Published in final edited form as:
*Stat Med*. 2008 August 15; 27(18): 3674–3688. doi:10.1002/sim.3267.

# Zero inflation in ordinal data: Incorporating susceptibility to response through the use of a mixture model

**Mary E. Kelley, PhD**[1] and **Stewart J. Anderson, PhD**[2]

[1] Department of Biostatistics, Rollins School of Public Health, Emory University, Atlanta, GA 30322

[2] Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15260

## Summary

The aim of the paper is to produce a methodology that will allow users of ordinal scale data to more accurately model the distribution of ordinal outcomes in which some subjects are susceptible to exhibiting the response and some are not (i.e., the dependent variable exhibits zero inflation). This situation occurs with ordinal scales in which there is an anchor that represents the absence of the symptom or activity, such as “none”, “never” or “normal”, and is particularly common when measuring abnormal behavior, symptoms, and side effects. Due to the unusually large number of zeros, traditional statistical tests of association can be non-informative. We propose a mixture model for ordinal data with a built-in probability of non-response that allows modeling of the range (e.g., severity) of the scale, while simultaneously modeling the presence/absence of the symptom. Simulations show that the model is well behaved and a likelihood ratio test can be used to choose between the zero-inflated and the traditional proportional odds model. The model, however, does have minor restrictions on the nature of the covariates that must be satisfied in order for the model to be identifiable. The method is particularly relevant for public health research such as large epidemiological surveys where more careful documentation of the reasons for response may be difficult.

### Keywords

zero inflation; proportional odds; ordinal data; mixture models

## 1. Introduction

One of the most common issues surrounding the analysis of clinical and utilization outcomes is what statisticians refer to as the excess zero or zero inflation problem. This occurs when an outcome is measured, e.g. counts of service utilization, and the resulting data collected contain many observations that are zero, i.e., did not use the service being measured. In the case of normal or lognormal data, there are a variety of applications where there are a considerable number of subjects with zero values, either due to the fact that they do not participate in the activity (e.g., drinking alcohol [1]) or that they are below a certain threshold for detection [2]. These cases have both been dealt with through the application of a mixture model with the appropriate continuous distribution and a built-in probability of non-response, also known as a two-part model [3,4]. These models provide a conditional test of association between the outcome and any predictor of interest, i.e., a test that removes the effect of the zero responses, in order to answer meaningful research questions [5].

The need for a conditional test of association may also apply to ordinal scales in which there are anchors that represent the absence of the symptom or activity, such as “none”, “never” or

"normal". A mixture distribution in this context would imply that there are patients who are "susceptible" and patients who are not. However, for discrete data the definition of "excess" zero is more complicated. The choice to be made is whether or not we wish to include some of the zeros in the conditional test of association. If it is assumed that all zeros are "true" zeros, then the second distribution in the mixture can be modified (e.g. truncated) to reflect the association with the probability of response [6]. However, given sampling uncertainty and measurement error, the usual approach has been to split the zeros into two distinct populations: a group of subjects that do not have a response, and a group of subjects that have a response, who nonetheless exhibit zero values. Farewell and Sprott [7] used the term "sampling zero" for those zeros not attributed to "cure". In the derivation for the zero inflated poisson (ZIP) model [8], the mixture probability was referred to as the probability of the "perfect state" and the second distribution referred to another state in which "defects are possible, but not inevitable", allowing for some zero counts in the conditional distribution. For the ordinal case presented here, a zero measure would indicate that the symptom is not present at the time of measurement but that the subject is still susceptible to the phenomenon being measured. This pattern of response is particularly evident in the case of low incidence sequelae such as side effects or suicidal ideation. For example, it is the case that pharmacological side effects, when they are present, are usually dose related. However, there are a considerable number of subjects who will not experience these side effects at any dose, while others will not experience the side effect at their current dose at the time of the assessment. If there are sampling zeros of this nature, we would expect an underlying association between dose and severity of the side effect in those patients who are susceptible. If the symptom is measured on an ordinal scale, this would coincide with a relationship that is consistent with the proportional odds assumption. We would also expect that dose may not be related to presence or absence of the side effect, but only related to the severity of the side effect when it occurs. Thus we have two distinct reasons for the application of a mixture model if there are, in fact, different populations of patients: 1) to explicitly model the clinical source of the zero inflation, and 2) to clarify the relationship with possible predictors.

The proportional odds (PO) model, when it applies, is unique among the ordinal regression models in that it is invariant to collapsing across categories, which is often needed to summarize results. More importantly to the current application, however, is its similarity to the results of a traditional linear regression on the underlying variable in that it allows for a test of association between a predictor and the outcome variable that is not category specific. Given that most ordinal scales are constructed with an underlying variable in mind, it would be desirable to the clinician to retain the ability to perform one test of association. It is assumed that for ordinal regression, the direct results of a mixture of responders and non-responders would be a deviation from the proportional odds assumption. When the proportional odds assumption is violated, another alternative is the partial proportional odds model (PPO) [9]. For this particular application, the constrained form of the model in which only the zero category was specified to have non-proportional odds would be the most likely analogue to the proposed model. It differs from the current model in that it does not allow for sampling zeros, and it does not incorporate the probability of response into the conditional (non-zero) distribution. However, we will consider this model further in our clinical example.

Our aim is to produce a methodology that will allow users of ordinal scale data to more accurately model the distribution of ordinal outcomes when it is assumed that not all patients are susceptible to the phenomenon being measured, and that this is the primary reason for any deviation from the proportional odds assumption. Since most applications we will consider are measures of symptoms or side effects, we will refer to the two aspects of the distribution as "incidence" and "severity" in order to simplify the discussion. Similar terms

that have also been used are “occurrence” and “intensity”. The model proposed allows modeling of the range (e.g., severity) of the scale, while simultaneously modeling the presence/absence of the symptom, and allows the predictors of incidence and severity to differ.

