Abstract

Date sets with excess zeros are frequently analyzed in many disciplines. A common framework used to analyze such data is the zero-inflated (ZI) regression model, which assumes a mixture of a degenerate distribution which all realizations are zero with a non-degenerate distribution. The model has two linear predictors, one for the mixing probabilities and the other for the mean of the non-degenerate distribution. While there are many proposed $R^2$ measures for generalized linear models, there has not yet been one that fits in a ZI framework. We extend the pseudo-$R^2$ measure of Cameron and Windmeijer (1997) to include this model class. If the data do not have excess zeros, then our measure reduces to Cameron and Windmeijer’s statistic. We also propose adjusted $R^2$-like versions of these quantities to avoid inflation of these statistics due to the inclusion of irrelevant covariates in the model. Additionally, we propose marginal ZI (MZI) models for cross-sectional and longitudinal data with a focus on count data. In the ZI framework developed by Lambert (1992), the effects of the covariates are on the mixing probabilities and mean of the latent distribution, which may be undesired. We propose a model where the interpretation of the regression coefficients are the effects of the covariates on the overall mean. This model allows for more straightforward interpretations and inference. We discuss different methods of maximizing the
loglikelihoods, including the EM algorithm. The simulations examine properties of the maximum likelihood estimators as well as commonly used inference methods. Several examples are given for the proposed pseudo-$R^2$ and MZI models that illustrate their uses in practical data analysis.

*Keywords:* Zero Inflation; Exponential Dispersion Family; Generalized Linear Models; Marginalized Models; EM Algorithm; Adjusted $R^2$; Clustered Data; Random Effects
Topics in Zero-inflated Count Regression:
Coefficients of Determination and Marginal Models

by

Jacob William Martin

B.S., University of West Georgia, Georgia 2010
M.S., University of Georgia, Georgia 2011

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2016
# Contents

1 Introduction and Literature Review \hfill 1

2 $R^2$ measures for zero-inflated regression models for count data \hfill 7

- 2.1 Introduction \hfill 8
- 2.2 $R^2$ Measures \hfill 9
- 2.3 $R^2$ Measures for Poisson Regression \hfill 12
- 2.4 $R^2$ Measures for ZI Regression \hfill 14
- 2.5 Simulations \hfill 16
- 2.6 Examples \hfill 22
- 2.7 Discussion \hfill 26

3 Marginal zero-inflated regression models for count data \hfill 34

- 3.1 Introduction \hfill 35
- 3.2 Zero Inflated Regression \hfill 37
- 3.3 Marginal Regression Models \hfill 40
  - 3.3.1 Model Formulation \hfill 40
  - 3.3.2 Special Cases \hfill 42
  - 3.3.3 Model interpretability \hfill 46
- 3.4 Simulation \hfill 48
- 3.5 Examples \hfill 54
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Bar plot for the DMFT scores of children at follow-up in the study of Men-donça and Böhning (1994)</td>
<td>26</td>
</tr>
<tr>
<td>3.1</td>
<td>Histograms for the apple root data set by photoperiod</td>
<td>56</td>
</tr>
</tbody>
</table>
List of Tables

2.1 Mean of $R^2$ measures when $\phi = 1$ .......................................................... 19
2.2 Mean of $R^2$ measures when $\phi = 2$. .......................................................... 20
2.3 Mean of $R^2$ measures when $\phi = 3$. .......................................................... 21
2.4 Frequency table for number of boating trips ................................................. 23
2.5 Parameter estimates for the models for trips data set ................................. 23
2.6 $R^2$ and other goodness of fit criterion for boating trips example, measured in
   percent. .............................................................................................................. 24
2.7 Parameter estimates for the ZIP models for the tooth decay data set .......... 27
2.8 $R^2$ and other goodness of fit criterion for tooth decay example, measured in
   percent. .............................................................................................................. 27
3.1 Simulation results from the MZIP and MZINB models. PRMB: Percent relative
   median bias, SSD: Simulation standard deviation, MBSE: Model based standard
   error, MBCP: Model based coverage probability .............................................. 49
3.2 Results for the MZIB model. PRMB: Percent relative median bias, SSD: Simula-
   tion standard deviation, MBSE: Model based standard error, MBCP: Model based
   coverage probability .......................................................................................... 50
3.3 Comparing the coverage probabilities(width of interval) of the ZIP, ZINB, and
   ZIB models to their marginal equivalences ...................................................... 51
3.4 The power of a .05 significance level test for $H_{01}: \log(R) = 0$ and $H_{02}: \gamma_1 = \beta_1 = 0$. ................................. 53

3.5 Estimates and standard errors of the three proposed models for the apple root data. ......................... 55

3.6 $\hat{\mu}_{ij}$ for each block and treatment combination. ........................................ 59

4.1 Simulation results from the MZIP model. PRMB: Percent relative median bias, SSE: Simulation standard deviation, MBSE: Model based standard error. ................ 82

4.2 Ratio of the estimated mean and true mean ($\hat{R}/R_0$) ........................................ 84

4.3 Information criterion for the model shown with the largest interaction terms in the linear component for the marginal mean ........................................ 85

4.4 Wald tests for the different treatment effects on the marginal mean ............................... 87
Chapter 1

Introduction and Literature Review

Count data—both bounded and unbounded counts—regularly occur in a variety of disciplines. However, the standard modeling frameworks for such data, logistic regression and Poisson loglinear models, often fail to adequately account for the observed variability, a phenomenon known as overdispersion. In the common scenario that the data exhibit greater spread throughout their range (fat tails) than standard distributions would predict, a dispersion parameter can be helpful, either as a component of a more general two-parameter count distribution such as the negative binomial or beta-binomial distribution, or as an extra variance parameter in a model based only on moment assumptions and analyzed via quasilikelihood. However, in other cases overdispersion may result from a more specific distributional violation: an excess frequency of zero values. The dissertation will focus on topics for models that are fitted to data that falls into the second scenario, which we refer to as excess zero (EZ) data. Many research disciplines encounter EZ data for both cross-sectional data such as manufacturing (Lambert [1992], horticulture (Hall [2000]), entomology (Yeşilova et al. [2010]), fish and wildlife conservation (Maunder and Punt [2004]), and psychology (Catty et al. [2008]), and longitudinal data like health (Yau and Lee [2001], Albert et al. [2014]), horticulture (van Iersel et al. [2000]), and sociology (Buu et al. [2012]).
There are two common models fit to EZ data. The first is the hurdle regression model described by Mullahy (1986), which is a two part regression model with a linear predictor fit to the probability of an realization being a zero and the other linear predictor fitted to a distribution truncated at zero. The second commonly used model is the zero-inflated (ZI) regression model Lambert (1992), which is similar to the hurdle model, but instead of fitting a linear predictor to the probability of an outcome being zero, it instead models the probability that an observation belongs to a degenerate distribution that guarantees the outcome to be zero. Given that the observation is not from the degenerate distribution, the second linear predictor estimates the mean of the non-degenerate distribution. While the hurdle model is more flexible (it can be fit to zero-deflated data as well as EZ data), if the data are EZ, the ZI regression model has a more natural interpretation. This dissertation will focus on topics related to ZI regression for EZ count data. In particular, the main contributions of the dissertation are (1) to propose an $R^2$ like quantity for ZI regression; (2) to develop marginal zero-inflated regression models; and (3) to extend marginal ZI models to the clustered (e.g., longitudinal) data context.

The topic discussed in chapter 1 is a coefficient of determination for ZI regression models. There has been extensive research on pseudo- (and generalized) $R^2$ measures for logistic regression (Mittlböck and Schemper 1996; Tjur 2009) and some work for $R^2$-like quantities for Poisson and negative binomial regression (Mittlböck and Waldhör 2000; Heinzl and Mittlböck 2003). Despite the growing use of ZI regression models, there has not been a $R^2$ quantity proposed for this class of models. We propose a pseudo-$R^2$ quantity for ZI regression models for count data, denoted $R^2_{ZI}$, by extending the work of Cameron and Windmeijer (1997). This proposed measure is based on a ratio of deviances, giving it an interpretation of the information lost from the saturated model by the fitted model. Using this ratio requires specifications of null and saturated ZI regression models, concepts that have not been addressed in the literature until now. E.g., to define a suitable coefficient
of determination we addressed the open questions, what are the values of \( \pi_i \) and \( \lambda_i \) for the saturated model? Should the null model be an intercept only ZI model or Poisson model?

However, like many \( R^2 \) measures, our proposed \( R^2_{ZI} \) statistic increases as additional predictors are added, even if they are irrelevant. To counteract this potential unnecessary inflation, an adjustment is calculated, which is an extension of the work by Mittlböck and Waldhör (2000). This adjustment is derived from the asymptotic distribution of the difference between the null and fitted model with the assumption none of the added covariates are significant. Additionally, Heinzel and Mittlböck (2003) note that if the data are assumed Poisson but the sample variance is less than or in excess of the estimated model mean, then the adjustment recommended by Mittlböck and Waldhör is insufficient. To remedy this issue, we recommend that a ZI negative binomial (ZINB) regression model is used instead of a ZI Poisson (ZIP) model. Simulation results show, as expected, that \( R^2_{ZI} \) increases with unrelated predictors and when the data are overdispersed, the rate of inflation increases.

The proposed adjustments for unneeded predictors and overdispersion adequately temper the inflation in \( R^2_{ZI} \) while not overpenalizing when the model is correct or overdispersion is not present. Two examples included in the chapter discuss boating recreation in Texas and tooth decay in Brazilian children. The examples demonstrate the usefulness of such a measure for ZI models in different disciplines while illustrating that \( R^2_{ZI} \) and measures of fit will not always agree and thus \( R^2 \) measures should not be treated as such. The paper in chapter has been accepted for publication (Martin and Hall, 2016).

Chapter 2.7 discusses marginal ZI regression for cross-sectional count data. This chapter formulates and explores an alternative to the ZI regression model in which the marginal mean of the response variable is modelled directly. For a ZI model, the natural parameters of the model are the non-degenerate distribution’s mean and the mixing probability. However, both of these are means for latent random variables rather than the observed response. This makes the interpretation of a ZI regression model awkward, especially if interest focuses on
a covariate’s direct effect on the mean response, as it often does. To address this issue, we propose a marginal ZI (MZI) regression model that retains the mixture structure of classical ZI regression, but parameterizes it with a linear predictor that links directly to the marginal mean response, making model interpretation more straightforward. Two methods used to fit this class of models are discussed, using a straightforward function optimizer (such as NLMIXED in SAS or nlm in R), and the EM algorithm. Unlike standard ZI regression, the complete data log-likelihood does not decompose into two parts that can be maximized separately. Nevertheless, it is still possible to decompose the M step into two parts: first, the parameters in the mixing probability are estimated via a function optimizer; second, the regression parameters in the linear predictor for the mean can be calculated with a weighted generalized linear model (GLM). Details for fitting MZI regression models are given for three common count distributions, the MZI Poisson (MZIP), MZI negative binomial (MZINB), and MZI binomial (MZIB). We discuss when the ZI and MZI models are equivalent and when certain estimates of a ZI model have marginal interpretations.

Additional benefits of the MZI model compared to the ZI model is the ease of inference on the overall mean. Confidence intervals and hypothesis tests for the overall mean or ratio of means can be performed directly from the model estimates and their standard errors; in contrast, inference on the marginal mean using the ZI model is more difficult and less accurate as it often requires methods that have additional calculations, such as the delta method. Our simulation results not only show that maximum likelihood parameter estimates for the MZI model are exhibit small and decreasing empirical bias as the sample size increases, but that, in contrast to corresponding inferences based on ZI models, confidence intervals and hypothesis tests have correct coverage and size for means and mean ratios. To demonstrate the usefulness of the model class, two different examples are given. The first analyzes unbounded count data with the response being shoots from apple roots grown under different conditions, including two different lengths of photoperiod. For the shorter photoperiod, the
number of shoots appears to follow a negative binomial distribution, whereas for the longer period the response exhibits a large degree zero inflation. The second example analyzes different methods of pesticide application on survival probabilities for whiteflies on greenhouse plants. Since the response is a bounded count, the MZIB model is fitted to the data.

Lastly, chapter [4] extends the work in chapter [2.7] to clustered data. To account for within-subject correlation, Hall (2000) extended the ZI Poisson regression model to include a subject specific random effect in the model for the mean of the latent Poisson distribution. Hur et al. (2002) and Wang et al. (2002) add subject-specific random effects to the linear predictor for the mixing probabilities as well. Analogously, we extend these papers to formulate mixed MZI (MMZI) models for clustered data. We describe maximum likelihood estimation in this class of models via the EM algorithm and explore the properties of the ML estimators via simulation. We also highlight the advantage of the MMZI model relative to the traditional ZI mixed-effect model for marginalizing over the distribution of the random effects. The former class of models yields parameters that have both subject specific (SS) and population averaged (PA) interpretations, giving them a distinct advantage over the ZI mixed-effect models. Although ZI models can be marginalized to yield PA parameter estimates, this process is more complicated and computationally expensive than for MMZI models. The simulation section shows that maximum likelihood estimators (MLEs) of the fixed effects are accurate for modest sample sizes when ordinary Gaussian quadrature is used with relatively few quadrature points. Variance components are more difficult to estimate, however, requiring larger sample sizes and more quadrature points for accurate estimation. The example discusses the whitefly data from chapter [2.7] but analyzes the unbounded counts of immature flies after the pesticide application, where the same plants 54 plants are used across twelve weeks, inducing a dependency among repeated observations from the same plant.
Chapters 1-4 are written as stand-alone papers. As such, they have their own introductions, literature reviews, and conclusions. This introduction is brief to avoid redundancy. The final chapter is a conclusion that discusses the main contributions of the papers and ideas for future work.
Chapter 2

$R^2$ measures for zero-inflated regression models for count data

---

Abstract

Generalized linear models are often used to analyze discrete data. There are many proposed $R^2$ measures for this class of models. For loglinear models for count data, Cameron and Windmeijer (1997) developed an $R^2$-like measure based on a ratio of deviances. This quantity has since been adjusted to accommodate both overspecification and overdispersion. While these statistics are useful for Poisson and negative binomial regression models, count data often include many zeros, a phenomenon that is often handled via zero-inflated (ZI) regression models. Building on Cameron and Windmeijer’s work, we propose $R^2$ statistics for the ZI Poisson and ZI negative binomial regression contexts. We also propose adjusted $R^2$-like versions of these quantities to avoid inflation of these statistics due to the inclusion of irrelevant covariates in the model. The properties of the proposed measures of fit are examined via simulation, and their use is illustrated on two data sets involving counts with excess zeros.

Keywords: Adjusted $R^2$; Poisson regression; negative binomial regression; overdispersion; deviance; zero inflation

2.1 Introduction

When analyzing unbounded count data, a Poisson distribution is often used. Frequently, however, this framework fits the data poorly due to extra-Poisson variability, a phenomenon known as overdispersion. In some cases the data may exhibit more large and small values (fat tails) than Poisson data should. Such generalized overdispersion can be accounted for by allowing a dispersion parameter, as in the negative binomial distribution. However, in other cases overdispersion may result from a more specific distributional violation: an excess frequency of zero values. This paper will focus on data that falls into the second scenario, known as zero-inflated (ZI) data.
There has been increasing attention to the need for models that can account for such situations. Many research disciplines encounter ZI data, such as manufacturing (Lambert, 1992), horticulture (Hall, 2000), fish and wildlife conservation (Maunder and Punt, 2004), and psychology (Catty et al., 2008). The most popular model for ZI data is a mixture model in which excess zeros are hypothesised to come from a degenerate distribution that is responsible for the non-zero data as well as some of the zeros. Such a mixture is known as a ZI distribution or, when the effects of covariates are incorporated, a ZI regression model. In the latter case, dependence on covariates is built in through generalized linear model-like specifications for the count mean and mixture probability. Prominent examples of ZI regression models are the ZI Poisson (ZIP) model, ZI negative binomial (ZINB) model, and ZI binomial (ZIB) model.

Although ZI regression methodology has developed rapidly, one tool that is lacking is a measure of model fit, such as a coefficient of determination. While there is extensive research on pseudo- (and generalized) $R^2$ measures for logistic regression models (Mittlböck and Schemper, 1996; Tjur, 2009) and some work on $R^2$-like quantities for Poisson regression (Mittlböck and Waldhör, 2000; Heinzl and Mittlböck, 2003), there is no work for ZI regression. Since a variety of fields are now using these models, a tool that quantifies model fit on an easy-to-understand 0 to 1 scale would be helpful to practitioners.

### 2.2 $R^2$ Measures

When analyzing data with regression models, a coefficient of determination is often used to assess the goodness of fit. The most frequently used coefficient of determination for linear models, denoted $R^2_O$ (‘O’ for ordinary least squares, or OLS), is
\[ R^2_O = \frac{SSR}{SST} \equiv R^2_{mod} \]
\[ = 1 - \frac{SSE}{SST} \equiv R^2_{res} \]
\[ = \sum_{i} (y_i - \bar{y})(\hat{y}_i - \bar{y}) \]
\[ \equiv R^2_{cor} \]  

(2.1)

where \( \hat{y}_i \) is the fitted value of \( y_i \), \( \bar{y} = \frac{1}{n} \sum_{i} y_i \), \( SSR = \sum_{i} (\hat{y}_i - \bar{y})^2 \), \( SSE = \sum_{i} (y_i - \hat{y}_i)^2 \), and \( SST = \sum_{i} (y_i - \bar{y})^2 \). The statistic \( R^2_O \) has many desirable properties, such as lying in \([0, 1]\), being nondecreasing as predictors are added, and offering easy interpretation.

However, these properties do not necessarily hold when \( R^2_O \) is applied to nonlinear regression models such as generalized linear models (GLMs) and ZI regression models, which are extensions of GLMs. Generalized linear models assume an exponential dispersion family for the response variable \( y_i \) with mean \( \mu_i = g^{-1}(x'_i \beta) \), where \( g(.) \) is a link function, \( x_i \) is a vector of predictors for observation \( i \), and \( \beta \) is the parameter vector. We denote the loglikelihood by \( \ell(\beta; y) \), which is maximized with respect to \( \beta \) in order to find the maximum likelihood estimate (MLE), \( \hat{\beta} \).

There are many \( R^2 \)-like statistics proposed for nonlinear models. To provide an appropriate and applicable measure, Kvalseth (1985) recommended eight properties that all ‘good’ \( R^2 \) statistics should have. In addition to those mention previously, he recommended that such measures should be dimensionless, broadly applicable, directly comparable across different models, not be tied to a specific estimation method, and should weigh positive and negative residuals equally.

There are two common approaches for defining \( R^2 \)-like quantities outside of the linear model context: as a function of fitted and observed values, or based on a comparison of
objective functions (likelihoods, typically). These approaches coincide in $R^2_O$ for the linear model, but lead to various distinct statistics otherwise.

As examples of the first approach, any of the formulas for $R^2_{mod}$, $R^2_{res}$ and $R^2_{cor}$ can be applied in nonlinear model contexts (with minor modifications), but the resulting statistics are no longer equivalent and each choice has drawbacks. Tjur (2009) compares these three definitions for logistic regression. He notes that these measures differ in this context, but they are asymptotically equivalent and typically take similar values in applications. His simulation studies found them to have Monte Carlo means and standard deviations that were quite close. However, he notes that $R^2_{res}$ can fall below zero, and therefore recommended the other two statistics as better choices. Tjur also proposes a new measure, named the coefficient of discrimination, given by

$$D = \bar{\hat{\pi}}_1 - \bar{\hat{\pi}}_2 = \frac{\sum_i (y_i - \bar{y})(\hat{\pi}_i - \bar{y})}{\sum_i (y_i - \bar{y})^2},$$

where $\hat{\pi}_i$ is the estimated probability for observation $i$. The advantages of this statistic is that it falls within the unit interval, with $D = 0$ indicating complete lack of fit and $D = 1$ indicating perfect fit, and has an appealing interpretation for binary data. It can decrease when predictors are added; however, in simulations, he found this behavior to occur very rarely. For other measures defined in terms of observed and fitted quantities, see Mittlböck and Schemper (1996) and Menard (2000).

The other common approach for defining an $R^2$ statistic is to base it on comparisons of the likelihood function or, more generally, the objective function. For OLS models, minimized objective functions for the fitted and null models are $SSE$ and $SST$, respectively, and $R^2_O$ can easily be shown to be equivalent to $R^2_{res}$, which is one minus the ratio of these quantities. McFadden (1974) defines a pseudo-$R^2$, denoted $R^2_M$, similarly, using the ratio of the maximized loglikelihoods for the fitted and null regression models:
\[ R^2_M = 1 - \ell(\hat{\beta}; y) / \ell(\hat{\beta}_0; y), \]

where \( \hat{\beta}_0 \) is the estimated parameter under the null model, an intercept only model. It should be noted that this \( R^2 \) measure is described as a pseudo-\( R^2 \) because it does not reduce to \( R^2_O \) in the linear model case. Cox and Snell (1989) developed a \( R^2 \) that does simplify to \( R^2_O \) (that is, a generalized \( R^2 \)) that is also a ratio of loglikelihoods,

\[ R^2_{CS} = 1 - \exp \left[ -\frac{2}{n} (\ell(\hat{\beta}; y) - \ell(\hat{\beta}_0; y)) \right]. \]

This measure has a disadvantage of having a maximum at \( 1 - \exp \left[ -\frac{2}{n} (\ell(\hat{\beta}_0; y)) \right] \) which is less that one. Nagelkerke (1991) suggests that this can be remedied by multiplying it by the reciprocal of its maximum, but this involves some loss of interpretability. A third example is that of Cameron and Windmeijer (1997) who developed an \( R^2 \) statistic based on deviances. Their approach is described and extended to ZI regression models in the sections to follow.

### 2.3 \( R^2 \) Measures for Poisson Regression

Cameron and Windmeijer (1997) developed an \( R^2 \)-like measure based on the deviances of the proposed model and the null model for Poisson regression. In the model they consider, the response vector \( y = (y_1, ..., y_n)' \) has independent elements, where \( y_i \sim \text{Poisson}(\mu_i) \), \( \mu = (\mu_1, ..., \mu_n)' = \exp(X\beta) \) for a model matrix \( X \), and \( \beta \) is a \( k + 1 \)-dimensional regression parameter. Their statistic, denoted \( R^2_D \), is given by

\[
R^2_D = 1 - D(\hat{\mu}) / D(\hat{\mu}_0) = 1 - \frac{[\ell(y; y) - \ell(\hat{\mu}; y)]}{[\ell(y; y) - \ell(\hat{\mu}_0; y)]}, \tag{2.2}
\]
where $\ell(\mu; y)$ denotes the loglikelihood for $\mu$, $\hat{\mu} = \mu(\hat{\beta})$, and $\hat{\mu}_0$ is the mean of the model with no covariates ($\hat{\mu}_0 = \bar{y}$ in the loglinear Poisson model). This statistic has the properties identified by Kvålseth and can be interpreted as the reduction in deviance using the fitted model compared to the intercept (null) model. Under normality and identity link, $R^2_D$ becomes $R^2_O$ if the variance estimate from the fitted model is used in all three likelihoods, which classifies it as a generalized $R^2$. If different variance estimates are used, this no longer holds true and it is a pseudo-$R^2$.

