



Faculty Publications

2013-3

A tutorial on count regression and zero-altered count models for longitudinal substance use data

Scott A. Baldwin
Brigham Young University - Provo

David C. Atkins

Cheng Zheng

Robert J. Gallop

Clayton Neighbors

Follow this and additional works at: <https://scholarsarchive.byu.edu/facpub>



Part of the [Psychology Commons](#)

BYU ScholarsArchive Citation

Baldwin, Scott A.; Atkins, David C.; Zheng, Cheng; Gallop, Robert J.; and Neighbors, Clayton, "A tutorial on count regression and zero-altered count models for longitudinal substance use data" (2013). *Faculty Publications*. 6061.

<https://scholarsarchive.byu.edu/facpub/6061>

This Peer-Reviewed Article is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Faculty Publications by an authorized administrator of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.

Published in final edited form as:

Psychol Addict Behav. 2013 March ; 27(1): 166–177. doi:10.1037/a0029508.

A tutorial on count regression and zero-altered count models for longitudinal substance use data

David C. Atkins,

Department of Psychiatry and Behavioral Sciences, University of Washington

Scott A. Baldwin,

Department of Psychology, Brigham Young University

Cheng Zheng,

Department of Biostatistics, University of Washington

Robert J. Gallop, and

Applied Statistics Program, West Chester University

Clayton Neighbors

Department of Psychology, University of Houston

Abstract

Critical research questions in the study of addictive behaviors concern how these behaviors change over time - either as the result of intervention or in naturalistic settings. The combination of count outcomes that are often strongly skewed with many zeroes (e.g., days using, number of total drinks, number of drinking consequences) with repeated assessments (e.g., longitudinal follow-up after intervention or daily diary data) present challenges for data analyses. The current article provides a tutorial on methods for analyzing longitudinal substance use data, focusing on Poisson, zero-inflated, and hurdle mixed models, which are types of hierarchical or multilevel models. Two example datasets are used throughout, focusing on drinking-related consequences following an intervention and daily drinking over the past 30 days, respectively. Both datasets as well as R, SAS, Mplus, Stata, and SPSS code showing how to fit the models are available on a supplemental website.

What is the impact of personalized normative feedback on drinking related problems in college students over time? How does weekend versus weekday drinking vary by gender and fraternity/sorority status when assessed on a daily basis? Many questions about alcohol and substance abuse focus on change across time, and the methods used to analyze these questions need to account for the longitudinal nature of the data. Generalized linear mixed models (GLMMs; Gelman & Hill, 2007; Hedeker & Gibbons, 2006; also called hierarchical [or multilevel] generalized linear modeling, Raudenbush & Bryk, 2002; Snijders & Bosker, 1999) are increasingly common analytic approaches for longitudinal data, given their flexible handling of unbalanced repeated measures (i.e., individual participants may have unique numbers and timings of assessments) and the widespread availability of software for estimating such models. Moreover, GLMMs are appropriate for continuous as well as discrete outcomes.

However, the distributions of alcohol and substance abuse outcomes have characteristic shapes: They are often positively skewed and bounded by zero. Moreover, there can be a

large stack of data points at zero, indicating individuals and/or occasions without drinking, use, or related problems. These distributions reflect that alcohol and substance abuse outcomes are often count data, representing a total number of something, be it drinks, days using, or number of problems. Except in special circumstances (e.g., specially selected samples with high drinking or drug use), statistical models that assume normally distributed residuals will provide poor fit to such data and will lead to incorrect confidence intervals and p -values. Instead, count regression approaches such as Poisson or negative binomial regression or zero-altered count models (e.g., zero-inflated or hurdle models) are much more appropriate for these types of data (Atkins & Gallop, 2007; Cox, West, & Aiken, 2009; Hilbe, 2011; Neal & Simons, 2007; Simons, Neal, & Gaher, 2006).

In the past, addictions researchers have often ignored (or not been aware of) violations of distributional assumptions or have attempted to deal with them in non-optimal ways. Count regression models are beginning to be applied to addictions data (e.g., Gaher & Simons, 2007; Lewis et al., 2010), but accessible resources on how to apply these models to longitudinal data are scarce. The present article provides a tutorial in analytic methods for count data from longitudinal studies, focusing on extensions to GLMMs for count outcomes. We use two examples from our research to illustrate the need for, and application of, longitudinal count models. Data and computer code to run the analyses in R, Mplus, SAS, Stata, and SPSS are available on a supplementary website, though note that not all software can run all models that are covered here at present time (<http://depts.washington.edu/cshrb/newweb/statstutorials.html>). The outline of the article is as follows: Introduction to example data and research questions, brief overview of count regression models, GLMMs for count regression models, analyses and interpretation of example data, and discussion of software and practical issues in using these methods. In addition, there is a technical appendix containing important, but more advanced, material. We assume that readers have a basic familiarity with linear mixed models (i.e., hierarchical linear or multilevel models assuming normally distributed errors) and count regression models, though both are introduced briefly here and introductory resources are highlighted throughout.

Motivating Examples

The first example dataset is drawn from an intervention study aimed at reducing problematic drinking in college students (Neighbors et al., 2010). The current paper focuses on gender differences across two years in alcohol-related problems, as measured by the Rutgers Alcohol Problem Index (RAPI; White & Labouvie, 1989). The dataset includes 3,616 repeated measures across five time points from 818 individuals. The second dataset involves intensive, daily assessment of drinking. The data come from a larger intervention study of event specific prevention (i.e., drinking related interventions for 21st birthdays and spring break), but the current data are observational. Specifically, these data record the number of drinks for each day over approximately the last 30 days for 980 individuals (23,992 total person-days), as measured by the timeline follow-back interview (TLFB). These data came from a survey study of 21st birthday drinking and consequently include some extreme drinking events relative to a random sample of student's drinking (Neighbors et al., 2011). Analyses focus on drinking differences by gender, greek status (i.e., fraternity/sorority membership), and weekend (Thurs-Sat) vs. weekday (Sun-Wed). (Note that Thursday was included as part of the "weekend" given that drinking on Thursday was more similar to Friday and Saturday drinking than other days of the week.)

Count Regression Models

Count variables are often positively skewed and often include many observations at zero. The top row of Figure 1 displays (unconditional) distributions of drinking (TLFB) and

alcohol problems (RAPI), which are strongly skewed with a mode of zero. The bottom row of Figure 1 shows histograms of residuals from regressing the RAPI on gender and time assuming normally distributed residuals (i.e., ordinary least squares [OLS] regression) – on the left without any transformation and on the right with a log transformation of the RAPI. These plots show that skewed count regression outcomes will rarely meet the distributional assumptions of OLS regression or linear mixed models. Moreover, count outcomes will also typically violate the equal variances (i.e., homoskedasticity) assumption of linear models as count outcomes have a direct relationship between their mean and variance, where higher levels of the outcome have greater variance. Although transforming the outcome is a commonly suggested strategy for skewed data, a stack of zeroes will not be smoothed out by a transformation. Moreover, focusing on the TLFB data, the distribution shown in Figure 1 is suggestive of two different types of associations or research questions: a) what is related to no drinking vs. any drinking (i.e., zero vs. non-zero), and b) what is related to the amount of drinking when there is drinking? Although the zero / non-zero aspects of the RAPI data are not quite as notable, a similar set of questions could be asked of the RAPI data. As will be described later, zero-inflated and hurdle models have submodels that focus on these two questions. Prior to introducing count regression models, we consider the qualities of the example data that are related to the questions just posed.

Consider how the proportion of individuals drinking versus not drinking and number of drinks on drinking days are related to covariates in the TLFB data. Figure 2 presents means and 95% confidence intervals for number of drinks on drinking days (top half of graph) and proportion of days drinking (bottom half of graph) by weekday vs. weekend, greek status, and gender.

There are strong differences in the proportion of people drinking for weekend vs. weekday, with less prominent differences for number of drinks when drinking. Moreover, from this descriptive view, the number of drinks is highly related to fraternity/sorority status, whereas proportion of people drinking does not appear as strongly related to greek status.

Figure 3 shows a similar set of plots for the RAPI by gender across time. As with the TLFB drinking data, patterns of association appear different across the two aspects of the outcome. Over the five assessments (and following intervention for most participants), the proportion of individuals with any alcohol problems is dropping, with a greater proportion of men consistently reporting more alcohol problems (right panel). When examining the number of alcohol problems (given some problems), we see evidence for divergence between the sexes, with women showing slight decreases in number of alcohol problems whereas men appear to show slight increases, though with notable variability as seen in the confidence intervals (left panel). Thus, these graphs serve to highlight that different associations can occur with outcomes with notable zeroes (i.e., whether there is any drinking [zero vs. not zero] and amount of drinking when any drinking).