## 2. The zero inflated proportional odds model (ZIPO)

### 2.1 Model specification

The development of a zero inflated proportional odds model is similar in derivation to previous zero inflated models [8,10], with the distribution component consisting of a multinomial distribution with the logit link used for the linear predictor.

Let $y_i$ be an ordinal measure, on the ith subject, i=1,…, n, with levels 0,1,…, J with a multinomial distribution:

$$y_i \sim MULT(1, \gamma_{0,i}, \ldots, \gamma_{J,i})$$

with cumulative probabilities defined such that:

$$\gamma_j = \Pr(y_i \leq J) \quad j = 0, 1, \ldots, J$$

For this adaptation, the response is distributed as a mixture of two distributions, a point mass at 0 and a multinomial:

$$\begin{aligned} y_i &\sim 0 \text{ with Probability } p_i \\ &\sim MULT(1, \gamma_{0,i}, \ldots, \gamma_{J,i}) \text{ with Probability } 1 - p_i \end{aligned}$$

so that

$$\begin{aligned} y_i &= 0 \text{ with probability } p_i + (1 - p_i)\gamma_{0,i} \\ y_i &= j \text{ with Probability } (1 - p_i)(\gamma_{j,i} - \gamma_{j-1,i}). \end{aligned}$$

For the ordinal regression, the parameters p and $(\gamma_0,\ldots,\gamma_J)$ will be modeled via a canonical logit link. If $\mathbf{x}_i$ denotes the full covariate vector, we can choose any subset, including $\mathbf{x}_i$ itself, for $\mathbf{G}_i$ and $\mathbf{B}_i$ to form the linear predictors for the probabilities of the perfect state and other state, respectively. The vectors of covariate effects corresponding to the perfect state and the multinomial state will be denoted by $\boldsymbol{\tau}$ and $\boldsymbol{\beta}$, respectively. In order to maintain notational consistency with proportional odds models, we use the parameter $\boldsymbol{\theta}$ for the intercept parameters in the ordinal regression, and we will identify the complete parameter vector for the ordinal part as $\boldsymbol{\eta}' = (\theta_0,\theta_1,\ldots,\theta_{J-1}, \boldsymbol{\beta})'$. The model equations will then be:

$$\begin{aligned} \text{logit}(p_i) &= \mathbf{G}_i\boldsymbol{\tau} \\ \gamma_{j,i} &= F(\theta_j - \mathbf{B}_i\boldsymbol{\beta}) \end{aligned}$$

with F being the cumulative distribution function for the logit:

$$F(\theta_j - \mathbf{B}_i\boldsymbol{\beta}) = \frac{\exp(\theta_j - \mathbf{B}_i\boldsymbol{\beta})}{1+\exp(\theta_j - \mathbf{B}_i\boldsymbol{\beta})}$$

This results in an observed-data log-likelihood, in terms of the regression parameters, of the form

$$l(\boldsymbol{\tau},\boldsymbol{\eta};\mathrm{Y}) = \sum_{i=1}^{n} \left\{ \begin{array}{l} \log\{\exp(\mathbf{G}_i\boldsymbol{\tau}) + F(\theta_0 - \mathbf{B}_i\boldsymbol{\beta})\}\,[I(y_i=0)] \\ + \left\{ \sum_{j=1}^{J} \log\left[F(\theta_j - \mathbf{B}_i\boldsymbol{\beta}) - F(\theta_{j-1} - \mathbf{B}_i\boldsymbol{\beta})\right][I(y_i=j)] \right\} \\ -\log(1+\exp(\mathbf{G}_i\boldsymbol{\tau})) \end{array} \right\} \tag{2.1}$$

If we could observe which data came from which part of the mixture in an indicator variable, $Z=(z_1,\ldots z_n)'$, i.e.,

$$z_i = \left\{ \begin{array}{ll} 1 & \text{if } y_i \text{ is from the}''\text{perfect state}''; \\ 0 & \text{otherwise.} \end{array} \right.$$

where the "perfect state" refers to the point mass at zero (i.e., those subjects that do not exhibit a response), then we could construct a complete data log-likelihood:

$$l(\boldsymbol{\tau},\boldsymbol{\eta};\boldsymbol{Y},\boldsymbol{Z}) = \sum_{i=1}^{n} z_i \{\log(\exp(\mathbf{G}_i\boldsymbol{\tau}) - \log(1+\exp(\mathbf{G}_i\boldsymbol{\tau})\} + \sum_{i=1}^{n}(1-z_i)\left\{\sum_{j=0}^{J}\log[F(\theta_j - \mathbf{B}_i\boldsymbol{\beta}) - F(\theta_{j-1} - \mathbf{B}_i\boldsymbol{\beta})][I(y_i=j)] - \log(1+\exp(\mathbf{G}_i\boldsymbol{\tau})\right\}. \tag{2.2}$$

Rearranging and simplifying, we then have:

$$\begin{aligned} l(\boldsymbol{\tau},\boldsymbol{\eta};\boldsymbol{Y},\boldsymbol{Z}) &= \sum_{i=1}^{n} z_i \{\mathbf{G}_i\boldsymbol{\tau} - \log(1+\exp(\mathbf{G}_i\boldsymbol{\tau})\} - \sum_{i=1}^{n}(1-z_i)\{\log(1+\exp(\mathbf{G}_i\boldsymbol{\tau})\} \\ &+ \sum_{j=0}^{J}\sum_{i=1}^{n}(1-Z_i)\left\{\log[F(\theta_j - \mathbf{B}_i\boldsymbol{\beta}) - F(\theta_{j-1} - \mathbf{B}_i\boldsymbol{\beta})][I(y_i=j)]\right\} \\ &= l_C[\boldsymbol{\tau}|\boldsymbol{Y},\boldsymbol{Z}] + l_C[\boldsymbol{\eta}|\boldsymbol{Y},\boldsymbol{Z}] = l_C[\boldsymbol{\tau},\boldsymbol{\eta}|\boldsymbol{Y},\boldsymbol{Z}] \end{aligned} \tag{2.3}$$

