2009

Logistic Models with Missing Categorical Covariates

Jeremiah Rounds
Utah State University

Follow this and additional works at: https://digitalcommons.usu.edu/gradreports

Part of the Mathematics Commons, and the Statistics and Probability Commons

Recommended Citation
https://digitalcommons.usu.edu/gradreports/1242
LOGISTIC MODELS
WITH MISSING CATEGORICAL COVARIATES

by

Jeremiah Rounds

A report submitted in partial fulfillment
of the requirements for the degree
of
MASTER OF SCIENCE
in
Statistics

UTAH STATE UNIVERSITY
Logan, Utah
2009
Abstract

Logistic Models
With Missing Categorical Covariates

by

Jeremiah Rounds, Master of Science
Utah State University, 2009

Major Professor: Dr. John R. Stevens
Department: Mathematics And Statistics

We present an EM based solution to missing categorical covariates in Binomial models with logit links using an assumption that experimental units are drawn from a Multinomial population of infinite size. We further address the problem of separation of points inducing large variances on parameter estimates by the use of a novel score-modification based on Firth’s bias-reduction score-modification. We simulate to address questions about estimate bias, distribution, and appropriate parameter coverage by Wald intervals.

(70 pages)
Acknowledgments

This work was supported by generous research assistantships from the Department of Mathematics and Statistics, Utah State University, under the advising of Dr. Stevens; for which, I am very grateful.
# Contents

Abstract .......................................................................................................................... iii
Acknowledgments ............................................................................................................ iv
List of Tables ................................................................................................................... vii
List of Figures .................................................................................................................. viii

## 1 Introduction

1.1 Phosphine Fumigant Dose-Response For *Rhyzopertha dominica* ......................... 2

## 2 Model

2.1 Multinomial Binomial Hierarchical Model ............................................................... 5
  2.1.1 Multinomial Model .......................................................................................... 5
  2.1.2 Binomial Model .............................................................................................. 6
2.2 Missing Categorical Covariates On The Failure Observations ............................... 7
2.3 Returning To *Rhyzopertha dominica* .................................................................. 7

## 3 Parameter Estimation with Expectation-Maximization ........................................ 9

3.1 Missing At Random ................................................................................................. 9
3.2 Expectation-Maximization Algorithm ...................................................................... 9
3.3 Expectation .............................................................................................................. 10
  3.3.1 Complete-Data Log-Likelihood And Its Expectation ........................................ 11
  3.3.2 Distribution Of $z_j | y, \theta$ And $z_{ij} | y, \theta$ .................................................... 11
  3.3.3 Details Of $E[\log(z_{ij}) | y, \hat{\theta}(j)]$ ................................................................. 13
3.4 Maximization .......................................................................................................... 14
  3.4.1 Maximization With Respect To $P$ .................................................................. 14
  3.4.2 Maximization With Respect To $\beta$ ................................................................. 14
3.5 Observed Information Via Direct Calculation ....................................................... 15
3.6 Apply EM To The *Rhyzopertha dominica* Data .................................................... 15

## 4 Balanced Firth Modified-Score Functions ............................................................... 19

4.1 Penalized Complete-Data Log-Likelihood ............................................................... 19
  4.1.1 Complete-Data Score Function ....................................................................... 19
  4.1.2 Actual Firth Modified-Score Function ............................................................ 20
  4.1.3 Balanced Firth Modified-Score Function ....................................................... 20
4.2 Maximizing A Balanced Firth Penalized Log-Likelihood ......................................... 22
  4.2.1 Partials Of $J$ .................................................................................................. 22
  4.2.2 Partial Of trace $\left( J^{-1} \frac{\partial J}{\partial \beta_s} \right)$ ...................................................... 23
  4.2.3 Observed Information Of A Firth-Like Penalized Log-Likelihood .................. 23
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Observed <em>Rhizopertha dominica</em> data. The first four rows describe the treatments. The last six rows describe observed response across levels of genotype locus crossings. Each of the entries in $y$ can be thought of as a binomial response where the binomial index($m_{ij}$) is unobserved. In further work, we will assume that experimental units are drawn at random from a infinite population with index $m_j$ for each treatment $j$ (columns).</td>
<td>4</td>
</tr>
<tr>
<td>3.1</td>
<td>Techniques of Chapter 3 applied to observed <em>Rhizopertha dominica</em> experimental data. We can see a complete separation of points by reviewing expectations and observations for genotype-locusA. This contributes to the overly-broad confidence intervals associated with estimates in Table 3.2.</td>
<td>17</td>
</tr>
<tr>
<td>3.2</td>
<td>Techniques of Chapter 3 used to create estimate $\hat{\theta}$ on <em>Rhizopertha dominica</em> data with standard errors. The first 5 estimates are population proportion ($\hat{P}$) and the remaining estimates are $\hat{\beta}$.</td>
<td>18</td>
</tr>
<tr>
<td>4.1</td>
<td>Techniques of Chapter 4 used to create estimate $\hat{\theta}$ on <em>Rhizopertha dominica</em> data with standard errors. The first 5 estimates are population proportion ($\hat{P}$) and the remaining estimates are $\hat{\beta}$.</td>
<td>28</td>
</tr>
<tr>
<td>4.2</td>
<td>Techniques of Chapter 4 applied to observed <em>Rhizopertha dominica</em> experimental data. A complete separation of points from Table 3.2 has been removed.</td>
<td>29</td>
</tr>
<tr>
<td>5.1</td>
<td>Observed indications of bias for our <em>Rhizopertha dominica</em> data simulation. The column on the right is the mean of the errors on the scale of a corresponding Wald interval with $\alpha = .05$ for each observation. The first 5 estimates are population proportion ($\hat{P}$), and the last are $\hat{\beta}$.</td>
<td>43</td>
</tr>
<tr>
<td>5.2</td>
<td>Observed coverage probabilities for Wald intervals with $\alpha = .05$. The first 5 estimates are population proportion ($\hat{P}$), and the last are $\hat{\beta}$.</td>
<td>47</td>
</tr>
<tr>
<td>6.1</td>
<td>An outline of our EM for Multinomial realization $j$. On the left is the observation with which we begin: the count of successes by cell and the count of failures grouped together as a marginal total. In the center, we have the prior probabilities of one observation being associated with each particular cell. Between the center and the right, we calculate the marginal probability associated with $z_{ij}$ as $P_{z_{ij}} = \sum_{i=1}^{n} P_i(1 - \pi_{ij})$. On the right, the expectations for the missing cell counts is the marginal count $z_{ij}$ multiplied by the proportion of the marginal probability that belongs to that cell. After taking that expectation, we fit new $\hat{P}$ and $\hat{\pi}$ on these new cell counts. We iterate creating new values for the center and right tables until convergence.</td>
<td>49</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Hindrance plots for a simple dose-response simulation. In these plots, we consider attributes intrinsic to the underlying true parameterization and experimental design that make accurate point estimates increasingly difficult to create. In (a) and (b), we have the expected proportion of our experimental units that will not have observed class. And in (c) and (d), we have the proportion of class 0 to class 1 that is observed in the survivors.</td>
<td>35</td>
</tr>
<tr>
<td>5.2</td>
<td>For Treatment Structure 1, trimmed-means of simulated point estimates for the class 1 effect.</td>
<td>38</td>
</tr>
<tr>
<td>5.3</td>
<td>For Treatment Structure 2, trimmed-means of simulated point estimates for the class 1 effect.</td>
<td>39</td>
</tr>
<tr>
<td>5.4</td>
<td>Trimmed-means of simulated standard errors for the class 1 effect. All plots are for simulated fittings using methods that have unobserved class covariates on dead beetles. Use of the balanced Firth ad hoc penalty can restrain standard errors.</td>
<td>40</td>
</tr>
<tr>
<td>5.5</td>
<td>Sampled estimates for $\hat{P}$ are standardized and then displayed overlaid in a normal-quantile plot.</td>
<td>45</td>
</tr>
<tr>
<td>5.6</td>
<td>Sampled estimates for $\hat{\beta}$ are standardized and then displayed overlaid in a normal-quantile plot.</td>
<td>45</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Suppose that we were to do an experiment with binomial outcomes on experimental units with intrinsic categorical covariates of interest, for example, dose-response experiments. But suppose we did not select experimental units to get specific treatments according to those categorical covariates of interest; rather, we allowed them to be selected from a larger population at random. And having selected the group at random from the larger population, we apply treatments to that entire selection. This would be an observational study. We are sure that everyone interested in this paper is well versed in their analysis.

But now suppose that the categorical covariates of interest for the failure observations were unobserved. We will show how to evaluate exactly that data, provided the logit link on the log-odds is an appropriate model.

We build on previous work by Dr. John R. Stevens of Utah State University and Dr. David Schilipalius of The University of Queensland which we summarize at the end of our chapters. Our original contribution is a generalizing of the algorithm to any binomial model with a logit link for which you can specify a model matrix, the construction of explicit expression for important partial derivatives, the weaving into the algorithm an ad hoc score-modification, the expression of a technique to fit that score-modification, and the implementation of general forms of these things in R.

We will finish in this Chapter by stating the motivating problem in detail; at the end of each Chapter, we will develop this example further and motivated the work of the next Chapter. In Chapter 2, we state the model precisely. We then describe in summary how we could maximize the observed likelihood via EM (Chapter 3) and create subsequent confidence intervals on our estimates. We learn in Chapters 3 and 5 that such a direct maximization can interact poorly with binomial logit models (likely due to quasi or complete
separation of points in some abstraction). So we propose score-modifications that might
direct the fit towards estimates in not so flat regions of the observed likelihood (Chapter
4). Finally, we begin to put these techniques through their paces in Chapter 5.

1.1 Phosphine Fumigant Dose-Response For Rhyzopertha dominica

In 2006, Dr. Stevens of Utah State University collaborated with Dr. Schilipalius of
The University of Queensland to analyze the dose-response of Rhyzopertha dominica (lesser
grain borer) after 48-hour exposures to various levels of phosphine fumigant. Their work
appears in a paper titled Dose-Response Modeling With Marginal Information On A Missing
Categorical Covariate [14]. Of interest was the relationship of dose-response to a crossing
of a specific genotype(+/-) and locus(A,H,B).

In the experiment, beetles were selected at random for each treatment from a larger
general population. Prior to treatments the genotype and locus associated with a beetle
is not known to the experimenters. 10798 beetles were exposed to poison, only 378 bee­
tles survived; a decision was made to only find the genotype and locus for the surviving
beetles. We note that the key characteristic of this data is that the binomial index for
each genotype-locus-dose crossing was not observed; all that we know are the genotype­
locus-dose covariates for the surviving beetles. What might we say about the dose-response
relationship of this beetle as it varies by genotype and locus given only the results in Table
1.1?

The techniques of this report are created to answer questions of this type.

A standard model from previous work is:

$$\log \frac{\pi_{ij}}{1 - \pi_{ij}} = \mu + C_i + D_j + (CD)_{ij}$$  \hspace{1cm} (1.1)$$

$\mu$ is a grand-mean. Class (C) has 6 levels indexed by $i$ representing the crossing of two levels
of genotype(+/-) and three levels of locus(A,H,B), so $i \in \{1..6\}$. Dose (D) was modeled as
a continuous covariate, but only 11 levels exist for it; $j \in \{1..11\}$. Organizing the model in
this way makes the levels of C understood as rows of $y$ in Table 1.1, and the levels of D are
understood as columns of $y$ in Table 1.1.

This model is impossible to fit directly because the cell proportions are not observed. We are interesting in creating estimates of this model’s covariates and their standard errors using additional assumptions.
<table>
<thead>
<tr>
<th>Phosphine Dosage (mg/L)</th>
<th>0.000</th>
<th>0.003</th>
<th>0.004</th>
<th>0.005</th>
<th>0.010</th>
<th>0.050</th>
<th>0.100</th>
<th>0.200</th>
<th>0.300</th>
<th>0.400</th>
<th>1.000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Receiving Dosage ($m_j$)</td>
<td>98</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>300</td>
<td>400</td>
<td>750</td>
<td>500</td>
<td>500</td>
<td>7850</td>
</tr>
<tr>
<td>Total Deaths ($z_j$)</td>
<td>0</td>
<td>16</td>
<td>68</td>
<td>78</td>
<td>77</td>
<td>270</td>
<td>383</td>
<td>740</td>
<td>490</td>
<td>492</td>
<td>7806</td>
</tr>
<tr>
<td>Total Survivors</td>
<td>98</td>
<td>84</td>
<td>32</td>
<td>22</td>
<td>23</td>
<td>30</td>
<td>17</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>44</td>
</tr>
</tbody>
</table>

Table 1.1: Observed *Rhizopertha dominica* data. The first four rows describe the treatments. The last six rows describe observed response across levels of genotype locus crossings. Each of the entries in $y$ can be thought of as a binomial response where the binomial index($m_{ij}$) is unobserved. In further work, we will assume that experimental units are drawn at random from a infinite population with index $m_j$ for each treatment $j$ (columns).
Chapter 2

Model

We will set up a Multinomial-Binomial hierarchical model for an experiment. The experimental units will be selected with the Multinomial model. After the experiment is done each unit will be associated with success or failure. We will count successes or failures in a Binomial random variable, but implicitly inside of the Binomial is one Bernoulli random variable for each experimental unit. When that Bernoulli outcome is a failure some categorical covariates possessed by the associated experimental unit will be unobserved. When the Bernoulli outcome is success all covariates for that experimental unit will be observed.