Before moving on to discuss GLMMs for count outcomes, we briefly introduce count regression models (more thorough introductions can be found in articles by Atkins & Gallop, 2007, and Cox, West, & Aiken, 2009, and the book by Hilbe, 2011). The basic count regression model is Poisson regression, which is one of the generalized linear models (McCullagh & Nelder, 1989). There are two critical differences between OLS regression and Poisson regression. First, the outcome (conditional on covariates) is assumed to be distributed as a Poisson random variable as opposed to a Normal random variable. The top row of Figure 4 depicts three different Poisson distributions, with varying means (denoted by the Greek letter μ).¹ This figure underscores that the Poisson distribution is a discrete distribution for non-negative integers, the exact qualities of count variables (i.e., a count variable can not be negative or fractional). Second, in Poisson regression the linear predictor

of the regression model (i.e., the right hand side of the regression equation) is connected to the outcome via a natural logarithm link function. Although this is not identical to transforming the outcome, it does mean that the regression coefficients from a Poisson model are on a log scale. Similar to logistic regression, raw coefficients are typically raised to the base e (i.e., exponentiated, the anti-log function) and interpreted as rate ratios. Like odds ratios, rate ratios are inversely proportion around one (i.e., a rate ratio of 3 is equal in strength but opposite in direction to a rate ratio of $1/3$). Rate ratio interpretation will be described in greater detail later, in the applications section.

The Poisson distribution (and regression) has an Achilles' heel of sorts in that it has the property that the mean equals the variance. In real data the variance often far exceeds the mean, and we would say that the data are over-dispersed relative to the Poisson distribution. Using descriptive statistics, dispersion is typically defined by the ratio of the variance to the mean. Thus, a Poisson distribution assumes a dispersion parameter of 1 (also called equi-dispersion). A dispersion parameter of three would indicate a variance value of three times the mean, which would (descriptively) indicate over-dispersion. Both of our example datasets show evidence of over-dispersion simply based on descriptive statistics (TLFB: $M = 1.2$, $Var = 9.3$; RAPI: $M = 6.3$, $Var = 82.9$), though over-dispersion can also be influenced by the longitudinal nature of these data and may be accounted for by covariates included in the model. Note that this description is meant to convey the intuitive ideas underlying over-dispersion, but formal tests can (and should) be used in regression modeling of count data (see, e.g., Hilbe, 2011).

The negative binomial model extends the Poisson model by allowing the mean and variance to be different. The lower row of Figure 4 presents three different negative binomial distributions. Note that each of these distributions has the same mean, but the dispersion varies, highlighting the primary difference between the Poisson and negative binomial distributions. The Poisson regression model is a special case of the negative binomial model, and when the mean equals the variance, the two will yield identical results. However, when the variance exceeds the mean, the negative binomial is more appropriate, and its standard errors will be reliably larger than those from the Poisson, reflecting the additional variance in the outcome. In practice, the Poisson model is rarely a good fitting model for exactly this reason, and selecting a Poisson model when the data are over-dispersed will yield overly liberal statistical tests (i.e., p -values will appear significant when they are not in reality, using a more appropriate model)².

As seen in Figure 4, the negative binomial regression model can fit highly skewed data, including data with a relatively large number of zeroes. However, when there is a clear stack of zeroes in the data and especially when the non-zero distribution is not a smooth extension from the zeroes, alternative models may be appropriate. Two closely related models explicitly handle count data with zeroes above and beyond what would be predicted by a negative binomial model: zero-inflated models and hurdle models (see e.g., Hilbe, 2011; Zeilis, Kleiber, & Jackman, 2008). Both models include a logistic regression for the zeroes in the data and a count regression (either Poisson or negative binomial) for the counts. However, zero-inflated and hurdle models take different approaches to dividing the data around zero. The predicted zeroes in a zero-inflated model come from both the count distribution and an extra mass of zeroes and are a type of mixture model in which the distribution at zero arises from two sources (i.e., count and logistic submodels). Thus, the

¹The six graphs in Figure 4 were created by simulating 1,000 random draws from the specified Poisson or negative binomial distributions. Thus, these are similar to frequency histograms, except there is no binning of multiple values.

²There are alternatives to the negative binomial distribution in terms of dealing with over-dispersed count data, including quasi-Poisson models, Poisson-normal models, and robust standard errors. Details on these models can be found in Hilbe (2011).

logistic regression model in a zero-inflated model is for “excess zeroes,” over and above what would be predicted by the count distribution. Hurdle models, on the other hand, model all the zeroes in the logistic regression, and non-zero counts are modeled by a truncated count regression (i.e., truncated because it does not include zero). Hurdle models are related to the more general class of models, often called “two-part models,” in which a logistic model for zero vs. non-zero is combined with a model for non-zero values. In the current manuscript we consider only Poisson and negative binomial for non-zero models, but this does not have to be the case (see, e.g., health care costs as in Buntin & Zaslavsky, 2004). In many instances zero-inflated models and hurdle models will yield similar results; however, hurdle models are more straightforward to interpret as all zeroes are handled in one portion of the model, and computationally, hurdle models are somewhat easier to fit as the two parts of the model can be fit independently of one another.

All of the count regression models introduced thus far assume independent observations. This assumption is violated in longitudinal data because repeated observations on the same individual will be correlated. To accommodate this non-independence of observations, mixed model extensions to count regressions can be used.

Generalized Linear Mixed Models for Count Outcomes

The top half of Figure 5 displays change across time in the RAPI for eight randomly selected individuals. As seen in that plot individuals started the study with a wide range of alcohol problems, and some individuals made notable changes in alcohol problems, whereas others did not. To illustrate the variability in intercepts (i.e., initial RAPI) and slopes (i.e., change over time in RAPI), Poisson regressions were fit to each individual's data (i.e., RAPI was regressed on time for each individual separately, using a Poisson regression; see Singer & Willett, 2003 for general discussion of this approach). The distributions of intercepts and slopes are plotted in the bottom of Figure 5. The means of these distributions are descriptive estimates of average intercepts ($M = 8.2$) and slopes ($M = 0.99$), and the distributions themselves highlight the variability across individuals in intercepts and slopes³.

GLMMs for longitudinal data extend this logic of individual growth curves by including subject-specific variability in one or more components of the model (e.g., intercepts, slopes) via additional variance terms that describe a distribution of regression coefficients across individuals within the study. GLMMs have been referred to as random coefficient models for precisely this reason. Using the systems of equations or hierarchical format with explicit level 1 and level 2 models, a Poisson GLMM for the RAPI data might be:

$$\begin{aligned}\text{Level 1: } \log(E[RAP I_{ti}]) &= \pi_{0i} + \pi_{1i} \text{Time} \\ \text{Level 2: } \pi_{0i} &= b_0 + b_1 \text{Male} + u_{0i} \\ \pi_{1i} &= b_2 + u_{1i}\end{aligned}\quad (1a)$$

which is identical to the composite form of the model by substituting the level 2 equations into their level 1 counterparts:

$$\log(E[RAP I_{ti}]) = b_0 + b_1 \text{Male} + b_2 \text{Time} + u_{0i} + u_{1i} \text{Time} \quad (1b)$$

where t indexes time, i indexes individuals, and the linear predictor (i.e., right hand side of equation) is connected to the mean of the outcome via a natural logarithm link function.⁴

³Note that the distribution of slopes appears somewhat odd, with spikes at extremely high or low values. This reflects that some individuals have incomplete data, and hence, slopes fit to two or three points of data can yield very extreme values. The subject-specific predictions from GLMMs pool information about the sample and the individual's data, leading to more sensible fits for individuals (see, e.g., discussion of empirical Bayes estimates in Raudenbush & Bryk, 2002).

Male is a dummy-variable for gender (female = 0, male = 1), and *Time* measures time in months since the start of the study. The two variance terms (u_{0i} , u_{1i}) describe the deviations of each participant's intercept and slope around the overall trajectory for the sample, defined by the fixed effects (i.e., $b_0 + b_1\text{Male} + b_2\text{Time}$). As is common with error terms, these variances are assumed (multivariate) normally distributed. Finally, the model assumes that the conditional distribution of the outcome given fixed and random-effects is Poisson distributed. It is important to note that both the fixed- and random-effects are connected to the outcome via the link function, which has implications for interpretation of the model (see discussion of marginal vs. conditional effects, below).

The results for the model in Equation 1 are presented in Table 1. Focusing first on the fixed-effects, the subject-specific rate ratios (RRs) are simply the exponentiated coefficients (e.g., for the intercept, $e^{1.45} = 4.28$). The intercept RR provides the estimated alcohol problems at baseline for women (i.e., when all covariates are equal to zero, as in all regression models), conditional on the random-effects. The RR of 1.29 for men indicates that their alcohol problems are 29% higher than women on average. Generally, the distance above or below 1 is interpreted as the percentage increase or decrease in the outcome for a one unit increase in the predictor. Similarly, the 0.97 RR for time implies a 3% reduction in RAPI for each one month increase. Let's consider the interpretation of RRs a bit more closely. The predicted regression lines based on the fixed-effects are shown in Figure 6, and the specific values are presented in the figure as well. If this were a linear mixed model or OLS regression, the predicted regression lines would be perfectly straight and parallel (i.e., because there is no interaction between *Male* and *Time*), but it is clear from the figure that the predicted regression lines for men and women from the Poisson GLMM are curved and are getting closer to one another over time. Poisson models with their log link functions are sometimes called multiplicative models, whereas OLS regression (and linear mixed models) are considered additive. For the latter, the regression coefficients tell us how much to "add" to our prediction for each one unit increase in a covariate.