Thus, the complete data log-likelihood is easily maximized due to the fact that $\boldsymbol{\tau}$ and $\boldsymbol{\eta}$ can be maximized separately. It is also useful to note that the part involving $\boldsymbol{\eta}$ can be estimated by using appropriate weights in the fit of a traditional proportional odds model. The part of the complete-data likelihood involving $\boldsymbol{\tau}$ is identical to the zero inflated poisson (ZIP) model, and can be solved using the method derived by Lambert [8], which involves augmenting the data with an indicator vector and then solving the resulting weighted logistic regression using the appropriate weights for each piece of the likelihood. The ZIPO models were estimated using S-PLUS® (Insightful Corporation, Seattle, WA) code adapted from the Design library[11] and modified by the authors to fit the proposed models.

## 2.2 Model estimation

The estimation process is a function of the ordinal link we use. We will show the derivation for the logit link; the others can be derived similarly. The main problem in our estimation of the zero-inflated model is that we don't in fact observe the $z_i$'s. We can, however, formulate the problem with the $z_i$'s as incomplete data and solve using the EM algorithm [12]. The EM algorithm maximizes the log-likelihood iteratively by estimating the $z_i$'s using the current estimates of the parameters, and then maximizing the log-likelihood with the $z_i$'s fixed at their estimated expected values. This process is repeated until the algorithm converges.

**E step—**For iteration *k*, estimate $E[z_i \mid y_i, \boldsymbol{\tau}^{(k)}, \boldsymbol{\eta}^{(k)}]$ by its posterior mean $z_i^{(k)}$ given the data y, and the current estimates of $\boldsymbol{\tau}^{(k)}$ and $\boldsymbol{\eta}^{(k)}$ (see Appendix for derivation):

$$\begin{aligned} z_i^{(k)} &= \Pr[perfect\ state | y_i, \boldsymbol{\tau}^{(k)}, \boldsymbol{\eta}^{(k)}] \\ &= \frac{\Pr[y_i | perfect\ state]\Pr[perfect\ state]}{\Pr[y_i | perfect\ state]\Pr[perfect\ state] + \Pr[y_i | MULT]P[MULT]} \\ &= \begin{cases} \frac{\exp(\mathbf{G}_i\boldsymbol{\tau}^{(k)})(1+\exp(\theta_0{}^{(k)} - \mathbf{B}_i\boldsymbol{\beta}^{(k)}))}{\exp(\mathbf{G}_i\boldsymbol{\tau}^{(k)})(1+\exp(\theta_0{}^{(k)} - \mathbf{B}_i\boldsymbol{\beta}^{(k)})) + \exp(\theta_0{}^{(k)} - \mathbf{B}_i\boldsymbol{\beta}^{(k)})} & if\ y_i = 0 \\ 0 & if\ y_i \neq 0 \end{cases} \end{aligned} \tag{2.4}$$

**M step—**Given that the expected values of the $z_i$'s are constants, the maximization of $Q = E_Z[l_C(\mathbf{Y}, \mathbf{Z} \mid \boldsymbol{\tau}^{(k)}, \boldsymbol{\eta}^{(k)})]$ is, in this instance, equivalent to maximizing the augmented logistic and the weighted ordinal regression, using the current estimates of $z_i = z_i^{(k)}$ as weights.

To initialize values for the conditional parameters (in this case $\boldsymbol{\eta}$), we used the standard proportional odds estimates. In order to point the algorithm towards the mixture solution, we calculated a z-score using the mean and variance of the non-zero data and used it to derive the expected proportion of zeros using the logistic cumulative distribution function., We then used the logit transform of this value as the initial $\tau_0$, with $\tau_1$ initialized at zero. For standard error estimates, we directly evaluated the observed information matrix using the appropriate derivatives of (2.1).

## 2.3 Tests of significance

Traditional likelihood ratio statistics can be computed for the parameters as well as any test statistics for nested hypotheses [13]. In addition, an hypothesis of the form $H_0$: p=0, can be tested using the method of Self and Liang [14] for hypothesis tests on the boundary of the parameter space. This provides a likelihood ratio test of the benefit of the zero inflated model over the proportional odds model, i.e., for the presence of "nonresponders".

## 2.4 Special considerations for estimation in the multinomial setting

**2.4.1 Identifiability—**It can be shown that a single binary predictor for both portions of the model will result in a non-unique solution. However, simple bounds on the predictors can ensure identifiability of the proposed model. Given that the sufficient statistics for the multinomial probabilities are the respective counts for that anchor, and these probabilities are mutually exclusive, the only difference in estimation from the regular multinomial problem is for the probability of a zero response. Given that the marginal distribution of any particular response category is Bernoulli, the subject-specific probability of a zero response will then be a mixture of two Bernoulli distributions. It has been shown that mixtures of two binomials (parameters n and p) are only identifiable when n > 1 (Teicher [15]). However, Follmann and Lambert [16] determined that finite mixtures of logistic regressions with a Bernoulli response can be identified, as long as the number of unique predictor combinations (covariates) is sufficient for identifiability. Specifically, they show that the number of components of the mixture, c, is constrained by

$$c \leq \sqrt{N_1+2} - 1 \tag{2.5}$$

where $N_1$ is the number of unique observed values of the covariate vector. Thus, a single binary covariate will only support one component distribution. Given Follmann and Lambert's theorem (2.5), a mixture of two Bernoulli distributions is identifiable if the number of unique combinations of the covariate vector is at least seven (i.e., seven is sufficient, but not necessary). Given the nature of most applications, this restriction will rarely be prohibitive. If the focus of inference is a comparison of groups, one can simply incorporate a continuous but nonsignificant predictor to ensure identifiability that would have little effect on the group comparison.