2.1 Multinomial Binomial Hierarchical Model

Presented here is essentially the model of an observational study, though we apply this model to experiments.

2.1.1 Multinomial Model

Assume our population of interest can be partitioned into $c$ classes with population proportions $\mathbf{P}$, a vector of $P_1, \ldots, P_c$ scalars; such that, $\forall i \in 1 \ldots c$, $0 < P_i < 1$, and $1 - \sum_{i=1}^{c-1} P_i = P_c$. Further, assume that population is infinite in size, so that each draw from this population is independent.

Each member of the population can only belong to one of these partitions. And we will assign membership according to the value(s) of categorical covariates belonging to individuals in that population. But we allow for models in which some categorical covariates possess missing observations, the ones associated with this multinomial realization, while other categorical covariates are fully observed (as in an observational study). To make this
distinction clear, we refer to unique combinations of categorical covariates associated with one multinomial partition as a “class”. Each member of the population is one of \( c \) classes.

We are interested in \( t \) unique covariate profiles (“treatments”); for each of which, we assume that an independent multinomial random draw occurred from a larger population.

Let \( \mathbf{M}_1, \ldots, \mathbf{M}_t \) be vector random variables with a multinomial distribution with indexes \( m_1, \ldots, m_t \) and realizations \( \mathbf{m}_1, \ldots, \mathbf{m}_t \) drawn from a population with proportions as above. So \( \mathbf{M}_j \sim \text{Multinomial}(m_j, \mathbf{P}) \), and \( m_{ij} \) represents the scalar count of class \( i \) that receive treatment \( j \), which is also entry \( i \) of vector realization \( \mathbf{m}_j \). The order of indexes is chosen so that we can view the treatment multinomial realizations as the columns of a matrix. The rows then align to class realizations. We work with \( \mathbf{m} = (\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_t) \), where \( \mathbf{m}_j \) are column vectors. And \( m_{ij} \) is the \( i \)th \( j \)th entry of that matrix.

### 2.1.2 Binomial Model

Now then let \( Y_{11}, \ldots, Y_{ct} \) be independent random variables with Binomial distributions having indexes \( m_{11}, \ldots, m_{ct} \), parameters \( \pi_{11}, \ldots, \pi_{ct} \), and realizations \( y_{11}, \ldots, y_{ct} \). Let \( \mathbf{\pi} \) and \( \mathbf{y} \) be matrices of the same quantities with organization to match \( \mathbf{m} \).

Let \( \mathbf{z} \) be a matrix such that \( \mathbf{m} = \mathbf{z} + \mathbf{y} \), so that \( \mathbf{z} \) is a matrix of failure counts. Then \( \mathbf{z} \) forms the missing data in our application, while \( \mathbf{y} \) is the observed data.

Though these binomial random variables are presented as 2-dimensional matrices as in the previous section, it is necessary to work with them in vector form during algebra related to binomial models. In this document, ‘arrow’ will signify the variables or realizations of interest in a vector form. Let \( n = c \times t \). Let \( \mathbf{\tilde{m}} = (\mathbf{m}_1^T, \ldots, \mathbf{m}_t^T) \) be a vector of the indexes for \( Y_{11}, \ldots, Y_{ct} \). Other quantities and variables will be put into a corresponding vector with organization to match \( \mathbf{\tilde{m}} \) as necessary.

In general, we take a cue from the number of subscripts on these variables and the context for determining whether the subscripts are referring to matrix or vector forms.

### GLM With A Logit Link

Much of the work for the Binomial model will be that of a GLM with a logit link
function. Henceforth in this document, we will assume that we have a model with matrix \( X \) of order \( n \times p \) such that:

\[
\eta_i = \log \left( \frac{\pi_i}{1 - \pi_i} \right) = \sum_{r=1}^{p} x_{ir} \beta_r
\]

(2.1)

where \( \pi_i \) is the \( i \)th element of \( \vec{\pi} \), where \( \beta_r \) is the \( r \)th element of \( \vec{\beta} \), a length \( p \) vector of unknown parameters, and where \( x_{ir} \) refers to the corresponding entry in the model matrix.

We will make use of a well known result later in the paper [10], namely

\[
\frac{\partial \pi_i}{\partial \beta_r} = \pi_i(1 - \pi_i)x_{ir}
\]

(2.2)

2.2 Missing Categorical Covariates On The Failure Observations

The chief difficulty in this work is that the class of the observations associated with failure \( (z) \) will be unobserved. The binomial index \( (m) \) matrix is then unobserved. To make matters worse the multinomial population proportions for class \( (P) \) are unknown. We must estimate these three quantities in some fashion then: \( \hat{P} \), \( \hat{\beta} \), and \( \hat{z} \).

2.3 Returning To *Rhyzopertha dominica*

Building on Section 1.1, most of this work in Chapter 2 has a straight-forward application to the *Rhyzopertha dominica* dose-response motivating example.

We see the number of classes \( (c) \) as the number of rows in \( y \) of Table 1.1. We see the number of treatments \( (t) \) as the number of columns in Table 1.1.

Also in Table 1.1, the row, *Total Receiving Dosage*, describes the multinomial indexes \( m_1, \ldots, m_t \) of Section 2.1.1. The original description of the experiment [14] validates the assumption that, for each treatment level, we might view the experimental units as drawn from a population with constant proportions across a number of classes \( (c) \).
The logit model of Equation 2.1 is consistent with how we described the original model (Equation 1.1). \( y \) is entirely observed and presented in Table 1.1 as it needs to be. As said before, left to do is describe how we might estimate \( \hat{P}, \hat{\beta}, \) and \( \hat{z}. \)
Chapter 3

Parameter Estimation with Expectation-Maximization

In this chapter, we will show how to use EM [4] to create parameter estimates for our model. There is no modification to the score function in this chapter. We are working with a straight maximum-likelihood problem. The subscript "non" is used occasionally to make clear that these functions are different than penalized versions which appear later in this report.

3.1 Missing At Random

A prerequisite to using EM is the property that class is "missing at random" or MAR. We use the following definition of the property [8];

If covariate x is MAR, this means that the probability of observing x (conditional on y and the other observed covariates) does not depend on x or any other unobserved covariate, but may depend on y and the other observed covariates.

We consider here the Bernoulli trials on experimental units underlying the Binomial experiment. The probability of observing class (x) once conditioned on the Bernoulli outcome(y) is either 1 or 0: 1 if the outcome was success and 0 if the outcome was failure. This implies the probability of observing class (x) conditioned on the Bernoulli outcome (y) does not depend on the realization of class (x) nor any other unobserved covariate, and MAR is satisfied.

3.2 Expectation-Maximization Algorithm

We follow the notation of Hastie [6]. This, of course, is a well established algorithm in the literature [4]. We only restate it here to introduce notation.
Let \( \theta = (P, \beta) \) be the parameters we are interested in estimating. Recall that \( y \) is our observed data, while \( z \) is our unobserved data. \( l_{\text{non}}(\theta' | y, z) \) is our complete-data log-likelihood (Equation 3.2). Following Hastie:

1. Initialize parameters \( \hat{\theta}^{(0)} \).

2. **Expectation Step:** at the jth step, compute

   \[
   Q_{\text{non}}(\theta', \hat{\theta}^{(j)}) = E[l_{\text{non}}(\theta' | y, z) | y, \hat{\theta}^{(j)}]
   \]  

   (3.1)

3. **Maximization Step:** determine the new estimate \( \hat{\theta}^{(j+1)} \) as the maximizer of \( Q_{\text{non}}(\theta', \hat{\theta}^{(j)}) \) over \( \theta' \).

   It is convenient to do this in two stages, creating maximizing estimates \( \hat{P}^{(j+1)} \) and \( \hat{\beta}^{(j+1)} \) one at a time and then declaring \( \hat{\theta}^{(j+1)} = (\hat{P}^{(j+1)}, \hat{\beta}^{(j+1)}) \).

4. Iterate steps 2 and 3 until convergence.

We now expound on the expectation and maximization calculations.

### 3.3 Expectation

In this section, we explore all the necessary details of the Expectation step.
3.3.1 Complete-Date Log-Likelihood And Its Expectation

The complete-data log-likelihood is given by:

\[
\ln \left( l_{\text{non}}(\theta|y, z) \right) = \sum_{i=1}^{c} \sum_{j=1}^{t} \log \left( \left( \frac{m_{ij}}{y_{ij}} \right)^{y_{ij}} \right) \left( 1 - \pi_{ij} \right)^{m_{ij} - y_{ij}} \\
+ \sum_{j=1}^{t} \log \left( \prod_{i=1}^{c} \prod_{i=1}^{t} m_{ij} \right)^{m_{ij}} \\
= \sum_{i=1}^{c} \sum_{j=1}^{t} \left( \pi_{ij} \log(1 - \pi_{ij}) + \log(\pi_{ij} + \pi_{ij} \log(1 - \pi_{ij})) \right) \\
+ \sum_{j=1}^{t} \log(m_{.j}) + \sum_{i=1}^{c} \sum_{j=1}^{t} (m_{ij} \log(P_i) - \log(m_{ij}!)) \\
= \sum_{i=1}^{c} \sum_{j=1}^{t} \left( \pi_{ij} \log(P_i) + \log(1 - \pi_{ij} P_i) \right) \\
- \sum_{i=1}^{c} \sum_{j=1}^{t} \log(y_{ij}) + \log(z_{ij}!)) \\
+ \sum_{j=1}^{t} \log(m_{.j}!) \\
\]

Many of these factors are implicitly functions of \( \theta, y, \) and \( z. \)

Now

\[
E[l_{\text{non}}(\theta'|y, z)|y, \hat{\theta}(j)] = \sum_{i=1}^{c} \sum_{j=1}^{t} \left( \pi_{ij} \log(P_i) + E[z_{ij}|y, \hat{\theta}(j)] \log((1 - \pi_{ij}) P_i) \right) \\
- \sum_{i=1}^{c} \sum_{j=1}^{t} \left( \log(y_{ij}) + E[\log(z_{ij})|y, \hat{\theta}(j)] \right) \\
+ \sum_{j=1}^{t} \log(m_{.j}!) \\
\]

So we have to calculate two expectations in particular detail.

3.3.2 Distribution Of \( z_{ij}|y, \theta \) And \( z_{ij}|y, \theta \)

It is conceptually simpler to talk about the conditional distributions of the unobserved quantities before proceeding to the expectations.
We will work in columns of matrix \( m, y, \) etc. The column in particular will be \( m_j, y_j, \) etc. Since the multinomial column realizations and subsequent binomial realizations were independent, we can discard the other columns when finding a conditional distribution for \( z_{ij}. \)

For the rest of this section, we suppress conditioning on \( \theta, \) as these parameters do not have distributions, they are considered known for these calculations, and appending their notation is unwieldy. We also add \( z_{.j} \) to our notation as the total number of failures associated with multinomial draw with index \( m_j. \) We define \( y'_{.j} = (y_j, z_j) \) as a length \( c+1 \) vector with the count of failures appended to the end of the successes by class. \( y'_{.j} \) represents everything fully observed from the multinomial draw \( j. \) \( y'_{.j} \) is implicitly given when \( y_j \) is given because \( m_j \) is assumed known.

We begin by trying to find the distribution of \( m_j | y'_{.j} \) with

\[
f(m_j | y'_{.j}) = \frac{f(y'_{.j} | m_j) f(m_j)}{f(y'_{.j})}
\]

We can find the distributions of these random variables by inspection. \( f(y'_{.j} | m_j) = \prod_{i=1}^c f_{ij}(y_{ij} | m_{ij}), \) because after we know the binomial index these were independent outcomes. \( m_j \sim \text{Multinomial}(m_j, \mathbf{P}) \) by our model. \( y'_{.j} \sim \text{Multinomial}(m_j, \mathbf{P}'_{.j}) \) where \( \mathbf{P}'_{.j} \) is a \( c+1 \) length vector given as \( \mathbf{P}'_{.j} = (P_{1\pi_{1j}}, \ldots, P_{c\pi_{cj}}, 1 - \sum_{i=1}^c P_i \pi_{ij}). \)

We have:

\[
f(m_j | y'_{.j}) = \frac{f(y'_{.j} | m_j) f(m_j)}{f(y'_{.j})} = \frac{\prod_{i=1}^c m_{ij}!}{y_{ij}! (m_{ij} - y_{ij})!} (1 - \pi_{ij})^{m_{ij} - y_{ij}} \pi_{ij}^{y_{ij}}} \prod_{i=1}^c m_{ij}! \prod_{i=1}^c P_i^{m_{ij}} \]

\[
= \frac{m_j!}{z_j! \prod_{i=1}^c y_{ij}!} (1 - \sum_{i=1}^c P_i \pi_{ij})^{z_j} \prod_{i=1}^c (P_i \pi_{ij})^{y_{ij}}
\]

\[
= \frac{z_j!}{\prod_{i=1}^c z_{ij}!} \prod_{i=1}^c \left[ \frac{P_i (1 - \pi_{ij})}{\sum_{i=1}^c P_i (1 - \pi_{ij})} \right]^{z_{ij}}
\]

From this we can infer \( z_{.j} | y'_{.j} \sim \text{Multinomial}(z_{.j}, \lambda_{.j}), \) where \( \lambda_{.j} \) is a length \( c \) vector with ith
entry \( \lambda_{ij} = \frac{P_i(1 - \pi_{ij})}{\sum_{c=1}^{C} P_i(1 - \pi_{ij})} \). We then conclude \( z_{ij} | y_j' \sim \text{Binomial} \left( z_{ij}, \frac{P_i(1 - \pi_{ij})}{\sum_{c=1}^{C} P_i(1 - \pi_{ij})} \right) \). This gives us conditional expectation as

\[
E[z_{ij} | y, \hat{\theta}^{(j)}] = z_{ij} \frac{\hat{P}_i(1 - \hat{\pi}_{ij})}{\sum_{k=1}^{K_i} \hat{P}_k(1 - \hat{\pi}_{kj})}
\]  

(3.6)

We introduce \( \tilde{z}_{ij} \) as referring to the above expectation. And the conditional variance as

\[
\text{Var}[z_{ij} | y, \hat{\theta}^{(j)}] = z_{ij} \frac{\hat{P}_i(1 - \hat{\pi}_{ij})}{\sum_{k=1}^{K_i} \hat{P}_k(1 - \hat{\pi}_{kj})} \left( 1 - \frac{\hat{P}_i(1 - \hat{\pi}_{ij})}{\sum_{k=1}^{K_i} \hat{P}_k(1 - \hat{\pi}_{kj})} \right)
\]  

(3.7)

3.3.3 Details Of \( E[\log(z_{ij}) | y, \theta] \)

The work in this section is entirely previous work of Stevens and Schlipalius [14]; though, we introduce the distributional notation.