The interpretation of multiplicative models is more complicated. First, we exponentiate the fixed effects in Equation 1b:

$$E[RAP I_{it}] = \exp(b_0 + b_1\text{Male} + b_2\text{Time}), \quad (1c)$$

where \exp is the exponentiating function, raising to the base e . We then re-express Equation 1c using a property of exponentiated sums:

$$E[RAP I_{it}] = \exp(b_0) * \exp(b_1\text{Male}) * \exp(b_2\text{Time}). \quad (1d)$$

Equation 1d shows why Poisson models are considered multiplicative: The exponentiated coefficients – the RRs – provide a multiplying factor for each unit change in the covariate. Thus, the RR of 1.29 for men indicates that men's predicted alcohol problems are 1.29 times women's predicted alcohol problems at each level of time, which is identical to saying that it is 29% higher (e.g., at time = 0, $4.28 \times 1.29 = 5.52$ and at time = 24 months, $2.03 \times 1.29 = 2.62$). For every 6 months, the model predicts alcohol problems are falling by $e^{-0.03 * 6} = 0.83$ (i.e., an RR for each 6 months change in time). We can similarly confirm that each 6-month change in predicted alcohol problems is 17% less (i.e., multiply by 0.83) than the preceding predicted value. Thus, for OLS regression and linear mixed models, coefficients

⁴The notation in equation 1 is slightly non-standard. In the statistical literature it is common to see GLMMs described in terms such as: $\eta = XB + Zb$, with $g(\mu) = \eta$, and $g() = \log$. Translating into words this means that η is the linear predictor, or right-hand side of the model, including both fixed and random-effects (i.e., XB and Zb , respectively). The linear predictor is connected to the mean of the outcome (i.e., μ) via some function. With the Poisson model, the link function is the log.

describe the amount added to predictions with each one unit covariate change, whereas Poisson and Poisson GLMM coefficients describe the amount multiplied.

One critical feature of these coefficients that has not been widely noted in the applied literature is that these coefficients (and their interpretation) are conditional on specific values of the random-effects distribution. In linear mixed models, fixed-effect coefficients can be interpreted as “averaging over” the random-effects (sometimes called marginal coefficients), but this is not true in GLMM with non-identity link functions (see, e.g., Heagerty & Zeger, 2000 and Raudenbush & Bryk, 2002, chapter 10). This issue of conditional vs. marginal coefficients (also called unit-specific vs. population-average) is discussed in the online supplemental material, along with additional resources.

There are several considerations in judging how well the model fits the data. First, similar to all regression models, we should consider whether we have the correct predictors included in the model and whether any terms have nonlinear associations with the outcome. These decisions should be informed by past research, theory, and thorough descriptive statistics and graphs. In fact, it is hard to over-emphasize the importance of thorough descriptive analyses to help guide decisions in building complicated models, such as those considered here. In addition, similar to many regression models, GLMMs assume that variance terms are normally distributed and homoskedastic, and plots of these assumptions can be found in the accompanying online materials (see eFigure 1 in online supplementary material). Second, similar to linear mixed models we should consider which terms in our model should have corresponding random-effects; with GLMMs decisions about which random-effects to include are usually determined by testing nested models via deviance tests (also called likelihood ratio chi-square tests; Molenberghs & Verbeke, 2005). Based on descriptive statistics and graphs as well as deviance tests for random-effects, the initial model for the RAPI was extended to include an interaction between gender and time and an additional per-observation random-effect (discussed below). These results are shown under Model 2 in Table 1.

Most (but not all) GLMMs are fit via a procedure called maximum likelihood, and when the iterative fitting procedure stops, a deviance statistic is reported (technically, $-2 \times \log$ -likelihood, and sometimes reported as $-2LL$). The difference between deviances from two nested models is distributed as a chi-square random variate and hence can be tested using a chi-square distribution with degrees of freedom equal to the difference in parameters between the models (see Singer & Willett, 2003, for further details). Deviance tests can also be used with count regression GLMMs, with one major caveat. Fitting GLMMs with non-normal outcomes (e.g., binary or count outcomes) is considerably more challenging, and there are a variety of optimization strategies and algorithms. Only certain algorithms have been shown to be accurate enough to allow deviance tests of nested models. A brief overview of algorithms and approaches is presented in the online supplemental material.

There are two final, critical aspects of model evaluation for GLMMs for count outcomes: over-dispersion and zeroes. As noted earlier, the Poisson GLMM assumes that, conditional on fixed and random-effects, the distribution of the outcome is distributed as a Poisson variable, with its property of equi-dispersion (i.e., mean equal to the variance). The most straightforward way to examine whether data are over-dispersed relative to the Poisson is to fit a model that allows for over-dispersed data. Logically, we might consider the negative binomial. For technical details that go beyond our focus here, the negative binomial model is challenging (though not impossible) to extend to random-effects (see, e.g., Hilbe, 2011, pp. 488–501), though some software packages are beginning to incorporate the negative binomial with random effects (e.g., Mplus, R, SAS). With GLMMs we can extend the Poisson model to include a per-observation error term, which captures over-dispersion. This

type of model is often called an over-dispersed Poisson model, which is functionally similar to a negative binomial model (see Rabe-Hesketh & Skrondal, 2008, pp. 389–390, or Ver Hoef & Boveng, 2007, for further discussion). Moreover, adding the extra error term is very similar to a residual error term, as in models for normally distributed data, and can be assessed via a deviance test. Model 2 in Table 1 fits an over-dispersed Poisson GLMM, incorporating this extra error term. A deviance test for assessing the improvement in fit due to the addition of the per-observation error term is highly significant ($\chi^2(1) = 2,032.8, p < .01$).

Earlier, we commented on the number of zeroes descriptively in the RAPI data, and a final component with count regression models is to examine how well the model replicates the number of zeroes, and more generally, the distribution of counts in the raw data; that is, what does the model predict the histogram of counts looks like, and specifically, what the zeroes look like?⁵ A figure included with the supplementary online material presents the raw distribution of the RAPI along with estimates from model two in Table 1 (see eFigure 2). Although model 2 appears quite accurate for counts greater than two, there is noticeable lack of fit for counts of zero, one, and two. Essentially, the Poisson distribution is not flexible enough to fit the steep drop of counts from zero to one, with much more moderate change for counts greater than one. As a result the raw RAPI data has 756 zeroes and 395 ones, whereas the final over-dispersed Poisson GLMM predicts 616 zeroes and 546 ones. This is a primary statistical motivation for considering a zero-inflated or hurdle mixed model.

Following our example with the RAPI, a hurdle mixed model would be:

$$\log\left(\frac{p}{1-p}\right) = b_{0l} + b_{1l}Male + b_{2l}Time + b_{3l}Male:Time + u_{0li} \quad (2a)$$

$$\log(E[RAPI_{it}|RAPI > 0]) = b_{0c} + b_{1c}Male + b_{2c}Time + b_{3c}Male:Time + u_{0ci} \quad (2b)$$

where t indexes time, i indexes individuals, l and c index logistic and count portions of the model, and p is the proportion of RAPI scores greater than zero. As noted earlier, zero-inflated and hurdle models have two submodels, one related to the zeroes and a second related to the counts. The key difference between hurdle and zero-inflated models is how they handle zeroes: Hurdle models cleanly divide the models, with all zeroes accounted for in the logistic regression, whereas zero-inflated models treat the zeroes as a mixture (i.e., both submodels in zero-inflated models contribute zeroes). As we saw with the over-dispersed Poisson GLMM, the hurdle GLMM presented above adds random-effects to the linear predictors, with the major difference now being that there are two linear predictors. The random intercept in the logistic model implies that there is variability across individuals in the likelihood of reporting any problems, whereas the random intercept and slope in the count regression models variability in the intercept and change across time in number of

⁵Estimating the predicted number of zeroes based on a GLMM takes us into somewhat more technical material. The Poisson GLMM can be represented by:

$$Pois(XB + Zb) \quad (4)$$

where X is the design matrix of the fixed-effects; B is a vector of fixed-effects coefficients; Z is the design matrix of the random-effects; and b is a vector of random-effects variances. The key thing to realize is that the formula within the parentheses defines the fitted values of the model. Thus, we can read this equation as: The data are distributed as a Poisson random variable with a mean structure defined by the fixed and random-effects of the GLMM. Similar to the discussion of conditional effects, the challenge in equation 4 is how best to handle the random-effects. The accompanying R code shows one method for simulating from the fitted model to estimate the predicted distribution of counts.

problems when there are problems reported. A random slope for time in the logistic portion would model individual variability over time in proportion of zeroes.