**2.4.2 Boundary solution for multinomial probability of zero response—**Because the multinomial distribution has mutually exclusive categories, it is a possibility that the estimate for the conditional probability of a zero response ($\gamma_0$) can approach zero, resulting in a corresponding regression coefficient ($\theta_0$) which tends to infinity. This would essentially result in a solution in which all the zeros were classified as "non-responders" and the non-zero data would be modeled using the resulting conditional probabilities. To provide an empirical estimate, we decreased the convergence criterion until there was clear separation between the valid solutions and those on the boundary. Then, in order to determine the extent of the impact of this boundary solution, we tracked the percentage of occurrences of this type of solution by considering any fitted model with $\theta_0 > 5.6$ as indicating a dataset with the solution on the boundary.

## 3. Model evaluation/simulation

### 3.1 Generation of zero inflated ordinal variates

In order to simulate data with the expected underlying mechanism, we began with a choice of regression coefficients ($\boldsymbol{\tau}$,$\boldsymbol{\eta}$) and proceeded to determine the resulting binomial and multinomial probabilities associated with a fixed covariate $\mathbf{x}$ (p, $\gamma_0$,$\gamma_1$,…$\gamma_J$). We assumed the same full linear predictor (i.e., $\mathbf{B}$=$\mathbf{G}$=$\mathbf{x}$) for both parts of the mixture, although this is not necessary for estimation. We then generated the ordinal variate, y, using the following process:

1. Choose a sample size, parameters ($\boldsymbol{\tau}$,$\boldsymbol{\eta}$) and a fixed covariate vector $\mathbf{x}$. Then, for each observation i:
2. Generate a set of probabilities from the specified linear predictor

$$p_i = \frac{\exp(\mathbf{x}_i\boldsymbol{\tau})}{1+\exp(\mathbf{x}_i\boldsymbol{\tau})}$$
$$\gamma_{j,i} = \frac{\exp(\theta_j - \mathbf{x}_i\boldsymbol{\beta})}{1+\exp(\theta_j - \mathbf{x}_i\boldsymbol{\beta})} \quad j=0,\ldots,J \tag{3.1}$$

3. Generate $z_i$ as a single random draw from a Bernoulli distribution with probability $p_i$
4. Generate a single uniform(0,1) variate, $u_i$, and assign values to the categorical variable $t_i$:

$$t_i = j \text{ if } \gamma_{j-1,i} \leq u_i < \gamma_{j,i}$$
$$\text{where } \gamma_{-1,i}=0 \text{ and } \gamma_{J,i}=1 \tag{3.2}$$

5. Generate $y_i$ by multiplication of $z_i$ and $t_i$:

$$y_i = z_i t_i \tag{3.3}$$

6. Repeat steps 2) – 5) S times.

### 3.2 Finite sample properties

The simulations were designed to test the asymptotic properties of the maximum likelihood estimates in finite sample sizes (50,100,200,500). We used fixed parameters as indicated below, and evaluated the following measures in S=2000 simulations (for definitions see the Appendix):

1. sample mean and variance
2. averaged standard deviation estimate from the observed information
3. averaged tail probabilities, and overall confidence interval coverage (normal theory)

### 3.3 Parameter choices

We created an example where the baseline (x=0) probability of nonresponse was 0.18, and the baseline cumulative probabilities for the ordinal categories 0,1,2,3 were equal to 0.1, 0.3, 0.6, 1, which corresponds to multinomial probabilities of 0.1, 0.2, 0.3, 0.4. These correspond to parameter values of $\tau_0 = -1.5$ and $\boldsymbol{\theta}$=(2.1972,0.8473, −0.8473, −2.1972). For the relationship with covariates we chose values of $\tau_1$=2.0, $\beta$=2.0, and the predictor x was given equally spaced values between 0 and 1. This provided a range of the probability of response (1−p) between 0.18 and 0.62 and the conditional probability of nonresponse ($\gamma_0$) between 0.015 and 0.1. The appropriate probabilities were calculated for each subject based on their linear predictor value.

### 3.4 Summary of simulation results

**3.4.1 Sources of error—**Some simulation data sets could not be used for assessment of the asymptotic properties due to particular properties of the data set. These properties are defined below and the frequency of their occurrence is documented in Table 1.

**Incomplete scale:** Because the proportional odds model has intercepts for each category of the scale, we eliminated simulated datasets that did not have all possible values of the dependent variable (e.g., 0,1,2,3,4), because the number of intercepts, as well as the values of the intercepts, would not correspond to the proposed values. However, this would not be a problem in practice, as the model fits only as many intercepts as are necessary for the data.

**Singular information:** In the evaluation of the ZIP model, Lambert discusses briefly that some data sets exhibit singular observed information and that removing these instances from the simulation results improves the estimation of the parameters considerably. In order to further clarify this source of error for the proposed model, we tracked the occurrence of singularity of observed information in the simulations. Singularity is clearly a function of the sample size [Table 1], with larger sample sizes less likely to exhibit singular information.

**Boundary solution:** The proportion of ordinal solutions on the boundary of the parameter space decreases as the sample size increases, as expected [Table 1]. Given that the true parameter value of $\theta_0$ was set to be in the interval (0.015–0.1) for this set of simulations, it is

likely that the low values of this conditional probability were the cause of this numerical problem, as the algorithm must choose which zeros are from which distribution. In those cases in which the probability of a sampling zero is low, either higher sample sizes can be used, such as with the clinical example presented here, or a two-part model [3] should be considered.

