Chapter 10

Count Data

In the previous chapter we generalised the linear model framework to the exponential family. GLM is often used for modelling count data, in these cases usually the Binomial, Poisson or Multinomial distributions are used.

Types of data and the distribution:

Distribution	Regressors	Response variables
Binomial	x_i	$\mathbf{Y}_i = (Y_i, N - Y_i) = (Y_{i,1}, Y_{i,2})$
Poission	x_i	$\mathbf{Y}_i = Y_i$
Multinomial	x_i	$\mathbf{Y_i} = (Y_{i,1}, Y_{i,2}, \ldots, Y_{i,m})$ $(\sum_j Y_{i,j} = N)$

Distribution	Probabilities
Binomial	$P(Y_{i,1} = k, Y_{i,2} = N - k) = \binom{N}{k}(1 - \pi(\beta' x_i))^{N-k}\pi(\beta' x_i)^k$
Poission	$P(Y_i = k) = \frac{\lambda(\beta' x_i)^k \exp(-\beta' x_i)}{k!}$
Multinomial	$P(Y_{i,1} = k_1, \ldots, Y_{i,m} = k_m) = \binom{N}{k_1,\ldots,k_m}\pi_1(\beta' x_i)^{k_1} \ldots \pi_m(\beta' x_i)^{k_m}$

In this section we will be mainly dealing with count data where the regressors tend to be ordinal (not continuous regressors). This type of data normally comes in the form of a contingency table. One of the most common type of contingency table is the two by two table, and we will consider this in the Section below.

Towards the end of this chapter we use estimating equations to estimate the parameters in overdispersed models.

10.1 Two by Two Tables

Consider the following 2×2 contingency table

	Male	Female	Total
Blue	25	35	60
Pink	15	25	40
Total	40	60	100

Given the above table, one can ask if there is an association between gender and colour preference. The standard method is test for independence. However, we could also pose question in a different way: are proportion of females who like blue the same as the proportion of males who like blue. In this case we can (equivalently) test for equality of proportions (this equivalance usually only holds for 2 by 2 tables).

There are various methods for testing the above hypothesis

- The log-likelihood ratio test.
- The Score test
- The Wald test.
- Through Pearson residuals (which is the main motivation of the chi-squared test for independence).

There can be so many tests for doing the same thing. But recall from Section 2.8.2 that asymptotically all of these tests are equivalent; for a large enough sample size their p-values are nearly the same.

We go through some examples in the following section.

10.1.1 Tests for independence

Approach 1: Pearson and log-likelihood ratio test

The chi-square test for independence is based upon the Pearson residuals:

$$T_1 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}},$$

where $O_{i,j}$ are the observed numbers and E_{ij} are the expected numbers under independence. We recall that by modelling the counts are a multinomial distribution we can show that the test statistic T_1 is asymptotically equivalent to the a log-likelihood ratio test.

Approach 2: Score test

Let us consider the altenative approach, testing for equality of proportions. Let π_M denote the proportion of males who prefer pink over blue and π_F the proportion of females who prefer pink over blue. Suppose we want to test that $H_0 : \pi_F = \pi_M$ against $H_0 : \pi_F \neq \pi_M$. On method for testing the above hypothesis is to use the test for equality of proportions using the Wald test, which gives the test statistic

$$T_2 = \frac{\hat{\pi}_F - \hat{\pi}_M}{I(\pi)^{-1/2}} = \frac{\hat{\pi}_F - \hat{\pi}_M}{\sqrt{\widehat{\pi}\left(\frac{1}{N_F} + \frac{1}{N_M}\right)}},$$

where

$$\widehat{\pi} = \frac{N_{\mathrm{M,P}} + N_{F,P}}{N_M + N_F}$$

and N_M N_F correspond to the number of males and females and $N_{M,P}$ and $N_{F,P}$ the number of males and females who prefer pink.

Approach 3: modelling

An alternative route for conducting the test, is to parameterise π_M and π_F and do a test based on the parametrisation. For example, without loss of generality we can rewrite π_M and π_F as

$$\pi_F = \frac{\exp(\gamma)}{1 + \exp(\gamma)} \qquad \pi_M = \frac{\exp(\gamma + \delta)}{1 + \exp(\gamma + \delta)}.$$

Hence using this parameterisation, the above test is equivalent to testing $H_0 : \delta = 0$ against $H_A : \delta \neq 0$. We can then use the log likelihood ratio test to do the test.

10.2 General contingency tables

Consider the following experiment. Suppose we want to know whether ethnicity plays a role in the number of children a females has. We interview a sample of women, where we

	1	2	3
Background A	20	23	28
Background B	14	27	23

determine her ethnicity and the number of children. The data is collected below in the form of a 3×2 contingency table.

How can such data arise? There are several ways this data could have been collected, and this influences the model we choose to fit to this data. Consider the general $R \times C$ table, with cells indexed by (i, j). Note that in the above example $R = 2$ and $C = 3$.

(a) The subjects arise at random, the study continues until a fixed time elapses. Each subject is categorised according to two variables. Suppose the number in cell (i, j) is Y_{ij}, then it is reasonable to assume $Y_{ij} \sim \text{Poisson}(\lambda_{ij})$ for some $\{\lambda_{ij}\}$, which will be the focus of study. In this case the distribution is

$$P(Y = y) = \prod_{i=1}^{C} \prod_{j=1}^{R} \frac{\lambda_{ij}^{y_{ij}} \exp(-\lambda_{ij})}{y_{ij}!}$$

(b) The total number of subjects is fixed at N, say. The numbers in cells follow a multinomial distribution: $(Y_{ij}) \sim M(N; (\pi_{ij}))$:

$$P(Y = y) = \frac{N!}{\prod_{i=1}^{C} \prod_{j=1}^{R} y_{ij}!} \prod_{i=1}^{C} \prod_{j=1}^{R} \pi_{ij}^{y_{ij}}$$

if $\sum_i \sum_j y_{ij} = N$.