Prior to fitting the hurdle mixed model, let us say a word about when to use zero-inflated or hurdle models. Similar to many model selection decisions, the choice between models should include statistical considerations, theoretical considerations, and parsimony. As noted earlier the over-dispersed Poisson GLMM of the RAPI data under-predicted zeroes (756 in the raw data vs. 616 predicted by the model), and with the TLFB data, the same type of model led to a highly non-normal distribution for the per-observation error term, based on a histogram and quantile-quantile plot of this error term. Thus, in both of these cases a zero-inflated or hurdle mixed model might be preferred on statistical grounds. However, the statistical motivations just noted do not automatically mean that covariates have important and distinct relationships across the logistic and count portions of a hurdle (or zero-inflated) model. That is, sometimes a hurdle or zero-inflated mixed model will fit the data better, but the conclusions are largely the same compared to a Poisson GLMM (typically because the covariates are primarily related to the nonzero counts). In this scenario, parsimony might suggest the simpler Poisson GLMM is adequate as it yields similar conclusions, though we strongly suggest fitting the more complex model to test whether the simpler model is adequate. Finally, theory may make clear predictions about the proportion of days with any drinking or the amount drunk on days with drinking. These would clearly prefer hurdle or zero-inflated models.

Results from the hurdle mixed model shown in equations 2a and 2b are presented in Table 2.⁶ In examining Table 2, we find that these results map on to the earlier descriptive graphs quite closely. In particular, the count portion of the model shows that men tend to have more alcohol problems when there are problems at the start of the study ($RR = 1.21$). Because of the coding of the gender dummy-variable, the main effect of time describes the change in women's alcohol problems across time. The RR of 0.98 implies that, conditional on the random-effects, women's RAPI (i.e., alcohol problems when there are problems) is dropping 2% with each successive assessment, whereas the interaction between gender and time implies that men's alcohol problems (when there are problems) are barely changing at all, which is confirmed by estimating the men's simple slope for time ($RR = 0.99$, 95% CI: 0.98, 1.00).

The logistic portion of the model describes the proportion of the sample reporting some alcohol problems. (Note that software for zero-inflated and hurdle models are not always consistent as to whether the logistic portion predicts zeroes as opposed to non-zeroes. Here it is predicting non-zeroes.) The odds ratios (OR) in Table 2 show that there are no differences between the sexes in the proportion of individuals reporting alcohol problems at the start of the study ($OR = 0.83$, 95% CI = 0.59, 1.18). For women, the odds of reporting any alcohol problems decreases by 3% per month ($OR = 0.97$, 95% CI = 0.95, 0.99), conditional on the random-effects. For men, their rate of change is somewhat slower ($OR = 1.02$, 95% CI = 0.99, 1.04).

TLFB Application

In turning to the TLFB data, the earlier plots showed that there was a very large stack of zeroes and that covariate effects (i.e., gender, weekend, and fraternity/sorority status) might

⁶We note in passing that the hurdle mixed models reported in the manuscript were fit using the `MCMCglmm` package (Hadfield, 2010) in R, which uses a Bayesian approach to model estimation. For many practical problems maximum likelihood estimation and Bayesian MCMC estimation will yield similar if not identical results, though there are basic differences in both inferential philosophy as well as estimation. A brief discussion is provided in the technical appendix of the online material. In this particular instance, the choice is quite pragmatic as `MCMCglmm` provides the most flexible option for fitting these types of models in R.

vary across submodels defined by likelihood of any drinking (i.e., zero vs. not zero) and amount of drinking when there is drinking. These considerations would point toward a zero-inflated or hurdle model; moreover, preliminary models using an over-dispersed Poisson GLMM showed that there was a very poor fit between predicted and observed count distribution, and the per-observation random-effect distribution was not close to being normally distributed. Given these considerations, a hurdle over-dispersed Poisson GLMM was fit to the TLFB data, fitting main and 2-way interaction effects in both submodels via three dummy variables: Men (0 – Women, 1 – Men), Weekend (0 – Sun/Wed, 1 – Thurs/Sat), and FratSor (0 – Not in fraternity/sorority, 1 – Fraternity/Sorority). A random intercept and random slope for weekend were included in both the logistic and truncated count submodels, as well as the per-observation random intercept in the truncated count submodel. Conditional (or unit-specific) fixed-effects are shown in Table 3. In the logistic portion of the model, there is a single, significant effect for weekend: Many more students drink, regardless the quantity, on the weekends as compared to the weekdays ($OR = 3.06$, 95% CI = 2.60, 3.54). In the count submodel, all three 2-way interactions are significant. One approach to interpreting interactions is to use “simple slopes” (see, e.g., Jaccard, 2001). However, in the present case, the model is estimating mean drinking on drinking days for eight specific groups – for each combination of gender, weekend, and fraternity / sorority status. To help interpret the findings of the count submodel, we estimated marginal (or population-average),⁷ predicted means for each of these eight groups, which are shown in Figure 7. The figure reveals several interesting findings. Not surprisingly, fraternity and sorority members drink more on drinking days than students who are not in fraternities and sororities, and fraternity and sorority members drinking is similar on weekdays vs. weekend – when they are drinking (i.e., the count submodel in a hurdle is only for non-zeroes). However, contrasts of weekend vs. weekday drinking for men and women who are not in fraternities and sororities shows that they do drink more on weekends (women: $RR = 1.07$, 95% CI = 0.99, 1.14, $p = .06$; men: $RR = 1.20$, 95% CI = 1.12, 1.28, $p < .01$). Thus, the hurdle mixed model provides a more nuanced view of drinking differences across these groups, parsing out drinking vs. not drinking and mean drinking on drinking days.

Software for GLMMs

Statistical software packages are steadily expanding their coverage of GLMMs and the models covered in the current paper. Table 4 presents an overview of SPSS, Stata, R, SAS, and Mplus, describing which of the models covered in the present paper are possible to fit. As seen there, the basic Poisson GLMM is implemented in all software, whereas negative binomial GLMM and zero-altered models are less widely available. The online materials provide computer code for fitting some example models in each of the packages mentioned. The online supplementary material covers additional details related to fitting GLMM, which is both more technical but still practically important. Not all software packages were able to fit all models described in the present article (e.g., SAS NLMIXED can fit zero-inflated mixed models but could not fit all the variance components describe in the TLFB example). Over the coming years, there will likely be greater software availability for fitting the types of models covered here as well as more robust fitting algorithms.

Summary

The present article discussed extensions to count regression and zero-altered count regression models to longitudinal data based on GLMM. We hope that this presentation, along with the appendix and available data and code, helps addiction researchers to learn

⁷To estimate marginal (or population-average) estimates from a GLMM, it is necessary to include both fixed and random-effects. The online supplementary material provides details and the example of how the present estimates were created.

and appropriately apply these models. Given the research designs and data that addiction researchers often collect, the models covered here are often the most appropriate analytic tools. At the same time, like other statistical models that are finding their way to the applied research literature (e.g., growth mixture models; Muthén & Muthén, 2000), multilevel count models are complex. The types of assumptions and general considerations in model fitting grow exponentially from linear models to generalized linear models to linear mixed models to generalized linear mixed models. We encourage researchers to explore and use these tools as their hypotheses and data merit, though also note that they require a serious commitment of time and study to be used appropriately.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Eun-Young Mun, Zac Imel, Isaac Rhew, Jennifer Kirk, and an anonymous reviewer for helpful feedback that improved the manuscript in numerous ways. The example data included in the article were collected via NIAAA grants AA016099 and AA014576 (Clayton Neighbors, PI), and David Atkins' time was supported in part by NIAAA grant AA019511 (Eun-Young Mun, PI).

References

- Atkins DC, Gallop RJ. Rethinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models. *Journal of Family Psychology*. 2007; 21:726–735. [PubMed: 18179344]
- Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JS. Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology and Evolution*. 2009; 24(3):127–135. [PubMed: 19185386]
- Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*. 1993; 88(421):9–25.
- Buntin MB, Zaslavsky AM. Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures. *Journal of Health Economics*. 2004; 23:525–542. [PubMed: 15120469]
- Coxe S, West SG, Aiken LS. The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of Personality Assessment*. 2009; 91:121–136. [PubMed: 19205933]
- Draper, D. Bayesian multilevel analysis and MCMC. In: de Leeuw, J.; Meijer, E., editors. *Handbook of Multilevel analysis*. New York, NY: Springer; 2008. p. 77-139.
- Gaher RM, Simons JS. Evaluations and expectancies of alcohol and marijuana problems among college students. *Psychology of Addictive Behaviors*. 2007; 21:545–554. [PubMed: 18072837]
- Gelman, A.; Hill, J. *Data Analysis Using Regression and Multilevel / Hierarchical Models*. New York, NY: Cambridge University Press; 2007.
- Heagerty PJ, Zeger SL. Marginalized multilevel models and likelihood inference. *Statistical Sciences*. 2000; 15:1–26.
- Hedeker, D.; Gibbons, RD. *Longitudinal data analysis*. Hoboken, NJ: John Wiley & Sons; 2006.
- Hilbe, JM. *Negative binomial regression*. 2nd ed.. New York, NY: Cambridge University Press; 2011.
- Lewis MA, Neighbors C, Geisner IM, Lee CM, Kilmer JR, Atkins DC. Examining the associations among severity of injunctive drinking norms, alcohol consumption, and alcohol-related negative consequences: The moderating roles of alcohol consumption and identity. *Psychology of Addictive Behaviors*. 2010; 24:177–189. [PubMed: 20565144]
- Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73:13–22.
- Lynch, S. *Introduction to applied Bayesian statistics and estimation for social scientists*. New York, NY: Springer; 2007.

