UNIVERSITY OF CALIFORNIA
RIVERSIDE

Goodness-of-Fit Tests for Autoregressive Logistic Regression Models and Generalized Linear Mixed Models

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Anne Mary Hansen

December 2012

Dissertation Committee:

Dr. Daniel Jeske, Chairperson
Dr. James Flegal
Dr. Kurt Schwabe

The Dissertation of Anne Mary Hansen is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

Thank you to Dr. Daniel Jeske for your excellent advising and guidance during this process. Also thank you to my family and friends for their many years of support. No one can accomplish this alone. Thank you also to the Statistics Department at UC Riverside, which has been like a second family.

ABSTRACT OF THE DISSERTATION

Goodness-of-Fit Tests for Autoregressive Logistic Regression Models
and Generalized Linear Mixed Models

by

Anne Mary Hansen

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, December 2012
Dr. Daniel Jeske, Chairperson

Goodness-of-fit is a very important concept in data analysis, as most statistical models make some underlying assumptions. When these assumptions are violated, any model inference can be suspect. Thus, a goodness-of-fit check is necessary in order to trust any conclusions drawn from the model. Herein we propose two goodness-of-fit tests, one that addresses autoregressive logistic regression (ALR) models and another that is appropriate for generalized linear mixed models (GLMMs).

Both GLMMs and ALR models are extensions of generalized linear models, a broad class of models that includes logistic regression and Poisson regression. ALR models go a step beyond typical generalized linear models by regressing upon past observations. In contrast, GLMMs go beyond the scope of generalized linear models by incorporating random effects.

For the ALR model, a chi-square test is proposed and the asymptotic distribution of the statistic is derived. General guidelines for a two-dimensional, dynamic binning strategy are provided, which make use of two types of maximum likelihood parameter estimates. For smaller sample sizes, a bootstrap p-value procedure is discussed. Simulation studies indicate that the procedure has the correct size and is sensitive to model misspecification. In particular, the test is very good at detecting the need for an additional lag. An application to a dataset relating to late-onset Alzheimer's disease is provided.

For GLMMs, we propose a Cramer-von-Mises omnibus test statistic, which extends upon a procedure applied to Poisson regression. Here, predictors of the random effects are plugged into the model to approximate a simpler, generalized linear model. The statistic is then calculated by making use of a probability integral transformation. Simulation studies

indicate that the test has good size and power for a Poisson GLMM. Some ideas for future research are also proposed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Background

1.1 Goodness-of-Fit

> "From the earliest days of statistics, statisticians have begun their analysis by proposing a distribution for their observations and then, perhaps with somewhat less enthusiasm, have checked on whether this distribution is true ... test procedures have appeared, and the study of these procedures has come to be known as goodness-of-fit"(D'Agostino & Stephens, 1986, Preface).

No introductory Statistics course would be complete without a discussion of simple linear regression, its underlying assumptions, and a goodness-of-fit analysis. In general, goodness-of-fit is a very important concept in data analysis, as most statistical models make some underlying distributional assumptions. When these assumptions are violated, any inference from the models can be biased or in some cases entirely misleading. Thus, a verification of these assumptions, commonly known as a goodness-of-fit check, is necessary in order to take a model's findings seriously. Goodness-of-fit encompasses graphical approaches as well as formal hypothesis tests of model adequacy.

In the case of linear regression, goodness-of-fit is often resolved by looking at plots of model residuals. For more complex linear models, such as those that make use of non-identity link functions or incorporate random effects, goodness-of-fit is still of great concern. However, how to go about evaluating the fit is often not as obvious as in the case of simpler models.

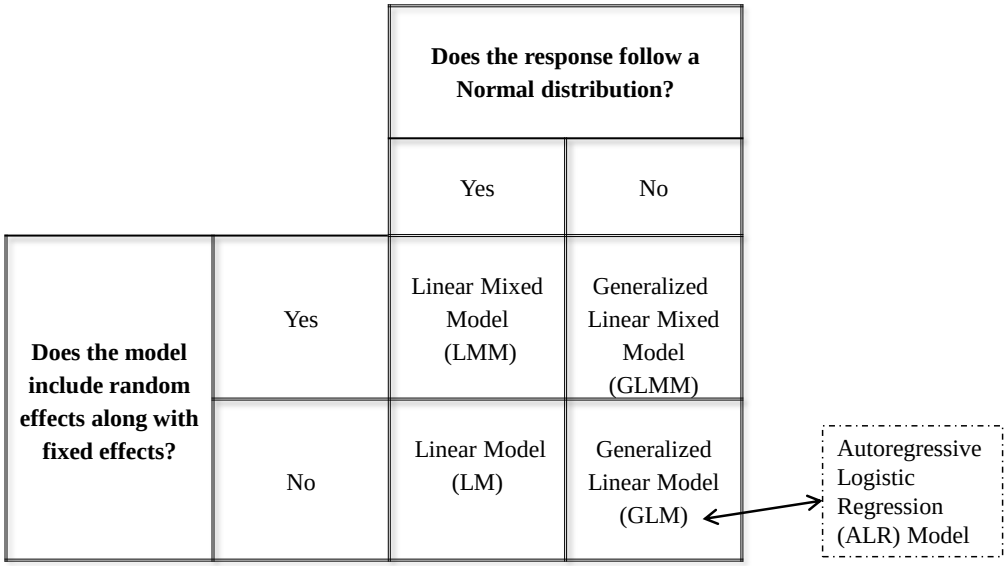
This dissertation will focus on assessing the goodness-of-fit for two types of models: Generalized Linear Mixed Models (GLMMs) and Autoregressive Logistic Regression (ALR) models. Both of these types of models could be considered extensions of a broad class of

models called Generalized Linear Models (GLMs). Similarly, GLMs are an extension of linear models (LMs), in that they "generalize" linear models for data that follow a non-Gaussian distribution. In contrast, linear mixed models (LMMs) are linear models that include random effects. Figure 1.1 provides a visual of the relationships between these model types.

ALR models can be thought of as a special extension of GLMs, since they regress upon past states but can be treated like a GLM in order to estimate parameters. A GLMM is a GLM that incorporates random effects.

Although linear models, linear mixed models, and generalized linear models are not the direct subject of this dissertation, these classes of models lay the groundwork for GLMMs and ALR Models.

Figure 1.1: A comparison of models.



1.2 Linear Models

Suppose we have n items of data, $Y_1, Y_2, ..., Y_n$. McCulloch, Searle and Neuhaus (2008) describe a linear model equation to be

$$E[y_i] = \mu_i \tag{1.1}$$

Equivalently, we might write equation (1.1) using vector notation:

$$E[Y] = \mu$$

where

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \text{ and } \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$$

Both vector notation and standard notation will be used in tandem in this dissertation.

Since Y is data, we can think of it as realized values of some random process. As such, linear models further assume $Y \sim (\mu, V)$ implying that Y has a mean μ and an nxn variance-covariance matrix V, where

$$V = \begin{bmatrix} v_{11} & v_{12} & ... & v_{1n} \\ v_{21} & v_{22} & ... & v_{2n} \\ ... & ... & ... & ... \\ v_{n1} & v_{n2} & ... & v_{nn} \end{bmatrix}$$

and $Cov(Y_i, Y_{i'}) = v_{ii'} = v_{i'i}$.

Equation (1.1) is extremely general. Note that there are n data values, but μ could potentially have up to n unique elements, and V might have $n(n+1)/2$ unique elements. In order to estimate μ and V, they must be modeled using less than n parameters. A good model will specify a μ and V in terms of k parameters ($k < n$) that are appropriate for the process under study and have good explanatory power.

In general, μ has the form $\mu = X\beta$ and $V = \sigma^2 I_n$, where X is a matrix of covariate values, β is a vector of parameters, σ^2 is the variance parameter, and I_n is an n x n identity matrix. Depending on the context, X is known as a design or covariate matrix. A linear regression model is the archetype example of a linear model. Let Y_i be the response of the ith subject $i = 1, ..., n$. Suppose also there are p associated covariates values $X_{i1}, X_{i2}, ..., X_{ip}$. Then, a linear regression is of the form:

$$\begin{aligned} E[Y_i] = \mu_i &= X_i'\beta \\ &= \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{k-1} X_{ip} \end{aligned} \tag{1.2}$$

where $\beta = (\beta_0, \beta_1, ..., \beta_p)$ is a vector of $(p+1)$ parameters and $X_i = (1, X_{i1}, X_{i2}, ..., X_{ip})$. Further, it is assumed that $Y_i \sim$ i.i.d. $N(\mu_i = X_i'\beta, \sigma^2)$. Parameter estimates, $\hat{\beta}$ and $\hat{\sigma}^2$,

can be found using the methods of least squares or maximum likelihood estimation. For linear regression models, parameter estimates have straightforward, closed forms such that

$$\hat{\beta} = (X'X)^{-1}X'Y, \quad \hat{\sigma}^2 = MSE = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{n - \#\text{parameters} - 1} \tag{1.3}$$

where $\hat{Y}_i = X_i'\hat{\beta}$. Other common linear models include both one-way and two-way Analysis of Variance (ANOVA) models.

1.2.1 Goodness-of-fit for Linear Models

Goodness-of-fit is focused on testing that the model assumptions are upheld by the data. For linear regression and other linear models, the main assumption is that $Y_i \sim$ i.i.d. $N(\mu_i, \sigma^2)$. Normality can be assessed by generating a quantile-quantile (QQ) plot of the residuals or errors, $\hat{e}_i = Y_i - \hat{Y}_i$. The QQ-plot compares the ordered residuals to quantiles of the standard Normal distribution. Further, model residuals are generally plotted versus fitted values $\hat{Y}_i$ and versus the individual covariates (columns of X) to check for independence and homogeneity of σ^2. There are also numerous, formal hypothesis tests for normality and constant variance. These include the Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors, and Anderson-Darling tests which can be applied to model residuals to check for normality (Razali & Wah, 2011), and the Brusch-Pagan and White tests for heteroscedasticity.

If some of the assumptions are violated, parameter estimates such as those given in (1.3) may be biased, perhaps extremely so. Further, any confidence intervals or hypothesis tests about the parameters may be misleading. Thus, it is important to check that the model's distributional assumptions are fulfilled.

1.2.2 Model Selection and Goodness-of-Fit

Along with checking goodness-of-fit, a modeler must choose which covariates to include in the model and what form they will take (e.g. quadratic terms, interactions). This process is generally known as "model selection" and deals with how to pick the best $X\beta$. We will briefly discuss model selection here, as many of the techniques can be applied to more complicated models (such as GLMMs and ALR models).

An ideal model will be parsimonious (i.e., have few parameters) and have good explanatory power. In the case of linear regression, the coefficient of determination R^2 can

be calculated, where $R^2 = 1 - \sum_i \hat{e}_i^2 / \sum_i (Y_i - \bar{Y})^2$. This coefficient describes what percentage of variation in the data can be explained by the model. For competing models, R^2 values can be directly compared. Further, for nested models, a likelihood ratio test can be performed to see if incorporating additional variables provides enough explanatory power to make them worth adding to the model. For non-nested models (as well as nested ones), Akaike's (1974) Information Criterion (or AIC) can be used to quickly compare different models. AIC is defined as

$$AIC = -2\text{log(likelihood)} + 2(\#\text{ of model parameters}).$$

AIC is essentially twice the negative log likelihood with a penalty for the number of parameters in the model. A small AIC value is preferable. Thus, models can be compared side-by-side for the smallest AIC. Similar to AIC, a Bayesian Information Criterion (BIC) has been developed for comparison between models, where BIC is

$$BIC = -2\text{log(likelihood)} + (\#\text{ of model parameters})log(n).$$

Compared to AIC, BIC takes the sample size n into account, and leans less quickly towards complex models as n increases. For additional details, see Raferty (1986).

One drawback of R^2 is that it will generally increase as the number of parameters increases. Thus, a related measure, adjusted R^2, can be calculated (Hocking, 1976) such that

$$R^2_{adj} = 1 - \frac{(n-1)}{(n - \#\text{ parameters})}(1 - R^2).$$

This R^2_{adj} will only increase if a new term enhances the model more than what is expected just by chance. Other criterion such as Mallow's C_p and the PRESS statistic can also be used for choosing between models (Mendenhall & Sincich, 2003).

Goodness-of-fit and model selection are not entirely unrelated concepts. For example, linear regression residual plots can identify the need for quadratic terms and spot the most influential covariates. A potential model identified by model selection techniques must also be checked for goodness-of-fit. Many of the model selection criteria, such as AIC, can easily be extended to more complex models, as long as those models have a likelihood. In contrast, goodness-of-fit usually becomes more difficult to test as the complexity of the model increases.

In some instances where data is not normally distributed, a function of the mean may instead be modeled as a linear combination of covariates. This function is called a link function, and the model is called a generalized linear model (GLM).

1.3 Generalized Linear Models

To construct a GLM, we must make three decisions:

1. What is the distribution of the response data?
2. What function of the mean will be modeled as linear in the predictors? (Link Component)
3. What will the covariates be? (Systematic Component)

For linear models such as linear regression, the first two questions are already answered for us (normal distribution, identity link). For a GLM, assume $Y_i \sim$ independent $f_{Y_i}(\cdot|\mu_i, \theta)$. The function $f_{Y_i}(\cdot|\mu_i, \theta)$ is a density function with a mean value of μ_i and θ represents any nuisance parameters, such as σ^2 in the case of the Normal distribution. Then, the general form of a GLM is given below (McCulloch et al., 2008):

$$\begin{aligned} E[Y_i] &= \mu_i \\ g(\mu_i) &= X_i'\beta. \end{aligned} \tag{1.4}$$

where X_i and β are defined as before and $g(\cdot)$ is a known link function. Examples of GLMs include models such as probit regression, logistic regression, and Poisson regression. In fact, all linear models could be classified as a special case of GLM with an identity link function.

Table 1.1: Some well-known GLMs.

GLM	Response Distribution	Model Equation	Link Function
Linear Model	$y_i \sim N(\mu_i, \sigma^2)$	$E[y_i] = \mu_i$	$g(\mu_i) = \mu_i$ (identity)
Probit	$y_i \sim Bernoulli(p_i)$	$E[y_i] = p_i$	$g(p_i) = \Phi^{-1}(p_i)$
Poisson (regular)	$y_i \sim Poisson(\lambda_i)$	$E[y_i] = \lambda_i$	$g(\lambda_i) = log(\lambda_i)$
(overdispersed)	$y_i \sim Neg.Binom.(\lambda_i, k)$	$E[y_i] = \lambda_i$	$g(\lambda_i) = log(\lambda_i)$
Logistic (binary)	$y_i \sim Bernoulli(p_i)$	$E[y_i] = p_i$	$g(p_i) = log\frac{p_i}{1-p_i}$

Parameter estimates for GLMs can be found using maximum likelihood estimation techniques. Generally closed form solutions for β do not exist, however iterative least square methods (such as the Fisher Scoring or Newton-Raphson method) can quickly zero in on parameter estimates that maximize the likelihood (Agresti, 2003).

A goodness-of-fit analysis of a GLM would address the adequacy of all three components of the model. Is the response distribution appropriate? Is the link function correct? Do

the covariates in the model adequately predict the response? There are a plethora of goodness-of-fit techniques for generalized linear models. Here, we will focus on a few of the goodness-of-fit tests for logistic regression and Poisson regression.

1.3.1 Logistic Regression Models

As Table 1.1 indicates, a logistic regression model employs a log odds link function to relate probability to a linear combination of covariates. Suppose $Y_1, Y_2, ..., Y_n$ take binary values (0 or 1). For simplicity, we will assume that there is a single covariate X_i for each subject and that $p_i = P(Y_i = 0|X_i)$. Then, we can describe a logistic model equation:

$$log\Big(\frac{p_i}{1-p_i}\Big) = \alpha + \beta X_i.$$

From this equation, we can solve for the individual p_i, i.e.,

$$\frac{p_i}{1-p_i} = e^{\alpha+\beta X_i} \Rightarrow p_i = (1-p_i)e^{\alpha+\beta X_i} \Rightarrow p_i = \frac{e^{\alpha+\beta X_i}}{1+e^{\alpha+\beta X_i}}$$

where the likelihood is given by

$$L(\alpha, \beta) = \prod_{i=1}^{n} p_i^{1-Y_i}(1-p_i)^{Y_i}.$$

Logistic regression models can also be used for polytomous (i.e., multi-state) observations that take a finite number of possible states. For example, suppose observations take one of three states: A, B, or C. As there are now three possible states, two logit equations must be set up to define the logistic regression. Define $p_{iA} = P(Y_i = A|X_i)$ and let this be the reference probability. Note that the choice of reference probability is arbitrary. Then, the model equations are:

$$log\Big(\frac{p_{iB}}{p_{iA}}\Big) = \alpha_1 + \beta_1 X_i$$
$$log\Big(\frac{p_{iC}}{p_{iA}}\Big) = \alpha_2 + \beta_2 X_i.$$

This implies

$$p_{iB} = p_{iA}e^{\alpha_1+\beta_1 X_i}$$
$$p_{iC} = p_{iA}e^{\alpha_2+\beta_2 X_i}$$
$$\Rightarrow 1 = p_{iA}(1 + e^{\alpha_1+\beta_1 X_i} + e^{\alpha_2+\beta_2 X_i})$$

such that

$$p_{iA} = \frac{1}{(1 + e^{\alpha_1+\beta_1 X_i} + e^{\alpha_2+\beta_2 X_i})},\ p_{iB} = \frac{e^{\alpha_1+\beta_1 X_i}}{(1 + e^{\alpha_1+\beta_1 X_i} + e^{\alpha_2+\beta_2 X_i})},$$

and

$$p_{iC} = \frac{e^{\alpha_2+\beta_2 X_i}}{(1 + e^{\alpha_1+\beta_1 X_i} + e^{\alpha_2+\beta_2 X_i})}$$

Here, the likelihood is given by

$$L(\lambda) = \prod_{i=1}^{n} (p_{iA})^{I[Y_i=A]} (p_{iB})^{I[Y_i=B]} (p_{iC})^{I[Y_i=C]}. \tag{1.5}$$

where $\lambda = (\alpha_1, \alpha_2, \beta_1, \beta_2)$ and $I[\cdot]$ is the indicator function.

Now suppose that A,B, and C are ordinal in nature. It makes sense to perhaps use cumulative logits. Here we would have

$$log\Big(\frac{P(Y_i = A|X_i)}{1 - P(Y_i = A|X_i)}\Big) = \alpha_1 + \beta X_i$$
$$log\Big(\frac{P(Y_i \leq B|X_i)}{1 - P(Y_i \leq B|X_i)}\Big) = \alpha_2 + \beta X_i.$$

This implies

$$P(Y_i = A|X_i) = \frac{e^{\alpha_1+\beta X_i}}{1 + e^{\alpha_1+\beta X_i}}$$
$$P(Y_i \leq B|X_i) = \frac{e^{\alpha_2+\beta X_i}}{1 + e^{\alpha_2+\beta X_i}}$$

Note that there is a common slope β. This is required so that no negative probabilities are obtained. For instance, suppose the two logits had different slopes, β_1 and β_2 respectively, then

$$P(Y_i = B) = P(Y_i \leq B) - P(Y_i = A) = \frac{e^{\alpha_2+\beta_2 X_i}}{1 + e^{\alpha_2+\beta_2 X_i}} - \frac{e^{\alpha_2+\beta_1 X_i}}{1 + e^{\alpha_2+\beta_1 X_i}}$$

which could be negative unless $\beta_1 = \beta_2 = \beta$. Along with a common slope, it can be shown that since $P(Y_i = B) \geq 0$ this implies that $\alpha_2 \geq \alpha_1$. Thus, ordinal logits require a common slope and a monotone ordering of the intercepts. Probabilities for each state can be found via algebra, such that

$$p_{iA} = \frac{e^{\alpha_1+\beta X_i}}{1 + e^{\alpha_1+\beta X_i}},\ p_{iB} = \frac{e^{\alpha_2+\beta X_i}}{1 + e^{\alpha_2+\beta X_i}} - \frac{e^{\alpha_1+\beta X_i}}{1 + e^{\alpha_1+\beta X_i}},\ p_{iC} = \frac{1}{1 + e^{\alpha_2+\beta X_i}}.$$

The likelihood equation is then as given in equation (1.5) where $\lambda = (\alpha_1, \alpha_2, \beta)$.

Logistic regression models can be easily extended to four or more states by increasing the number of model equations. If observations can take one of K states, then $K-1$ distinct logit equations are required.

1.3.2 General (Nominal) versus Ordinal Logits

If possible states are ordinal in nature, it makes sense to consider using ordinal logits. The advantage of this is that there are generally fewer parameters when using ordinal logits. However, the use of ordinal logits requires a common slope and the monotonicity of the intercepts. This forces what is known as a proportional odds relationship. For example, if we look at the log odds ratios of a one unit change in X for the 3-state ordinal model, we observe the following for the two logit equations:

$$log\left\{\frac{\frac{P(Y_i=A|X_i=1)}{1-P(Y_i=A|X_i=1)}}{\frac{P(Y_i=A|X_i=0)}{1-P(Y_i=A|X_i=0)}}\right\} = (\alpha_2+\beta)-\alpha_2 = \beta$$

$$log\left\{\frac{\frac{P(Y_i\leq B|X_i=1)}{P(Y_i>B|X_i=1)}}{\frac{P(Y_i\leq B|X_i=0)}{P(Y_i>B|X_i=0)}}\right\} = (\alpha_1+\beta)-\alpha_1 = \beta.$$

The log odds ratio of a one unit change in X for both the first and second logit equations is equal to β. The same cannot be said when using the nominal logits. For example, when using the nominal logits

$$log\left\{\frac{\frac{P(Y_i=A|X_i=1)}{1-P(Y_i=A|X_i=1)}}{\frac{P(Y_i=A|X_i=0)}{1-P(Y_i=A|X_i=0)}}\right\} = log\left\{\frac{\frac{1}{e^{\alpha_1+\beta_1}+e^{\alpha_2+\beta_2}}}{\frac{1}{e^{\alpha_1+\alpha_2}}}\right\} = log\left\{\frac{e^{\alpha_1+\alpha_2}}{e^{\alpha_1+\beta_1}+e^{\alpha_2+\beta_2}}\right\}$$

$$log\left\{\frac{\frac{P(Y_i\leq B|X_i=1)}{P(Y_i>B|X_i=1)}}{\frac{P(Y_i\leq B|X_i=0)}{P(Y_i>B|X_i=0)}}\right\} = log\left\{\frac{\frac{1+e^{\alpha_1+\beta_1}}{e^{\alpha_2+\beta_2}}}{\frac{1+e^{\alpha_1}}{e^{\alpha_2}}}\right\} = log\left\{\frac{e^{\alpha_2}+e^{\alpha_1+\alpha_2+\beta_1}}{e^{\alpha_2+\beta_2}+e^{\alpha_1+\alpha_2+\beta_2}}\right\}.$$

Thus, we have symmetry in the proportional odds when using ordinal logits. If a dataset is ordinal in nature and a proportional odds relationship is reasonable, then ordinal logits are a good choice. Additionally, a likelihood ratio test can be performed to assess nominal versus ordinal logits, since ordinal logits are the most general parameterization and ordinal logits represent a constrained version of the general model, where $\alpha_1 \leq \alpha_2$ and $\beta_1 = \beta_2 = \beta$.

1.3.3 Goodness-of-Fit Tests for Logistic Regression Models

A number of goodness-of-fit tests exist for logistic regression models, such as the well-known Hosmer-Lemeshow (1980) test, which is available in most software packages and is appropriate for a binary response variable. This test works by evenly partitioning observations into at least three, but no more than ten, groups based upon the model's predicted probabilities for each outcome (e.g. $P(Y = 1)$), and then constructing a Pearson chi-square

statistic across those cells. It has been pointed out that a flaw of the Hosmer-Lemeshow procedure is that observations with very different covariate patterns may be grouped together into the same cell. This can be overcome by instead creating one group for each unique covariate pattern and then constructing the statistic (Tsiatis, 2002). Extensions of the Hosmer-Lemeshow test have been made to logistic regression with continuous covariates (Pulkstenis & Robinson, 2002), as well as to multinomial logistic regression models (Fagerland, Hosmer, & Bofin, 2008). Additionally, a tree-based model checking procedure has also been proposed (Su, 2007) using classification and regression trees (CART), which can shed light on the source of why a model may not be fitting properly.

1.3.4 Poisson Regression Models

Poisson regression models are a type of GLM that relate a discrete response variable Y, assumed to have a Poisson distribution, to a linear combination of covariates through a log-link function. Suppose $Y_1, Y_2, ..., Y_n$ take non-negative, discrete values and that it is reasonable to think that $Y_i \sim Poisson(\lambda_i)$, $i = 1, 2, ...n$. For simplicity, we can again assume that there is a single covariate X_i for each subject. Then, a Poisson regression model is given by

$$log(\lambda_i) = \alpha + \beta X_i$$

From this equation, we can solve for the λ_i such that $\lambda_i = e^{\alpha+\beta X_i}$. The likelihood is then given by

$$L(\alpha, \beta) = \prod_i^n \frac{e^{-\lambda_i} \lambda_i^{Y_i}}{Y_i!}.$$

Recall that for the Poisson distribution, the mean and variance are equal. In some circumstances, the Poisson distribution is not adequate, as the data may have a variance larger than its mean. In this case, we can model the observations as Negative Binomial, which incorporates an additional dispersion parameter.

1.3.5 Goodness-of-Fit for Poisson Regression Models

Agresti (2003) describes a number of goodness-of-fit tests that can be applied to a Poisson regression setting to test the appropriateness of the Poisson distribution. A generalized Pearson χ^2 statistic has been used

$$P = \sum_{i=1}^{n} \frac{(Y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i}$$

which under the null and some regularity conditions has been shown to have a χ^2_{n-k} distribution where k is the number of parameters. Similarly, a deviance statistic, $G^2 = 2\sum_{i=1}^{n} y_i log(y_i/\hat{\lambda}_i)$, has been proposed that also follows a χ^2_{n-k} distribution under the null. Additionally, a number of tests have been proposed to test for the specific alternative of overdispersion. One approach is to fit the model to a Negative Binomial distribution, and then test for the absence of dispersion (Lawless, 1987). Dean and Lawless (1989) explore a test for overdispersion by fitting data to a mixed model, where $v_1, ..., v_n$ are i.i.d. random variables such that given X_i and v_i, $Y_i \sim Poisson(v_i\mu_i)$. Assuming that the v_i's have first and second moments and that $E[v_i] = 1$ and $Var(v_i) = \tau$, it follows that $Var(Y_i|X_i) = \mu_i + \tau\mu_i^2$. Thus, the Poisson model can be tested against any extra-Poisson variation via a score test of $H_0 : \tau = 0$ versus $H_a : \tau > 0$.