(c) One margin is fixed: say $\{y_{+j} = \sum_{i=1}^{C} y_{ij}\}$ for each $j = 1, 2, \ldots, R$. In each column, we have an independent multinomial sample

$$P(Y = y) = \prod_{j=1}^{R} \left(\frac{y_{+j}!}{\prod_{i=1}^{C} y_{ij}!} \prod_{i=1}^{C} \rho_{ij}^{y_{ij}} \right)$$

where ρ_{ij} is the probability that a column-j individual is in row i (so $\rho_{+j} = \sum_{i=1}^{C} \rho_{ij} = 1$).

Of course, the problem is without knowledge of how the data was collected it is not possible to know which model to use. However, we now show that all the models are

closely related, and with a suitable choice of link functions, diffferent models can lead to the same conclusions. We will only show the equivalence between cases (a) and (b), a similar argument can be extended to case (c).

We start by show that if π_{ij} and λ_{ij} are related in a certain way, then the log-likelihoods of both the poisson and the multinomial are effectively the same. Define the following log-likelihoods for the Poisson, Multinomial and the sum of independent Poissons as follows

$$\begin{aligned}
\mathcal{L}_P(\lambda) &= \sum_{i=1}^{C}\sum_{j=1}^{R}\Big(y_{ij}\log\lambda_{ij} - \lambda_{ij} - \log y_{ij}!\Big) \\
\mathcal{L}_M(\pi) &= \log\frac{N!}{\prod_{i=1}^{C}\prod_{j=1}^{R} y_{ij}!} + \sum_{i=1}^{C}\sum_{j=1}^{R} y_{ij}\log\pi_{ij} \\
\mathcal{L}_F(\lambda_{++}) &= N\log\lambda_{++} - \lambda_{++} - \log N!.
\end{aligned}$$

We observe that $\mathcal{L}_P$ is the log distribution of $\{y_{i,j}\}$ under Poisson sampling, $\mathcal{L}_M$ is the log distribution of $\{y_{i,j}\}$ under multinomial sampling, and $\mathcal{L}_F$ is the distribution of $\sum_{ij} Y_{ij}$, where Y_{ij} are independent Poisson distributions each with mean λ_{ij}, $N = \sum_{ij} Y_{ij}$ and $\lambda_{++} = \sum_{ij}\lambda_{ij}$.

Theorem 10.2.1 *Let $\mathcal{L}_P, \mathcal{L}_M$ and $\mathcal{L}_F$ be defined as above. If λ and π are related through*

$$\pi_{ij} = \frac{\lambda_{ij}}{\sum_{s,t}\lambda_{st}} \quad \lambda_{ij} = \lambda_{++}\pi_{ij},$$

where λ_{++} is independent of (i,j). Then we have that

$$\mathcal{L}_P(\lambda) = \mathcal{L}_M(\pi) + \mathcal{L}_F(\lambda_{++}).$$

PROOF. The proof is straightforward. Consider the log-likelihood of the Poisson

$$\begin{aligned}
\mathcal{L}_P(\lambda) &= \sum_{i=1}^{C}\sum_{j=1}^{R}\Big(y_{ij}\log\lambda_{ij} - \lambda_{ij} - \log y_{ij}!\Big) \\
&= \sum_{i=1}^{C}\sum_{j=1}^{R}\Big(y_{ij}\log\lambda_{++}\pi_{ij} - \lambda_{++}\pi_{ij} - \log y_{ij}!\Big) \\
&= \Big[\sum_{i=1}^{C}\sum_{j=1}^{R} y_{ij}\log\pi_{ij} + \log N! - \sum_{i=1}^{C}\sum_{j=1}^{R}\log y_{ij}!\Big] + \sum_{i=1}^{C}\sum_{j=1}^{R}\Big(y_{ij}\log\lambda_{++} - \lambda_{++} - \log N!\Big) \\
&= \mathcal{L}_M(\pi) + \mathcal{L}_F(\lambda_{++}).
\end{aligned}$$

Which leads to the required result. □

Remark 10.2.1 *The above result means that the likelihood of the independent Poisson conditioned on the total number of participants is N, is equal to the likelihood of the multinomial distribution where the relationship between probabilities and means are given above.*

By connecting the probabilities and mean through the relation

$$\pi_{ij} = \frac{\lambda_{ij}}{\sum_{s,t} \lambda_{st}} \quad \text{and} \quad \lambda_{ij} = \lambda_{++} p_{ij},$$

it does not matter whether the multinomial distribution or Possion distribution is used to do the estimation. We consider a few models which are commonly used in categorical data.