- McCullagh, P.; Nelder, JA. Generalized Linear Models. 2nd ed.. Boca Raton, FL: Chapman & Hall / CRC; 1989.
- Molenberghs, G.; Verbeke, G. Models for discrete longitudinal data. New York, NY: Springer; 2005.
- Muthén B, Muthén L. Integrating person-centered and variable-centered analysis: growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research*. 2000; 24:882–891.
- Neal DJ, Simons JS. Inference in regression models of heavily skewed alcohol use data: A comparison of ordinary least squares, generalized linear models, and bootstrap resampling. *Psychology of Addictive Behaviors*. 2007; 21:441–452. [PubMed: 18072826]
- Neighbors C, Lewis MA, Atkins DC, Jensen MM, Walter T, Fossos N, et al. Efficacy of web-based personalized normative feedback: A two-year randomized controlled trial. *Journal of Consulting and Clinical Psychology*. 2010; 78:898–911. [PubMed: 20873892]
- Neighbors C, Atkins DC, Lewis MA, Lee CM, Kaysen D, Mittmann A, Fossos N, Rodriguez LM. Event specific drinking among college students. *Psychology of Addictive Behaviors*. 2011; 25:702–707. [PubMed: 21639597]
- Rabe-Hesketh, S.; Skrondal, A. Multilevel and longitudinal modeling using Stata. College Station, TX: Stata Press; 2008.
- Raudenbush, SW. Many Small Groups. In: de Leeuw, J.; Meijer, E., editors. *Handbook of Multilevel Analysis*. New York, NY: Springer; 2008. p. 207-236.
- Raudenbush, SW.; Bryk, AS. Hierarchical linear models: Applications and data analysis methods. 2nd ed.. Thousand Oaks, CA: Sage Publications, Inc.; 2002.
- Rodriguez G, Goldman N. Improved estimation procedures for multilevel models with binary response: A case-study. *Journal of the Royal Statistics Society, Series A*. 2001; 164:339–355.
- Simons JS, Neal DJ, Gaher RM. Risk for marijuana-related problems among college students: An application of zero-inflated negative binomial regression. *American Journal of Drug and Alcohol Abuse*. 2006; 32:41–53. [PubMed: 16450642]
- Singer, JD.; Willett, JB. *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press; 2003.
- Snijders, TAB.; Bosker, RJ. *An Introduction to Basic and Advanced Multilevel Modeling*. Thousand Oaks, CA: Sage Publications; 1999.
- White HR, Labouvie EW. Towards the assessment of adolescent problem drinking. *Journal of Studies on Alcohol*. 1989; 50:30–37. [PubMed: 2927120]
- Ver Hoef JM, Boveng PL. Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*. 2007; 88:2766–2772. [PubMed: 18051645]
- Zeileis A, Kleiber C, Jackman S. Regression models for count data in R. *Journal of Statistical Software*. 2008; 27(8):1–25.

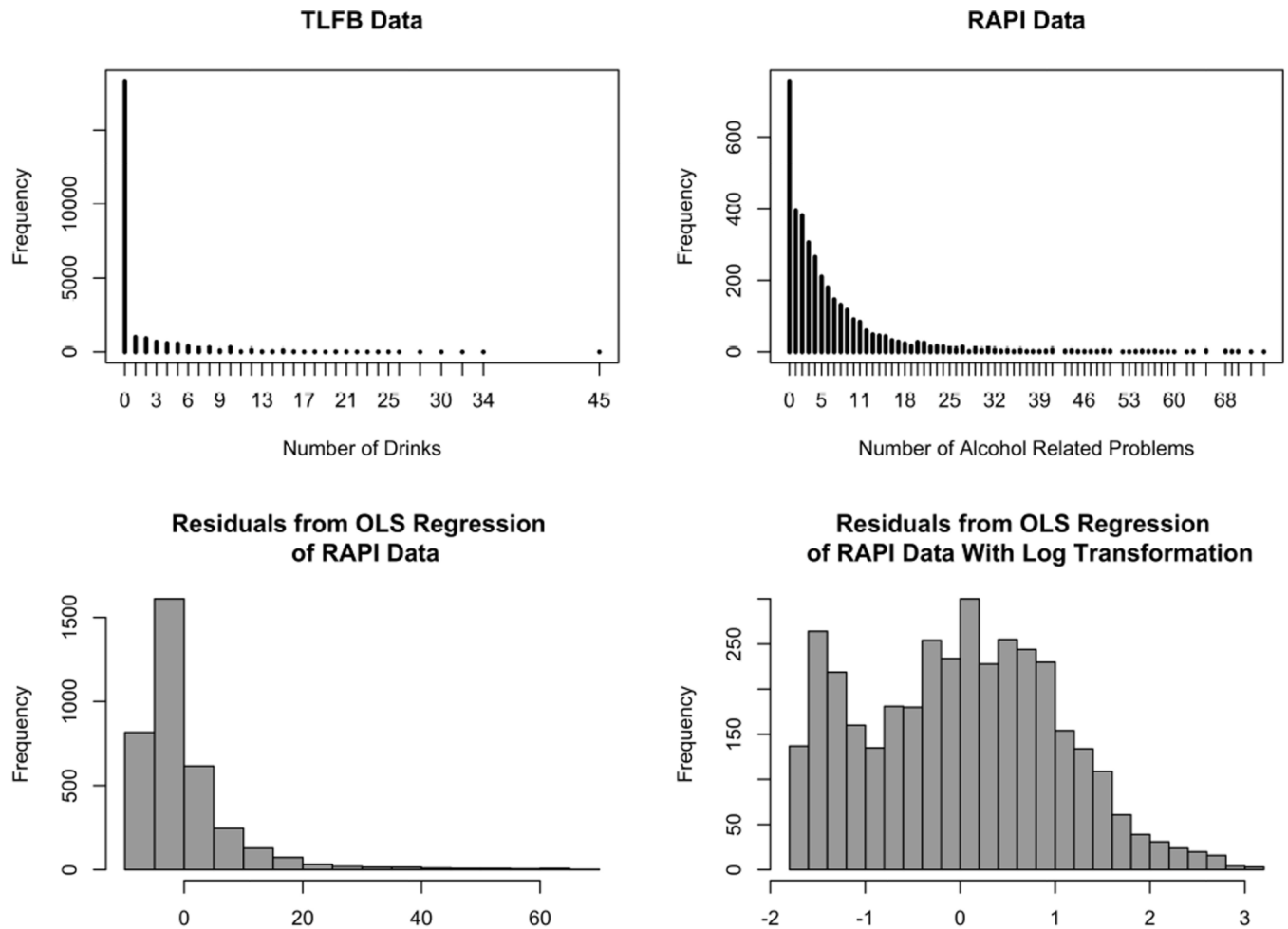


Figure 1. Plots of Frequency Counts of Daily Drinks from Timeline Follow-Back (TLFB; upper left) and Rutgers Alcohol Problems Index (RAPI; upper right). Residuals from fitting an ordinary least squares regression to the RAPI or log-transformed RAPI are in lower left and lower right, respectively.