**3.4.2 Asymptotic properties—**Assessment of the valid simulations revealed that while estimates of the linear predictor parameters are variable [Table 2], the nature of the logistic function is such that the convergence of the underlying probabilities was reasonably accurate [Table 3]. As with the ZIP model [8] the linear predictor for the probability of a "perfect" zero is not estimated well. We assume that this is partly due to the indeterminate nature of assigning the zeros to the two distributions, and perhaps also due to the over-parameterization of what is essentially a point mass. We suspect that this is a property of all discrete zero inflated models, and possibly the continuous two-part models as well. Unfortunately, simulations of this nature are not available for the majority of other zero-inflated models so we have no way of examining this proposition without additional investigation. In this multinomial extension of the problem, it appears that this over-parameterization also effects the estimation of the probability of a sampling zero, which is consistently underestimated.

It is encouraging, however, that the covariate parameters have the desired qualities. Wald confidence intervals were surprisingly accurate for the ordinal parameter at all sample sizes [Table 4]. In contrast, the corresponding confidence intervals for the predictor of non-response ($\tau_1$) are somewhat conservative for the smaller sample sizes ($n< 200$). This is an interesting result and requires further investigation using other zero-inflated models to determine if an adjustment could be made that would correct this problem. Although the proposed method is much more suitable for larger ($n \geq 200$) sample sizes, analyses of smaller data sets using this technique could still provide more accurate results if it is assumed that there is a mixture of subjects in the sample.

## 4. Motivating clinical example: Alcohol consumption data

The Alcohol Use Disorders Identification Test (AUDIT) is a frequently used measure of the level of consumption as well as some of the negative consequences resulting from the consumption of alcohol [17]. Recently, a focus of primary care services research has been to identify hazardous drinkers, i.e., those patients who drink at a level that may cause health and legal repercussions but have not yet developed clinically diagnosable alcoholism [18]. In studying these patients, many of the traditional measures used for alcoholics have been used despite the fact that some items may not directly apply to this population. Many primary care patients do not drink at all, resulting in an inflation of "never" responses. This provides a setting for the mixture model proposed in order to correctly determine the predictors of severity. A similar approach was used by Olsen and Schafer [1], who used a continuous mixture model to assess predictors of alcohol consumption in teenagers. The data used for the example come from a study of Early Lifestyles Management (ELM), which identified possible hazardous drinkers through the use of a screening evaluation of alcohol consumption in the waiting room of 12 primary care clinics in the Pittsburgh area [19]. For the main predictor variable, we chose gender, which is a general demographic also collected on a large number of screens. There is considerable evidence of gender differences in consumption [20,21], with males commonly reporting more drinks per week than females.

We included an additional covariate, age group, for two purposes: 1) to force the model to be identifiable and 2) because age was a significant confounder as indicated by its association with both gender and consumption. For our sample, the males were considerably

older than the females (grouped means of 55 and 45 years, respectively) and those who reported never drinking alcohol tended to be older than those who drink (grouped means of 54 vs 48 years). Unfortunately, due to the brief nature of the survey, age was not a continuous variable but a categorical variable with nine categories. We chose to keep age group as a continuous variable in the analysis, however, to simplify the interpretation and keep the focus on our primary variable of interest, gender. We did this by substituting the midpoint age value for each category and treating the resulting variable as continuous. The corresponding frequency distribution for the 11,492 completed screening data forms with valid age group and gender is shown in Figure 1. It can be seen that a large proportion of the subjects chose the "never" option. However, the next anchor, monthly or less, may be more frequent than some occasional drinkers were willing to admit, making it possible that not all zeros were true abstainers. In addition, we have sufficient reason to believe that the predictors of abstinence may be quite different from predictors of consumption in drinkers. This provides initial justification for use of the new model.

### 4.1 Existing method results

Although the observed proportion of non-drinkers is nearly identical in male and female patients [40.3% vs. 40.5%, $\chi^2$ =0.041, p=0.840], the test of association between gender and consumption level is highly significant in the ordinal regression, indicating that males consume more alcohol [Table 5]. This apparent contradiction leads us to pursue a different model from a clinical perspective. One way of reconciling the discrepant results is if we consider abstinence as an additional factor. Our clinical objective would then be to determine if there is a gender difference in prevalence of abstinence and subsequently if there is a gender difference in consumption in those who are not abstinent. The proposed model tests this explicitly, as stated in the introduction; however we will first consider possible constraints for the PPO model that might allow us to incorporate abstinence in a similar way. If we fit equal slopes to all categories but the first (zero), we get the results in Table 5. The interpretation of the gender effect for this model is similar to the PO model in that males have a lower probability of being in the "never" category only the difference is less pronounced for the constrained PPO model due to the significant offset to the parameter in that group. However, the data are not consistent with this result as the proportions in the "never" category are nearly identical.

### 4.2 New model results

If we incorporate the concept of abstinence as an additional source of "never" responses that is separate from the level of consumption, the results indicate that males have a slightly higher proportion of abstainers than females [Table 6], which could then account for the fact that there is no gender difference in the observed proportions in the "never" group. Coincidentally, both information criteria (AIC and BIC) as well as a comparison of the likelihoods show a preference for the zero-inflated model over the PO and PPO models for these data. A likelihood ratio test indicating the improvement with the mixture can be constructed using a boundary test as discussed by Self and Liang [14]. This test is based on a reference distribution of a 50:50 percent mixture of $\chi^2$ distributions with 0 and 1 degrees of freedom respectively. This result is highly significant [$\chi^2$=428.94, df=0.5, $p < 0.0005$] although in the case of the current model, it is somewhat less informative as it cannot indicate if the lack of fit is due to the need for a mixture or the assumption of proportional odds.