Example 10.2.1 *Let us consider suitable models for the number of children and ethnicity data. Let us start by fitting a multinomial distribution using the logistic link. We start y modelling $\beta' x_i$. One possible model is*

$$\beta' x = \eta + \alpha_1 \delta_1 + \alpha_2 \delta_2 + \alpha_3 \delta_3 + \beta_1 \delta_1^* + \beta_2 \delta_2^*,$$

where $\delta_i = 1$ if the female has i children and zero otherwise, $\delta_1^* = 1$ if female belongs to ethnic group A and zero otherwise, $\delta_2^* = 1$ if female belongs to ethnic group B and zero otherwise. The regressors in this example are $x = (1, \delta_1, \ldots, \delta_2^*)$. Hence for a given cell (i, j) we have

$$\beta' x_{ij} = \eta_{ij} = \eta + \alpha_i + \beta_j.$$

One condition that we usually impose when doing the estimation is that $\sum_{i=1}^{3} \alpha_i = 0$ and $\beta_1 + \beta_2 = 0$. These conditions mean the system is identifiable. Without these conditions you can observe that there exists another $\{\tilde{\alpha}_i\}$, $\{\tilde{\beta}_i\}$ and $\tilde{\eta}$, such that $\eta_{ij} = \eta + \alpha_i + \beta_j = \tilde{\eta} + \tilde{\alpha}_i + \tilde{\beta}_j$.

Now let understand what the above linear model means in terms of probabilities. Using the logistic link we have

$$\pi_{ij} = g^{-1}(\beta' x_{ij}) = \frac{\exp(\eta + \alpha_i + \beta_j)}{\sum_{s,t} \exp(\eta + \alpha_s + \beta_t)} = \frac{\exp(\alpha_i)}{\sum_s \exp(\alpha_s)} \times \frac{\exp(\beta_j)}{\sum_t \exp(\beta_t)},$$

where π_{ij} denotes the probability of having i children and belonging to ethnic group j and x_{ij} is a vector with ones in the appropriate places. What we observe is that the above

model is multiplicative, that is

$$\pi_{ij} = \pi_{i+}\pi_{+j}$$

where $\pi_{i+} = \sum_j \pi_{ij}$ *and* $\pi_{+j} = \sum_i \pi_{i+}$. *This means by fitting the above model we are assuming independence between ethnicity and number of children. To model dependence we would use an interaction term in the model*

$$\beta' x = \eta + \alpha_1\delta_1 + \alpha_t\delta_2 + \alpha_3\delta_3 + \beta_1\delta_1^* + \beta_2\delta_1^* + \sum_{i,j}\gamma_{ij}\delta_i\delta_j^*,$$

hence

$$\eta_{ij} = \eta + \alpha_i + \beta_j + \gamma_{ij}.$$

However, for $R \times C$ *tables an interaction term means the model is saturated (i.e. the MLE estimator of the probability* π_{ij} *is simply* y_{ij}/N*). But for* $R \times C \times L$*, we can model interactions without the model becoming saturated. These interactions may have interesting interpretations about the dependence structure between two variables. By using the analysis of deviance (which is a effectively the log-likelihood ratio test, we can test whether certain interaction terms are significant - similar things were done for linear models).*

We transform the above probabilities into Poisson means using $\lambda_{ij} = \gamma\pi_{ij}$. *In the case there is no-interaction the mean of Poisson at cell* (i, j) *is* $\lambda_{ij} = \gamma \exp(\eta + \alpha_i + \beta_j)$.

In the above we have considered various methods for modelling the probabilities in a mulitnomial and Poisson distributions. In the theorem we show that so long as the probabilities and Poisson means are linked in a specific way, the estimators of β will be identical.

Theorem 10.2.2 (Equivalence of estimators) *Let us suppose that* π_{ij} *and* μ_{ij} *are defined by*

$$\pi_{ij} = \pi_{ij}(\beta) \quad \lambda_{ij} = \gamma\pi_{ij}(\beta),$$

where γ *and* $\beta = \{\alpha_i, \beta_j\}$ *are unknown and* $C(\beta)$ *is a known function of* β *(such as*

$\sum_{i,j} \exp(\alpha_i + \beta_j)$ *or* 1*). Let*

$$\begin{aligned}
\mathcal{L}_P(\beta,\gamma) &= \sum_{i=1}^{C}\sum_{j=1}^{R}\Big(y_{ij}\log\gamma\pi_{ij}(\beta) - \gamma\pi_{ij}(\beta)\Big) \\
\mathcal{L}_M(\beta) &= \sum_{i=1}^{C}\sum_{j=1}^{R} y_{ij}\log\pi_{ij}(\beta) \\
\mathcal{L}_F(\beta,\gamma) &= N\log\gamma - \gamma,
\end{aligned}$$

which is the log-likelihoods for the Multinomial and Poisson distributions without unnecessary constants (such as $y_{ij}!$*). Define*

$$\begin{aligned}
(\hat{\beta}_P,\hat{\gamma}_P) &= \arg\max \mathcal{L}_P(\beta,\gamma) \\
\hat{\beta}_B &= \arg\max \mathcal{L}_M(\beta) \quad \hat{\gamma}_F = \arg\max \mathcal{L}_F(\beta,\gamma).
\end{aligned}$$

Then $\widehat{\beta}_P = \widehat{\beta}_M$ *and* $\widehat{\gamma}_P = \hat{\gamma}_M = N/C(\widehat{\beta}_M)$*.*

PROOF. We first consider $\mathcal{L}_P(\beta,\gamma)$. Since $\sum_{i,j} p_{i,j}(\beta) = 1$ and $\sum_{i,j} y_{i,j} = 1$ we have

$$\begin{aligned}
\mathcal{L}_P(\beta,\gamma) &= \sum_{i=1}^{C}\sum_{j=1}^{R}\Big(y_{ij}\log\gamma C(\beta)\pi_{ij}(\beta) + \gamma C(\beta)\pi_{ij}(\beta)\Big) \\
&= \sum_{i=1}^{C}\sum_{j=1}^{R}\Big(y_{ij}\log\pi_{ij}(\beta)\Big) + N\log\gamma C(\beta) - C(\beta)\gamma.
\end{aligned}$$