Beyond these tests, an omnibus goodness-of-fit test is proposed for Poisson regression by Spinelli, Lockhart, and Stephens (2002). This paper is of particular interest because we have extended this procedure to GLMMs. The authors propose a Cramer-von-Mises (CVM) test of fit that makes use of a probability integral transformation.

To perform the test, first estimate the model parameters, such that each Y_i has an associated predicted value $\hat{\lambda}_i$, $i = 1, 2, ..., n$. Then, make the transformation

$$V_i = P_i(Y_i) \text{ where } \begin{cases} P_i(0) = 0 \\ P_i(j) = P(Y_i \leq j-1) = \sum_{k=0}^{j-1} \frac{\hat{\lambda}_i^k e^{-\hat{\lambda}_i}}{k!} \end{cases}$$

It follows that each V_i has a distribution function, that is, $F_i(t) = P(V_i \leq t)$, for $j = 0, 1, ...,$

$$F_i(t) = P_i(j+1), \quad P_i(j) \leq t < P_i(j+1)$$

At this point define $\tilde{F}_n(t)$ as the empirical distribution function (edf) of the set of V_i's. Also, define the average of the estimated distribution functions as $F_{ave}(t) = n^{-1}\sum_{i=1}^{n} F_i(t)$. Then, using the edf and the average function, define the residual process

$$Z_n(t) = \sqrt{n}\{\tilde{F}_n(t) - F_{ave}(t)\}$$

This residual process $Z_n(t)$ is used to calculate the Cramer-von-Mises test statistic:

$$W_n^2 = n\int_0^1 Z_n^2(t)dt.$$

Since both $\tilde{F}_n(t)$ and $F_{ave}(t)$ are step functions, the integral can be evaluated by summing the squared distances between the functions over the steps.

If the model has been correctly specified, the W_n^2 statistic will follow a distribution that is a mixture of χ^2 random variables, whose coefficients are dependent upon data values. However, the null distribution can also be obtained via bootstrap simulation. This Cramer-von-Mises test has also been applied to non-homogeneous Poisson processes (Jeske, Lockhart, Stephens, & Zhang 2008). We will expand this idea to GLMMs in Chapter 7.

1.4 Linear Mixed Models

A linear model that incorporates one or more random effects is known as a linear mixed model (LMM). In contrast to a fixed effect (e.g. the β's in a linear model), a random effect follows a distribution that is governed by one or more parameters. As before, the observations of interest are thought to follow a normal distribution, only now the observations are conditional upon the random effect(s).

Let Y_i, X_i, and β be defined as in equation (1.2). A LMM has the following form:

$$E[Y_i|s] = X_i'\beta + Z_i s$$

where Z_i is the ith row of the random effect design matrix Z and s is a nx1 vector of random effects. Usually, the first and second moments of s are specified or s is assigned a distribution (generally a multivariate normal) such that

$$s \sim (0, D), \text{ where } E[s] = 0 \text{ and } Var(s) = D$$

Define $Var(Y|s) = R$. Then it follows that

$$Y \sim MVN(X\beta, ZDZ' + R).$$

Let $V = ZDZ' + R$ such that $Y \sim MVN(X\beta, V)$. Estimating the parameters in a mixed model is more difficult than in a linear model, since we have β as well as V, which is composed of unknown parameters in D and R. Suppose that V were known, then we can estimate β via generalized least squares where,

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$$

Of course, V is generally unknown, thus an estimate of V must be used in place of V. In order to get a reasonable estimate of V, a maximum likelihood (ML) or a restricted

maximum likelihood (REML) approach is generally taken. The log likelihoods are provided below:

$$L_{ML}(D,R) = -\frac{1}{2}log|V| - \frac{1}{2}r/V^{-1}r - \frac{n}{2}log(2\pi)$$
$$L_{REML}(D,R) = -\frac{1}{2}log|V| - \frac{1}{2}log|X'V^{-1}X| - \frac{1}{2}r/V^{-1}r - \frac{n-p}{2}log(2\pi)$$

where r is a vector of residuals where $r = y - (X'V^{-1}X)^{-1}X'V^{-1}y$ and p is the rank of X. Solutions to these equations can be found by a Newton-Raphson approach or by using the Expectation-Maximization (EM) algorithm, although the Newton-Raphson approach is generally preferred (Lindstrom & Bates, 1988) and used as the default by most statistical software such as SAS (SAS, 2012).

While this approach is successful at estimating β and V, it is sometimes useful to predict the realized values of the random effects s. To obtain a prediction for s, the standard approach is to solve the mixed model equations (Henderson, 1984):

$$\begin{bmatrix} X'\hat{R}^{-1}X & X'\hat{R}^{-1}Z \\ Z'\hat{R}^{-1} & Z'\hat{R}^{-1}Z + \hat{D}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{s} \end{bmatrix} = \begin{bmatrix} X'\hat{R}^{-1}y \\ Z'\hat{R}^{-1}y \end{bmatrix}.$$

The solutions to these equations are:

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$$
$$\hat{s} = \hat{D}Z'\hat{V}^{-1}(y - X\hat{\beta}).$$

1.4.1 Fixed versus Random Effects

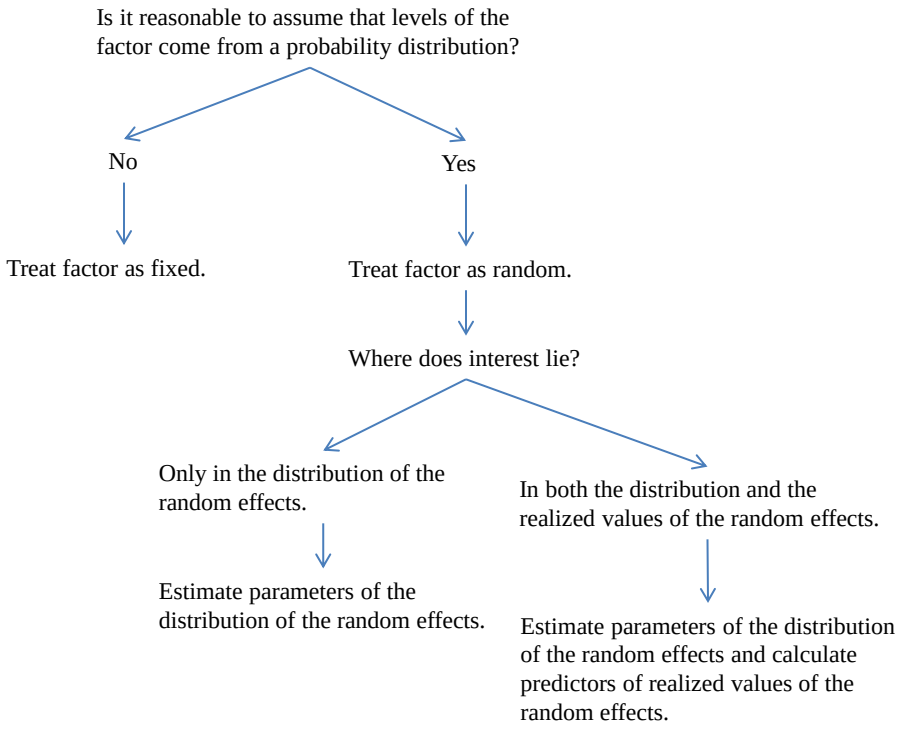
The distinction between fixed and random effects is very important since the analysis and interpretation of fixed and random effects is quite different. Fixed effects can be thought of as levels of a factor that are deliberately chosen to be in a study because they are of interest. For example, suppose we want to study the taste and texture of loaves of bread prepared under the same conditions and then baked at 350, 400 and 450 degrees Fahrenheit. These temperature values were chosen because they are of interest to the researchers and are fixed effects. In the case of fixed effects, we usually want to make direct comparisons between different levels of the effect.

In contrast, random effects are thought to be generated by some underlying random phenomena or process, and the properties of that process (such as the mean or variance) are of interest. For example, in a study of a new drug, patients might be given the new drug

(treatment) or a placebo (control) and monitored every week for four weeks. Some patients will consistently report better results; others will consistently report worse results. Thus, each individual patient's inherent status can be thought of as a random effect that comes from an underlying distribution of patient statuses. This distribution can be described by parameters. Additionally, the realized values of the random effects may be of interest. Calculating these values, which are known as predictors, is addressed in Section 5.4.

McCulloch et al. (2008) provides a useful decision tree for fixed and random effects, which we have recreated in Figure 1.2.

Figure 1.2: A decision tree for fixed and random effects.



1.4.2 Goodness-of-Fit for Linear Mixed Models

Many of the graphical approaches used for linear models can be modified for mixed models to check the normality of the error terms. Calvin and Sedransk (1991) offer two methods for checking the normality assumption of the error terms. The first method consists of premultiplying the responses Y by the inverse of the estimated variance matrix $\hat{V}$ of the response variables. This leads to residuals that are approximately standard normal.

The second approach makes use of best linear unbiased predictors (BLUPs, discussed in Section 5.4.1) of the random effects and then computes residuals of the form $Y - X\hat{\beta} - Z\hat{s}$. However, these residuals are correlated. A similar approach to the first method is proposed by Jacqmin-Gadda, Sibillot, Proust, Molina and Thiebaut (2007). They obtain residuals by multiplying $Y - X\hat{\beta}$ by the Cholesky square root of the covariance matrix, which can then used in a QQ-plot to check for normality

Beyond these graphical approaches, formal tests for normality of the response distribution have been explored. Most approaches attempt to transform the correlated error residuals into uncorrelated residuals, and then apply classical tests for normality (e.g. Shapiro-Wilk test). For example, Hwang and Wei (2006) apply a transformation to a two-stage cluster design. This mixed model has the form

$$Y_{jk} = \mu_j + s_j + e_{jk}, \quad j = 1, ..., m, k = 1, ..., n_j$$

where e_{jk} and s_j are independent random variables with expected values of zero and variances σ_e^2 and σ_s^2, respectively. Assuming normality for the error terms and the random effects, a transformation on Y_{jk} is constructed that results in uncorrelated normal random variables. These can then be tested for univariate normality. However, when this test rejects it is not clear whether the error terms or the random effects are misspecified.

Jiang (2001) also provides an omnibus test of normality for both the random effects and the error terms. The authors constructs a χ^2-like statistic, comparing observed cell counts to the estimated expected cell counts under the null, which are calculated by plugging in REML estimators of the fixed effects and variance components. However, the resulting test statistic does not have an exact χ^2 distribution.

Claskens and Hart (2009) focus on assessing the distributional assumption of the random effects, which are generally assumed to be normally distributed (i.e., $H_0 : s \sim N_d(\mu_s, \Sigma_s)$). The authors use a semi-nonparametric estimator for the distribution of the random effects s. The estimator is based upon a Hermite expansion of the unknown density of s around the standard normal density. First, s is reparameterized, where $s = \mu_s + GU$ and $\Sigma_s = GG'$. It then becomes sufficient to test if $U \sim N_d(0, I)$ where I is the identity matrix. The test statistic is constructed by making use of an Edgeworth expansion of the density of U around the normal density ϕ. For the one dimensional case (i.e., $d = 1$), this looks like

$$f_U(u) = \phi(u)\{1 + k_3 H_3(u) + k_4 H_4(u) + \cdots\} \tag{1.6}$$

where k_3, k_4 are related to the cumulants of U, and the Hermite polynomials satisfy

$H_j(u)\phi(u) = (-1)^j \frac{d^j\phi(u)}{du^j}$. For example, $H_3(u) = u^3 - 3u$ and $H_4(u) = u^4 - 6u^2 + 3$. By reordering the terms in the expansion, the infinite series in (1.6) can be approximated by a semi-nonparametric density,

$$\hat{f}_{U,M}(u) = P_M^2(u)\phi(u) \tag{1.7}$$

where P_M is a d-variable polynomial, such that

$$P_M(u) = \sum_{|\lambda| \leq M} a_\lambda u^\lambda$$

where $\lambda = (\lambda_1, ..., \lambda_d)$, $|\lambda| = \sum_{l=1}^d \lambda_l$, $u^\lambda = u_1^{\lambda_1}, ..., u_d^{\lambda_d}$, and the coefficients a^λ ensure that $\hat{f}_{U,M}$ is equal to 1. The positive integer M is the order of the polynomial. The log-likelihood function can then be written making use of (1.7). If the random effects have a d-variate normal distribution, then $f(Y_i|u,\theta)\phi(u) = g(U|Y_i)g(Y_i|\theta)$, where $g(Y_i|\theta)$ is the marginal density of Y_i under the null and $g(u|Y_i)$ is the conditional density of the random effects, given Y_i. Then, the log-likelihood can be written as

$$\sum_{i=1}^n log\ g(Y_i|\theta) + \sum_{i=1}^n log(E_{u_i|Y_i,\theta}[P_M^2(u_i)]) \tag{1.8}$$

This method offers a closed-form likelihood and ML estimates of the parameters can be found directly, which now include the fixed effects, variance components, and the polynomial coefficients of P_M. An informal test would let M take the values 0,1, or 2, and calculate the AIC for each value of M. It the smallest AIC corresponds to $M \geq 1$, this indicates that a more complex distribution is needed for the random effects. The authors also propose some test statistics that look at the distance between (1.8) and the likelihood under the null.

1.5 Structure of the Dissertation

Thus far, we have provided a brief overview of linear models, generalized linear models and linear mixed models along with some goodness-of-fit methods. The remainder of the dissertation is as follows. Part I of the dissertation contains chapters 2-4 and focuses on goodness-of-fit for Autoregressive Logistic Regression models. Chapter 2 provides some background on ALR models, ALR goodness-of-fit, and discusses a motivating example. Chapter 3 proposes a chi-square goodness-of-fit test, complete with a simulation study. Chapter 4 provides a real life application using a dataset relating to Alzheimer's disease from Loma Linda University in Loma Linda, CA

Part II of the dissertation contains chapters 5-7 and addresses goodness-of-fit for Generalized Linear Mixed Models. Chapter 5 provides a general description of GLMMs and discusses the two main examples we will use for our analysis: the Randomized Clinical Trial model and the Spatial model. Chapter 6 is a review of current goodness-of-fit procedures for GLMMs. Then, in Chapter 7 we will discuss a Cramer-von-Mises based goodness-of-fit procedure and related simulation study. Finally, Chapter 8 summarizes this dissertation and some ideas for future research are examined.

Part I

Goodness-of-Fit for Autoregressive Logistic Regression Models

Chapter 2

Autoregressive Logistic Regression

An autoregressive logistic regression (ALR) model employs a logit link function to relate a binary (or multi-state) response to a linear combination of covariates and past responses. The covariates may be fixed or continually measured over time and past responses are included either by direct plug-in or, when necessary, through the use of dummy variables.

By including past responses, an ALR model can describe the strength of the dependency between repeated measurements on subjects while controlling for other covariates. As there are no measured past responses at the first time point(s) in a series, special care must be taken when handling initial past responses. Although no longer a traditional generalized linear model (GLM) as described in McCulloch et al. (2008), an ALR model can handle longitudinal data with a complex covariance structure.

ALR models are a natural extension of what Bonney (1987) describes as "regressive logistic regression," or logistic regression for a series of dependent binary responses. Subjects are assumed to be independent, while the repeated measures upon subjects are thought to be correlated in some way.

In this chapter we will first discuss logistic regression for dependent binary responses, which provide a natural segue for ALR models, which will be discussed in the second section. The third section deals with current goodness-of-fit procedures for ALR models. The final section discusses a motiving example that will be used to illustrate our goodness-of-fit procedure.

2.1 Foundations for ALR: Logistic Regression for Dependent Binary Responses

Traditional binary logistic regression assumes that the outcome series $(Y_1, Y_2, ..., Y_J)$ are independent observations from a random process which take the values 0 or 1. However, suppose that $(Y_1, Y_2, ..., Y_J)$ is a dependent series of observations, such that past observations are in some way correlated with future observations.

For the case of independent observations, the indexing set $j = 1, 2, ..., J$ is used to collect the observations together. The order of the observations is not important. For example, perhaps J independent subjects were observed, so j serves to relate Y_j back to the jth subject. The same cannot be said for dependent observations. Observations can no longer be treated as interchangeable, as the indexing set provides a fixed, natural ordering between the observations. Thus, we can think of Y_j as being observed at the jth time (or space) point, before Y_{j+1} and after Y_{j-1}. Further, we make the assumption that the time (or distance) between j and $j+1$ is fixed for all j. Although this assumption can sometimes be relaxed, it makes the interpretation of model parameters much more straightforward.

As there are many types of dependencies that could exist for a dataset, several different approaches have been explored to handle this problem, as discussed by Bonney (1987). This expository paper explores the idea of "regressive" logistic regression, or logistic regression for a dependent binary series.

Consider a set of J dependent binary variables $Y = (Y_1, Y_2, ..., Y_J)$ where each Y_j has an associated explanatory variable X_j, $i = 1, ..., J$. Then, the probability of Y given $X = (X_1, X_2, ..., X_J)$ can be decomposed as a product of n conditional probabilities:

$$\begin{aligned} P(Y|X) &= P(Y_1, Y_2, ..., Y_J|X) \\ &= P(Y_1|X_1)P(Y_2|Y_1, X_2) \cdots P(Y_J|Y_1, Y_2, ..., Y_{J-1}, X_J) \end{aligned}$$

A jth logit θ_j can then be defined:

$$\theta_j = log \frac{P(Y_j = 0|Y_1, Y_2, ..., Y_{j-1}, X_j)}{P(Y_j = 1|Y_1, Y_2, ..., Y_{j-1}, X_j)}$$

where θ_j is modeled as a linear combination of past states $Y_1, Y_2, ..., Y_{j-1}$ and X_j. Describing the logit in this way creates a regression where the response Y_j is binary, but the number of explanatory variables changes with j, as shown in Table 2.1(a). This issue can be overcome by instead considering the regression of Y_j on $\{Z_{j1}, Z_{j2}..., Z_{j,j-1}, X_j\}$, such that $Z_{jk} =$

$Z_{jk}(Y_k)$ are known linear functions of Y. Although many definitions are possible, Bonney recommends defining Z in the following manner:

$$Z_{jk} = \begin{cases} 2Y_k - 1, & \text{if } k < j, \\ 0 & \text{if } k \geq j \end{cases} \tag{2.1}$$

where j=1, 2, ...J, and k=1, 2, ..., J-1, such that Z_{jk} takes the values -1, 0, or 1. Additional choices for Z are discussed in Section 2.1.3. Then, the logit can be written

$$\begin{aligned} \theta_j &= \alpha + \gamma_1 Z_{j1} + \gamma_2 Z_{j2} + \cdots + \gamma_{j-1} Z_{j,j-1} + \beta X_j \\ &= \alpha + \gamma_1 Z_{j1} + \gamma_2 Z_{j2} + \cdots + \gamma_{j-1} Z_{j,j-1} + \gamma_j 0 + \cdots + \gamma_{J-1} 0 + \beta X_j \\ &= \alpha + \sum_{k=1}^{J-1} \gamma_k Z_{jk} + \beta X_j \end{aligned} \tag{2.2}$$

where α, β and the γ's are parameters. Table 2.1(a) is now replaced with Table 2.1(b) (or equivalently, 2.1(c)). Thus, our model has been modified into a univariate logistic regression for n independent binary observations with the same set of explanatory variables for each response.

We can also describe model (2.2) using vector notation:

$$\theta = [\theta_1 \ \theta_2 \ \cdots \ \theta_J]',$$

$$\lambda = [\alpha \ \gamma_1 \ \gamma_2 \ \cdots \ \gamma_{J-1} \ \beta]',$$

$$\text{and } A = \begin{bmatrix} 1 & Z_{11} & Z_{12} & \cdots & Z_{1J-1} & X_1 \\ 2 & Z_{21} & Z_{22} & \cdots & Z_{2J-1} & X_2 \\ \vdots & & & & & \\ J & Z_{J1} & Z_{J2} & \cdots & Z_{J,J-1} & X_J \end{bmatrix}$$

such that $\theta = A\lambda$.

Table 2.1: Covariates for regressive logistic regression.

	Response	Explanatory Variables				
(a)	Y_j	Y_1	Y_2	$\cdots$	Y_{J-1}	X_j
	Y_1	$-$	$-$	$\cdots$	$-$	X_1
	Y_2	Y_1	$-$	$\cdots$	$-$	X_2
	Y_3	Y_1	Y_2	$\cdots$	$-$	X_3
	$\vdots$					
	Y_J	Y_1	Y_2	$\cdots$	Y_{J-1}	X_J
(b)	Y_j	Y_1	Y_2	$\cdots$	Y_{J-1}	X_j
	Y_1	0	0	$\cdots$	0	X_1
	Y_2	Z_{21}	0	$\cdots$	0	X_2
	Y_3	Z_{31}	Z_{32}	$\cdots$	0	X_3
	$\vdots$					
	Y_J	Z_{J1}	Z_{J2}	$\cdots$	$Z_{J,J-1}$	X_J
(c)	Y_j	Y_1	Y_2	$\cdots$	Y_{J-1}	X_j
	Y_1	Z_{11}	Z_{12}	$\cdots$	Z_{1J-1}	X_1
	Y_2	Z_{21}	Z_{22}	$\cdots$	Z_{2J-1}	X_2
	Y_3	Z_{31}	Z_{32}	$\cdots$	Z_{2J-1}	X_3
	$\vdots$					
	Y_J	Z_{J1}	Z_{J2}	$\cdots$	$Z_{J,J-1}$	X_J

2.1.1 Maximum Likelihood Estimation

The joint likelihood for model (2.2) is given by

$$L(\lambda) = \prod_{j=1}^{J} (\frac{1}{1+e^{\theta_j}})^{Y_j} (\frac{e^{\theta_j}}{1+e^{\theta_j}})^{1-Y_j}$$

where $\lambda = (\alpha, \gamma_1, \gamma_2, ..., \gamma_{J-1}, \beta)$. Note that in its current form, maximum likelihood estimates for the parameters cannot be found as this model is oversaturated (i.e., there are J observations but $J+1$ parameters). However, this "full" model has many smaller, reduced models that can illustrate a variety of different types of dependencies in the data.

2.1.2 Modeling Equally Predictive Observations

Logistic regression for dependent observations is quite flexible in its ability to specify a number of different types of dependencies between the observations. For instance, it may be the case that past states $Y_1, Y_2..., Y_{j-1}$ have equal and additive predictive effects on a

future state Y_j, i.e.,

$$\gamma_1 = \gamma_2 = \cdots = \gamma_{J-1} = \gamma.$$

Let $S_m = Z_1 + Z_2 + \cdots + Z_m$ denote the mth partial sum of the Z's as defined in (2.1), and let $S_0 = 0$. Then, a model with equally predictive observations would take the form

$$\theta_j = \alpha + \gamma S_{j-1} + \beta X_j,\ j = 1, ..., J. \tag{2.3}$$

The matrices A and λ can be easily modified to reflect this change.

Further, it may be the case that past successes or failures, respectively, have a different but equally predictive effect on future responses. Define S_m^+ to be the number of 1's among the first m outcomes, likewise let S_m^- to be the number of 0's among the first m outcomes, where $S_0^+ = S_0^- = 0$. Then, $S_m = S_m^+ - S_m^-$. A model with equally predictive, but separate, effects for successes and failures is given by

$$\theta_j = \alpha + \gamma^+ S_{j-1}^+ + \gamma^- S_{j-1}^- + \beta X_j,\ j = 1, ..., J.$$

A regression on just the cumulative sum of preceding successes or failures can be specified by $\gamma^- = 0$ or $\gamma^+ = 0$, respectively. Thus, Z is very useful to specify a great variety of model relationships. In addition to equally predictive observations, serial correlation can be described in a logistic regression model for dependent outcomes.

2.1.3 The Choice of Z

Bonney recommends that past states be addressed by the variable Z as defined in (2.1). Bonney also mentions that other definitions for Z are possible, although the paper does not provide any. The advantage of Bonney's Z is that it lends itself to an straightforward interpretation of the logit equations associated with Model (2.2). For instance, if past state $Y_k = 1$ $(k < j)$, this increases the odds of $Y_j = 0$ by e^{γ_k}, while if $Y_k = 0$ the odds are decreased by e^{γ_k}. Further, Z can form cumulative sums, which can be used when fitting models with equally predictive outcomes.

Z should be chosen thoughtfully in order to accurately express the model relationship of interest. Bonney's Z is quite flexible to to describe a variety of model relationships. However, one flaw of the Z's as currently defined is that there are no Z's to represent past states prior to Y_1. Further, as there are only two possible states in a binary series, it is not always necessary to use a transformation on the past states. For instance, when modeling

serial correlation on a binary series, it makes sense just to plug in the past state rather than first transforming the past state.

2.1.4 Serially Dependent Observations and the Initial Stage Problem

Serial dependence can also be specified in a logistic regression for dependent observations. For instance, serial dependence of order 1 implies that future responses are conditional upon the most recent past response, i.e.,

$$P(Y|X) = \prod_{j=1}^{J} P(Y_j|Y_{j-1}, X_j).$$

This can be represented in the model

$$\theta_j = \alpha + \gamma Y_{j-1} + \beta X_j, \ j = 1, ..., J.$$

This type of model is also referred to as a 1-lag model, since we include 1 past response in the model equation. Note that this model does not quite land under the umbrella of model (2.2) which does not regress upon any past states beyond Y_1. Here, we face an issue of how to deal with the lagged response of the initial observation, which we can call Y_0. Before the first observation Y_1, there are no recorded observations. This does not necessarily mean that Y_0 did not exist nor that it does not have predictive power for future observations.