Binet’s Approximation [12] gives:

\[
\log(z_{ij}) \approx g(z_{ij}) = (z_{ij} + 0.5)\log(z_{ij} + 1) - (z_{ij} + 1) + 0.5\log2\pi
\]  

(3.8)

We employ a second-order Taylor series approximation taken about \( \tilde{z}_{ij} \) of \( g(z_{ij}) \). It is easy to see that \( g(z) < (z + 1)^2 \) for all \( z \in \mathbb{Z}^+ \). Also

\[
g''(z_{ij}) = \frac{z_{ij} + 1.5}{(z_{ij} + 1)^2}
\]  

(3.9)

Now resolving the conditional expectations taking care to note that \( z_{ij} \) is a random variable with a known conditional distribution (binomial), we have

\[
E[\log(z_{ij}) | y, \theta] \approx E[g(\tilde{z}_{ij}) + g'(\tilde{z}_{ij})(z_{ij} - \tilde{z}_{ij}) + 0.5g''(\tilde{z}_{ij})(z_{ij} - \tilde{z}_{ij})^2 | y, \theta] \\
= g(\tilde{z}_{ij}) + 0.5g''(\tilde{z}_{ij})\text{Var}[z_{ij} | y, \theta]
\]  

(3.10)

This is simply the Delta Method [2]. We have no further simplifications, so we conclude by noting all these terms in Equation 3.10 now are completely specified for calculation via Equations 3.6, 3.7, 3.8 and 3.9.
3.4 Maximization

In this section, we explore all the necessary details of the Maximization step. We maximize in two steps. First, we maximize with respect to $P$ and then we maximize with respect to $\beta$.

3.4.1 Maximization With Respect To $P$

We introduce notation $\tilde{m}_i = \sum_{j=1}^{t}(y_{ij} + \tilde{z}_{ij})$.

The terms of Equation 3.3 that need be considered while maximizing with respect to $P$ are:

$$Q_P = \sum_{i=1}^{c} \sum_{j=1}^{t} (y_{ij} + \tilde{z}_{ij}) \log P_i = \sum_{i=1}^{c} \tilde{m}_i \log P_i$$

(3.11)

Using the method of Lagrange multipliers to maximize $Q_P$ subject to constraint $\sum_{i=1}^{c} P_i = 1$ results in MLEs of $\hat{P}_i = \frac{\tilde{m}_i}{m_\cdot}$, where $m_\cdot$ is the sum of all the multinomial indexes (a relatively intuitive estimate).

3.4.2 Maximization With Respect To $\beta$

The terms of Equation 3.3 that need be considered while maximizing with respect to $\beta$ are:

$$Q_\beta = \sum_{i=1}^{c} \sum_{j=1}^{t} \left( y_{ij} \log(\pi_{ij}) + E[\tilde{z}_{ij} | y_i, \hat{\theta}^{(j)}] \log(1 - \pi_{ij}) \right)$$

$$= \sum_{i=1}^{c} \sum_{j=1}^{t} (y_{ij} \log(\pi_{ij}) + (\tilde{m}_{ij} - y_{ij}) \log(1 - \pi_{ij}))$$

(3.12)

This is the same maximization problem as when treating the $\tilde{m}_{ij}$ as observed, so we proceed by maximizing the binomial likelihood via standard methods as if everything had been observed.
3.5 Observed Information Via Direct Calculation

After EM has converged to the observed maximum likelihood solution $\hat{\theta}$, we will want to calculate the covariance matrix of $\hat{\theta}$ via inverting the observed information.

For this matrix calculation, we omit consideration of the constrained $P_c$, and when we speak of vector $P$ inside of $\theta = (P, \beta)$ it is a length $c - 1$ vector. Wherever $P_c$ appears, we consider it a function equal to $1 - \sum_{i=1}^{c-1} P_i$.

We use the methods of Oakes for calculating the observed information matrix [11]:

$$\frac{\partial^2 l_{\text{non}}(\hat{\theta})}{\partial \theta \partial \theta} = \left[ \frac{\partial^2 Q_{\text{non}}(\theta, \hat{\theta})}{\partial \theta \partial \theta} + \frac{\partial^2 Q_{\text{non}}(\theta, \hat{\theta})}{\partial \theta \partial \theta} \right]_{\theta = \hat{\theta}} \quad (3.13)$$

This is two matrices summed together to create a third. In the first term on the right, we have a Hessian, during the calculation of which, we treat the conditional expected values as observed. In the second term on the right, we have a Hessian in which the hat factors are injected by the conditional expected values. For this second Hessian, terms without conditional expected values as functions of those particular hat variables need not be considered.

Construction of these matrices is summarized in Appendix A, and the covariance matrix from the EM algorithm is taken as the inverse of the negative of Equation 3.13.

3.6 Apply EM To The *Rhyzopertha dominica* Data

Returning to the motivating example outlined in Sections 1.1 and 2.3, we applied the techniques of this Chapter to the data in Table 1.1. Our model matrix was a bit more expressive than Equation 1.1 would suggest. We create $X$ to be consistent with effects parameterization

$$\log \frac{\pi_{ijk}}{1 - \pi_{ijk}} = \mu + D_i + G_j + L_k + (DG)_{ij} + (DL)_{ik} + (GL)_{jk} + (DGL)_{ijk} \quad (3.14)$$

Dose (D) was modeled as a continuous covariate. But only 11 levels exist for it, so $i \in$
{1..11}. Genotype(G) has two levels (+/−). L has three levels (A,B,H). The interaction effects have the usual interpretations.

The original effects model was designed to be consistent with the presentation of y (rows and columns indexed by i and j in both the effects model and observed data matrix). The mapping of this effects model (Equation 3.14) to rows of y is through the crossing of levels of $G_j$ and $L_k$, but our having to keep track of these details is now abstracted away: there is no issue here when we construct model matrix $X$ to have row order consistent with that of the y in Table 1.1.

A presentation of the cell probabilities and expectations created using the techniques of this chapter is in Table 3.1. We have provided both the expected $\bar{m}$ and the observed y (repeated from Table 1.1). If we direct our attention to the expectations and observations for genotype-locusA, we find a complete separation of points. For phosphine dosages less than or equal to 0.010 (mg/L), genotype-locusA is predicted to always survive, and for phosphine dosages greater than or equal to 0.050 (mg/L), genotype-locusA is predicted to always succumb to the poison.

The associated parameter estimates and standard errors are in Table 3.2. Before we interpret the estimates, we should direct our attention to the standard errors which are overly-broad. The population proportions ($\hat{P}$) are estimated sensibly, but for the binomial model parameters ($\hat{\beta}$), we see standard errors we might associate with a separation of points in a fully-observed logistic model.

In Chapter 4, we consider how we might apply a penalty to the log-likelihood to create estimates with some confidence.
<table>
<thead>
<tr>
<th>Phosphine Dosage (mg/L)</th>
<th>0.000</th>
<th>0.003</th>
<th>0.004</th>
<th>0.005</th>
<th>0.010</th>
<th>0.050</th>
<th>0.100</th>
<th>0.200</th>
<th>0.300</th>
<th>0.400</th>
<th>1.000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Receiving Dosage ($m_j$)</td>
<td>98</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>300</td>
<td>400</td>
<td>750</td>
<td>500</td>
<td>500</td>
<td>7850</td>
</tr>
</tbody>
</table>

| genotype-locusA | 10.00 | 10.00 | 3.00 | 7.00 | 9.00 | 24.27 | 31.73 | 59.95 | 39.64 | 39.76 | 624.00 |
| genotype+locusA | 4.00 | 4.08 | 4.15 | 2.13 | 0.13 | 5.43 | 7.67 | 11.78 | 11.64 | 10.24 | 163.34 |
| genotype-locusB | 31.00 | 23.69 | 44.62 | 40.74 | 37.79 | 110.65 | 144.68 | 59.95 | 39.64 | 39.76 | 624.00 |
| genotype+locusB | 6.00 | 6.00 | 4.00 | 2.00 | 8.00 | 16.18 | 21.15 | 39.96 | 26.43 | 26.51 | 416.00 |
| genotype-locusH | 27.00 | 35.80 | 23.69 | 44.62 | 40.74 | 37.79 | 110.65 | 144.68 | 59.95 | 39.64 | 39.76 | 624.00 |
| genotype+locusH | 20.00 | 20.43 | 7.83 | 6.79 | 5.90 | 31.64 | 48.54 | 88.70 | 58.80 | 58.97 | 925.57 |

| genotype-locusA | 10 | 10 | 3 | 7 | 9 | 0 | 0 | 0 | 0 | 0 | 0 |
| genotype+locusA | 4 | 4 | 4 | 2 | 0 | 5 | 7 | 10 | 10 | 8 | 44 |
| genotype-locusB | 31 | 18 | 10 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| genotype+locusB | 6 | 6 | 4 | 2 | 8 | 5 | 0 | 0 | 0 | 0 | 0 |
| genotype-locusH | 27 | 26 | 4 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| genotype+locusH | 20 | 20 | 7 | 6 | 5 | 20 | 10 | 0 | 0 | 0 | 0 |

| genotype-locusA | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| genotype+locusA | 0.944 | 0.943 | 0.943 | 0.943 | 0.943 | 0.933 | 0.920 | 0.886 | 0.842 | 0.784 | 0.269 |
| genotype-locusB | 1.000 | 0.762 | 0.223 | 0.025 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| genotype+locusB | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.309 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| genotype-locusH | 0.987 | 0.594 | 0.280 | 0.094 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| genotype+locusH | 0.952 | 0.945 | 0.943 | 0.940 | 0.927 | 0.676 | 0.181 | 0.002 | 0.000 | 0.000 | 0.000 |

Table 3.1: Techniques of Chapter 3 applied to observed *Rhizopertha dominica* experimental data. We can see a complete separation of points by reviewing expectations and observations for genotype-locusA. This contributes to the overly-broad confidence intervals associated with estimates in Table 3.2.
<table>
<thead>
<tr>
<th>No Score-Modification</th>
<th>Estimate</th>
<th>Std. Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>genotype-locusA</td>
<td>0.080</td>
<td>0.012</td>
</tr>
<tr>
<td>genotype+locusA</td>
<td>0.021</td>
<td>0.007</td>
</tr>
<tr>
<td>genotype-locusB</td>
<td>0.363</td>
<td>0.045</td>
</tr>
<tr>
<td>genotype+locusB</td>
<td>0.053</td>
<td>0.010</td>
</tr>
<tr>
<td>genotype-locusH</td>
<td>0.366</td>
<td>0.045</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>28.432</td>
<td>11541.838</td>
</tr>
<tr>
<td>genotype+locusB</td>
<td>-25.614</td>
<td>11541.839</td>
</tr>
<tr>
<td>locusB</td>
<td>-20.033</td>
<td>11541.835</td>
</tr>
<tr>
<td>locusH</td>
<td>-24.075</td>
<td>11541.839</td>
</tr>
<tr>
<td>dose</td>
<td>-981.773</td>
<td>235113.878</td>
</tr>
<tr>
<td>genotype+:locusB</td>
<td>41.996</td>
<td>17183.778</td>
</tr>
<tr>
<td>genotype+:locusH</td>
<td>24.241</td>
<td>11541.839</td>
</tr>
<tr>
<td>genotype+:dose</td>
<td>977.956</td>
<td>235113.878</td>
</tr>
<tr>
<td>locusB:dose</td>
<td>-1430.217</td>
<td>235113.779</td>
</tr>
<tr>
<td>locusH:dose</td>
<td>-343.583</td>
<td>235114.078</td>
</tr>
<tr>
<td>genotype+:locusB:dose</td>
<td>922.349</td>
<td>322311.443</td>
</tr>
<tr>
<td>genotype+:locusH:dose</td>
<td>302.475</td>
<td>235114.078</td>
</tr>
</tbody>
</table>

Table 3.2: Techniques of Chapter 3 used to create estimate $\hat{\theta}$ on *Rhizopertha dominica* data with standard errors. The first 5 estimates are population proportion ($\hat{P}$) and the remaining estimates are $\hat{\beta}$. 
Chapter 4
Balanced Firth Modified-Score Functions

In this chapter, we will present an ad hoc modification to a Binomial score function. We will show how this modification can be maximized (not necessarily the fastest known algorithm), and then we will show how to augment our EM procedure of Chapter 3 to use this new penalty. In a later simulation, we will discover this balanced Firth modification appears to perform better than the unmodified score functions in EM.