Now we consider the partial derivatives of $\mathcal{L}_P$ to obtain

$$\begin{aligned}
\frac{\partial\mathcal{L}_P}{\partial\beta} &= \frac{\partial\mathcal{L}_M}{\partial\beta} + \gamma\frac{\partial C(\beta)}{\partial\beta}\left(\frac{N}{\gamma C(\beta)} - 1\right) = 0 \\
\frac{\partial\mathcal{L}_P}{\partial\gamma} &= \left(\frac{N}{\gamma} - C(\beta)\right) = 0.
\end{aligned}$$

Solving the above we have that $\hat{\beta}_P$ and $\hat{\gamma}_P$ satisfy

$$\hat{\gamma}_P = \frac{N}{\widehat{C}(\beta)} \qquad \frac{\partial\mathcal{L}_M}{\partial\beta}\rfloor_{\beta=\hat{\beta}_P} = 0. \tag{10.1}$$

Now we consider the partial derivatives of $\mathcal{L}_M$ and $\mathcal{L}_C$

$$\frac{\partial\mathcal{L}_M}{\partial\beta} = 0 \qquad \frac{\partial\mathcal{L}_F}{\partial\gamma} = \left(\frac{N}{\gamma} - C(\beta)\right) = 0. \tag{10.2}$$

Comparing the estimators in (10.1) and (10.2) it is clear that the maximum likelihood estimators of β based on the Poisson and the Binomial distributions are the same. □

Example 10.2.2 *Let us consider fitting the Poisson and the multinomial distributions to the data in a contingency table where π_{ij} and λ_{ij} satisfy*

$$\lambda_{ij} = \exp(\eta + \beta' x_{ij}) \text{ and } \pi_{ij} = \frac{\exp(\beta' x_{ij})}{\sum_{s,t} \exp(\beta' x_{s,t})}.$$

Making a comparison with $\lambda_{ij}(\beta) = \gamma C(\beta)\pi_{ij}(\beta)$ we see that $\gamma = \exp(\eta)$ and $C(\beta) = \sum_{s,t} \exp(\beta' x_{s,t})$. Then it by using the above theorem the estimator of β is the parameter which maximises

$$\sum_{i=1}^{C}\sum_{j=1}^{R}\left(y_{ij} \log \frac{\exp(\beta' x_{ij})}{\sum_{s,t} \exp(\beta' x_{s,t})}\right),$$

and the estimator of γ is the parameter which maximises

$$N \log \exp(\eta) C(\hat{\beta}) - \exp(\eta) C(\hat{\beta}),$$

which is $\eta = \log N - \log(\sum_{s,t} \exp(\hat{\beta}' x_{s,t}))$.

10.3 Overdispersion

The binomial and Poisson distributions have the disadvantage that they are determined by only one parameter (π in the case of Binomial and λ in the case of Possion). This can be a disadvantage when it comes to modelling certain types of behaviour in the data. A type of common behaviour in count data is overdispersed, in the sense that the variance appears to be larger than the model variance.

Checking for overdispersion

- First fit a Poisson model to the data.
- Extract the Pearson residuals from the data (see Section 9.3.5), for the Poisson it is

  $$r_i = \frac{(Y_i - \hat{\mu}_i)}{\phi^{1/2} V(\mu_i)^{1/2}} = \frac{Y_i - \hat{\mu}_i}{\sqrt{\mu_i}}.$$

  If the model is correct, the residuals $\{r_i\}$ should be 'close' to a standard normal distribution. However, in the case of overdispersion it is likely that the estimated variance of r_i will be greater than one.
- Plot r_i against μ_i.

10.3.1 Modelling overdispersion

Modelling overdispersion can be done in various ways. Below we focus on Poisson-type models.

Zero inflated models

The number of zeros in count data can sometimes be more (inflated) than Poisson or binomial distributions are capable of modelling (for example, if we model the number of times a child visits the dentist, we may observe that there is large probability the child will not visit the dentist). To model this type of behaviour we can use the inflated zero Poisson model, where

$$P(Y=k) = \begin{cases} (1-p)(1-\exp(-\lambda)) = 1-p+p\exp(-\lambda) & k=0 \\ p\frac{\exp(-\lambda)\lambda^k}{k!} & k>0 \end{cases}.$$

We observe that the above is effectively a mixture model. It is straightforward to show that $\mathrm{E}(Y) = p\lambda$ and $\mathrm{var}(Y) = p\lambda(1+\lambda(1-p))$, hence

$$\frac{\mathrm{var}(Y)}{\mathrm{E}(Y)} = (1+\lambda(1-p)).$$

We observe that there is more dispersion here than classical Poisson where $\mathrm{var}(Y)/\mathrm{E}(Y) = 1$.

Modelling overdispersion through moments

One can introduce overdispersion by simply modelling the moments. That is define a psuedo Poisson model in terms of its moments, where $\mathrm{E}(Y) = \lambda$ and $\mathrm{var}(Y) = \lambda(1+\delta)$ ($\delta \geq 0$). This method does not specify the distribution, it simply places conditions on the moments.

