7

The Multilevel Generalized Linear Model for Categorical and Count Data

When outcome variables are severely non-normal, the usual remedy is to try to normalize the data using a non-linear transformation, to use robust estimation methods, or a combination of these (see Chapter Four for details). Then again, just like dichotomous outcomes, some types of data will always violate the normality assumption. Examples are ordered (ordinal) and unordered (nominal) categorical data, which have a uniform distribution, or counts of rare events. These outcomes can sometimes also be transformed, but they are preferably analyzed in a more principled manner, using the generalized linear model introduced in Chapter Six. This chapter describes the use of the generalized linear model for ordered categorical data and for count data.

7.1 ORDERED CATEGORICAL DATA

There is a long tradition, especially in the social sciences, of treating ordered categorical data as if they were continuous and measured on an interval scale. A prime example is the analysis of Likert scale survey data, where responses are collected on ordered response categories, for example ranging from 1=totally disagree to 5=totally agree. Another example is a physician's prognosis for a patient categorized as 'good', 'fair' and 'bad'.

The consequences of treating ordered categorical data as continuous are well known, both through analytical work (Olsson, 1979) and simulations (e.g., Dolan, 1994; Muthén & Kaplan, 1985). The general conclusion is that if there are at least five categories, and the observations have a symmetric distribution, the bias introduced by treating categorical data as continuous is small (Bollen & Barb, 1981). With seven or more categories, the bias is very small. If there are four or fewer categories, or the distribution is skewed, both the parameters and their standard error tend to have a downward bias. When this is the case, a statistical method designed for ordered data is needed. Such models are discussed by, a.o., McCullagh and Nelder (1989) and Long (1997). Multilevel extensions of these models are discussed by Goldstein (1995, 2003), Raudenbush & Bryk (2002), and Hedeker and Gibbons (1994). This chapter treats the cumulative regression model, which is frequently used in practice; see Hedeker (2008) for a discussion of other multilevel models for ordered data.

7.1.1 Cumulative Regression Models for Ordered Data

A useful model for ordered categorical data is the cumulative ordered logit or probit model. It is common to start by assigning simple consecutive values to the ordered categories, such as 1…*C* or 0…*C*-1. For example, for a response variable *Y* with three categories such as 'never', 'sometimes' and 'always' we have three response probabilities:

$$\text{Prob}(Y=1) = p_1$$
$$\text{Prob}(Y=2) = p_2$$
$$\text{Prob}(Y=3) = p_3$$

The cumulative probabilities are given by

$$p_1^* = p_1$$
$$p_2^* = p_1 + p_2$$
$$p_2^* = p_1 + p_2 + p_3 = 1$$

where p_3 is redundant. With C categories, only C-1 cumulative probabilities are needed. Since p_1 and p_2 are probabilities, generalized linear regression can be used to model the cumulative probabilities. As stated in chapter Six, a generalized linear regression model consists of three components:

1. an outcome variable y with a specific error distribution that has mean μ and variance σ^2,
2. a linear additive regression equation that produces a predictor η of the outcome variable y,
3. a *link function* that links the expected values of the outcome variable y to the predicted values for η: $\eta=f(\mu)$.

For a logistic regression we have the logit link function

$$\eta_c = \text{logit}\left(p_c^*\right) = \log\left(\frac{p_c^*}{1-p_c^*}\right), \tag{7.1}$$

and for probit regression the inverse normal link

$$\eta_c = \Phi\left(p_c^*\right)^{-1}, \tag{7.2}$$

for c=1…C-1. Assume we specify an intercept-only model for the cumulative probabilities, written as

$$\eta_{ic} = \theta_c. \tag{7.3}$$

Equation 7.3 specifies a different intercept θ_c for each of the estimated probabilities. These intercepts are called thresholds, because they specify the link between the latent variable η and the observed categorical outcome. The position on the latent variable determines which categorical response is observed. Specifically,

$$y_i = \begin{cases} 1, & \text{if } \eta_i \le \theta_1 \\ 2, & \text{if } \theta_1 < \eta_i \le \theta_2 \\ 3, & \text{if } \theta_2 < \eta_i, \end{cases}$$

where y_i is the observed categorical variable, η_i is the latent continuous variable, and θ_1 and θ_2 are the thresholds. Note that a dichotomous variable only has one threshold, which becomes the intercept in a regression equation. Figure 7.1 illustrates the relations between the thresholdsθ, the unobserved response variable η, and the observed responses.