The MLE, \( \hat{\pi} \), defined as the maximization of the complete-data log-likelihood. Equation 3.2, with respect to \( \pi \) is biased. In 1993, Firth published a bias-reducing score-modification for GLMs [5]. In 2002, Heinze recast this same score modification as a guarantee of finite variance on estimates [7], which is our primary interest. We would desire to place Firth’s modification into our EM procedure; however, a necessary expectation has proven to be intractable.

Ultimately, our ad hoc modification is to simply drop the binomial indexes from Firth’s bias-reducing score modification.

4.1 Penalized Complete-Data Log-Likelihood

4.1.1 Complete-Data Score Function

Define \( U(\beta_r) \) as the score function for the complete-data log-likelihood. Equation 3.2, with respect to \( \beta_r \). Constructing this result with Equation 2.2,

\[
U(\beta_r) = \frac{\partial}{\partial \beta_r} \ln \text{non}(\theta | y, z)
\]

\[
= \sum_{i=1}^{c} \sum_{j=1}^{t} (y_{ij} x_{(ij)} r - y_{ij} \bar{\pi}_{ij} x_{(ij)} r - \varepsilon_{ij} \pi_{ij} x_{(ij)} r)
\]

(4.1)
20

where \( x_{(ij)r} \) describes the \( r \)th entry of a model matrix row corresponding to the \( y_{ij} \) and \( z_{ij} \) observations.

### 4.1.2 Actual Firth Modified-Score Function

The actual Firth modified-score function is:

\[
U_A(\beta_r) = U(\beta_r) + \frac{1}{2} \text{trace} \left( J^{-1}_A \frac{\partial J_A}{\partial \beta_r} \right)
\]

(4.2)

Where \( J_A = X^T W_A X \), with \( X \) equal to the model matrix, \( W_A = \text{diag}\{m_i \pi_i (1 - \pi_i)\} \), and subscript \( i \) refers to the \( i \)th entry of \( \vec{m} \) and \( \vec{\pi} \).

There is a high-quality body of evidence supporting the use of \( U_A(\beta_r) \) [5] [7]. The corresponding penalized complete-data log-likelihood is the non-penalized complete-data log-likelihood, Equation 3.2, plus another term:

\[
l^A_{\text{pen}}(\theta|y, z) = l_{\text{non}}(\theta|y, z) + \frac{1}{2} \log |J_A|
\]

(4.3)

In EM, we would be required to maximize \( Q^A_{\text{pen}}(\theta', \hat{\theta}^{(j)}) = E[l^A_{\text{pen}}(\theta'|y, z)|y, \hat{\theta}^{(j)}] \). \( J_A \) is a function of the missing data \( (z) \) through \( m = y + z \), and inside of \( Q^A_{\text{pen}}(\theta', \hat{\theta}^{(j)}) \), we encounter the intractable \( E[ \log \det(J_A)|y, \hat{\theta}^{(j)}] \).

We looked for another way to penalize these likelihood functions; so that the maximization over the complete-data still tends towards estimates corresponding to Hessian determinants away from zero, while also being amenable to the maximization step for the incomplete-data in EM.

### 4.1.3 Balanced Firth Modified-Score Function

Suppose that the actual Firth modified-score function (Equation 4.2) was that for a balanced Binomial experiment. We would have \( \forall m_i \in \vec{m}, m_i = m \) where \( m \) is some integer 1 or greater. In this case, \( W_A = m W_B \) with \( W_B = \text{diag}\{\pi_i (1 - \pi_i)\} \), and the actual Firth
modified-score function is

\[
U_A(\beta_r) = U(\beta_r) + \frac{1}{2} \text{trace} \left( (X^T W_A X)^{-1} \frac{\partial}{\partial \beta_r} (X^T W_A X) \right) = U(\beta_r) + \frac{1}{2} \text{trace} \left( (X^T (mW_B) X)^{-1} X^T \frac{\partial (mW_B)}{\partial \beta_r} X \right) = U(\beta_r) + \frac{1}{2} \text{trace} \left( (X^T W_B X)^{-1} X^T \frac{\partial W_B}{\partial \beta_r} X \right)
\] (4.4)

From this, we see that for a balanced Binomial experiment the actual binomial index (m) has no part to play in the modification of the score function. Specifically, the binomial indexes could be 1 or 1000 and the modification to the score is the same. The capacity of \( \tilde{m} \) to decrease or increase the score-modification is through the relative differences in elements of \( m \).

Suppose then the actual Binomial indexes were not balanced. What would happen if we continued to use the balanced Firth modified-score function? To be clear, term \( U(\beta_r) \) changes appropriately as the Binomial indexes change; it is the score-modification term that is restrained. We don’t hope for bias-reduction in general, but rather the new score function might be well behaved in terms of bias-inducing variance-reducing penalty functions.

We will see in Section 4.3 that there is no longer difficulty in maximizing the expectation of this penalty with respect to the observed data: this new penalty is not a function of the missing data.

We define a balanced Firth modified-score function as

\[
U^B_{\text{pen}}(\beta_r) = U(\beta_r) + \frac{1}{2} \text{trace} \left( J_B^{-1} \frac{\partial J_B}{\partial \beta_r} \right)
\] (4.5)

Where \( J_B = X^T W_B X \). We define a corresponding balanced Firth penalized log-likelihood as

\[
l^B_{\text{pen}}(\theta|y, z) = l_{\text{non}}(\theta|y, z) + \frac{1}{2} \log |J_B|
\] (4.6)
4.2 Maximizing A Balanced Firth Penalized Log-Likelihood

In this section, we turn for the moment away from the hierarchical model and the problem of unobserved data. We are just speaking of a complete-data logit link Binomial model. We will show how to maximize a corresponding balanced Firth penalized log-likelihood.

We preferred to do this work in a great deal of generality to maximize the range of penalties we could consider in exploratory work. We defined $J$ to be any positive definite matrix of the form $X^T W X$ such that $W = \text{diag}(m_i' \pi_i (1 - \pi_i))$, and where $\mathbf{m}'$ is any arbitrary vector of constants. We relax everything about the penalty, but its basic form, invertability of $J$, and dependence on $\mathbf{m}$ (in truth, our original solution even relaxed $X$ to not necessarily be the model matrix). We also introduce $\alpha$ as an arbitrary constant.

We define this more general form of Firth-like modified-score function as

$$U_{pen}(\beta) = U(\beta) + \alpha \text{trace} \left( J^{-1} \frac{\partial J}{\partial \beta} \right)$$

(4.7)

With a corresponding general form of Firth-like penalized log-likelihood as

$$l_{pen}(\theta | y, z) = l_{non}(\theta | y, z) + \alpha \log|J|$$

(4.8)

We can recover Equations 4.5 and 4.6 with $\alpha = \frac{1}{2}$ and $\mathbf{m}' = 1$.

4.2.1 Partials Of $J$

Utilizing Equation 2.2, we derive

$$\frac{\partial J}{\partial \beta_r} = X^T \text{diag}(m_i' \pi_i (1 - \pi_i)) X$$

(4.9)

$$= X^T \text{diag}(m_i' x_{ir} (2\pi_i^3 - 3\pi_i^2 + \pi_i)) X$$

(4.10)

which allows from the calculation of a gradient $U_{pen}(\beta)$ using the model matrix, $\beta$, and Equation 4.7.
Likewise, the second partials are given by
\[
\frac{\partial^2 J}{\partial \beta_r \partial \beta_s} = X^T \text{diag}\{m'_i \frac{\partial^2 \pi_i (1 - \pi_i)}{\partial \beta_r \partial \beta_s}\} X \\
= X^T \text{diag}\{m'_i x_{ir} x_{is} (6 \pi_i^2 - 6 \pi_i + 1) \pi_i (1 - \pi_i)\} X
\]
(4.11)

4.2.2 Partial Of trace \(J^{-1} \frac{\partial J}{\partial \beta_s}\)

This calculation is required in the observed information of the next subsection.
\[
\frac{\partial}{\partial \beta_r} \text{trace} \left(J^{-1} \frac{\partial J}{\partial \beta_s}\right) = \text{trace} \left(\frac{\partial J^{-1}}{\partial \beta_r} \frac{\partial J}{\partial \beta_s} + J^{-1} \frac{\partial^2 J}{\partial \beta_r \partial \beta_s}\right) \\
= \text{trace} \left(-J^{-1} \frac{\partial J}{\partial \beta_r} J^{-1} \frac{\partial J}{\partial \beta_s} + J^{-1} \frac{\partial^2 J}{\partial \beta_r \partial \beta_s}\right) \\
= \text{trace} \left(J^{-1} \frac{\partial^2 J}{\partial \beta_r \partial \beta_s} - J^{-1} \frac{\partial J}{\partial \beta_r} J^{-1} \frac{\partial J}{\partial \beta_s}\right)
\]
(4.12)

The important part of this calculation is while the inverses are not well regarded, this calculation is completely specified with Equations 4.10 and 4.11 given \(\beta\) and the model matrix \(X\). In application, we invoke symmetry and use eigenvalue decompositions to invert \(J\). This is sufficient for our current needs.

4.2.3 Observed Information Of A Firth-Like Penalized Log-Likelihood

We indicate the non-penalized Binomial model information as
\[
I^\text{bin} = X W_A X
\]
(4.13)

\(W_A\) is not present in Equation 4.13 due to the construction of this particular penalty. It is present in this calculation because the actual Firth modified-score function is defined from the information of the binomial model. To avoid creating yet another diagonal matrix notation, we reuse the already presented notation.
Assuming regularity conditions are met, we have

\[
I_{rs} = -\frac{\partial U_{\text{pen}}(\beta)}{\partial \beta_r} \\
= t_{rs}^{\text{bin}} - \alpha \frac{\partial}{\partial \beta_r} \text{trace} \left( J^{-1} \frac{\partial J}{\partial \beta_s} \right) \\
= t_{rs}^{\text{bin}} - \alpha \text{trace} \left( J^{-1} \frac{\partial^2 J}{\partial \beta_r \partial \beta_s} - J^{-1} \frac{\partial J}{\partial \beta_r} J^{-1} \frac{\partial J}{\partial \beta_s} \right)
\]  

(4.14)

with the \( rs \) referring to the element in row \( r \) and column \( s \) of the matrix \( I \), so we can calculate this using Equations 4.12 and 4.13.

### 4.2.4 Maximization Of The Complete-Data Penalized Log-Likelihood

Since the gradients of the penalized log-likelihoods in this chapter are specified by Equation 4.10 and the informations are specified by Equation 4.14, we can find maximums of Equation 4.8 by applying Fisher Scoring Iterations until convergence. In application, criteria to determine if a maximum has been discovered can be applied at run time (all of the gradient entries are non-positive and the Jacobian is positive).

### 4.3 Balanced Firth Score-Modification In EM

In this section, we present how to augment our previous EM work from Chapter 3 to fit the balanced Firth modified-score function with missing observations. Section 3.1 continues to apply. The EM overview and notation in Section 3.2 continues to apply when “non” is changed to “pen”, and we are applying EM to a penalized maximum likelihood.

In Chapter 5, we will show a simulation that provides evidence that this balanced Firth adjustment is superior to our straight EM. For our continued work in this section, we will show the penalized conditional expectation, the maximization of that expectation, and then the necessary modification to the information. These will not be difficult, because the balanced Firth score-modification has been chosen to make them all conveniently calculated.
4.3.1 $Q_{\text{pen}}(\theta', \hat{\theta}^{(j)})$ And Conditional Expectation

Starting from Equation 4.8 and aiming to augment Equation 3.3 of Section 3.3.1, we have

\[
Q_{\text{pen}}(\theta', \hat{\theta}^{(j)}) = E[l_{\text{pen}}(\theta'|y, z)|y, \hat{\theta}^{(j)}] \\
= E[l_{\text{non}}(\theta'|y, z)|y, \hat{\theta}^{(j)}] + \alpha E[\log|J||y, \hat{\theta}^{(j)}] \\
= E[l_{\text{non}}(\theta'|y, z)|y, \hat{\theta}^{(j)}] + \alpha \log|J| \tag{4.15}
\]

The balanced Firth score-modification is not a function of the unobserved data, so in the log-likelihood the conditional expectation for the penalty term is identically itself. This would not be true if we had used a penalty that depends on unobserved data, such as the actual Firth penalized log-likelihood. The idea, in general, does emit other possibilities. For example, we might have used the expected $\hat{m}_i$ in place of $m_i$ in the actual Firth score-modification, and then Equation 4.15 would continue to be true (though we will learn the information is different than with this balanced Firth score-modification).

4.3.2 Maximization

As in Section 3.4, again we maximize with respect to $P$ and then with respect to $\beta$. And since the modification to the log-likelihood does not involve $P$, Section 3.4.1 still applies in full.