Due to the nature of the cumulative logit models, both the PO and the PPO model predict higher incidence of "never" responses among females due to the fact that males drink more at the higher levels of consumption [Table 6]. However with the mixture model, the data indicate that for these data males are abstainers in higher proportions, but when they do

drink, they also have higher levels of consumption [Table 6]. This interpretation of the current sample fits better with existing data on gender and alcohol use. The National Health Interview Survey of drinking [20] differentiated between total abstainers and what they called “lifetime infrequent drinkers”, which would correspond to the two populations of zeros we suspect. However, the national data shows that females have higher proportions in both groups, thus we would not expect to see equal proportions of “never” responses across gender. The only group of non-drinkers with higher proportions in males is “previous drinkers”, or those who abstain most likely due to prior issues with drinking. Thus, the current data could be explained if there were a significant proportion of previous drinkers in the current sample. It turns out that the recruitment strategy for this study focused on areas in which high levels of drinking were expected and thus, may also have higher incidence of previous drinkers.

## 5. Conclusions

In this article, we have adapted the use of mixture models for zero inflation in categorical data such as the ZIP and zero inflated binomial (ZIB) [10] models to include ordinal level variables fit using the multinomial distribution. The extension to the multinomial distribution requires a few more restrictions on the nature of the predictors in the model; however, these restrictions can easily be met in most modeling applications. The model, however, does tend to require larger sample sizes ($n \geq 200$) due to the number of additional parameters involved. This should not be a problem for large survey applications, but for smaller samples any opportunities to collapse across categories should be considered carefully in order reduce the number of parameters under consideration. Although our model can potentially be used for any type of ordinal variable, it appears to apply best to ordinal scales where there is an anchor that represents a lack of response so that there is a justification for the existence of true nonresponders and thus can be easily interpreted from a clinical perspective.

## Acknowledgments

This work was funded in part by a seed grant (PI MK) from the VISN-4 Mental Illness Research, Education, and Clinical Center (MIRECC), Department of Veterans Affairs, Pittsburgh, PA and Public Health Service Grants No. U10CA-69974 and U10CA-69651 from the National Cancer Institute and the Department of Health and Human Services.

The authors would like to thank Steve A. Maisto, PhD, Syracuse University, and Joseph Conigliaro, MD, University of Kentucky for kindly providing the alcohol screening data. This work was funded in part by a seed grant (PI MK) from the VISN-4 Mental Illness Research, Education, and Clinical Center (MIRECC), Department of Veterans Affairs, Pittsburgh, PA and by Public Health Service Grants No. U10CA-69974 and U10CA-69651 from the National Cancer Institute and the Department of Health and Human Services..

## References

1. Olsen MK, Schafer JL. A two-part random effects model for semicontinous longitudinal data. JASA. 2001; 96(454):730–745.
2. Moulton LH, Halsey NA. A mixture model with detection limits for regression analyses of antibody response to vaccine. Biometrics. 1995; 51:1570–1578. [PubMed: 8589241]
3. Lachenbruch PA. Analysis of data with excess zeros. Statistical methods in medical research. 2002; 11:297–302. [PubMed: 12197297]
4. Mullahy J. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. Journal of Health Economics. 1998; 17:247–281. [PubMed: 10180918]
5. Boos DD, Brownie C. Mixture models for continuous data in dose-response studies when some animals are unaffected by treatment. Biometrics. 1991; 47:1489–1504. [PubMed: 1786327]

6. Heilbron DC. Zero-altered and other regression models for count data with added zeros. Biom J. 1994; 36(5):531–547.
7. Farewell VT, Sprott DA. The use of a mixture model in the analysis of count data. Biometrics. 1988; 44:1191–1194. [PubMed: 2466491]
8. Lambert D. Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics. 1992; 34(1):1–14.
9. Peterson B, Harrell FE. Partial proportional odds models for ordinal response variables. Appl Statist. 1990; 39(2):205–217.
10. Hall DB. Zero-inflated poisson and binomial regression with random effects: A case study. Biometrics. 2000; 56:1030–1039. [PubMed: 11129458]
11. Harrell, FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis, Springer series in statistics. New York: Springer; 2001.
12. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Statist Soc B. 1977; 39(1):1–38.
13. McCullagh P. Quasi-likelihood functions. Ann Statist. 1983; 11(1):59–67.
14. Self SG, Liang K-Y. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. JASA. 1987; 82(398):605–610.
15. Teicher H. Identifiability of finite mixtures. Ann Math Stat. 1963; 34(4):1265–1269.
16. Follmann DA, Lambert D. Identifiability of finite mixtures of logistic regression models. J Stat Planning Inference. 1991; 27:375–381.
17. Babor, TF.; de la Fuente, JR.; Saunders, J.; Grant, M. The Alcohol Use Identification Test: Guidelines for use in primary health care. Geneva, Switzerland: World Health Organization; 1989.
18. Sanchez-Craig M, Wilkinson A, Davila R. Empirically based guidelines for moderate drinking: one-year results from three studies with problem drinkers. Am J Pub Health. 1995; 85:823–828. [PubMed: 7762717]
19. Maisto SA, Conigliaro J, McNeil M, Kraemer K, Conigliaro RL, Kelley ME. Effects of two types of brief intervention and readiness to change on alcohol use in hazardous drinkers. J Stud Alcohol. 2001; 62:605–614. [PubMed: 11702799]
20. Dawson DA, Archer L. Gender differences in alcohol consumption. Br J Addict. 1992; 87:119–123. [PubMed: 1543934]
21. Olenick NL, Chalmers DK. Gender-specific drinking styles in alcoholics and nonalcoholics. J Stud Alcohol. 1991; 52:325–330. [PubMed: 1875705]