As with many $R^2$ statistics, $R^2_D$ increases as irrelevant predictors are introduced into the model, especially for small sample sizes and situations in which the response mean is small. To correct this behavior, Mittlböck and Waldhöhr (2000) proposed a modification to $R^2_D$ similar in purpose to the adjusted $R^2$ statistics of linear regression. Under the assumption $H_0: \beta_1 = \ldots = \beta_k = 0$, $2(\ell(\hat{\mu}; y) - \ell(\hat{\mu}_0; y)) \sim \chi^2_k$; that is, when none of the predictors are associated with the response, the difference between the asymptotic means of the numerator and the denominator in the second term of (2.1) is $k/2$. Therefore, their adjusted $R^2$ is

$$R^2_{D,adj} = 1 - \frac{[\ell(y, y) - \ell(\hat{\mu}, y) + k/2]}{[\ell(y, y) - \ell(\hat{\mu}_0, y)]}.$$ 

This adjustment corrects the increase in $R^2_D$ well, but for small sample sizes and low values of $R^2$, $R^2_{D,adj}$ can fall below 0.

The Poisson distribution is often used when analyzing count data; however over- or underdispersion is commonly present; that is, it is often the case that the sample variance is less than or, more commonly, in excess of the estimated model mean. Often, a quasi-likelihood approach is taken to deal with this phenomenon, where a dispersion parameter is incorporated, allowing the variance of $y_i$ to be $\phi \mu_i$. Heinzl and Mittlböck (2003) report that in the presence of over- or underdispersion, the adjustment proposed previously is no
longer adequate. Therefore they suggest multiplying the adjustment by an estimate of $\phi$. This leads to their statistic,

$$R^2_{D,adj} = 1 - \left[ \ell(y, y) - \ell(\hat{\mu}, y) + \phi k/2 \right] / \left[ \ell(y, y) - \ell(\hat{\mu}_0, y) \right],$$

where $\hat{\phi}$ can be estimated either with the deviance divided by the residual degrees of freedom, denoted $\hat{\phi}_D$, or using the generalized Pearson statistic divided by the degrees of freedom, denoted $\hat{\phi}_P$. Simulation results of Heinzl & Mittlböck favor the use of $\hat{\phi}_P$.

### 2.4 $R^2$ Measures for ZI Regression

Lambert [1992] introduced ZIP regression to address zero inflation in unbounded count data. The ZIP regression model assumes the data are generated by a mixture of a distribution degenerate at 0 and a Poisson($\mu_i$) distribution, with mixing probabilities $\pi_i$ and $1 - \pi_i$, respectively. The probability mass function (pmf) for a ZIP distribution is

$$P(Y_i = k) = \begin{cases} 
\pi_i + (1 - \pi_i) \exp(-\mu_i) & \text{if } k = 0; \\
(1 - \pi_i) \exp(-\mu_i) \mu_i^k/k! & \text{if } k = 1, 2, 3, \ldots
\end{cases}$$

It follows that the ZIP regression loglikelihood is

$$\ell(\gamma, \beta; y) = \sum_{i=1}^{n} \left\{ u_i \log(\pi_i(\gamma) + (1 - \pi_i(\gamma)) \exp(-\mu_i(\beta))) \\
+ (1 - u_i) \left[ \log(1 - \pi_i(\gamma)) - \mu_i(\beta) + y_i \log \mu_i(\beta) \right] \right\}$$

where $u_i = I(y_i = 0)$ and

$$\pi_i(\gamma) = \exp(g_i^{\prime}(\gamma))/(1 + \exp(g_i^{\prime}(\gamma))) \quad (2.3)$$

$$\mu_i(\beta) = \exp(x_i^{\prime}\beta)$$
To define the saturated ZIP model, we set $\pi_i = u_i$ and $\mu_i = y_i$. Plugging these values into the ZIP loglikelihood, we recover the saturated Poisson regression loglikelihood function, which matches the intuition that the saturated version of the ZIP and Poisson regression models should fit the data equally well. In addition, we take the null model to be the simplest non-degenerate model in the ZIP class: a Poisson regression model with only an intercept in the linear predictor, which leads to $\mu_i = \bar{y}$ for all $i$. Note that this choice of null model is also appealing because it leads to a $R^2$ measures that reduces to the $R^2_D$ statistic proposed by Cameron and Windmeijer (1997) when $\pi_i = 0$ for all $i$. Let $\pi_i(\hat{\gamma}) = \hat{\pi}_i$ and $\mu_i(\hat{\beta}) = \hat{\mu}_i$, then $R^2_D$ for a ZIP regression model is

$$R^2_{ZIP} = 1 - \frac{\sum_{i=1}^{n} \{-u_i \log((1-\hat{\pi}_i) \exp(\hat{\mu}_i)) + (1-u_i)[y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i) - \log(1-\hat{\pi}_i)]\}}{\sum_{i=1}^{n} \{y_i \log(y_i/\bar{y}) - (y_i - \bar{y})\}}.$$

Like other $R^2$ measures, this statistic tends to increase with the addition of irrelevant covariates to the model, so an adjustment is needed. Following the work of Mittlböck and Waldhör (2000), the adjustment is derived from the hypothesis $H_0 : \gamma_1 = \ldots = \gamma_p = \beta_1 = \ldots = \beta_k = 0, \gamma_0 = -\infty$. Because this hypothesis sets $\gamma_0$ on a boundary value, $2(\ell(\hat{\pi}, \hat{\mu}; y) - \ell(0, \bar{y}; y))$ is not asymptotically $\chi^2_{p+k+1}$. To determine the correct adjustment, the difference in loglikelihoods can be written as the sum of two differences, $2(\ell(\hat{\pi}, \hat{\mu}; y) - \ell(\hat{\pi}, \hat{\mu}; y))$ and $2(\ell(\hat{\pi}, \hat{\mu}; y) - \ell(0, \bar{y}; y))$. The first difference is asymptotically distributioned $\chi^2_{p+k}$ and the second is asymptotically distributed as an even mixture of $\chi^2_1$ and $\chi^2_0$; therefore the adjustment is $p + k + \frac{1}{2}$.

As for count data, ZI count data are often overdispersed. Therefore we consider $R^2$ measures for zero inflated negative binomial (ZINB) regression models as well. In ZINB
regression we assume

\[ Y_i \sim \begin{cases} 
0 & \text{with probability } \pi_i; \\
NB(\mu_i, \alpha) & \text{with probability } 1 - \pi_i.
\end{cases} \]

where \( NB(\mu_i, \alpha) \) denotes a negative binomial distribution with mean \( \mu_i \) and dispersion parameter \( \alpha \). As in ZIP regression, loglinear and logistic models are assumed for \( \mu_i \) and \( \pi_i \), relating these quantities to covariates as in 2.3. Note that according to the ZINB model,

\[ E(Y_i) = (1 - \pi_i)\mu_i \quad \text{and} \quad \text{var}(Y_i) = (1 - \pi_i)\mu_i(1 + \mu_i(\pi_i + 1/\alpha)). \]

Let the loglikelihood of a ZINB regression model be denoted by \( \ell_1(\pi, \mu, \alpha; y) \) and the loglikelihood of an ordinary negative binomial as \( \ell_2(\mu, \alpha; y) \), then the coefficient of determination for ZINB regression is

\[ R^2_{ZINB} = 1 - \frac{\ell_1(u, y; \hat{\alpha}) - \ell_1(\hat{\pi}, \hat{\mu}, \hat{\alpha}; y)}{\ell_1(\hat{u}, y; \hat{\alpha}) - \ell_2(\bar{y}; \hat{\alpha}; y)}, \]

where \( \hat{\alpha} \) is the estimate of the fitted model. The adjustment is calculated similarly as the adjustment for \( R^2_{ZIP} \), except that \( \alpha \) is fixed at \( \hat{\alpha} \) when evaluating \( \ell_2 \), so the adjustment becomes \( k + p + 1.5 \).

### 2.5 Simulations

A simulation study was performed to determine the properties of the different \( R^2 \) measures. Datasets with sample size \( n \) were generated, involving a response following either a ZIP or ZINB regression model with specifications as follows: \( \logit(\pi) = X\gamma \), \( \log(\mu) = X\beta \). Here \( X \) is the design matrix for a main-effects only two-factor balanced model; that is,
\[
X = \begin{bmatrix}
1 & 1 & 1 \\
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & 0 & 0
\end{bmatrix} \otimes j_{n/4}
\]

where \( j_d \) is a \( d \)-dimensional vector of ones. For the case of no overdispersion, the data were generated from a ZIP model and when the data are overdispersed, a ZINB model was used with \( \alpha \) values chosen so that \( \phi = 1 + \bar{\mu}/\alpha = 2, 3 \).

In addition to varying \( \phi \), the degree of extra-Poisson variability in the non-degenerate component of the model, we also considered four sample sizes, \( n = 64, 128, 256 \) and 512; and four ‘true’ values of \( R^2_D \), denoted \( R^2_0 \): 0.00, 0.25, 0.50, and 0.75. To generate data with a given value of \( R^2_0 \), we follow the procedure used by Mittlböck and Waldhör (2000) in which \( R^2_0 \) for a given model is assumed to be the value of \( R^2_D \) obtained from an enormous data set generated from that model. In particular, we fixed \( \beta_0 = \log(2) \) and \( \gamma_0 = \text{logit}(0.2) \) and then calibrated the values of the other parameters so that the \( R^2_D \) value calculated from a correctly specified model for a data set of size \( n = 2^{21} \) generated under those parameter values was equal to each choice of \( R^2_0 \). Then from each sample data set, four models of increasing complexity were fit, where model \( k \) has the form: \( \text{logit}(\pi_i) = \gamma_0 + \sum_{j=1}^{k} \gamma_{ij} x_{ij}, \) \( k = 1, 2, 3, 4 \), with the linear predictor for \( \log(\mu_i) \) taking the same form. Here \( x_{i1} \) and \( x_{i2} \) are elements of the second and third columns of \( X \), the design matrix of the data generating model, and \( x_{i3} \) and \( x_{i4} \) are two additional dichotomous covariates, arranged as in a balanced factorial design.

**Results**

For data with no overdispersion, the results are displayed in Table 2.1. For this case, ZIP and ZINB models give very similar parameter estimates provided that the models are not underspecified, so we expect similar values of \( R^2_{ZIP} \) and \( R^2_{ZINB} \) and similar values of \( R^2_{ZIP,adj} \) and \( R^2_{ZINB,adj} \) for \( k > 1 \). Indeed, this is the case. In addition, while \( R^2_{ZIP} \) and \( R^2_{ZINB} \) exhibit
the usual behavior of increasing with model complexity, the adjusted $R^2$ statistics effectively compensate for the inclusion of unrelated covariates in the model. All $R^2$ values appear to be unbiased under the correctly specified model ($k = 2$) and the adjusted statistics remain nearly constant as $k$ grows. Although the adjustment in $R^2_{ZIP,\text{adj}}$ and $R^2_{ZINB,\text{adj}}$ results in slightly negative values for $R^2_0 = 0.00$, this effect diminishes with $n$, indicating that the adjustment, which is based on an asymptotic result, tends to improve with the sample size. This phenomenon can also be seen in the degree to which the adjustment stays constant as $k$ grows beyond 2, which improves with $n$ for all values of $R^2_0$. Although potential for negative values of $R^2_{ZIP,\text{adj}}$ and $R^2_{ZINB,\text{adj}}$ is somewhat undesirable, such a result occurs in models with essentially no predictive power, and replacing negative values by 0, seems a simple and appropriate convention.

The results for the data with overdispersion are displayed in Tables 2.2 and 2.3. The results show that there is a large, positive bias in $R^2_{ZIP}$ that increases as $\phi$ changes from 2 to 3. This bias is not present in ZINB regression model as there is little difference in $R^2_0$ and $R^2_{ZINB}$ for $k = 2$ (there is a notable decrease when the model is underspecified that grows as $\phi$ increases). Additionally, there is little difference in the averages when $\phi$ increases for $k > 1$, indicating that $R^2_{ZINB}$ appropriately accounts for potential overdispersion. However, as $k$ increases, $R^2_{ZINB,\text{adj}} > R^2_0$, but this discrepancy is small, especially for larger values of $R^2_0$ and decreases quickly as $n$ grows. The results from this and the previous paragraph indicate that when overdispersion is not present, there is little difference in the parameter estimates for a ZIP and ZINB model, which is also true for the values of $R^2_{ZIP}$ and $R^2_{ZINB}$. When overdispersion is present, $R^2_{ZIP}$ does not perform well, but this is but one of a host of problems that can be expected under this sort of model misspecification. Overdispersion is common relative to both Poisson regression models and ZIP models, and modifying the model to account for it, when present, is crucial for valid statistical inference.
<table>
<thead>
<tr>
<th>size</th>
<th>measure</th>
<th>$R^2 = 0.00$</th>
<th>$R^2 = 0.25$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>64</td>
<td>$R^2_{ZIP}$</td>
<td>0.023</td>
<td>0.047</td>
</tr>
<tr>
<td></td>
<td>$R^2_{ZIP,adj}$</td>
<td>-0.012</td>
<td>-0.017</td>
</tr>
<tr>
<td></td>
<td>$R^2_{ZINB}$</td>
<td>0.022</td>
<td>0.046</td>
</tr>
<tr>
<td></td>
<td>$R^2_{ZINB,adj}$</td>
<td>-0.014</td>
<td>-0.019</td>
</tr>
<tr>
<td>128</td>
<td>$R^2_{ZIP}$</td>
<td>0.012</td>
<td>0.024</td>
</tr>
<tr>
<td></td>
<td>$R^2_{ZIP,adj}$</td>
<td>-0.005</td>
<td>-0.007</td>
</tr>
<tr>
<td></td>
<td>$R^2_{ZINB}$</td>
<td>0.011</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td>$R^2_{ZINB,adj}$</td>
<td>-0.007</td>
<td>-0.008</td>
</tr>
<tr>
<td>256</td>
<td>$R^2_{ZIP}$</td>
<td>0.007</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>$R^2_{ZIP,adj}$</td>
<td>-0.002</td>
<td>-0.002</td>
</tr>
<tr>
<td></td>
<td>$R^2_{ZINB}$</td>
<td>0.006</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>$R^2_{ZINB,adj}$</td>
<td>-0.002</td>
<td>-0.003</td>
</tr>
<tr>
<td>512</td>
<td>$R^2_{ZIP}$</td>
<td>0.003</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>$R^2_{ZIP,adj}$</td>
<td>-0.001</td>
<td>-0.001</td>
</tr>
<tr>
<td></td>
<td>$R^2_{ZINB}$</td>
<td>0.003</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>$R^2_{ZINB,adj}$</td>
<td>-0.001</td>
<td>-0.001</td>
</tr>
</tbody>
</table>

Table 2.1: Mean of $R^2$ measures when $\phi = 1$. 
<table>
<thead>
<tr>
<th>size</th>
<th>measure</th>
<th>$R^2_0 = 0.00$</th>
<th>$R^2_0 = 0.25$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 2 3 4</td>
<td>1 2 3 4</td>
</tr>
<tr>
<td>64</td>
<td>$R_{ZIP}$</td>
<td>0.096 0.120 0.146 0.175</td>
<td>0.318 0.382 0.401 0.422</td>
</tr>
<tr>
<td></td>
<td>$R_{ZIP,adj}$</td>
<td>0.076 0.084 0.094 0.106</td>
<td>0.302 0.354 0.361 0.369</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB}$</td>
<td>0.042 0.073 0.107 0.145</td>
<td>0.217 0.320 0.348 0.378</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB,adj}$</td>
<td>-0.001 0.009 0.022 0.040</td>
<td>0.182 0.273 0.285 0.300</td>
</tr>
<tr>
<td>128</td>
<td>$R_{ZIP}$</td>
<td>0.086 0.097 0.109 0.121</td>
<td>0.316 0.370 0.380 0.389</td>
</tr>
<tr>
<td></td>
<td>$R_{ZIP,adj}$</td>
<td>0.076 0.079 0.083 0.087</td>
<td>0.309 0.357 0.360 0.364</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB}$</td>
<td>0.024 0.039 0.053 0.075</td>
<td>0.194 0.287 0.301 0.316</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB,adj}$</td>
<td>0.002 0.005 0.012 0.019</td>
<td>0.176 0.262 0.268 0.274</td>
</tr>
<tr>
<td>256</td>
<td>$R_{ZIP}$</td>
<td>0.080 0.086 0.091 0.097</td>
<td>0.315 0.364 0.368 0.373</td>
</tr>
<tr>
<td></td>
<td>$R_{ZIP,adj}$</td>
<td>0.075 0.077 0.079 0.080</td>
<td>0.311 0.358 0.359 0.361</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB}$</td>
<td>0.012 0.020 0.029 0.038</td>
<td>0.178 0.265 0.272 0.279</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB,adj}$</td>
<td>0.001 0.003 0.005 0.008</td>
<td>0.168 0.253 0.255 0.257</td>
</tr>
<tr>
<td>512</td>
<td>$R_{ZIP}$</td>
<td>0.077 0.079 0.082 0.085</td>
<td>0.317 0.364 0.367 0.369</td>
</tr>
<tr>
<td></td>
<td>$R_{ZIP,adj}$</td>
<td>0.074 0.075 0.079 0.080</td>
<td>0.316 0.361 0.362 0.363</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB}$</td>
<td>0.006 0.010 0.014 0.019</td>
<td>0.173 0.260 0.264 0.267</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB,adj}$</td>
<td>0.001 0.001 0.002 0.004</td>
<td>0.168 0.254 0.255 0.256</td>
</tr>
<tr>
<td></td>
<td>$R^2_0 = 0.50$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$R^2_0 = 0.75$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 2 3 4</td>
<td>1 2 3 4</td>
</tr>
<tr>
<td>64</td>
<td>$R_{ZIP}$</td>
<td>0.406 0.573 0.579 0.586</td>
<td>0.564 0.787 0.796 0.806</td>
</tr>
<tr>
<td></td>
<td>$R_{ZIP,adj}$</td>
<td>0.402 0.565 0.568 0.571</td>
<td>0.550 0.778 0.784 0.790</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB}$</td>
<td>0.263 0.520 0.529 0.538</td>
<td>0.443 0.774 0.784 0.796</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB,adj}$</td>
<td>0.246 0.504 0.507 0.510</td>
<td>0.417 0.756 0.761 0.767</td>
</tr>
<tr>
<td>128</td>
<td>$R_{ZIP}$</td>
<td>0.414 0.583 0.597 0.612</td>
<td>0.549 0.780 0.784 0.789</td>
</tr>
<tr>
<td></td>
<td>$R_{ZIP,adj}$</td>
<td>0.405 0.567 0.574 0.582</td>
<td>0.547 0.775 0.778 0.781</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB}$</td>
<td>0.285 0.540 0.558 0.579</td>
<td>0.428 0.762 0.767 0.772</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB,adj}$</td>
<td>0.253 0.508 0.515 0.525</td>
<td>0.414 0.753 0.755 0.757</td>
</tr>
<tr>
<td>256</td>
<td>$R_{ZIP}$</td>
<td>0.406 0.571 0.574 0.578</td>
<td>0.548 0.776 0.779 0.781</td>
</tr>
<tr>
<td></td>
<td>$R_{ZIP,adj}$</td>
<td>0.404 0.567 0.569 0.570</td>
<td>0.547 0.774 0.776 0.777</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB}$</td>
<td>0.253 0.513 0.517 0.521</td>
<td>0.420 0.756 0.758 0.760</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB,adj}$</td>
<td>0.244 0.505 0.506 0.507</td>
<td>0.413 0.751 0.752 0.753</td>
</tr>
<tr>
<td>512</td>
<td>$R_{ZIP}$</td>
<td>0.404 0.567 0.568 0.570</td>
<td>0.546 0.775 0.776 0.778</td>
</tr>
<tr>
<td></td>
<td>$R_{ZIP,adj}$</td>
<td>0.403 0.565 0.566 0.566</td>
<td>0.546 0.774 0.775 0.776</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB}$</td>
<td>0.245 0.505 0.507 0.509</td>
<td>0.415 0.754 0.755 0.756</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB,adj}$</td>
<td>0.240 0.501 0.501 0.502</td>
<td>0.411 0.752 0.752 0.753</td>
</tr>
</tbody>
</table>

Table 2.2: Mean of $R^2$ measures when $\phi = 2$. 
<table>
<thead>
<tr>
<th>size</th>
<th>measure</th>
<th>$R^2_0 = 0.00$</th>
<th></th>
<th>$R^2_0 = 0.25$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>64</td>
<td>$R_{ZIP}$</td>
<td>0.170</td>
<td>0.196</td>
<td>0.217</td>
<td>0.240</td>
</tr>
<tr>
<td></td>
<td>$R_{ZIP,adj}$</td>
<td>0.155</td>
<td>0.171</td>
<td>0.187</td>
<td>0.200</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB}$</td>
<td>0.038</td>
<td>0.009</td>
<td>0.126</td>
<td>0.165</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB,adj}$</td>
<td>0.016</td>
<td>0.027</td>
<td>0.042</td>
<td>0.052</td>
</tr>
<tr>
<td>128</td>
<td>$R_{ZIP}$</td>
<td>0.164</td>
<td>0.175</td>
<td>0.185</td>
<td>0.195</td>
</tr>
<tr>
<td></td>
<td>$R_{ZIP,adj}$</td>
<td>0.156</td>
<td>0.161</td>
<td>0.166</td>
<td>0.170</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB}$</td>
<td>0.029</td>
<td>0.047</td>
<td>0.065</td>
<td>0.084</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB,adj}$</td>
<td>0.006</td>
<td>0.013</td>
<td>0.020</td>
<td>0.027</td>
</tr>
<tr>
<td>256</td>
<td>$R_{ZIP}$</td>
<td>0.161</td>
<td>0.166</td>
<td>0.171</td>
<td>0.176</td>
</tr>
<tr>
<td></td>
<td>$R_{ZIP,adj}$</td>
<td>0.157</td>
<td>0.160</td>
<td>0.162</td>
<td>0.164</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB}$</td>
<td>0.016</td>
<td>0.025</td>
<td>0.033</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB,adj}$</td>
<td>0.005</td>
<td>0.007</td>
<td>0.009</td>
<td>0.012</td>
</tr>
<tr>
<td>512</td>
<td>$R_{ZIP}$</td>
<td>0.158</td>
<td>0.160</td>
<td>0.163</td>
<td>0.166</td>
</tr>
<tr>
<td></td>
<td>$R_{ZIP,adj}$</td>
<td>0.156</td>
<td>0.157</td>
<td>0.158</td>
<td>0.160</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB}$</td>
<td>0.007</td>
<td>0.012</td>
<td>0.016</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td>$R_{ZINB,adj}$</td>
<td>0.001</td>
<td>0.002</td>
<td>0.004</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>$R^2_0 = 0.50$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$R^2_0 = 0.75$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3: Mean of $R^2$ measures when $\phi = 3$. 
2.6 Examples

Boating trips example

Cameron and Trivedi (1998) present data from a survey conducted by Seller et al. (1985) of 2000 registered leisure boat owners in eastern Texas. The subset analyzed here contains 659 self reports of the number of boating trips the respondent made in 1980, along with related covariate information. The descriptive statistics and frequency distribution of the response and the predictors we consider are presented by Cameron and Trivedi (1998, p. 209). The frequency distribution for the response shows a high number of zeroes (63.3%) as well as a long right tail, indicating that the data may be zero-inflated and overdispersed. Additionally, the distribution shows some evidence of digit preference, with multiples of 5 and some even values occurring especially often. This phenomenon is common in self-reported counts and can contribute to over- or under-dispersion relative to a Poisson or ZIP distribution.