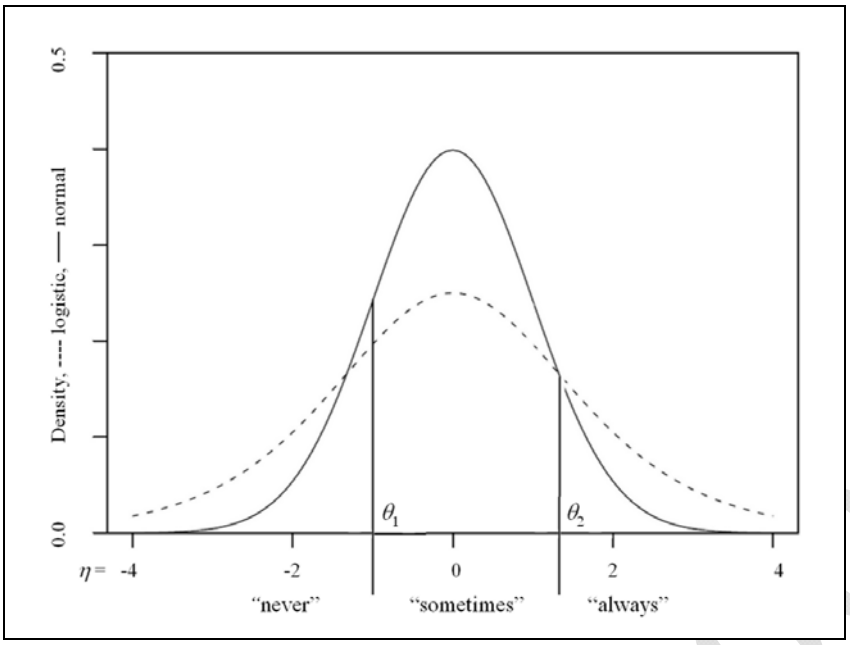

Figure 7.1 Thresholds and observed responses for logit and probit model.

The model in 7.1 is often called a proportional odds model because it is assumed that the effect of predictor variables in the regression is that the entire structure is shifted. This implies that the predictors have the same effect on the odds for each category *c*. The assumption of proportional odds is equivalent to the assumption of parallel regression lines; when the structure is shifted the slope of the regression lines do not change. This assumption is also made in the probit model. An informal test of the assumption of parallel regression lines is made by transforming the ordered categorical variable into a set of dummy variables, following the cumulative probability structure. Thus, for an outcome variable with *C* categories, *C*-1 dummies are created. The first dummy variable equals 1 if the response is in category 1, and 0 otherwise. The second dummy variable equals 1 if the response is in category 2 or 1, and 0 otherwise. And so on until the last dummy variable which equals 1 if the response is in category *C*-1 or lower, and 0 of the response is in category *C*. Finally, independent regressions are carried out on all dummies, and the null-hypothesis of equal regression coefficients is informally assessed by inspecting the estimated regression coefficients and their standard errors. Long (1997) gives and example of this procedure and describes a number of formal statistical tests.

7.1.2 Cumulative Multilevel Regression Models for Ordered Data

Just as in multilevel generalized linear models for dichotomous data, the linear regression model is constructed on the underlying logit or probit scale. Both have a mean of zero, the variance of the logistic distribution is $\pi^2/3$ (standard deviation 1.81), and the standard normal distribution for the probit has a variance of 1. As a consequence, there is no lowest level error term e_{ij}, similar to its absence in generalized linear models for dichotomous data. In fact, dichotomous data can be viewed as ordered data with only two categories. The results for the logit and the probit formulation are generally very similar, but due to the larger variance on the logit scale both the regression coefficients and their standard errors tend to be approximately 1.6 times larger on that scale (cf. Gelman & Hill, 2007).