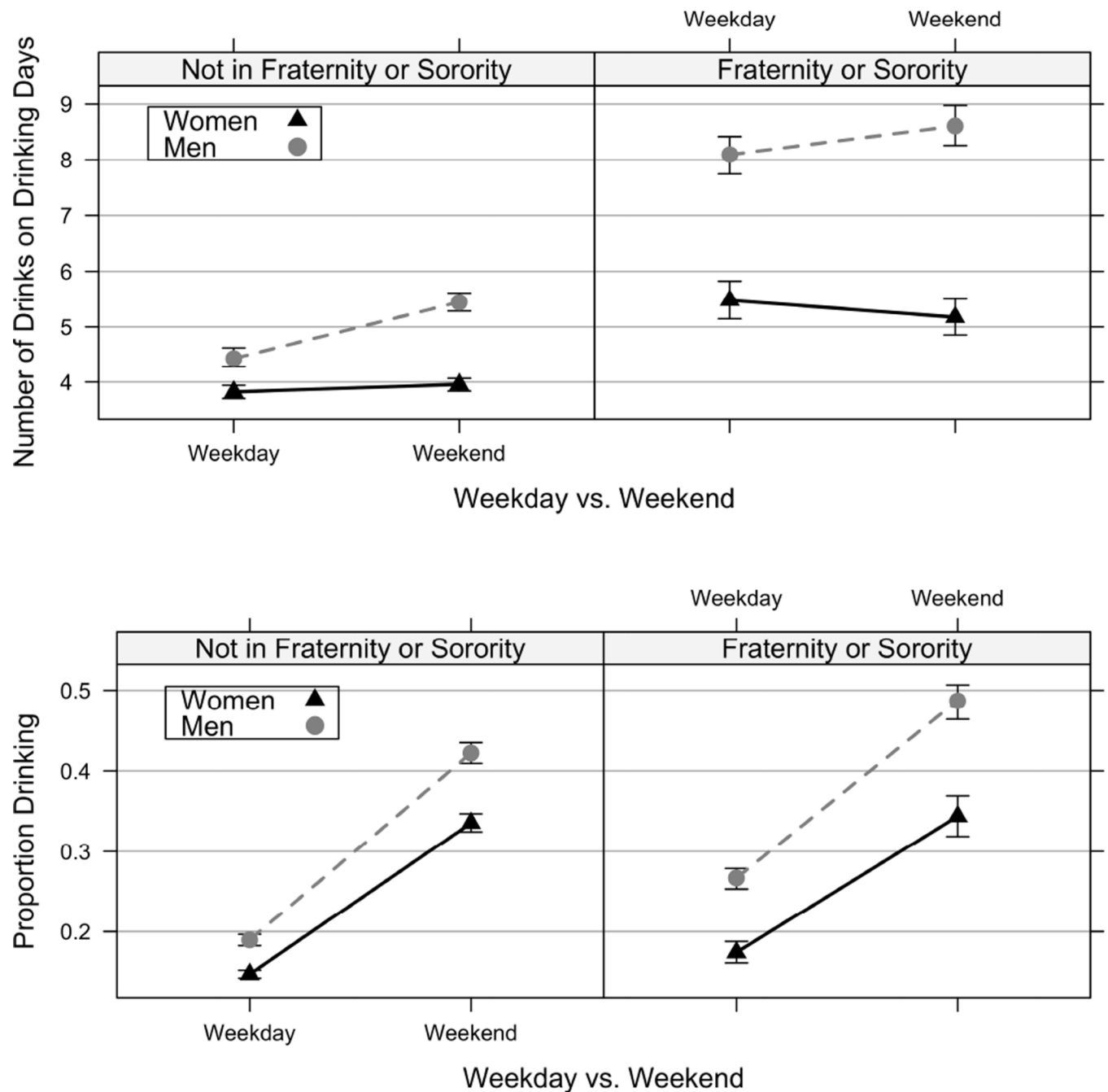


Figure 2. Means and 95% Confidence Intervals for Drinking on Drinking Days (top row) and Proportion of Individuals Drinking (bottom row). Means are stratified by weekend vs. weekday, male vs. female, and fraternity / sorority member or not.

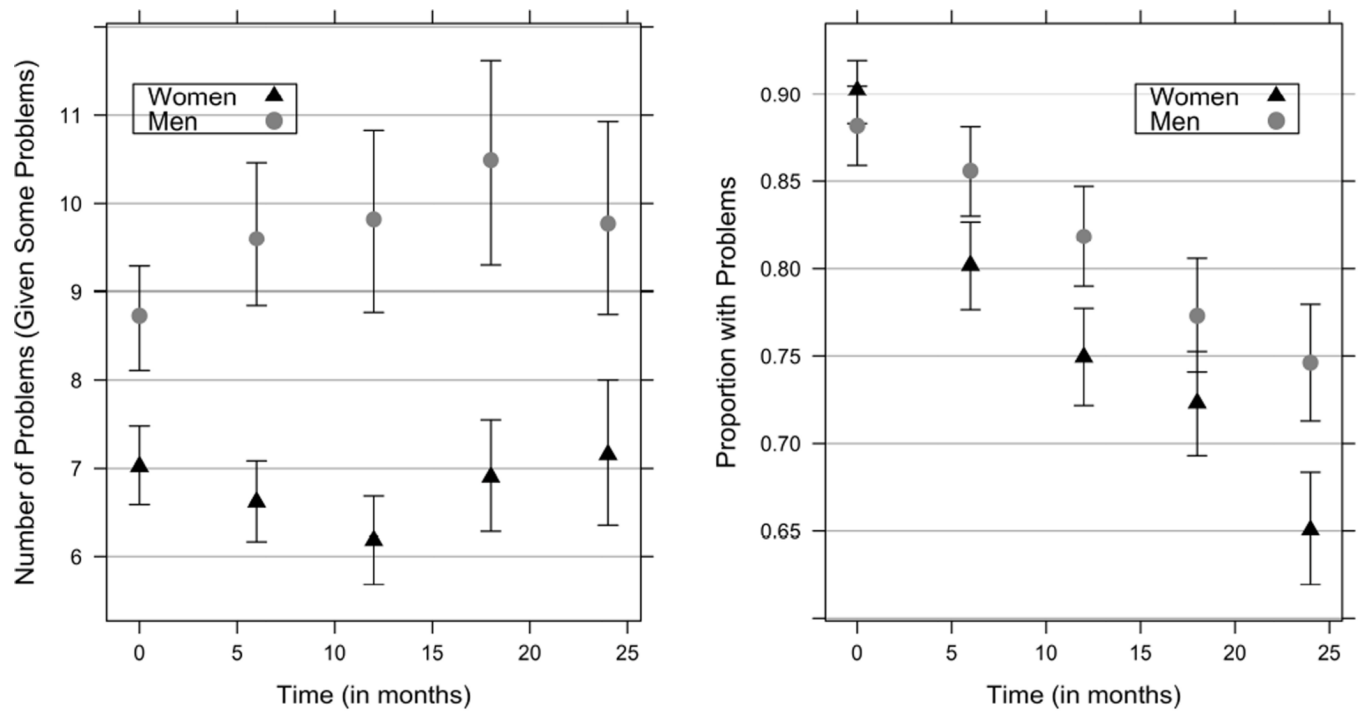


Figure 3. Means and 95% Confidence Intervals for Number of Problems when there are any problems (left) and Proportion of Individuals reporting any drinking-related problems (right). Means are stratified by assessment period and gender.

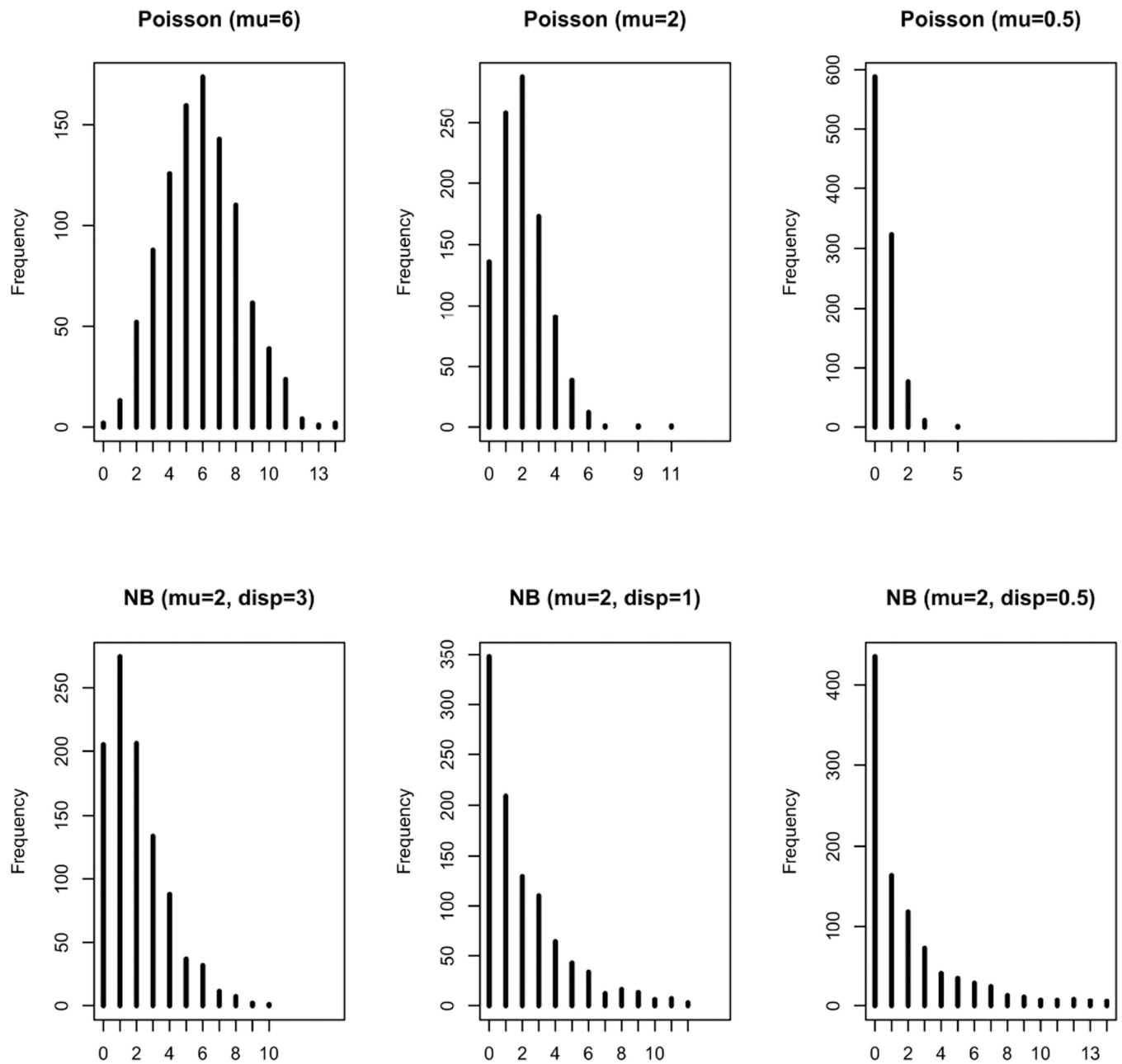
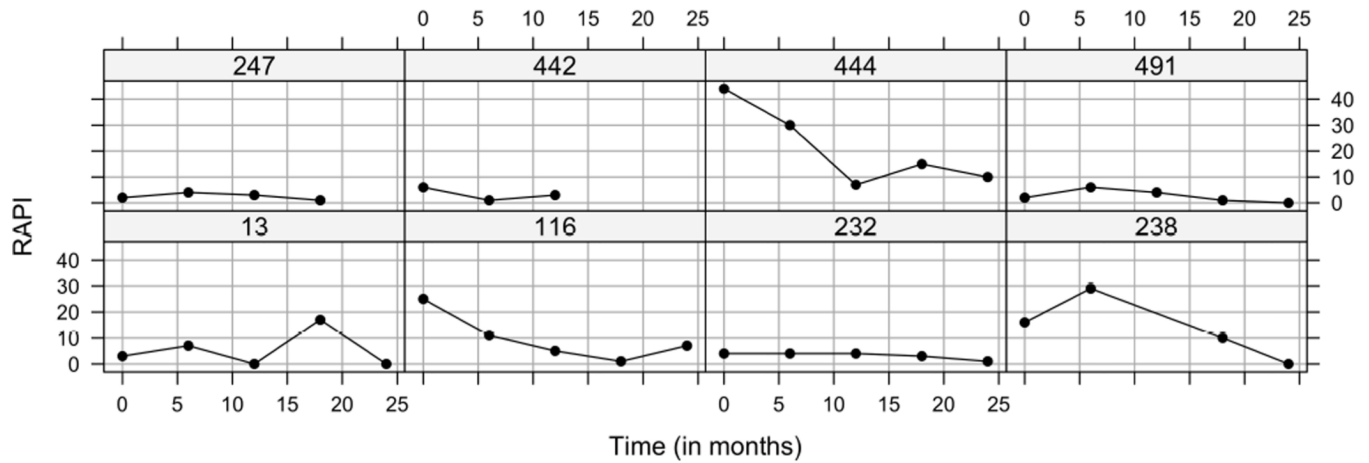