Depending upon the process under study, different approaches can be taken to handle this "initial stage problem." In some cases, Y_1 can be treated as a given constant. As such, the likelihood would lose θ_1. If cyclic conditions are appropriate, we can replace $P(Y_1|Y_0, X_1)$ with $P(Y_1|Y_J, X_1)$, adjusting θ_1 appropriately. Another approach, which we recommend, is to set up a tiered system of logits. For a 1-lag model this would look like:

$$\theta_1 = \alpha_1 + \beta_1 X_1$$
$$\theta_j = \alpha + \gamma Y_{j-1} + \beta X_j, \ j = 2, ..., J.$$

Here, a separate (regular) logistic regression is set up for the first time point. If only one series of observations is available, it would not be possible to estimate α_1 and β_1 from a single observation. However, it is often the case than multiple dependent series are available, so estimation is possible. For serial dependence of higher orders, we can easily extend the model by incorporating additional γ parameters in the logits. A model with serial dependence of order D would include D lags into the model.

Many specialized patterns of dependence beyond what we have described here are possible (see Bonney, 1987; Cox, 1970). Modeling dependent binary series using logistic regression allows for great flexibility to describe the nature of the relationship between dependent outcomes. Interactions between variables can also be easily specified. More than one X covariate could be included. This now becomes an issue of variable selection.

2.1.5 Comparing Different Dependencies

A likelihood ratio test can be used to test independence against a specified pattern of dependence. Note that for any pattern of dependence, independence can be represented by a reduced model (i.e., $\forall\gamma = 0$). For example, in model (2.2) independence corresponds to $\gamma_1 = \cdots = \gamma_{n-1} = 0$. In the case of non-nested models, AIC or BIC can be used to compare different models.

This section has looked at a single series of dependent observations. However, it is often the case that a group of independent subjects might each have their own series of dependent observations. This situation is addressed by autoregressive logistic regression models. In the next section, we will address ALR models with a binary response.

2.2 Binary Autoregressive Logistic Regression

All of the dependencies that are possible for a single series of observations can similarly be expressed for multiple independent series of dependent observations. Here we will focus primarily on ALR models with lagged variables.

Define Y_{ij} be the response of the ith subject at the jth time point, $i = 1, ..., n$, $j = 1, ..., J_i$, where responses are independent between subjects. Thus, we obtain observations as in Table 2.2.

Table 2.2: Schematic dataset for an ALR model.

Independent Subjects	Observations
1	$\underbrace{Y_{11}, Y_{12}, ..., Y_{1J_1}}_{\text{Dependent Series}}$
2	$\underbrace{Y_{21}, Y_{22}, ..., Y_{2J_2}}_{\text{Dependent Series}}$
$\vdots$	$\vdots$
n	$\underbrace{Y_{n1}, Y_{n2}, ..., Y_{nJ_n}}_{\text{Dependent Series}}$

Suppose also that $Y_{ij} \sim Bernoulli(p_{ij})$ where $p_{ij} = P(Y_{ij} = 0)$. Then, a binary ALR model with D lags, $D \in \mathbb{Z}^+$, can be defined by the following model equation:

$$log\frac{P(Y_{ij} = 0|Y_{ij-1}, ..., Y_{ij-D}, X_{ij})}{P(Y_{ij} = 1|Y_{ij-1}, ..., Y_{ij-D}, X_{ij})} = \alpha + X_{ij}\beta + \sum_{d=1}^{D} \gamma_d Y_{ij-d}$$

where X_{ij} is a vector of covariates for the ith patient at the jth time point, β is the associated parameter vector, and α is the intercept parameter. Y_{ij-d} represents the past response of the ith patient at the dth lag, $d = 1, ..., D$, and γ_d's are the associated parameters for each lag. This model is quite similar to a binary logistic regression model, with the exception of the $\sum_{d=1}^{D} \gamma_d Y_{ij-d}$ term.

For simplicity, from now on we will write

$$\theta_{ij}^{D} \text{ in place of } log\frac{P(Y_{ij} = 0|Y_{ij-1}, ..., Y_{ij-D}, X_{ij})}{P(Y_{ij} = 1|Y_{ij-1}, ..., Y_{ij-D}, X_{ij})}$$

The ij subscript is necessary to reflect that θ is the log odds of probabilities related to Y_{ij}, and the D superscript indicates that the logit is conditional on past observations $Y_{ij-1}, Y_{ij-2}, ..., Y_{ij-D}$.

Now suppose D=1, such that only the previous response is taken into account. Then, a binary ALR model with 1-lag might be given by the following:

$$\theta_{ij}^{1} = \alpha + X_{ij}\beta + \gamma_1 Y_{ij-1} \tag{2.4}$$

This logit is conditional on the most recent previous state and γ_1 represents the strength of the dependency between adjacent responses. Here we directly plug in the past responses. Thus, if $Y_{ij-1} = 1$ this would indicate an increase in the odds by e^{γ_1}. While if $Y_{ij-1} = 0$, essentially e^{γ_1} would not contribute to the odds.

However, equation (2.4) comes with a caveat in that there are no previously observed responses before the first time point, i.e., $Y_{10}, Y_{20}, ..., Y_{n0}$ are all unobserved but used in the logit equation. Depending upon the process under study, several different approaches are possible, as discussed in Section 2.1.4. We propose replacing the first logit with a logistic regression model dependent upon X_{ij}, such that

$$\theta_{i1} = \alpha_0 + X_{i1}\beta$$
$$\theta^1_{ij} = \alpha_1 + X_{ij}\beta + \gamma_1 Y_{ij-1} \text{ for } j \geq 2$$

where θ_{i1} is a logit conditional only upon X_{i1}, i.e., $\theta_{i1} = logit(P(Y_{i1} = 0|X_{i1}))$.

2.2.1 Binary ALR Model with D-Lags

Similarly, a Binary ALR Model with D-lags can be described using a set of conditional logits, incorporating as many lags as possible at each subsequent time point. For instance, at $j = 1$ no past observations exist to be incorporated. At $j = 2$ one past observation, Y_{i1}, can be observed, at $j = 3$ up to 2 past observations can be observed for each subject, and so on. Keeping this in mind, logits can be constructed to absorb as much past information as possible at each subsequent time point.

For example, if D=3 lags, then there would be four distinct logits, one each for $j = 1, 2, 3$ and one to represent $j \geq 4$.

$$\theta_{i1} = \alpha_0 + X_{ij}\beta$$
$$\theta^1_{i2} = \alpha_1 + X_{i2}\beta + \gamma_{11} Y_{ij-1}$$
$$\theta^2_{i3} = \alpha_2 + X_{i3}\beta + \gamma_{21} Y_{ij-1} + \gamma_{22} Y_{ij-2}$$
$$\theta^3_{ij} = \alpha + X_{ij}\beta + \gamma_1 Y_{ij-1} + \gamma_2 Y_{ij-2} + \gamma_3 Y_{ij-3} \text{ for } j \geq 4$$

Of course, to incorporate $D = 3$ lags into a model, some of the subjects must have observations that cover at least four time points ($J_i \geq 4$). A model with D-lags can be written as:

$$\begin{array}{llll} \theta_{ij} & = \alpha_0 & +X_{ij}\beta & \\ \theta^{j-1}_{ij} & = \alpha_{j-1} & +X_{ij}\beta \quad +\sum_{d=1}^{j-1} \gamma_{j-1,d} Y_{ij-d} & \text{for } j = 2, ..., D-1 \\ \theta^{D}_{ij} & = \alpha & +X_{ij}\beta \quad +\sum_{d=1}^{D} \gamma_{d} Y_{ij-d} & \text{for } j \geq D \end{array}$$

2.2.2 Maximum Likelihood Estimation

The likelihood for any of the binary lagged models mentioned is given by:

$$L(\lambda) = \prod_{i=1}^{n} \prod_{j=1}^{J_i} \Big(\frac{1}{1 + e^{\theta_{ij}^{min\{j-1,D\}}}} \Big)^{Y_{ij}} \Big(\frac{e^{\theta_{ij}^{min\{j-1,D\}}}}{1 + e^{\theta_{ij}^{min\{j-1,D\}}}} \Big)^{1-Y_{ij}}$$

Maximum likelihood estimates can then be found using the usual methods.

2.2.3 A Bayesian Mixture Model for Finding Parameter Estimates

Beyond using the tiered model, Chan (2000) discusses a Bayesian approach of modeling Y_{i0} by using a Beta prior. The resulting model is a two-point mixture model such that the outcome $Y_{i0} = k$ occurs with probability π_k, $k = 0, 1$ and $\pi_0 + \pi_1 = 1$. A simulation study found that the mixture model had more precise parameter estimates, with smaller mean square errors and relative biases for most of the covariates. The improvement is greatest for those datasets where patients have been observed for shorter lengths of time.

2.3 Multinomial Autoregressive Logistic Regression

Now suppose Y_{ij} can take one of K values (say, 0, 1, ..., K-1) at the jth time point such that $Y_{ij} \sim$ indep. $Multinomial(p_{ij0}, ..., p_{ijK-1})$ and that there is a dependency between responses for D lags. Much like multinomial logistic regression, we require multiple logits to represent this relationship. Also, because there are more than two possible past states, we can no longer directly plug-in past states into the model equations and must use some dummy variables. To explore these issues, let's first look at an multinomial ALR with three possible states and 1-lag.

2.3.1 A Multinomial ALR with 1-Lag that takes Three Possible States

Suppose Y_{ij} takes the values 0,1, and 2 and that $Y_{ij} \sim$ indep. $Multinomial(p_{ij0}, p_{ij1}, p_{ij2})$. Further, define $Z_{ij}(d) = [Z_{ij1}(d), Z_{ij2}(d)]$, $1 \leq d \leq D$ denote the vectors of dummy variables that encode the previous d lagged states of the ith subject at their jth visit. Table 2.3 illustrates the case of D=3 lags. To interpret this table, it helps to think of state 0 corresponding to the pair (0,0), state 1 corresponding to (1,0), and state 2 corresponding to (0,1). Thus, if a past state takes the value 1, then, its dummy variables Z_{ij1} and Z_{ij2} would take the two values in the pair, 1 and 0, respectively.

Table 2.3: Dummy variables for previous D=3 lagged states.

Time(j)	Y_{ij}	$Z_{ij1}(1)$	$Z_{ij2}(1)$	$Z_{ij1}(2)$	$Z_{ij2}(2)$	$Z_{ij1}(3)$	$Z_{ij2}(3)$
1	0	-	-	-	-	-	-
2	1	0	0	-	-	-	-
3	1	1	0	0	0	-	-
4	2	1	0	1	0	0	0
5	0	0	1	1	0	1	0
6	2	0	0	0	1	1	0

A multinomial 1-lag ALR model is given by the following equations:

$$log\big(\frac{P(Y_{ij}=0|Y_{ij-1})}{1-P(Y_{ij}=0|Y_{ij-1})}\big) = \alpha_0 + X_{ij}\beta + \gamma_{11}Z_{ij1}(1) + \gamma_{12}Z_{ij2}(1) \tag{2.5}$$
$$log\big(\frac{P(Y_{ij}\leq 1|Y_{ij-1})}{1-P(Y_{ij}\leq 1Y_{ij-1})}\big) = \alpha_1 + X_{ij}\beta + \gamma_{11}Z_{ij1}(1) + \gamma_{12}Z_{ij2}(1)$$

Note that because we have three possible states, there must be two logit equations. Here the logits are ordinal, and the parameters are the same in both equations with the exception of the intercepts. If ordinal logits are not appropriate for the dataset, a nominal logit model could instead be set up with respect to a reference probability. Here, let's have the reference probability be $P(Y_{ij}=2)$.

$$log\big(\frac{P(Y_{ij}=0|Y_{ij-1})}{P(Y_{ij}=2|Y_{ij-1})}\big) = \alpha_0 + X_{ij}\beta_0 + \gamma_{11}Z_{ij1}(1) + \gamma_{12}Z_{ij2}(1)$$
$$log\big(\frac{P(Y_{ij}=1|Y_{ij-1})}{P(Y_{ij}=2|Y_{ij-1})}\big) = \alpha_1 + X_{ij}\beta_1 + \gamma_{11}Z_{ij1}(1) + \gamma_{12}Z_{ij2}(1)$$

The choice between ordinal and nominal logits is left to the practitioner. Refer to Section 1.3.1 for further discussion.

Again, we now must confront the issue of how to deal with those previous, unobserved states. As before, we can set up a tiered set of conditional logits, thus model (2.5) is rewritten as:

$$\text{If } j=1, \quad log\big(\frac{P(Y_{i1}=0|Y_{i1-1})}{1-P(Y_{i1}=0)}\big) = \alpha_{10} + X_{i1}\beta$$
$$log\big(\frac{P(Y_{i1}\leq 1)}{1-P(Y_{i1}\leq 1)}\big) = \alpha_{11} + X_{ij}\beta$$
$$\text{If } j\geq 2, \quad log\big(\frac{P(Y_{ij}=0|Y_{ij-1})}{1-P(Y_{ij}=0|Y_{ij-1})}\big) = \alpha_{20} + X_{ij}\beta + \gamma_{11}Z_{ij1}(1) + \gamma_{12}Z_{ij2}(1)$$
$$log\big(\frac{P(Y_{ij}\leq 1|Y_{ij-1})}{1-P(Y_{ij}\leq 1|Y_{ij-1})}\big) = \alpha_{21} + X_{ij}\beta + \gamma_{11}Z_{ij1}(1) + \gamma_{12}Z_{ij2}(1).$$

Using these model equations, we can set up a likelihood equation.

2.3.2 Maximum Likelihood Estimation for a Three State ALR with 1-Lag

Using the properties of logits, we can obtain probabilities from the logit equations for the first time point:

$$
\begin{aligned}
p_{i10} &= P(Y_{i1} = 0|X_{i1}) = \frac{e^{\alpha_{10}+X_{i1}\beta}}{1+e^{\alpha_{10}+X_{i1}\beta}} \\
p_{i11} &= P(Y_{i1} = 1|X_{i1}) = \frac{e^{\alpha_{11}+X_{i1}\beta}}{1+e^{\alpha_{11}+X_{i1}\beta}} - \frac{e^{\alpha_{10}+X_{i1}\beta}}{1+e^{\alpha_{10}+X_{i1}\beta}} \\
p_{i12} &= P(Y_{i1} = 2|X_{i1}) = \frac{1}{1+e^{\alpha_{11}+X_{i1}\beta}}
\end{aligned}
$$

and for all subsequent time points:

$$
\begin{aligned}
p_{ij0} &= P(Y_{ij} = 0|Y_{ij-1}, X_{ij}) = \frac{e^{\alpha_{20}+X_{ij}\beta+\gamma_{11}Z_{ij1}(1)+\gamma_{12}Z_{ij2}(1)}}{1+e^{\alpha_{20}+X_{ij}\beta+\gamma_{11}Z_{ij1}(1)+\gamma_{12}Z_{ij2}(1)}} \\
p_{ij1} &= P(Y_{ij} = 1|Y_{ij-1}, X_{ij}) = \frac{e^{\alpha_{21}+X_{ij}\beta+\gamma_{11}Z_{ij1}(1)+\gamma_{12}Z_{ij2}(1)}}{1+e^{\alpha_{21}+X_{ij}\beta+\gamma_{11}Z_{ij1}(1)+\gamma_{12}Z_{ij2}(1)}} \\
&\qquad - \frac{e^{\alpha_{20}+X_{ij}\beta+\gamma_{11}Z_{ij1}(1)+\gamma_{12}Z_{ij2}(1)}}{1+e^{\alpha_{20}+X_{ij}\beta+\gamma_{11}Z_{ij1}(1)+\gamma_{12}Z_{ij2}(1)}} \\
p_{ij2} &= P(Y_{ij} = 2|Y_{ij-1}, X_{ij}) = \frac{1}{1+e^{\alpha_{21}+X_{ij}\beta+\gamma_{11}Z_{ij1}(1)+\gamma_{12}Z_{ij2}(1)}}
\end{aligned}
$$

where $P(Y_{ij} = 0) + P(Y_{ij} = 1) + P(Y_{ij} = 2) = 1$ for all i, j. Then, a likelihood equation can be set up:

$$
P(\lambda) = \prod_{i=1}^{n}\prod_{j=1}^{J_i}(p_{ij0})^{I[Y_{ij}=0]}(p_{ij1})^{I[Y_{ij}=1]}(p_{ij2})^{I[Y_{ij}=2]}
$$

where λ is a vector of parameters that includes $(\alpha_{10}, \alpha_{11}, \alpha_{20}, \alpha_{21}, \gamma_{11}, \gamma_{12})$ and the β's and $I[\cdot]$ is the indicator function.

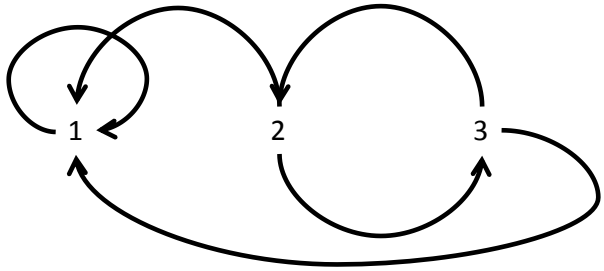
2.3.3 A General Multinomial ALR Model with D-Lags

An ALR model with three or more states and D-lags can be described by expanding upon the ideas discussed previously. If observations can take one of K possible states, then $K-1$ logits are required at each time point until the logit can be conditioned on all lags. If a model has D lags, then $D+1$ tiers of logits are needed. So, a multinomial ALR model with D lags that takes K possible states would require $(K-1)(D+1)$ unique logits. Further, if there are K possible past states, then $K-1$ dummy variables representing past states would need to be incorporated into the model.

2.3.4 Absorbing States

As subjects move across states over time, it sometimes appropriate to have one (or more) absorbing states in a model. For example, suppose a response takes three possible states: 1, 2, or 3, where 1 is an absorbing state, which perhaps corresponds to a subject dropping out of the study. An illustration of this situation is given in Figure 2.1. Absorbing states can be included in an ALR model by modifying the logit equations such that $P(\text{being in the aborbing state} \mid \text{previously entering absorbing state}) = 1$.

Figure 2.1: Three possible response states: 1, 2, or 3, where 1 is an absorbing state.



2.4 Current Goodness-of-Fit Diagnostics for ALR Models

ALR models are applicable in a number of fields where longitudinal data is abundant (de Vries, Fidler, Kuipers & Hunick, 1998; Mueller, Voelkle, & Hattrup, 2011; Slud & Kedem, 1994). In particular, ALR models are very useful for modeling chronic disease status, where patients may fluctuate between certain fixed states over time. However, there are very few goodness-of-fit diagnostics for autoregressive logistic regression models. A graphical comparison of observed and predicted marginal probabilities (de Vries et al., 1998) has been used as a measure of goodness-of-fit for ALR models. Additionally, Slud and Kedem (1994) have proposed a test for binary ALR models based upon Schoenfeld residuals.

Herein we propose an omnibus goodness-of-fit test for autoregressive logistic regression models, based upon a Pearson statistic that makes use of patients' unique paths through time. We will initially explore this test for a specific ALR model with fixed, binary covariates.

2.5 Motivating Example: Claudication Paper

Much of our initial inspiration to create a goodness-of-fit procedure came from a 1998 paper "Fitting Multistate Models with Autoregressive Logistic Regression: Supervised Exercise in Intermittent Claudication" (de Vries et al.). In this paper, an ALR model was used to analyze patient responses to a walking therapy program used to treat intermittent claudication, or severe leg pain due to peripheral arterial disease. The model sought to identify which patient characteristics led to success in the walking program, and conversely, which characteristics led to a worsening of symptoms and an inability to proceed in the study. Data was collected from 329 patients over four visits every two months. At each time point, patients were observed as 1 (moderate improvement in symptoms) or 2 (great improvement in symptoms). Some patients also dropped out of the study due to an increase in leg pain, a state denoted by 0. Once a patient dropped out of the study, they could not return. Thus, we might observe patient responses like the ones in Table 2.4.

Table 2.4: Schematic dataset from the claudication paper.

Visit	Subject 1	Subject 2	Subject 3	Subject 4	...	Subject n
1	1	1	2	1	...	2
2	2	0	2	1	...	1
3	2	.	1	2	...	0
4	2	.	1	0	...	.

Other patient variables were recorded at baseline such as age (years), gender, diabetes status (yes/no), smoking status (yes/no), number of symptomatic limbs (1 to 21), season (fall, winter, spring, summer), duration of the disease (months), and thigh/ankle brachial index (ABI, 0 to 1.5). Using Akaike's information criterion as a basis for comparison between different ALR models, an autoregressive logistic regression model with two lags and three covariates (ABI, age, and duration) was selected as the best predictive model. Finally, the authors provided a graphical goodness-of-fit assessment, plotting as a function of time the observed and the predicted fraction of patients with a particular response for three disjoint subsets of the covariates.

Although we were unable to obtain the original data used in the cladication paper, we have constructed an ALR model that is similar in nature. This model will be used to illustrate our general approach to goodness-of-fit for ALR models.

2.5.1 The Model

Suppose we have three response states: -1, 0, and 1, that are ordinal in nature and that -1 is an absorbing state, which corresponds to dropping out of the study. Thus, subjects begin the study either in state 0 or 1. Although the numbers used to describe the states are different from those in the cladication paper, describing them in this way allows a direct plug-in of the past states, since past states can really only take the values 0 or 1. If there were $K \geq 3$ possible past states, then $K-1$ dummy variables would be needed for each lag.

Suppose also there are two binary covariates for each subject, V_i and W_i. Let Y_{it} be the state of the ith patient $i = 1, ..., n$ at the t-th time point, $t = 1, ..., T_i$, where $t = 1$ is a "baseline" state and T_i is the number of scheduled visits for the ith patient. Patients are observed T_i times unless they enter the absorbing state before the T_i-th scheduled visit. It is also assumed that patients have equally spaced visits and no missed visits. An ALR model with two lags can be expressed by the following set of logit equations:

$$
\begin{aligned}
&\text{If t} = 1,\ \text{logit}(P(Y_{i1} = 0)) = \alpha_0 + \alpha_1 V_i + \alpha_2 W_i \\
&\text{If t} = 2,\ \text{logit}(P(Y_{i2} = -1 | Y_{i1} = y_{i1})) = \beta_{01} + \beta_1 V_i + \beta_2 W_i + \beta_3 y_{i1} \\
&\qquad\qquad \text{logit}(P(Y_{i2} \leq 0 | Y_{i1} = y_{i1})) = \beta_{02} + \beta_1 V_i + \beta_2 W_i + \beta_3 y_{i1} \\
&\text{If t} \geq 3,\ \text{logit}(P(Y_{it} = -1 | Y_{it-1} = y_{it-1}, Y_{it-2} = y_{it-2})) \\
&\qquad\qquad = \begin{cases} \gamma_{01} + \gamma_1 V_i + \gamma_2 W_i + \gamma_3 y_{it-1} + \gamma_4 y_{it-2} & \text{if } \{y_{it-1}, y_{it-2}\} \in \{0, 1\} \\ \infty & \text{otherwise} \end{cases} \\
&\qquad\qquad \text{logit}(P(Y_{it} \leq 0 | Y_{it-1} = y_{it-1}, Y_{it-2} = y_{it-2})) \\
&\qquad\qquad = \begin{cases} \gamma_{02} + \gamma_1 V_i + \gamma_2 W_i + \gamma_3 y_{it-1} + \gamma_4 y_{it-2} & \text{if } \{y_{it-1}, y_{it-2}\} \in \{0, 1\} \\ \infty & \text{otherwise} \end{cases}
\end{aligned}
\tag{2.6}
$$

where $\beta_{01} \leq \beta_{02}$ and $\gamma_{01} \leq \gamma_{02}$. Here ordinal logits are used, as discussed in Section 1.3.2. We can do this because our response is ordinal in nature, but this is only recommended if a proportional odds relationship is reasonable (Agresti, 2003). An advantage of using ordinal logits is that there are generally fewer parameters versus using nominal logits. For instance, model (2.6) has a total of 14 parameters. An equivalent nominal logit model would have 21 parameters. Let θ represent the vector of model parameters for $t \geq 2$, such that

$$
\theta = (\beta_{01}, \beta_{02}, \beta_1, \beta_2, \beta_3, \gamma_{01}, \gamma_{02}, \gamma_1, \gamma_2, \gamma_3, \gamma_4). \tag{2.7}
$$

Chapter 3

A Chi-Square Test for Autoregressive Logistic Regression

The previous chapter described the claudication paper, which provides an example that will both motivate and highlight our test procedure. Although we were unable to obtain the original data used in the paper, we have constructed an ALR model that is similar in nature. This model will be used to illustrate our general approach to goodness-of-fit for ALR models.

3.1 A Goodness-of-Fit Statistic that Makes Use of Unique Paths

An omnibus goodness-of-fit procedure will test

H_o: the ALR model is a good fit versus H_a: the ALR model is a poor fit

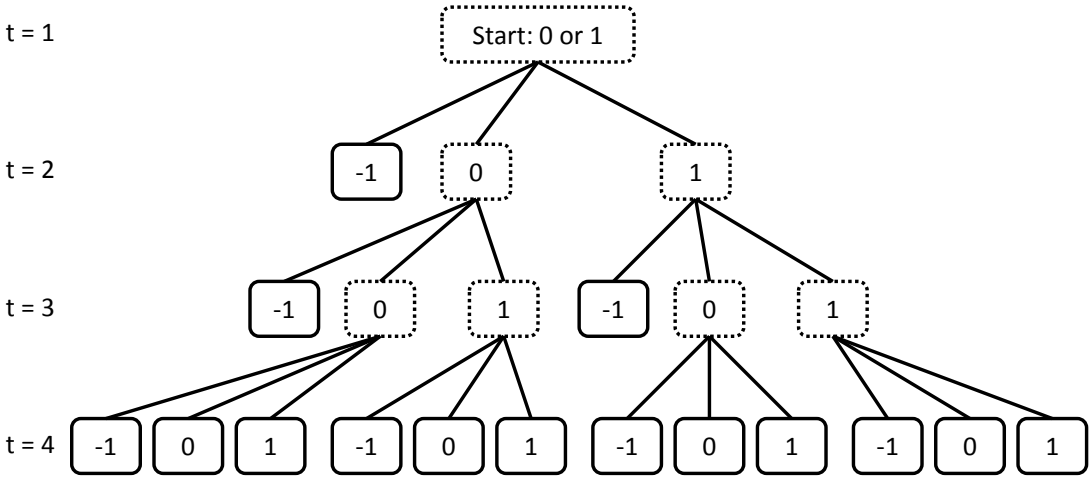
The construction of an appropriate statistic for this hypothesis and its asymptotic distribution can be described for model (2.6), and is based upon tabulating unique path probabilities. The procedure can be adapted to other ALR models.