In order to test if there is evidence of zero inflation, a Poisson loglinear model with all seven predictors was fit along with the corresponding ZIP model with constant mixture probability, that is, \( \logit(\pi_i) = \gamma_0 \). The loglikelihood ratio test has a test statistic of 196.5, which strongly indicates that the ZIP model is a better fit to the data. Three nested ZIP models were considered, differing in the specification of the linear predictor for the Poisson component: Model (1) \( so + ski \), Model (2) \( so + ski + c1 + c3 + c4 \), and Model (3) \( so + ski + c1 + c3 + c4 + fc3 + i \). Following Cameron and Trivedi’s analysis, only the covariate \( so \) is included in the logistic component. The resulting parameter estimates, standard errors, and p-values are given in Table 2.5, where regression parameters for the mixing probabilities are denoted with the subscript \( \gamma \). Since the data are skewed right, ZINB models with the same combination of predictors are also reported.

For the ZIP model, neither \( so \) nor \( c1 \) are statistically significant for the three models considered, while all other predictors are significant at level 0.05. For the ZINB model,
Table 2.4: Frequency table for number of boating trips.

<table>
<thead>
<tr>
<th>Number of Trips</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>417</td>
</tr>
<tr>
<td>1</td>
<td>68</td>
</tr>
<tr>
<td>2</td>
<td>38</td>
</tr>
<tr>
<td>3</td>
<td>34</td>
</tr>
<tr>
<td>4</td>
<td>17</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of Trips</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td>16</td>
<td>13</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
</tr>
<tr>
<td>26</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>3</td>
</tr>
<tr>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td>88</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2.5: Parameter estimates for the models for trips data set.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>ZIP</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est SE p-val</td>
<td>Est SE p-val</td>
<td>Est SE p-val</td>
<td>Est SE p-val</td>
</tr>
<tr>
<td>Intercept</td>
<td>1.711 0.081 0.000</td>
<td>1.902 0.092 0.000</td>
<td>2.097 0.111 0.000</td>
<td></td>
</tr>
<tr>
<td>so</td>
<td>0.009 0.022 0.678</td>
<td>0.021 0.023 0.369</td>
<td>0.034 0.024 0.160</td>
<td></td>
</tr>
<tr>
<td>ski</td>
<td>0.135 0.053 0.011</td>
<td>0.419 0.057 0.000</td>
<td>0.471 0.058 0.000</td>
<td></td>
</tr>
<tr>
<td>c1</td>
<td>0.002 0.003 0.521</td>
<td>0.002 0.004 0.529</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c3</td>
<td>-0.040 0.002 0.000</td>
<td>-0.038 0.002 0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c4</td>
<td>0.027 0.003 0.000</td>
<td>0.025 0.003 0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>-0.098 0.020 0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fc3</td>
<td>0.610 0.079 0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept, so</td>
<td>2.973 0.223 0.000</td>
<td>3.072 0.247 0.000</td>
<td>3.111 0.259 0.000</td>
<td></td>
</tr>
<tr>
<td>so, fc3</td>
<td>-1.546 0.111 0.000</td>
<td>-1.795 0.159 0.000</td>
<td>-1.898 0.196 0.000</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predictor</th>
<th>ZINB</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est SE p-val</td>
<td>Est SE p-val</td>
<td>Est SE p-val</td>
<td>Est SE p-val</td>
</tr>
<tr>
<td>Intercept</td>
<td>1.350 0.239 0.000</td>
<td>0.878 0.215 0.000</td>
<td>1.088 0.257 0.000</td>
<td></td>
</tr>
<tr>
<td>so</td>
<td>0.072 0.066 0.279</td>
<td>0.176 0.054 0.001</td>
<td>0.169 0.053 0.001</td>
<td></td>
</tr>
<tr>
<td>ski</td>
<td>0.128 0.152 0.398</td>
<td>0.492 0.134 0.000</td>
<td>0.501 0.135 0.000</td>
<td></td>
</tr>
<tr>
<td>c1</td>
<td>0.044 0.015 0.003</td>
<td>0.040 0.015 0.005</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c3</td>
<td>-0.072 0.008 0.000</td>
<td>-0.066 0.008 0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c4</td>
<td>0.022 0.010 0.031</td>
<td>0.021 0.010 0.044</td>
<td></td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>-0.067 0.044 0.125</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fc3</td>
<td>0.542 0.283 0.055</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>log(α)</td>
<td>-0.339 0.097 0.001</td>
<td>0.159 0.112 0.154</td>
<td>0.021 0.010 0.044</td>
<td></td>
</tr>
<tr>
<td>Intercept, so</td>
<td>4.923 0.713 0.000</td>
<td>4.628 0.719 0.000</td>
<td>4.651 0.718 0.000</td>
<td></td>
</tr>
<tr>
<td>so, fc3</td>
<td>-7.269 1.602 0.000</td>
<td>-8.275 3.863 0.032</td>
<td>-8.308 3.986 0.037</td>
<td></td>
</tr>
</tbody>
</table>

neither so nor ski are significant for Model 1, but become significant once the additional predictors are added in Model 2. Neither i nor fc3 are statistically significant in Model 3 at the .05 significance level, although fc3 is close at p-value=.055.

The $R^2$ measures discussed in the previous section as well as $-2\ell$ and AIC for each model are given in Table 2.6. For the ZIP model, Model 1 has an $R^2_D$ of 54.1, indicating that
there is about a 54% decrease in the deviance using Model 1 rather than the null model. Both $R_D^2$ and $R_{D,adj}^2$ increase for Models 2 and 3, indicating that the predictors reduce the deviance by an amount greater than expected if the added predictors were not related to the response. The AIC for the models agree with the conclusion reached using $R_{D,adj}^2$, as it decreases as the models become more complex. There is a large decrease in both $-2\ell$ and the AIC when the model changes from ZIP to ZINB, indicating that the latter model class fits the data better. While $R_{D,adj}^2$ ‘agrees’ with the AIC, it is not recommended as a model selection criterion across different model classes, only as a measure of model fit. When the nonsignificant predictors in Model 3 are included, $R_D^2$ and $R_{D,adj}^2$ increase by only .03 and .02, respectively, indicating that their inclusion does not contribute much to the reduction of the deviance.

<table>
<thead>
<tr>
<th></th>
<th>ZIP</th>
<th></th>
<th></th>
<th>ZIP</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>$-2\ell$</td>
<td>AIC</td>
<td>$R_D^2$</td>
<td>$R_{D,adj}^2$</td>
<td>$-2\ell$</td>
<td>AIC</td>
</tr>
<tr>
<td>Model 1</td>
<td>2977.2</td>
<td>2987.2</td>
<td>54.1</td>
<td>54.1</td>
<td>1572.8</td>
<td>1584.8</td>
</tr>
<tr>
<td>Model 2</td>
<td>2439.9</td>
<td>2455.9</td>
<td>65.2</td>
<td>65.1</td>
<td>1450.8</td>
<td>1468.8</td>
</tr>
<tr>
<td>Model 3</td>
<td>2361.8</td>
<td>2381.8</td>
<td>66.8</td>
<td>66.7</td>
<td>1444.6</td>
<td>1466.6</td>
</tr>
</tbody>
</table>

Table 2.6: $R^2$ and other goodness of fit criterion for boating trips example, measured in percent.

**Tooth decay example**

The decayed, missing, and filled teeth (DMFT) index is a measure of a person’s overall dental health that is widely used in dental epidemiology. It is a count of the number of teeth an individual has that meets one of the namesake conditions. Böhning et al. (1999) analyzed data from a prospective study of school children to determine the effect of different dental hygiene methods. The children, who were seven years old when the survey began,
were recruited from an urban area of Belo Horizonte, Brazil. The five different treatments along with a control were randomly assigned to six different schools in the area. The DMFT scores from the eight deciduous molars at baseline and two years later for each of the 797 children were recorded, along with each child’s gender, ethnicity, and school. The data and definitions for the variables can be found at www.blackwellpublishing.com/rss.

A bar plot for the DMFT scores after two years, given in Figure 2.1, shows a spike at zero, which is typical in this type of study. Due to the large zero frequency, a Poisson distribution fits poorly while a ZIP model approximates the data well (Böhminger et al., 1999). In order to analyze the effect of the treatment for each school, three different models were analyzed: Model (1) $\log(DMFT1)$, Model (2) $\log(DMFT1) + school$, and Model (3) $\log(DMFT1) + school + gender + ethnic$. The term $\log(DMFT1)$ ($ldmft1$) is included to account for baseline differences in DMFT among the subjects. It is also the only predictor used in the logistic component of the model because this choice yields the lowest AIC for each specification of the linear predictor for the Poisson mean. Additionally, the corresponding ZINB models were fit, however, the results are not included due to the estimates being identical to those of the ZIP model, as there is no evidence of overdispersion ($\alpha > 400$ in each case).

The parameter estimates for the three models are found in Table 2.7. The strongest indicator of a child’s DMFT score after two years is their DMFT score at baseline, which has a p-value $< .001$ in all three models. For models (2) and (3), the treatment for oral hygiene (school 5) had the largest effect on the mean compared to the control group. Both ricebran (school 3) and the combined interventions (school 2) have p-values between .05 and .10, which is marginally suggestive that these interventions had real effects on the log mean. The addition of gender and ethnicity in Model (3) shows some (non-significant) evidence of a gender effect, but very little suggestion of ethnicity differences. These are reflected in the statistics in Table 2.8. There is no difference in the loglikelihoods or $R^2_D$ between the ZIP and ZINB models, as expected since there is no evidence of overdispersion. The AIC
indicates that Model (2) is the least complex model that still fits the data well among the three candidates, while $R_{D,adj}^2$ sees little increase between the three models. This indicates that, as in classical linear models, $R_{D,adj}^2$ and model selection criteria like AIC will not always agree.

2.7 Discussion

In this paper, we have generalized Cameron and Windmeijer’s $R_D^2$ to ZI regression models for count data that have excess zeros. The proposed $R^2$-measures are calculated from the saturated, fitted, and null models. For ZI models, the null is assumed to be the intercept only Poisson model with $\pi = 0$ and the saturated model is equivalent to the saturated model for the corresponding non-ZI regression model. To adjust for the tendency for $R_{D}^2$
Table 2.7: Parameter estimates for the ZIP models for the tooth decay data set.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est</td>
<td>SE</td>
<td>p-val</td>
</tr>
<tr>
<td>Intercept</td>
<td>1.160</td>
<td>0.034</td>
<td>0.000</td>
</tr>
<tr>
<td>ldfmt1</td>
<td>0.310</td>
<td>0.029</td>
<td>0.000</td>
</tr>
<tr>
<td>school1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>school2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>school3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>school4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>school5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gender1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ethnic1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ethnic2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept,γ</td>
<td>-1.146</td>
<td>0.118</td>
<td>0.000</td>
</tr>
<tr>
<td>ldfmt1,γ</td>
<td>-1.716</td>
<td>0.157</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 2.8: $R^2$ and other goodness of fit criterion for tooth decay example, measured in percent.

<table>
<thead>
<tr>
<th>Model</th>
<th>$-2\ell$</th>
<th>AIC</th>
<th>$R^2_D$</th>
<th>$R^2_{D,adj}$</th>
<th>$-2\ell$</th>
<th>AIC</th>
<th>$R^2_D$</th>
<th>$R^2_{D,adj}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>3177.4</td>
<td>3185.4</td>
<td>40.7</td>
<td>40.6</td>
<td>3177.4</td>
<td>3187.4</td>
<td>40.7</td>
<td>40.5</td>
</tr>
<tr>
<td>Model 2</td>
<td>3166.7</td>
<td>3184.7</td>
<td>41.2</td>
<td>40.9</td>
<td>3166.7</td>
<td>3186.7</td>
<td>41.2</td>
<td>40.8</td>
</tr>
<tr>
<td>Model 3</td>
<td>3161.9</td>
<td>3185.9</td>
<td>41.5</td>
<td>41.0</td>
<td>3161.9</td>
<td>3187.9</td>
<td>41.5</td>
<td>40.9</td>
</tr>
</tbody>
</table>

Our simulation studies show the mean of the resulting $R^2$ values are very close to the corresponding population values. An exception occurs when the a ZIP model is fit to data for which a ZINB model is more appropriate. In that case, $R^2_D$ can be upwardly biased by the failure to account for overdispersion. As one would expect, the use of an over-specified model (a ZINB model when a ZIP model is adequate) leads to no such bias.
Though there are other $R^2$-related statistics that could be explored in a ZI regression context, we have chosen to concentrate on $R^2_D$ because of its tractability, good performance in simpler contexts (i.e., Poisson regression [Mittlböck and Waldhör 2000]), and its relatively straight-forward definition and interpretation. This approach is also promising for ZI models for clustered data such as ZI mixed effect models [Hall 2000], marginal ZI regression models [Hall and Zhang 2004], and for hurdle models. We intend to pursue such extensions in future work.
References


Chapter 3

Marginal zero-inflated regression models for count data

Abstract

Data sets with excess zeroes are frequently analyzed in many disciplines. A common framework used to analyze such data is the zero-inflated regression model. It mixes a degenerate distribution with point mass at zero with a non-degenerate distribution. The estimates from the zero inflated models quantify the effects of covariates on the means of latent random variables, which are often not the quantities of primary interest. Recently, marginal zero-inflated Poisson (MZIP; Long et al., 2014) and negative binomial (MZINB; Presser et al., 2015) models have been introduced that model the mean response directly, yielding covariate effects that have simpler interpretations that are more appealing for many applications. This paper outlines a general framework for marginalized zero-inflated models where the latent distribution is a member of the exponential dispersion family, focusing on common distributions for count data. In particular, our discussion includes the marginal zero inflated binomial (MZIB) model, which has not been discussed previously. The details of maximum likelihood estimation via the EM algorithm are presented and the properties of the estimators as well as Wald and likelihood ratio-based inference are examined via simulation. Two examples presented illustrate the advantages of MZIP, MZINB, and MZIB models for practical data analysis.

Keywords: exponential dispersion family; generalized linear models; zero inflation; marginalized models; EM algorithm

3.1 Introduction

Count data—both bounded and unbounded counts—regularly occur in a variety of disciplines. However, the standard modeling frameworks for such data, logistic regression and Poisson loglinear models, often fail to adequately account for the observed variability, a phenomenon known as overdispersion. In the common scenario that the data exhibit greater
spread throughout their range (fat tails) than standard distributions would predict, a dispersion parameter can be helpful, either as a component of a more general two-parameter count distribution such as the negative binomial or beta-binomial distribution, or as an extra variance parameter in a model based only on moment assumptions and analyzed via quasilikelihood. However, in other cases overdispersion may result from a more specific distributional violation: an excess frequency of zero values. This paper will focus on data that falls into the second scenario, which we refer to as excess zero (EZ) data. Many research disciplines encounter EZ data, such as manufacturing (Lambert, 1992), horticulture (Hall, 2000), entomology (Yeşilova et al., 2010), fish and wildlife conservation (Maunder and Punt, 2004), and psychology (Catty et al., 2008). The most common type of model for EZ data is the class of zero-inflated (ZI) regression models. Such models include the ZI Poisson (ZIP) and ZI negative binomial (ZINB) models for unbounded counts, and the ZI binomial (ZIB) model for bounded counts. Continuous data with many zeros occur as well. While such data, known as semi-continuous data, can often be analyzed more easily than discrete EZ data, there are situations for which ZI models such as the zero inflated tobit (ZIT) model prove useful (e.g., longitudinal semi-continuous data, data with both true zeros and zeros due to a detection limit). ZI models are based on a mixture structure in which excess zeros are hypothesized to come from a degenerate distribution that produces zeros with probability one, mixed with a standard distribution that is responsible for the non-zero data and, possibly, some of the zeros. Often these models include covariate effects, in which they become ZI regression models. In these cases, the dependence on covariates is built in through generalized linear model-like specifications for the non-degenerate distribution’s mean and the mixing probability.

The goal of this paper is to formulate and explore an alternative to the ZI regression model in which the marginal mean of the response variable is modeled directly. In a ZI model, the natural parameters of the model are the non-degenerate distribution’s mean
and the mixing probability. However, both of these are means for latent random variables rather than the observed response. This makes the interpretation of a ZI regression model awkward, especially if interest focuses on a covariate’s direct effect on the mean response, as it often does. To address this issue, we propose a marginal ZI model that retains the mixture structure of classical ZI regression, but parameterizes it with a linear predictor that links directly to the marginal mean response, making model interpretation more straightforward.

### 3.2 Zero Inflated Regression

Zero-inflated probability distributions have been in the statistical literature for more than a half century (Cohen 1963), but regression formulations were proposed in the 1980s and 1990s. Covariates were introduced in what Mullahy (1986) called a hurdle model, which assumes the data have probability mass

$$P(Y_i = k) = \begin{cases} 
\pi(g'_i \gamma) & \text{if } k = 0; \\
\{1 - \pi(g'_i \gamma)\} \frac{g[k; \lambda(x'_i \beta)]}{1 - g[0; \lambda(x'_i \beta)]} & \text{if } k > 0.
\end{cases}$$

This model assigns 0 with a probability $\pi$ and non-zero values according to an, often discrete, truncated distribution with mean $\lambda$. Heilbron (1994) recommended using a complementary log-log link for the zero component when the truncated distribution is Poisson, since $P(Y = 0) = \exp\{-\exp(x'_i \beta)\}$ and the aforementioned link is the inverse for the right side of this expression. He also recommended that the two pieces be related, using what he called a compatible model. Instead of having two unrelated parameter vectors for $\pi$ and $\lambda$, the model is parameterized so that $\beta_\pi = \gamma_2 \beta_\lambda + \log(\gamma_1)$. The benefit of the compatible model is it is simple to test for zero inflation or deflation. If $\gamma_1 = \gamma_2 = 1$, then a standard model is sufficient. Holding $\gamma_2 = 1$, if $\gamma_1 < 1$ then there is evidence of zero inflation, while $\gamma_1 > 1$ indicates zero deflation.
The drawback to the hurdle model is it assumes both the structural and distributional zeros can be adequately modeled by one regression. Elements of \( \gamma \) therefore describe covariate effects on a probability that pertains to two separate data-generating mechanisms, which is unappealing from an interpretation stand-point. Alternatively, Lambert (1992) introduced the zero-inflated regression model, which has the probability mass function (pmf),

\[
P(Y_i = k) = \begin{cases} 
\pi(g'_i \gamma) + [1 - \pi(g'_i \gamma)]g[0, \lambda(x'_i \beta)] & \text{if } k = 0; \\
\{1 - \pi(g'_i \gamma)}g[k, \lambda(x'_i \beta)] & \text{if } k > 0.
\end{cases}
\]

In this distribution, \( \pi \) is no longer the probability of observing a 0, but the probability of the observation being generated from a “perfect” state where only zeros occur. This distinction no longer requires truncating a distribution (Lambert focused on the Poisson distribution). Now a zero can belong to either the “perfect” degenerate state or the count distribution, but it is typically impossible with certainty to determine from which state it belongs.

The ZIP model allows the parameters \( \lambda \) and \( \pi \) to depend on covariates through the relationships

\[
\log \left[ \frac{\pi(g'_i \gamma)}{1 - \pi(g'_i \gamma)} \right] = g'_i \gamma \\
\log [\lambda(x'_i \beta)] = x'_i \beta,
\]

where \( x_i \) and \( g_i \) are the predictor vectors for the count mean and zero probability, respectively, and \( \beta \) and \( \gamma \) are the corresponding regression parameters. The coefficients are typically estimated using maximum likelihood, which can be implemented via direct optimization of the loglikelihood via, for example, gradient methods, or via the EM algorithm, which leads to iterative fitting of standard GLMs (Lambert 1992; Hall 2000).

Hall (2000) extended ZI regression to accommodate binomial counts and clustered data.

In the case of EZ bounded count data, the regression structures of the model take the form
\[
\log \left( \frac{\pi_i}{1 - \pi_i} \right) = g'_i \gamma
\]
\[
\log \left( \frac{\lambda_i}{\lambda_i - \lambda_j} \right) = x'_i \beta,
\]
where \( \lambda_i \) is the probability of a “success” for a binomial distribution. Hall recommends using the EM algorithm since maximizing the complete loglikelihood over \((\gamma', \beta')'\) can be done by iteratively using logistic regression for \(\gamma\) and weighted logistic regression for \(\beta\).

To model longitudinal or, more generally, clustered data, Hall proposed mixed effect versions of the ZIP and ZIB models. For clustered unbounded counts, the so-called ZIP-mixed model assumes that conditional on a subject-specific random effect \(b_i\), \(y_{ij}\), the \(j\)th response from the \(i\)th cluster, follows a ZIP distribution, with conditional mixing probability \(\pi_{ij}\) and Poisson mean \(\lambda_{ij}\) which are related to covariates and the random effects via

\[
\log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = g'_i \gamma
\]
\[
\log(\lambda_{ij}) = x'_{ij} \beta + z'_{ij} D^{t/2} b_i,
\]
where \(b_i \sim N_q(0, I)\) and \(D(\theta) = D^{t/2}D^{1/2}\) is a variance-covariance matrix parameterized by \(\theta\), and \(z_{ij}\) contains covariates/design variables for the random effects. The likelihood to be maximized is

\[
\ell(\gamma, \beta, \theta; y) = \sum_{i=1}^{K} \log \int_{R^q} \left[ \prod_{j=1}^{n_i} \Pr(Y_{ij} = y_{ij} | b_i) \right] \phi(b_i) db_i.
\]

In this case, maximization of the loglikelihood is complicated by the integration over the random effects distribution. Hall showed how the integration can be incorporated into the EM algorithm by using a Gaussian quadrature approximation. Alternatively, numerical or Monte Carlo integration techniques can be combined with gradient methods to maximize the loglikelihood, as implemented in PROC NLMIXED in SAS.
3.3 Marginal Regression Models

3.3.1 Model Formulation

One problem with ZI regression is that, in general, the resulting parameter estimates describe covariate effects on the latent count distribution, not the effects on the marginal mean, $E(Y|X)$. In many applications, interest focuses on the marginal mean, and while it is possible to compute the mean based on a standard ZI model, a covariate’s effect comes in through both $\pi$ and $\lambda$, complicating the interpretation of the model. Instead it is desirable to model a covariate’s effect on the mean directly while also accounting for excess zeroes. To do so, we reparameterize the ZI model. Let $\pi_i$ and $\lambda_i$ be the probability of the response being generated from the degenerate distribution and mean of the distribution, respectively, for the $i^{th}$ individual. Instead of a GLM-like specification for $\lambda_i$, as in standard ZI regression models, we assume $g(\mu_i) = x'_i \beta$ for $g$ a link suitable for the range of $\mu_i$. In addition, a logistic model, $\log(\pi_i/(1-\pi_i)) = g'_i \gamma$ is assumed for the mixing probability as usual. The resulting class of models, which we refer to as marginal ZI (MZI) regression models, are still ZI mixtures of a degenerate distribution at zero with a standard exponential family distribution, but the mean of the non-degenerate distribution $\lambda_i$, and thus also its canonical parameter $\theta_i$, now depends upon both $\beta$ and $\gamma$ through the relationship $\lambda_i = \mu_i/(1 - \pi_i)$. Writing the non-degenerate distribution’s density function in standard exponential dispersion family form, the probability density function for subject $i$ becomes

$$f(y_i; \gamma, \beta) = \left\{1 - \pi_i(\gamma) \right\} \left[ \frac{\pi_i(\gamma)}{1 - \pi_i(\gamma)} + \exp \left\{ \frac{b[\theta_i(\gamma, \beta)]}{a(\phi)} + c(0, \phi) \right\} \right]^{u_i} \left[ \exp \left\{ \frac{y_i \theta_i(\gamma, \beta) + b[\theta_i(\gamma, \beta)]}{a(\phi)} + c(y_i, \phi) \right\} \right]^{1-u_i}$$

(3.1)

where $u_i = 1$ if $y_i = 0$ and 0 otherwise.
The joint loglikelihood for the model is the sum of log densities of the form (3.1). A function optimizer can be used to find the MLE. However, this can be computationally taxing if the parameter dimension is large. Therefore we also outline an EM algorithm that is similar to the one used for standard ZI models, although it is not as convenient in the MZI case because the complete data likelihood does not factor cleanly into separate components for $\gamma$ and $\beta$.