Assuming individuals *I* nested in groups *j*, and distinguishing between the different cumulative proportions, we write the model for the lowest level as follows:

$$\begin{aligned} \eta_{1ij} &= \theta_{1j} + \beta_{1j} X_{ij} \\ \eta_{2ij} &= \theta_{2j} + \beta_{1j} X_{ij} \end{aligned}, \qquad (7.4)$$

where the thresholds θ_1 and θ_2 are the intercepts for the two response outcomes. The model given by 7.4 is problematic, because we have two intercepts or thresholds that can vary across groups. The interpretation of such variation is that the groups differ in how the values of the underlying η variable are translated into response categories. If this is the case, there is no measurement equivalence between different groups, and it is impossible to make meaningful comparisons. Therefore the model is rewritten by subtracting the value from the first threshold from all thresholds. Thus, the first threshold becomes zero, and is effectively removed from the model. It is replaced by an overall intercept, which is allowed to vary across groups. Thus, the lowest level model is

$$\begin{aligned} \eta_{1ij} &= \beta_{0j} + \beta_{1j} X_{ij} \\ \eta_{2ij} &= \theta_2 + \beta_{0j} + \beta_{1j} X_{ij} \end{aligned}, \tag{7.5}$$

where in 7.5 the threshold θ_2 is equal to θ_2 - θ_1 in 7.4. Obviously, the value for β_0 in 7.5 will be equal to -θ_1 in 7.4. Note that the transformed threshold θ_2 does not have a subscript for groups; it is assumed to be fixed to maintain measurement invariance across the groups. To keep the notation simple, we will continue to use $\theta_2 \ldots \theta_C$ to refer to the thresholds in the 7.5 parameterization, where the first threshold is constrained to zero to allow an intercept in the model, and the other thresholds are all shifted.

From this point, the multilevel generalized model for ordinal observations is constructed following the accustomed procedures. Thus, the intercept β_{0j} and the slope β_{1j} are modeled using a set of second level regression equations

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01} Z_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} Z_j + u_{1j} \end{aligned}. \tag{7.6}$$

The single equation version of the model is

$$\begin{aligned} \eta_{1ij} &= \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} X_{ij} Z_j + u_{0j} + u_{1j} X_{ij} \\ \eta_{2ij} &= \theta_2 + \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} X_{ij} Z_j + u_{0j} + u_{1j} X_{ij} \end{aligned}, \tag{7.7}$$

or, in a simpler notation

$$\eta_{cij} = \theta_c + \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} X_{ij} Z_j + u_{0j} + u_{1j} X_{ij}, \tag{7.8}$$

with the condition that θ_1 is zero. Using the empty model

$$\eta_{cij} = \theta_c + \gamma_{00} + u_{0j}, \tag{7.9}$$

we obtain estimates the variance of the residual errors u_0 that can be used to calculate the intraclass correlation. The total variance is equal to $\pi^2/3$ for the logit and 1 for the probit scale. Note that the ICC is defined on the underlying scale, and not on the observed categorical response scale. Just as in the dichotomous case, the underlying scale is rescaled in each model, and the regression coefficients from different models can not be compared directly. Fielding (2004) discusses techniques that allow comparisons between different models in multilevel generalized linear models.

Modeling the cumulative probabilities p_1, $p_1 + p_2$, $p_1 + p_2 + \ldots + p_{C-1}$ makes the last response category the reference category. As a result, the regression coefficients in the cumulative regression model will have a sign that is the opposite of the sign given by an ordinary linear regression. This is confusing, and most model and software writers solve this effectively by writing the regression

equation e.g. for 7.8 as

$$\eta_{cij} = -1\left(\theta_c + \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}X_{ij}Z_j + u_{0j} + u_{1j}X_{ij}\right), \qquad (7.10)$$

which restores the signs to the direction they would have in a standard linear regression. However, this is not universally done, and software users should understand what their particular software does.