Modelling overdispersion with another distribution (latent variable)

Another method for introducing overdispersion into a model is to include a 'latent' (unobserved) parameter ε. Let us assume that ε is a positive random variable where $\mathrm{E}(\varepsilon) = 1$ and $\mathrm{var}(\varepsilon) = \xi$. We suppose that the distribution of Y conditioned on ε is Poisson, i.e. $P(Y=k|\varepsilon) = \frac{(\lambda\varepsilon)^k \exp(-\lambda\varepsilon)}{k!}$. The introduction of latent variables allows one to generalize

several models in various directions. It is a powerful tool in modelling. For example, if one wanted to introduce dependence between theY_is one can do this by conditioning on a latent variable which is dependent (eg. the latent variable can be a time series).

To obtain the moments of Y we note that for any random variable Y we have

$$\begin{aligned}\mathrm{var}(Y) &= \mathrm{E}(Y^2) - \mathrm{E}(Y)^2 = \mathrm{E}\Big(\mathrm{E}(Y^2|\varepsilon) - \mathrm{E}(Y|\varepsilon)^2\Big) + \mathrm{E}\big(\mathrm{E}(Y|\varepsilon)^2\big) - \mathrm{E}\big(\mathrm{E}(Y|\varepsilon)\big)^2 \\ &= \mathrm{E}\Big(\mathrm{var}(Y|\varepsilon)\Big) + \mathrm{var}(\mathrm{E}(Y|\varepsilon)),\end{aligned}$$

where we note that $\mathrm{var}(Y|\varepsilon) = \sum_{k=0}^{\infty} k^2 P(Y=k|\varepsilon) - (\sum_{k=0}^{\infty} kP(Y=k|\varepsilon))^2$ and $\mathrm{E}(Y|\varepsilon) = \sum_{k=0}^{\infty} kP(Y=k|\varepsilon)$. Applying the above to the conditional Poisson we have

$$\begin{aligned}\mathrm{var}(Y) &= \mathrm{E}(2(\lambda\varepsilon) - (\lambda\varepsilon)) + \mathrm{var}(\lambda\varepsilon) \\ &= \lambda + \lambda^2\xi = \lambda(1+\lambda\xi) \\ \text{and } \mathrm{E}(Y) &= \mathrm{E}(\mathrm{E}(Y|\varepsilon)) = \lambda.\end{aligned}$$

The above gives an expression in terms of moments. If we want to derive the distribution of Y, we require the distribution of ε. This is normally hard in practice to verify, but for reasons of simple interpretation we often let ε have a Gamma distribution $f(\varepsilon;\nu,\kappa) = \frac{\nu^\kappa \varepsilon^{k-1}}{\Gamma(\kappa)}\exp(-\nu\varepsilon)$, where $\nu=\kappa$, hence $\mathrm{E}(\varepsilon)=1$ and $\mathrm{var}(\varepsilon) = 1/\nu(=\xi)$. Therefore in the case that ε is a Gamma distribution with density $f(\varepsilon;\nu,\nu) = \frac{\nu^\nu \varepsilon^{\nu-1}}{\Gamma(\nu)}\exp(-\nu\varepsilon)$ the distribution of Y is

$$\begin{aligned}P(Y=k) &= \int P(Y=k|\varepsilon) f(\varepsilon;\nu,\nu)d\varepsilon \\ &= \int \frac{(\lambda\varepsilon)^k \exp(-\lambda\varepsilon)}{k!}\frac{\nu^\nu\varepsilon^{k-1}}{\Gamma(\nu)}\exp(-\nu\varepsilon)d\varepsilon \\ &= \frac{\Gamma(k+\nu)}{\Gamma(\nu)k!}\frac{\nu^\nu\lambda^k}{(\nu+\lambda)^{\nu+\lambda}}.\end{aligned}$$

This is called a negative Binomial (because in the case that ν is an integer it resembles a regular Binomial but can take infinite different outcomes). The negative binomial only belongs to the exponential family if ν is known (and does not need to be estimated). Not all distributions on ε lead to explicit distributions of Y. The Gamma is popular because it leads to an explicit distribution for Y (often it is called the conjugate distribution).

A similar model can also be defined to model overdispersion in proportion data, using a random variable whose conditional distribution is Binomial (see page 512, Davison (2002)).

Remark 10.3.1 (Using latent variables to model dependence) *Suppose Y_j conditioned on $\{\varepsilon_j\}$ follows a Poisson distribution where $P(Y_j = k|\varepsilon_j) = \frac{(\lambda\varepsilon_j)^k \exp(-\lambda\varepsilon_j)}{k!}$ and $Y_i|\varepsilon_i$ and $Y_j|\varepsilon_j$ are conditionally independent. We assume that $\{\varepsilon_j\}$ are positive continuous random variables with correlation $\mathrm{cov}[\varepsilon_i, \varepsilon_j] = \rho_{i,j}$. The correlations in ε_j induce a correlation between Y_j through the relation*

$$\begin{aligned}\mathrm{cov}\left[Y_i, Y_j\right] &= \mathrm{E}\left(\underbrace{\mathrm{cov}\left[Y_i, Y_j|\varepsilon_i, \varepsilon_j\right]}_{=0(a.s.)}\right) + \mathrm{cov}\left(\underbrace{\mathrm{E}[Y_i|\varepsilon_i]}_{=\lambda\varepsilon_i}, \underbrace{\mathrm{E}[Y_j|\varepsilon_j]}_{=\lambda\varepsilon_j}\right) \\ &= \lambda^2 \mathrm{cov}\left(\varepsilon_i, \varepsilon_j\right) = \lambda^2\rho_{ij}.\end{aligned}$$

10.3.2 Parameter estimation using estimating equations

We now consider various methods for estimating the parameters. Some of the methods described below will be based on the Estimating functions and derivations from Section 9.3.1, equation (9.10).