The complete data loglikelihood is

$$
\ell_c(\gamma, \beta; y, z) = \sum_{i=1}^n \left\{ z_i \log \left( \frac{\pi_i}{1 - \pi_i} \right) + \log(1 - \pi_i) + (1 - z_i) \log h(y_i; \Psi) \right\},
$$

where $h(y_i; \Psi)$ is the non-degenerate distribution, $\Psi$ is the vector of parameters, and

$$
z_i = \begin{cases} 
1, & \text{if } y_i \text{ is from the degenerate distribution;} \\
0, & \text{otherwise.}
\end{cases}
$$

At the $r + 1^{th}$ EM iteration the E step is taken by estimating the posterior probabilities,

$$
z_i^{(r+1)} = \frac{\hat{\pi}_i^{(r)} u_i}{\hat{\pi}_i^{(r)} + \left( 1 - \hat{\pi}_i^{(r)} \right) h \left( y_i; \hat{\pi}_i^{(r)}, \hat{\mu}_i^{(r)} \right)}. \tag{3.2}
$$

Then the parameters, $\Psi$, can be estimated iteratively via a two-part M step. The first part uses a generic function optimizer (e.g., the nlm() function in R) to maximize $\ell_c(\gamma, \hat{\beta}^{(r)}; y, z^{(r+1)})$ to find $\hat{\gamma}^{(r+1)}$. After finding $\hat{\gamma}^{(r+1)}$, $\hat{\beta}^{(r+1)}$ is calculated by fitting a weighted GLM of the form $g^* \left( \lambda_i; (1 - \hat{\pi}_i^{(r+1)}) \right) = g \left( (1 - \hat{\pi}_i^{(r+1)}) \lambda_i \right) = x_i^T \beta$ with and weights $\omega_i^{(r+1)} = 1 - z_i^{(r+1)}$. Here, note that the link $g^*$ for this step is a function of $\lambda_i$ but also depends on $1 - \hat{\pi}^{(r+1)}$, which is treated as fixed. Its form matches that of $g$, the link function for $\mu_i$ in the MZI model specification, evaluated at $1 - \hat{\pi}^{(r+1)} \lambda_i$. The E and two-part M steps are iterated until the algorithm converges.
The benefit of using this approach is there is a reduction in the parameter dimension that is maximized using the function optimizer. Instead of maximizing $\ell(\gamma, \beta; y)$ with respect to $(\gamma', \beta')'$ using unstructured optimization, only $\ell^c(\gamma, \beta^{(r)}; y, z^{(r+1)})$ needs to be optimized with respect to $\gamma$ in this manner. The dimension of $\beta$ is often larger than that of $\gamma$, leading to a large decrease in computational complexity at each iteration of the fitting algorithm, although with the EM algorithm we can expect many iterations to be necessary. The information matrix can be found via direct differentiation of the observed data loglikelihood and is given in the appendix. At convergence, an estimated asymptotic variance-covariance matrix can be obtained from the inverse information matrix evaluating at the MLE $\hat{\Psi}$.

3.3.2 Special Cases

In this section we provide some estimation details for the most prominent special cases within the MZI model class.

MZIB Model

For the MZIB model, the non-degenerate component is $\text{Bin}(m_i, \lambda_i)$ and the observed data loglikelihood is

$$
\ell(\gamma, \beta; y, m) = \sum_{i=1}^{n} \left\{ \log(1 - \pi_i) + u_i \log \left[ \frac{\pi_i}{1 - \pi_i} + (1 - \pi_i)^{-m_i}(1 - \pi_i - \mu_i)^{m_i} \right] 
+ (1 - u_i) \left[ y_i \log \mu_i + (m_i - y_i) \log(1 - \pi_i - \mu_i) - m_i \log(1 - \pi_i) \right] \right\},
$$

where $m_i$ is the number of trials and $\mu_i$ is the marginal probability of success for the $i^{th}$ observation. In this model the link function $g$ relating $\mu$ to $x'\beta$ typically will be the logit, but other links such as the probit and complementary log-log functions are also permissible. The complete data log likelihood, on which the EM algorithm is based, takes the form
\[
\ell^c(\gamma, \beta; y, m) = \sum_{i=1}^{n} \{ \log(1 - \pi_i) + z_i \log \left( \frac{\pi_i}{1 - \pi_i} \right) + y_i \log \mu_i \\
+ (1 - z_i) [(m_i - y_i) \log(1 - \pi_i - \mu_i) - m_i \log(1 - \pi_i)] \},
\]

For the \( r + 1 \)th iteration of the EM algorithm, the steps are

1. (E Step) Calculate the posterior probabilities, \( z_i^{(r+1)} \), using (3.2).

2. (M Step) Update the estimates for \( \gamma \) and \( \beta \):
   
   (a) \( \hat{\gamma}^{(r+1)} \) is found by maximizing \( \ell^c(\gamma, \beta; y, m) \) with \( \beta \) fixed at \( \beta^{(r)} \).
   
   (b) \( \hat{\beta}^{(r+1)} \) is calculated using a GLM with:
      
      • link \( g^* \{ \lambda_i; (1 - \hat{\pi}_i^{(r+1)}) \} = g \left\{ \left( 1 - \hat{\pi}_i^{(r+1)} \right) \hat{\lambda}_i \right\} \)
      
      • weights \( \omega_i = 1 - z_i^{(r+1)} \)

3. Repeat steps 1 and 2 until convergence.

MZIP Model

For the MZIP model, the non-degenerate component takes the form Pois(\( \lambda_i \)), leading to an observed data loglikelihood of the form

\[
\ell(\gamma, \beta; y) = \sum_{i=1}^{n} \{ \log(1 - \pi_i) + u_i \log \left[ \frac{\pi_i}{1 - \pi_i} + \exp \left( \frac{-\mu_i}{1 - \pi_i} \right) \right] \\
+ (1 - u_i) \left( y_i \log \mu_i - \frac{\mu_i}{1 - \pi_i} \right) \},
\]

where again \( \mu_i \) is the marginal mean for the \( i \)th observation, and a log link for \( \mu \) is assumed.

In this case, the complete data log likelihood is
\[
\ell^c(\gamma, \beta; y) = \sum_{i=1}^{n} \left\{ z_i \log \frac{\pi_i}{1 - \pi_i} - (y_i - 1) \log(1 - \pi_i) + \frac{1 - z_i}{1 - \pi_i} \left[ y_i (1 - \pi_i) \log(\mu_i) - \mu_i \right] \right\}.
\]

For the \( r + 1 \)th iteration of the EM algorithm, the steps are

1. (E Step) Calculate the posterior probabilities, \( z^{(r+1)}_i \), using (3.2).

2. (M Step) Update the estimates for \( \gamma \) and \( \beta \):
   
   (a) \( \hat{\gamma}^{(r+1)} \) is found by maximizing \( \ell^c(\gamma, \beta; y) \) with \( \beta \) fixed at \( \beta^{(r)} \).
   
   (b) \( \hat{\beta}^{(r+1)} \) is calculated using a GLM with:
       
       - link \( g(\mu_i) = x'_i \beta \)
       - response \( y^c_i \equiv y_i \left( 1 - \hat{\pi}^{(r+1)}_i \right) \)
       - weights \( \omega_i = \frac{1 - z^{(r+1)}_i}{1 - \hat{\pi}^{(r+1)}_i} \).

3. Repeat steps 1 and 2 until convergence.

### MZINB Model

For the MZINB model, the non-degenerate component is NB(\( \lambda_i, \phi \)), and the observed data loglikelihood is

\[
\ell(\gamma, \beta, \phi; y) = \sum_{i=1}^{n} \left\{ \log(1 - \pi_i) + u_i \log \left[ \frac{\pi_i}{1 - \pi_i} + \left( \frac{(1 - \pi_i)\phi}{\mu_i + (1 - \pi_i)\phi} \right)^\phi \right] \right\} 
+ (1 - u_i) \left\{ y_i \log \mu_i + \phi \log[(1 - \pi_i)\phi] - (\phi + y_i) \log[\mu_i + (1 - \pi_i)\phi] + \log \frac{\Gamma(\phi + y)}{\Gamma(\phi)} \right\},
\]

44
As in the MZIP case, a log-linear model for $\mu$ is assumed. In this case, the complete data log likelihood is

$$\ell_c(\gamma, \beta, \phi; y) = \sum_{i=1}^{n} \left\{ \log(1 - \pi_i) + z_i \log \left( \frac{\pi_i}{1 - \pi_i} \right) + \frac{1 - z_i}{1 - \pi_i} [y_i(1 - \pi_i) \log \mu_i + (1 - \pi_i)\phi \log((1 - \pi_i)\phi)] - [(1 - \pi_i)\phi + y_i(1 - \pi_i)] \log[\mu_i + (1 - \pi_i)\phi] + \log \left( \frac{\Gamma(\phi + y_i)}{\Gamma(\phi)} \right) \right\}.$$  

For the $r + 1$th iteration of the EM algorithm, the steps are

1. (E Step) Calculate the posterior probabilities, $z_i^{(r+1)}$, using (3.2).

2. (M Step) Update the estimates for $\gamma$ and $\beta$:
   
   (a) $((\hat{\gamma}^{(r+1)})', \hat{\phi}^{(r+1)})'$ is found by maximizing $\ell_c(\gamma, \beta, \phi; y)$ with $\beta$ fixed at $\beta^{(r)}$.
   
   (b) $\hat{\beta}^{(r+1)}$ is calculated using a glm with:
      
      • link $g(\mu_i) = x_i'\beta$
      • response $y_i^c \equiv y_i \left( 1 - \hat{\pi}_i^{(r+1)} \right)$
      • weights $\omega_i = \frac{1 - z_i^{(r+1)}}{1 - \hat{\pi}_i^{(r+1)}}$
      • fixed dispersion parameter $\phi_i^c \equiv \hat{\phi}^{(r+1)} \left( 1 - \hat{\pi}_i^{(r+1)} \right)$

3. Repeat steps 1 and 2 until convergence.

Here, the quantity $\phi_i^c$ is treated as a fixed dispersion parameter in the iteratively reweighted least squares (IRLS) fitting algorithm used for $\beta$. This can be done in some GLM software and IRLS algorithms, or can be programmed easily from scratch.

Alternatively, $\phi$ can be grouped with $\beta$ in the M step and profiled during the M step for $(\beta', \phi)'$. This would fix $\phi$ over each value in grid of values in its parameter space, allowing $\beta$
to be updated via a standard GLM algorithm. This is akin to the approach is some negative
binomial regression software (e.g., the glm.nb function in the MASS package for R), but
offers no computational advantage over numerically optimizing $\ell^c(\gamma, \beta, \phi; y)$ with respect
to both $\gamma$ and $\phi$ as described above, which we find to be simpler to implement and faster
computationally.

### 3.3.3 Model interpretability

Because of the functional relationship between $\mu_i$ and $\pi_i$, one should take care to use sens-
able specifications for these two quantities in an MZI model. In particular, while it is not
incompatible to include a covariate in the model for $\pi_i$ but exclude its effect from the linear
predictor for $\mu_i$, such models are best avoided. The reason for this recommendation is that
if a covariate $x_i$ is specified to have a direct effect on $\pi_i$ then it will necessarily also have an
effect on $\mu_i$ through the relationship $\mu_i = (1 - \pi_i)\lambda_i$. Thus, omitting $x_i$ from the model for
$\mu_i$ implies that $\lambda_i$ changes with $x_i$ in just such a way as to nullify this covariate’s effect on
$\pi_i$. Such a model seems highly implausible; therefore, for model interpretability’s sake we
recommend that, in general, covariates in $G$ be a subset of those in $X$.

Also worth noting is the relationship between covariate effects on $\lambda_i$ and $\mu_i$ in ZI and
MZI models involving log link. In particular, consider a ZI model with specifications

$$\logit(\pi_i) = g_i'\gamma$$
$$\log(\lambda_i) = x'_i\beta.$$ 

The relationship between $\mu_i$, $\lambda_i$, and $\pi_i$ implies $\log(\mu_i) = \log(\lambda_i) + \log(1 - \pi_i) = x'_i\beta - \nu_i$, 
where $\nu_i = \log(1 + \exp(g'_i\gamma))$. Therefore, if the predictor $x_{ik}$ is not included in $g_i$, then $e^{\beta_k}$

46
quantifies the multiplicative effect of $x_{ik}$ on both $\lambda$ and $\mu$:

$$
\log\left(\frac{\mu'_i}{\mu_i}\right) = (c + 1)\beta_k + x_{i2}'\beta_2 - \nu_i - (c\beta_k + x_{i2}'\beta_2 - \nu_i)
$$

$$
= \beta_k = \log\left(\frac{\lambda'_i}{\lambda_i}\right),
$$

where, without loss of generality, $x_i = (x_i, x_{i2})'$, and where $\mu'_i = E(y_i|x_{i2}, x_{ik} = c + 1)$, $\mu_i = E(y_i|x_{i2}, x_{ik} = c)$, and $\lambda'_i, \lambda_i$ are defined similarly, conditional on $z_i = 1$.

In particular, when the mixing probability is constant, then the ZI and MZI models are reparameterizations of each other. In this case, the regression coefficients $\beta$ in an MZIP or MZINB model differ from those in the corresponding ZI model only through the intercept, $\beta_0$, which is shifted by the constant $\nu = \log(1 - \pi)$. The equivalence extends to models in which the mixing probability differs only across groups of subjects (that is, for an anova-like specification for $g'(\gamma)$), provided that the group structure is also included in the specification of $x_i'\beta$. E.g., in a ZI model with $\logit(\pi_{ij}) = \gamma_j$, and $\log(\lambda_{ij}) = \beta_j + x_{ij}'\beta$, then a marginal group mean for group $j$ at fixed $x_i$ is given by $E(y_{ij}|x_i) = \exp\{(\beta_j - \nu_j) + x_{ij}'\beta\}$. For cases when the mixing probability depends on continuous covariates or for models that lack a log link (e.g., for bounded count data), however, the ZI and MZI model classes are not equivalent.

When MZI and ZI models are distinct, the advantage of the former class is that its parameters can be interpreted as covariate effects or differences in group means on the marginal mean response (on a link-transformed scale, as in all GLMs). In contrast, ZI model parameters pertain to the mean of the latent non-degenerate variable in the mixture underlying both model classes. While effects on the marginal mean can be estimated from ZI model parameters, there are problems with such quantities. First, in a ZI model there is typically no scale on which a covariate effect is constant; it depends on the values of other explanatory variables in the model. This vastly diminishes the interpretability of a ZI regression model when interest focuses on marginal mean effects. Secondly, when
marginal mean effects are computed in a ZI model, these quantities are, in general, nonlinear functions of both \( \hat{\beta} \) and \( \hat{\gamma} \). While asymptotic normality for these quantities can often be justified via the delta method, the resulting asymptotic standard errors and normal-theory reference distributions for Wald inference can be expected to be poor relative to the more direct inference on marginal mean effects that is available for an MZI model. The latter consideration is not a concern when likelihood ratio inference is used, but Wald inference is often much easier in practice for these models because, for example, Wald confidence intervals are much less demanding computationally than profile likelihood intervals and formulating and/or fitting the model under a null hypothesis of interest is sometimes difficult. We investigate the inference advantages of MZI models over traditional ZI regression in section 3.4, but it is worth emphasizing the fact that the ZI model often does not yield constant covariate effects at all, over problems with inference on those effects in a ZI model.

### 3.4 Simulation

Simulation studies were performed to demonstrate the advantages of the MZI model over a ZI model, and to examine the properties of ML estimators and Wald and likelihood ratio-based inferences in these models. Following a similar formulation of Long et al. (2014), data were generated from MZI models, all of which take the form

\[
\begin{align*}
\text{logit}(\pi_i) &= \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i}, \\
g(\mu_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}
\end{align*}
\]

where \( x_1 \) is a vector of an equal number of zeros and ones and \( x_2 \sim \text{Uniform}(-1,1) \). Table 3.1 reports results for data generated from MZIP and MZINB models. In these models \( \gamma = (0.60, -2, 0.25)' \) and \( \beta = (0.25, \log(1.5), 0.25)' \) and, for the MZINB model, \( \phi = 3 \). Results for MZIB data appear in Table 3.2, in this case \( \gamma = (-1.386, 0.539, 0.25) \) and \( \beta = (-0.405, 0.811, -0.25) \). Properties of the estimates from correctly specified models fit to 10,000 data sets of size \( n = 100, 200, \) and 1,000 are reported. These properties include percent relative median bias (PRMB), simulation standard devia-
tion (SSD), the model based standard error (MBSE), and model-based coverage probability (MBCP) of 95% Wald confidence intervals.

<table>
<thead>
<tr>
<th></th>
<th>MZIP</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>MZINB</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Size</td>
<td>$\gamma_0$</td>
<td>$\gamma_1$</td>
<td>$\gamma_2$</td>
<td>$\beta_0$</td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
<td>$\gamma_0$</td>
<td>$\gamma_1$</td>
<td>$\gamma_2$</td>
<td>$\beta_0$</td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
<td>$\phi$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>-1.07</td>
<td>3.21</td>
<td>4.57</td>
<td>-6.35</td>
<td>0.86</td>
<td>-0.75</td>
<td>100</td>
<td>0.353</td>
<td>0.446</td>
<td>0.436</td>
<td>0.257</td>
<td>0.304</td>
<td>0.299</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PRMB</td>
<td>200</td>
<td>-0.77</td>
<td>1.55</td>
<td>3.82</td>
<td>-4.30</td>
<td>-0.23</td>
<td>200</td>
<td>0.232</td>
<td>0.320</td>
<td>0.295</td>
<td>0.175</td>
<td>0.236</td>
<td>0.211</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>0.60</td>
<td>0.28</td>
<td>-0.21</td>
<td>-1.36</td>
<td>0.35</td>
<td>1000</td>
<td>0.101</td>
<td>0.130</td>
<td>0.125</td>
<td>0.075</td>
<td>0.097</td>
<td>0.087</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SSD</td>
<td>100</td>
<td>0.334</td>
<td>0.409</td>
<td>0.414</td>
<td>0.244</td>
<td>0.290</td>
<td>100</td>
<td>0.950</td>
<td>0.944</td>
<td>0.946</td>
<td>0.944</td>
<td>0.947</td>
<td>0.950</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>0.229</td>
<td>0.310</td>
<td>0.291</td>
<td>0.169</td>
<td>0.233</td>
<td>200</td>
<td>0.951</td>
<td>0.949</td>
<td>0.951</td>
<td>0.946</td>
<td>0.950</td>
<td>0.951</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>0.101</td>
<td>0.129</td>
<td>0.126</td>
<td>0.075</td>
<td>0.097</td>
<td>1000</td>
<td>0.954</td>
<td>0.949</td>
<td>0.953</td>
<td>0.950</td>
<td>0.950</td>
<td>0.955</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MBSE</td>
<td>100</td>
<td>0.45</td>
<td>1.03</td>
<td>-2.11</td>
<td>-4.28</td>
<td>-0.95</td>
<td>100</td>
<td>0.56</td>
<td>0.23</td>
<td>-1.14</td>
<td>-2.06</td>
<td>0.35</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>0.45</td>
<td>0.364</td>
<td>0.338</td>
<td>0.198</td>
<td>0.276</td>
<td>200</td>
<td>0.118</td>
<td>0.157</td>
<td>0.150</td>
<td>0.089</td>
<td>0.116</td>
<td>0.107</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>0.456</td>
<td>0.673</td>
<td>0.550</td>
<td>0.297</td>
<td>0.361</td>
<td>1000</td>
<td>0.120</td>
<td>0.158</td>
<td>0.152</td>
<td>0.089</td>
<td>0.116</td>
<td>0.109</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MBSE</td>
<td>100</td>
<td>0.960</td>
<td>0.958</td>
<td>0.949</td>
<td>0.935</td>
<td>0.939</td>
<td>100</td>
<td>0.960</td>
<td>0.955</td>
<td>0.949</td>
<td>0.948</td>
<td>0.945</td>
<td>0.943</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>0.956</td>
<td>0.955</td>
<td>0.949</td>
<td>0.948</td>
<td>0.945</td>
<td>200</td>
<td>0.954</td>
<td>0.950</td>
<td>0.950</td>
<td>0.950</td>
<td>0.950</td>
<td>0.955</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>0.948</td>
<td>0.953</td>
<td>0.944</td>
<td>0.951</td>
<td>0.948</td>
<td>1000</td>
<td>0.948</td>
<td>0.948</td>
<td>0.948</td>
<td>0.948</td>
<td>0.948</td>
<td>0.960</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: Simulation results from the MZIP and MZINB models. PRMB: Percent relative median bias, SSD: Simulation standard deviation, MBSE: Model based standard error, MBCP: Model based coverage probability.

For the MZIP and MZINB results, the observed relative bias values for the sample size of 100 are much higher than those for sample sizes 200 and 1000 while the PRMB for the MZIB model is comparatively low for all three sample sizes. While there is some inaccuracy
Table 3.2: Results for the MZIB model. PRMB: Percent relative median bias, SSD: Simulation standard deviation, MBSE: Model based standard error, MBCP: Model based coverage probability in the MBSEs at the smaller sample sizes, this diminishes with increasing sample size, and the Wald coverage probabilities are accurate for all choices.