The estimation issues discussed in modeling dichotomous outcomes and proportions also apply to estimating ordered categorical models. One approach is to use Taylor series linearization, using either the Marginal Quasi Likelihood (MQL) or the Penalized Quasi Likelihood (PQL). PQL is generally considered more accurate, but in either case the approximation to the likelihood is not accurate enough to permit deviance difference tests. The other approach is to use numerical methods, which is more computer-intensive and more vulnerable to failure. When numerical methods are used, their performance can be improved by paying careful attention to the explanatory variables. Outliers and using variables that differ widely in scale increase the risk of failure. In addition, centering explanatory variables with random slopes to make sure that zero is an acceptable value is important. The next section presents an example where these issues are present.

7.1.3 Example of ordered categorical data

Assume that we undertake a survey to determine how characteristics of streets affect feelings of unsafety in people walking these streets. A sample of 100 streets is selected, and on each street a random sample of 10 persons is asked how often they feel unsafe while walking that street. The safety is asked using three answer categories: 1 = never, 2 = sometimes, 3 = often. Predictor variables are age and gender, street characteristics are an economic index (standardized *Z*-score) and a rating of the crowdedness of the street (7point scale).

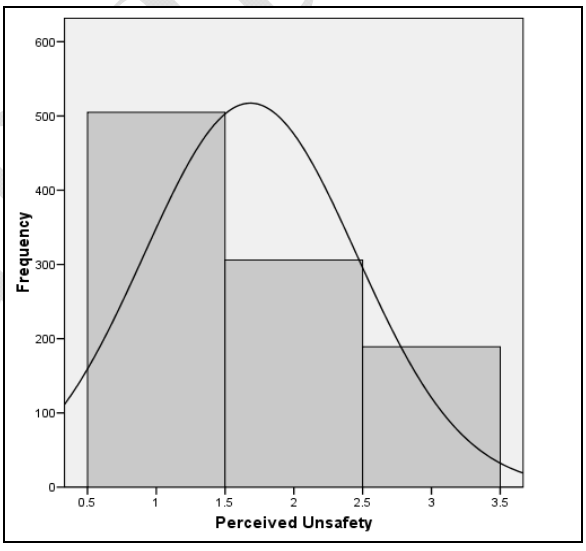

Figure 7.2 Histogram of outcome variable Feeling Unsafe

Figure 7.2 shows the distribution of the outcome variable. In addition to having only three categories, it is clear that the distribution is not symmetric. Treating this outcome as continuous is not proper. The data have a multilevel structure with people nested in streets.

These data have some characteristics that make estimation unnecessarily difficult. The respondents' age is recorded in years, and ranges from 18–72. This range is much different from the range of the other variables. In addition, zero is not a possible value in age and crowdedness. To deal with these issues, the variable age is divided by 10, and all explanatory variables (including

sex) are centered. Using the exploration strategy put forward earlier (Chapter 4), it turns out that all explanatory variables are significant, and that Age has significant slope variation across streets. However, this variation cannot be explained by economic status or crowdedness. Table 7.1 presents the results of the final model, once for the logit and once for the probit model. Estimates were made with Full ML, using numerical approximation in SuperMix.

Table 7.1 Results logit and probit model for unsafety data

Model:	**Logit**		**Probit**	
Fixed part				
predictor	coeff.	s.e.	coeff.	s.e.
intercept	-.02	.09	-.02	.06
threshold 2	2.02	.11	1.18	.06
age/10	0.46	.07	0.27	.04
sex	1.23	.15	0.72	.09
economic	-.72	.09	-.42	.05
crowded	-.47	.05	-.27	.03
Random part				
intercept	0.26	.12	.10	.04
age/10	0.20	.07	.07	.02
int/age	-0.01	.07	-.00	.02
Deviance	1718.58		1718.08	

In terms of interpretation the two models are equivalent. On average, the coefficients and the standard errors in the logit model are on average 1.71 times larger than in the probit model. The variances and their standard errors are 2.73 times larger, which is approximately 1.65 squared. The probit model is simple to interpret, since the underlying scale has a standard deviation of 1. So, an increase in age by 10 years increases the feelings of unsafety by approximate $1/4^{th}$ of a standard deviation, which is a relatively small effect. On the other hand, the difference between men and women on the underlying scale is about 0.7^{th} of a standard deviation, which is a large effect. On the logit scale, the interpretation is often in terms of the odds ratio. Thus, the odds ratio corresponding to the regression coefficient of 0.46 for age/10 is $e^{0.46}$=1.59. Thus, a difference of 10 years results in an odds ratio for being in response category *c* compared to *c*-1 that is 1.59 times larger. Note that the change in odds ratio is independent of the response category, which follows from the proportional odds assumption.