# Appendix

## 1. Simulation statistics

S=number of simulations; β represents the parameter of interest; $v_{jj}$ is the appropriate diagonal of the observed information matrix; I is the generic indicator function

Sample mean and variance were defined as

$$\text{mean} = \sum_{s=1}^{S} \widehat{\beta}_j^{(s)} / S$$

$$\text{variance} = \sum_{s=1}^{S} \left[ \widehat{\beta}_j^{(s)} - \sum_{r=1}^{S} \widehat{\beta}_j^{(r)} / S \right] / (S - 1)$$

Averaged standard deviation estimate from the observed information

$$\sqrt{\sum_{s=1}^{S} v_{jj}^{(s)}/S}$$

Averaged tail probabilities, and overall confidence interval coverage (normal theory):

$$\widehat{\beta}_j^{(s)} + Z_{1-\alpha/2}\sqrt{v_{jj}^{(s)}} = upper_j^{(s)}$$
$$\widehat{\beta}_j^{(s)} - Z_{1-\alpha/2}\sqrt{v_{jj}^{(s)}} = lower_j^{(s)}$$
$$total\ coverage = \sum_{s=1}^{S} I(\beta_j \in (lower_j^{(s)}, upper_j^{(s)}))$$
$$lower\ tail\ prob = \sum_{s=1}^{S} I(\beta_j < lower_j^{(s)})$$
$$upper\ tail\ prob = \sum_{s=1}^{S} I(\beta_j > upper_j^{(s)})$$

## 2. Derivation of $z_i^{(k)}$

Given that $z_i=0$ when $y_i \neq 0$ (because (Pr[perfect state=0]) the calculation is only for when $y_i=0$

the calculation then equals

$$\frac{1 \times p_i}{1 \times p_i + (1 - p_i) \times \gamma_{0,i}}$$

given that

$$p_i = \frac{\exp(\mathbf{G_i}\boldsymbol{\tau})}{1+\exp(\mathbf{G_i}\boldsymbol{\tau})}$$

and

$$\gamma_{0,i} = \frac{\exp(\theta_0 - \mathbf{B_i}\boldsymbol{\beta})}{1+\exp(\theta_0 - \mathbf{B_i}\boldsymbol{\beta})}$$

then $z_i$ is

$$= \frac{\exp(\mathbf{G_i}\boldsymbol{\tau})}{1+\exp(\mathbf{G_i}\boldsymbol{\tau})} \div \left[ \frac{\exp(\mathbf{G_i}\boldsymbol{\tau})}{1+\exp(\mathbf{G_i}\boldsymbol{\tau})} + \frac{1}{1+\exp(\mathbf{G_i}\boldsymbol{\tau})} \times \frac{\exp(\theta_0 - \mathbf{B_i}\boldsymbol{\beta})}{1+\exp(\theta_0 - \mathbf{B_i}\boldsymbol{\beta})} \right]$$

$$= \frac{\exp(\mathbf{G_i}\boldsymbol{\tau})}{1+\exp(\mathbf{G_i}\boldsymbol{\tau})} \div \frac{\exp(\mathbf{G_i}\boldsymbol{\tau})(1+\exp(\theta_0 - \mathbf{B_i}\boldsymbol{\beta})) + \exp(\theta_0 - \mathbf{B_i}\boldsymbol{\beta})}{(1+\exp(\mathbf{G_i}\boldsymbol{\tau}))(1+\exp(\theta_0 - \mathbf{B}} \mathbf{i}\boldsymbol{\beta})) \quad cmyk0000$$

$$= \frac{\exp(\mathbf{G_i}\boldsymbol{\tau})(1+\exp(\theta_0 - \mathbf{B_i}\boldsymbol{\beta}))}{\exp(\mathbf{G_i}\boldsymbol{\tau})(1+\exp(\theta_0 - \mathbf{B}} \mathbf{i}\boldsymbol{\beta})) + \exp(\theta_0 - \mathbf{B_i}\boldsymbol{\beta}) \quad cmyk0000$$

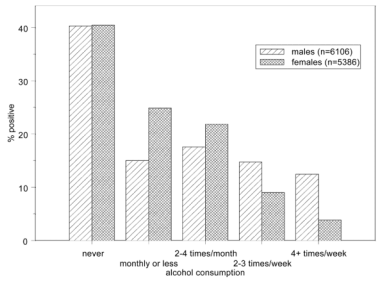


**Figure 1.**
Legend: Alcohol consumption as measured by the first item of the Alcohol Use Disorders Identification Test in a primary care screening sample [17]

**Table 1**

Sources of error in the simulated datasets by sample size (% of 2000 simulations):

| Sample size | % incomplete scale | % singular information | % valid sims (n) | % boundary solution (of valid) |
|---|---|---|---|---|
| 50 | 1.8 | 6.0 | 92.4 (1847) | 16.5 |
| 100 | 0.0 | 1.5 | 98.6 (1971) | 12.8 |
| 200 | 0.0 | 0.2 | 99.8 (1996) | 10.0 |
| 500 | 0.0 | 0.0 | 100.0 (2000) | 7.3 |

Legend: Incomplete scale=dataset did not have all possible values of y; Singular information = datasets for which the information matrix was not invertible; boundary solution= dataset in which the estimate for the probability of a conditional zero approached zero, resulting in infinite estimates;

