



Two-level negative binomial model

Contents

3.1	INTRODUCTION	1
3.1.1	<i>Negative binomial distribution</i>	1
3.1.2	<i>Adaptive versus non-adaptive quadrature</i>	2
3.1.3	<i>The data</i>	2
3.1.4	<i>A 2-level negative binomial model with 2 predictors</i>	4

3.1 Introduction

A count variable counts the number of discrete occurrences of a characteristic of interest that takes place during a time interval. Examples are the occurrence of cancer cases in a hospital during a given period of time, the number of cars that pass through a toll station per day, and the phone calls at a call center. The most common distribution for a count variable is the Poisson distribution. Besides the Poisson distribution, negative binomial distributions may also be used to describe the properties of count variables. In this chapter, models for count data, utilizing both the Poisson and negative binomial distributions, are discussed.

3.1.1 Negative binomial distribution

The negative binomial distribution is a probability distribution used to describe a certain number of failures and successes in a series of independent and identically distributed Bernoulli trials. Specifically, for $k+r$ Bernoulli trials with success probability p , the negative binomial gives the probability of k failures and r successes, with success on the last trial. In other words, the negative binomial distribution is the probability distribution of the number of failures before the r^{th} success in a Bernoulli process, with probability p of success on each trial.

The negative binomial distribution can be expressed as

$$f(y_i) = \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i + 1)\Gamma(1/\alpha)} \times \frac{(\alpha\mu_i)^{y_i}}{(1 + \alpha\mu_i)^{y_i + 1/\alpha}}$$

with $\Sigma(y_i) = \mu_i + \alpha\mu_i^2$, where $\Gamma(\cdot)$ is the gamma function or generalized factorial from advanced calculus, and where α denotes an additional parameter and it can no longer be assumed that the variance is a known function of the mean. In the example to follow, α is assumed to have a fixed value.

3.1.2 Adaptive versus non-adaptive quadrature

Ordinary quadrature is a numeric method for evaluating multi-dimensional integrals. For mixed-effect models with count and categorical outcomes, the log-likelihood function is expressed as the sum of the logarithm of integrals, where the summation is over higher-level units, and the dimensionality of the integrals equals the number of random effects.

A problem with ordinary quadrature is that it assumes a common location and scale for each level-2 unit. This assumption often requires the use of a large number of quadrature points to calculate the log-likelihood and derivatives to an acceptable level of accuracy. To overcome this problem with ordinary quadrature, SuperMix also offers a numeric integration procedure called adaptive quadrature. The adaptive quadrature procedure uses the empirical Bayes means and covariances, updated at each iteration to essentially shift and scale the quadrature locations of each higher-level unit in order to place them under the peak of the corresponding integral. To distinguish between the two quadrature methods, SuperMix uses the terminology non-adaptive quadrature (ordinary quadrature) and adaptive quadrature. The model fitted here will use non-adaptive quadrature.

3.1.3 The data

The data set is from the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC), which was designed to be a longitudinal survey with its first wave fielded in 2001–2002. This data file has been used in some of the examples in Chapter 3, and contains information on the occurrences of major depression, family history of major depression and dysthymia of 2339 dysthymia respondents. After listwise deletion, the sample size is 1981.

	(A) PSU	(B) FINWT	(C) CONC	(D) AGE D	(E) N DEP
1	1011	7256.15	0	51	1
2	1011	3476.67	1	48	1
3	1011	3052.10	1	59	1
4	1011	1182.03	1	36	2
5	1011	3041.05	1	17	1
6	1011	8342.94	0	16	1
7	1011	6767.06	1	29	1
8	1011	3460.29	1	43	1
9	1015	3167.29	1	55	1
10	1018	1014.56	0	37	1

The variables of interest are:

- PSU denotes the Census 2000/2001 Supplementary Survey (C2SS) primary sampling unit.
- FINWT represents the NESARC weights sample results used to form national level estimates. The final weight is the product of the NESARC base weight and other individual weighting factors.
- CONC_DEP contains the information captured in field S4CQ3A6 of the NESARC data. It represents the response to the statement "Often had trouble concentrating/keeping mind on things," with 1 indicating "Yes," and 0 indicating "No."
- AGE_DEP is based on field S4CQ7AR of the NESARC data. It represents the age at onset of first episode.
- N_DEP is recoded from field S4CQ6A of the NESARC data, and gives the number of depression/dysthymia episodes. This is the count variable we would like to use as outcome variable in the examples to follow.