To gain some insight in the effect of different estimation strategies, Table 7.2 presents the same results for the logit model only, where the estimation methods are varied. The first column contains the estimates produced using Taylor series expansion (using HLM). The second column contains the estimates using numerical integration with SuperMix, the third column contains the estimates using numerical integration with Mplus, which uses a different estimation algorithm.

Table 7.2 Results unsafety data with different estimation methods

Estimation:	**Taylor series** (HLM)		**Numerical** (SuperMix)		**Numerical** (Mplus)	
Fixed part						
predictor	coeff.	s.e.	coeff.	s.e.	coeff.	s.e.
intercept/thresh	-.01	.09	-.02	.09	0.02	.09

threshold 2	1.96 .10	2.02 .11	2.04 .12
age/10	-.42 .06	0.46 .07	0.46 .07
sex	-1.15 .14	1.23 .15	1.22 .14
economic	0.68 .09	-.72 .09	-.72 .09
crowded	0.44 .05	-.47 .05	-.47 .05
Random part			
intercept	0.21	0.26 .12	0.26 .07
age/10	0.16	0.20 .07	.20 .07
int/age	-.01	-0.01 .07	-.01 .07
Deviance		1718.58	1718.59
AIC		1736.58	1736.59
BIC		1780.75	1780.76

All estimates in Table 7.2 are close. The Taylor series linearization in HLM produces estimates that are a bit smaller than the numerical integration methods do. For dichotomous data it has been shown that the Taylor series approach tends to have a small negative bias (Breslow & Lin, 1995; Raudenbush, Yang & Yosef, 2000; Rodriguez & Goldman, 1995). The estimates in Table 7.2 suggest that the same bias occurs in modeling ordered data. Nevertheless, the estimates produced by Taylor series approximation for the unsafety data are very close to the other estimates, and the differences would not lead to a different interpretation. The estimates produced by the numerical integration in SuperMix and Mplus are essentially identical. HLM does not give standard errors for the random part, but the chi-square test on the residuals (see Chapter 3 for details) shows both the intercept and the slope variance is significant.

Table 7.2 also illustrates the effect of different choices for the model parameterization. HLM used the proportional odds model as presented in equation 7.8. This models the probability of being in category *c* or lower against being in the last category *c=C*. Thus, the regression coefficients have a sign that is opposite to the sign in an ordinary regression model). SuperMix and Mplus use the model as presented in equation 7.10, where the linear predictor in the generalized regression model is multiplied by -1 to restore the signs of the regression coefficients. A small difference between SuperMix and Mplus is that SuperMix transforms the thresholds as described above, and Mplus does not. So the first row in the fixed part shows the intercept for SuperMix, and threshold 1 for Mplus. If we subtract 0.02 from both thresholds in the Mplus column, the first becomes 0 and the second becomes identical to threshold 2 in the SuperMix column. All these model parameterizations are equivalent, but Table 7.2 shows that it is important to know exactly what the software at hand actually does.

7.2 COUNT DATA

Frequently the outcome variable of interest is a count of events. In most cases count data do not have a nice normal distribution. A count can not be lower than zero, so count data always have a lower bound at zero. In addition, there may be a number of extreme values, which result in a long tail at the right and hence skewness. When the outcome is a count of events that occur frequently, these problems can usually be solved by taking the square root or in more extreme cases the logarithm. On the other hand, when the counts are of relatively rare events, it is commonly assumed that they follow a Poisson distribution, and they are modeled using a generalized linear model. Examples of such events are frequency of depressive symptoms in a normal population, traffic accidents on specific road stretches, or conflicts in stable relationships.