What changes is how we maximize $Q_{\text{pen}}(\theta', \hat{\theta}^{(j)})$ with respect to $\beta$, but in actual fact, the theme of the original unpenalized maximization remains the same.
Maximization With Respect To $\beta$

The portion of $Q_{pen}(\theta', \hat{\theta}^{(j)})$ that need be considered while maximizing with respect to $\beta$ is:

$$
Q_{\beta} = \sum_{i=1}^{c} \sum_{j=1}^{t} (y_{ij} \log(\pi_{ij}) + \mathbb{E}[z_{ij} | y_{ij}, \hat{\theta}^{(j)}] \log(1 - \pi_{ij})) + \frac{1}{2} \log(||X^tW_BX||)
$$

$$
= \sum_{i=1}^{c} \sum_{j=1}^{t} (y_{ij} \log(\pi_{ij}) + (\tilde{m}_{ij} - y_{ij}) \log(1 - \pi_{ij})) + \frac{1}{2} \log(||X^tW_BX||)
$$

This is the same maximization problem as when treating the $\tilde{m}_{ij}$ as observed, so we proceed by maximizing the binomial likelihood with a balanced Firth-like penalized likelihood via our method outlined in Section 4.2 as if everything had been observed.

4.3.3 Observed Information With A Balanced Firth Modified-Score

This is a simple modification of our unpenalized case for this particular penalty; however, there is a point to make clear in the larger picture.

By Oakes [11] again,

$$
\frac{\partial^2 l_{pen}(\hat{\theta})}{\partial \theta \partial \hat{\theta}} = \left[ \frac{\partial^2 Q_{pen}(\theta, \hat{\theta})}{\partial \theta \partial \hat{\theta}} + \frac{\partial^2 Q_{pen}(\theta', \hat{\theta})}{\partial \theta \partial \hat{\theta}} \right]_{\theta = \hat{\theta}}
$$

$$
= \left[ \frac{\partial^2 Q_{non}(\theta, \hat{\theta})}{\partial \theta \partial \hat{\theta}} + \frac{\partial^2 Q_{non}(\theta', \hat{\theta})}{\partial \theta \partial \hat{\theta}} \cdot \frac{1}{2} \log(||X^tW_BX||) + \frac{\partial^2 Q_{non}(\theta, \hat{\theta})}{\partial \theta \partial \hat{\theta}} + \frac{\partial^2 Q_{non}(\theta', \hat{\theta})}{\partial \theta \partial \hat{\theta}} \cdot \frac{1}{2} \log(||X^tW_BX||) \right]_{\theta = \hat{\theta}}
$$

Now in this particular case, $\frac{\partial^2}{\partial \theta \partial \hat{\theta}} \frac{1}{2} \log(||X^tW_BX||)$ is identically a matrix of zeros as our particular $W_B$ is not a function of conditional expectations. But it is not hard to create variants where that is not true while the expectation and maximization steps are no more difficult than what is here. For example, the diagonal of $W$ might be the actual Firth penalty except with the unobserved $m_i$ replaced with their expected values, $\tilde{m}_i$.

Past work on the topic of penalized likelihoods in EM was done by Segal, Bachetti, and Jewell from Berkeley [13]. They write (using different notation): "The observed data
asymptotic covariance matrix of \( \tilde{\theta} - \theta \) is

\[ V = \{ I_0(\tilde{\theta}|Y_{\text{obs}}) + \lambda D^2 J(\tilde{\theta}) \}^{-1}. \]

Here "\( D^2 J(\theta) \) is the matrix of second partial derivatives of \( J(\theta) \) with respect to \( \theta \)" and \( \lambda \) is a constant. Their \( \lambda J(\theta) \) is an arbitrary penalty function subtracted from the actual log-likelihood being considered. But that result only holds if \( J(\theta) \) has a nominal expected value with respect to the observed data. If that is not the case then, for example, \( \frac{\partial^2}{\partial \theta \partial \theta} \log(X^t W X) \) is not necessarily a matrix of zeros, and what they wrote is an error.

None of that is a concern in this case; working with Equations 4.17 and 3.13, we have,

\[ \frac{\partial^2 l_{\text{pen}}(\tilde{\theta})}{\partial \tilde{\theta} \partial \tilde{\theta}} = \frac{\partial^2 l_{\text{non}}(\tilde{\theta})}{\partial \tilde{\theta} \partial \tilde{\theta}} + \left[ \frac{\partial^2}{\partial \tilde{\theta} \partial \tilde{\theta}} \frac{1}{2} \log(X^t W_B X) \right]_{\tilde{\theta}=\hat{\theta}} \]  

(4.18)

Since \( \log(X^t W_B X) \) is not a function of \( P \), here we are just describing an extra piece to be added unto \( \frac{\partial^2 Q}{\partial \beta \partial \beta} \) described in Appendix A. And in particular, we can perform that calculation using Equation 4.12 to arrive at the new penalized observed information.

4.4 Balanced Firth Penalty Applied To The Rhyzopertha dominica Data

In this section, we begin where we left off in Section 3.6 considering the motivating example introduced in Section 1.1. In Table 4.1, we have a fit of the Rhyzopertha dominica data using the balanced Firth penalized-likelihood. In Table 4.2, we can see the problem of separation of points has been removed by the techniques of this chapter. We have standard errors that could provide usable intervals around our parameter estimates.

The question that remains is what evidence do we have that a Wald confidence interval constructed with these standard errors is meaningful? These are maximized observed likelihood equations, but the rates of convergence are not clear when the number of experimental units with missing covariates to be estimated is on average a fraction of the experimental units. We simulate in the next chapter (Section 5.2) to consider whether or not asymptotic statements about maximized likelihoods apply to the fit for Rhyzopertha dominica data (assuming it were the truth), and in general, we simulate to address whether or not, there are cases where we can trust Wald intervals based on these smaller penalized standard errors.
<table>
<thead>
<tr>
<th></th>
<th>Balanced Firth Score-Modification</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Std. Err.</td>
</tr>
<tr>
<td>genotype-locusA</td>
<td>0.086</td>
<td>0.017</td>
</tr>
<tr>
<td>genotype+locusA</td>
<td>0.024</td>
<td>0.007</td>
</tr>
<tr>
<td>genotype-locusB</td>
<td>0.357</td>
<td>0.043</td>
</tr>
<tr>
<td>genotype+locusB</td>
<td>0.056</td>
<td>0.012</td>
</tr>
<tr>
<td>genotype-locusH</td>
<td>0.355</td>
<td>0.043</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>2.941</td>
<td>1.415</td>
</tr>
<tr>
<td>genotype+</td>
<td>-1.264</td>
<td>1.944</td>
</tr>
<tr>
<td>locusB</td>
<td>5.107</td>
<td>2.746</td>
</tr>
<tr>
<td>locusH</td>
<td>1.210</td>
<td>1.754</td>
</tr>
<tr>
<td>dose</td>
<td>-142.411</td>
<td>43.542</td>
</tr>
<tr>
<td>genotype+:locusB</td>
<td>-3.830</td>
<td>3.429</td>
</tr>
<tr>
<td>genotype+:locusH</td>
<td>-0.277</td>
<td>2.383</td>
</tr>
<tr>
<td>genotype+:dose</td>
<td>139.543</td>
<td>43.557</td>
</tr>
<tr>
<td>locusB:dose</td>
<td>-2167.420</td>
<td>573.599</td>
</tr>
<tr>
<td>locusH:dose</td>
<td>-1112.528</td>
<td>267.315</td>
</tr>
<tr>
<td>genotype+:locusB:dose</td>
<td>2093.190</td>
<td>574.882</td>
</tr>
<tr>
<td>genotype+:locusH:dose</td>
<td>1074.049</td>
<td>267.433</td>
</tr>
</tbody>
</table>

Table 4.1: Techniques of Chapter 4 used to create estimate $\hat{\theta}$ on *Rhizophora dominica* data with standard errors. The first 5 estimates are population proportion ($\hat{P}$) and the remaining estimates are $\hat{\beta}$. 
<table>
<thead>
<tr>
<th>Phosphine Dosage (mg/L)</th>
<th>0.000</th>
<th>0.003</th>
<th>0.004</th>
<th>0.005</th>
<th>0.010</th>
<th>0.050</th>
<th>0.100</th>
<th>0.200</th>
<th>0.300</th>
<th>0.400</th>
<th>1.000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Receiving Dosage ($m_j$)</td>
<td>98</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>300</td>
<td>400</td>
<td>750</td>
<td>500</td>
<td>500</td>
<td>7850</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Balanced Firth Score-Modification</th>
<th>Expected Binomial Index $\hat{m}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>genotype-locusA</td>
<td>10.41 3.92 7.95 10.60 25.96 34.54 65.18 43.08 43.19 678.63</td>
</tr>
<tr>
<td>genotype+locusA</td>
<td>4.24 4.48 2.44 0.40 6.31 8.93 14.54 13.69 12.48 189.53</td>
</tr>
<tr>
<td>genotype-locusB</td>
<td>31.52 43.94 40.18 36.69 108.78 142.60 269.06 177.83 178.28 2801.29</td>
</tr>
<tr>
<td>genotype+locusB</td>
<td>6.22 4.46 2.45 8.58 17.17 22.24 42.32 27.97 28.04 440.66</td>
</tr>
<tr>
<td>genotype-locusH</td>
<td>27.50 35.00 39.84 37.47 108.16 141.79 267.54 176.83 177.27 2785.44</td>
</tr>
<tr>
<td>genotype+locusH</td>
<td>20.59 8.20 7.14 6.25 33.63 49.90 91.35 60.59 60.74 954.45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observed Surviving $y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>genotype-locusA</td>
</tr>
<tr>
<td>genotype+locusA</td>
</tr>
<tr>
<td>genotype-locusB</td>
</tr>
<tr>
<td>genotype+locusB</td>
</tr>
<tr>
<td>genotype-locusH</td>
</tr>
<tr>
<td>genotype+locusH</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Balanced Firth Score-Modification</th>
<th>Probability Of Survival $\hat{\pi}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>genotype-locusA</td>
<td>0.950 0.925 0.915 0.903 0.820 0.015 0.000 0.000 0.000 0.000 0.000</td>
</tr>
<tr>
<td>genotype+locusA</td>
<td>0.842 0.841 0.841 0.839 0.822 0.801 0.751 0.693 0.629 0.233</td>
</tr>
<tr>
<td>genotype-locusB</td>
<td>1.000 0.754 0.233 0.029 0.000 0.000 0.000 0.000 0.000 0.000 0.000</td>
</tr>
<tr>
<td>genotype+locusB</td>
<td>0.950 0.938 0.934 0.929 0.899 0.289 0.009 0.000 0.000 0.000 0.000</td>
</tr>
<tr>
<td>genotype-locusH</td>
<td>0.984 0.595 0.295 0.107 0.000 0.000 0.000 0.000 0.000 0.000 0.000</td>
</tr>
<tr>
<td>genotype+locusH</td>
<td>0.931 0.923 0.920 0.917 0.900 0.632 0.179 0.003 0.000 0.000 0.000</td>
</tr>
</tbody>
</table>

Table 4.2: Techniques of Chapter 4 applied to observed *Rhizopertha dominica* experimental data. A complete separation of points from Table 3.2 has been removed.
Chapter 5
Simulation And Application

In the first example, we consider the simplest model that carries the core features of the motivating dose-response example. In the second example, we consider the motivating dose-response application.

5.1 Simple Dose-Response

In this section, we conduct a simulation involving a simple dose-response scenario. The purpose of this simulation is to provide evidence that our ad hoc balanced Firth score-modification has value.

As a point of clarification for this section, when we refer to experiments 1 and 2 those are not simulations 1 and 2. Experiments 1 and 2 refer to a different treatment structures that we will compare. Treatment Structure 1 is the same as Treatment Structure 2 except there will be a dose level where almost all the beetles survive to get their class viewed. We will learn that Treatment Structure 1 is a much better design for employing this EM algorithm.

To further ease vocabulary in this example, we will imagine we are dealing with a beetle-like pest. The pest comes in two classes (0 and 1) with no other covariates of interest. The beetles are randomly assigned to levels of a poison dose (treatment factor). We do not model poison dose as a continuous covariate because this is a simple abstraction. Each experiment has 600 beetles to estimate the class effect. We will not be interested in the poison effect for this section.

We do our work with two treatment structures. In the first treatment structure, we have three levels of poison dose: “low”, “mid”, and “high”. The second treatment structure just has two levels: “medium” and “high”. So that the first treatment structure has 200
beetles per a level of poison dose, and the second treatment structure has 300 beetles per a level of poison dose.

To make the comparison as direct as possible, we simulated the class of 600 beetles and left those class assignments unchanged as we moved across treatment structures. For each simulated set of 600 beetles, we first simulated what would happen in Treatment Structure 1 and then what would happen in Treatment Structure 2 before drawing another simulated set of 600 beetles. For the first experiment having drawn 600 random beetles, we partition at 201 and 401 to break apart the result and assign them to levels of dose. In the second experiment, we broke apart the class vector at 301 to assign them to two levels of poison.

5.1.1 Models

Unbalanced populations seemed to be more realistic, so we set population proportions on class 0 at .75 while class 1 has a population proportion of .25.

Our model for treatment structure 1 and 2 is:

\[
\log \frac{\pi_{ij}}{1 - \pi_{ij}} = \mu + C_i + D_j
\]

(5.1)

\(\mu\) is the grand mean. \(C_i\) is the effect due to the ith class, and \(i \in \{0, 1\}\) is the class effect index. And \(D_j\) is the effect due to the jth level of dose; treatment structure 1 has \(j \in \{0, 1, 2\}\) while treatment structure 2 has \(j \in \{1, 2\}\).