3.1.1 Unique Paths

Suppose the number of scheduled visits is $T_i = 4$ for all subjects in model (2.6). Recall that there are three ordinal states -1, 0, and 1 where -1 is an absorbing state. Then, for a

given start state (i.e., 0 or 1), there are only 15 unique paths a subject might take, as seen in Figure 3.1. Each subject can follow only one possible path.

Figure 3.1: The number of unique paths is equal to 15 (solid rectangles) given a Start state when the number of scheduled visits is 4 and there are three possible states: -1, 0, and 1, where -1 is an absorbing state.



Although we can model $P(Y_{i1} = 0)$ (and thus $P(Y_{i1} = 1)$) at the first time point, start state is in some sense conditional, i.e., patients "appear" in the study with some start value, S_i (either 0 or 1). Likewise, patients have some fixed covariate values V_i (0 or 1) and W_i (0 or 1). For our model, there are $2 \cdot 2 \cdot 2 = 8$ combinations of S, V, and W. Each combination represents a cohort of patients with similar baseline characteristics.

Define $p_\theta(u|SVW) = P_\theta(\text{taking the uth path}|\text{ start state S and covariates V, W})$, $u = 1, ..., 15$, $S = 0, 1$, $V = 0, 1$, and $W = 0, 1$. Thus, there are eight disjoint groups of patients at baseline and each group has a set of 15 conditional path probabilities, as illustrated in Table 3.1.

Table 3.1: All possible (S, V, W) combinations and associated conditional path probabilities.

Start	V	W	Conditional Probabilities
0	0	0	$p_\theta(1\|000), ..., p_\theta(15\|000)$
0	0	1	$p_\theta(1\|001), ..., p_\theta(15\|001)$
0	1	0	$p_\theta(1\|010), ..., p_\theta(15\|010)$
0	1	1	$p_\theta(1\|011), ..., p_\theta(15\|011)$
1	0	0	$p_\theta(1\|100), ..., p_\theta(15\|100)$
1	0	1	$p_\theta(1\|101), ..., p_\theta(15\|101)$
1	1	0	$p_\theta(1\|110), ..., p_\theta(15\|110)$
1	1	1	$p_\theta(1\|111), ..., p_\theta(15\|111)$

These conditional probabilities can be found from the logit equations in model (2.6). For example, suppose u represents the path $S \to 0 \to 1 \to 1$. Then,

$$\begin{aligned}
p_\theta(u|SVW) &= P_\theta(Y_{i4} = 1, Y_{i3} = 1, Y_{i2} = 0|S, V, W) \\
&= P_\theta(Y_{i4} = 1|Y_{i3} = 1, Y_{i2} = 0, S, V, W) \cdot P_\theta(Y_{i3} = 1|Y_{i2} = 0, S, V, W) \\
&\quad \cdot P_\theta(Y_{i2} = 0|S, V, W) \\
&= P_\theta(Y_{i4} = 1|Y_{i3} = 1, Y_{i2} = 0, V, W) \cdot P_\theta(Y_{i3} = 1|Y_{i2} = 0, S, V, W) \\
&\quad \cdot P_\theta(Y_{i2} = 0|S, V, W) \\
&= \frac{1}{1 + e^{\gamma_{02} + \gamma_1 V + \gamma_2 W + \gamma_3}} \cdot \frac{1}{1 + e^{\gamma_{02} + \gamma_1 V + \gamma_2 W + \gamma_4 S}} \\
&\quad \cdot \left(\frac{e^{\beta_{02} + \beta_1 V + \beta_2 W + \beta_3 S}}{1 + e^{\beta_{02} + \beta_1 V + \beta_2 W + \beta_3 S}} - \frac{e^{\beta_{01} + \beta_1 V + \beta_2 W + \beta_3 S}}{1 + e^{\beta_{01} + \beta_1 V + \beta_2 W + \beta_3 S}} \right)
\end{aligned}$$

To get an estimate of $p_\theta(u|SVW)$, plug in an estimate for the parameters $\hat{\theta}$, giving $p_{\hat{\theta}}(u|SVW)$.

3.1.2 The Construction of the Statistic

Suppose n_{SVW} subjects belong to each (S, V, W) group. We can construct a chi-square statistic in the following manner:

1. Estimate the expected counts for each path,

$$e_{\hat{\theta}}(u|SVW) = n_{SVW} \cdot p_{\hat{\theta}}(u|SVW) \tag{3.1}$$

2. For each (S, V, W) sort and bin the 15 expected counts, obtaining bins $B_1, ..., B_{K_{SVW}}$ where the estimated value of the kth bin, $k = 1, ..., K_{SVW}$ is

$$b_{\hat{\theta}}(k|SVW) = n_{SVW} \cdot s_{\hat{\theta}}(k|SVW) \tag{3.2}$$

where

$$s_{\hat{\theta}}(k|SVW) = \sum_{u=1}^{15} I_k\{u\} p_{\hat{\theta}}(u|SVW) \tag{3.3}$$

and $I_k\{u\} = 1$ if uth path $\in$ kth bin, 0 otherwise.

3. After sorting and binning, let

$$o(k|SVW) = (\#\text{ of observed paths out of } n_{SVW} \text{ in the kth bin}) \tag{3.4}$$

such that

$$X^2_{SVW} = \sum_{k=1}^{K_{SVW}} \frac{(o(k|SVW) - b_{\hat{\theta}}(k|SVW))^2}{b_{\hat{\theta}}(k|SVW)} \tag{3.5}$$

4. Then, the chi-square statistic for the whole dataset is

$$X^2 = \sum_{S=0}^{1} \sum_{V=0}^{1} \sum_{W=0}^{1} \chi^2_{SVW} \tag{3.6}$$

which under H_0 and some regularity conditions will approximately follow a χ^2 distribution with degrees of freedom

$$df = \sum_{S=0}^{1} \sum_{V=0}^{1} \sum_{W=0}^{1} (K_{SVW} - 1) - (\text{some adjustment for parameter estimation})$$

.

3.1.3 Distribution of the Chi-Square Statistic under the Null Hypothesis

The chi-square statistic could be constructed by plugging in the standard "raw data" maximum likelihood parameter estimates, denoted by $\hat{\theta}_{RD}$. These can be found by maximizing the joint likelihood, which can be written by expanding upon the definition of conditional probability:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} L_i(\theta) \\ &= \prod_{i=1}^{n} P_\theta(Y_{iT_i} = y_{iT_i}, ..., Y_{i2} = y_{i2} | S_i, V_i, W_i) \\ &= \prod_{i=1}^{n} P_\theta(Y_{iT_i} = y_{iT_i} | Y_{iT_{i-1}} = y_{iT_{i-1}}, Y_{iT_{i-2}} = y_{iT_{i-2}}, V_i, W_i) \\ &\quad \cdot P_\theta(Y_{iT_{i-1}} = y_{iT_{i-1}} | Y_{iT_{i-2}} = y_{iT_{i-2}}, Y_{iT_{i-3}} = y_{iT_{i-3}}, V_i, W_i) \\ &\quad \cdots P_\theta(Y_{i2} = y_{i2} | S_i, V_i, W_i). \end{aligned} \tag{3.7}$$

However, the resulting degrees of freedom is difficult to characterize (Chernoff & Lehmann, 1954). In fact, the general rule that the degrees of freedom is the "number of bins minus one, minus the number of estimated parameters" applies only when the parameter estimates are found by maximizing a product of multinomial pmfs based upon the binning structure. We can solve for these "bin-based" parameter MLEs, denoted by $\hat{\theta}_B$, by maximizing the following:

$$L(\theta) = \prod_{S=0}^{1} \prod_{V=0}^{1} \prod_{W=0}^{1} \prod_{k=1}^{K_{SVW}} s_\theta(k|SVW)^{o(k|SVW)} \tag{3.8}$$
$$\longrightarrow \hat{\theta}_B = (\hat{\beta}_{01}, \hat{\beta}_{02}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\gamma}_{01}, \hat{\gamma}_{02}, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3, \hat{\gamma}_4)$$

Since our statistic is constructed conditional on the start state S, V, and W, we do not need to estimate the three α parameters associated with $t = 1$. Thus, for our model the degrees of freedom is

$$df = \sum_S \sum_V \sum_W (K_{SVW} - 1) - 11 \tag{3.9}$$

only if the bin-based parameter MLEs are plugged in to construct the statistic. Appropriate parameter estimates could also be found by using minimum chi-square estimators or modified minimum chi-square estimators (Moore, 1986).

Additionally, the bins must be selected in such a way that a positive number of degrees of freedom is obtained while also upholding some regularity conditions. These require that the expected counts within each bin must be sufficiently large. Cochran's chi-square "rule of thumb" addresses these issues (Cochran, 1954). This rule states that all expected bin counts must be at least 1, and that at least 80% of the total bins should have an expected count of 5 or more (i.e., $b_{\hat{\theta}}(k|SVW) \geq 5$). If the expected bin counts are not sufficiently large, the test statistic may not converge to the χ^2 distribution. Research has shown that Cochran's rule is sometimes too conservative under certain conditions (Moore, 1986). A general binning strategy is described in the next section.

3.1.4 A Note on Binning

Bins must be constructed in such a way to obtain a positive number of degrees of freedom while also upholding the regularity conditions. In the case of an ALR model, there may not be enough patients in a particular (S,V,W) combination (i.e., n_{SVW} is small) to justify binning only within that combination. We suggest a two-dimensional binning strategy,

where "rows" of disjoint covariate patterns are set up with corresponding "columns" of expected path counts, which are then tabulated. Table 3.1 illustrates such a two-dimensional table, although unique path probabilities are shown rather than path counts.

Row Binning. For our illustrative model, recall that there are eight distinct rows of (S, V, W) and each row has an associated n_{SVW}. First, check if the smallest n_{SVW} exceeds some row threshold, say 10. This threshold value should be large enough so that small n_{SVW} can be absorbed, but small enough to not force unnecessary merging between large cohorts.

If the smallest count n_{SVW} is less than or equal to the row threshold, sort the counts, then add the two smallest and take the union of their associated (S, V, W) as a new category. Then, repeat the process until all categories exceed the threshold. An example is provided in Table 3.2.

Table 3.2: An example of row binning for $n = 200$ patients with a row threshold= 10. Once the rows have been merged, the next step is to bin the expected path counts within each row.

SVW	n_{SVW}		SVW	n_{SVW}		SVW	n_{SVW}		SVW	n_{SVW}
000	5		000	5		$000 \cup 011$	12		110	7
001	11	sort	011	7	merge	110	7	sort	001	11
010	15	$\rightarrow$	110	7	$\rightarrow$	001	11	$\rightarrow$	$000 \cup 011$	12
011	7		001	11		010	15		010	15
100	62		010	15		111	43		111	43
101	50		111	43		101	50		101	50
110	7		101	50		100	62		100	62
111	43		100	62						

	SVW	n_{SVW}	Conditional Probabilities
	$110 \cup 001$	18	$p_\theta(1\|110 \cup 001), ..., p_\theta(15\|110 \cup 001)$
merge	$000 \cup 011$	12	$p_\theta(1\|000 \cup 011), ..., p_\theta(15\|000 \cup 011)$
$\longrightarrow$	010	15	$p_\theta(1\|010), ..., p_\theta(15\|010)$
	111	43	$p_\theta(1\|111), ..., p_\theta(15\|111)$
	101	50	$p_\theta(1\|101), ..., p_\theta(15\|101)$
	100	62	$p_\theta(1\|100), ..., p_\theta(15\|100)$

Note that it is straightforward to solve for the conditional path probabilities (and thus the estimated counts) given a union of one or more (S,V,W) combinations. For example,

$$\begin{aligned} p_\theta(u|SVW = 000 \text{ or } SVW = 011) &= p_\theta(u|000 \cup 011) \\ &= \frac{P_\theta(u \cap (000 \cup 011))}{P(000 \cup 011)} \\ &= \frac{P_\theta(u \cap 000) + P_\theta(u \cap 011)}{P(000 \cup 011)} \\ &= \frac{p_\theta(u|000)n_{000} + p_\theta(u|011)n_{011}}{n_{000} + n_{011}}. \end{aligned}$$

Column Binning. Once the rows have been merged (if necessary), we can bin the expected counts associated with the paths in each row. First, use the raw data MLEs found by maximizing (3.7) to solve for the expected path counts, i.e., $e_{\hat{\theta}_{RD}}(u|SVW) = n_{SVW} \cdot p_{\hat{\theta}_{RD}}(u|SVW)$. Then, sort these expected counts from largest to smallest. Since our goal is that at least 80% of the bins have an expected value of 5 or more, bin the counts in such a way that each $b_{\hat{\theta}_{RD}}(k|SVW)$ is greater than or equal to some column cutoff that reinforces the rule of thumb, say 5 or larger. If the expected count of the first sorted path is at least 5, it becomes a bin. If the expected count is less than 5, merge it with the next largest and so on, until the expected bin count equals 5 or more. After creating the first bin, the next largest expected counts should be merged (if necessary) to get the next bin. Binning continues in this fashion until all the expected path counts are incorporated into a bin. An example is given in Table 3.3.

By using the raw data MLEs $\hat{\theta}_{RD}$ to construct the bins, we resolve the issue of needing an a priori binning structure to solve for the bin-based MLEs. However, once the binning structure is determined, the expected bin counts for the chi-square statistic are calculated by plugging in the bin-based MLEs $\hat{\theta}_B$ found by maximizing (3.8), obtaining $b_{\hat{\theta}_B}(k|SVW)$. Under the null hypothesis, this statistic will follow a chi-square distribution with degrees of freedom described in (3.9).

Table 3.3: An example of column binning for a particular SVW combination where $n_{SVW} = 30$ patients, with a column cutoff= 5. The expected count for each path is the product of the total row count and the predicted conditional path probability, i.e., $n_{SVW} \cdot p_{\hat{\theta}_{RD}}(u|SVW)$. Here, five bins are generated as observed by the alternating bold/non-bold text in the last rows of the table.

Path	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Expected Count	.50	6	10	.02	1	1	2	.01	1	4	.01	.02	4	.04	.40
							sort ↓								
Path	3	2	10	13	7	5	5	9	1	15	14	4	12	8	11
Expected Count	10	6	4	4	2	1	1	1	.50	.40	.04	.02	.02	.01	.01
							bin ↓								
Path	**3**	2	**10**	**13**	7	5	5	**9**	**1**	**15**	**14**	**4**	**12**	**8**	**11**
Expected Count	**10**	6	**4**	**4**	2	1	1	**1**	**.50**	**.40**	**.04**	**.02**	**.02**	**.01**	**.01**

Both the rows and the columns make use of some minimum threshold or cutoff to guide the construction of bins. However, the rows employ a bottom-up approach, while the columns make use of a top-down approach. For the row binning, the bottom-up approach can quickly identify any smaller rows which need to be merged. Further, this approach ensures that all row counts will exceed the chosen row threshold. In contrast, the column binning top-down strategy ensures that most (as opposed to all) of the bins will exceed the column cutoff value. Such an approach is appropriate since regularity conditions only require that 80% of expected bin counts exceed the value 5 or larger.

3.2 Size and Power Study

The properties of the proposed X^2 statistic are explored in this section for model (2.6). Let $T_i = 4\ \forall i$, such that all patients are seen for four visits unless they have entered the absorbing state previously. As before, patients are assumed to have equally spaced visits with no missed visits. Three sample sizes were selected: $n = 200, 400$ and 2000. Also, two distributions of (S, V, W) combinations, denoted by "Balanced" and "Unbalanced," were simulated as shown in Table 3.4. In particular, the Unbalanced case was selected so that some row merging would have to occur at the smaller sample sizes $n = 200$ and $n = 400$,

where the smallest row combinations contained one and two observations, respectively. Additionally, three sets of parameter values were chosen to exhibit a range of possible relationships among the variables, and are denoted by "Mild", "Moderate", and "Severe," reflecting the expected percent of patients who have dropped out (i.e., entered the absorbing state) by the fourth visit. The parameter values are provided in Table 3.5. This table also displays the percent of patients who are expected to have dropped out under the Balanced and Unbalanced cases. These percentages were found by simulating 5,000 datasets of size $n = 200$ and calculating the average percentage of subjects who had entered the absorbing state by the fourth visit.

Table 3.4: Two distributions of patients at baseline.

Start	V	W	Balanced	Unbalanced
0	0	0	.125	.120
0	0	1	.125	.050
0	1	0	.125	.250
0	1	1	.125	.100
1	0	0	.125	.225
1	0	1	.125	.100
1	1	0	.125	.150
1	1	1	.125	.005

Table 3.5: Parameter values for the simulation study and associated dropout rates for the Balanced and Unbalanced cases by the fourth visit.

Label	Parameter Values	Expected Dropout	
		Balanced	Unbalanced
Mild	$\beta_{01} = -1.8, \beta_{02} = -0.3, \beta_1 = -2.2, \beta_2 = 0.5, \beta_3 = -1.7$ $\gamma_{01} = -0.4, \gamma_{02} = 1.9, \gamma_1 = -2.5, \gamma_2 = -1.2, \gamma_3 = 0.4, \gamma_4 = -1.9$	21.4%	18.4%
Moderate	$\beta_{01} = -3.2, \beta_{02} = -1.7, \beta_1 = 1.5, \beta_2 = 2.5, \beta_3 = -3.1$ $\gamma_{01} = -2.3, \gamma_{02} = -0.3, \gamma_1 = -1.8, \gamma_2 = -1.5, \gamma_3 = -0.5, \gamma_4 = 2.7$	42.5%	45.2%
Severe	$\beta_{01} = -0.9, \beta_{02} = -0.3, \beta_1 = 2.2, \beta_2 = -1.7, \beta_3 = 0.8$ $\gamma_{01} = 0.8, \gamma_{02} = 1.5, \gamma_1 = -0.4, \gamma_2 = 2.1, \gamma_3 = -1.6, \gamma_4 = -0.9$	86.9%	87.1%

3.2.1 Size

To evaluate the size of our proposed test, 10,000 datasets were generated from model (2.6) under different conditions using R software version 2.14.2. Then, the raw data MLEs

$\hat{\theta}_{RD}$ were found for each dataset using the optim() function to minimize the negative log of the joint likelihood given in (3.7).

Using these MLEs to solve for the expected path counts, the binning structure was then dynamically determined using a row threshold of 10 and a column cutoff of 5. Once the bins were set, bin-based MLEs $\hat{\theta}_B$ were found by minimizing the negative log of the joint multinomial likelihood given in (3.8). These bin-based MLEs were plugged into (3.1)-(3.6) to calculate the value of the statistic. For each dataset, a p-value was generated based on the statistic relative to the chi-square distribution with the appropriate degrees of freedom as given in (3.9).

Table 3.6 displays the proportion of p-values out of 10,000 less than or equal to 0.10, 0.05, and 0.01. It appears that the test statistic has adequate size under all of the different settings. In some instances, the size is slightly lower than ideal for the 0.10 case, however since data is being generated from the null hypothesis, this is an acceptable result. Along with size, three power alternatives were looked at and are discussed in the next sections.

Table 3.6: Proportion of p-values out of 10,000 less than or equal to the listed cutoffs.

	n=200		n=400		n=2000	
Mild	Bal	Unb	Bal	Unb	Bal	Unb
≤ 0.01	0.015	0.016	0.013	0.011	0.011	0.011
≤ 0.05	0.050	0.049	0.052	0.047	0.054	0.046
≤ 0.10	0.089	0.087	0.099	0.094	0.103	0.091
Moderate	Bal	Unb	Bal	Unb	Bal	Unb
≤ 0.01	0.015	0.013	0.012	0.012	0.010	0.014
≤ 0.05	0.055	0.049	0.052	0.052	0.050	0.054
≤ 0.10	0.104	0.092	0.100	0.101	0.099	0.102
Severe	Bal	Unb	Bal	Unb	Bal	Unb
≤ 0.01	0.012	0.012	0.015	0.012	0.012	0.013
≤ 0.05	0.046	0.042	0.051	0.046	0.051	0.049
≤ 0.10	0.088	0.082	0.101	0.093	0.101	0.096

3.2.2 Power Alternative: Subjects have Random Intercepts

Here, the alternative of subject random intercepts was explored by generating 1,000 datasets from model (2.6) with the additional feature that each patient had a randomly generated intercept $r_i \sim N(0, \sigma^2)$, for $\sigma = 0.5, 1.0, 1.5$, and 2.0. The test statistic and associated p-value was then calculated using the same procedure as in the size study.

Table 3.7 displays the percentage of p-values out of 1,000 less than or equal to 0.10, 0.05, and 0.01 for different values of σ. As the results from all three of the different parameter sets were quite similar, only the Moderate parameter case is presented in the table. There are two main patterns present in the table. As expected, the power of the test increases as sigma increases. Similarly, we also see an improvement in power as sample size increases.

Table 3.7: Percentage of p-values out of 1,000 less than or equal to the listed cutoffs for a random intercept alternative, moderate parameter case only.

	n=200		n=400		n=2000	
$\sigma = 0.5$	Bal	Unb	Bal	Unb	Bal	Unb
≤ 0.01	1.3%	1.3%	0.6%	1.3%	2.9%	2.1%
≤ 0.05	5.8%	5.2%	5.1%	5.1%	8.8%	7.0%
≤ 0.10	10.0%	9.7%	11.6%	10.0%	16.0%	13.4%
$\sigma = 1.0$	Bal	Unb	Bal	Unb	Bal	Unb
≤ 0.01	2.4%	0.6%	2.5%	2.9%	36.6%	34.5%
≤ 0.05	7.4%	4.4%	9.1%	10.7%	62.1%	61.3%
≤ 0.10	13.7%	10.4%	14.6%	18.0%	72.4%	73.6%
$\sigma = 1.5$	Bal	Unb	Bal	Unb	Bal	Unb
≤ 0.01	4.3%	2.3%	25.8%	4.8%	98.6%	99.1%
≤ 0.05	9.2%	8.5%	47.8%	14.9%	99.5%	99.9%
≤ 0.10	14.2%	15.8%	61.7%	25.4%	99.8%	100%
$\sigma = 2.0$	Bal	Unb	Bal	Unb	Bal	Unb
≤ 0.01	2.8%	1.2%	16.3%	30.5%	100%	100%
≤ 0.05	9.8%	4.2%	37.0%	53.1%	100%	100%
≤ 0.10	16.8%	8.2%	51.1%	66.2%	100%	100%

3.2.3 Power Alternative: Omitted Covariate

Here, model (2.6) includes an additional baseline binary covariate, Z. Values of Z were generated either from a Binomial(0.5) or a Binomial(0.75). To describe the effect of this new covariate, two new parameters β_4 and γ_5 are included in the logit equations corresponding to $t = 2$ and $t \geq 3$, respectively. The power study looked at two sets of parameter values: $\beta_4 = 2.5$, $\gamma_5 = 1.3$ and $\beta_4 = 0.9$, $\gamma_5 = -1.8$. These values were chosen to reflect different relationships between the covariate Z and the response.

Table 3.8 displays the percentage of p-values out of 1,000 less than or equal to 0.10, 0.05, and 0.01 for different values of β_4, γ_5 and Z distributions. The table displays results from

the Moderate parameter case only, as results from the Mild and Severe parameter values were quite similar. As expected, we see a general trend of power increasing as n increases. Also, power seems to be better for the Binomial(0.5) versus the Binomial(0.75) case. The test is not good at detecting the missing covariate at the smaller sample sizes.

Table 3.8: Percentage of p-values out of 1,000 less than or equal to the listed cutoffs for a omitted covariate alternative, moderate parameter case only.

	n=200		n=400		n=2000	
$\beta_4 = 2.5, \gamma_5 = 1.3$, Z∼Bin(0.5)	Bal	Unb	Bal	Unb	Bal	Unb
≤ 0.01	1.9%	2.5%	4.0%	1.9%	30.2%	45.1%
≤ 0.05	6.8%	7.2%	11.1%	8.3%	55.4%	68.1%
≤ 0.10	11.7%	12.0%	18.7%	15.3%	68.5%	78.8%
$\beta_4 = 2.5, \gamma_5 = 1.3$, Z∼Bin(0.75)	Bal	Unb	Bal	Unb	Bal	Unb
≤ 0.01	1.6%	1.2%	1.9%	1.4%	20.1%	28.7%
≤ 0.05	5.5%	5.5%	7.5%	6.1%	41.5%	51.0%
≤ 0.10	10.6%	10.6%	12.8%	12.9%	55.5%	65.5%
$\beta_4 = 0.9, \gamma_5 = -1.8$, Z∼Bin(0.5)	Bal	Unb	Bal	Unb	Bal	Unb
≤ 0.01	1.1%	1.0%	1.5%	1.7%	29.9%	22.6%
≤ 0.05	4.3%	5.3%	7.9%	7.5%	55.2%	47.8%
≤ 0.10	8.2%	9.7%	12.6%	13.9%	68.3%	62.4%
$\beta_4 = 0.9, \gamma_5 = -1.8$, Z∼Bin(0.75)	Bal	Unb	Bal	Unb	Bal	Unb
≤ 0.01	1.2%	1.1%	1.4%	1.5%	15.1%	14.8%
≤ 0.05	5.4%	4.4%	4.6%	6.0%	31.4%	31.7%
≤ 0.10	9.8%	8.9%	9.1%	12.2%	45.2%	44.7%

3.2.4 Power Alternative: Misspecified Lag

One of the features of an ALR model is being able to observe to what degree past responses affect future responses. Here, a simulation was set up to test if the chi-square statistic could detect a need for an additional lag. This was done by fitting a 1-lag model to data that was generated from a 2-lag model.