7.2.1 The Poisson model for count data

In the Poisson distribution, the probability of observing *y* events (*y*=0,1,2,3,…) is

$$\Pr(y) = \frac{\exp(-\lambda)\lambda^{y}}{y!}, \tag{7.11}$$

where **exp** is the inverse of the natural logarithm. Just like the binomial distribution, the Poisson distribution has only one parameter, the event rate λ (lambda). The mean and variance of the Poisson distribution are both equal to λ. As a result, with an increasing event rate, the frequency of the higher counts increases, and the variance of the counts also increases, which introduces heteroscedasticity. An important assumption in the Poisson distribution is that the events are independent. For example, counting how many days a pupil has missed school is probably not a Poisson variate, because one may miss school because of an illness, and if this lasts several days these counts are not independent. The number of typing errors on randomly chosen book pages is probably a Poisson variate.

A generalized linear regression model consists of three components:

1. an outcome variable *y* with a specific error distribution that has mean μ and variance σ^2,
2. a linear additive regression equation that produces a predictor η of the outcome variable *y*,
3. a *link function* that links the expected values of the outcome variable *y* to the predicted values for η: $\eta=f(\mu)$.

For counts, the outcome variable is commonly assumed to follow a Poisson distribution with event rate λ. The Poisson model assumes that the length of the observation period is fixed in advance (constant exposure), the events occur at a constant rate, and that the number of events in disjoint intervals are statistically independent. The multilevel Poisson model deals with certain kinds of dependence. The model can be further extended by including an varying exposure rate *m*. For instance, if book pages have different numbers of words, the distribution of typing errors would be Poisson with exposure rate the number of words on a page. In some software the exposure variable just needs to be specified. If this is not possible, the exposure variable is added to the model, including a log transformation LN(*m*) to put it on the same scale as the latent outcome variable η. Such a term is often called the *offset* in the linear model.

The multilevel Poisson regression model for a count Y_{ij} for person *i* in group *j* can be written as:

$$Y_{ij} \,|\, \lambda_{ij} = \text{Poisson}\left(m_{ij}, \lambda_{ij}\right). \tag{7.12}$$

The standard link function for the Poisson distribution is the logarithm, and

$$\eta_{ij} = \log\left(\lambda_{ij}\right). \tag{7.13}$$

The level-1 and level-2 model is constructed as usual, so

$$\eta_{ij} = \beta_{0j} + \beta_{1j} X_{ij}, \tag{7.14}$$

and

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01} Z_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} Z_j + u_{1j} \end{aligned}, \tag{7.15}$$

giving

$$\eta_{cij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} X_{ij} Z_j + u_{0j} + u_{1j} X_{ij}. \tag{7.16}$$

Since the Poisson distribution has only one parameter, specifying an expected count implies a specific variance. Hence, the first-level equations do not have a lowest level error term. In actual practice, we often find that the variance exceeds its expected value. In this case we have overdispersion. In more rare cases we may have underdispersion. Underdispersion often indicates a misspecification of the model, such as the omission of large interaction effects. Overdispersion can occur if there are extreme outliers, or if we omit an entire level in a multilevel model. In binomial models, very small group sizes (around three or less) also lead to overdispersion (Wright, 1997), this is likely to be also the case in Poisson models. A different problem is the problem of having many more zero counts than expected. This problem is dealt with later in this chapter.

Example of count data

Skrondahl and Rabe-Hesketh (2004) discuss an example where 59 patients who suffer from epileptic seizures are followed on four consecutive visits to the clinic. There is a baseline count of the number of epileptic seizures in the two weeks before the treatment starts. After the baseline count, the patients are randomly assigned to a treatment (drug) and a control (placebo) condition. One additional variable is the patients' age. The baseline count and age are log transformed, and centered around their grand mean.

