 4.4 shows that the relative efficiencies are mixed. However, in this case we may generate artifacts due to the phenomenon of all rats in a SET developing tumors. When this occurs, `glm()` produces large but finite estimates of the SET parameter although the MLE should be infinite. When the sample sizes are increased by a factor of 4, the conditional logistic estimators are less efficient by anywhere from 1% to 7%, as

Table 4.4: *Relative Efficiency of coefficient estimators: (Variance from glm) / (Variance from clogist); 43 SETs, for* $20 \leq n_{ij} \leq 40$.

Coefficient	glm variance estimate	clogist variance estimate	efficiency
TFA2	7.810434×10^{-4}	8.212567×10^{-4}	95.10%
KCA2	3.779182×10^{-4}	3.275509×10^{-4}	115.38%
$(\text{TFA2})^2$	1.319375×10^{-6}	1.388029×10^{-6}	95.05%

Table 4.5: *Relative Efficiency: (Variance from glm) / (Variance from clogist); 43 SETs, for* $80 \leq n_{ij} \leq 160$.

Coefficient	glm variance estimate	clogist variance estimate	efficiency
TFA2	1.959070×10^{-4}	2.035563×10^{-4}	96.24%
KCA2	7.962770×10^{-5}	8.029840×10^{-5}	99.16%
$(\text{TFA2})^2$	3.252529×10^{-7}	3.507202×10^{-5}	92.74%

shown in Table 4.5.

The asymptotic theory for GLMM calls for the number of levels of the random effect to increase to infinity. Therefore we simulate data from either 43 SETs (104 groups) or 86 SETs (208 groups), with normally distributed random effects. Table 4.6 shows efficiency losses of 14%–17% for the coefficients of TFA2 and $(\text{TFA2})^2$ but 17% higher efficiency for KCA2, while Table 4.7 shows that conditional logistic regression is less efficient by anywhere from 8% to 12%.

Table 4.6: *Relative Efficiency: (Variance from glmer) / (Variance from clogist); 43 SETs, for* $20 \leq n_{ij} \leq 40$.

Coefficient	glmer	clogist	efficiency
TFA2	7.074027×10^{-4}	8.212567×10^{-4}	86.14%
KCA2	3.849120×10^{-4}	3.275509×10^{-4}	117.51%
$(\text{TFA2})^2$	1.157775×10^{-6}	1.388029×10^{-6}	83.41%

Table 4.7: *Relative Efficiency: (Variance from glmer) / (Variance from clogist); 86 SETs, for* $20 \leq n_{ij} \leq 40$.

Coefficient	glmer	clogist	efficiency
TFA2	3.326970×10^{-4}	3.796903×10^{-4}	87.62%
KCA2	1.944651×10^{-4}	2.103535×10^{-4}	92.45%
$(\text{TFA2})^2$	5.455341×10^{-7}	6.177444×10^{-7}	88.31%

For smaller samples with either random effects (Table 4.6) or fixed effects (Table 4.4), we note that conditional logistic regression is more efficient than the full likelihood estimates for the KCA2. Previously we have seen that there are strong correlations between the KCA2 coefficient and the SET effects in the fixed effect models. We conjecture that the anomalous relative efficiency results for KCA2 are related to the fact that in most SETs only one value of KCA2 is used in all groups, and typically that one value is the imputed value of 52.5. It appears that in such SETs, the SET effect, whether fixed or random, is confounded with the KCA2 effect, and this is somehow related to the efficiencies in smaller samples.

Chapter 5

Conclusions and Future Research

We performed a variety of analyses on Freedman's meta-analysis data and conducted several simulation studies to address issues related to our findings and to check the reliability of large sample approximations associated with our analyses.

Our analyses of the rat tumor data led to the following findings:

1. The data were analyzed using a fixed effect model, a random effect model, and conditional logistic regression. Estimates of the logistic regression coefficients based on these three methods were nearly the same. The standard errors were also nearly the same.

2. Freedman's original model was a fixed-effect model with three terms: SET, TFA2, and KCA2. We found that the squared term, $(\text{TFA2})^2$, is also significant according to the Wald test. Adding the $(\text{TFA2})^2$ to the model also reduces the deviance from 91 to 81, which represents a significant likelihood ratio test for the hypothesis that the coefficient of $(\text{TFA2})^2$ is zero. No other second degree terms improved the fit significantly.

3. Slud's GLMM software was used to test the hypothesis that the random effect variance is zero. This produced an approximate likelihood ratio test which rejected the null hypothesis. We also generated 1000 bootstrap replications to produce confidence bounds for the variance of the random effects. The two

procedures agree with each other, so we conclude that the random effects are significant.

Simulation studies were performed to check bias and standard errors of estimates in problems resembling the original rat data. We also used simulations to check on the validity of large sample theory, as either as the group sizes or the number of SETs became large. The important findings were as follows:

1. Under the fixed effect model, the regression coefficients were slightly biased. The bias was reduced if the number of observations per group, n_{ij}, were all increased. However, if the number of SETs was increased while the n_{ij} were unchanged, the bias was not improved. Similar results were observed when the data were modeled as a GLMM.

2. In fixed-effect models, when $Y_{ij}/n_{ij} = 1$ for all groups in SET i, the estimates of SET effect produced by `glm()` are unreliable. The ratio, (number of nuisance parameters)/(number of observations), determines the quality of the estimates of SET effect instead of the number of nuisance parameters.