Here, the 1-lag model corresponds to a modified model (2.6), where the logit equations associated with $t \geq 3$ are absent, and the logit equations for $t = 2$ now correspond to all time points $t \geq 2$. For each of the parameter settings, $1,000$ datasets were generated from the 2-lag model, which corresponds to model (2.6). Then, raw data MLEs were fit as if a 1-lag model was appropriate and using the same row and column cutoffs as in the size study,

a binning structure was determined. Using this binning structure, bin-based MLEs were found corresponding to a 1-lag model. These were plugged into (3.1)-(3.6) to calculate the chi-square statistic. The degrees of freedom was determined by $df = \sum_{rows}(K_{SVW} - 1) - 5$, where "rows" indicates a summation over all of the rows that resulted from the binning strategy. Further, 5 is subtracted since there are only 5 parameters in the 1-lag model. Table 3.9 displays the percentage of p-values out of 1,000 less than or equal to 0.10, 0.05, and 0.01. The power is excellent for the larger sample sizes $n = 400$ and $n = 2000$, as the departure is detected 100% of the time. Power is still quite good for $n = 200$.

Table 3.9: Percentage of p-values out of 1,000 less than or equal to the listed cutoffs for a 2-lag alternative under a 1-lag null.

	n=200		n=400		n=2000	
Mild	Bal	Unb	Bal	Unb	Bal	Unb
≤ 0.01	86.6%	88.7%	100%	100%	100%	100%
≤ 0.05	95.3%	96.0%	100%	100%	100%	100%
≤ 0.10	97.5%	98.0%	100%	100%	100%	100%
Moderate	Bal	Unb	Bal	Unb	Bal	Unb
≤ 0.01	100%	100%	100%	100%	100%	100%
≤ 0.05	100%	100%	100%	100%	100%	100%
≤ 0.10	100%	100%	100%	100%	100%	100%
Severe	Bal	Unb	Bal	Unb	Bal	Unb
≤ 0.01	71.0%	86.1%	100%	100%	100%	100%
≤ 0.05	87.5%	95.8%	100%	100%	100%	100%
≤ 0.10	92.0%	98.3%	100%	100%	100%	100%

3.2.5 Bootstrap Correction for Small Sample Sizes

The chi-square test for ALR models, like many other goodness-of-fit tests, is primarily a large sample test. In the simulation study, size began to deteriorate for some of the parameter settings with sample sizes significantly less than 200. For example, 10,000 datasets were generated using Mild parameters with Balanced covariates for $n = 100$. Then, under the same conditions as in the size study, bins were formed and a chi-square statistic calculated. Of the p-values generated by this procedure, 12.1% of the resulting p-values were less than 0.10, 6.9% were less than 0.05, and 3.1% were less than 0.01. Using the chi-square statistic under these conditions might lead to an overly high rejection rate.

Further, a sample size could be too small in the sense that using cutoff values of 10 and

5 results in a negative degrees of freedom. To adjust for this, we recommend lowering the column cutoff when binning so that a positive degrees of freedom is obtained. Although the rule of thumb for the chi-square test may be violated, a bootstrap p-value could still be constructed. Thus, for smaller sample sizes, it is recommended to construct a bootstrap p-value.

A bootstrap p-value could be found in the following manner:

1. For a proposed ALR model, calculate the chi-square statistic X^2 for some choice of row threshold and column cutoff.

2. Then, using the raw data MLEs $\hat{\theta}_{RD}$ from step 1, generate $m = 1000$ null datasets from ALR model, treating the MLEs as the true parameter values.

3. Maintain the same binning structure from step 1 and calculate 1000 X^{2*} statistics.

4. The bootstrap p-value is then the number of $(X^{2*} > X^2)/m$.

As a check of the bootstrap procedure, 1,000 datasets were generated from model (2.6) for $n = 100$ patients with Balanced (S, V, W) and Mild parameters (this is the same case mentioned previously where size was inflated). For each dataset, $m = 1,000$ bootstrap datasets were generated and a single bootstrap p-value calculated (Steps 2-4). Of this set of bootstrap p-values, 10.4% were less than 0.10, 4.8% were less than 0.05, and 1.2% were less than 0.01. Thus, we see an improvement in using the bootstrap p-value.

3.2.6 A Computing Aspect

The vast majority of the computing time required by the simulation study (or to obtain a bootstrap p-value) is spent on finding parameter estimates, both raw data or bin-based. The time required to find parameter estimates will also be a concern to the practitioner. Solving for the raw data MLEs requires a maximization of (3.7), which as written requires optimizing a function to solve for 11 parameters and takes a fair amount of computing time. However, due to the tiered model structure, the joint likelihood can be split into two smaller likelihoods, each of which could then be maximized.

Recall, θ as given in (2.7) represents our vector of model parameters for $t \geq 2$. Let θ_2 represent the parameters associated with $t = 2$, such that $\theta_2 = (\beta_{01}, \beta_{02}, \beta_1, \beta_2, \beta_3)$. Likewise, let θ_3 contain all the γ parameters associated with $t \geq 3$, such that $\theta_3 = (\gamma_{01}, \gamma_{02}, \gamma_1, \gamma_2, \gamma_3, \gamma_4)$. Then, the joint likelihood of θ can be rewritten as a product of

two independent likelihoods for θ_2 and θ_3:

$$\begin{aligned}
L(\theta) =& \prod_{i=1}^{n} L_i(\theta) \\
=& \prod_{i=1}^{n} P_\theta(Y_{iT_i} = y_{iT_i}, ..., Y_{i2} = y_{i2} | S_i, V_i, W_i) \\
=& \prod_{i=1}^{n} P_{\theta_3}(Y_{iT_i} = y_{iT_i} | Y_{iT_{i-1}} = y_{iT_{i-1}}, Y_{iT_{i-2}} = y_{iT_{i-2}}, V_i, W_i) \\
& \cdot P_{\theta_3}(Y_{iT_{i-1}} = y_{iT_{i-1}} | Y_{iT_{i-2}} = y_{iT_{i-2}}, Y_{iT_{i-3}} = y_{iT_{i-3}}, V_i, W_i) \cdots P_{\theta_2}(Y_{i2} = y_{i2} | S_i, V_i, W_i) \\
=& \prod_{i=1}^{n} \left(\prod_{t=3}^{T_i} P_{\theta_3}(Y_{it} = y_{it} | Y_{it-1} = y_{it-1}, Y_{it-2} = y_{it-2}, V_i, W_i) \cdot P_{\theta_2}(Y_{i2} = y_{i2} | S_i, V_i, W_i) \right) \\
=& \left(\prod_{i=1}^{n} \prod_{t=3}^{T_i} P_{\theta_3}(Y_{it} = y_{it} | Y_{it-1} = y_{it-1}, Y_{it-2} = y_{it-2}, V_i, W_i) \right) \left(\prod_{i=1}^{n} P_{\theta_2}(Y_{i2} = y_{i2} | S_i, V_i, W_i) \right)
\end{aligned}$$

Rather than maximizing $L(\theta)$ to obtain 11 parameter estimates, two smaller likelihoods can instead be maximized solving for 6 and 5 parameters, respectively. This approach significantly sped up the time of the simulations. For example, on a Toshiba laptop computer (Dual Core CPU 2.0 GHz, 4.0GB RAM) the time to find the raw data parameter estimates went from 20 seconds down to 5 seconds when $n = 200$. While both times seem relatively small, these would need to be repeated at least 1,000 times when finding a bootstrap p-value or when evaluating size or power. By saving those 15 seconds, this reduces the simulation time needed to find raw data MLEs from 5.56 hours down to 1.39 hours, which is a vast improvement.

3.3 Extension of the Goodness-of-Fit Procedure to General ALR Models

Thus far, we have provided an omnibus goodness-of-fit test for an ALR model with fixed, binary covariates and where each patient has been scheduled for an equal number of visits. However, ALR models are not restricted to having only binary covariates, and could have multi-state or continuous covariates. Also, the model could incorporate time-varying covariates. Further, ALR models might be uneven in the sense that the number of scheduled visits is not the same for all subjects (i.e., T_i vary between subjects). We will briefly address how the chi-square test might be modified to handle the following situations: 1) multi-state or continuous covariates, 2) time-varying covariates, and 3) unequal T_i.

For an ALR model with a multi-state covariate, rows could be set up making use of all of the covariate's possible values. For example, suppose model (2.6) had a multi-state covariate V_i that takes four possible values: 1, 2, 3, and 4. Then, row bins could be set up based upon combinations of (S, V, W) as before, although there would now be $2 \cdot 4 \cdot 2 = 16$ unique rows. Then, binning and calculating the expected path counts could occur as outlined and the statistic calculated.

Suppose V_i is still a multi-state covariate but with a larger number of possible values, say 10 possible values. One approach is to proceed as before, although there would now be $2 \cdot 10 \cdot 2 = 40$ disjoint rows. Many of these rows might have very few observations. Rather than having to merge many rows, V_i might instead be further discretized into a smaller number of categories. For example, suppose V_i takes the integer values 1 through 10. Define a new variable V_i^* where

$$V_i^* = \begin{cases} 1 & \text{if } V_i = 1, 2, \text{or } 3, \\ 2 & \text{if } V_i = 4, 5, \text{or } 6, \\ 3 & \text{if } V_i = 7, 8, 9, \text{or } 10. \end{cases}$$

Then, set up the rows based upon values of (S, V^*, W). Here there would be $2 \cdot 3 \cdot 2 = 12$ rows. Then for a particular subject, path probabilities are constructed given the original V covariate values along with W and the start state,

$$p_\theta(u|S_i V_i W_i), \text{ for } u = 1, ..., U.$$

Here, the subscript i for the covariates is necessary because path probabilities in the same row are no longer necessarily identical. Then, expected path counts for the uth path given a (S, V^*, W) row could be found by summing up individual path probabilities

$$e_\theta(u|SV^*W) = \sum_{i=1}^{n} I\{i\} p_\theta(u|S_i V_i W_i)$$

where $I\{i\} = 1$ if $(S_i, V_i^*, W_i) = (S, V^*, W)$, 0 otherwise. Then, any row merging and column binning could occur as described before and the statistic calculated.

The case of a continuous covariate could be handled in a similar manner, in that possible values of the covariate could be discretized into two to four categories when setting up the rows. For example, a continuous covariate could be split into four categories with category boundaries at the 25th, 50th, and 75th percentiles. For an ALR model with only continuous covariates, each variable would have to be discretized in some fashion to form disjoint rows.

To address time-varying covariates, one possible approach is to calculate the path probabilities conditional on all covariate values over time. For example, suppose model (2.6) has a time-varying binary covariate $V_{it},\ i = 1, ..., n,\ t = 1, ..., 4$ that takes the values 0 or 1 at each time point. Then, row bins could be set up based upon combinations of $(S, V_1, V_2, V_3, V_4, W)$. This would create a large number of row combinations (64 in fact), many of which might be sparse. Then, row merging and column binning could occur as outlined previously and the statistic calculated.

If the sheer number of rows generated by using all of the time-varying covariates is too overwhelming, row binning could occur just upon the initial covariate values. For example, suppose V_{it} is a time-varying binary covariate as before. Rather than starting with 64 rows, only 6 rows are needed to describe all possible combinations of (S, V_1, W). For a particular subject Y_i, path probabilities are then constructed given all of the covariates, denoted by

$$p_\theta(1|S_i V_{i1} V_{i2} V_{i3} V_{i4} W_i), ..., p_\theta(U|S_i V_{i1} V_{i2} V_{i3} V_{i4} W_i)$$

As before, these path probabilities are calculated as a product of conditional probabilities defined by the logit equations with the appropriate covariate values plugged in. Expected path counts could be calculated for each unique path $u = 1, ..., U$ by summing up certain individual path probabilities

$$e_\theta(u|SV_1W) = \sum_{i=1}^{n} I\{i\} p_\theta(u|S_i V_{i1} V_{i2} V_{i3} V_{i4} W_i) \tag{3.10}$$

where $I(i) = 1$ if $(S_i, V_{i1}, W_i) = (S, V_1, W)$, 0 otherwise. Then, a table with rows given by combinations of (S, V_1, W) and columns of expected path counts could be created. From this point, binning and calculating the goodness-of-fit statistic would be straightforward.

A disadvantage of setting up rows based only on initial covariate values is that it reduces the within row covariate homogeneity. The same could be said when a multi-state or continuous variable is further discretized to form disjoint rows.

Finally, the test procedure could also be expanded to uneven data sets, where subjects may have been observed for different numbers of visits. One way to do this might be by partitioning the subjects into disjoint sets based upon a common T_i, or the number of scheduled visits. Then, for all subjects with a distinct T_i, a chi-square statistic could be calculated using the methods previously described. This chi-square statistic can be thought of as a measure of goodness-of-fit for the model for those subjects who have observations for T_i time points. These individual chi-square tests could provide insight as to where the

model breaks down. Then, an overall model goodness-of-fit statistic could be constructed by summing the individual chi-square statistics.

In the next section, we will evaluate the goodness-of-fit of a 1-lag ALR model for a relatively small, uneven dataset with time-varying covariates.

Chapter 4

An Application to an Alzheimer's Disease Study

The following dataset was provided by researchers at Loma Linda University in Loma Linda, CA.

4.1 The Dataset

Over 1200 older adults were screened and an eventual 59 selected to participate in a longitudinal mental health study of late-onset dementia. Patients began the study classified as either Normal or as Mildly Cognitively Impaired (MCI) based upon a series of psychological exams. Patients were evaluated every year for at least three years, and at each evaluation characterized as either Normal (1) , MCI (0), or as having developed Alzheimer's disease (-1). Thus, patients take one of three states: 1, 0, and -1, where -1 could be thought of as an absorbing state.

At the time of their psychological evaluations, the subjects also underwent a battery of MRI scans to detect iron content in the brain. At each MRI session, brain images were taken over 14 regions of interest on both hemispheres, for a total of 28 regions. Two types of analyses were done on each brain region image.

The first was a phase measurement value provided by an MRI technology called Susceptibility Weighted Imaging (SWI). These values are inversely associated with regional iron content. Additionally, counts of brain microbleeds (BMBs), were recorded by trained image readers for each region. Demographic variables such as age, gender, and years of education

were recorded upon enrollment in the study. Of the 59 patients in the study, 28 made three scheduled visits, 23 made a maximum of four visits, five patients made a maximum of five visits, two made at most six visits, and one patient was evaluated seven times. Such uneven datasets are quite common in many studies, as patient recruitment may occur over many years. Researchers at Loma Linda hoped to determine if iron buildup in certain regions of the brain could be used to indicate the progression of late-onset Alzheimer's disease among older adults. Additional information about the study can be found in (Kirsch et al., 2009).

4.2 The 1-Lag ALR Model

We found that an ALR model with 1-lag and three covariates best describes the relationships in the data, as this model had the smallest AIC value among other 1-lag models. Let Y_{it} be the state of the ith patient at the t-th time point, $t = 1, ..., T_i$. Age_i ranges from 57 to 81 years ($s = 7.43$). BMB_{it} is the total number of BMBs in the right hemisphere for the ith patient at the t-th time point, and takes discrete values between 0 and 10 ($s = 2.20$). SWI_{it} is an standardized average of four brain region phase values for the ith patient at the t-th time point, and varies between -4.15 and 2.24 ($s = 1.00$). Parameter values were fitted using the optim() function to maximize the likelihood in R and were also verified using proc logistic in SAS.

The proposed 1-lag ALR model for $t \geq 2$ is given by:

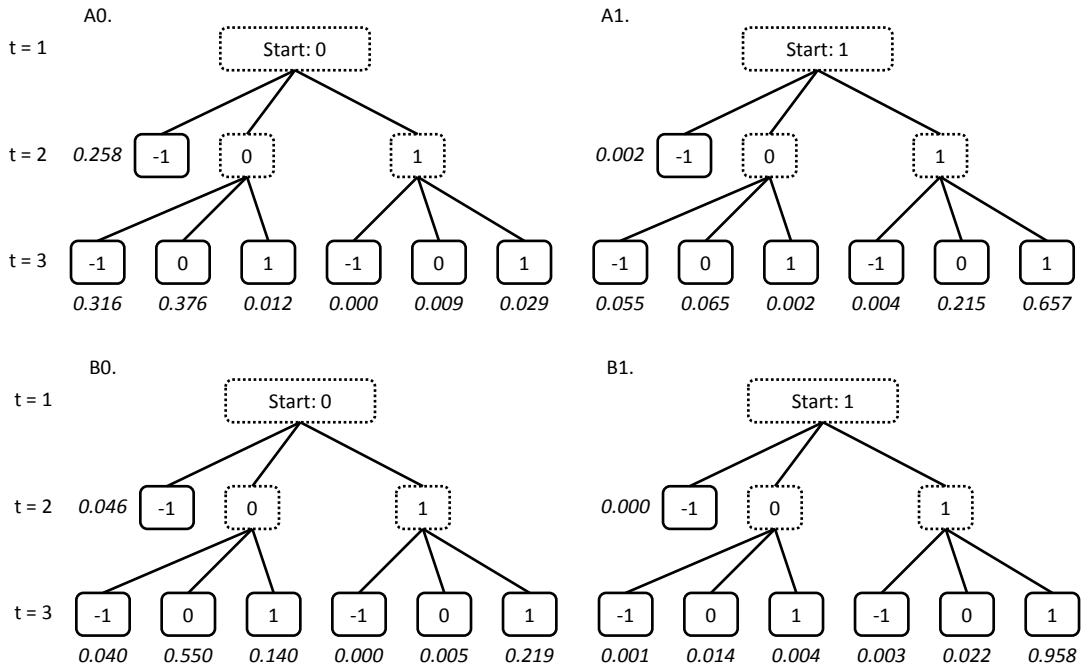
$$\begin{aligned}
&\text{logit}(P(Y_{it} = -1 | Y_{it-1} = y_{it-1})) \qquad (4.1)\\
&= \begin{cases} -9.434 + 0.094 Age_i + 0.192 BMB_{it} - 0.940 SWI_{it} - 5.175 y_{it-1} & \text{if } y_{it-1} \in \{0, 1\} \\ \infty & \text{otherwise} \end{cases}\\
&\text{logit}(P(Y_{it} \leq 0 | Y_{it-1} = y_{it-1}))\\
&= \begin{cases} -5.154 + 0.094 Age_i + 0.192 BMB_{it} - 0.940 SWI_{it} - 5.175 y_{it-1} & \text{if } y_{it-1} \in \{0, 1\} \\ \infty & \text{otherwise} \end{cases}
\end{aligned}$$

Figure 4.1 provides a graphical representation of the model's possible paths and associated path probabilities for two disparate sets of covariate values, where $T_i = 3$. Note that for both sets of covariates, there is a striking difference between progressing from 0 at the start to -1 by the third time point. In A, the probability is about 0.57, while for B the probability is about 0.09. If a patient starts in state 1, they are extremely likely to remain there by the third time point for both sets of covariates.

Figure 4.1: Possible paths and associated path probabilities for two sets of covariate values, $T_i = 3$.

A: Age=80, $BMB_2 = 1$, $BMB_3 = 3$, $SWI_2 = -0.7$, $SWI_3 = -1.2$.

B: Age=65, $BMB_2 = 0$, $BMB_3 = 1$, $SWI_t = -0.3\ \forall t$.



We would like to do a goodness-of-fit test for the model given in (4.1), which has six parameters. Recall also that the dataset is uneven. However, in this case it is not possible to apply the chi-square procedure to the data subsetted by a common T_i, since sample sizes are small. As there are six parameters to estimate, there would not be enough bins to estimate all of the parameters.

Instead, we can use all 59 patients by testing the goodness-of-fit of the model for just the first two time points after the initial visit, $t = 2$ and $t = 3$. This was done by setting up rows based upon different combinations of start state (0 or 1) and BMB_1, which has been further discretized into three categories. Recall that BMB_1 is the number of brain microbleeds at $t = 1$ and takes values from 0 to 10. These possible values were divided into three categories: 0, 1-2, or 3+. Thus, there are $2 \cdot 3 = 6$ disjoint row combinations, as shown in Table 4.1. Note that this is not the only way rows could have been formed. Age or SWI variables could have also been incorporated by discretizing their ranges to form

disjoint categories.

Table 4.1: Initial rows based on start state S and BMB_1, which has been further discretized into three categories.

S	BMB_1	Count
0	0	18
0	1-2	10
0	3+	9
1	0	14
1	1-2	5
1	3+	3

There are seven unique paths a patient might take through $t = 3$, as shown in Figure 4.1. For each patient, seven conditional path probabilities could be calculated. Binning was done using a row threshold of 10 and a column cutoff of 4, resulting in a test statistic value $X^2 = 8.45$. Table 4.2 displays the binning structure used in the calculation of the statistic. A bootstrap p-value based on 1,000 simulated datasets was found to be 0.291. We can be confident in the model fit through the first two time points.

Table 4.2: Row and column bins used in the calculation of the Loma Linda data statistic ($X^2 = 8.45$) are shown in bold. Row binning was done based upon a row threshold of 10. Column merging used a cutoff of 4, and bins were formed by combining appropriate $e_{\hat{\theta}_{RD}}(u|\cdot)$. Expected bin counts $b_{\hat{\theta}_B}(k|\cdot)$ are displayed below each column bin. The observed counts for each bin are also provided in parentheses.

Row Bins $(\mathbf{S}, \mathbf{BMB_1})$	Count	**Column Bins**			
(1, 3+) ∪ (0,3+) ∪ (1, 1-2)	17	**Path 7**	**Path 1, Path 3**	**Path 2, Path 4, Path 5, Path 6**	
		4.700 (6)	8.135 (6)	4.163 (5)	
(0, 1-2) ∪ (1, 0)	24	**Path 7**	**Path3**	**Path 1, Path 4, Path 6**	**Path 2, Path 5**
		14.008 (15)	5.143 (4)	2.344 (1)	2.505 (4)
(0,0)	18	**Path 3**	**Path 2, Path 4**	**Path 1, Path 6, Path 7**	**Path 5**
		11.283 (9)	3.337 (7)	2.645 (2)	0.735 (0)

Path 1: $Start \to -1$, Path 2: $Start \to 0 \to -1$, Path 3: $Start \to 0 \to 0$, Path 4: $Start \to 0 \to 1$,
Path 5: $Start \to 1 \to -1$, Path 6: $Start \to 1 \to 0$, Path 7: $Start \to 1 \to 1$

Part II

Goodness-of-Fit for Generalized Linear Mixed Models

Chapter 5

Generalized Linear Mixed Models

A GLM that incorporates random effects is known as a Generalized Linear Mixed Model (GLMM). This model has four components that must be addressed:

1. What is the distribution of the response data?
2. What function of the mean will be modeled as linear in the predictors? (Link Component)
3. What fixed covariates will be included into the model? (Systematic Component)
4. What type of random effects will be included into the model? (Random Component)

By incorporating random effects, GLMMs can model many complex covariance structures. By using a link function, GLMMs are not restricted to data that follows a normal distribution. Thus, GLMMs are very flexible and useful models for a variety of datasets.

5.1 A General Model

Under a GLMM framework, the data (conditional on the random effect(s)) are assumed to be independent observations from a distribution. Thus, let

$$Y_i|s_i \sim \ indep \ f_{Y_i}(\cdot|\mu_i, \theta)$$

where as before, μ_i is the mean value of the distribution and θ represents any nuisance parameters. Then, a GLMM employs a link function $g(\cdot)$ such that

$$\begin{aligned} E[Y_i|s_i] &= \mu_i \\ g(\mu_i) &= X_i'\beta + Z_i's \end{aligned}$$

where Z_i' represent a row of a design matrix Z and s is a vector of random effects. As in the LMM case, we assume that

$$s \sim \ f_S(\cdot|D)$$

with mean 0 and unknown variance D.

5.2 Maximum Likelihood Estimation

Parameter estimates for GLMMs are more difficult to obtain than in GLMs due to the complex nature of the likelihood function,

$$L(\beta, D, \theta) = \int \cdots \int \Big\{ \prod_i f_{Y_i|s_i}(y_i) \Big\} f_{S_i}(s_i) ds_i.$$

This type of likelihood is known as an integrated likelihood. However, several methods have been developed to find estimates of the parameters such as quadrature and quasi-likelihood approaches. We will briefly discuss some of these methods here. Of course if the response distribution conditional on the random effects follows a normal distribution, then linear mixed model approaches can be used.

5.2.1 Quadrature

Quadrature was first developed as a mathematical technique for evaluating integrals. Suppose that we have a GLMM with a single independent, normally distributed random effect, that is $s_i \sim$ i.i.d. $N(0, \sigma_s^2)$. Then, the likelihood for this model can be written as a function of the random effects s_i, such that

$$L = \prod_i \int_{-\infty}^{\infty} f_{Y_i|s_i}(y_i) \frac{e^{-s_i^2/(2\sigma_s^2)}}{\sqrt{2\pi\sigma_s^2}} ds_i$$

Thus, the likelihood is a product of one-dimensional integrals of the form

$$\int_{-\infty}^{\infty} h(s) \frac{e^{-s^2/(2\sigma_s^2)}}{\sqrt{2\pi\sigma_s^2}} ds.$$

Upon a change of variables $s = \sqrt{2}\sigma_u v$, this can be re-written as

$$\int_{-\infty}^{\infty} h^*(v) e^{-v^2} dv \tag{5.1}$$

where $h^*(\cdot) = h(\sqrt{2}\sigma_s \cdot)/\sqrt{\pi}$. For smooth functions $h^*(\cdot)$ multiplied by e^{-v^2}, Gauss-Hermite quadrature is available and approximates the integral (5.1) as a weighted sum:

$$\int_{-\infty}^{\infty} h^*(v)e^{-v^2}dv \approx \sum_{k=1}^{d} h^*(x_k)w_k,$$

where weights w_k and the points of evaluation x_k can be calculated by most statistical software since

$$x_k = \text{ith zero of } H_n(x)$$
$$w_k = \frac{2^{n-1}n!\sqrt{\pi}}{n^2[H_{n-1}(x_k)]^2}$$

where $H_n(x)$ is the Hermite polynomial of degree n. Tables are also available in Abramowitz and Stegun (1964) along with details about calculating the weights and evaluation points. By using quadrature of a high enough degree (say, $d \geq 30$), accurate approximations of the likelihood can be calculated.

Numerical quadrature is limited to smooth functions, and is easily applied to models with a single, independent random effect. Additional random effects (such as an interaction effect) would increase the dimension of the integral. Gauss-Hermite quadrature is limited to integrals that can be put into the form of (5.1).

5.2.2 The EM Algorithm

An EM algorithm approach treats the random effects s as if it were missing data, such that the complete data is $w' = (Y', s')$. The EM algorithm then proceeds by forming the log-likelihood of the complete data, calculating its expectation with respect to the conditional distribution of $s|Y$ and then maximizing with respect to its parameters. The algorithm is iterative as the expectation of the log-likelihood must be recalculated for the complete data given a new set of parameter estimates.