Table 3.3 shows the coverage probabilities of Wald 95% confidence intervals of $\mu_i = g^{-1}(\beta_0 + \beta_1 + \beta_2 x_{2i})$. Ten thousand data sets were generated from the MZIP and MZINB models where $\text{logit}(\pi_i) = 0.6 - 2x_{1i} + 0.25x_{2i}$, $\log(\mu_i) = 0.25 + \log(1.5)x_{1i} + 0.5x_{2i}$, and $\phi = 3$ and from a MZIB model with $\logit(\pi_i) = -1.386 + 0.539x_{1i} + 0.25x_{2i}$ and $\logit(\mu_i) = -0.405 + 0.405x_{1i} + 0.25x_{2i}$, where $x$ is defined previously. Both MZI and ZI models were fitted to these data sets and the confidence interval for $\mu_i$ was calculated by $g^{-1}\{\log(\hat{\mu}_i) \pm 1.96 \times se(\hat{\mu}_i)\}$. For the ZI models, the delta method was used to find the standard error of $\hat{\mu}_i$. Three different values of $x_{2i}$ were chosen when calculating the confidence interval for $\mu_i$, $-1$, $0$, and $1$. As expected, due to the linear relationship between the estimators for the ZI and MZI models with a log link when $x_{2i} = 0$, the ZIP and MZIP models gave almost identical coverage probabilities, and the same is true for the ZINB and MZINB models. Results are
also similar when \( x_{2i} = 0 \) for the ZIB and MZIB models, but to a lesser degree because the MZI and ZI model equivalence does not hold true for the logit link. When \( x_{2i} \neq 0 \) (most noticeably for \( x_{2i} = -1 \)), while the MZI model-based Wald intervals have close-to-nominal coverage, intervals for the ZI model have lower coverage, especially for sample sizes of 1000.

As noted by Long et al. (2014), the delta method for a ZI model strongly depends on the values of extraneous covariates and the larger the value of these covariates, the greater the effect on estimation of the standard error, with the effect increasing with the sample size.

<table>
<thead>
<tr>
<th>( x_{2i} )</th>
<th>Size</th>
<th>ZIP</th>
<th>MZIP</th>
<th>ZINB</th>
<th>MZINB</th>
<th>ZIB</th>
<th>MZIB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>0.944(0.941)</td>
<td>0.948(0.971)</td>
<td>0.938(1.134)</td>
<td>0.942(1.254)</td>
<td>0.945(0.212)</td>
<td>0.949(0.213)</td>
</tr>
<tr>
<td>-1</td>
<td>200</td>
<td>0.934(0.689)</td>
<td>0.953(0.836)</td>
<td>0.934(0.902)</td>
<td>0.945(1.014)</td>
<td>0.931(0.152)</td>
<td>0.947(0.154)</td>
</tr>
<tr>
<td>0</td>
<td>1000</td>
<td>0.853(0.290)</td>
<td>0.948(0.351)</td>
<td>0.915(0.376)</td>
<td>0.950(0.423)</td>
<td>0.873(0.068)</td>
<td>0.947(0.069)</td>
</tr>
<tr>
<td>0</td>
<td>200</td>
<td>0.949(0.668)</td>
<td>0.950(0.670)</td>
<td>0.945(0.832)</td>
<td>0.945(0.826)</td>
<td>0.945(0.136)</td>
<td>0.950(0.136)</td>
</tr>
<tr>
<td>0</td>
<td>1000</td>
<td>0.942(0.296)</td>
<td>0.943(0.299)</td>
<td>0.946(0.370)</td>
<td>0.947(0.370)</td>
<td>0.944(0.061)</td>
<td>0.950(0.061)</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>0.949(3.157)</td>
<td>0.944(3.245)</td>
<td>0.933(3.560)</td>
<td>0.943(3.581)</td>
<td>0.947(0.203)</td>
<td>0.950(0.209)</td>
</tr>
<tr>
<td>1</td>
<td>200</td>
<td>0.936(1.888)</td>
<td>0.946(1.973)</td>
<td>0.946(2.500)</td>
<td>0.944(2.522)</td>
<td>0.943(0.138)</td>
<td>0.951(0.142)</td>
</tr>
<tr>
<td>1</td>
<td>1000</td>
<td>0.903(0.815)</td>
<td>0.951(0.853)</td>
<td>0.929(1.070)</td>
<td>0.951(1.085)</td>
<td>0.914(0.062)</td>
<td>0.951(0.063)</td>
</tr>
</tbody>
</table>

Table 3.3: Comparing the coverage probabilities (width of interval) of the ZIP, ZINB, and ZIB models to their marginal equivalences.

Table 3.4 contains the results of simulations that examine tests of hypotheses that compare marginal means across two treatment groups. Here, we generated data from MZI models with \( g(\mu_i) = \beta_0 + \beta_1 \text{trt}_i + \beta_2 x_i, \ i = 1, ..., n \), where \( \text{trt}_i \) is a 0-1 indicator of which subjects received the treatment. In addition, we set \( \logit(\pi_i) = \gamma_0 + \gamma_1 \text{trt}_i + \gamma_2 x_i \) and the values of \( \gamma \) and \( \beta \) are the same as the simulations of Table 3.3 except \( \beta_1 \) takes three different values: 0, log(1.5), and log(2). Define \( R \) as the mean response ratio in the two treatment groups; we are interested in testing the hypothesis \( H_0 : \log(R) = 0 \). In MZI models this quantity does not depend upon \( x_i \) and the hypothesis can be formulated and computed easily using either a Wald or likelihood ratio (LR) test of \( H_{01} : \beta_1 = 0 \). In this context ZI regression models do not provide a convenient framework for inference on \( R \). Nevertheless, when available explanatory
variables include a treatment indicator and continuous covariate it is natural to consider ZI regression models of the form \( g(\lambda_i) = \beta_0^* + \beta_1^* \text{trt}_i + \beta_2^* x_i \), logit(\( \pi_i \)) = \( \gamma_0^* + \gamma_1^* \text{trt}_i + \gamma_2^* x_i \). Unlike the corresponding MZI model, here the ratio of means depends upon \( x_i \). In particular, the hypothesis \( \beta_1^* = 0 \) no longer addresses whether the ratio of marginal means is 0 and is not a suitable substitute for \( H_0 \). Instead we consider two hypothesis tests that one might conduct in the ZI model if one were interested in whether the marginal mean was equal across treatment groups. First, using the relationship \( \mu_i = (1 - \pi_i)\lambda_i \) one can always estimate \( R_i = E(\text{y}_i|\text{trt}_i = 1, x_{2i})/E(\text{y}_i|\text{trt}_i = 0, x_{2i}) \) and conduct a Wald test of the hypothesis \( H_{02} : \log(R_i) = 0 \) using the delta method to estimate the required nonlinear functions of the model’s regression parameters and their standard errors. Note that in the ZI model the ratio of marginal means depends on \( x_{2i} \), hence the notation \( R_i \) instead of \( R \). Thus \( H_{02} \) is not equivalent to \( H_0 \) and its target of inference varies with \( x_{2i} \). Despite this non-equivalence, we examine the ZI regression-based Wald test of \( H_{02} \) at \( x_{2i} = -1, 0, 1 \). Second, a natural way to assess the overall treatment effect in a ZI model is to consider models with and without the treatment factor in the set of explanatory variables in the model. That is, for ZI models we consider Wald and LR tests of \( H_{03} : \gamma_1^* = \beta_1^* = 0 \). Note that, again, \( H_{03} \) is not equivalent to \( H_0 \) because while \( H_{03} \) implies \( H_0 \), the converse is not true. In particular, it is possible for the treatment to have positive effects on both \( \lambda \) and \( \pi \) that offset to give a net null effect on the marginal mean response.

In Table 3.4 data are generated from MZI models with \( R = 1, 1.5, \) and 2 by setting \( \beta_1 \) to 0, log(1.5) and log(2), respectively. In the first case, \( H_{01} \) is true, \( H_{02} \) is true provided that \( x_{2i} = 0 \), and \( H_{03} \) is false. In this setting, all of the tests for \( H_{01} \), as well as tests for \( H_{02} \) when \( x_{2i} = 0 \), have empirical sizes that are reasonably close to nominal (0.05) at the largest sample size; the Wald and LR tests of \( H_{01} \), which are those that are based on fitted MZI models, show the best performance overall. When \( x_{2i} \neq 0 \), the Wald test of \( H_{02} \) based on a ZI model often rejects at a much higher rate than .05, and its rejection rate increases
with the sample size. The non-equivalence of $H_0$ and $H_{03}$ can be seen by the high rejection rates for tests of the latter hypothesis based on ZI models, when the data are generated with $\log(R) = 0$. Thus none of the tests based on ZI models that attempt to address this hypothesis are suitable. When $H_0$ is false, the power of the Wald and LRTs of $H_{01}$ are similar, with neither test dominating the other.

<table>
<thead>
<tr>
<th>$x_{2i} = -1$</th>
<th>MZIP</th>
<th>MZINB</th>
<th>MZIB</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$ Size</td>
<td>Wald LRT</td>
<td>Wald LRT</td>
<td>Wald LRT</td>
</tr>
<tr>
<td>100</td>
<td>0.909 0.906 0.909 0.908 0.909</td>
<td>0.906 0.905 0.905 0.904 0.904</td>
<td>0.903 0.902 0.902 0.901 0.901</td>
</tr>
<tr>
<td>1</td>
<td>0.048 0.048 0.048 1.000 1.000</td>
<td>0.057 0.056 0.056 0.992 0.992</td>
<td>0.063 0.062 0.062 0.831 0.831</td>
</tr>
<tr>
<td>1000</td>
<td>0.949 0.947 0.948 1.000 1.000</td>
<td>0.956 0.956 0.956 1.000 1.000</td>
<td>0.959 0.959 0.959 1.000 1.000</td>
</tr>
<tr>
<td>1.5</td>
<td>0.468 0.463 0.489 0.585 0.608</td>
<td>0.379 0.367 0.346 0.374 0.393</td>
<td>0.220 0.215 0.178 0.320 0.320</td>
</tr>
<tr>
<td>200</td>
<td>0.720 0.714 0.700 0.898 0.898</td>
<td>0.577 0.573 0.595 0.611 0.629</td>
<td>0.384 0.372 0.294 0.961 0.961</td>
</tr>
<tr>
<td>1000</td>
<td>1.000 1.000 1.000 1.000 1.000</td>
<td>0.996 0.996 0.997 1.000 1.000</td>
<td>0.954 0.953 0.889 1.000 1.000</td>
</tr>
<tr>
<td>2</td>
<td>0.924 0.921 0.942 0.991 0.991</td>
<td>0.827 0.821 0.799 0.835 0.849</td>
<td>0.607 0.593 0.495 0.992 0.992</td>
</tr>
<tr>
<td>1000</td>
<td>1.000 1.000 1.000 1.000 1.000</td>
<td>0.983 0.982 0.982 0.995 0.995</td>
<td>0.855 0.852 0.761 1.000 1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$x_{2i} = 0$</th>
<th>MZIP</th>
<th>MZINB</th>
<th>MZIB</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$ Size</td>
<td>Wald LRT</td>
<td>Wald LRT</td>
<td>Wald LRT</td>
</tr>
<tr>
<td>100</td>
<td>0.056 0.059 0.060 0.997 0.998</td>
<td>0.060 0.059 0.061 0.878 0.965</td>
<td>0.061 0.061 0.059 0.507 0.517</td>
</tr>
<tr>
<td>1</td>
<td>0.049 0.047 0.046 1.000 1.000</td>
<td>0.061 0.060 0.061 0.994 0.998</td>
<td>0.053 0.052 0.045 0.831 0.836</td>
</tr>
<tr>
<td>1000</td>
<td>0.063 0.066 0.066 1.000 1.000</td>
<td>0.062 0.058 0.064 1.000 1.000</td>
<td>0.062 0.061 0.060 1.000 1.000</td>
</tr>
<tr>
<td>1.5</td>
<td>0.347 0.382 0.320 0.986 0.991</td>
<td>0.264 0.275 0.234 0.751 0.932</td>
<td>0.416 0.405 0.413 1.000 1.000</td>
</tr>
<tr>
<td>200</td>
<td>0.609 0.633 0.508 1.000 1.000</td>
<td>0.480 0.496 0.490 0.974 1.000</td>
<td>0.659 0.653 0.660 1.000 1.000</td>
</tr>
<tr>
<td>1000</td>
<td>0.997 0.997 0.997 1.000 1.000</td>
<td>0.995 0.995 0.992 1.000 1.000</td>
<td>1.000 1.000 1.000 1.000 1.000</td>
</tr>
<tr>
<td>2</td>
<td>0.814 0.835 0.803 0.984 0.988</td>
<td>0.657 0.670 0.648 0.834 0.967</td>
<td>0.734 0.732 0.715 1.000 1.000</td>
</tr>
<tr>
<td>1000</td>
<td>0.990 0.990 0.990 1.000 1.000</td>
<td>0.937 0.940 0.935 0.984 1.000</td>
<td>0.967 0.964 0.961 1.000 1.000</td>
</tr>
<tr>
<td>2</td>
<td>1.000 1.000 1.000 1.000 1.000</td>
<td>1.000 1.000 1.000 1.000 1.000</td>
<td>1.000 1.000 1.000 1.000 1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$x_{2i} = 1$</th>
<th>MZIP</th>
<th>MZINB</th>
<th>MZIB</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$ Size</td>
<td>Wald LRT</td>
<td>Wald LRT</td>
<td>Wald LRT</td>
</tr>
<tr>
<td>100</td>
<td>0.059 0.056 0.054 0.997 1.000</td>
<td>0.065 0.069 0.061 0.890 0.961</td>
<td>0.062 0.058 0.055 0.492 0.506</td>
</tr>
<tr>
<td>1</td>
<td>0.068 0.064 0.051 1.000 1.000</td>
<td>0.050 0.045 0.045 0.991 0.998</td>
<td>0.036 0.036 0.047 0.826 0.828</td>
</tr>
<tr>
<td>1000</td>
<td>0.054 0.057 0.117 1.000 1.000</td>
<td>0.049 0.047 0.120 1.000 1.000</td>
<td>0.044 0.044 0.097 1.000 1.000</td>
</tr>
<tr>
<td>1.5</td>
<td>0.335 0.364 0.254 0.987 0.991</td>
<td>0.230 0.242 0.175 0.791 0.952</td>
<td>0.374 0.366 0.559 1.000 1.000</td>
</tr>
<tr>
<td>200</td>
<td>0.614 0.641 0.549 1.000 1.000</td>
<td>0.457 0.469 0.433 0.987 0.999</td>
<td>0.664 0.662 0.848 1.000 1.000</td>
</tr>
<tr>
<td>1000</td>
<td>0.999 0.999 1.000 1.000 1.000</td>
<td>0.996 0.997 0.997 1.000 1.000</td>
<td>1.000 1.000 1.000 1.000 1.000</td>
</tr>
<tr>
<td>2</td>
<td>0.821 0.838 0.629 0.977 0.986</td>
<td>0.653 0.667 0.484 0.805 0.958</td>
<td>0.731 0.721 0.905 1.000 1.000</td>
</tr>
<tr>
<td>1000</td>
<td>1.000 1.000 1.000 1.000 1.000</td>
<td>1.000 1.000 1.000 1.000 1.000</td>
<td>1.000 1.000 1.000 1.000 1.000</td>
</tr>
</tbody>
</table>

Table 3.4: The power of a .05 significance level test for $H_{01} : \log(R) = 0$ and $H_{02} : \gamma_1 = \beta_1 = 0$. 

53
3.5 Examples

3.5.1 Apple roots

In Ridout et al. (1998), data from an experiment in which the number of roots produced by shoots of an apple tree cultivar were analyzed. Growing conditions were determined by the $2 \times 4$ factorial combinations of two factors: length of photoperiod (8 and 16h) and concentration of the cytokinin BAP (2.2, 4.4, 8.8, and 17.6 µM). The design was unbalanced, with a total of 267 shoots propagated, where two treatments had 40 shoots, one had 39 shoots, four had 30 shoots, and one had 28 shoots propagated. The response variable is the number of roots grown per shoot. The data set contains a large number of zeros, but histograms of the response by photoperiod suggest that zero inflation is primarily of concern under the 16 hour photoperiod (see Figure 3.1).

Let $y_{ijk}$ be the number of roots for the $k^{th}$ shoot at the $i^{th}$ photoperiod and $j^{th}$ BAP concentration. Also, $\pi_{ijk}$, $\lambda_{ijk}$, and $\mu_{ijk}$ are the mixing probability, latent mean, and overall mean for $y_{ijk}$, respectively. Three different ZIP and ZINB models for the data were fitted and compared:

Model 1: $\text{logit}(\pi_{ijk}) = \gamma_{1i}$, \quad $\text{log}(\lambda_{ijk}) = \beta_{1i} + \beta_{2}\text{log}(\text{BAP}_{ij})$;

Model 2: $\text{logit}(\pi_{ijk}) = \gamma_{1i}$, \quad $\text{log}(\lambda_{ijk}) = \beta_{1i} + \beta_{2i}\text{log}(\text{BAP}_{ij})$;

Model 3: $\text{logit}(\pi_{ijk}) = \gamma_{1i} + \gamma_{2}\text{log}(\text{BAP}_{ij})$, \quad $\text{log}(\lambda_{ijk}) = \beta_{1i} + \beta_{2i}\text{log}(\text{BAP}_{ij})$.

In addition, corresponding MZIP and MZINB models were also fit that have linear predictors of the same form as Models 1–3. The parameter estimates for each model are found in Table 3.5.

Because LR tests and AIC statistics favor the NB-based models over the corresponding Poisson-based models, we restrict further attention to ZINB and MZINB models. From Model 1, it would appear that $\text{log}(\text{BAP})$ has little effect on the mean since the standard error is greater than $\hat{\beta}_{2}$. However, when separated into two different slopes for each photoperiod,
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZIP</td>
<td>MZIP</td>
<td>ZIP</td>
</tr>
<tr>
<td>γ_{11}</td>
<td>-4.275</td>
<td>0.747</td>
<td>-4.256</td>
</tr>
<tr>
<td>γ_{12}</td>
<td>-0.071</td>
<td>0.178</td>
<td>-0.073</td>
</tr>
<tr>
<td>γ_{2}</td>
<td>-0.011</td>
<td>0.217</td>
<td>-0.011</td>
</tr>
<tr>
<td>β_{11}</td>
<td>1.933</td>
<td>0.075</td>
<td>1.794</td>
</tr>
<tr>
<td>β_{12}</td>
<td>1.669</td>
<td>0.087</td>
<td>1.011</td>
</tr>
<tr>
<td>β_{2}</td>
<td>0.019</td>
<td>0.035</td>
<td>0.019</td>
</tr>
<tr>
<td>β_{21}</td>
<td>-0.084</td>
<td>0.179</td>
<td>-0.084</td>
</tr>
<tr>
<td>β_{22}</td>
<td>0.019</td>
<td>0.044</td>
<td>0.019</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZINB</td>
<td>MZINB</td>
<td>ZINB</td>
</tr>
<tr>
<td>γ_{11}</td>
<td>-4.481</td>
<td>0.925</td>
<td>-4.371</td>
</tr>
<tr>
<td>γ_{12}</td>
<td>-0.084</td>
<td>0.179</td>
<td>-0.084</td>
</tr>
<tr>
<td>γ_{2}</td>
<td>-0.026</td>
<td>0.220</td>
<td>-0.025</td>
</tr>
<tr>
<td>β_{11}</td>
<td>1.944</td>
<td>0.094</td>
<td>1.794</td>
</tr>
<tr>
<td>β_{12}</td>
<td>1.676</td>
<td>0.106</td>
<td>1.024</td>
</tr>
<tr>
<td>β_{2}</td>
<td>0.012</td>
<td>0.044</td>
<td>0.012</td>
</tr>
<tr>
<td>β_{21}</td>
<td>0.000</td>
<td>0.052</td>
<td>0.000</td>
</tr>
<tr>
<td>β_{22}</td>
<td>0.000</td>
<td>0.078</td>
<td>0.000</td>
</tr>
<tr>
<td>log(φ)</td>
<td>2.553</td>
<td>0.330</td>
<td>2.688</td>
</tr>
<tr>
<td>AIC</td>
<td>1251.116</td>
<td>1242.832</td>
<td>1244.829</td>
</tr>
</tbody>
</table>

Table 3.5: Estimates and standard errors of the three proposed models for the apple root data.
Figure 3.1: Histograms for the apple root data set by photoperiod

log(BAP) does have a significant effect on the mean, one that is positive for photoperiod 8 and negative for photoperiod 16. Since by Wald and likelihood ratio tests $\hat{\gamma}_2$ is not statistically different from zero in Model 3 and Model 2 has the lowest AIC, we conclude that log(BAP) has a negligible effect on the mixing probabilities and settle on the MZINB and ZINB versions of Model 2 for further analyses.

Because Model 2 has an anova structure for $\pi$, the MZINB and ZINB models are equivalent as described in section 3.3.3. In particular, for this model we have
This relationship is satisfied by the parameter estimates in Table 3.5. Note that the estimated probability of zero inflation for photoperiod=8 is so small ($\hat{\pi}_1 = \log^{-1}(\hat{\gamma}_{11}) = 0.012$) that the difference in the estimates of $\beta_{11}$ is negligible for the MZINB and ZINB models. In contrast (and as seen in Figure 1), much greater evidence of zero inflation was found for photoperiod=16 ($\hat{\pi}_2 = 0.479$) leading to a much more substantial shift in $\hat{\beta}_{12}$ for the MZINB model relative to the ZINB model. Note also that while MZI and ZI results are quite similar for Model 3, these models are not equivalent because of the presence of a covariate effect in the model for $\pi$. The close similarity of the parameter estimates and AIC statistics for the two model types is simply because the estimate of $\gamma_2$ is very near 0.

### 3.5.2 Whitefly data

van Iersel et al. (2000) report the results of a horticultural experiment to examine the effect of six methods of applying pesticide to control silverleaf whiteflies on greenhouse-raised poinsettia. Fifty four plants were used in a randomized complete block design where experimental units containing three plants were randomly assigned to each treatment in three complete blocks. Each week a collection of whiteflies was placed on each plant (confined inside a cage clipped to a leaf of the plant) and then the number of surviving whiteflies was measured two days later. It was of interest to compare the survival probabilities across plants treated with different methods of pesticide application. Despite the large frequency of zero-valued responses in these data, the ZIB model is not the best framework for analyzing them because its estimates of $\beta$ quantify treatment effects on the conditional probability of survival given that the response comes from the binominal component of the mixture. In the
context of the problem one might interpret this as the conditional probability of survival given that the treatment is not certain to kill all whiteflies. In contrast, the $\beta$ coefficients in an MZIB model provide a more straight-forward and desirable interpretation as direct treatment effects on the log odds of survival.

Although the original experimental design included repeated measures over time, for simplicity we analyze only the data from week 6, the middle week of the study. Extensions of MZI models to the longitudinal/clumped data setting will be considered elsewhere. The six treatments used in the experiment included four methods of sub-irrigated pesticide following 4, 2, 1 and 0 days without water (treatments 1-4 respectively), a treatment involving pesticide application via hand-watering (treatment 5), and a control treatment in which no pesticide was used (treatment 6).