**Table 2**

Simulation results - estimates of linear predictors and measures of error for varying sample size (valid simulations)

| | n | $\tau_0$ | $\tau_1$ | $\theta_0$ | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\beta$ |
|---|---|---|---|---|---|---|---|---|
| Asymptotic value | Sample size | **−1.500** | **2.000** | **2.197** | **0.847** | **−0.847** | **−2.197** | **2.000** |
| Sample Mean | 50 | −3.313 | 3.920 | 2.631 | 0.529 | −1.123 | −2.532 | 2.301 |
| | 100 | −2.600 | 3.276 | 2.554 | 0.640 | −1.015 | −2.382 | 2.153 |
| | 200 | −2.012 | 2.608 | 2.535 | 0.701 | −0.948 | −2.299 | 2.101 |
| | 500 | −1.761 | 2.307 | 2.502 | 0.744 | −0.916 | −2.259 | 2.059 |
| Sample Std. Dev. | 50 | 4.834 | 6.108 | 2.514 | 0.818 | 0.721 | 0.863 | 1.465 |
| | 100 | 3.028 | 3.390 | 2.141 | 0.645 | 0.507 | 0.591 | 1.017 |
| | 200 | 1.077 | 1.277 | 1.885 | 0.528 | 0.358 | 0.395 | 0.652 |
| | 500 | 0.582 | 0.706 | 1.629 | 0.397 | 0.245 | 0.260 | 0.424 |
| Info Std. Dev | 50 | 2.959 | 3.509 | 7.161 | 1.001 | 0.758 | 0.853 | 1.396 |
| | 100 | 1.729 | 2.056 | 5.031 | 0.755 | 0.527 | 0.582 | 0.961 |
| | 200 | 0.997 | 1.226 | 3.791 | 0.596 | 0.383 | 0.412 | 0.665 |
| | 500 | 0.600 | 0.741 | 2.643 | 0.434 | 0.258 | 0.267 | 0.416 |

Legend: definition of statistics included in the Appendix

**Table 3**

Simulation results – estimates of distribution parameters for varying sample size (valid simulations)

| | n | p | $\pi_0$ | $\pi_1$ | $\pi_2$ | $\pi_3$ | $\pi_4$ |
|---|---|---|---|---|---|---|---|
| **Asymptotic value at x=0** | **Sample size** | **0.18** | **0.10** | **0.20** | **0.40** | **0.20** | **0.10** |
| | 50 | 0.04 | 0.07 | 0.30 | 0.38 | 0.17 | 0.07 |
| | 100 | 0.07 | 0.07 | 0.27 | 0.39 | 0.18 | 0.08 |
| | 200 | 0.12 | 0.07 | 0.26 | 0.39 | 0.19 | 0.09 |
| | 500 | 0.15 | 0.08 | 0.25 | 0.39 | 0.19 | 0.09 |

**Table 4**

95% Wald (normal theory) confidence intervals (Z=1.96) for the model covariate parameters ($\tau_1$, $\beta$)

| | $Tau_1$ | | | Beta | | |
|---|---|---|---|---|---|---|
| **Sample size (n)** | **Total coverage(%)** | **Lower tail rejection probability** | **Upper tail rejection probability** | **Total coverage(%)** | **Lower tail rejection probability** | **Upper tail rejection probability** |
| 50 | 98.3 | 0.008 | 0.009 | 95.4 | 0.028 | 0.018 |
| 100 | 97.4 | 0.012 | 0.014 | 94.3 | 0.037 | 0.021 |
| 200 | 96.5 | 0.020 | 0.015 | 95.4 | 0.032 | 0.015 |
| 500 | 95.3 | 0.021 | 0.027 | 94.3 | 0.037 | 0.020 |

**Table 5**

New method results compared to ordinal regression: Alcohol consumption data

| | PO | Constrained PPO | ZIPO |
|---|---|---|---|
| Model LL | −16570.98 | −16456.825 | −16354.66 |
| AIC | 2.885 | 2.865 | 2.845 |
| BIC | −74245.32 | −74454.93 | −74649.92 |
| | Beta(se) | Beta (se) | Beta (se) |
| **Incidence:** | | | |
| Intercept | | | −1.826 (0.135) |
| Gender* | | −0.442 (0.035) | −0.304 (0.060) |
| Age (group median) | | −0.003 (0.001) | 0.031 (0.002) |
| **Severity:** | | | |
| Intercepts: y>=0 | −1.362 (0.057) | −1.574 (0.064) | 4.439 (2.258) |
| y>=1 | −0.527 (0.055) | −0.338 (0.060) | 0.385 (0.097) |
| y>=2 | 0.467 (0.056) | 0.660 (0.060) | −1.038 (0.080) |
| y>=3 | 1.511 (0.061) | 1.710 (0.065) | −2.241 (0.082) |
| Gender | 0.597 (0.036) | 0.808 (0.040) | 1.004 (0.052) |
| Age (group median) | −0.025 (0.001) | −0.024 (0.001) | −0.004 (0.002) |

*coding for gender: 0=female, 1=male

**Table 6**

Predicted percentages for ordinal models compared to the observed percentages

| | Observed | | PO | | Constrained PPO | | ZIPO | |
|---|---|---|---|---|---|---|---|---|
| | **Males** | **Females** | **Males** | **Females** | **Males** | **Females** | **Males** | **Females** |
| % abstinent | | | | | | | 40.0 | 39.6 |
| Never | 40.4 | 40.5 | 33.5 | 47.7 | 39.3 | 41.3 | 0.3 | 0.8 |
| Monthly or Less | 15.0 | 24.9 | 20.2 | 20.1 | 14.7 | 25.7 | 13.8 | 26.2 |
| 2–4 times/month | 17.6 | 21.8 | 22.1 | 17.3 | 21.5 | 17.4 | 19.5 | 19.5 |
| 2–3 times/week | 14.7 | 9.0 | 14.1 | 9.1 | 14.1 | 9.5 | 14.9 | 8.9 |
| 4+ times/week | 12.5 | 3.9 | 10.1 | 5.8 | 10.4 | 6.1 | 11.4 | 5.0 |