There is no interaction term because we want to simulate on the simplest informative case.
We parameterized the model matrix for these two treatment structures as:

\[
X_1 = \begin{pmatrix}
1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
1 & 1 & 1 & 0 \\
1 & 0 & 0 & 1 \\
1 & 1 & 0 & 1
\end{pmatrix} \quad X_2 = \begin{pmatrix}
1 & 0 & 0 \\
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & 1 & 1
\end{pmatrix}
\]

(5.2)

The subscripts refer to Treatment Structures 1 and 2. For \(X_1\), the intercept applies to the low dose level 0 effect. The columns for dose then apply the multiple of the coefficient for the difference between dose levels 1 and 2 from level 0. For \(X_2\), the intercept applies to the mid dose level 1 effect. The column for dose 2 then applies to the coefficient for the difference between dose levels 2 and 1.

**Coefficients**

The simulations over the two treatment structures were not done on one parameterization of coefficients, but rather a family of coefficients that was conceived as a grid of dose 2 effect versus class 1 effect, so that we could view information about the results in a contour plot.

**Dose Effects**

The effect of the low dose level is fixed in this family. Conceptually it represents a dose where almost all the beetles survive to get their class viewed. Consistent with that it was chosen to give class 0 beetles a .90 probability of surviving dose 0, and so \(D_0 = 2.19731\).

The effect of the mid dose level is fixed in this family. It represents a dose approximately LD50 for at least one of the classes: it was selected to be the LD50 dose for class 0. This makes the mid dose level effect equal to 0, and so \(D_1 = 0\).
The effect of the high dose level varies. The range of values for dose 2 begin at $-1$ and end at $-6$. This family only considers 10 values in that range, so for all simulations

$$D_2 \in \{-1.00, -1.56, -2.11, -2.67, -3.22, -3.78, -4.33, -4.89, -5.44, -6.00\}$$

Class Effects

The effect of class 0 is fixed with $C_1 = 0$. The effect of class 1 varies, but is always non-negative. The range of values for class 1 begin at 0 and end at 5, and again this family only considers 10 values in that range. For all simulations

$$C_1 \in \{0.00, 0.56, 1.11, 1.67, 2.22, 2.78, 3.33, 3.89, 4.44, 5.00\}$$

5.1.2 All Simulations

The crossing of the two effects that are allowed to vary creates 100 different true parameterizations. Each of these got approximately 236 simulations on each of the two treatment structures. The goal was to assess how accurately we would fit the class effect without viewing the class of the beetles that died. In each simulation, we fit exact the true model, our techniques of Chapters 2 through 4 only had to overcome finding the correct parameterizing of the model.

5.1.3 Results

For the plots related to this section, we, quite extensively, used R's contour routine with default smoothing. The red tick marks in the contour plots indicate the grid of parameters used for simulation.

Hindrance

It is worth reviewing new abstract ideas that are working against successfully estimating model parameters because of missing observations. We only can find two new issues due to missing data, and we think of them as a source of hindrance to creating nearly correct
estimates.

The amount of unobserved data is certainly a source of difficulty in estimation, but also, the proportion of observed classes being different than the infinite population from which we pulled experimental units is going to be a hindrance. The first, the amount of missing data defines the character of these problems, but by themselves unobserved covariates are not particularly troublesome. If we were somehow guaranteed that the missing data was distributed according to a multinomial population for which we knew the population proportion than all we would have from this process is a little extra variance in the Binomial coefficients. But we are doing more than that. We are reconstructing the population proportions from a view of the data that, depending on the treatments, is increasingly distorted from reality.

We offer contour maps of hindrance (expected missing data and distorted perspective) in Figure 5.1.

In Figures 5.1a and 5.1b, we see expected proportions of experimental units that will have unobserved class (this is the proportion of beetles not expected to survive their treatments). In Figure 5.1b, Treatment Structure 2 possesses quite a bit more expected missing data for most treatment covariates; but Treatment Structure 1 still is hindered by 34 to 52 percent expected unobserved class.

In Figures 5.1c and 5.1d, we have the observed relative proportion of class 0 to class 1. In reality this ratio is 3 for the larger simulated population. Where $C_1 = 0$, we see what we expect, which is the ratio to be 3. But as the $C_1$ effect increases a disportionate proportion of class 1 beetles survive their treatments. We will speculate that eventually the distorted perspective about the population proportions overwhelms the machinery of our EM algorithm.

**Point Estimates**

In this and subsequent sections, we use a trimmed-mean of all the simulations at a particular true covariate profile (point) to create our plots. The trimming is very strong with the lower quarter and upper quarter of observations moved to the .25 and .75 percentiles
Fig. 5.1: Hindrance plots for a simple dose-response simulation. In these plots, we consider attributes intrinsic to the underlying true parameterization and experimental design that make accurate point estimates increasingly difficult to create. In (a) and (b), we have the expected proportion of our experimental units that will not have observed class. And in (c) and (d), we have the proportion of class 0 to class 1 that is observed in the survivors.
in value before the mean is calculated. The point of this is not to obscure anything; in fact, it does not matter to the shape of the contours we are presenting. The point of this is to rule out the possibility of extreme observations creating the "bad plots", and thus making the better plots look better by contrast. There is nothing of that sort going on here.

What we are assessing in this section is bias. We will see if the trimmed-means for the point estimate for the $C_1$ effect is consistently low or high.

**Treatment Structure 1**

In Figure 5.2, we have a four plot of trimmed-means of simulation point estimates for the $C_1$ effect using Treatment Structure 1. This was the treatment structure with a near zero level of dose that becomes nearly fully observed.

Method 1 is a fit on the simulated data for Treatment Structure 1 with all class data completely observed using no penalty to create the likelihood fit. These contours are plotted in Figure 5.2a. This method recovers the $C_1$ effect with no particular bias.

Method 2 is a fit on the simulated data for Treatment Structure 1 with all class data completely observed but using the ad hoc balanced Firth penalized likelihood fit. These contours are plotted in Figure 5.2b. This method recovers the $C_1$ effect with no particular bias.

Method 1 and 2 together illustrate that the use of this penalty in this particular experiment is irrelevant if the data are fully observed.

Method 3 is a fit on the simulated data for Treatment Structure 1 with the class of the dead beetles unobserved using no modification to the score function. These contours are plotted in Figure 5.2c. This is not an impressive recovering of the underlying $C_1$ effect for large parts of the plot. In Figure 5.4a, we can see the areas of poor fit correspond to massive standard errors. The techniques used for this plot are those of Chapter 3.

Method 4 is a fit on the simulated data for Treatment Structure 1 with the class of the dead beetles unobserved using the balanced Firth modification to the score function. These contours are plotted in Figure 5.2d. The techniques of this plot are those of Chapter 4. This is actually a quite nice recovering of the underlying $C_1$ effect for most of the plot:
impressive given the circumstances. There is a little unexplained curve in the lower right of the plot that likely has to do with the bias induced by the balanced Firth penalty as the class 1 beetles almost always live. Overall when this plot is compared to 5.2c, there is good evidence that we are justified in using this penalty, and in particular, we can see in Figure 5.4b that this technique is not troubled by the huge standard errors for certain covariate profiles as is Method 3.

**Treatment Structure 2**

In Figure 5.3, we have a four plot of trimmed-means of simulation point estimates for the $C_1$ effect using Treatment Structure 2. This was the treatment structure without a nearly fully observed level of dose.

As before with Treatment Structure 1 referring to Figures 5.3a and 5.3b, there is no particular difference in the complete observation estimates with or without penalty for Treatment Structure 2. We will call them Method 1 and Method 2 to maintain parity with the previous subsection.

Method 3 is a fit on the simulated data for Treatment Structure 2 with the class of the dead beetles unobserved using no modification to the score function. These contours are plotted in Figure 5.3c. This is a particularly poor estimation of $C_1$ through large regions. Again just like with Treatment Structure 1 without a penalized likelihood, we can see in Figure 5.4c that the regions of poor estimation correspond to regions with large standard errors.

Method 4 is a fit on the simulated data for Treatment Structure 2 with the class of the dead beetles unobserved using a balanced Firth modification to the score function. These contours are plotted in Figure 5.3d. This method has a profound bias in much of the same area that the non-penalized fits are also poor. We know in this region that the Hessian of the unpenalized log-likelihood is describing a nearly flat curve because this is how we get the large standard errors of Figure 5.4c. But then this is also the region where the actual Firth penalty term would be the largest as it is the log of a near zero determinant of that non-penalized Hessian. The balanced Firth penalty term would be increased in much the
Fig. 5.2: For Treatment Structure 1, trimmed-means of simulated point estimates for the class 1 effect.
(a) Complete Observations Without Penalty

(b) Complete Observations With Balanced Firth Penalty

(c) Missing Observations Without Penalty

(d) Missing Observations With Balanced Firth Penalty

Fig. 5.3: For Treatment Structure 2, trimmed-means of simulated point estimates for the class 1 effect.
Fig. 5.4: Trimmed-means of simulated standard errors for the class 1 effect. All plots are for simulated fittings using methods that have unobserved class covariates on dead beetles. Use of the balanced Firth ad hoc penalty can restrain standard errors.
same way, but now the concern is this is also region where the balanced Firth penalty can be most deviated from the actual Firth bias-reducing penalty in some systematic way.

We speculate for Treatment Structure 2, we are seeing the consequences of choosing a score-modification that was an ad hoc deconstruction of the actual bias-reducing score modification in 5.3d, but from a larger perspective, the lesson is that this treatment structure should be avoided when using these techniques.

5.1.4 Remarks

We can see from the simulation study for Treatment Structure 1 that we are justified in considering the balanced Firth ad hoc modification to the score function, but we also learn that the actual design of the experiment can have a profound effect on whether these methods work well.

When we reflect on the hindrance plots in Figure 5.1 as they might relate to the point estimate plots, it is interesting that both Treatment Structure 1 and 2 seem to have trouble in their point-estimate bias when the observed class 0 to class 1 ratio is under 1.6. We will learn in the next simulation in Section 5.2 that these methods tend to overestimate the population of class that is most adept at surviving.

A distorted perspective about population proportion would seem to dominate “missingness”, for explaining bias. Consider the regions of the point estimate plots (Figures 5.2 and 5.3) most biased were also the regions of the hindrance plots (Figure 5.1) with the least amount of missing class observations. As the class 1 effect increases, the extra non-missing observations are increasing only associated with class 1.

From this simulation study, we offer that our method is best suited to experiments that have at least one treatment in which most of the experiment units have their class observed (survive) and are estimating subtle effects. The second criteria is somewhat surprising. This simulation study suggests this method is at its best for smaller differences among class effects; the method is at its best in a sweet spot where the numerical mechanics are not overwhelmed by an increasingly skewed estimation of the population combined with a separation of observed outcomes partitioned by class.
5.2 Applied Dose-Response With *Rhyzopertha dominica*

Changing topics from the previous section, we seek in simulation to find evidence that we can trust the standard errors of Table 4.2, find evidence about any asymptotic normality in our fit coefficient distribution, and expose bias.

5.2.1 Simulation

We assume that our modified-score function fit from Section 4.4 is the truth. The hope is that things we discover while assuming it is true for simulation apply in a neighborhood around this particular parameterization.

We simulated from this model with no change to the parameterization 4995 times. 1 of these simulations are removed due to not converging in 10000 iterations. 69 more are removed due to numerical issues (NAs where there should be numbers), leaving 4885 simulated fits to inspect.

In light of the fact that this simulation can (very rarely) cause our methods to numerically struggle, we elect to remove the 99.9th upper and 0.1th lower percentile from the data set. This is a modest pruning of the most extreme observations. If any simulation had a fit parameter in the lowest 5 or highest 5 observations, the entire simulation is removed. After this process, we have 4781 simulated fits.

5.2.2 Bias

In Table 5.1, we have collected observations of bias. For each element of $\theta$, we have the sample mean of $\hat{\theta}$ (column 2). We believe that after this many simulations the expected value is well approximated by the sample mean. Bias (column 3) is formed as the true value minus the sample mean.