Following van Iersel et al. (2000) we summed the responses for each trio of plants and analyzed a single bounded count per experimental unit. Let $\mu_{ij}$ represent the probability of survival for the $n_{ij}$ insects from the experimental unit assigned to the $j$th treatment in the $i$th block, which was modeled by $\text{logit}(\mu_{ij}) = \beta_0 + \beta_{1i} + \beta_{2j}$ where $\beta_0$ is the intercept, $\beta_{1i}, i = 1, 2, 3$, is the effect of the $i$th block, and $\beta_{2j}, j=1,2,...,6$, is the effect of the $j$th treatment. Two models with this specification for $\mu_{ij}$ were fit to the data, differing in their specification of the mixing probability: (a) $\text{logit}(\pi_{ij}) = \gamma_0$, and (b) $\text{logit}(\pi_{ij}) = \gamma_0 + \gamma_{1j}$. Models with both treatment and block effects on the mixing probability were not considered due to limited degrees of freedom and the sparse occurrence of zeros across blocks. A Wald test of $H_0: \gamma_{11} = \cdots = \gamma_{16} = 0$ strongly favors model (b) ($P(\chi^2_5 > 60.27) \approx 0$).

Table 3.6 shows the estimated marginal probabilities for each block and treatment combination. Based on model (b), there are several sets of contrasts among the six treatment effects that one might choose to test. We tested $H_{0A} : \frac{1}{5} \sum_{j=1}^{5} \beta_{2j} - \beta_{26}$, $H_{0B} : \frac{1}{4} \sum_{j=1}^{4} \beta_{2j} - \beta_{25}$, and $H_{0C} : \beta_{21} = \cdots = \beta_{24}$ to compare active treatments to the control ($P(\chi^2_1 > 30.61) \approx 0$), sub-irrigation to hand watering ($P(\chi^2_1 > 8.29) = 0.004$), and the 4 levels of restricting wa-
ter prior to subirrigation-based pesticide application \( P(\chi^2_3 > 3.71) = 0.294 \), respectively. These hypotheses have straight-forward interpretations, and we conclude that the odds of survival are significantly different (lower) when pesticide is applied, significantly different (lower) when subirrigation is used for pesticide delivery, and that the survival odds do not differ significantly depending on the length of water restriction prior to subirrigation with pesticide.

<table>
<thead>
<tr>
<th>Block</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.011</td>
<td>0.009</td>
<td>0.000</td>
<td>0.039</td>
<td>0.160</td>
<td>0.742</td>
</tr>
<tr>
<td>2</td>
<td>0.045</td>
<td>0.040</td>
<td>0.000</td>
<td>0.148</td>
<td>0.453</td>
<td>0.926</td>
</tr>
<tr>
<td>3</td>
<td>0.021</td>
<td>0.018</td>
<td>0.000</td>
<td>0.073</td>
<td>0.273</td>
<td>0.850</td>
</tr>
</tbody>
</table>

Table 3.6: \( \hat{\mu}_{ij} \) for each block and treatment combination.

An advantage of the MZIB model compared to the ZIB model is that it estimates survival odds ratios for distinct treatments that do not depend on block. Given the additive structure of the model, this feature is desirable. E.g., using the MZIB model, we may estimate the ratio of survival odds for treatments 2 and 5 as \( \exp(\hat{\beta}_{25} - \hat{\beta}_{22}) = 20.08 \). In contrast, a ZIB model with linear predictors of the same form as our chosen MZIB model yields an estimated survival odds ratio for treatments 2 and 5 of \( \frac{\hat{\mu}_5 \cdot 1 - \hat{\mu}_2}{1 - \hat{\mu}_5 \cdot \hat{\mu}_2} = \frac{(1 - \hat{\pi}_i \hat{\lambda}_5, 1 - (1 - \hat{\pi}_i \hat{\lambda}_2)}{1 - (1 - \hat{\pi}_i \hat{\lambda}_5 \hat{\lambda}_2)} \). Note that this quantity depends upon \( i \), so the model, despite the additive structure of its linear predictors, essentially induces an interaction between treatments and blocks. The ratio of survival odds for treatments 2 and 5 for blocks 1, 2, and 3 for a ZIB model are 27.29, 19.88, and 25.53, respectively. Note that the average of the odd ratios from the ZIB model is not especially close to the odds ratio found from the MZIB model, which highlights the fact that these models are not merely reparameterizations of each other and should not be expected to agree closely.
3.6 Conclusion

ZI regression is a useful and appealing framework for analyzing count data with excess zeros, but it is awkward to obtain inferences on the marginal mean from such models. Often there is no scale on which covariate effects are constant, precluding succinct statements about the relationship between the mean response and a covariate of interest. In many contexts an MZI model will prove more convenient and interpretable. Such models retain the basic mixture structure of ZI regression, but provide direct estimates of marginal mean effects.

In addition to providing direct estimates of interpretable quantities, MZI models allow more straightforward inferences, avoiding the need for delta-method confidence intervals and tests. This is not only an advantage of convenience but, as shown in our simulation results, produces more accurate coverage and error rates. As the simulations show, the coverage probabilities for confidence intervals derived from the ZI models decrease as the sample size increases while the coverage of the intervals formed from the MZI models remain close to the desired level, which is consistent with the results from Long et al. (2014). When testing for a treatment effect on the marginal mean, only those using the results of the MZI model have the correct size and therefore the ZI model should not be used for inference on the marginal mean of the response.

While special cases of MZI regression have been considered recently by other authors using maximum likelihood estimation (Long et al., 2014; Preisser et al., 2015) and moment-based estimation (Staub and Winkelmann, 2013), we formulate a broader class of models and describe a general EM-based estimation framework. Both straightforward loglikelihood maximization and the EM algorithm require using a function optimizer to find the estimates for the mixing probabilities. The advantage of the EM algorithm is the estimation of \( \Psi = (\gamma', \beta')' \) using an optimizer can be computational cumbersome, the EM algorithm only needs
to use such a function to estimate $\gamma$, which often has a much small dimension than $\Psi$, which can reduce the computation time when the dimension of $\Psi$ is large.

While in this paper we have focused on models for independent data (e.g., a cross-sectional study design), extensions to accommodate longitudinal/clustered data are desirable. For this purpose a generalization of the MZI regression model class that is natural and tractable incorporates random cluster-specific random effects. Such an extended mixed-effect MZI class is the subject of ongoing research and will appear elsewhere. While this paper gives details on several special cases of MZI models (MZIP, MZIB, MZINB), these are other examples we have not discussed. For instance an MZI tobit model may be useful for semi-continuous data and can be seen as an alternative to the ZI tobit model (e.g., Moulton et al., 2002).
References


severe mental illness: results of an international six-centre randomised controlled trial. 


65


**Appendix**

Denote the link functions and linear predictors of the MZI regression model as

\[
g(\pi_i) = \delta_i = g'_i \gamma \\
h(\mu_i) = \eta_i = x'_i \beta
\]

and let \(d(\theta_i, \phi) = \frac{-b(\theta_i)}{a(\phi)} + c(0, \phi)\) (cf. equation (3.2)). It follows that

\[
P(Y_i = 0) = \frac{\pi_i}{1-\pi_i} + \exp\{d(\theta_i, \phi)\}
\]

\[
P(Y_i \neq 0) = \frac{\eta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi),
\]
and the loglikelihood of the regression model can be written

\[ \ell(\gamma, \beta, \phi; y) = \sum_{i=1}^{n} \{ \log(1 - \pi_i) + u_i \log P(Y_i = 0) + (1 - u_i)P(Y_i \neq 0) \} \]

The information matrix is given by

\[
I(\gamma, \beta) = -E \left( \begin{array}{cccc}
\frac{\partial^2 \ell(\gamma, \beta, \phi; y)}{\partial \gamma \partial \gamma'} & \frac{\partial^2 \ell(\gamma, \beta, \phi; y)}{\partial \gamma \partial \beta'} & \frac{\partial^2 \ell(\gamma, \beta, \phi; y)}{\partial \gamma \partial \phi'} & \frac{\partial^2 \ell(\gamma, \beta, \phi; y)}{\partial \phi \partial \gamma'} \\
\frac{\partial^2 \ell(\gamma, \beta, \phi; y)}{\partial \beta \partial \gamma'} & \frac{\partial^2 \ell(\gamma, \beta, \phi; y)}{\partial \beta \partial \beta'} & \frac{\partial^2 \ell(\gamma, \beta, \phi; y)}{\partial \beta \partial \phi'} & \frac{\partial^2 \ell(\gamma, \beta, \phi; y)}{\partial \phi \partial \beta'} \\
\frac{\partial^2 \ell(\gamma, \beta, \phi; y)}{\partial \phi \partial \gamma'} & \frac{\partial^2 \ell(\gamma, \beta, \phi; y)}{\partial \phi \partial \beta'} & \frac{\partial^2 \ell(\gamma, \beta, \phi; y)}{\partial \phi \partial \phi'} & \frac{\partial^2 \ell(\gamma, \beta, \phi; y)}{\partial \phi \partial \phi'} \\
\end{array} \right),
\]

where

\[
\frac{\partial^2 \ell(\gamma, \beta, \phi)}{\partial \Psi \partial \nu'} = \sum_{i=1}^{n} \left\{ u_i \left[ \frac{\partial^2 P(Y_i = 0)}{\partial \Psi \partial \nu'} P(Y_i = 0) - \frac{\partial P(Y_i = 0)}{\partial \Psi} \frac{\partial P(Y_i = 0)}{\partial \nu'} \right] P(Y_i = 0)^{-2} \right.
+ \left. (1 - u_i) \frac{\partial^2 P(Y_i \neq 0)}{\partial \Psi \partial \nu'} + \frac{\partial^2 \log(1 - \pi_i)}{\partial \Psi \partial \nu'} \right\},
\]

for \( \Psi, \nu \in \{\gamma, \beta, \phi\} \).

Let \( r_{\pi_i} = \frac{\partial \theta_i}{\partial \pi_i}, r_{\mu_i} = \frac{\partial \theta_i}{\partial \mu_i}, t_i = \frac{\partial \theta_i}{\partial \pi_i}, \frac{\partial \theta_i}{\partial \mu_i}, \) and \( s_i = \frac{\partial \mu_i}{\partial \eta_i}, \frac{\partial \eta_i}{\partial \mu_i}, \) note that if the canonical links are used, \( r_{\pi_i} = r_{\mu_i} = s_i t_i = 1 \). The partial derivatives needed to compute the information matrix are as follows:
\[
\frac{\partial P(Y_i = 0)}{\partial \beta} = -a(\phi)^{-1} \exp\{d(\theta_i, \phi)\} \frac{\partial b(\theta_i)}{\partial \theta_i} r_{\mu_i} x_i'
\]

\[
\frac{\partial P(Y_i = 0)}{\partial \gamma} = \left[ (1 - \pi_i)^{-2} - \exp\{d(\theta_i, \phi)\} a(\phi)^{-1} \frac{\partial b(\theta_i)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \pi_i} \right] \frac{\partial \pi_i}{\partial \theta_i} g_i'
\]

\[
\frac{\partial P(Y_i = 0)}{\partial \phi} = -\exp\{d(\theta_i, \phi)\} \left[ \frac{a'(\phi)}{a(\phi)^2} b(\theta_i) + c'(0, \phi) \right]
\]

\[
\frac{\partial^2 P(Y_i = 0)}{\partial \beta \partial \beta'} = a(\phi)^{-1} \exp\{d(\theta_i, \phi)\} \left[ \left( \frac{\partial b(\theta_i)}{\partial \theta_i} \right)^2 \frac{r_{\mu_i}}{a(\phi)} - \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} r_{\mu_i} - \frac{\partial b(\theta_i)}{\partial \theta_i} \frac{\partial^2 \theta_i}{\partial \mu_i^2} \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2 - \frac{\partial b(\theta_i)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial^2 \mu_i}{\partial \eta_i^2} \right] x_i x_i'
\]

\[
\frac{\partial^2 P(Y_i = 0)}{\partial \gamma \partial \gamma'} = \left[ 2(1 - \pi_i)^{-3} - a(\phi)^{-1} \exp\{d(\theta_i, \phi)\} \left\{ -a(\phi)^{-1} \left( \frac{\partial b(\theta_i)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \pi_i} \right)^2 + \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} \left( \frac{\partial \theta_i}{\partial \pi_i} \right)^2 \right\} \right] + \frac{\partial^2 b(\theta_i)}{\partial \theta_i \partial \theta_i} \left( \frac{\partial \pi_i}{\partial \delta_i} \right)^2 + \left( 1 - \pi_i^{-2} - \exp\{d(\theta_i, \phi)\} a(\phi)^{-1} \frac{\partial b(\theta_i)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \pi_i} \right) \frac{\partial^2 \pi_i}{\partial \delta_i^2} g_i g_i'
\]

\[
\frac{\partial^2 P(Y_i = 0)}{\partial \gamma' \partial \beta} = a(\phi)^{-1} \exp\{d(\theta_i, \phi)\} \left[ \left\{ \left( \frac{\partial b(\theta_i)}{\partial \theta_i} \right)^2 a(\phi)^{-1} - \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} \right\} t_i - \frac{\partial b(\theta_i)}{\partial \theta_i} \frac{\partial^2 \theta_i}{\partial \mu_i \partial \pi_i} \right] s_i g_i x_i'
\]

\[
\frac{\partial^2 P(Y_i = 0)}{\partial \phi^2} = \exp\{d(\theta_i, \phi)\} \left[ \left( \frac{a'(\phi)}{a(\phi)^2} b(\theta_i) + c'(0, \phi) \right) \right]^2 + \frac{a''(\phi) a(\phi) - 2a'(\phi)^2 b(\theta_i) + c''(0, \phi)}{a(\phi)^3}
\]

\[
\frac{\partial^2 P(Y_i = 0)}{\partial \phi \partial \gamma} = a(\phi)^{-1} \exp\{d(\theta_i, \phi)\} \left[ \frac{a'(\phi)}{a(\phi)^2} b(\theta_i) - \frac{a'(\phi)}{a(\phi)} + c'(0, \phi) \right] \frac{\partial b(\theta_i)}{\partial \theta_i} r_{\pi_i} g_i'
\]

\[
\frac{\partial^2 P(Y_i = 0)}{\partial \phi \partial \beta} = a(\phi)^{-1} \exp\{d(\theta_i, \phi)\} \left[ \frac{a'(\phi)}{a(\phi)^2} b(\theta_i) - \frac{a'(\phi)}{a(\phi)} + c'(0, \phi) \right] \frac{\partial b(\theta_i)}{\partial \theta_i} r_{\mu_i} x_i'
\]

\[
\frac{\partial^2 P(Y_i = 0)}{\partial \beta' \beta} = a(\phi)^{-1} \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} r_{\mu_i}^2 x_i x_i'
\]

\[
\frac{\partial^2 P(Y_i = 0)}{\partial \phi \partial \gamma} = a(\phi)^{-1} \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} r_{\pi_i}^2 g_i g_i'
\]

\[
\frac{\partial^2 P(Y_i = 0)}{\partial \gamma' \beta} = a(\phi)^{-1} \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} t_i s_i g_i x_i'
\]

\[
\frac{\partial^2 P(Y_i = 0)}{\partial \phi^2} = \left[ E(Y_i) \theta_i - b(\theta_i) \right] \frac{2a'(\phi)^2 - a(\phi)a''(\phi)}{a(\phi)^3} + c''(y_i, \phi)
\]

\[
\frac{\partial^2 P(Y_i = 0)}{\partial \phi \partial \gamma} = 0
\]

\[
\frac{\partial^2 P(Y_i = 0)}{\partial \phi \partial \beta} = 0.
\]
Chapter 4

Marginal zero-inflated models for clustered data

\footnote{Marginal zero-inflated models for clustered data. J. Martin and D.B. Hall. To be submitted to Journal of Applied Statistics.}
Abstract

A common regression model to analyze data with excess zeros is the zero-inflated regression model (Lambert, 1992), which mixes two distributions, a degenerate distribution with point mass zero and a non-degenerate distribution with mean $\lambda_i$. These models have since been extended to include clustered or longitudinal data by including random effects in either a single component or both (Hall, 2000; Hur et al., 2002). The estimates from this model quantify the effects of the covariates on the mixing probabilities and the mean of the non-degenerate distribution. However, often the effect of the covariates on the overall mean of the response is of interest. Marginal zero-inflated (MZI) models have been introduced that accomplishes this aim (Long et al., 2014; Martin and Hall, 2016). This paper extends the MZI model framework to include random effects in both of the linear predictors when the non-degenerate distribution is a member of the exponential dispersion family. Described therein are methods used to estimate the parameters, either via direct function optimization or the EM algorithm, properties of the maximum likelihood estimators, and and advantages of the mixed MZI (MMZI) model compared to the ZI mixed model. An example is presented to illustrate the advantages of the MMZI model in practical data analysis.

Keywords: exponential dispersion family; generalized linear models; zero inflation; marginalized models; EM algorithm; longitudinal data; random effects

4.1 Introduction

A common phenomenon in count data is an excess of observed zero values relative to standard distributions such as the Poisson, binomial and negative binomial. In recent years there has been much research focused on regression models for such excess zero (EZ) data. The two most common frameworks are the zero-inflated (ZI) regression model (Lambert, 1992; Hall, 2000) and the hurdle regression model (Mullahy, 1986). Both assume a mixture of
two probability distributions; one a degenerate distribution at zero and the other a standard probability distribution (ZI regression) or a truncated probability distribution (hurdle regression). The dependence on covariates for both models is built in through generalized linear model (GLM)-like specifications for the non-degenerate distribution’s mean and the mixing probability. However, both of these are means for latent random variables rather than the observed response, which is often the main interest. Recently, Martin and Hall (2016), building on work of Long et al. (2014) and Preisser et al. (2015), introduced the marginal ZI (MZI) regression model, which relates the marginal mean response directly to covariates rather than modeling the mean of the latent non-degenerate random variable in the mixture. This approach offers advantages in terms of interpretation and inference when scientific interest focuses on the mean response (e.g., when one wants to compare the mean response across treatments). In this paper we extend the MZI model to the clustered data context.

Clustered data (e.g., longitudinal, repeated measures, panel data) with EZs occur in many disciplines, such as health (Yau and Lee 2001, Albert et al. 2014), horticulture (van Iersel et al. 2000), and sociology (Buu et al. 2012). To account for within-subject correlation, Hall (2000) extended Lambert’s ZI Poission (ZIP) regression model to include a subject specific random slope in the model for the latent Poisson distribution. Hur et al. (2002) and Wang et al. (2002) add subject-specific random effects to the linear predictor for the mixing probabilities as well. Analogously, the goal of this paper is to formulate mixed MZI models for clustered data. We describe maximum likelihood estimation in this class of models via the EM algorithm and explore the properties of the ML estimators via simulation. We also highlight the advantage of a marginal mixed-effect ZI model relative to a traditional mixed-effect ZI (MMZI) model when it comes to marginalizing over the distribution of the random effects. For many purposes, the former class of models leads to more interpretable subject-specific and population averaged inferences than the latter class.
4.2 Marginal Zero-Inflated Regression

4.2.1 Independent Data

We first consider the marginal ZI regression model in the context of independent (e.g., cross-sectional) data. Suppose response data \( y_1, ..., y_n \) are available corresponding to independent random variables \( Y_i, i = 1, ..., n \), each of which follows a mixture distribution depending on covariate data \( x_i \). In particular, we assume that, conditional on \( x_i \),

\[
Y_i \sim \begin{cases} 
0 & \text{with probability } \pi_i \\
F_{ND} & \text{with probability } 1 - \pi_i
\end{cases} \tag{4.1}
\]

where \( F_{ND} \) is a non-degenerate (ND) distribution in the exponential dispersion family with corresponding density function \( f_{ND}(y_i; \theta_i, \sigma) \) with canonical location parameter \( \theta_i \), mean \( \lambda_i \), and scale parameter \( \sigma \). As in ZI regression, we assume that the mixing probabilities \( \{\pi_i\} \) may depend on covariates through a GLM-like specification (e.g., a logit-linear form), but instead of modeling covariate effects on \( \lambda_i \), we link covariates directly to \( \mu_i \equiv E(Y_i|x_i) \), which is often of greater interest.

Given this mixture structure, the probability mass function (pmf) of the MZI model is

\[
f_{MZI}(y_i; \gamma, \beta, \delta) = \{1 - \pi_i(\gamma)\} \left[ \frac{\pi_i(\gamma)}{1 - \pi_i(\gamma)} + \exp \left\{ \frac{b[\theta_i(\gamma, \beta)]}{a_i(\sigma)} + c(0, \sigma) \right\} \right]^{u_i} \\
\left[ \exp \left\{ \frac{y_i \theta_i(\gamma, \beta) + b[\theta_i(\gamma, \beta)]}{a_i(\sigma)} + c(y_i, \sigma) \right\} \right]^{1-u_i} \tag{4.2}
\]

where

\[
g_\pi[\pi_i(\gamma)] = g_i \gamma \\
g_\mu[\mu_i(\beta)] = x_i \beta,
\]

72
$g_\pi(.)$ and $g_\mu(.)$ are suitable link functions, and $u_i = 1$ if $y_i = 0$ and 0 otherwise. Here our notation $\pi_i(\gamma)$ and $\mu_i(\beta)$ indicates the dependence of these quantities on regression parameters $\gamma$ and $\beta$, respectively, which quantify effects of covariates in the vectors $g_i$ and $x_i$. For interpretability’s sake (Martin and Hall, 2016), we assume covariates in the former vector are a subset of those in $x_i$, though this may be relaxed. Note that $\theta_i$ is a function of $\lambda_i$, the mean of $f_{\text{ND}}(y_i; \theta_i, \sigma)$, so it is a function of both $\gamma$ and $\beta$ through the relationship between $\lambda_i$ and $\mu_i$, which is $\mu_i = (1 - \pi_i)\lambda_i$.