To set up the EM algorithm, first note that the distribution of the complete data can be factored as $f_{Y,s} = f_{Y|s}f_s$, such that the complete log likelihood is,

$$\begin{aligned} logL &= logf_{Y|s} + logf_s \\ &= \sum_{i=1}^{n} logf_{Y_i|s} + logf_s \end{aligned}$$

Note that the parameters β and θ enter only the first part of the log likelihood, and D enters only through f_s. The EM algorithm then takes the following form (McCulloch et al., 2008):

1. Choose starting values β^0, θ^0, and D^0. Set m=0.

2. Then, perform the following:

   a. Calculate β^{m+1} and θ^{m+1} to maximize $E[log f_{Y|s}(Y|s,\beta,\theta)|y]$, where the expectation is evaluated under the mth iteration parameter values.

   b. Similarly, calculate D^{m+1} to maximize $E[log f_s(s|D)|y]$.

   c. Set $m = m + 1$.

3. If convergence is achieved, declare the current values to be the MLEs; else return to step # 2 and repeat.

In general, the expectations in step #2 cannot be computed in closed form. However, this approach can be combined with Monte Carlo approximations, which will obtain the required expectations. We discuss this in the next section.

5.2.3 Markov Chain Monte Carlo (MCMC) Metropolis Algorithm

A Metropolis algorithm generates a Markov chain sequence of values that eventually stabilizes to draws from a candidate distribution. To specify a Metropolis algorithm, a candidate distribution, $h_s(s)$, must be selected, from which potential new values are drawn. An acceptance function gives the probability of accepting a new value (versus keeping a previous value) is given by

$$A_k(s^*, s) = min\left\{1, \frac{f_{s|Y}(s^*|y,\beta,\theta,D)h_s(s)}{f_{s|Y}(s^*|y,\beta,\theta,D)h_s(s^*)}\right\}$$

where $s^* = (s_1, s_2, ..., s_{k-1}, s_k^*, s_{k+1}, ..., s_q)'$, which is the candidate new value and has all the entries equal to the previous value except the kth. Looping over k, a sample s is obtained from the candidate distribution.

In the context of parameter estimation, we can choose $h_s = f_s$, and our ratio becomes

$$\begin{aligned}\frac{f_{s|Y}(s^*|y,\beta,\theta,D)h_s(s)}{f_{s|Y}(s|y,\beta,\theta,D)h_s(s^*)} &= \frac{\prod_{i=1}^n f_{Y_i|s}(y_i|s^*,\beta,\theta,D)f_s(s^*|D)f_s(s|D)}{\prod_{i=1}^n f_{Y_i|s}(y_i|s,\beta,\theta,D)f_s(s|D)f_s(s^*|D)} \\ &= \frac{\prod_{i=1}^n f_{Y_i|s}(y_i|s^*,\beta,\theta,D)}{\prod_{i=1}^n f_{Y_i|s}(y_i|s,\beta,\theta,D)}\end{aligned}$$

which only involves the "GLM" part of the model. This Metropolis step can be incorporated into the EM algorithm. A Monte Carlo EM (MCEM) algorithm is then given by

1. Choose starting values β^0, θ^0, and D^0. Set m=0.

2. Generate M values, $s^{(1)}, s^{(2)}, ..., s^{(M)}$, from the conditional distribution of s given Y using the Metropolis algorithm.

   a. Calculate β^{m+1} and θ^{m+1} to maximize a Monte Carlo estimate of $E[log f_{Y|s}(Y|s, \beta, \theta)|y]$, that is choose values to maximize $(1/M)\sum_{k=1}^{M} f_{Y|s}(Y|s^{(k)}, \beta, \theta)$.

   b. Calculate D^{m+1} to maximize $(1/M)\sum_{k=1}^{M} log f_s(s^{(k)}|D)$.

   c. Set $m = m + 1$.

3. If convergence is reached, declare the current values to be the MLEs; else return to step # 2.

This approach is perhaps computationally intensive, but feasible for many GLMM models.

5.2.4 Other Methods for Finding ML Estimates

McCulloch et al. (2008) describes other techniques to obtain ML parameter estimates, including a Monte Carlo Newton-Raphson method, a stochastic approximation (SA) algorithm, and a simulated maximum likelihood, where simulation is done to maximize the likelihood directly (as opposed to the log likelihood).

5.3 Penalized Quasi-likelihood and Laplace Approximation

It is sometimes the case that the distribution of the response is not known with certainty or less restrictive assumptions are desired. Not knowing the distribution makes it impossible to construct a likelihood. However, it would be useful to have a method that works almost as well as maximum likelihood but without making specific distributional assumptions. Quasi-likelihood allows the derivation of a likelihood-like quantity under less specific constraints. This is done by mimicking the properties of the derivative of the log likelihood (a.k.a. the score function).

One quasi-likelihood approach makes use of the Laplace approximation, which is based

upon a second-order Taylor series expansion which takes the form

$$log \int_{\Re^q} e^{h(s)} ds \approx h(s_0) + \frac{q}{2} log 2\pi - \frac{1}{2} log \left| -\frac{d^2 h(s)}{ds ds'} \right|_{s=s_0} \tag{5.2}$$

where s_0 is the solution to $dh(s)/ds = 0$. This result can be used to approximate the log-likelihood of a GLMM

$$\begin{aligned} l &= log \int f_{Y|s} f_s ds \\ &= log \int e^{log f_{y|s} + log f_s} \\ &= log \int e^{h(s)} ds \end{aligned}$$

To construct a Laplace approximation, s_0 must be found and an expression for the second derivative in (5.2) is needed. These can be found via differentiation along with the use of some clever approximations (see McCulloch et al. for further details). This approximated likelihood can then be differentiated with respect to β. In the end, we are left with two equations that must be solved simultaneously to estimate β and s_0. Once the β estimates have been found, another approach must be taken to solve for D, the covariance matrix of s. Methods for solving these equations are known as penalized quasi-likelihood (PQL) methods. However, many PQL methods lead to estimates that are asymptotically biased.

5.4 Finding Predictors for Random Effects

Along with finding parameter estimates, it is sometimes of interest to predict the values of the random effects themselves. We will refer to these values as predictors. Some authors describe the predicted values of the random effects as "estimators", but that term is problematic as it blurs the distinction between estimating fixed and random effects.

For linear mixed models, solving for predictors is outlined in Section 1.4, and makes use of the mixed model equations.

5.4.1 Best Predictors (BPs) and Best Linear Predictors (BLPs)

To obtain predictors of random effects, we should start by looking at a general prediction problem. Suppose we would like to estimate w based on y. In the context of GLMMs (and

LMMs), w is a vector of random effects and y is the available data.

$$\begin{bmatrix} y \\ w \end{bmatrix} \sim \left[\begin{pmatrix} \mu_y \\ \mu_w \end{pmatrix}, \begin{pmatrix} V_{yy} & V_{yw} \\ \cdot & V_{ww} \end{pmatrix} \right]$$

The minimum mean square error predictor of w, based upon y, is $E[w|y]$, often referred to as the Best Predictor (BP) of w. Further, suppose that we wish to use a linear function of y (i.e., $a + b'y$) to predict w. The best linear predictor (BLP) of w, based on y is $BLP(w) = \mu_w + V'_{yw}V_{yy}^{-1}(y - \mu_y)$. Also, it can be shown that the best linear predictor is identical to the best predictor under normality.

As an example, we will show how to obtain the BLP for the following Beta-Binomial model given below, $i = 1, 2, ..., m$, $j = 1, 2, ..., n_i$,

$$\begin{aligned} Y_{ij}|p_i &\sim indep.\ Bernoulli(p_i) \\ E[Y_{ij}|p_i] &= p_i \\ p_i &\sim i.i.d.\ Beta(\alpha, \beta) \end{aligned}$$

This model is not a traditional GLMM and is known as a random effects model, since there are only random effects (and no fixed effects) in the model formulation. However, it will serve as a tractable example to highlight best linear prediction. Suppose that we want to find the BLP of the p_i, based upon $Y_{i\cdot} = (Y_{i1}, Y_{i2}, ..., Y_{in_i})'$. Using the formula,

$$BLP(p_i) = E[p_i] + Cov(Y_{i\cdot}, p_i)'[Var(Y_{i\cdot})]^{-1}[Y_{i\cdot} - E[Y_{i\cdot}]]$$

It follows from the Beta distribution that

$$\begin{aligned} E[p_i] &= \frac{\alpha}{\alpha + \beta}, \\ Var(p_i) &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \end{aligned}$$

Then using properties of expectations and variances,

$$\begin{aligned} E[Y_{ij}] &= E[E[Y_{ij}|p_i]] = E[p_i] = \frac{\alpha}{\alpha + \beta}, \\ Var(Y_{ij}) &= E[Var(Y_{ij}|p_i)] + Var(E[Y_{ij}|p_i]) \\ &= E[p_i(1 - p_i)] + Var(p_i) \\ &= E[p_i] - E[p_i^2] + E[p_i^2] - E[p_i]^2 \\ &= E[p_i](1 - E[p_i]) \\ &= \frac{\alpha}{\alpha + \beta}\frac{\beta}{\alpha + \beta} = \frac{\alpha\beta}{(\alpha + \beta)^2}, \end{aligned}$$

and for $j \neq j'$,

$$\begin{aligned} Cov(Y_{ij}, Y_{ij'}) &= Cov(E[Y_{ij}|p_i], E[Y_{ij'}|p_i]) + E[Cov(Y_{ij}, Y_{ij'}|p_i)] \\ &= Cov(p_i, p_i) + 0 \\ &= Var(p_i). \end{aligned}$$

The covariance of Y_{ij} and p_i is also straightforward

$$\begin{aligned} Cov(Y_{ij}, p_i) &= Cov(E[Y_{ij}|p_i], E[p_i|p_i]) + E[Cov(Y_{ij}, p_i|p_i)] \\ &= Cov(p_i, p_i) + 0 \\ &= Var(p_i). \end{aligned}$$

Now, we can find the best linear predictor of p_i, provided we can solve for the inverse of $Var(Y_{i\cdot})$.

$$\begin{aligned} BLP(p_i) &= E[p_i] + Cov(Y_{i\cdot}, p_i)'[Var(Y_{i\cdot})]^{-1}[Y_{i\cdot} - E[Y_{i\cdot}]] \\ &= \frac{\alpha}{\alpha+\beta} \\ &\quad + [Var(p_i) \cdots Var(p_i)] \begin{bmatrix} Var(Y_{i1}) & Var(p_i) & \cdots & Var(p_i) \\ Var(p_i) & Var(Y_{i2}) & \cdots & Var(p_i) \\ \vdots & \vdots & & \vdots \\ Var(p_i) & Var(p_i) & \cdots & Var(Y_{in_i}) \end{bmatrix}^{-1} \begin{bmatrix} Y_{i1} - \frac{\alpha}{\alpha+\beta} \\ Y_{i2} - \frac{\alpha}{\alpha+\beta} \\ \vdots \\ Y_{in_i} - \frac{\alpha}{\alpha+\beta} \end{bmatrix} \end{aligned}$$

Although the calculations are a bit tedious, a closed form BLP for this model is possible. Of course, parameter values such as α and β are generally unknown. Thus, we can obtain the empirical best linear predictors (or eBLPs) by plugging in ML estimates for the unknown parameters.

5.4.2 Best Linear Unbiased Predictors (BLUPs) and Empirical Best Linear Unbiased Predictors (eBLUPs) for LMMs

Now, suppose that $\mu_y = X\beta$ and $\mu_w = \lambda'\beta$ for some matrix X, vector λ and parameter β.

$$\begin{bmatrix} y \\ w \end{bmatrix} \sim \left[\begin{pmatrix} X\beta \\ \lambda'\beta \end{pmatrix}, \begin{pmatrix} V_{yy} & V_{yw} \\ \cdot & V_{ww} \end{pmatrix} \right]$$

It follows that,

$$BLP(w) = \lambda'\beta + V_{yw}'V_{yy}^{-1}(y - X\beta)$$

But suppose β is unknown. Let $\hat{\beta}$ denote the generalized least squares estimator of β, i.e., $\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$. It can be shown that the best linear unbiased predictor (BLUP) of w, based upon y, is

$$BLUP(w) = \lambda'\hat{\beta} + V'_{yw}V_{yy}^{-1}(y - X\hat{\beta})$$

Now suppose that $\beta = \beta(\theta)$ and

$$\begin{pmatrix} V_{yy} & V_{yw} \\ \cdot & V_{ww} \end{pmatrix} = \begin{pmatrix} V_{yy}(\theta) & V_{yw}(\theta) \\ \cdot & V_{ww}(\theta) \end{pmatrix}$$

Then, the BLUP is

$$BLUP(w) = \lambda'\hat{\beta}(\theta) + V_{yw}(\theta)'V_{yy}^{-1}(\theta)(y - X\hat{\beta}(\theta))$$

but if θ is unknown, the BLUP cannot be used. However, an empirical best linear predictor (eBLUP) is

$$eBLUP(w) = \lambda'\hat{\beta}(\hat{\theta}) + V_{yw}(\hat{\theta})'V_{yy}^{-1}(\hat{\theta})(y - X\hat{\beta}(\hat{\theta}))$$

where $\hat{\theta}$ is an estimator based on y of θ. For example, $\hat{\theta}$ could be a ML or REML estimator.

5.4.3 Empirical Bayes Prediction

Another approach to prediction is empirical Bayes prediction. Suppose that Y is our data, and θ represents all of the model parameters. A Bayesian approach to parameter estimation assumes that parameters are random variables, with a prior distribution $\Pi(\theta)$. Then, by properties of conditional probability

$$f(y, \theta) = f(y|\theta)\Pi(\theta)$$

and we can obtain an "updated" distribution of θ called the posterior distribution,

$$\Pi(\theta|y) = \frac{f(y|\theta)\Pi(\theta)}{f(y)} = \frac{f(y|\theta)\Pi(\theta)}{\int f(y|\theta)\Pi(\theta)d\theta}$$

and the mean of $\theta|y$ derived from this density is the Bayes estimator of θ.

Now suppose that $\Pi(\theta)$ and $\Pi(\theta|y)$ involve some parameters ϕ so that

$$f(y) = \frac{f(y|\theta)\Pi(\theta|\phi)}{\Pi(\theta|y, \phi)}$$

Suppose ϕ can be estimated as a function of y. Then, $E[\theta|y]$ derived from $\Pi(\theta|y)$ with the estimates of ϕ substituted in is known as the empirical Bayes estimate of θ.

As an example, suppose we have the following model for $i = 1, 2, ..., m$ and $j = 1, ..., n$ where

$$\begin{aligned} Y_{ij}|a &\sim indep.\ Bernoulli(p_{ij}) \\ logit(p_{ij}) &= \mu + a_i \\ a_i &\sim i.i.d. N(0, \sigma_a^2) \end{aligned}$$

The log likelihood is given below, which can easily be put into a form appropriate for quadrature

$$\begin{aligned} logL &= log \prod_i \int f_{Y|a_i}(Y|a_i) f_{a_i}(a_i) da_i \\ &= \sum_i log \int e^{(\mu+a_i)Y_{i\cdot} - nlog(1+e^{\mu+a_i})} \frac{e^{-a_i^2/2\sigma_a^2}}{\sqrt{2\pi\sigma_a^2}} da_i \\ &\approx \sum_i log \left(\sum_k e^{(\mu+x_k)Y_{i\cdot} - nlog(1+e^{\mu+x_k})} w_k/\sqrt{\pi} \right). \end{aligned}$$

Then, the empirical Bayes estimators, which are also the best predicted values, for the model are

$$\begin{aligned} E[a_i|Y] &= \int a_i f_{a_i|Y}(a_i|Y) da_i \\ &= \int a_i f_{Y|a_i}(Y|a_i) f_{a_i}(a_i)/f_Y(Y) da_i \\ &= \frac{\int a_i f_{Y|a_i}(Y|a_i) f_{a_i}(a_i) da_i}{\int f_{Y|a_i}(Y|a_i) f_{a_i}(a_i) da_i} \end{aligned}$$

which can also be solved via quadrature.

5.5 A GLMM Example: The Randomized Clinical Trial Model

Suppose Y_{ijk} represents the number of recovery days needed for the kth patient in the ith clinic under the jth treatment arm, where $i = 1, 2, ..., I$ clinics, $j = 1, 2$ treatments and $k = 1, 2, ..., n_{ij}$. We can define the following Randomized Clinical Trial (RCT) model as defined in Li, Jeske and Klein (2011) as follows:

$$\begin{aligned} &Y_{ijk}|s_i \sim \text{indep. } Poisson(\lambda_{ijk}) \\ &log(\lambda_{ijk}) = \theta + \beta X_j + s_i \\ &s_i \sim N(0, \sigma^2) \end{aligned}$$

where $X_j = 0$ if a patient is assigned to the control arm and 1 for patients assigned to the treatment arm, and s_i is the random effect for the ith clinic. This random effect implies that responses from all patients in the ith center have some correlation. The integrated likelihood for this model can be described:

$$\begin{aligned}
L(\theta, \beta, \sigma^2) &= \prod_{i=1}^{I} \int_{-\infty}^{\infty} \prod_{j=1}^{2} \prod_{k=1}^{n_{ij}} \frac{e^{-(\theta+\beta X_{ij}+s_i)}(\theta + \beta X_{ij} + s_i)^{Y_{ijk}}}{Y_{ijk}!} \frac{e^{-s_i^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} ds_i \\
&= \prod_{i=1}^{I} \int_{-\infty}^{\infty} \Big(\prod_{k=1}^{n_{i1}} \frac{e^{-(\theta+\beta+s_i)}(\theta + \beta + s_i)^{Y_{i1k}}}{Y_{i1k}!} \Big) \Big(\prod_{k=1}^{n_{i2}} \frac{e^{-(\theta+s_i)}(\theta + s_i)^{Y_{i2k}}}{Y_{i2k}!} \Big) \frac{e^{-s_i^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} ds_i \\
&= \prod_{i=1}^{I} \int_{-\infty}^{\infty} \Big(\frac{e^{-n_{i1}(\theta+\beta+s_i)}(\theta + \beta + s_i)^{Y_{i1\cdot}}}{\prod_k Y_{i1k}!} \Big) \Big(\frac{e^{-n_{i2}(\theta+s_i)}(\theta + s_i)^{Y_{i2\cdot}}}{\prod_k Y_{i2k}!} \Big) \frac{e^{-s_i^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} ds_i.
\end{aligned}$$

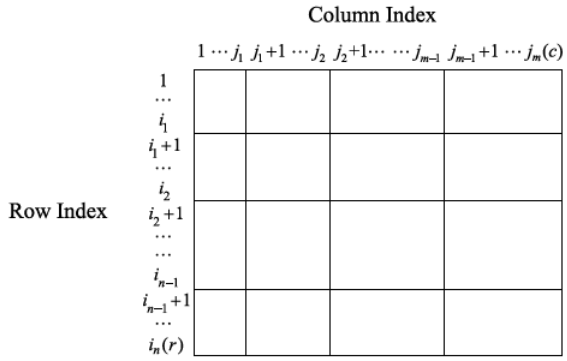
We will use this model to highlight our goodness-of-fit ideas.

5.6 A GLMM Example: The Spatial Model

Along with the RCT model, a spatial, co-clustering model has been defined by Zhang et al. (2012) in relation to estimating pest density in an orchard setting. Co-clustering, also known as bivariate clustering, has been applied in many settings. Generally, data is arranged in a matrix of rows and columns, where each cell in the matrix is represented by a real number. Rather than identifying similar rows and columns independently, co-clustering seeks to take advantage of underlying dependencies and simultaneously cluster rows and columns. See Zhang et al. for more information on co-clustering.

Consider an r x c spatial grid, where each point on the grid is a potential sampling site. These r rows and c columns can be divided into contiguous, disjoint groups like a checkerboard. A co-clustering with n groups of rows and m groups of columns is represented in Figure 5.1. Thus, a specific design with n rows and m columns is denoted by $(i_1, i_2 - i_1, ..., i_n - i_{n-1})$ x $(j_1, j_2 - j_1, ..., j_m - j_{m-1})$.

Figure 5.1: A checkerboard co-cluster structure with $n = 4$ rows and $m = 4$ columns.



Zhang et al. propose the following GLMM for co-clustering in this checkerboard structure:

$$
\begin{aligned}
&Y_{j(i)}|s \text{ indep. } \sim Negative\ Binomial(\theta_i, \kappa), i = 1, 2, ..., nm, j = 1, 2, ..., n_i;\\
&log(\theta_i) = \mu + s_i;\\
&s = (s_1, s_2, ..., s_{nm})' \sim MVN(0, \sigma^2 I),
\end{aligned}
$$

where $Y_{j(i)}$ is the count number from the jth sampling unit in the ith co-cluster, n_i is the total number of sampling units in the ith co-cluster, θ_i is the conditional mean of counts associated with the ith co-cluster, κ is the dispersion parameter, and s_i is the random effect associated with each co-cluster.

Note that here we are using what is known as the "Entomologist" parameterization of the Negative Binomial, which is often used to model pest density. To demonstrate this parameterization, let $X =$ a discrete, non-negative integer random variables such that $X \sim$ Negative Binomial(μ, κ) where μ is the mean and κ is known as the dispersion parameter. The pmf of X is given by

$$
\begin{aligned}
P(X = x) &= \frac{\Gamma(x + \kappa)}{\Gamma(x + 1)\Gamma(\kappa)} \left(\frac{\kappa}{\mu + \kappa}\right)^{\kappa} \left(\frac{\mu}{\mu + \kappa}\right)^{x}, \quad x = 0, 1, ...\\
&= \binom{x + \kappa - 1}{\kappa - 1} \left(\frac{\kappa}{\mu + \kappa}\right)^{\kappa} \left(\frac{\mu}{\mu + \kappa}\right)^{x}, \quad x = 0, 1, ...
\end{aligned}
$$

and $EX = \mu$ and $Var(X) = \mu(1 + \mu/\kappa)$. This is in contrast to a typical Negative Binomial, where $X = \#$ of trials before achieving the rth success and $X \sim$ Negative Binomial(r, p) where r is the number of successes of interest and p is the probability of success. Here, the

pmf is given by

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad x = r, r+1, ...$$

where $EX = r/p$ and $Var(X) = rp/p^2$. Additional alternate parameterizations of the Negative Binomial exist, such as when $X = \#$ of failures before observing the rth success, and the Johnson and Kotz parameterization. We will use the Entomologist parameterization of the Negative Binomial throughout this dissertation.

Zhang et al. go on to develop a heuristic optimization algorithm to select the best m x n design that maximizes the likelihood, given by

$$l(\mu, \sigma^2, \kappa) = \prod_{i=1}^{nm} \int_{-\infty}^{\infty} \left(\prod_{j=1}^{n_i} \Big(\frac{\Gamma(Y_{j(i)} + \kappa)}{\Gamma(Y_{j(i)} + 1)\Gamma(\kappa)} \Big) \Big(\frac{\kappa}{e^{\mu+s_i} + \kappa} \Big)^{\kappa} \Big(\frac{e^{\mu+s_i}}{e^{\mu+s_i} + \kappa} \Big)^{Y_{j(i)}} \right) \frac{e^{-s_i^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} ds_i$$

We will use this Spatial model, along with the RCT model, to illustrate our goodness of fit methodology.

GLMMs are becoming increasingly relevant and are extremely useful for modeling complex datasets. With improved computing power, fitting GLMMs is no longer as burdensome as it once was. Many statistical software packages have methods to fit these models (e.g. SAS proc glimmix). For GLMMs, we have to correctly specify the response distribution and regression model, as well as the distribution of the random effects. However, since GLMMs can be quite complex, it is sometimes difficult to determine if the model fits the data reasonably. In the next chapter, we will discuss goodness-of-fit GLMMs.

Chapter 6

Literature Review of Current Goodness-of-Fit Methods for GLMMs

A goodness-of-fit analysis for a GLMM could focus on any of the model components described in Chapter 2. Is the response distribution (conditional on the random effects) appropriate? Is the link function correct? Are the random effects correctly specified? A goodness-of-fit test for GLMMs should be able to address at least one of these questions. There is also a fourth question: are the covariates included in the model appropriate? Often, this can be answered by model selection techniques, which we will discuss in Section 6.3.

The development of goodness-of-fit procedures for GLMMs is challenging because the existence of random effects not only complicates any theoretical derivations, but also imposes computational challenges. There have been very few omnibus goodness-of-fit tests developed for GLMMs. An omnibus test is one in which the null hypothesis is fully specified and there is no alternative. A good omnibus test would have high power to detect any deviation from the null.

6.1 Consequences of a Misspecified GLMM

A paper by Heagerty and Kurland (2001) explores the consequences of misspecifying the random effects distribution of a GLMM. The authors generated data from the following

logistic GLMM:

$$g(\mu_{ij}) = \beta_0 + \beta_1 X_{ij,1} + \beta_2 X_{ij,2} + \beta_3 X_{ij,1} X_{ij,2} + b_{ij}$$

where $b_{ij} = \sigma(a_i - \lambda)/\sqrt{\lambda}$ and $a_i \sim Gamma(\lambda, 1)$ for different values of λ. $X_{ij,1}$ and $X_{ij,2}$ are indicator variables. They then found the ML estimates for this model via quadrature, where a Gaussian distribution was assumed for the random effects (*i.e.*, $b_{ij} \sim N(0, \sigma_b^2)$) The authors found significant bias in the estimators of $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \widehat{\sigma}_b^2)$. Thus, in a GLMM, the incorrect specification of the random effects has some negative effects on parameter estimations. This is in contrast to results for linear mixed models, where Verbeke and Lesaffre (1997) showed that misspecified random effects can still produce consistent and asymptotically normal ML parameter estimates under certain settings.

Along with the random effects, a misspecified response distribution or link function can also lead to biased parameter estimates. In the next section, we will discuss a few formal goodness-of-fit tests that have been developed from GLMMs.