Let us suppose that $\{Y_i\}$ are overdispersed random variables with regressors $\{x_i\}$ and $\mathrm{E}(Y_i) = \mu_i$ with $g(\mu_i) = \beta' x_i$. The natural way to estimate the parameters β is to use a likelihood method. However, the moment based modelling of the overdispersion does not have a model attached (so it is not possible to use a likelihood method), and the modelling of the overdispersion using, say, a Gamma distribution, is based on a assumption that is hard in practice to verify (that the latent variable is Gaussian). An alternative approach is to use moment based/estimating function methods which are more robust to misspecification than likelihood methods. In the estimation we discuss below we will focus on the Poisson case, though it can easily be generalised to the non-Poisson case.

Let us return to equation (9.10)

$$\sum_{i=1}^{n} \frac{(Y_i - \mu_i)x_{ij}}{\phi V(\mu_i) g'(\mu_i)} = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)x_{ij}}{\phi V(\mu_i)} \frac{d\mu_i}{d\eta_i} = 0 \qquad 1 \leq j \leq p. \tag{10.3}$$

In the case of the Poisson distribution, with the log link the above is

$$\sum_{i=1}^{n} (Y_i - \exp(\beta' x_i))x_{ij} = 0 \qquad 1 \leq j \leq p. \tag{10.4}$$

We recall if $\{Y_i\}$ are Possion random variables with mean $\exp(\beta' x_i)$, then variance of the limiting distribution of β is

$$(\widehat{\beta} - \beta) \approx N_p(0, (X^T W X)^{-1}),$$

since the Fisher information matrix can be written as

$$(I(\beta))_{jk} = E\left(-\frac{\partial^2 \mathcal{L}_n(\beta)}{\partial\beta_j \partial\beta_k}\right) = \mathrm{E}\left(-\sum_{i=1}^{n} \frac{d^2\ell_i}{d\eta_i^2} x_{ij}x_{ik}\right) = (X^T W X)_{jk}.$$

where

$$\begin{aligned} W &= \mathrm{diag}\left(\mathrm{E}(-\frac{\partial^2 \ell_1(\eta_1)}{\partial \eta_1^2}), \ldots, \mathrm{E}(-\frac{\partial^2 \ell_n(\eta_n)}{\partial \eta_n^2})\right) \\ &= \mathrm{diag}\left(\exp(\beta' x_1), \ldots, \exp(\beta' x_n)\right). \end{aligned}$$

However, as we mentioned in Section 9.3.1, equations (10.3) and (10.4) do not have to be treated as derivatives of a likelihood. Equations (10.3) and (10.4) can be viewed as estimating equation, since they only use the first and second order moments of $\{Y_i\}$. Hence they can be used as the basis of the estimation scheme even if they are not as efficient as the likelihood. In the overdispersion literature the estimating equations (functions) are often called the Quasi-likelihood.

Example 10.3.1 *Let us suppose that $\{Y_i\}$ are independent random variables with mean $\exp(\beta' x_i)$. We use the solution of the estimating function*

$$\sum_{i=1}^{n} g(Y_i;\beta) = \sum_{i=1}^{n} (Y_i - \exp(\beta' x_i))x_{ij} \quad = \quad 0 \qquad 1 \leq j \leq p.$$

to estimate β. Using Theorem 8.2.2 we derive the asymptotic variance for two models:

(i) $\mathrm{E}(Y_i) = \exp(\beta' x_i)$ **and** $\mathrm{var}(Y_i) = (1+\delta)\exp(\beta' x_i)$ **$(\delta \geq 0)$.**

Let us suppose that $\mathrm{E}(Y_i) = \exp(\beta' x_i)$ and $\mathrm{var}(Y_i) = (1+\delta)\exp(\beta' x_i)$ $(\delta \geq 0)$. Then if the regularity conditions are satisfied we can use Theorem 8.2.2 to obtain the limiting variance. Since

$$\begin{aligned} \mathrm{E}\left(\frac{-\partial \sum_{i=1}^{n} g(Y_i;\beta)}{\partial\beta}\right) &= X^T \mathrm{diag}\left(\exp(\beta' x_1), \ldots, \exp(\beta' x_n)\right) X \\ \mathrm{var}\left(\sum_{i=1}^{n} g(Y_i;\beta)\right) &= (1+\delta) X^T \mathrm{diag}\big(\exp(\beta' x_1), \ldots, \exp(\beta' x_n)\big) X, \end{aligned}$$

the limiting variance is

$$(1+\delta)(X^T W X)^{-1} = (1+\delta)(X^T \mathrm{diag}\left(\exp(\beta' x_1), \ldots, \exp(\beta' x_n)\right) X)^{-1}.$$

Therefore, in the case that the variance is $(1+\delta)\exp(\beta' x_i)$*, the variance of the estimator using the estimating equations* $\sum_{i=1}^n g(Y_i;\beta)$*, is larger than for the regular Poisson model. If* δ *is quite small, the difference is also small. To estimate* δ *we can use*

$$\sum_{i=1}^n \frac{(Y_i - \exp(\widehat{\beta}' x_i))^2}{\exp(\widehat{\beta}' x_i)}.$$

(ii) $\mathrm{E}(Y_i) = \exp(\beta' x_i)$ **and** $\mathrm{var}(Y_i) = \exp(\beta' x_i)(1+\xi\exp(\beta' x_i))$.