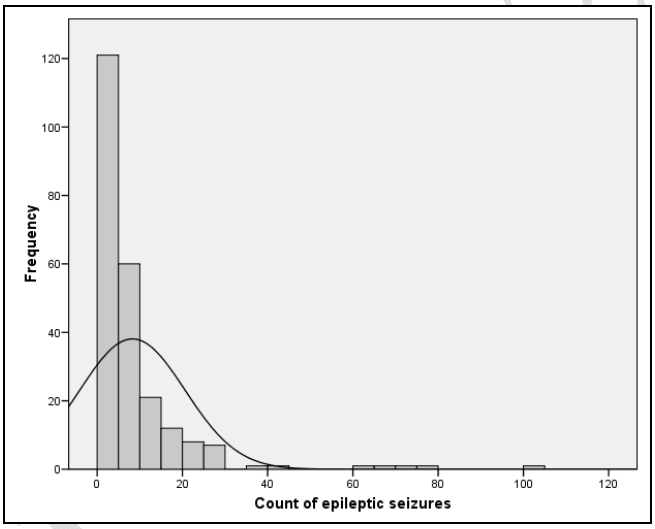

Figure 7.3 Frequency distribution of epileptic seizures

Figure 7.3 shows the frequency distribution of the seizures, which is evidently not normally distributed. The mean number of seizures is 8.3 and the variance is 152.7, which casts some doubt on the applicability of the Poisson model. However, the histogram also shows some extreme outliers. Skrondal and Rabe-Hesketh (2004) discuss these data in more detail, pointing out how inspection of residuals and related procedures provide information about the model fit.

Table 7.3 presents the results of a multilevel Poisson regression analysis of the epilepsy data. We omit the age variable, which is not significant. Given the heterogeneity in the data, robust standard errors are used where available. HLM does not give standard errors for the random part, but the chi-square test on the residuals (see Chapter 3 for details) shows that the intercept variance is significant. The three different analyses result in very similar estimates. A fourth analysis with SuperMix (not presented here), which uses a different type of numerical approximation than HLM results in virtually the same estimates as the HLM numerical approach presented in the last column. Note that HLM does not allow overdispersion with the numerical approximation. SuperMix also does not allow an overdispersion parameter in the model, but it can also estimate models for count data using a negative binomial model. This is an extension of the Poisson model that allows extra variance in the counts.

Table 7.3 Results epilepsy data with different estimation methods					
Estimation:	**Taylor series** (HLM)		**Taylor series** (HLM)		**Numerical** (HLM)
Fixed part					
predictor	coeff.	s.e.[r]	coeff.	s.e.[r]	coeff. s.e.
intercept	1.82	.09	1.83	.09	1.80 .13
log baseline	1.00	.10	1.00	.11	1.01 .10
treatment	-.33	.15	-.30	.15	-.34 .16
Random part					
intercept	0.28		0.26		0.27
overdispersion			1.42		

With all estimation methods the baseline measurement has a strong effect, and the treatment effect is significant at the 0.05 level. To interpret the results in Table 7.3, we need to translate the estimates on the log scale to the observed events. The log baseline is centered, and the control group is coded 0, so the intercept refers to the expected event rate for the control group. Using the estimates in the last column, we take $Y=e^{1.8}=6.05$ as the event rate in the control group. In the experimental group we take $Y=e^{(1.8-0.34}=4.31$ as the event rate. On average, the drug lowers the event rate by 28.8% of the event rate in the untreated control group.

7.2.3 The negative binomial model for count data

In the Poisson model, the variance of the outcome is to the mean. When the observed variance is larger than expected under the Poisson model, we have overdispersion. One way to model overdispersion is to add an explicit error term to the model. Thus, for the Poisson model we have the link function (see 7.13 $\eta_{ij}=\mathbf{log}(\lambda_{ij})$, and the inverse is $\lambda_{ij}=\mathbf{exp}(\eta_{ij})$, where η_{ij} is the outcome predicted by the linear regression model. The negative binomial adds an explicit error term ε to the model, as follows:

$$\lambda_{ij} = \exp\left(\eta_{ij} + \varepsilon_{ij}\right) = \exp\left(\eta_{ij}\right)\exp\left(\varepsilon_{ij}\right) \qquad (7.17)$$

The error term in the model increases the variance produces by the Poisson model. This is close to the dispersion parameter estimated in a Poisson model, a detailed description is given by Long (1997). When the data are analyzed with the negative binomial model, the estimates are very close to the results in the last column of Table 7.3. The dispersion parameter is 0.14 (s.e.=0.04, p=0.001). The subject-level variance is a bit lower at 0.24, which is reasonable given that in the negative binomial model there is more variance at the event level than in the Poisson model presented in Table 7.3.