3. In the fixed-effect model, the group sizes of the current data seem too small to rely on the large sample theory of the deviance statistic. The possibility of all rats developing tumors has some effect on the distribution of the deviance. As a result, the deviance statistic is stochastically larger than the theoretical χ^2 variable and the deviance test of fit is biased. The deviance statistic produces excessive Type I errors.

4. For GLMM, estimation of fixed effect parameters is robust against misspecification of the distribution of the random effects, but estimation of the random effect variance is sensitive to misspecification of the random effect distribution.

5. When the sample sizes are increased by a factor of 4, the conditional logistic estimators are less efficient by anywhere from 1% to 7%, compared to those in the fixed-effect model. When the number of SETs are increased by a factor of 2, the conditional logistic regression is less efficient by anywhere from 8% to 12%.

Our conclusion:

Conditional logistic regression avoids the possibility of bias when the number of studies is very large in a GLM analysis and also avoids effects of misspecification of the random effect distribution in a GLMM analysis, but at the cost of some information loss. When the number of studies is very large, it is worth while to try a GLMM analysis. If the number of observations within a group is large, one might want to use a GLM analysis.

5.1 Topics for Future Research

The rat data consist of many SETs, each with an unknown nuisance parameter. If the data are to be analyzed as a fixed effect model and the number of rats per group, n_{ij} were very small, we might expect that maximum likelihood estimates of the logistic regression coefficients might be inconsistent. However, in our problem,

the sample sizes in each group are moderate ($20 \leq n_{ij} \leq 40$). We saw small biases in our simulations, but the biases were not substantial. The natural question to study is: if the number of nuisance parameters grows, how fast must the n_{ij} grow to assure consistency of the MLE's?

If $Y_{ij} = n_{ij}$ in several groups, finite maximum likelihood estimates of the group specific intercepts may not exist. How should one proceed if the intercept parameters are of scientific interest? This problem might arise if the intercept parameters represent effects of some treatment factor, for example.

All of the analyses presented in this thesis were based on frequentist methods. Bayesian methods might also be applied, but one must devise reasonable prior distributions and analyze the performance of the resulting estimates. The behavior of Bayes estimates when the number of nuisance parameters is large is of particular interest.

Bibliography

[1] Agresti, A., (2002) *Categorical Data Analysis* (2nd ed.). New Jersey: Wiley

[2] Albert, A., and Anderson, A. 1984. On the Existence of Maximum Likelihood Estimates in the Logistic models. *Biometrika* **71**, 1-10

[3] Andersen, E. B. (1970). Asymptotic Properties of Conditional Maximum-Likelihood Estimators. *Journal of the Royal Statistical Society B* Vol. **32**, 283-301.

[4] Andersen, E. B. (1980). *Discrete Statistical Models with Social Science Applications*. Amsterdam: North-Holland.

[5] Best, N. G., Spiegelhalter, D. J., Thomas, A. and Brayne, C. E. G. (1996). Bayesian analysis of realistically complex models. *Journal of the Royal Statistical Society, Series A*, **159**, 323-342.

[6] Cox, D. R. (1990). Role of Models in Statistical Analysis. *Statistical Science* Vol. **5**, No.**2**, 169-174.

[7] Diggle, P. J., Liang, K. -Y.,and Zeger, S. L. (1994) *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

[8] Freedman, L. S., Clifford, C., and Messina, M. (1990). Analysis of Dietary Fat, Calories, Body Weight, and the Development of Mammary Tumors in Rats and Mice: A Review. *Cancer Research*, **50**, 5710-5719.

[9] Freedman, L. S. (1994). Meta-Analysis of Animal Experiments on Dietary Fat Intake and Mammary Tumors. *Statistics in Medicine*, **13**, 709-718.

[10] Fay, M. P., Graubard, B. I., Freedman, L. S., and Midthune, D. N. (1998). Conditional logistic regression with sandwich estimators: Application to a meta-analysis. *Biometrics* **54**, 195-208.

[11] Glass, G. V. (1976). Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher*, **5-10**, 3-8.

[12] Hedges, L. V., and Olkin, I. (1985). *Statistical Methods for Meta-analysis*. New York: Academic.

[13] Kauermann, G., and Carroll, R. J. (2001). A Note on the Efficiency of Sandwich Covariance Matrix Estimation. *J. Amer. Statist. Assoc.* **96**, 1387-1397.

[14] Neuhaus, J. M. and Jewell, N. P. (1990b). The effect of retrospective sampling on binary regression models for clustered data. *Biometrics* **46**, 977-990.

[15] Neuhaus, J. M. and Lesperance, M. L. (1996). Estimation efficiency in a binary mixed-effect model setting. *Biometrika* **83**, 441-446.

[16] R Development Core Team, R: *A language and environment for statistical computing*. Vienna: R Foundation for Stat. Comp., http://WWW.R-project.org, 2008.

[17] Self, S. and Liang, K, -Y. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Test Under Nonstandard Conditions. *J. Amer. Statist. Assoc.*, **82**, 605-610.

[18] Slud E. V., Personal communication.

[19] Stram, D. O. and Lee, J. W. (1994). Variance Components Testing in the Longitudinal Effects Model. *Biometrics*, **50**, 1171-1177.

[20] Venables, W. N., and Ripley, B. D. (1999) *Modern Applied Statistics with S-Plus*. New York: Springer.

[21] Yusuf, S., Peto, R., Lewis, J., Collins, R., and Sleight, P. (1985). Beta blockade during and after myocardial infarction: An overview of the randomized trials. *Progress in Cardiovascular Diseases*, **27**, 335-371.