6.2 Formal Goodness-of-Fit Tests for GLMMs

In contrast to GLMs and LMMs, there have been few formal goodness-of-fit tests developed for GLMMs. Pan and Lin (2005) have developed a series of goodness-of-fit tests for GLMMs based upon cumulative sums of residuals, which are targeted towards certain types of model departures. They also provide an omnibus goodness-of-fit test. Gu (2008) also provides an omnibus goodness-of-fit test for GLMMs with nested or crossed random effects, based upon a modified chi-square statistic. In contrast to these omnibus tests, Abad and Litiere (2010) have developed procedures to evaluate misspecification of the random effects in GLMMs using model information matrix tests (IMTs). We will briefly discuss these papers in the following sections.

6.2.1 Tests for Model Misspecification using Cumulative Sums

Cumulative sums (or CUSUMs) are quite commonly used in statistical process control to detect when a process is out of control. Here, Pan and Lin use specialized cumulative sums of GLMM residuals to detect certain types of model mispecification. The authors consider a longitudinal GLMM, where Y_{ij} is the response of the ith subject on the jth occasion, and let X_{ij} and Z_{ij} be the p x 1 and q x 1 vectors associated with the fixed and

random effects, respectively. The GLMM takes the form

$$g\{E[y_{ij}|u_i]\} = X'_{ij}\beta + Z'_{ij}u_i \; i = 1, ..., n; \; j = 1, ..., t_i \tag{6.1}$$

where $g(\cdot)$ is a known differentiable link function, β is a p x 1 vector of unknown regression parameters and u_i is a q x 1 vector of random effects for the ith subject. Under this model, the marginal means of y_{ij} are given by

$$E(y_{ij}) = E[E(y_{ij}|u_i)] = \int g^{-1}(X'_{ij}\beta + Z'_{ij}u_i) f_{u_i}(u_i) du_i$$

Denote these marginal means by $m_{ij}(\theta)$, where θ is the vector of parameters. Given the estimates $\hat{\theta}$, the residuals are defined as $e_{ij} = y_{ij} - m_{ij}(\hat{\theta})$.

While these individual residuals could be plotted, such plots are difficult to interpret since the variability of the residuals is unknown. However, it is possible to aggregate these residuals in such a way that they follow a known distribution, provided that the GLMM has been correctly specified.

Here, two cumulative sum processes can be specified:

$$W(x) = n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{t_i} I(X_{ij} \leq x) e_{ij} \tag{6.2}$$

$$W_g(r) = n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{t_i} I(\hat{m}_{ij} \leq r) e_{ij} \tag{6.3}$$

where $x = (x_i, ..., x_p)' \in \Re^p$. Note that $I[\cdot]$ is the indicator function where $I(X_{ij} \leq x) = I(X_{1ij} \leq x_1, ..., X_{pij} \leq x_p)$. Using these CUSUMs, it is possible to test the goodness-of-fit of the functional form of the fixed effects, the link function, and the overall mean response function.

Define H_0 as the correct specification of the GLMM model as given in equation (6.1). Under the null, these cumulative sum process $W(x)$ can be shown to converge in distribution to zero-mean Gaussian processes. Thus, under the null these processes should fluctuate around 0. A large value of $sup|W(x)|$ or $sup|W_g(r)|$ would indicate misspecification, and p-values can be obtained via simulation.

The null distribution of W(x) is unknown, but it can be approximated. Define

$$\hat{W}(x) = n^{-1/2} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{t_i} I(X_{ij} \leq x) e_{ij} + \eta'(x; \hat{\theta}) \mathcal{I}^{-1}(\hat{\theta}) \right\} G_i$$

where $(G_1, ..., G_n)$ are i.i.d. standard Normal variables that are independent of the data (y_{ij}, X_{ij}, Z_{ij}) and

$$\eta'(x; \hat{\theta}) = -n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{t_i} I(X_{ij} \leq x) \frac{dm_{ij}(\theta)}{d\theta}$$

The conditional distribution of $\hat{W}(x)$ given the data (y_{ij}, X_{ij}, Z_{ij}) is the same in limit as $W(x)$ under H_0 as shown by Pan and Li. Thus, to approximate a null distribution for $W(x)$, a large number of realizations of $\hat{W}(x)$ can be obtained by repeatedly generating a normal random sample $(G_1, ..., G_n)$ while fixing the data (y_{ij}, X_{ij}, Z_{ij}). Likewise, the distribution of $W_g(r)$ can be approximated by $\hat{W}(x)$, where $\hat{W}_g(r)$ is obtained from $\hat{W}(x)$ by replacing $I(X_{ij} \leq x)$ with $I(\hat{m}_{ij} \leq r)$.

To check the functional form of a fixed effect, consider the following CUSUM process:

$$W_k(x) = n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{t_i} I(X_{kij} \leq x) e_{ij}$$

where $x \in \mathfrak{R}$. This is a special case of $W(x)$ with $x_l = \infty$ for all $l \neq k$. Thus, the null distribution of $W_k(x)$ can also be approximated by the $\hat{W}$ process. As for the statistics given in (6.2) and (6.3), these can be used as checks of the link function and as an overall omnibus goodness-of-fit test, respectively.

The authors carried out simulation studies to investigate the behavior of the three CUSUM tests. All three had good size. The functional form test was found to have excellent power against misspecification of the functional forms of covariates, while the link function and omnibus tests had reasonable power at the larger sample sizes.

6.2.2 Omnibus Goodness-of-Fit Using a Modified Chi-Square Statistic

Gu (2008) proposes two modified chi-square goodness-of-fit tests for GLMMs with either nested or crossed random effects. A nested effect is defined as a source of variation that is nested within a factor. More specifically, factor B is nested within factor A if every level of B only appears within a level of A. We can observe a nested random effect in the following GLMM model equation for $i = 1, ..., m$ subjects and $j = 1, ..., n_i$ observations on the ith subject:

$$g(\theta_{ij}) = \mu + \alpha_i + b_{j(i)}$$

Here, the nested random effect $b_{j(i)}$ contributes to the variance of the main (fixed) effect α. Gu proposes an omnibus chi-square test, based upon a partitioning of the observed data's

sample space into M disjoint cells, $E_1, ..., E_M$. The test statistic has the following form:

$$\hat{\chi}^2 = \frac{1}{m} \sum_{k=1}^{M} |N_k - g_k(\hat{\theta})|^2$$

where N_k are the number of observations in the kth cell, and

$$\begin{aligned} g_k(\theta) &= E_\theta(N_k) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n_i} p_{ijk}(\theta) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n_i} P(Y_{ij} \in E_k). \end{aligned}$$

such that $g_k(\hat{\theta})$ is the estimated value with minimum chi-square (MCE) parameter estimates plugged in (as opposed to ML estimates). Provided the subjects are independent, it can be shown that the asymptotic distribution of the $\hat{\chi}^2$ statistic is a mixture of chi-square random variables $\sum_{k=1}^{M} d_k Z_k^2$ where $Z_1, ..., Z_M$ are i.i.d. standard normal random variables and $d_1 \geq \cdots \geq d_M$ are the eigenvalues of $\Sigma = m^{-1} \sum_{i=1}^{m} Var(Y_i)$. To obtain the critical value of the statistic, the eigenvalues are replaced with their estimates, and H_0 : the model is a good fit is rejected if $\hat{\chi}^2$ exceeds the critical value of $\sum_{k=1}^{M} \hat{d}_k Z_k^2$.

The second proposed test is for GLMMs with crossed random effects. Crossed effects exist independently of each other. For example, in a study about the number of heart attacks, there might be a random subject effect and a random effect for brand of medical devices. We can observe a GLMM with crossed random effects in the following model equation for $i = 1, ..., n$ patients and $j = 1, ..., m$ medical devices.

$$g(Y_{ij}) = \mu + l_i + g_j$$

For this nested effect GLMM, the proposed test statistic is

$$\hat{\chi}^2 = \frac{1}{\sqrt{m^3}} \sum_{k=1}^{M} |N_k - E_{\hat{\theta}} N_k|^2$$

where $E_{\hat{\theta}} N_k$ is the expected number of observations in the kth cell plugging in parameter estimates found by the method of simulated moments (MSM). An asymptotic distribution is also provided for this statistic. Simulation studies show that both statistics have adequate size and power against alternatives.

6.2.3 Tests for Misspecified Random Effects Distributions

One of the key assumptions for GLMMs relates to the distribution of the random effects. In practice, random effects are often assumed to be normally distributed, as this often improves the tractability of the likelihood function. For LMMs, Verbeke and Lesaffre (1997) showed that the estimation of fixed effects and variance components is to some extent unaffected by a misspecified random effects distribution. However, research by Agresti, Caffo and Strickland (2004) has indicated that the same cannot be said for GLMMs. Further, Litiere, Alonso, and Molenberghs (2008) showed that variance component estimates were extremely biased under a misspecified random effects distribution. Abad and Litiere (2010) propose two diagnostic tests designed to detect GLMM misspecification of the random effects based upon a modified information matrix test.

6.3 Model Selection

Many model selection techniques, such as AIC, BIC, and likelihood ratio tests (for nested models), can be easily applied to GLMMs. Many of these techniques are described in Section 1.2.2. Along with these, Vonesh, Chinchilli, and Pu (1996) have developed a concordance statistic that is similar in interpretation to the coefficient of determination R^2 for linear regression.

Chapter 7

A Cramer-von-Mises Goodness-of-Fit Procedure for GLMMs

In this chapter, we aim to develop a omnibus goodness-of-fit test procedure for a GLMM. In Section 1.3.5, a Cramer-von-Mises (CVM) test is described for a Poisson regression model. We extend this procedure to two GLMMs, the Randomized Clinical Trial model and the Spatial model. These models have been previously discussed in Sections 5.5 and 5.6, respectively.

The main idea behind extending the test is to treat predictors of random effects as constants, which are then plugged into the GLMM model equations to approximate a GLM. Then, using this approximate GLM, the goodness-of-fit test can be performed. A simulation study is performed for the RCT model along with some discussion about how this procedure can be applied to other GLMMs.

7.1 The Proposed Goodness-of-Fit Test Statistic

An omnibus goodness-of-fit procedure will test H_0 : the GLMM model is a good fit versus H_a : the GLMM model is a poor fit. For a particular GLMM model, the test statistic can be calculated in the following manner:

First, fit the model parameters of the GLMM and use these to calculate the estimated predictors, $\hat{s}$. Then, approximate the GLMM with a GLM by substituting in the predictors

for the random effects. Our model then becomes

$$g(\mu_i) = x_i^{'}\hat{\beta} + z_i^{'}\hat{s} = x_i^{*'}\hat{\beta}^*$$

and $\hat{\mu_i} = g^{-1}(x_i^{*'}\hat{\beta}^*)$. Following Spinelli et al. (2002), use $\hat{\mu}_i$ and make the probability integral transformation on Y_i, call this V_i, where

$$V_i = P_i(Y_i) \text{ where } \begin{cases} P_i(0) = 0 \\ P_i(j) = P(Y_i \leq j-1), \text{ for } j \geq 1. \end{cases} \tag{7.1}$$

For example, if our proposed distribution was a Poisson (λ_i) then $P_i(j) = P(Y_i \leq j - 1) = \sum_{k=0}^{j-1} \frac{\hat{\lambda}_i^k e^{-\hat{\lambda}_i}}{k!}$. (If our proposed distribution was a Binomial$(1, p_i)$ then $p_i(0) = 0$ and $p_i(1) = 1 - \hat{p}_i$, and so on for any choice of response distribution.) Then, the V_i's have a distribution function, that is, $F_i(t) = P(V_i \leq t)$, is, for $j = 0, 1, ...,$

$$F_i(t) = P_i(j+1), \quad P_i(j) \leq t < P_i(j+1)$$

Suppose $\tilde{F}_n(t)$ is the edf of the set V_i, and suppose the average of the estimated distribution functions is $F_{ave}(t) = n^{-1}\sum_{i=1}^n F_i(t)$. A special case is when $t = 0$, then $F_{ave}(0) = n^{-1}\sum_{i=1}^n F_i(0)$. We can then define the residual process

$$Z_n(t) = \sqrt{n}\{\tilde{F}_n(t) - F_{ave}(t)\}$$

This residual process $Z_n(t)$ is used to calculate our test statistic:

$$W_n^2 = n\int_0^1 Z_n^2(t)dt$$

If the model has been correctly specified, this statistic follows a null distribution that can be obtained via bootstrap simulation.

The W_n^2 test statistic, while written as an integral, can be calculated as a summation since the functions $\tilde{F}_n(t)$ and $F_{ave}(t)$ are both step functions. To illustrate this, as well as to obtain a better understanding of how to construct the statistic, we will look at a toy example of calculating the test statistic for $n = 2$ observations.

A Toy Example. Suppose that we have two observations from a process, $Y_1 = 0$ and $Y_2 = 2$, that is thought to be best modeled by a Poisson GLMM. After fitting the parameter values and plugging in the predictors, we obtain the estimates $\hat{\lambda}_1 = 1$ and $\hat{\lambda}_2 = 2$. Using these estimates we can easily tabulate the probability mass functions for both observations, as listed in Table 7.1.

Table 7.1: PMFs for $Y_1 \sim Poisson(\hat{\lambda}_1 = 1)$ and $Y_2 \sim Poisson(\hat{\lambda}_2 = 2)$.

k	0	1	2	3	4	5	6	7	8	9	...
$P(Y_1 = k)$	0.368	0.368	0.183	0.061	0.015	0.003	0.001	0.000	0.000	0.000	...
$P(Y_2 = k)$	0.135	0.271	0.271	0.180	0.090	0.036	0.012	0.003	0.001	0.000	...

Based on the data, the realized values of $V_1 = 0$ and $V_2 = 0.406$, so the empirical distribution function $\tilde{F}_n(t)$ for our toy example has the simple form:

$$\tilde{F}_n(t) = \begin{cases} \frac{1}{2} & 0 \leq t < 0.406 \\ 1 & 0.406 \leq t \leq 1 \end{cases}$$

Further, V_1 has a distribution function $F_1(t)$ and V_2 has a distribution function $F_2(t)$, $0 \leq t \leq 1$, and both are step functions that can be tabulated as in Table 7.2.

Table 7.2: The distributions of V_1 and V_2.

V_1 takes values:	0	0.368	0.736	0.920	0.981	0.996	0.999	0.9999	... 1
w/ prob.	0.368	0.368	0.183	0.061	0.015	0.003	0.001	0.000	... 0
cum. prob.	0.368	0.736	0.920	0.981	0.996	0.999	0.9999	0.99999	... 1
V_2 takes values:	0	0.135	0.406	0.677	0.857	0.947	0.983	0.995	... 1
w/ prob.	0.135	0.271	0.271	0.180	0.090	0.036	0.012	0.003	... 0
cum. prob.	0.135	0.406	0.677	0.857	0.947	0.983	0.995	0.999	... 1

We have plotted $F_1(t)$ and $F_2(t)$ in Figure 7.1. From this point, it would be straightforward to calculate $F_{ave}(t)$, which would be the average of the two distribution functions. Both $\tilde{F}_n(t)$ and $F_{ave}(t)$ are plotted side-by-side in Figure 7.2.

Figure 7.1: Step Functions $F_1(t)$ and $F_2(t)$ on the left and right, respectively.

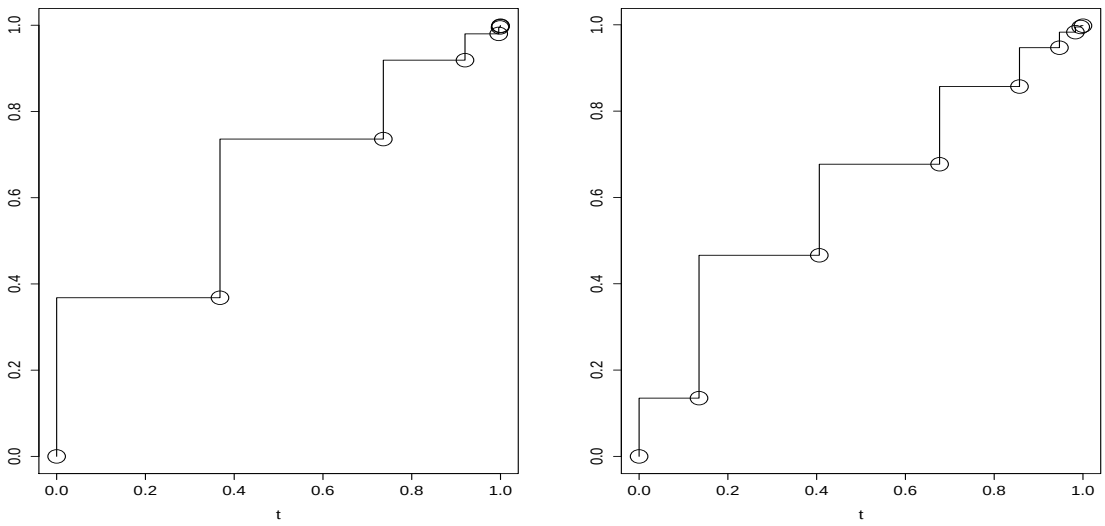
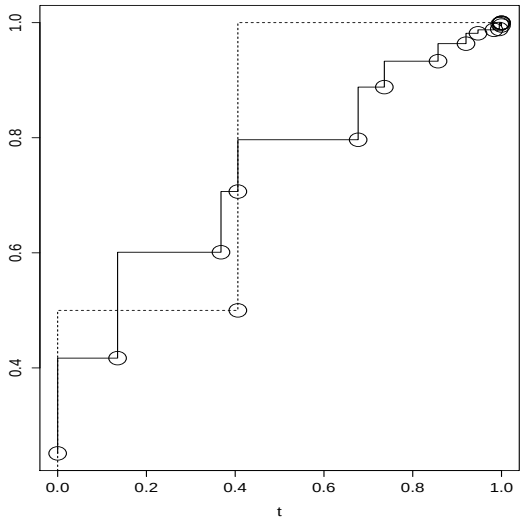


Figure 7.2: Empirical distribution function $\tilde{F}_n(t)$ (dashed line) and average estimated distribution function $F_{ave}(t)$ (solid line) for the toy example.



From Figure 7.2 it is clear that $F_{ave}(t)$ can only take a step each time that an individual $F_i(t)$ takes a step. The first step is at $t = 0$. Further, the edf $\tilde{F}_n$ must take steps up at one of the existing steps of $F_{ave}(t)$.

If the null distribution is appropriate, we should expect to see the two step functions very close to each other. These two lines are not very close to each other since we only have a sample size of $n = 2$. However, from this example we can get an idea of how to construct the statistic, which is based upon the difference in the two functions. It is clear that the difference between the two functions can be calculated as a summation over values of t. At some point, the largest V_i will put the value of $\tilde{F}_n = 1$, while the average function will still be increasing in tiny increments indefinitely. However, this distance can be easily bounded. For example, suppose that $t = 0.999$ and that $\tilde{F}_n(t)$ has reached 1 and that $F_{ave}(t) = 0.9998$. Then, the remaining distance between the two functions certainly must be less than $(width)(height) = (1-0.999)(1-0.9998) = 0.0000002$. In our simulation study we use a cutoff of 1.0 x 10^{-6} for the remaining distance.

7.2 Simulation Study: CVM for the RCT Model

We can apply the CVM procedure to the RCT Model discussed in Section 5.5. Recall,

$$log(\lambda_{ijk}) = \theta + \beta X_j + s_i$$

where $Y_{ijk}|\lambda_{ijk} \sim Poisson(\lambda_{ijk})$ and $s_i \sim$ iid $N(0, \sigma^2)$ $i = 1, ..., I$ clinics, $j = 1, 2$ treatments and $k = 1, ..., K$ patients per clinic x treatment. $X_j = 1$ for treatment, 0 otherwise.

Data can be generated from this model at any combination of fixed values of I,J,K and θ, β, σ. We chose design values: $I = 10$, $J = 2$, $K = 4$, 8, or 20 (for a total of 80, 160, or 400 observations). Further, two sets of parameters were selected,

A: $\theta = 2.5$, $\beta = -1.5$, and s_i generated from a $N(0, 0.2^2)$ and
B: $\theta = -0.4$, $\beta = 2.1$, and s_i generated from a $N(0, 0.3^2)$.

These parameter values were chosen in order to create datasets that might reasonably reflect recovery time from a certain procedure/disease.

For the RCT model, we are proposing an omnibus test of

$$H_0 : Y_{ijk}|\lambda_{ijk} \sim Poi(\lambda_{ijk}).$$

We can evaluate the performance of the CVM test using the following bootstrap procedure:

1. Generate a vector of random effects $s = (s_1, ..., s_I)'$ for a set of design values.

2. Generate a dataset D from the RCT model for fixed θ, β and I,J,K.

3. Fit the model parameters via quadrature (glmer() function in R), to obtain $\hat{\theta}, \hat{\beta}, \hat{\sigma}$.

4. Using the parameter estimates, obtain predictors for the random effects: $\hat{s} = (\hat{s_1}, ..., \hat{s_I})'$. Here, we used eBLPs to find our predictors. A closed form equation for the eBLPs is provided in Li, Jeske, and Klein (2011).

5. Calculate the CVM statistic T using $\hat{\lambda}_{ijk} = exp(\hat{\theta} + \hat{\beta}X_j + \hat{s_i})$.

This allows us to obtain a single test statistic value. To get a bootstrap p-value requires further simulation:

6. Fix the $\hat{s} = (\hat{s_1}, ..., \hat{s_I})$ as the random effects.

7. Generate m=1000 null datasets for fixed $\hat{\theta}$, $\hat{\beta}$ and I,J,K

8. From these datasets, 1000 CVM statistics T^*s can be calculated.

9. The p-value is the number of $(T^* > T)/m$.

This generates a single bootstrap p-value. To explore the size and power of the test, many bootstrap p-values need to be generated. The results of our simulation study are in the next section.

7.2.1 Size

To evaluate the size of the proposed CVM test, 5,000 datasets were generated from the RCT model under different conditions using R software 2.14.2. Simulation study results are shown in Table 7.3, for two sets of parameter values (A and B) and for three values of K=4, 8, and 20. We found that the CVM test procedure is slightly conservative at the smaller sample sizes. However, as K increases the test has adequate size. Along with size, four power alternatives are discussed in the next section.

Table 7.3: Proportion of p-values out of 5,000 less than or equal to the listed cutoffs for two sets of parameter values (A or B) and three values of K.

A	K=4	K=8	K=20	B	K=4	K=8	K=20
≤ 0.01	0.009	0.008	0.013	≤ 0.01	0.007	0.007	0.009
≤ 0.05	0.044	0.045	0.052	≤ 0.05	0.040	0.039	0.048
≤ 0.10	0.092	0.087	0.099	≤ 0.10	0.084	0.088	0.098

7.2.2 Power Alternative: Overdispersed Poisson

Here, we generated Y_{ijk} values from a Negative Binomial with mean=λ_{ijk} and variance=$\lambda_{ijk}+\lambda_{ijk}^2/\kappa$. The percentage of p-values under the cutoffs are given in Table 7.4. The test procedure appears to have excellent power for this alternative. As expected, the power increases as κ decreases (i.e., the variance increases).

Table 7.4: Percentage of p-values out of 1,000 less than or equal to the listed cutoffs for the Negative Binomial alternative.

A, K=4	κ=100	κ=10	κ=8	κ=5	κ=1	κ=0.1
≤ 0.01	0.6%	22.2%	32.7%	65.7%	100%	100%
≤ 0.05	3.1%	41.9%	54.9%	85.7%	100%	100%
≤ 0.10	6.9%	53.8%	66.8%	91.4%	100%	100%
A, K=8	κ=100	κ=10	κ=8	κ=5	κ=1	κ=0.1
≤ 0.01	0.4%	65.8%	81.0%	98.5%	100%	100%
≤ 0.05	4.6%	83.2%	91.9%	99.7%	100%	100%
≤ 0.10	9.0%	90.5%	95.1%	100%	100%	100%
A, K=20	κ=100	κ=10	κ=8	κ=5	κ=1	κ=0.1
≤ 0.01	1.3%	96.6%	100%	100%	100%	100%
≤ 0.05	7.1%	100%	100%	100%	100%	100%
≤ 0.10	14.7%	100%	100%	100%	100%	100%
B, K=4	κ=100	κ=10	κ=8	κ=5	κ=1	κ=0.1
≤ 0.01	0.2%	1.9%	4.0%	9.9%	94.0%	100%
≤ 0.05	4.0%	8.3%	12.3%	25.2%	98.2%	100%
≤ 0.10	9.3%	15.0%	19.7%	37.5%	99.2%	100%
B, K=8	κ=100	κ=10	κ=8	κ=5	κ=1	κ=0.1
≤ 0.01	0.4%	8.0%	15.5%	43.9%	100%	100%
≤ 0.05	3.3%	24.6%	34.8%	65.5%	100%	100%
≤ 0.10	8.2%	37.8%	49.5%	77.1%	100%	100%
B, K=20	κ=100	κ=10	κ=8	κ=5	κ=1	κ=0.1
≤ 0.01	1.9%	43.8%	59.1%	94.7%	100%	100%
≤ 0.05	7.2%	68.6%	81.0%	99.1%	100%	100%
≤ 0.10	12.0%	76.6%	87.2%	99.7%	100%	100%

7.2.3 Power Alternative: Missing Covariate- Gender

For this alternative, we assume that there is a missing variable, say Gender, that has great explanatory power in the model. To test if our procedure can detect such a missing

covariate, data was generated from the following model:

$$log(\lambda_{ijk}) = \theta + \beta X_j + s_i + \omega V_{ijk}$$

where $V_{ijk} \sim Bernoulli(p = 0.5)$ for some weight ω, as if the patients have a 50/50 chance of being male or female. We can fix $\omega = 0.1, 0.5$, or 1. The percentage of p-values less-than-or-equal-to the cutoffs is given in Table 7.5. Again, we see good power as the strength of relationship increases.

Table 7.5: Percentage of p-values out of 1,000 less than or equal to the listed cutoffs for the Missing Covariate-Gender alternative.