In this case we have

$$\mathrm{E}\left(\frac{-\partial \sum_{i=1}^n g(Y_i;\beta)}{\partial\beta}\right) = X^T W X \quad \textit{and} \quad \mathrm{var}(\sum_{i=1}^n g(Y_i;\beta)) = X^T \tilde{W} X,$$

where

$$\begin{aligned} W &= \mathrm{diag}\Big(\exp(\beta' x_1),\ldots,\exp(\beta' x_n)\Big) \\ \tilde{W} &= \mathrm{diag}\Big(\exp(\beta' x_1)(1+\xi\exp(\beta' x_1)),\ldots,\exp(\beta' x_n)(1+\xi\exp(\beta' x_n))\Big). \end{aligned}$$

Hence the limiting variance is

$$(X^T W X)^{-1}(X^T \tilde{W} X)(X^T W X)^{-1}.$$

We mention that the estimating equation can be adapted to take into count the overdispersion in this case. In other words we can use as an estimator of β*, the* β *which solves*

$$\sum_{i=1}^n \frac{(Y_i - \exp(\beta' x_i))}{(1+\xi\exp(\beta' x_i))} x_{ij} = 0 \qquad 1 \le j \le p.$$

Though we mention that we probably have to also estimate ξ *when estimating* β*.*

10.4 A worked problem

(1) (a) Suppose that U is a Poisson distributed random variable with mean λ. Then for $k \ge 0$,

$$P(U=k) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{10.5}$$

(i) Let us suppose $U_1, \ldots, U_n$ are independent, identically distributed random variables from a Poisson distribution. What is the maximum likelihood estimator of λ?

(ii) For several count data sets it has been observed that there is an excessive number of zeros. To model 'inflated-zero' count data, the zero-inflated Poisson distribution model was proposed, where the observations are modelled as

$$Y = \delta U,$$

where δ and U are independent random variables, δ takes on value either zero or one with $P(\delta = 0) = p$, $P(\delta = 1) = (1 - p)$, and U has a Poisson distribution as defined as in (10.5).

Briefly explain why this model can account for an excessive number of zeros.

(iii) Show that the estimator defined in (i) is a *biased* estimator of λ when the observations come from a zero-inflated Poisson distribution.

(b) In this part of the question we consider the zero-inflated Poisson regression model, proposed in Lambert (1992), which is defined as

$$Y_j = \delta_j U_j,$$

where δ_j and U_j are independent random variables, $P(\delta_j = 0) = p$, $P(\delta_j = 1) = (1 - p)$, and U_j has a Poisson distribution with mean $\lambda_j = e^{\beta x_j}$ and x_j is a *fixed* covariate value. Our objective is to first construct crude estimators for p and β and to use these estimates as the initial values in an iterative scheme to obtain the maximum likelihood estimator.

(i) *Estimation of* β. What is the distribution of Y_j conditioned on $Y_j > 0$ and x_j?

Argue that, for each $k = 1, 2, \ldots$,

$$P(Y_j = k | Y_j > 0) = \frac{e^{-\lambda_j} \lambda_j^k / k!}{(1 - e^{-\lambda_j})}. \tag{10.6}$$

Let $\boldsymbol{Y}^+$ be the vector of all the non-zero Y_js. Use result (10.6) to define a *conditional* log-likelihood for $\boldsymbol{Y}^+$ given that all the Y_js in $\boldsymbol{Y}^+$ are positive.

Determine the derivative of this conditional log-likelihood, and explain how it can be used to determine an estimate of β. Denote this estimator as $\widehat{\beta}$.

(ii) *Estimation of p.* Define the dummy variable

$$Z_j = \begin{cases} 0 & \text{if } Y_j = 0 \\ 1 & \text{if } Y_j > 0. \end{cases}$$

Use $Z_1, \ldots, Z_n$ to obtain an explicit estimator of p in terms of $Y_1, \ldots, Y_n$, $x_1, \ldots, x_n$ and $\widehat{\beta}$.

Hint: One possibility is to use estimating equations.

(iii) We may regard each δ_j as a missing observation or latent variable. What is the full log-likelihood of (Y_j, δ_j), $j = 1, \ldots, n$, given the regressors $x_1, \ldots, x_n$?

(iv) Evaluate the conditional expectations $E[\delta_j | Y_j = k]$, $k = 0, 1, 2, \ldots$.

(v) Use your answers in part (iii) and (iv) to show how the EM-algorithm can be used to estimate β and p. (You need to state the criterion that needs to be maximised and the steps of the algorithm).

(vi) Explain why for the EM-algorithm it is important to use good initial values.

Reference: Zero-inflated Poisson Regression, with an application to defects in manufactoring. **Diane Lambert**, *Technometrics*, vol 34, 1992.

Solution

(1) (a) Suppose that U is a Poisson distributed random variable, then for $k \geq 0$,

$$P(U = k) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{10.7}$$

(i) Let us suppose $\{U_j\}$ independent, identically distributed random variables from a Poisson distribution. What is the maximum likelihood estimator of λ?

It is clear that it is the sample mean, $\widehat{\lambda} = \frac{1}{n}\sum_{j=1}^{n} U_j$

(ii) For several count data sets it has been observed that there is an excessive number of zeros. To model 'inflated-zero' count data, the zero-inflated

Poisson distribution model was proposed where the observations are modelled as

$$Y = \delta U,$$

δ and U are random variables which are independent of each other, where δ is random variable taking either zero or one, $P(\delta = 0) = p$, $P(\delta = 1) = (1 - p)$ and U is a Poisson random variable as defined as in (10.7) Briefly explain why this model can is able to model excessive number of zeros.