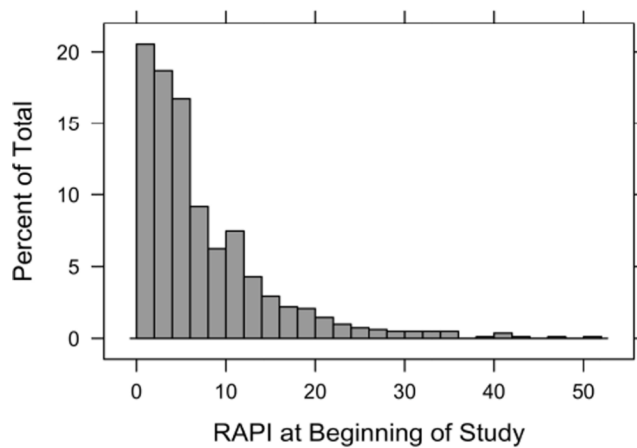
Figure 4.

Plots of frequency counts of data simulated from Poisson distributions with three different means (top row). The bottom row contains frequency counts of data simulated from negative binomial distributions with the same mean but varying dispersion.

Trajectories for 8 Randomly Selected Individuals



Distribution of Intercepts



Distribution of Slopes

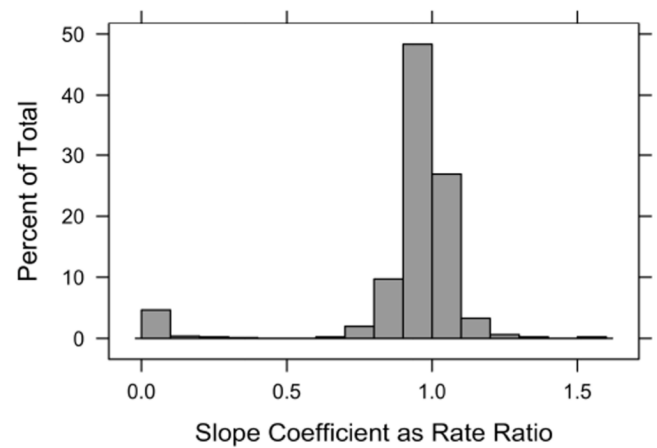


Figure 5.

RAPI scores plotted over time for 8 randomly selected individuals (top row). Distribution of intercepts (bottom left) and slopes (bottom right) from fitting separate Poisson regressions to each individual's data.

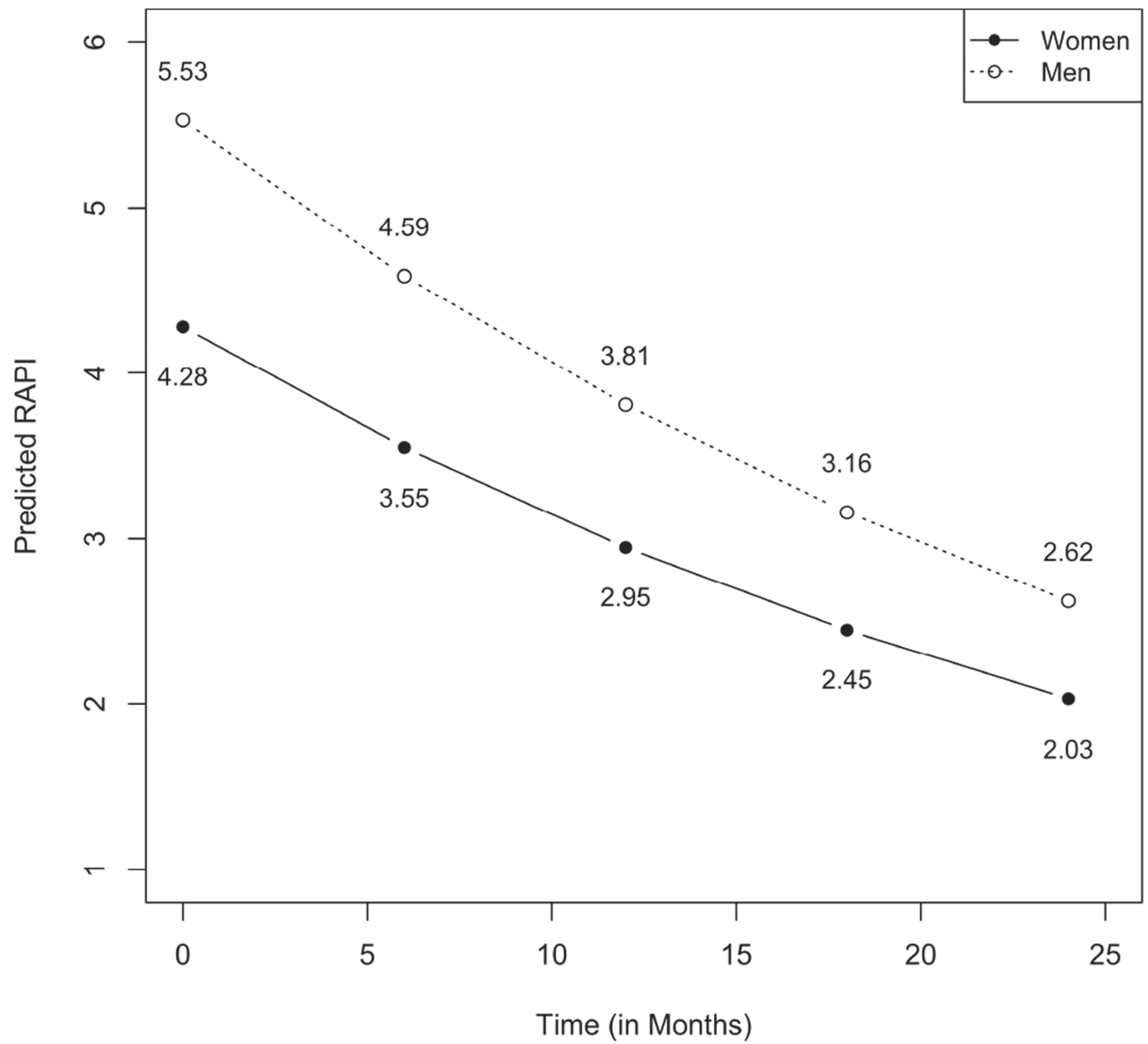


Figure 6. Predicted (unit-specific) RAPI scores from Poisson GLMM for men and women over time, with specific values noted in text.