A, K=4	ω=0.1	ω=0.5	ω=1	A, K=8	ω=0.1	ω=0.5	ω=1	A, K=20	ω=0.1	ω=0.5	ω=1
≤ 0.01	0.7%	18.1%	100%	≤ 0.01	0.5%	60.4%	100%	≤ 0.01	0.8%	99.5%	100%
≤ 0.05	3.2%	37.0%	100%	≤ 0.01	3.2%	79.6%	100%	≤ 0.01	4.6%	99.8%	100%
≤ 0.10	7.6%	48.4%	100%	≤ 0.01	7.4%	87.1%	100%	≤ 0.01	8.7%	100%	100%
B, K=4	ω=0.1	ω=0.5	ω=1	B, K=8	ω=0.1	ω=0.5	ω=1	B, K=20	ω=0.1	ω=0.5	ω=1
≤ 0.01	0.4%	1.5%	69.1%	≤ 0.01	0.4%	6.9%	100%	≤ 0.01	0.7%	32.2%	100%
≤ 0.05	3.5%	7.6%	83.3%	≤ 0.01	4.2%	21.0%	100%	≤ 0.01	4.4%	57.7%	100%
≤ 0.10	8.9%	14.2%	92.3%	≤ 0.01	8.5%	30.1%	100%	≤ 0.01	8.7%	68.3%	100%

7.2.4 Power Alternative: Missing Covariate- Over the Counter

Here, we again assume there is a missing variable; however this missing covariate can only vary at a clinic-treatment level. This can be described in the following model:

$$log(\lambda_{ijk}) = \theta + \beta X_j + s_i + \omega V_{ij}$$

where $V_{ijk} \sim Bernoulli(p = 0.5)$ for some weight ω, such that all the patients within each treatment x clinic receive the same form of treatment, such as over the counter medicine. We can fix $\omega = 0.1, 0.5$, or 1. Table 7.6 displays the simulation results. Compared to the Gender alternative, the power is not as strong. However, we do see power improve as the sample size increases.

Table 7.6: Percentage of p-values out of 1,000 less than or equal to the listed cutoffs for the Missing Covariate-Over the Counter alternative.

A, K=4	ω=0.1	ω=0.5	ω=1	A, K=8	ω=0.1	ω=0.5	ω=1	A, K=20	ω=0.1	ω=0.5	ω=1
≤ 0.01	1.1%	1.0%	38.7%	≤ 0.01	0.7%	2.7%	70.8%	≤ 0.01	0.5%	12.3%	93.1%
≤ 0.05	4.2%	5.0%	55.7%	≤ 0.01	4.0%	9.5%	83.4%	≤ 0.01	2.8%	25.8%	95.5%
≤ 0.10	8.9%	10.5%	63.1%	≤ 0.01	8.5%	16.6%	88.0%	≤ 0.01	7.1%	36.4%	96.8%
B, K=4	ω=0.1	ω=0.5	ω=1	B, K=8	ω=0.1	ω=0.5	ω=1	B, K=20	ω=0.1	ω=0.5	ω=1
≤ 0.01	1.3%	1.2%	1.2%	≤ 0.01	0.5%	2.5%	6.2%	≤ 0.01	1.2%	1.0%	21.5%
≤ 0.05	3.6%	3.7%	7.0%	≤ 0.01	2.8%	7.5%	16.5%	≤ 0.01	7.1%	5.8%	36.3%
≤ 0.10	8.6%	8.9%	12.1%	≤ 0.01	7.9%	11.4%	26.4%	≤ 0.01	11.3%	11.8%	46.8%

7.2.5 Power Alternative: Full Interaction Model

For this alternative, we assume that there is an interaction effect present. We generate data from the following alternative model:

$$log(\lambda_{ijk}) = \theta + \beta_0 X_j + s_i + t_{ij}$$

where t_{ij} is a random interaction effect between clinic and treatment, $i = 1, .., I$, $j = 1, 2$. The t_{ij}'s are i.i.d. $N(0, \sigma_t^2)$ where $\sigma_t = 0.01, 0.2$ or 0.8. Results are listed in Table 7.7. Power increase as σ_t increases. Also, power increases with sample size.

Table 7.7: Percentage of p-values out of 1,000 less than or equal to the listed cutoffs for the Full Interaction model alternative.

A, K=4	σ_t=0.01	σ_t=0.2	σ_t=0.8	B, K=4	σ_t=0.01	σ_t=0.2	σ_t=0.8
≤ 0.01	0.7%	0.6%	56.6%	≤ 0.01	0.9%	0.2%	8.8%
≤ 0.05	4.1%	3.5%	71.4%	≤ 0.01	4.8%	4.0%	17.3%
≤ 0.10	7.3%	8.4%	77.5%	≤ 0.01	8.5%	8.4%	25.1%
A, K=8	σ_t=0.01	σ_t=0.2	σ_t=0.8	B, K=8	σ_t=0.01	σ_t=0.2	σ_t=0.8
≤ 0.01	0.7%	1.1%	84.1%	≤ 0.01	0.6%	0.7%	23.0 %
≤ 0.05	4.1%	3.8%	90.7%	≤ 0.01	4.3%	3.1%	38.1%
≤ 0.10	10.0%	8.0%	93.3%	≤ 0.01	10.1%	10.3%	42.0%
A, K=20	σ_t=0.01	σ_t=0.2	σ_t=0.8	B, K=20	σ_t=0.01	σ_t=0.2	σ_t=0.8
≤ 0.01	1.0%	2.5%	97.2%	≤ 0.01	1.8%	0.8%	52.2%
≤ 0.05	5.1%	9.1%	98.4%	≤ 0.01	5.8%	6.3%	66.7%
≤ 0.10	10.7%	16.4%	99.2%	≤ 0.01	10.5%	10.2%	73.9%

7.3 Size Study: CVM for the Spatial Model

We can also apply the CVM procedure to the Spatial model discussed in Section 5.6. Recall,

$$\begin{aligned} Y_{j(i)}|s \text{ indep. } &\sim Negative\ Binomial(\theta_i, \kappa), i = 1, 2, ..., nm, j = 1, 2, ..., n_i; \\ log(\theta_i) &= \mu + s_i; \\ s = (s_1, s_2, ..., s_{nm})' &\sim MVN(0, \sigma^2 I), \end{aligned}$$

where $Y_{j(i)}$ is the count number from the jth sampling unit in the ith co-cluster, n_i is the total number of sampling units in the ith co-cluster, θ_i is the conditional mean of counts associated with the ith co-cluster, κ is the dispersion parameter, and s_i is the random effect associated with each co-cluster.

Data can be generated from this model at any combination of fixed values of n, m and n_i and parameters μ, dispersion parameter κ, and σ. We selected $m = 4$ rows and $n = 4$ columns, for a total of $4 \cdot 4 = 16$ separate co-clusters. For simplicity, we also set $n_i = 5$ or $n_i = 10$ for all co-clusters. Two sets of parameter values were chosen:

$$\begin{aligned} &\text{A: } \mu = 0.5, \kappa = 2.0, \text{ and } s_i \text{ generated from a } N(0, 1^2) \\ &\text{B: } \mu = 1.2, \kappa = 0.4, \text{ and } s_i \text{ generated from a } N(0, 1^2) \end{aligned}$$

These parameter values were chosen in order to create datasets that might reasonably reflect pest density in an orchard.

For this Spatial model, we are proposing an omnibus test of

$$H_0 : Y_{j(i)}|\theta_i \sim Negative\ Binomial(\theta_i).$$

We evaluate the performance of the CVM test using a similar procedure as outlined for the RCT model. The bootstrap procedure is as follows:

1. Generate a vector of random effects $s = (s_1, ..., s_I)'$ for a given σ.
2. Generate a dataset D from the Spatial model for fixed μ, κ and nm, n_i.
3. Fit the model parameters via quadrature (optim() function in R), to obtain $\hat{\mu}, \hat{\kappa}, \hat{\sigma}$.
4. Using these parameter estimates, obtain best predictors (BPs) for the random effects: $\hat{s} = (\hat{s_1}, ..., \hat{s_I})'$, which also requires quadrature.

5. Calculate the CVM statistic T using $\hat{\theta}_i = exp(\hat{\mu} + \hat{s}_i)$.

This allows us to obtain a single test statistic value. To get a bootstrap p-value requires further simulation:

6. Fix the $\hat{s} = (\hat{s_1}, ..., \hat{s_I})$ as the random effects.

7. Generate m=1000 null datasets for fixed $\hat{\mu}$, $\hat{\kappa}$ and nm, n_i.

8. From these datasets, 1000 CVM statistics T^*s can be calculated.

9. The p-value is the number of $(T^* > T)/m$.

This generates a single bootstrap p-value. To explore the size or power of the test, many bootstrap p-values need to be generated.

To evaluate the size of the CVM test, 2,000 datasets were generated from the Spatial model under different conditions using R software 2.14.2. Size results are shown for two sets of parameter values in Table 7.8. As in the case of the RCT model, we found the CVM test to be slightly conservative.

Table 7.8: Proportion of p-values out of 2,000 less than or equal to the listed cutoffs for two sets of parameter values where $n_i = 5$ or 10.

	A , $n_i = 5$	A, $n_i = 10$	B, $n_i = 5$	B, $n_I = 10$
≤ 0.01	0.8%	0.8%	0.7%	0.9%
≤ 0.05	5.1%	4.5%	4.2%	4.8%
≤ 0.10	8.9%	8.6%	9.9%	10.3%

Currently, simulation study has not yet been done to demonstrate power for the Spatial model. One alternative of interest is a situation where the random effects are correlated with each other, and that the degree of correlation is proportional to the distance between the midpoints of each pair of co-clusters. Another alternative of interest is a missing covariate, such as a spot treatment effect. Here, we could assume that there is a missing binary explanatory variable that corresponds to a co-cluster being "spot treated" for pests at an earlier time. Thus, those co-clusters which have been spot treated should experience lower pest density.

7.4 The CVM Test Applied to Other GLMMs

We have seen the CVM procedure perform well for the RCT model. The test is based upon an initial probability integral transformation (PIT) of the data. This property makes the CVM procedure extremely flexible, since any GLMM response distribution (conditional on the random effects) can have an associated PIT. We believe this procedure can be applied to the entire class of GLMMs, provided that reasonable predictors can be obtained for the random effects in the model. We will begin by discussing appropriate PITs for a variety of situations.

Recall, in the case of the RCT and Spatial models, the transformation was given by

$$V_i = P_i(Y_i) \text{ where } \begin{cases} P_i(0) = 0 \\ P_i(j) = P(Y_i \leq j-1). \end{cases}$$

This same transformation would also be appropriate for a logistic GLMM. For example, suppose we were looking at a logistic GLMM where the response $Y_i|s$ takes a multinomial distribution with four possible values: 0, 1, 2, or 3. Unlike in the case of the Poisson and Negative Binomial, the multinomial has a finite number of possible values. This actually slightly simplifies the CVM statistic calculations.

As before, $\tilde{F}_n(t)$ is the empirical distribution function of the V_i's and can be calculated in a straightforward manner. However, the average distribution function $F_{ave}(t)$ will stop increasing and reach the value of 1 at $t = sup_i|V_i|$. Thus, we avoid the problem of an ever shrinking, right-hand-side gap between the two distributions that we face in the Poisson and Negative Binomial case.

Although the Poisson distribution can take an infinite number of values, there is a definite "anchor" value at 0. This features prominently into the definition of V_i, since $V_i = 0$ if $Y_i = 0$. However, 0 may not always be the anchoring value of the response. Further, there may not even be an anchoring value. Suppose that a GLMM response distribution takes possible values from the set of integers. Although this set is countable, there is no smallest value among them. Consequently, consider placing an artificial anchoring value into the transformation, as guided by the data. Let $Y^* < min(Y_1, Y_2, ..., .Y_n)$ be some value below the smallest observed value. Then, we can define the PIT in the following manner:

$$V_i = P_i(Y_i) \text{ where } \begin{cases} P_i(j) = 0, \quad j \leq Y^* \\ P_i(j) = P(Y_i \leq j-1), \quad j > Y^*. \end{cases}$$

From this point, the calculation of the statistic is straightforward.

Having such an anchoring value simplifies the calculation of the statistic, however it is not required. Just as there is an ever shrinking, right-hand-side gap between the edf and the average function, there will be a similar left-hand-side gap without a defined anchor value. However, as in the case of the right-hand-side, this distance can be bounded.

Some GLMMs follow continuous distributions. Here, the definition of V_i as described is not appropriate, since the data no longer takes values that are one unit apart from each other. Suppose $Y_i|s$ are independent realizations from the continuous distribution $f_{Y_i}(\cdot|\mu_i, \theta)$. A probability integral transformation can be defined such that

$$V_i = \int_{-\infty}^{Y_i} f_{Y_i}(x|\hat{\mu}_i, \hat{\theta})dx = F(Y_i|\hat{\mu}_i, \hat{\theta}) \tag{7.2}$$

where $F(\cdot)$ is the distribution function associated with $f_{Y_i}(\cdot)$ and $\hat{\mu}_i = x_i'\hat{\beta} + z_i'\hat{s}$, where $\hat{s}$ are the random predictors.

Then, to construct the CVM statistic for a continuous distribution, first make the PIT defined in (7.2), transforming $Y_1, Y_2, ..., Y_n$ into $V_1, V_2, ..., V_n$. Using the V_i, the edf $\tilde{F}_n(t)$ can be constructed. Let $\tilde{F}_n(0) = 0$. Note that this will still be a step function.

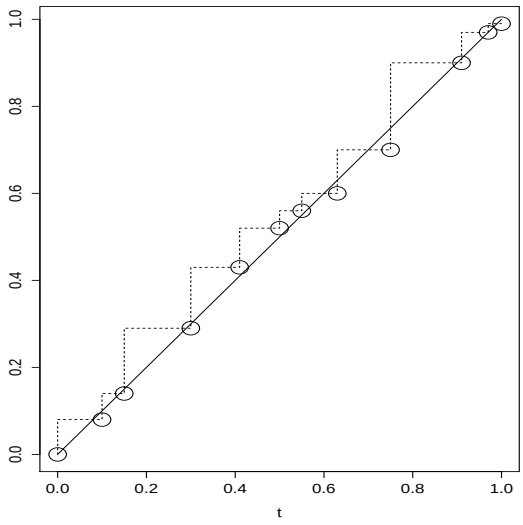
In contrast, the distributions $F_i(t)$ are no longer step functions since the responses are continuous. Yet by properties of the PIT given in (7.2) for continuous distributions, the V_i's will all have an approximate Uniform(0,1) density under the null. This implies that $F_i(t) = t$ for $0 \leq t \leq 1$ and 0 otherwise. Since $F_{ave}(t)$ is an average of these distributions, it follows that $F_{ave}(t) = F_i(t) = t$ for $0 \leq t \leq 1$ and 0 otherwise. A hypothetical illustration of the $\tilde{F}_n(t)$ and $F_{ave}(t)$ plotted side-by-side is provided in Figure 7.3.

Recall, our statistic $W_n^2 = n\int_0^1 Z_n^2(t)dt$ where $Z_n(t) = \sqrt{n}\{\tilde{F}_n(t) - F_{ave}(t)\}$. This integral of this residual process $Z_n^2(t)$ can be calculated as a sum of disjoint integrals. Suppose that $\tilde{F}_n(t)$ takes a jump at $t_1, ..., t_m$. Then, it follows that

$$\int_0^1 Z_n^2(t)dt = \int_0^{t_1} n(0-t)^2dt + \int_{t_1}^{t_2} n(\tilde{F}_n(t_1) - t)^2dt + \cdots + \int_{t_m}^1 n(1-t)^2dt$$

This approach is quite similar to a Kolmogorov-Smirnov test, where a PIT is applied to a dataset and then tested to see if it could reasonably come from a Uniform distribution.

Figure 7.3: A hypothetical empirical distribution function $\tilde{F}_n(t)$ (dashed line) and average estimated distribution function $F_{ave}(t)$ (solid line) for a GLMM with a continuous response.



Chapter 8

Summary and Future Work

Herein we have derived an asymptotically correct chi-square goodness-of-fit test for ALR models that relies upon the construction unique path probabilities over time. We have developed a two-dimensional, dynamic binning strategy to obtain cell counts in a proper way. We have also demonstrated size and power with a simulation study.

The examples provided have made use of ordinal logits, but the test can be easily extended to ALR models with nominal logits. Further, the procedure can be applied to any multi-state ALR model with or without an absorbing state (or states), as long as unique paths can be articulated. We have also provided a real world application to an Alzheimer's disease dataset provided by Loma Linda University.

In addition to this, we have presented a Cramer-von-Mises goodness-of-fit test for GLMMs. We have shown that it performs well for the RCT model, in terms of having adequate size and detecting misspecification, and looks promising for the Spatial model. This analysis has opened the door for future research.

Clearly, there is an opportunity to determine the asymptotic distribution of the CVM statistic for GLMMs. This would add significantly to the usefulness of the test statistic. Additional simulation study is planned for the Spatial model. Further, the CVM test procedure should also be compared to other omnibus GLMM goodness-of-fit tests, such as those proposed by Pan and Lin (2005) or Gu (2008), discussed in Sections 6.2.1 and 6.2.2.

The CVM test uses predictors of random effects to approximate a GLM. This same approach might also be applied to other goodness-of-fit tests for GLMs, thereby making them appropriate for GLMMs. Take for example the chi-square test for Poisson regression, which could be modified for the RCT GLMM. Recall, we have $log(\lambda_{ijk}) = \theta + \beta X_j + s_i$

where $Y_{ijk}|\lambda_{ijk} \sim Poisson(\lambda_{ijk})$ and $s_i \sim$ iid $N(0, \sigma^2)$ $i = 1, ..., I$ clinics, $j = 1, 2$ treatments and $k = 1, ..., K$ patients per clinic-treatment, and $X_j = 1$ for treatment 1, 0 otherwise.

For the ij-th clinic-treatment, we can compute a χ^2_{ij}. Let $A_{ij1}, A_{ij2}, ..., A_{ijL}$ be a collection of disjoint sets that cover the positive integers (zero inclusive) and let

$$p_{ijl} = P_{\lambda_{ijk}}\{Y_{ijk} \in A_{ijl}\} > 0$$

where $\lambda_{ijk} = e^{\theta+\beta X_j+s_i}$ and where $P_{\lambda_{ijk}}$ is the probability distribution of $Poisson(\lambda_{ijk})$. Let $Z_{ijl} = \#$ of $Y_{ijk} \in A_{ijl}$. Then, under some regularity conditions, it is clear that

$$X^2_{ij} = \sum_{l=1}^{L} \frac{Z_{ijl} - Kp_{ijl}}{Kp_{ijl}} \longrightarrow \chi^2_{L-1}$$

Then, for the whole dataset, we can get one chi-square statistic by adding up the independent chi-squares:

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} X^2_{IJ} \longrightarrow \chi^2_{IJ(L-1)}$$

The above test can be used as an overall test for specific values of θ, β, and the s_i's, although this type of test is probably not very useful.

However, it is reasonable that an omnibus goodness-of-fit test for the RCT GLMM could be constructed by plugging in estimates $\hat{\theta}, \hat{\beta}$ along with random predictors $\hat{s}_i$. Some care would need to be taken in finding the parameter estimates and the predictors, and in forming the bins. Also, the statistic's degrees of freedom will have to be adjusted appropriately to account for the parameter estimation.

Bibliography

[1] Abad, A.A., & Litiere S. (2010). Testing for misspecification in generalized linear mixed models. Biostatistics, 11(4), 771-786.

[2] Abramowitz, M., & Stegun, I.(Eds.). (1964). Handbook of mathematical functions. National Bureau of Standards, Washington, D.C.

[3] Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions of Automatic Control, 19, 716-723.

[4] Agresti, A. (2003). Categorical data analysis. Hoboken, New Jersey: John Wiley & Sons.

[5] Agresti, A., Caffo, B., & Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. Computational Statistics and Data Analysis, 47, 639-653.

[6] Bonney, G.E.(1987). Logistic regression for dependent binary observations. Biometrics, 43, 951-973.

[7] Calvin, J.A., & Sedransk, J. (1991). Bayesian and frequentist predictive inference for the patterns of care studies. Journal of the American Statistical Association, 86, 36-48.

[8] Chan, J.S.K. (2000). The initial state problem in autoregressive binary regression. Journal of the Royal Statistical Society Series D (The Statistician), 49(4), 495-502.

[9] Chernoff, H. & Lehmann, E. L. (1954). The use of maximum likelihood estimates in χ^2 tests for goodness of fit. Annals of Mathematical Statistics, 25, 579-586.

[10] Christensen, R., Pearson, L. M., & Johnson, W. (1992). Case-deletion diagnostics for mixed models. Technometrics, 34, 38-45.

[11] Claeskens, G., & Hart, J.D. (2009). Goodness-of-fit tests in mixed models. Test, 18, 213-239.

[12] Cochran, W. G. (1954). Some methods for strengthening the common chi-square test. Biometrics,10, 417-451.

[13] Cox, D.R. (1970). The analysis of binary data. London: Methuen.

[14] D'Agostino, R.B., & Stephens M.A. (Eds.). (1986). Goodness-of-fit techniques. New York: Marcel Dekker, Inc.

[15] Dean, C., & Lawless, J.F. (1989). Testing for overdispersion in poisson regression models. Journal of the American Statistical Association, 84, 467-472.

[16] de Vries, S., Fidler, V., Kuipers, W., & Hunink, M. (1998). Fitting multistate models with autoregressive logistic regression: supervised exercise in intermittent claudication. Journal of Medical Decision Making, 18, 52-60.

[17] Fagerland, M., Hosmer D., & Bofin, A. (2008). Multinomial goodness-of-fit tests for logistic regression models. Statistics in Medicine, 27, 4238-4253.

[18] Gu, Z. (2008). Model diagnostics for generalized linear mixed models. (Doctoral dissertation, University of California, Davis, 2008).

[19] Heagerty, P.J., & Kurland, B.F. (2001). Misspecified maximum likelihood estimates and generalized linear mixed models. Biometrika, 88, 973-985.

[20] Henderson, C. R. (1984). Applications of linear models in animal breeding. University of Guelph.

[21] Hocking, R. R. (1976). The analysis and selection of variables in linear regression. Biometrics, 32, 1-50.

[22] Hosmer, D., & Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. Communications in Statistics, A10, 1043-1069.

[23] Hwang, Y.T., & Wei, P.F. (2006). A novel method for testing normality in a mixed model of a nested classification. Computational Statistics and Data Analysis, 51, 1163-1183.

[24] Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J.M., & Thiebaut, R. (2007). Robustness of the linear mixed model to misspecified error distribution. Computational Statistics and Data Analysis, 51, 5142-5154.

[25] Jeske, D.R., Lockhart R.A., Stephens, M.A., & Zhang, Q. (2008). Cramer-von-mises tests for the compatibility of two software operating environments. Technometrics, 50(1), 53-63.

[26] Jiang, J. (2001). Goodness-of-fit for mixed model diagnostics. The Annals of Statistics, 29(4), 1137-1164.

[27] Kirsch, W., McAuley, G., Holshouser, B., Petersen, F., Ayaz, M., Vinters, H., et al. (2009). Serial susceptibility weighted MRI measures brain iron and microbleeds in dementia. Journal of Alzheimer's Disease, 17, 599-609.

[28] Lawless, J.F. (1987). Negative binomial and mixed poisson regression. Canadian Journal of Statistics, 15, 209-255.

[29] Li, J.X., Jeske, D.R., & Klein, J.A. (2012). Sequential analysis methodology for a poisson glmm with applications to multicenter randomized clinical trials. Journal of Statistical Planning and Inference, 142(12), 3225-3234.

[30] Litiere, S., Alonso, A., & Molenberghs, G. (2008). The impact of a misspecified random-effects distribution on maximum likelihood estimation in generalized linear mixed models. Statistics in Medicine, 27, 3125-3144.

[31] Lindstrom, M. J., & Bates, D. M. (1988). Newton-raphson and EM agorithms for linear mixed-effects models for repeated-measures data. Journal of the American Statistical Association, 83, 1014-1022.

[32] McCulloch, C.E., Searle, S.R., & Neuhaus J.M. (2008). Generalized, linear, and mixed models. Hoboken, New Jersey: John Wiley & Sons.

[33] Mendenhall, W., & Sincich, T. (2003). A second course in statistics regression analysis. Upper Saddle River, New Jersey: Pearson Education, Inc.

[34] Moore, D. (1986). Tests of chi-square type. In D'Agostino R. & Stephens M. (Eds.), Goodness-of-fit techniques (pp. 63-95). New York: Marcel Dekker, Inc.

[35] Mueller, K., Voelkle, M.C., & Hattrup, K. (2011). On the relationship between job satisfaction and non-response in employee attitude surveys: a longitudinal field study. Journal of Occupational and Organizational Psychology, 84, 780-798.

[36] Nagelkerke, N., Smits, J., le Cessie, S., & van Houwelignen, H. (2005). Testing goodness-of-fit of the logistic regression model in case-control studies using sample reweighting. Statistics in Medicine, 24, 121-130.

[37] Pulkstenis, E., & Robinson, T. (2002). Two goodness-of-fit tests for logistic regression models with continuous covariates. Statistics in Medicine, 21, 79-93.

[38] Raftery, A.E. (1986). Choosing models for cross-classifications. American Sociological Review, 51(1), 145-146.

[39] Razali, N.M., & Wah, Y.B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and and anderson-darling tests. Journal of Statistical Modeling and Analytics, 2(1), 21-33.

[40] Pan, Z., & Lin, D.Y. (2005). Goodness-of-fit methods for generalized linear mixed models. Biometrics, 61, 1000-1009.

[41] SAS. (2012). User's Guide 12.1.

[42] Slud, E., & Kedem, B. (1994). Partial likelihood analysis of logistic regression and autoregression. Statistica Sinica, 4, 89-106.

[43] Spinelli, J.J., Lockhart, R.A., & Stephens, M.A. (2002). Tests for the response distribution in a poisson regression model. Journal of Statistical Planning and Inference, 108, 137-154.

[44] Su, X. (2007). Tree-based model checking for logistic regression. Statistics in Medicine, 26, 2154-2169.

[45] Tsiatis, A. (2002). A note on a goodness-of-fit for the logistic regression model. Biometrika, 67, 250-251.

[46] Verbeke, G., & Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in a linear mixed model for longitudinal data. Computational Statistics and Data Analysis, 23, 541-556.

[47] Vonesh, E.F., Chinchilli, V.M., & Pu, K. (1996). Goodness-of-fit in generalized nonlinear mixed effects models. Biometrics, 52, 572-587.

[48] Zhang, Z., Jeske, D.R., Cui, X., & Hoddle, M. (2012). Co-clustering spatial data using a generalized linear mixed model with application to the integrated pest management. Journal of Agricultural, Biological, and Environmental Statistics, 17, 265-282.