7.2.4 Too many zeros: the zero inflated model

When the data show an excess of zeros compared to the expected number under the Poisson distribution, it is often assumed that there are two processes that produce the data. Some of the zeros are produced by the Poisson process. Other zeros are produced by a binary process. The assumption is that our data actually include two populations, one that always produces zeros and a second that produces counts following a Poisson model. For example, assume that we study risky behavior, such as using drugs or having unsafe sex. One population never shows this behavior, it is simply not part of the behavior repertoire. These individuals will always report a zero. The other population consists

of individuals who do have this behavior in their repertoire. These individuals can report on their behavior, and these reports can also contain zeros. An individual may sometimes use drugs, but just did not do this in the time period surveyed. Models for such mixtures are referred to as Zero Inflated Poisson or ZIP models. For the count part of the model we use a standard Poisson regression model, and for the probability of being in the population that can produce only zeros we use a standard logit model. Both models are estimated simultaneously. Table 7.4 presents the results for a multilevel Poisson and a multilevel zero inflated Poisson (ZIP) model for the epilepsy data (using Mplus).

Table 7.4 Results epilepsy data: Poisson and Zero Inflated Poisson

Estimation:	**Poisson**		**ZIP**	
Fixed part				
predictor	coeff.	s.e.[r]	coeff.	s.e.[r]
Intercept	1.80	.09	1.87	.09
log baseline	1.01	.11	0.99	.11
treatment	-.34	.15	-.35	.15
Inflation intercept			-3.08	.49
Random part				
intercept	0.28	.07	0.25	.06
Deviance	1343.20		1320.29	
AIC	1351.20		1330.29	
BIC	1365.05		1347.61	

Both the AIC and the BIC indicate that the ZIP model is better, although the parameter estimates for the Poisson model change very little. There is an extra parameter; the intercept of the inflation part. In the ZIP model reported in Table 7.4, there are no explanatory variables that predict the probability of being in the always zero class. As a result, the intercept indicates the average probability of being in that class. A large value of the intercept indicates a large fraction of 'always zero'. The model for the inflation part is a logistic model, so the intercept value of -3.08 is on the underlying logit scale. Translating it to a proportion using the inverse of the logit transformation (introduced in Chapter Six), we find

$$\hat{p} = \frac{e^{3.08}}{1+e^{3.08}} = 0.044\,, \qquad (7.18)$$

which shows that the fraction of 'always zero' in the epilepsy data is very small. In the epilepsy data set, 9.7% of the subjects reports zero seizures. Using 7.17), we can now estimate that 4.4% have no seizures, meaning that their epilepsy is totally suppressed, and 5.3% of the subjects merely happen to have no seizures in the period surveyed.

The ZIP model reported in Table 7.4 does not include predictors for the inflation part. It is possible to expand the inflation model, which is a logistic model similar to the models discussed in Chapter Six, by including predictors. In this particular data set, the available predictors do not predict the zero inflation, and the AIC and BIC indicate that the ZIP model without predictors is preferable.

The negative binomial model can also be extended to include inflated numbers of zeros, in a manner analogous to the Poisson model outlined above. In the epilepsy example data, this turns out to be superfluous, the latent class of extra zeros is estimated as very small, and the AIC and BIC indicate that the negative binomial model without zero inflation is preferable.

7.3 THE EVER CHANGING LATENT SCALE, AGAIN

Just like in logistic and probit regression, the scale of the latent outcome variable implicitly changes when the model is changed. The lowest level residuals are in each separate model scaled to the variance of the standard distribution. This variance is $\pi^2/3 \approx 3.29$ in the logistic distribution, and 1 in the normal distribution. In the Poisson distribution, the variance is equal to the (predicted) mean. In the negative binomial distribution, an extra error term is added to the Poisson variance. With some changes in calculating the lowest level residual variance, all procedures discussed in section 6.5 also apply to the ordered and count data discussed in this chapter.