3.1.3.1 Exploring the data

Inspecting the distribution of the intended outcome variable, N_DEP, before starting with the model is important. In the case of a count variable, this can easily be done by producing a bar chart of the observed frequencies of occurrence captured by the count variable. Select the **File, Data-based Graph, Univariate** option from the main SuperMix window and request a bar chart before clicking the **Plot** button.

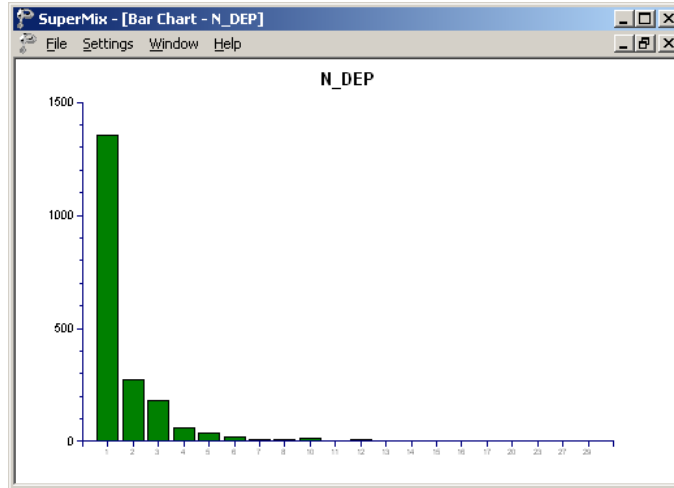


Figure 5.3: Bar chart for count variable N_DEP

The frequency bar chart for the count variable N_DEP shown in Figure 5.3 is obtained. We note that the number of depression episode ranges from 1 to 29, with most respondents having a small number of reported episodes of depression.

3.1.4 A 2-level negative binomial model with 2 predictors

3.1.4.1 The model

In a previous example, a Poisson model was fitted to the data. It was also noted that a Poisson distribution has an important property: the mean number of occurrences is equal to the variance. The negative binomial distribution is an alternative distribution that may also be used to describe the properties of count variables. If the assumption of a Poisson distribution is reasonable, one would expect a model using a negative binomial distribution with a very small dispersion parameter to produce results that correspond closely to those obtained for the Poisson model. In this section, we fit a negative binomial model, utilizing the same predictors and a small dispersion parameter, to the NESARC data. Again, adaptive quadrature is used as the method of optimization.

Recall that the negative binomial distribution can be expressed as

$$f(y_i) = \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i + 1)\Gamma(1/\alpha)} \times \frac{(\alpha\mu_i)^{y_i}}{(1 + \alpha\mu_i)^{y_i + 1/\alpha}}$$

with $\Sigma(y_i) = \mu_i + \alpha\mu_i^2$ where α denotes an additional parameter and it can no longer be assumed that the variance is a known function of the mean. We assume α to be a fixed parameter.

The model fitted to the data explores the relationship between N_DEP and the variables indicating concentration (or lack thereof) and age, as represented by the variables CONC_DEP and AGE_DEP.

The level-1 model is

$$\log\left[E\left(N_DEP_{ij}\right)\right] = \beta_0 + \beta_1 \times CONC_DEP_{ij} + \beta_2 \times AGE_DEP_{ij}$$

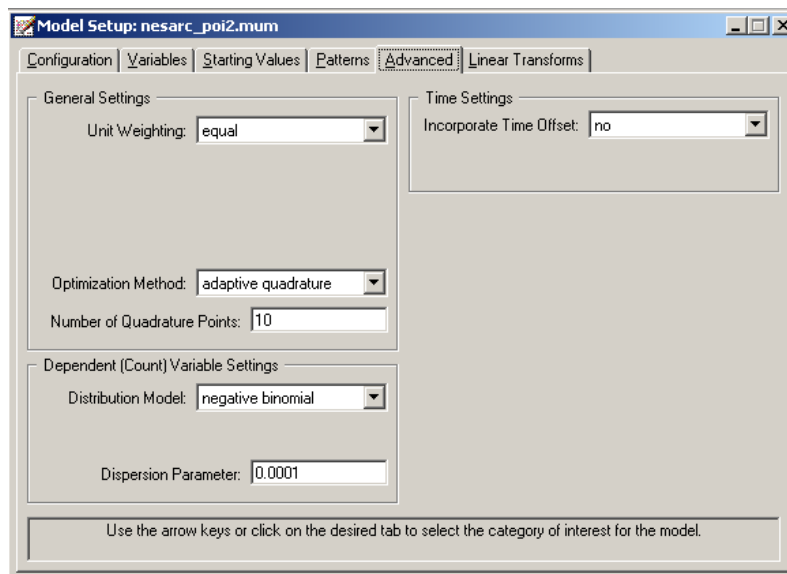
The level-2 model is

$$\beta_0 = b_{00} + v_{i0}, \beta_1 = b_{10} \text{ and } \beta_2 = b_{20}.$$

3.1.4.2 Setting up the analysis

Using the SuperMix spreadsheet **nesarc_poi.ss3** and model specification file **nesarc_poi1.mum** from the previous section, we now set up a negative binomial model for these data.