But what makes a bias large? We first considered column 4 as bias divided by true value, though we preferred to state the proportion of the sample mean created by the bias, which is in column 4. Even still, this is a measurement not much related to how we reason with statistics, so in column 5, we have the sample mean of the error for an observation (fit minus truth) divided by the standard error for that observation. This is the bias on
<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Estimate</th>
<th>Mean Estimate</th>
<th>Bias</th>
<th>Mean Error Proportion Of Mean Est.</th>
<th>Mean Error Proportion Of Std. Err.</th>
<th>Mean Error Proportion Of Wald Int.</th>
</tr>
</thead>
<tbody>
<tr>
<td>genotype-locusA</td>
<td>0.086</td>
<td>0.093</td>
<td>0.007</td>
<td>0.075</td>
<td>0.209</td>
<td>0.053</td>
</tr>
<tr>
<td>genotype+locusA</td>
<td>0.024</td>
<td>0.026</td>
<td>0.002</td>
<td>0.077</td>
<td>0.064</td>
<td>0.016</td>
</tr>
<tr>
<td>genotype-locusB</td>
<td>0.357</td>
<td>0.356</td>
<td>-0.001</td>
<td>-0.003</td>
<td>-0.048</td>
<td>-0.012</td>
</tr>
<tr>
<td>genotype+locusB</td>
<td>0.056</td>
<td>0.060</td>
<td>0.004</td>
<td>0.067</td>
<td>0.139</td>
<td>0.035</td>
</tr>
<tr>
<td>genotype-locusH</td>
<td>0.355</td>
<td>0.343</td>
<td>-0.012</td>
<td>-0.035</td>
<td>-0.277</td>
<td>-0.071</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>2.941</td>
<td>2.478</td>
<td>-0.463</td>
<td>-0.187</td>
<td>-0.408</td>
<td>-0.104</td>
</tr>
<tr>
<td>genotype+</td>
<td>-1.264</td>
<td>-1.006</td>
<td>0.258</td>
<td>-0.256</td>
<td>0.113</td>
<td>0.029</td>
</tr>
<tr>
<td>locusB</td>
<td>5.107</td>
<td>5.657</td>
<td>0.550</td>
<td>0.097</td>
<td>0.047</td>
<td>0.012</td>
</tr>
<tr>
<td>locusH</td>
<td>1.210</td>
<td>2.115</td>
<td>0.905</td>
<td>0.428</td>
<td>0.345</td>
<td>0.088</td>
</tr>
<tr>
<td>dose</td>
<td>-142.411</td>
<td>-152.096</td>
<td>-9.685</td>
<td>0.064</td>
<td>0.420</td>
<td>0.107</td>
</tr>
<tr>
<td>genotype+:locusB</td>
<td>-3.830</td>
<td>-4.802</td>
<td>-0.972</td>
<td>0.202</td>
<td>-0.189</td>
<td>-0.048</td>
</tr>
<tr>
<td>genotype+:locusH</td>
<td>-0.277</td>
<td>-0.786</td>
<td>-0.509</td>
<td>0.648</td>
<td>-0.139</td>
<td>-0.036</td>
</tr>
<tr>
<td>genotype+:dose</td>
<td>139.543</td>
<td>149.360</td>
<td>9.817</td>
<td>0.066</td>
<td>-0.415</td>
<td>-0.106</td>
</tr>
<tr>
<td>locusB:dose</td>
<td>-2167.420</td>
<td>-2177.005</td>
<td>-9.585</td>
<td>0.004</td>
<td>0.234</td>
<td>0.060</td>
</tr>
<tr>
<td>locusH:dose</td>
<td>-1112.528</td>
<td>-1194.887</td>
<td>-82.359</td>
<td>0.069</td>
<td>-0.006</td>
<td>-0.002</td>
</tr>
<tr>
<td>genotype+:locusB:dose</td>
<td>2093.190</td>
<td>2111.567</td>
<td>18.377</td>
<td>0.009</td>
<td>-0.217</td>
<td>-0.055</td>
</tr>
<tr>
<td>genotype+:locusH:dose</td>
<td>1074.049</td>
<td>1154.149</td>
<td>80.100</td>
<td>0.069</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 5.1: Observed indications of bias for our *Rhyzopertha dominica* data simulation. The column on the right is the mean of the errors on the scale of a corresponding Wald interval with \( \alpha = .05 \) for each observation. The first 5 estimates are population proportion (\( \hat{P} \)), and the last are \( \hat{\beta} \).
the scale of the calculated standard error averaged over every observation. This is closer to
being meaningful. In fact, for large sample asymptotics this is the sample mean z-statistic
for a known mean.

This is closer to what we want, but we typically do not reason on the z-value scale. We
often reason on the p-value scale, but those don't apply here. From a z-value scale one can
calculate the proportion of the Wald interval that the error would be for each simulated
parameter estimate. Here the Wald interval is created with \( \alpha = .05 \). We have averaged over
these measures of error to form column 6. If there is no bias this column should be entirely
0.

In column 6, we have the sample means of the error on the scale of the Wald interval.
This has a fairly intuitive interpretation. We picture the Wald interval, and then we picture
shifting it up or down by the proportion of its size indicated. If the Wald interval is
appropriate, this is the average effect of the bias on any inference we might make.

While that is interpretable, it is not the entire story. Means can be shifted by location
parameters being altered, but also means can be shifted by skewing the distribution.

Consider the population proportion for genotype-locusA and the effects for \( \text{dose} \) and
genotype+dose. The population proportion has average error of .05 of its Wald intervals,
and the two effect estimates have average error of .10 of their Wald intervals. We will see
evidence that it is skewing away from normal distributions that causes these location shifts.

5.2.3 Distribution

Figures 5.5 and 5.6 represent the observed distributions of balanced Firth penalized
EM parameter estimates.

To construct Figures 5.5 and 5.6, we first standardized each observation of a particular
estimate, and then we plotted them over-layed on a normal-quantile plot. To standardize
these data, for each parameter estimate we treat them as sampled data with several thousand
observations. We center them and divide by the sample standard deviation before plotting
them on a normal-quantile plot. The black line is a perfectly normal distribution.
Fig. 5.5: Sampled estimates for $\hat{P}$ are standardized and then displayed overlayed in a normal-quantile plot.

Fig. 5.6: Sampled estimates for $\hat{\beta}$ are standardized and then displayed overlayed in a normal-quantile plot.
What we are addressing here is whether or not an assumption that these coefficients are normally distributed around their mean is appropriate. Most of these are well approximated by normality. These are quantiles out to extreme observations. There are three notable departures.

In Figure 5.5, we have the normal-quantile curve for genotype-locus\textsuperscript{A} showing a distinct right skewing. This was also a slightly biased estimate. We believe this is part of a larger picture, in which this technique tends to over-estimate the population of anything that strongly over-represents itself in the extremes of the treatment structure.

In Figure 5.6, we have two problematic normal-quantile curves for dose and genotype+dose. These are both biased through skewing while also exhibiting mirrored departures from normality. We have no explanation.

5.2.4 Coverage Probabilities

We ask the question how often would a Wald interval constructed with $\alpha = .05$ centered on the estimate cover the truth. The resulting observed coverage probabilities are in Table 5.2. We can see from the coverage probabilities that these Wald intervals conservatively cover the truth quite often even though we don’t always have true normality.
<table>
<thead>
<tr>
<th>Coverage Probability</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>genotype-locusA</td>
<td>0.989</td>
</tr>
<tr>
<td>genotype+locusA</td>
<td>0.997</td>
</tr>
<tr>
<td>genotype-locusB</td>
<td>0.951</td>
</tr>
<tr>
<td>genotype+locusB</td>
<td>0.986</td>
</tr>
<tr>
<td>genotype-locusH</td>
<td>0.951</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>0.985</td>
</tr>
<tr>
<td>genotype+ dose</td>
<td>1.000</td>
</tr>
<tr>
<td>locusB</td>
<td>0.996</td>
</tr>
<tr>
<td>locusH</td>
<td>0.992</td>
</tr>
<tr>
<td>dose</td>
<td>0.947</td>
</tr>
<tr>
<td>genotype+:locusB</td>
<td>1.000</td>
</tr>
<tr>
<td>genotype+:locusH</td>
<td>0.998</td>
</tr>
<tr>
<td>genotype+:dose</td>
<td>0.949</td>
</tr>
<tr>
<td>locusB:dose</td>
<td>0.948</td>
</tr>
<tr>
<td>locusH:dose</td>
<td>0.977</td>
</tr>
<tr>
<td>genotype+:locusB:dose</td>
<td>0.950</td>
</tr>
<tr>
<td>genotype+:locusH:dose</td>
<td>0.976</td>
</tr>
</tbody>
</table>

Table 5.2: Observed coverage probabilities for Wald intervals with $\alpha = .05$. The first 5 estimates are population proportion ($\hat{P}$), and the last are $\hat{\beta}$. 
Chapter 6
Discussion

In most applied statistics papers the authors display an elaborate model but retain the luxury of stating they used a publicly available statistical package to do the actual work. It is a luxury because while stating their model they are free to work with just the big picture knowing that if any details are left out that a quick reference to how they generated the actual numbers will imply everything that anyone might need out of the small details.

We do not have that luxury in this work. All calculations are done exactly with the equations offered in the paper. Even the level of generality of the paper is exactly that of the implementation. If it seems that we passed over the big picture with an unusual brevity it is because we had so many small details to cover to be consistent with the goal of recreatability from just what is in the report. While we believe we have met that goal, if anyone feels different the source code will be made available on request.

6.1 On EM With Missing Categorical Covariates

Our intuition about this method gelled nicely around Table 6.1. We start with a large multinomial with 2 times the outcomes represented in $m_j$. A group of multinomial outcomes together, the failures, are observed only as a marginal total. Each multinomial cell outcome has an estimated probability. We take the estimated cell probabilities for all the failure results and rescale them so they sum to one. That informs us how we should expect the unobserved classes to be parcelled up among the failures.

What is exciting about this intuition is it was never about success or failure or complete binomial counts partitioned against incomplete binomial observations. It was about knowing what cells your missing observations might belong to and being able to form a marginal estimated probability on that group.
<table>
<thead>
<tr>
<th>Class</th>
<th>Observed Values</th>
<th>Prior To Expected Values</th>
<th>After Expected Values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Successes</td>
<td>Failures</td>
<td>Cell Probabilities</td>
</tr>
<tr>
<td>1</td>
<td>$y_{1j}$</td>
<td></td>
<td>$\hat{P}<em>1^{(k)} \pi</em>{1j}$</td>
</tr>
<tr>
<td>2</td>
<td>$y_{2j}$</td>
<td></td>
<td>$\hat{P}<em>2^{(k)} \pi</em>{2j}$</td>
</tr>
<tr>
<td>3</td>
<td>$y_{3j}$</td>
<td></td>
<td>$\hat{P}<em>3^{(k)} \pi</em>{3j}$</td>
</tr>
<tr>
<td>4</td>
<td>$y_{4j}$</td>
<td></td>
<td>$\hat{P}<em>4^{(k)} \pi</em>{4j}$</td>
</tr>
<tr>
<td>5</td>
<td>$y_{5j}$</td>
<td></td>
<td>$\hat{P}<em>5^{(k)} \pi</em>{5j}$</td>
</tr>
<tr>
<td>6</td>
<td>$y_{6j}$</td>
<td></td>
<td>$\hat{P}<em>6^{(k)} \pi</em>{6j}$</td>
</tr>
</tbody>
</table>

Table 6.1: An outline of our EM for Multinomial realization $j$. On the left is the observation with which we begin: the count of successes by cell and the count of failures grouped together as a marginal total. In the center, we have the prior probabilities of one observation being associated with each particular cell. Between the center and the right, we calculate the marginal probability associated with $z_{j}$ as $P_{z_{j}} = \sum_{i=1}^{c} P_{i} (1 - \pi_{ij})$. On the right, the expectations for the missing cell counts is the marginal count $z_{j}$ multiplied by the proportion of the marginal probability that belongs to that cell. After taking that expectation, we fit new $\hat{P}$ and $\hat{\pi}$ on these new cell counts. We iterate creating new values for the center and right tables until convergence.
There are generalization of this work for arbitrary configurations of missing data, and also for groups of cells with partial missing data provided the researcher can show MAR is valid. \( z_j \) represents arbitrary groupings that happen to be a particular column in this example. Referring to Table 6.1, with some mechanism to account for the probability of an observation being missing conditioned on the cell of interest there is no need for the column to be entirely missing, be an entire column, or even be a single group of missing observations. It might be that the column is partially observed or that there are two groups with missing observations (such as successes and failures partially observed), etc. In either case, the expectation assigned to each cell would be the marginal total of missing observations for that group multiplied by the proportion of the total marginal probability associated with that group that belongs to a particular cell. The reason we did not present a more general form, is that, we are still working on a simple expression for the generalization.

Once we have created those generalizations, we will, with full thunder of boisterous ignorance, be treading on the well researched domain of multiple imputation techniques. The consequence is that we have a much larger and well known body of literature to review.

For the observed information calculations in Appendix A, we vetted those equations via Maple. We symbolically encoded a specific complex example in Maple where each \( \eta_i \) (Equation 2.1) parameterization from the binomial model was coded by hand. We also provided symbolic forms of expectation calculations etc. From there, we had Maple symbolically differentiate \( Q \) to create the observed information matrix; the symbolic forms Maple uses are generally not viewable though as they are quite long. Maple had no particular simplifications with which to work. We applied the fit parameters to these symbolic forms to get the numerical observed information matrix which we could export. Numerically the result from Maple matched our implementation in R to the 6th significant digit.

6.2 On Balanced Firth Modified-Score Functions

Desirable would be bias calculations on the balanced Firth modified-score functions before applying ad hoc deconstructions of score functions. We were not able to complete that work. Firth does layout bias calculations in terms of cumulants that apply to these
score functions (if the "Null Expectation" of the score modification is 0) [5] which is a tantalizing beginning, but the work is quite difficult. Interested readers might find a paper by Anderson and Richardson [1], a gentler introduction to related topics. The earliest elements of these topics that we encountered was in a 1968 paper by Cox and Snell [3].

It is interesting that we have nowhere else found a version of our observed information calculation for the actual Firth score-modification (Equation 4.14). Heinze notes that its first-order approximation is just the normal Binomial information [7]. Implementations we encountered in R, such as brglm [9] (biased-reduction glm), fit the actual Firth score-modification, and then give the actual Binomial variance-covariance matrix when it is requested.