A function optimizer can be used to find the maximum likelihood estimate (MLE) of $\Psi = (\gamma', \beta', \sigma)'$. This process can be computationally taxing if the dimension of $\Psi$ is large. Alternatively, the EM algorithm can be used to find $\hat{\Psi}$. Let $z_i$ be the unobserved variable indicating that $y_i$ is from the degenerate distribution. Let $f_{\text{ND}}(y_i; \Psi)$ be the latent distribution, then the complete data loglikelihood for the MZI model is

$$
\ell_1^c(\Psi; y, z) = \sum_{i=1}^n \left\{ z_i \log \left( \frac{\pi_i}{1 - \pi_i} \right) + \log(1 - \pi_i) + (1 - z_i) \log f_{\text{ND}}(y_i; \Psi) \right\}.
$$

Because this quantity is linear in the missing data $z$, the expression to be maximized iteratively in the EM algorithm is simply $Q(\Psi|\Psi^{(r)}) \equiv \ell_1^c(\Psi; y, z^{(r+1)})$ where $z^{(r+1)} = E(z|y; \Psi^{(r)})$ and $\Psi^{(r)}$ is the parameter estimate after the $r$th step of the algorithm. Unlike the expected complete data loglikelihood for the ordinary ZI regression model, $Q(\Psi|\Psi^{(r)})$ does not decompose into two terms that can be maximized separately with respect to each regression parameter. However, a profiled optimization procedure still offers some computational convenience and efficiency. In particular, if we first maximize $Q$ with respect to $\gamma$ for fixed $\beta$ and plug the resulting estimate back into $Q$, then the remaining maximization with respect to $\beta$ can be accomplished by fitting a standard GLM (see Martin and Hall, 2016, for specific cases).
4.2.2 Clustered Data

The MZI models can be extended to include random effects that account for the dependence of the observations within clusters. Let \( y_{ij} \) be the response for the \( j^{th} \) observation of the \( i^{th} \) cluster for \( i = 1, \ldots, m \) and \( j = 1, \ldots, n_i \) where \( N = \sum_{i=1}^{m} n_i \), the total number of observations. Similarly, \( g_{ij} \) and \( x_{ij} \) are the vectors of covariates for the mixing probabilities \( \pi_{ij} \) and overall mean \( \mu_{ij} \), respectively. As in section 4.2.1, \( \gamma \) and \( \beta \) are the vectors of parameters with lengths \( p \) and \( r \) that are associated with \( g_{ij} \) and \( x_{ij} \). The MMZI model is specified similarly as in (4.1), conditional on cluster specific random effects \( b_i \); that is, a degenerate mixture as in (4.1) defines the conditional distribution of \( Y_{ij} \mid b_i \). This leads to a probability density function (pdf) of the form

\[
\begin{align*}
\mathcal{f}_{Y_{ij} \mid b_i}(y_{ij} \mid b_i; \Psi) &= \{1 - \pi_{ij}(\gamma)\} \left[ \frac{\pi_{ij}(\gamma)}{1 - \pi_{ij}(\gamma)} + \exp \left\{ \frac{b[\theta_{ij}(\gamma, \beta)]}{a(\delta)} + c(0, \delta) \right\} \right]^{u_{ij}} \\
& \quad \left[ \exp \left\{ \frac{y_{ij}\theta_{ij}(\gamma, \beta) + b[\theta_{ij}(\gamma, \beta)]}{a(\delta)} + c(y_{ij}, \delta) \right\} \right]^{1-u_{ij}},
\end{align*}
\]

where

\[
\begin{align*}
\eta_{\pi_{ij}} &= g_{\pi}(\pi_{ij}(\gamma)) = g'_{ij}\gamma + d'_{1ij} \Sigma_1^{\frac{1}{2}} b_{1i} \\
\eta_{\mu_{ij}} &= g_{\mu}(\mu_{ij}(\beta)) = x'_{ij}\beta + d'_{2ij} \Sigma_{12}^{\frac{1}{2}} b_{1i} + d'_{3ij} \Sigma_2^{\frac{1}{2}} b_{2i},
\end{align*}
\]

(4.3)

where \( \Psi \) is the vector of all model parameters. The random effects \( b_i = (b'_{1i}, b'_{2i})' \) are of length \( q = q_1 + q_2 \) and are independently normally distributed \( b_i \sim N(0, I) \). Vectors \( d'_{1ij}, d'_{2ij}, \) and \( d'_{3ij} \) are the rows of the model matrices for the random effects. This parameterization implies a covariance structure on \( \eta_{\pi_{ij}} \) and \( \eta_{\mu_{ij}} \) such that
\[
\begin{align*}
\text{Var}(\eta_{\pi ij}) &= d_{1ij}^\prime \Sigma_1 d_{1ij} \\
\text{Var}(\eta_{\mu ij}) &= d_{2ij}^\prime \Sigma_{12} d_{2ij} + d_{3ij}^\prime \Sigma_2 d_{3ij} \\
\text{Cov}(\eta_{\pi ij}, \eta_{\mu ij}) &= d_{1ij}^\prime \Sigma_1^2 \Sigma_{12}^2 d_{2ij}.
\end{align*}
\]

We assume $\Sigma_1, \Sigma_2, \Sigma_{12}$ are parameterized by vector-valued parameters $\sigma_1, \sigma_2$ and $\sigma_{12}$, respectively, so that the combined vector of model parameters is $\Psi = (\gamma', \beta', \sigma')'$, where $\sigma = (\sigma_1', \sigma_2', \sigma_{12}')'$. For more details, see Hall and Wang (2005).

The parameters $\Psi$ can be estimated using the maximum likelihood (ML) method. The density of $y_i | b_i$ is the product of that of each of the $n_i$ observations in the $i^{th}$ cluster

\[
f_{Y_i | b_i}(y_i | b_i; \Psi) = \prod_{j=1}^{n_i} f_{Y_{ij} | b_i}(y_{ij} | b_i; \Psi).
\]

Integrating over the random effects distribution yields the marginal probability density for $y_i$,

\[
f_{Y_i}(y_i; \Psi) = \int_{b_i} f_{Y_i | b_i}(y_i | b_i; \Psi) \phi(b_i) db_i,
\]

where $\phi(.)$ is the $q$-dimensional multivariate standard normal density. The full loglikelihood is the sums of the marginal loglikelihoods $\ell_2(\Psi; y) = \sum_{i=1}^{m} \log f_{Y_i}(y_i)$. There are many different techniques that can be used to numerically approximate the integral; a relatively simple approach is Gauss-Hermite quadrature. Maximizing $\ell_2$ provides the maximum likelihood estimates, which can be done with any one of a variety of different function optimization methods.

A particularly convenient way to do the optimization is with the EM algorithm. This approach was described by Martin and Hall (2016) for the MZI regression case, and we
extend it now for the MMZI version of the model. Again, let $z_{ij} = 1$ if $y_{ij}$ is from the degenerate distribution and 0 otherwise. To formulate the EM algorithm, consider $z_{ij}, b_i$ for all $i, j$ as missing data. The complete data loglikelihood is

$$f_c^2(\Psi; y, b, z) = \sum_{i=1}^{m} \log \phi(b_i) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log f_{Y_{ij}, z_{ij}|b_i}(y_{ij}, z_{ij}|b_i; \Psi),$$

where $\log f_{Y_{ij}, z_{ij}|b_i}(y_{ij}, z_{ij}|b_i; \Psi) = \log(1 - \pi_{ij}) + z_{ij}\logit(\pi_{ij}) + (1 - z_{ij}) \log f_{ND}(y_{ij}|b_i; \Psi)$ where $f_{ND}$ is the non-degenerate component of the mixture model for $Y_{ij}|b_i$. Apart from terms that do not depend on model parameters, the M step objective function at the $r + 1$st step of the EM algorithm is

$$Q(\Psi|\Psi^{(r)}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} E \left\{ E \left[ \log f_{Y_{ij}, z_{ij}|b_i}(y_{ij}, z_{ij}|b_i, \Psi)|y, b, \Psi^{(r)} \right] |y, \Psi^{(r)} \right\}. $$

Here, the inner expectation is with respect to $z_{ij}$, which enters the quantity linearly, so this step can be accomplished by replacing $z_{ij}$ with its conditional expectation,

$$z_{ij}^{(r+1)}(b_i) = E(z_{ij}|y, b, \Psi^{(r)}) = \frac{\pi_{ij}^{(r)} u_{ij}}{\pi_{ij}^{(r)} + (1 - \pi_{ij}^{(r)}) f_{ND}(y_{ij}|b_i; \Psi^{(r)})}. $$

The outer expectation has form
\[
Q(\Psi | \Psi^{(r)}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \int_{\mathcal{R}} \log f_{Y_{ij}, Z_{ij}}^c(y_{ij}, z_{ij}^{(r+1)} | b; \Psi) f_{Y_i | b_i}(y_i | b_i; \Psi^{(r)}) \phi(b_i) d\mathbf{b}
\]

\[
\int_{\mathcal{R}} f_{Y_i | b_i}(y_i | b_i; \Psi^{(r)}) \phi(b_i) d\mathbf{b},
\]

where the integral is calculated using \(l\)-point Gaussian quadrature and then maximized with respect to \(\Psi\). The form of \(Q(\Psi | \Psi^{(r)})\) can be separated into two parts which allows this maximization to be done in two steps. First \(Q(\Psi | \Psi^{(r)})\) is maximized with respect to \(\gamma, \sigma_1\) using a generic optimizer to calculate \(\hat{\gamma}^{(r+1)}, \hat{\sigma}_1^{(r+1)}\). Then \(\hat{\beta}^{(r+1)}, \hat{\sigma}_2^{(r+1)}, \hat{\sigma}_{12}^{(r+1)}\) are calculated via a weighted GLM. To fix ideas, in the next section we provide details of the implementation in the case in which there are random intercepts in each linear predictor. The general case follows in a straightforward, but notationally tedious way from the case given below.

**Random intercepts**

Assume a model where only random intercepts are included, i.e. \(\eta_{\pi_{ij}} = g'_{ij} \gamma + \sigma_1 b_{1i}\) and \(\eta_{\mu_{ij}} = x'_{ij} \beta + \sigma_1 b_{1i} + \sigma_2 b_{2i}\). Using \(l\)-point Gaussian quadrature, \(\mathbf{b}_i = (b_{1i}, b_{2i})'\) are replaced with abscissas \(\mathbf{b}_\ell = (b_{1\ell}, b_{2\ell})'\) which have weights \(w_\ell = (w_{\ell_1}, w_{\ell_2})'\); then \(Q(\Psi | \Psi^{(r)})\) can be approximated as

\[
Q(\Psi | \Psi^{(r)}) \approx \sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{\ell_1=1}^{l} \sum_{\ell_2=1}^{l} \log f_{Y_{ij}, Z_{ij}}^c(y_{ij}, z_{ij}, \mathbf{b}_\ell; \Psi) f_{Y_i | b_i}(y_i | b_i; \Psi^{(r)}) w_{\ell_1} w_{\ell_2}
\]

\[
\sum_{\ell_1=1}^{l} \sum_{\ell_2=1}^{l} f_{Y_i | b_i}(y_i | b_i; \Psi^{(r)}) w_{\ell_1} w_{\ell_2}
\]
M Step for $\gamma, \sigma_1$. The objective function is maximized using a numerical optimizer with respect to $(\gamma', \sigma_1')$ while $(\beta', \sigma_{12}, \sigma_2')$ are held constant at $(\hat{\beta}^{(r)}, \hat{\sigma}_{12}^{(r)}, \hat{\sigma}_2^{(r)})$ to calculate $\hat{\gamma}^{(r+1)}$ and $\hat{\sigma}_1^{(r+1)}$.

M Step for $\beta, \sigma_{12}, \sigma_2$. Let $\pi^*, \mu^*$ be

\[
\begin{align*}
g_\pi(\pi^*) &= G^* \gamma^{(r+1)} + \sigma_1^{(r+1)} b_1^* \\
g_\mu(\mu^*) &= X^* \beta^{(r)} + \sigma_{12}^{(r)} b_1^* + \sigma_2^{(r)} b_2^*,
\end{align*}
\]

and $y^* = y \otimes 1_{l_2}$, $G^* = G \otimes 1_{l_2}$, $X^* = X \otimes 1_{l_2}$, and

\[
b^* = (b_1^*, b_2^*) = 1_N \otimes (1_l \otimes (b_{\ell_1}, b_{\ell_2}, ..., b_{\ell_t})', (b_{\ell_1}, b_{\ell_2}, ..., b_{\ell_t})' \otimes 1_l),
\]

where $\otimes$ is the Kronecker product. Maximization of $Q(\Psi | \Psi^{(r)})$ over $\beta$, $\sigma_{12}$, and $\sigma_2$ can be done by fitting a weighted GLM with response $ay^*$, design matrix $(X^*, b_1^*, b_2^*)$, and weights,

\[
\omega_{ij\ell_1\ell_2} = \xi_{ij\ell_1\ell_2} (r+1) \frac{\prod_{j=1}^{n_l} f_{Y_{ij} | b_{ij}} (y_{ij} | b_{ij}; \Psi^{(r)}) w_{\ell}}{\sum_{\ell_1=1}^{l_1} \sum_{\ell_2=1}^{l_2} \prod_{j=1}^{n_l} f_{Y_{ij} | b_{ij}} (y_{ij} | b_{ij}; \Psi^{(r)}) w_{\ell}},
\]

where $\xi_{ij\ell_1\ell_2} (r+1)$ is a function of $z_{ij\ell_1\ell_2}^{(r+1)}$ and $\pi_{ij\ell_1\ell_2}^{(r+1)}$, the form of which depends on the choice of non-degenerate distribution. Like the weights, the specific forms of $a$ and $g_\mu(\mu^*)$ also depend on the choice of non-degenerate distribution (see Martin and Hall, 2016, for specifics).

The E step and both M steps are then repeated until convergence. This approach can be extended for additional random effects, such as random slopes, but of course the computational demands grow quickly with $q$. \[78\]
Subject Specific vs. Population Average Parameter Interpretations

To this point we have used the term “marginalized” in the sense of Long et al. (2014) and Martin and Hall (2016), where the marginalization involves averaging across the finite mixture of the degenerate and non-degenerate distributions. However, the terms “marginal” and “marginalized” are also used in the literature on generalized linear models for clustered data. In particular, a distinction is made between mixed-effect models, which are specified conditionally on a (possibly vector-valued) random effect $b_i$, and which therefore involve covariate effects on a conditional mean $E(y_i|b_i, x_i)$, and marginal models, which make assumptions directly on the marginal moments $E(y_i|x_i)$ and $\text{var}(y_i|x_i)$. As noted by several authors, beginning with Zeger et al. (1988), regression parameters have different interpretations in these two model classes. For mixed models, regression effects have a subject specific (SS) interpretation and for marginal models the regression parameters have a population averaged (PA) interpretation. A PA coefficient $\beta_k$ is interpreted as the effect of $x_{kij}$ between two different populations; e.g., the coefficient on smoking status in a marginal logistic model corresponds to the log odds ratio of lung disease for smokers compared to non-smokers. A PA interpretation can be obtained from a mixed effect model by marginalizing over the random effects distribution via the convolution equation

$$E(y_i|x_i) = E\{E(y_i|b_i, x_i)\}. \quad (4.5)$$

However, for models that do not specify a linear relationship between $E(y_i|b_i, x_i)$ and the covariate effects, this equation may be difficult to solve. In addition, constant effects (on the link scale) in a mixed model do not necessarily lead to constant effects after marginalization. This has led to a large literature on so-called marginalized mixed effect models (Heagerty, 1999; Heagerty et al., 2000) in which a model is specified conditionally as a mixed effect model, and where (4.5) is used to marginalize the model and obtain a regression coefficient
with a PA interpretation from a mixed effect model. Typically this marginalization involves complex and computationally expensive calculations. However, with both the identity link and log link, marginalization is very simple. In the log link cases, a model for \( \log(\mu_{ij}) \) that is specified conditionally on random effects as in (4.4), leads to marginal mean \( E(Y_i) = \exp\{x_i'\beta + \frac{1}{2}(d_{2ij}'\Sigma_{12}d_{2ij} + d_{3ij}'\Sigma_{2}d_{3ij})\} \). Thus, in an MMZI model with log link for \( g_{\mu} \), \( \beta \) has both a PA and SS interpretation.

Mixed effect models for longitudinal ZI data (e.g., ZIP-mixed models, mixed effect hurdle models), like other mixed models with a log link for the mean of the non-degenerate distribution, can provide parameter estimates with both SS and PA interpretations. However, if these models were to provide an marginal interpretation of the estimates for the overall mean, marginalizing over the random effects and the finite mixture, the task becomes complex as the relationship between the model parameters and the overall mean is \( E(y_{ij}|b_i, x_{ij}) = [1 - \pi_{ij}(b_{1i})] \mu_{ij}(b_{2i}) \). For the mixed effect version of the MZI models with a log link for \( \mu_{ij} \) the regression coefficients have dual SS and PA interpretations without the need to marginalize over the random effects distribution or even the finite mixture. This is a key advantage of the marginal MZI formulation.

### 4.3 Simulation Study

The first goal of the simulation study is to quantify the bias and variance of the MLEs. We let the covariate vector be \( x_{ij} = (1, x_{1i}, x_{2ij})' \) where \( x_{1i} \) is binary and \( x_{2ij} \sim \text{Uniform}(-1, 1) \). The data were generated from a MMZIP model in which \( \pi_{ij} = \logit^{-1}(x_{ij}'\gamma + b_{1i}\sigma_1) \) and \( \mu_{ij} = \exp(x_{ij}'\beta + b_{1i}\sigma_{12} + b_{2i}\sigma_2) \), where \( \gamma = (\logit(.2), -.5, 0)' \), \( \beta = (\log(2.5), .75, -.5)' \), and \( \sigma = (\sigma_1, \sigma_{12}, \sigma_2)' = (.3, -.25, .3) \). As before, \( \{b_i\} \) were bivariate standard normal. Five hundred data sets each with cluster sizes \( n_i = n = 5 \) were generated. The procedure was repeated for two choices of \( m \), \( m = 50 \) and \( m = 100 \). In half of the clusters we set \( x_{1i} = 1 \)
and in the other half \( x_{1i} = 0 \). A correctly specified mixed MZIP model was fit and four different methods were used to numerically calculate the integral: Gaussian-Hermite with quadrature points \( l = 5, 10 \), GH quadrature with data-driven choice of \( l \), as implemented in NLMIXED in SAS (denoted \( l_{DD} \) in Table 4.1), and adaptive Gaussian quadrature (denoted AG). For the last two quadrature methods SAS starts at \( l = 1 \) and calculates the objective function at the starting values, then increases the number of quadrature points by one until the difference between the current and previous value is below a certain threshold. The starting values for PROC NLMIXED were chosen to be the true values from which the data were generated.

From these models the point estimates and standard errors were recorded for \( \Psi = (\gamma, \beta, \sigma)' \). The percent median relative bias (PMRB) was calculated as \( 100 \times (\hat{\Psi}_i - \Psi_i) / \Psi_i \), simulated standard error (SSE), and model based standard error (MBSE) are recorded in Table 4.1. For \( \gamma_2 \), the PMRB is calculated as \( 100 \times (\hat{\gamma}_2 - \gamma_2) \).

The average number of quadrature points chosen by NLMIXED for \( l_{DD} \) for \( m = 50 \), 100 was 6.96 (sd=2.07) and 6.35 (sd=2.13), respectively. Similarly for AG, the average number of quadrature points was 1.03 (sd=.25) for \( m = 50 \) and 1 (sd=.09) for \( m = 100 \). From the PRMB values in Table 4.1, it seems that SAS chooses too few quadrature points to accurately estimate the variance components. The percent bias for the fixed effects is small for both sample sizes and quadrature points, and the empirical bias becomes smaller as \( m \) increases but increasing the number of quadrature points seem to have little effect. However, the accuracy of the estimates for the variance components is poor when either \( m = 50 \) or \( l = 5 \). If the accuracy of these parameters is important, a large sample and high number of quadrature points are necessary. For the SSE and MBSE, Table 4.1 shows that the standard errors calculated from the model tend to slightly underestimate the standard error of the parameter estimates calculated from the simulation. Additionally, the number of quadrature
| Parameters |  |  |  |  |  |  |  |  |  |  |  |  |  |
|------------|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **PRMB**  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 5 2.51 | 7.30 | 0.14 | 0.63 | 0.58 | 0.46 | 16.70 | 14.90 | 20.60 |
| 100 5 1.29 | 5.43 | 0.40 | 0.41 | 0.05 | 0.35 | 9.14 | 10.70 | 9.51 |
| 50 10 3.59 | 6.01 | 0.12 | 0.28 | 1.45 | 0.31 | 12.20 | 5.57 | 7.75 |
| 100 10 1.56 | 5.26 | 0.07 | 0.27 | 0.64 | 0.12 | 1.81 | 1.14 | 4.37 |
| 50 l_DD 3.37 | 6.98 | 0.72 | 0.68 | 0.19 | 0.50 | 12.40 | 4.10 | 12.00 |
| 100 l_DD 1.63 | 4.91 | 0.13 | 0.30 | 0.71 | 0.23 | 0.50 | 6.39 | 6.45 |
| 50 AG 4.32 | 6.26 | 0.67 | 0.64 | 0.13 | 0.43 | 19.70 | 26.60 | 97.00 |
| 100 AG 3.28 | 4.59 | 0.40 | 0.29 | 0.11 | 0.01 | 21.00 | 11.60 | 13.10 |
| **SSE**  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 5 0.31 | 0.41 | 0.40 | 0.11 | 0.15 | 0.07 | 0.33 | 0.17 | 0.14 |
| 100 5 0.22 | 0.29 | 0.24 | 0.07 | 0.10 | 0.05 | 0.27 | 0.14 | 0.13 |
| 50 10 0.32 | 0.41 | 0.40 | 0.11 | 0.15 | 0.07 | 0.37 | 0.15 | 0.12 |
| 100 10 0.22 | 0.29 | 0.24 | 0.08 | 0.10 | 0.05 | 0.28 | 0.12 | 0.09 |
| 50 l_DD 0.31 | 0.41 | 0.40 | 0.11 | 0.15 | 0.07 | 0.35 | 0.16 | 0.14 |
| 100 l_DD 0.22 | 0.29 | 0.24 | 0.07 | 0.10 | 0.05 | 0.28 | 0.12 | 0.10 |
| 50 AG 0.32 | 0.41 | 0.40 | 0.11 | 0.14 | 0.07 | 0.37 | 0.18 | 0.16 |
| 100 AG 0.23 | 0.29 | 0.25 | 0.07 | 0.10 | 0.05 | 0.31 | 0.15 | 0.17 |
| **MBSE**  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 5 0.30 | 0.41 | 0.35 | 0.10 | 0.13 | 0.07 | 0.33 | 0.11 | 0.12 |
| 100 5 0.21 | 0.28 | 0.24 | 0.08 | 0.10 | 0.05 | 0.26 | 0.10 | 0.10 |
| 50 10 0.30 | 0.40 | 0.35 | 0.10 | 0.13 | 0.07 | 0.34 | 0.09 | 0.08 |
| 100 10 0.21 | 0.29 | 0.24 | 0.08 | 0.10 | 0.05 | 0.26 | 0.08 | 0.06 |
| 50 l_DD 0.30 | 0.41 | 0.35 | 0.10 | 0.13 | 0.07 | 0.26 | 0.08 | 0.09 |
| 100 l_DD 0.20 | 0.28 | 0.24 | 0.08 | 0.10 | 0.05 | 0.26 | 0.08 | 0.07 |
| 50 AG 0.31 | 0.41 | 0.35 | 0.10 | 0.14 | 0.07 | 0.26 | 0.07 | 0.34 |
| 100 AG 0.22 | 0.29 | 0.24 | 0.08 | 0.10 | 0.05 | 0.22 | 0.09 | 0.21 |

Table 4.1: Simulation results from the MZIP model. PRMB: Percent relative median bias, SSE: Simulation standard deviation, MBSE: Model based standard error
points seem to have no effect on the standard errors for the fixed effects and a minor effect for the variance components in \( \mu_{ij} \).