Start by saving the model specification file under the new name **nesarc_poi2.mum** using the **File, Save As** option. Next, click on the **Advanced** tab of the **Model Setup** window. This is the only tab on which modifications have to be made to change the previously specified Poisson model to a negative binomial model. Set the **Distribution Model** to **negative binomial**, and the **Dispersion Parameter** to 0.0001 to obtain an **Advanced** tab as shown below.



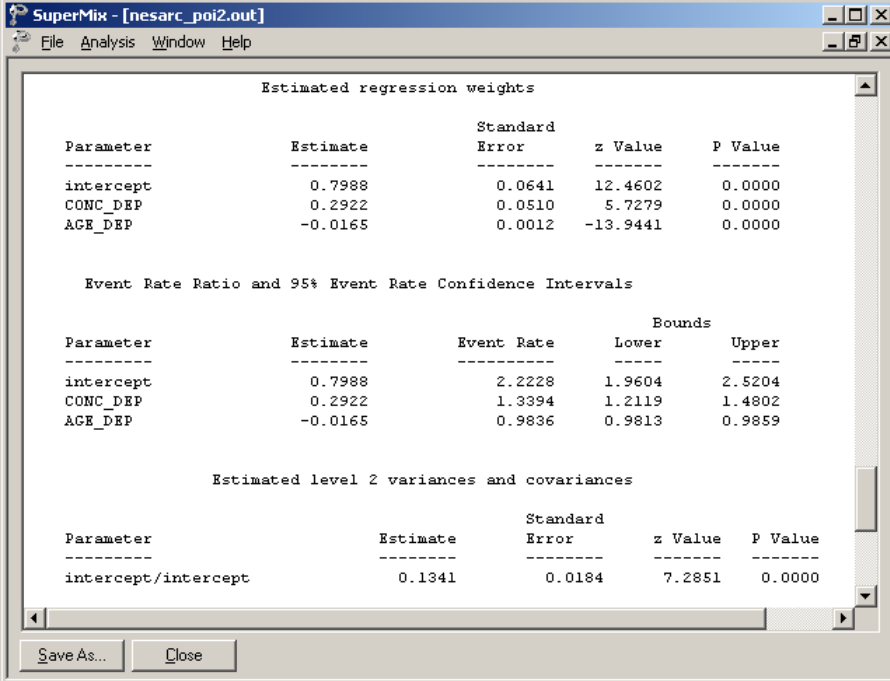
Save the revised model specification file, and click the **Analysis, Run** option to start the iterative process.

3.1.4.3 Discussion of results

Portions of the output file `nesarc_poi.out` are shown below.

Fixed and random effect results

The estimated regression coefficients for fixed effects in the model are shown below. Recall that the estimated coefficients of the intercept, `CONC_DEP`, and `AGE_DEP` under the Poisson model in Section 5.2.2 were 0.7982, 0.2922, and -0.0165 respectively. The estimated variation in the average estimated `N_DEP` at level-2 was 0.1347, and highly significant. The similarity of the results obtained under these two models indicate that the specification of a Poisson distribution model is reasonable for this data.



Estimated regression weights

Parameter	Estimate	Standard Error	z Value	P Value
intercept	0.7988	0.0641	12.4602	0.0000
CONC_DEP	0.2922	0.0510	5.7279	0.0000
AGE_DEP	-0.0165	0.0012	-13.9441	0.0000

Event Rate Ratio and 95% Event Rate Confidence Intervals

Parameter	Estimate	Event Rate	Bounds	
			Lower	Upper
intercept	0.7988	2.2228	1.9604	2.5204
CONC_DEP	0.2922	1.3394	1.2119	1.4802
AGE_DEP	-0.0165	0.9836	0.9813	0.9859

Estimated level 2 variances and covariances

Parameter	Estimate	Standard Error	z Value	P Value
intercept/intercept	0.1341	0.0184	7.2851	0.0000

Save As... Close