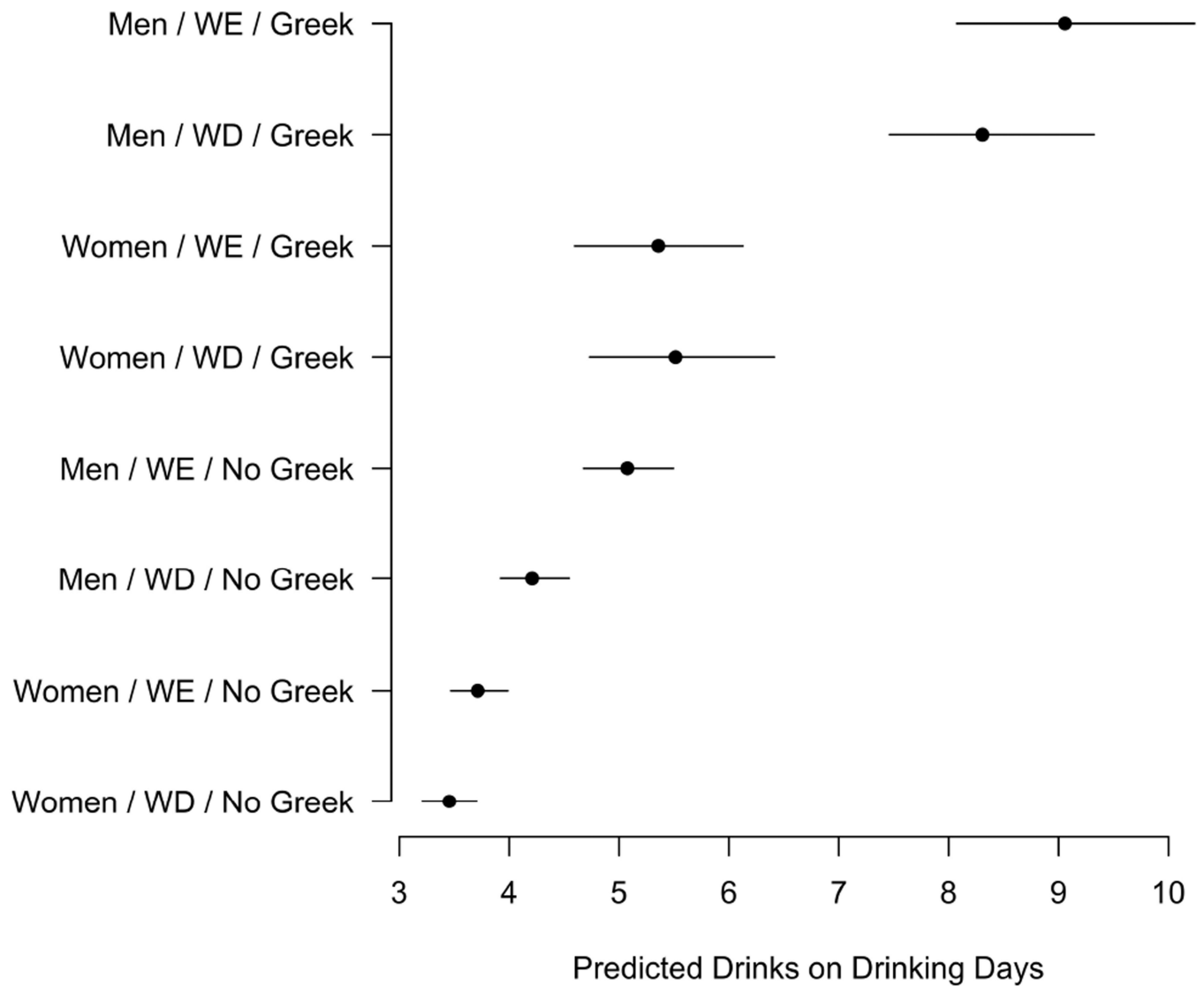


Figure 7. Predicted (marginal) drinks for each unique combination of gender, weekend (WE) vs. weekday (WD), and fraternity / sorority status from count submodel of over-dispersed Poisson hurdle mixed model.

Table 1

GLMM Poisson Results for RAPI Data

<i>Model 1</i>				
	RR ^a	B	95% CI for RR	
			<i>Lower</i>	<i>Upper</i>
Intercept	4.28	1.45	3.90	4.70
Male	1.29	0.26	1.12	1.49
Time	0.97	−0.03	0.96	0.97
<i>Model 2</i>				
	RR ^a	B	95% CI for RR	
			<i>Lower</i>	<i>Upper</i>
Intercept	4.00	1.39	3.64	4.40
Male	1.22	0.20	1.06	1.41
Time	0.96	−0.04	0.96	0.97
Male × Time	1.02	0.02	1.01	1.03

Note. B = Coefficient on linear-predictor scale (i.e., log of outcome); RR = Rate ratio; 95% CI = 95% confidence interval.

^aRRs and 95% CI are unit-specific (or conditional) estimates, as opposed to population-average (or marginal) estimates. See Appendix for further details.

Table 2

Over-dispersed Poisson Hurdle Mixed Model Results for RAPI Data

<i>Count Sub-model</i>				
	RR ^a	B	95% CI for RR	
			<i>Lower</i>	<i>Upper</i>
Intercept	4.72	1.55	4.27	5.15
Male	1.21	0.19	1.08	1.40
Time	0.98	−0.02	0.98	0.99
Male × Time	1.01	0.01	1.00	1.02
<i>Logit Sub-model</i>				
	OR ^a	B	95% CI for OR	
			<i>Lower</i>	<i>Upper</i>
Intercept	4.76	1.56	3.74	6.47
Male	0.83	−0.19	0.59	1.18
Time	0.97	−0.03	0.95	0.99
Male × Time	1.02	0.02	0.99	1.04

Note. B = Coefficient on linear-predictor scale (i.e., log of outcome); RR = Rate ratio; OR = Odds ratio; 95% CI = 95% confidence interval.

^aRRs, ORs, and 95% CI are unit-specific (or conditional) estimates, as opposed to populationaverage (or marginal) estimates. See Appendix for further details.

Table 3

Over-dispersed Poisson Hurdle Mixed Model Results for TLFB Data

<i>Count Submodel</i>				
	RR ^a	B	95% CI	
			<i>Lower</i>	<i>Upper</i>
Intercept	2.62	0.98	2.49	2.87
Weekend	1.08	0.07	0.99	1.16
Male	1.22	0.20	1.09	1.33
Greek	1.63	0.47	1.37	1.84
Weekend × Male	1.12	0.12	1.03	1.23
Weekend × Greek	0.90	−0.10	0.82	1.00
Male × Greek	1.19	0.21	1.01	1.49

<i>Logit Sub-model</i>				
	OR ^a	B	95% CI	
			<i>Lower</i>	<i>Upper</i>
Intercept	0.09	−2.40	0.08	0.10
Weekend	3.06	1.11	2.60	3.55
Male	1.10	0.10	0.89	1.32
Greek	0.82	−0.23	0.59	1.12
Weekend × Male	0.99	0.00	0.81	1.25
Weekend × Greek	0.87	−0.12	0.68	1.11
Male × Greek	0.99	0.06	0.71	1.62

Note. B = Coefficient on linear-predictor scale (i.e., log of outcome); RR = Rate ratio; OR = Odds ratio; 95% CI = 95% confidence interval.

^aRRs, ORs, and 95% CI are unit-specific (or conditional) estimates, as opposed to population-average (or marginal) estimates. See Appendix for further details.

Table 4
Comparison of Software for Fitting Generalized Linear Mixed Models (GLMM)

	SPSS	Stata	R ^a	SAS	Mplus
Poisson	GENLINMIXED	xtmepoisson	lme4 glmmADMB MCMCglmm	GLIMMIX	Count is outcome (p)
OD Poisson	GENLINMIXED	xtmepoisson	lme4 glmmADMB MCMCglmm	GLIMMIX	
Negative Binomial	GENLINMIXED		glmmADMB	GLIMMIX	Count is outcome (nb)
ZIP			glmmADMB ^b MCMCglmm	NLMIXED	Count is outcome (pi)
ZINB			glmmADMB MCMCglmm ^c	NLMIXED	Count is outcome (nbi)
HUP			glmmADMB ^d MCMCglmm	NLMIXED	
HUNB			glmmADMB MCMCglmm	NLMIXED	Count is outcome (nbh)

Note. OD Poisson = Over-dispersed Poisson; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; HUP = Hurdle Poisson; HUNB = Hurdle negative binomial

^aR packages are listed that include functions for fitting the models described, and the supplementary R code provides examples of specific commands. Because R has many user-contributed packages, there are additional packages that can fit some of the described models.

^bR package glmmADMB does not currently allow covariates or random-effects in the logit submodel (as of glmmADMB_0.7.2.12).

^cR package MCMCglmm fits an OD Poisson model as opposed to a negative binomial model for zero-inflated and hurdle models.

^dFor hurdle models and glmmADMB, it is necessary to manually create the two outcomes (i.e., logit and truncated count) and the fit the two submodels as separate models.