The second goal of the simulation study is to show when the random effects for the mixing probability are misspecified, the MLEs for parameters related to the marginal mean in the MMZI model have less bias. Three different models were used to generate the data sets:

- **Model 1**: \( \logit(\pi_{ij}) = g'_{ij} \gamma + b_{1i} \sigma_1 + b_{3i} x_{2ij} \sigma_3 \), \( \log(\mu_{ij}) = x'_{ij} \beta + b_{1i} \sigma_{12} + b_{2i} \sigma_2 \)
- **Model 2**: \( \logit(\pi_{ij}) = g'_{ij} \gamma + b_{1i} \sigma_1 + b_{3i} x_{2ij} \sigma_3 \), \( \log(\lambda_{ij}) = x'_{ij} \beta + b_{1i} \sigma_{12} + b_{2i} \sigma_2 \)
- **Model 3**: \( \text{probit}(\pi_{ij}) = \delta_{ij} + b_{1i} \sigma_1 + b_{3i} x_{2ij} \sigma_3 \), \( \log(\lambda_{ij}) = x'_{ij} \beta + b_{1i} \sigma_{12} + b_{2i} \sigma_2 \)

where \( \delta_{ij} = \sqrt{1 + \sigma_1^2 + \sigma_2^2 x_{2ij}^2} \Phi^{-1}[\expit(g'_{ij} \gamma)] \), the covariate vectors \( g_{ij} = x_{ij} = (1, x_{1i}, x_{2ij})' \), and \( x_{1i} \) was binary and \( x_{2ij} \sim \text{Unif}(-1, 1) \). The choice of link function for Model 3 is in order to make a more direct comparison, calculating the population average value of the mixing probability, \( \pi_{ij}^m = \expit(g'_{ij} \gamma) \). There is no closed form solution between \( \pi_{ij} \) and \( \pi_{ij}^m \) when both linear predictors use a logit link (as in Model 2), while a closed form solution does exist when the link functions probit and logit are used for \( \pi_{ij} \) and \( \pi_{ij}^m \), respectively (Griswold and Zeger, 2004). As mentioned in section 4.2.2, when the log link is used, \( \beta \) for the MMZIP regression model has both a SS and PA interpretation.

The values for \( \beta \), \( \sigma_1 \), \( \sigma_{12} \), and \( \sigma_2 \) were the same as the previous simulation and \( \gamma = (-1.386, -0.5, .5) \) and \( \sigma_3 = 2.5 \). Five hundred data sets with cluster size 50, 100, and 500 \( (N=250, 500, 2500) \) were generated from these models. Three models were fit to their corresponding data sets, where the fitted models were correctly specified except for the omission of term involving the random effect \( b_{3i} \). From the estimates of the fitted models, the median of the ratio \( \hat{R} = \hat{\mu}_1^*/\hat{\mu}_2^* \) for both models was calculated when \( \hat{\mu}_1^* \) is calculated at \( x_{1i} = x_{2ij} = 1 \) while \( \hat{\mu}_2^* \) is calculated at \( (x_{1i}, x_{2ij}) = (0, 1) \). For the MMZIP models, \( \hat{R} = \exp(\hat{\beta}_1) \), which does not depend on the value of \( \hat{\pi}_1 \) or the variance components. For the
ZIP mixed-effect model, $\hat{R} = \frac{\hat{\pi}_m}{1 - \hat{\pi}_m} \exp(\hat{\beta}_1)$. The ratios of each model were divided by the ratio of $\mu_1^*$ and $\mu_2^*$ for the true model ($R_0$). The results are in Table 4.2.

<table>
<thead>
<tr>
<th>$m$</th>
<th>50</th>
<th>100</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>1.013</td>
<td>1.002</td>
<td>1.000</td>
</tr>
<tr>
<td>Model 2</td>
<td>0.989</td>
<td>0.970</td>
<td>0.983</td>
</tr>
<tr>
<td>Model 3</td>
<td>0.983</td>
<td>0.994</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 4.2: Ratio of the estimated mean and true mean ($\hat{R}/R_0$)

For the misspecified model, the median of $\hat{R}/R_0$ is close to 1 for all three models, but as the sample size grows, the bias for Model 2 does not decrease as it does for the other two models. These results suggest that, with respect to estimation of the marginal mean response, the MMZI model and marginalized ZI-mixed model are more robust to model misspecification than the ordinary ZI-mixed model.

4.4 Whitefly example

van Iersel et al. (2000) report the results of a horticultural experiment to examine the effect of six methods of applying pesticide to control silverleaf whiteflies on greenhouse-raised poinsettia. For each of twelve consecutive weeks a collection of whiteflies confined inside a cage was clipped on a leaf on each of 54 plants. Three weeks after the cages were removed, the number of immature whiteflies was recorded with the goal of determining which of the six treatments had the fewest immature whiteflies. Trios of plants formed the experimental unit, and these units were randomly assigned to treatments in three complete blocks. Because the three plants in each experimental unit constituted subsamples (or pseudo-replicates), the response variable, $y_{ijkl}$, was summed over these three plants, $y_{ij} = \sum_{k=1}^{3} y_{ijkl}$, where $i = 1, \ldots, 6$ is the treatment, $j = 1, 2, 3$ is the block, $k = 1, 2, 3$ is the plant, and $l = 1, \ldots, 12$ is the week. The six treatments used in the experiment included four methods of sub-irrigated pesticide
following 4, 2, 1 and 0 days without water (treatments 1-4 respectively), a treatment involving pesticide application via hand-watering (treatment 5), and a control treatment in which no pesticide was used (treatment 6).

Due to the high percentage of zeros and because the scientific interest centers on determining treatment effects on the overall mean, a MMZIP model is preferred over a ZIP mixed model. Let $\pi_{ijl}$ and $\mu_{ijl}$ be the mixing probability and marginal mean, respectively, for the $i^{th}$ treatment and $j^{th}$ block during the $l^{th}$ week. The model for the marginal mean with only main effects is $\log(\mu_{ijl}) = \beta_0 + \beta_{1i} + \beta_{2j} + \beta_{3l} + \beta_{4} \log(s_{ijl}) + \sigma_{1i} b_{1i} + \sigma_{2i} b_{2i}$, where $l$ is the week, $s_{ijl}$ is the number of live whiteflies placed on the plant, and $b_{1i}$ and $b_{2i}$ are independent standard normal random effects. The linear predictor for $\logit(\pi_{ijl})$ has the same main effects and only $\sigma_{1i} b_{1i}$ as a random effect. Table 4.3 below shows the results of including interaction terms for the linear predictor for $\mu_{ijl}$, where $p$ is the number of fixed effects in the model.

<table>
<thead>
<tr>
<th>Highest Interaction</th>
<th>p</th>
<th>-2 Log Likelihood</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block × trt + Block × week + trt × week</td>
<td>37</td>
<td>2914.5</td>
<td>2988.5</td>
<td>3062.1</td>
</tr>
<tr>
<td>Block × trt + Block × week</td>
<td>32</td>
<td>3054.6</td>
<td>3118.6</td>
<td>3182.2</td>
</tr>
<tr>
<td>Block × trt + trt × week</td>
<td>35</td>
<td>2949.6</td>
<td>3019.6</td>
<td>3089.2</td>
</tr>
<tr>
<td>Block × week + trt × week</td>
<td>27</td>
<td>2923.2</td>
<td>2977.2</td>
<td>3030.9</td>
</tr>
<tr>
<td>trt × week</td>
<td>25</td>
<td>2977.4</td>
<td>3027.4</td>
<td>3077.1</td>
</tr>
<tr>
<td>Block × week</td>
<td>22</td>
<td>3069.3</td>
<td>3113.3</td>
<td>3157.1</td>
</tr>
<tr>
<td>Block × trt</td>
<td>30</td>
<td>3089.1</td>
<td>3149.1</td>
<td>3208.8</td>
</tr>
<tr>
<td>None</td>
<td>20</td>
<td>3073.4</td>
<td>3113.4</td>
<td>3153.1</td>
</tr>
</tbody>
</table>

Table 4.3: Information criterion for the model shown with the largest interaction terms in the linear component for the marginal mean

The model with the lowest AIC and BIC includes both interaction terms for week and it will be the model used for the remaining analyses. Thus the model used in the hypothesis testing below is
\[
\begin{align*}
\text{logit}(\pi_{ijl}) &= \gamma_0 + \gamma_{1i} + \gamma_{2j} + \gamma_k + \sigma_1 b_{1i} \\
\log(\mu_{ijl}) &= \beta_0 + \beta_{1i} + \beta_{2j} + \beta_4 \log(s_{ijl}) + (\beta_3 + \beta_{5i} + \beta_{6j}) l + \sigma_{12} b_{1i} + \sigma_{2} b_{2i}.
\end{align*}
\]

Analyzing the plant specific random effects, we test to determine if each of the random effects are significant separately. The first null hypothesis is \( H_{0A} : \sigma_{12} = \sigma_2 = 0 \), the second is \( H_{0B} : \sigma_{12} = \sigma_1 = 0 \), and the test statistics and p-values are in Table 4.4. Since the variance components \( \sigma_1 \) and \( \sigma_2 \) are set to a boundary point, the test statistic follows a 50:50 mixture of a \( \chi^2_1 \) and \( \chi^2_2 \). In order to estimate the p-value from the test statistic, 1,000,000 values were generated from each distribution and the p-value is calculated as the proportion of these values that is above the test statistic. The results show that both are significant and that there is plant to plant variability for both the mixing probability and overall mean. The estimated covariance components are \( \hat{\sigma}' = (\hat{\sigma}_1, \hat{\sigma}_{12}, \hat{\sigma}_2) = (0.248, 0.234, 0.228) \), so the plant specific variability is fairly large and the correlation between the linear predictors for \( \pi_{ijl} \) and \( \mu_{ijl} \) is .716.

The main interest to the researchers is how the different methods of pesticide application affect the overall mean number of live immature whitefly larvae on the plants. Treatment 6, which has no pesticide, is set as the baseline \( (\beta_{16} = 0) \). All of the estimates for the effects of treatments 1 through 5 are strongly negative (-3.01 to -4.89), the slope for week is positive but weak (.07), and the interaction effects between treatment and week are small but positive (.17 to .24). These results suggest that the pesticide inhibits whitefly reproduction, but becomes somewhat less effective as time passes. There are several hypotheses that could be of interest: are there significant interaction effects between treatment and week \( (H_{0C}) \)? Among treatments involving pesticide, are the slopes on week all the same \( (H_{0D}) \)? Averaged
over week, is there a difference in the mean response between the subirrigation methods (on average) as compared with hand-watering ($H_{0E}$)? And, among the sub-irrigation methods, is the time since last watering important ($H_{0F}$)? The results for Wald tests of these four hypotheses may be found in Table 4.4.

<table>
<thead>
<tr>
<th>Null Hypothesis</th>
<th>Distribution</th>
<th>Test Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_{0A}$: $\sigma_{12} = \sigma_2 = 0$</td>
<td>$1/2(\chi^2_1 + \chi^2_2)$</td>
<td>97.99</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>$H_{0B}$: $\sigma_1 = \sigma_{12} = 0$</td>
<td>$1/2(\chi^2_1 + \chi^2_2)$</td>
<td>68.20</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>$H_{0C}$: $\frac{1}{5} \sum_{i=1}^{5} \beta_{5i} - \beta_{56} = 0$</td>
<td>Normal</td>
<td>9.25</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>$H_{0D}$: $\frac{1}{4} \sum_{i=1}^{4} \beta_{5i} - \beta_{55} = 0$</td>
<td>Normal</td>
<td>0.23</td>
<td>0.817</td>
</tr>
<tr>
<td>$H_{0E}$: $\frac{1}{4} \sum_{i=1}^{4} \beta_{1i} - \beta_{15} = 0$</td>
<td>Normal</td>
<td>-14.27</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>$H_{0F}$: $\beta_{11} = \beta_{12} = \beta_{13} = \beta_{14}$</td>
<td>$\chi^2_3$</td>
<td>22.07</td>
<td>&lt; .001</td>
</tr>
</tbody>
</table>

Table 4.4: Wald tests for the different treatment effects on the marginal mean.

The hypothesis test for the interaction effect ($H_{0C}$) agrees with the information criterion used to select the model, with a p-value close to 0. The next hypothesis test indicates that the slopes for the treatments with pesticide are not significantly different (p-value=.817) from one another. Since there is no significant interaction with week involving treatments 1 through 5, the model was refit with two different slopes for week: one for the control treatment and another for the treatments involving pesticide. Testing hand watering vs sub-irrigation ($H_{0E}$) as the method of pesticide application shows that there is a decrease in the mean number of live whiteflies when sub-irrigation is used. Lastly, a test determined that the time between watering is significant ($H_{0F}$), as treatment 3 was significantly lower than treatments 1 (p-value< .001) and 2 (p-value= .0015), and treatment 4 was significantly lower than treatment 1 (p-value= .0029). This suggests that the longer the time since the last watering, the fewer live immature whiteflies were on the plant.
4.5 Conclusion

This paper proposes a likelihood-based approach to estimate marginal zero-inflated regression models with random effects in the linear predictors for both the mixing probability and overall mean. Building on the MZI models for cross-sectional count data from Long et al. (2014), Preisser et al. (2015), and Martin and Hall (2016), we build a MZI model for clustered data with the inclusion of random effects and allow a correlation between the two components. This flexible framework allows for models to have a straightforward interpretations in terms of the marginal mean response while accounting for zero-inflation and within-cluster correlation. Additionally, we discussed two different methods of fitting the model, either via straightforward function optimization or using the EM algorithm, and for models with a high dimension parameter vector for the marginal mean, the EM algorithm tends to be faster than generic function optimization. An advantage of the MMZI model compared to the ZI mixed model is the regression parameter estimates for the marginal mean have both a subject-specific and population average interpretation with any choice of link function for the mixing probabilities.

The simulation results show that for Gauss-Hermite quadrature, the estimates for fixed effects exhibit little bias even for moderate sample size and few quadrature points, but accurate variance component estimation requires higher sample sizes and more quadrature points. An exploration of the theoretical properties of MLEs in this class of model compared to those of a ZI mixed model would be worthwhile to understand important factors that affect the estimation, especially for misspecified models. This inquiry might be able to explain the results in Table 4.1 that shows that adaptive Gaussian quadrature is not as accurate as Gauss-Hermite quadrature.

This paper primarily examines models where the log link is used for the linear predictor of the overall mean. This link has a closed form relationship between the conditional and
population averaged means giving the parameter estimates interpretations in terms of both quantities. If a different link is used, such as a logit link for the marginal probability of a MMZI binomial (MMZIB) model, such a straightforward relationship no longer holds true. However, if a probit link is used for the conditional probability and the logit link is used for the marginal probability, an equation linking the fixed effects of the conditional probability and the parameters of the marginal probability exists and allows for the parameters to be estimated without using a convolution equation. That is, a model where the regression parameters \( \beta \) have PA interpretations are of the form
\[
\text{logit}(\pi_{ij}) = g'_{ij} \gamma + d'_{1i} \Sigma_1 b_{1i},
\]
\[
\text{probit}(\mu_{ij}) = \delta_{ij} + d'_{2i} \Sigma_2 b_{2i},
\]
and
\[
\text{logit}(\mu_{ij}^m) = x'_{ij} \beta,
\]
where \( \mu_{ij} \) is the marginal probability of a success and \( \delta_{ij} = \sqrt{1 + d'_{2i} \Sigma_2 d_{2i}} + d'_{3i} \Sigma_3 b_{3i} \Phi^{-1}[\expit(x'_{ij} \beta)].\]
Chapter 5

Conclusion

In this dissertation, two main topics for zero-inflated models have been developed. The first is generalizing a coefficient of determination for ZI models from Cameron and Windmeijer’s (1997) work on Poisson regression models. The pseudo-$R^2$ measure we propose is a function of the deviances of the current and null models, which for exponential family distributions, are estimates of the Kullback-Leibler divergence. The $R^2_{ZI}$ term can be interpreted as the reduction in information from the full model to the model fitted to the data. This measure has many desirable properties and behaves similarly to the $R^2$ for linear regression. One of these traits is that it increases as predictors are added, even if they are unrelated to the response. We then extended the work of Mittlböck and Waldhör (2000) by adding a penalty for lack of parsimony in the fitted model which is based on the asymptotic distribution of the likelihood ratio test comparing the null and current models. As data are often overdispersed relative to a Poisson distribution, which artificially inflates the value of $R^2_{ZI}$ (Heinzl and Mittlböck 2003), we suggest using a ZINB model instead of a ZIP. The simulation results show that when the data are overdispersed, the ZINB model correctly compensates and the value of $R^2_{ZI}$ is near the generated population value, but when the data
are not overdispersed the value of $R^2_{ZI}$ is the nearly the same as the value of the equivalent ZIP model.

Although commonly used by practitioners, coefficients of determination are of limited use as measures of model fit, and are not suitable as general model selection criteria. Prior to undertaking the work summarized in this dissertation, we first intended to conduct research on model selection criteria for ZI regression. Initial simulations that we conducted indicated that there was room for improving on the performance of traditional criteria such as AIC and BIC for the ZI regression context. One approach that initially appeared promising is that of Naik et al. (2007) who introduced an information criterion akin to the corrected AIC (AICc) suitable for finite mixture regression models. They focused primarily on finite mixtures of normal regression models, but briefly discussed finite mixtures of generalized linear models, where they formulated such mixtures using quasilikelihood. Traditional methods such as AIC equally penalize the predictors in each component regardless of the probability that an observation was generated by that component. Additionally, these methods do not differentiate between model complexity arising from an additional component versus the inclusion of more covariates in the regression function. These two issues indicate that their criterion, known as the MRC, can be improved. In principle, the MRC is applicable to the zero inflation regression context, but Naik et al. only consider the case when the mixing probability is constant, whereas in the zero inflated regression the mixing probability is often assumed to depend upon covariates. Using methods similar to those underlying the AICc, Naik et al. suggested a criterion of the following form

$$MRC = \sum_{k=1}^{K} \left[ \hat{n}_k \log(\hat{\sigma}_k^2) + \frac{\hat{n}_k (\hat{n}_k + p_k)}{\hat{n}_k - p_k - 2} - 2 \hat{n}_k \log(\hat{\alpha}_k) \right],$$
where $K$ is the number of components, $p_k$ is the number of predictors for the mean in component $k$, $\hat{n}_k$ is the estimated number of observations from component $k$, and $\hat{\alpha}_k$ is the estimated mixture probability.

Naik et al. suggested that the MRC can be applied to non-Gaussian mixture models, such as GLMs, and used a quadratic approximation in their derivation. Following similar steps, we derived a criterion for a two component mixture model with a regression structure for the mixing probabilities, which is

$$
MRC_{nc} = \sum_{k=1}^{2} \left[ \hat{n}_k \log(\hat{\sigma}_k^2) + \frac{\hat{n}_k(\hat{n}_k + p_k)}{\hat{n}_k - p_k - 2} - 2 \sum_{i=1}^{n} \hat{\tau}_{ik} \log(\hat{\alpha}_{ik}) \right] + p_{\gamma},
$$

where $\hat{\sigma}_k^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_{ik})^2}{\text{Var}(y_i)}$, $\hat{\tau}_{ik}$ is the estimated posterior probability for observation $i$ being in the $k^{th}$ cluster, and $p_{\gamma}$ is the number of covariates in the mixing probability. This formulation reduced to the MRC when the mixing probabilities are constant, which is a very desirable result. However, when we tried empirically comparing this criterion to AIC and BIC, it fared no better in the simulations. We then designed and ran a simulation for the MRC for a mixture of two GLMs with constant mixing probabilities, which had similar results as those of the criterion we developed. This called into question the usefulness of the MRC as a model selection criterion for mixtures of GLMs, in general, and we concluded that it was perhaps not the best basis of deriving a model selection for ZI regression, in particular.

Since the criterion for two component mixtures with nonconstant mixing probabilities did not perform better than existing methods, the direction of the dissertation changed to focus on formulating a zero-inflated regression model for count data with excess zeroes where the predictors estimate the effect on the overall mean of the response variable. Building on the work of Long et al. (2014) and Preisser et al. (2015), we formulate a model where the mean
of the count distribution, $\lambda_i$, is replaced by $\mu_i/(1 - \pi_i)$, where $\mu_i$ is the mean of the response. In ZI regression, most models do not allow a direct analysis of the effect predictors have the overall mean. Not only can the MZI formulation estimate the effect of a continuous variable has on the overall mean of the response, but it allows direct inference on these quantities as well. Fitting the model via the EM algorithm is not as simple for the MZI as it is for the ZI model as the complete data loglikelihood does not separate into two different sums that can be maximized independently during the M step. However, for models with a small amount of parameters for the mixing probability compared to the parameters estimated in the linear predictor for overall mean, the EM algorithm can be computationally faster, especially for data with large sample sizes. Confidence intervals and hypotheses tests on the marginal mean for ZI models can be difficult to formulate and calculate since the relationship between the parameter estimates and marginal mean is a function of both the mixing probabilities and conditional mean. The simulation results not only demonstrate the ease of performing inference on the marginal mean for MZI models compared to ZI models, but the coverage probability and size of the tests are correct for the MZI model, while the coverage probability and size for the ZI model were smaller and larger than desired.

Lastly we extended the work in Martin and Hall (2016) to include clustered data by forming a MMZI model with random effects in both linear predictors. We mention several advantages of the MMZI model compared to the ZI mixed-effect model. One of these advantages is marginalizing across the distribution of the random effects in the linear predictor for the mean. In order to obtain population averaged interpretations related to the marginal mean in a ZI mixed-effect model, it must be marginalized via a convolution equation because a closed form relationship between the conditional and marginal parameters does not exist. However, since the MMZI model relates the predictors to the marginal mean directly, when a log link is used a closed form relationship between the marginal and conditional mean exists such that regression parameters other than the intercept have both SS and PA
interpretations. When a log link is not used, this relationship no longer holds, but MMZI models remain easier to marginalize than corresponding ZI-mixed models. E.g., a MMZIB model that can be marginalized via a convolution equation that has closed form solution can be formulated using a probit link. Our simulation results for MMZI models demonstrate that MLEs of the fixed effects have small empirical bias when Gaussian quadrature is used for modest sample sizes and few quadrature points. However, for accurate estimates of the variance components, larger sample sizes and more quadrature points are needed. The simulations also show that, as compared to ZI-mixed models, the MMZI model is more robust to misspecification of the linear predictor for the mixing probabilities.

**Future research**

There are advancements and applications that can be made for both the pseudo-$R^2$ and MZI models. While beyond the scope of this dissertation, there are several promising avenues of future research.

Pseudo $R^2$ terms can be developed for other models for EZ data. Other than standard ZI models, $R^2$-like quantities will be of interest in ZI models for clustered data such as ZI-mixed regression ([Hall 2000](#)) and marginal ZI regression ([Hall and Zhang 2004](#)). In addition, such statistics will be of interest for hurdle models and for the MZI and MMZI models we proposed in this dissertation. In principle the approach that we took in Chapter 2 can be pursued for all of these contexts, but details of the statistics and their properties remain to be explored.

One of the main challenges we encountered in trying to develop a suitable model selection criterion for ZI regression was to obtain a bias-corrected estimate of the KL divergence in this context. This remains an open problem.

There has not been any work toward developing marginal hurdle models for cross-sectional and clustered data. In principle such models can be formulated in a manner similar to the MZI and MMZI models discussed in chapters 2.7 & 4. However, we are skeptical whether
the hurdle model provides the most natural and interpretable framework for the analysis of EZ data.

The development of non-likelihood approaches to model fitting in models for the EZ data is desirable. The methods used in chapters 2.7 and 4 assume fully parametric mixture model. Semiparametric formulations of ZI models that are not fully specified and which could be fit with, for example, quasi-likelihood, are desirable for parameter estimation that is robust to model assumptions.

A semi-parametric formulation of the MZI model is desirable in which smooth functions of one or more covariates are introduced in the linear predictor for the marginal mean. For instance, a penalized spline version of the MZI model would be useful for capturing complex nonlinear relationships between the mean response and one or more covariates.
References