6.3 On Application

In the beetle dose-response example, there initially was not such a great issue with using the unmodified score function to fit the data. In 2006, Stevens fit this data using EM with Maple doing much of the analytic work on confidence intervals. The code was using a different model matrix, fitting explicitly the probability of mortality, and using $P_1$ instead of $P_6$ as the population proportion indexed by all the rest. In those fits, only a handful of parameters had issues with infinite variances; in addition, the actual statements of variances were large, but not jaw dropping large.

When we worked towards generality in R, it became natural to drop the probability of mortality and start working with the probability of success. The change to the last population proportion being the dependent one was natural because of the way they are handled in for loops and placed into Hessians (starting at index 1,1 in the Hessian). But after we changed all these things the fit was dramatically different and every standard error was showing signs of a flat Hessian.

This caused us to worry, and we recoded Maple to offer its own version of the Hessian consistent with this new model matrix and dependent $P$. It turned out our work in R was correct though. I believe the lesson learned is that once several parameters in your model start showing signs of being influenced by a flat Hessian, you should be extremely
suspicious of making any inference with that model. In the end, once we were sure it was not from a bug in our code, it became a beautiful backdrop to introduce the balanced Firth score-modification.

When we look around for further applications, the thing that comes to mind most is situations where the data censors itself as part of a Binomial event, but the larger usefulness of this report is laying the groundwork for solutions to problems with more general specifications of missing data. This field is dominated by multiple imputation. As we submit this work, we remain optimistic for its utility.
References


Appendices
Appendix A

Observed Information For Non-Penalized EM

Continuing from Equation 3.13 will resolve these two matrices by applying first semester Calculus. This is a fairly blue-collar solution involving no matrix algebra. To introduce some notation consider

\[
\frac{\partial^2 Q_{\text{non}}(\theta, \hat{\theta})}{\partial \theta \partial \theta} = \begin{bmatrix}
\frac{\partial^2 Q}{\partial P \partial P} & \frac{\partial^2 Q}{\partial P \partial \beta} \\
\frac{\partial^2 Q}{\partial \beta \partial P} & \frac{\partial^2 Q}{\partial \beta \partial \beta}
\end{bmatrix}
\] (A.1)

and

\[
\frac{\partial^2 Q_{\text{non}}(\theta, \hat{\theta})}{\partial \theta \partial \hat{\theta}} = \begin{bmatrix}
\frac{\partial^2 Q}{\partial P \partial \hat{P}} & \frac{\partial^2 Q}{\partial P \partial \hat{\beta}} \\
\frac{\partial^2 Q}{\partial \beta \partial \hat{P}} & \frac{\partial^2 Q}{\partial \beta \partial \hat{\beta}}
\end{bmatrix}
\] (A.2)

We are decomposing these larger matrices into sectors, and for each we can express explicit partial derivatives. We show how to compute each in turn; starting with the first matrix and going across rows, one by one. In each computation, we will discard terms which clearly have partial derivatives equal to zero without comment. Also, our convention is to use subscripts \(u\) and \(v\) to represent the \(u,v\) entry of the corresponding Hessian.

In these sections when you see a summation, we follow that summation by asking ourselves what happens for specific indexes. Once \(u\) and \(v\) are given, usually there is just a couple of specific indexes of importance with many of the terms in the summation contributing zero to the overall result.
A.1 Compute $\frac{\partial^2 Q}{\partial P \partial P}$

This is a $(c - 1)$ by $(c - 1)$ matrix. We need only consider

$$Q_{l} = \sum_{i=1}^{c} \tilde{m}_{i} \log P_{i}$$  \hfill (A.3)

So

$$\frac{\partial Q_{l}}{\partial P_{u} \partial P_{v}} = \frac{\partial}{\partial P_{u}} \frac{\partial}{\partial P_{v}} \sum_{i=1}^{c} \tilde{m}_{i} \log P_{i}$$  \hfill (A.4)

$\tilde{m}_{i}$ is a constant with respect to these non-estimate $P$.

Recalling $P_{c} = 1 - \sum_{i=1}^{c-1} P_{i}$, for all $u$ and $v$.

$$\frac{\partial \tilde{m}_{c} \log P_{c}}{\partial P_{c}} = -\frac{\tilde{m}_{c}}{P_{c}}$$  \hfill (A.5)

And then

$$\frac{\partial}{\partial P_{u}} \frac{\partial}{\partial P_{v}} \tilde{m}_{c} \log P_{c} = -\frac{\tilde{m}_{c}}{P_{c}^2}$$  \hfill (A.6)

The other piece of $Q_{l}$ not zero under these two partials occurs if $u = v$ and not otherwise. Assuming $u = v = i$ we have:

$$\frac{\partial \tilde{m}_{u} \log P_{u}}{\partial P_{u} \partial P_{v}} = -\frac{\tilde{m}_{u} \log P_{u}}{P_{u}^2}$$  \hfill (A.7)

So there are two cases for these partials

$$\frac{\partial Q_{l}}{\partial P_{u} \partial P_{v}} = \begin{cases} \frac{-\tilde{m}_{c}}{P_{c}^2} & u \neq v \\ \frac{-\tilde{m}_{u}}{P_{u}^2} - \frac{\tilde{m}_{u}}{P_{u}^2} & u = v \end{cases}$$  \hfill (A.8)

A.2 Compute $\frac{\partial^2 Q}{\partial P \partial \beta}$

This is a $(c - 1)$ by $p$ matrix of zeros.
A.3 Compute $\frac{\partial^2 Q}{\partial \beta \partial \beta}$

This is a $p$ by $(c - 1)$ matrix of zeros.

A.4 Compute $\frac{\partial^2 Q}{\partial \beta \partial \beta}$

This is a $p$ by $p$ matrix. We need only consider

$$Q_{IV} = \sum_{i=1}^{c} \sum_{j=1}^{t} (y_{ij} \log(\pi_{ij}) + \hat{z}_{ij} \log(1 - \pi_{ij}))$$

$$= \sum_{i=1}^{c} \sum_{j=1}^{t} (y_{ij} \log(\pi_{ij}) + (\hat{m}_{ij} - y_{ij}) \log(1 - \pi_{ij}))$$

$$\text{(A.9)}$$

The Hessian of this is the well known result from binomial models with logistic links. So that

$$\frac{\partial^2 Q}{\partial \beta \partial \beta} = -X^T W X$$

\text{(A.10)}

where the $X$ is our model matrix and $W$ is a $n$ by $n$ diagonal matrix with $W_{ii} = m_i \pi_i(1 - \pi_i)$. $m_i$ refers to the $i$th entry of $\hat{m}$ which for this computation is filled with expected values from $\hat{m}$.

A.5 Compute $\frac{\partial z_{ij}}{\partial \hat{P}_v}$

This is a partial derivative required for upcoming work.

$$\frac{\partial z_{ij}}{\partial \hat{P}_v} = \frac{\partial}{\partial \hat{P}_v} \left( z_{ij} \frac{\hat{P}_i(1 - \hat{\pi}_{ij})}{\sum_{k=1}^{c} \hat{P}_k(1 - \hat{\pi}_{kj})} \right)$$

$$= \frac{\partial}{\partial \hat{P}_v} \left( z_{ij} \frac{f_{ij}}{g_j} \right)$$

\text{(A.11)}

$z_{ij}$ is constant with respect to $\hat{P}_v$. Also $f_{ij}$ and $g_j$ have been added as terse aliases for their respective terms.
Now the derivative of $f_{ij}$ with respect to $\hat{P}_v$ has three cases, assume $v \neq c$,

$$\frac{\partial f_{ij}}{\partial \hat{P}_v} = \begin{cases} \hat{\pi}_{ij} - 1 & i = c \\ 0 & i \neq c, i \neq v \\ 1 - \hat{\pi}_{ij} & i = v \end{cases} \quad (A.12)$$

Again noting that $P_c$ depends on the others, for $g_j$,

$$\frac{\partial g_j}{\partial \hat{P}_v} = \hat{\pi}_{cj} - \hat{\pi}_{vj} \quad (A.13)$$

By the quotient rule,

$$\frac{\partial \hat{z}_{ij}}{\partial \hat{P}_v} = z_j \frac{g_j \frac{\partial f_{ij}}{\partial \hat{P}_v} - f_{ij} \frac{\partial g_j}{\partial \hat{P}_v}}{g_j^2} \quad (A.14)$$

Which is fully specified by the cases above.

**A.6 Compute $\frac{\partial^2 Q}{\partial P \partial \hat{P}}$**

For this calculation, we need only consider

$$Q_V = \sum_{i=1}^{c} \sum_{j=1}^{t} \hat{z}_{ij} \log P_i \quad (A.15)$$

And

$$\frac{\partial \log P_i}{\partial \hat{P}_u} = \begin{cases} -\frac{1}{P_c} & , i = c \\ \frac{1}{P_u} & , i = u \\ 0 & , \text{otherwise} \end{cases} \quad (A.16)$$
Now by using the above and Equation A.14, the following is complete:

\[
\frac{\partial^2 Q}{\partial P_u \partial P_v} = \frac{1}{P_u} \sum_{j=1}^{t} \frac{\partial z_{ij}}{\partial P_v} - \frac{1}{P_v} \sum_{j=1}^{t} \frac{\partial z_{ij}}{\partial P_v} \quad \text{(A.17)}
\]

A.7 Compute \( \frac{\partial^2 Q}{\partial P \partial \beta} \)

In application, we used the transpose of \( \frac{\partial^2 Q}{\partial \beta \partial P} \) which is theoretically the same by symmetry.

A.8 Compute \( \frac{\partial^2 Q}{\partial \beta \partial \hat{P}} \)

For this calculation, we need only consider

\[
Q_{VII} = \sum_{i=1}^{c} \sum_{j=1}^{t} \hat{z}_{ij} \log(1 - \pi_{ij}) \quad \text{(A.18)}
\]

The calculation for \( \frac{\partial \hat{z}_{ij}}{\partial \hat{P}_v} \) was previously specified in Equation A.14.

\[
\frac{\partial \log(1 - \pi_{ij})}{\partial \beta_u} = \frac{1}{1 - \pi_{ij}} \frac{\partial \pi_{ij}}{\partial \beta_u} = \frac{1}{1 - \pi_{ij}} \frac{\pi_{ij}(1 - \pi_{ij}) x_{(ij)_u}}{1 - \pi_{ij}} = -\pi_{ij} x_{(ij)_u} \quad \text{(A.19)}
\]

Where \( x_{(ij)_u} \) refers to a row specified by \((ij)\) in the model matrix and column \( u \) corresponding to our \( \beta_u \).

And so,

\[
\frac{\partial^2 Q}{\partial \beta_u \partial \hat{P}_v} = - \sum_{i=1}^{c} \sum_{j=1}^{t} \pi_{ij} x_{(ij)_u} \frac{\partial \hat{z}_{ij}}{\partial \hat{P}_v} \quad \text{(A.20)}
\]

Which is now a completely specified computation.
A.9 Compute $\frac{\partial z_{ij}}{\partial \beta_v}$

This partial differentiation is required for an upcoming result.

$$\frac{\partial z_{ij}}{\partial \beta_v} = \frac{\partial}{\partial \beta_v} \left( z_j \frac{\hat{P}_i(1 - \hat{\pi}_{ij})}{\sum_{k=1}^{c} \hat{P}_k(1 - \hat{\pi}_{kj})} \right)$$

$$= \frac{\partial}{\partial \beta_v} \left( z_j \frac{f_{ij}}{g_j} \right) \tag{A.21}$$

$z_j$ is constant with respect to $\hat{\beta}_v$. Also $f_{ij}$ and $g_j$ have been added as terse aliases for their respective terms.

$f_{ij}$ has partial derivative

$$\frac{\partial f_{ij}}{\partial \beta_v} = -\hat{P}_i \hat{\pi}_{ij}(1 - \hat{\pi}_{ij}) x_{(ij)v} \tag{A.22}$$

Where $x_{(ij)v}$ refers to the row in the model matrix described by $(ij)$ and the column described by $v$.

For $g_j$ this calculation is very similar to the one for $f_{ij}$ only with a summation.

$$\frac{\partial g_j}{\partial \beta_v} = -\sum_{k=1}^{c} \hat{P}_k \hat{\pi}_{kj} (1 - \hat{\pi}_{kj}) x_{(kj)v} \tag{A.23}$$

So finally by the quotient rule,

$$\frac{\partial z_{ij}}{\partial \beta_v} = z_j \frac{g_j \frac{\partial f_{ij}}{\partial \beta_v} - f_{ij} \frac{\partial g_j}{\partial \beta_v}}{g_j^2} \tag{A.24}$$

Which may or may not have a simplification. But nevertheless it is now fully computable.
A.10 Compute $\frac{\partial^2 Q}{\partial \beta \partial \hat{\beta}}$

We need only consider

$$Q_{VIII} = \sum_{i=1}^{c} \sum_{j=1}^{l} \hat{z}_{ij} \log(1 - \pi_{ij})$$  \hspace{1cm} (A.25)

Using Equation A.19 and Equation A.24, we can arrive at and calculate

$$\frac{\partial^2 Q}{\partial \beta_u \partial \hat{\beta}_v} = -\sum_{ij} \pi_{ij} x_{(ij)u} \frac{\partial \hat{z}_{ij}}{\partial \hat{\beta}_v}$$  \hspace{1cm} (A.26)