The probability of zero is $P(Y = 0) = p + (1 - p)e^{-\lambda}$. Thus if p is sufficiently large, the chance of zeros is larger than the usual Poisson distribution (for a given λ).

(ii) Show that the estimator defined in (i) is a *biased* estimator of λ when the observations come from an zero-inflated Poisson distribution.

$\mathrm{E}[\widehat{\lambda}] = (1 - p)\lambda$**, thus when** $p > 0$**,** $\widehat{\lambda}$ **underestimates** λ**.**

(b) In this part of the question we consider the zero-inflated poisson regression model, proposed in Lambert (1992), which is defined as

$$Y_j = \delta_j U_j$$

where δ_j and U_j are random variables which are independent of each other, δ_j is an indicator variable, where $P(\delta_j = 0) = p$ and $P(\delta_j = 1) = (1 - p)$ and U_j has a Poisson regression distribution with

$$P(U_j = k|x_j) = \frac{\lambda_j^k e^{-\lambda_j}}{k!}$$

where $\lambda_j = e^{\beta x_j}$ and x_j is an observed regressor. Our objective is to first construct initial-value estimators for p and β and then use this to estimate as the initial values in when obtaining the maximum likelihood estimator.

(i) *Estimation of* β First obtain the distribution of Y_j conditioned on $Y_j > 0$ and x_j.

We note that $P(Y_j > 0) = P(\delta_j = 1, U_j > 0) = P(\delta_j = 1)P(U_j > 0) = (1 - p)(1 - e^{\lambda_j})$**. Similarly** $P(Y_j = k, Y_j > 0) = P(U_j = k, \delta_j = 1) = (1 - p)P(U_j = k)$**. Thus**

$$P(Y_j = k|Y_j > 0) = \frac{\lambda_j^k \exp(-\lambda_j)}{(1 - e^{\lambda_j})k!}$$

Let $\underline{Y}^+ = \{Y_j > 0\}$ (all the non-zero Y_j). Obtain the conditional log-likelihood of $\underline{Y}_+$ *conditioned* on $Y_j > 0$ and $\underline{x} = (x_1, \ldots, x_n)$. Derive the score equation and explain how β can be estimated from here. Denote this estimator as $\widehat{\beta}$.

The log-conditional likelihood is proportional to

$$\begin{aligned}\mathcal{L}_C(\beta) &= \sum_{Y_j>0} \left[Y_j \log \lambda_j - \lambda_j - \log(1 - e^{\lambda_j})\right] \\ &= \sum_{Y_j>0} \left\{\beta Y_j x_j - e^{\beta x_j} - \log(1 - e^{\beta x_j})\right\}.\end{aligned}$$

Thus to estimate β we differentiate the above wrt β (giving the score) and numerically solve the folling equation wrt β

$$\sum_{Y_j>0} Y_j x_j = \sum_{Y_j>0} x_j e^{\beta x_j} \left\{1 - \frac{1}{1 - e^{\beta} x_j}\right\}.$$

(ii) **Estimation of p** Define the dummy variable

$$Z_j = \begin{cases} 0 & \text{if } Y_j = 0 \\ 1 & \text{if } Y_j > 0. \end{cases}$$

Use $\{Z_j\}$ to obtain an explicit estimator of p in terms of $\underline{Y}$, $\underline{x}$ and $\widehat{\beta}$.
Hint: One possibility is to use estimating equations.

We solve the estimating equation

$$\sum_{j=1}^{n} [Z_j - \mathrm{E}(Z_j)] = 0,$$

wrt p. It is clear that $\mathrm{E}(Z_j) = P(Z_j = 1) = (1 - P(Z_j = 0)) = (1 - p)(1 - e^{-\lambda_j})$**. Thus the estimating equation is**

$$\sum_{j=1}^{n} \left[Z_j - (1 - p)(1 - e^{-\lambda_j})\right] = 0.$$

Replacing λ_j with $\hat{\lambda}_j = e^{\hat{\beta} x_j}$ and solving for p yields the estimator

$$\hat{p} = 1 - \frac{\sum_{j=1}^{n} Z_j}{\sum_{j=1}^{n} [1 - \exp(-e^{\hat{\beta} x_j})]}.$$

(iii) What is the complete log-likelihood of $\{Y_j, \delta_j; j = 1, \ldots, n\}$ (acting as if the variable δ_j is observed) given the regressors $\{x_j\}$?

The distribution of (Y_j, δ_j) is

$$P(Y_j = k, \delta_j) = [P(U_j = k)P(\delta_j = 1)]^{\delta_j} [P(\delta_j = 0)]^{1-\delta_j}.$$

Thus the log-likelihood of $\{Y_j, \delta_j; j = 1, \ldots, n\}$ is

$$\mathcal{L}_F(p, \beta) = \sum_{j=1}^{n} \delta_j [Y_j \log \lambda_j - \lambda_j + \log(1-p)] + \sum_{j=1}^{n} (1-\delta_j) \log p.$$

(iv) Evaluate the conditional expectations $\mathrm{E}[\delta_j | Y_j > 0]$ and $\mathrm{E}[\delta_j | Y_j = 0]$.

$\mathrm{E}[\delta_j | Y_j > 0] = 1$ **(since if $Y_j > 0$ then the only choice is $\delta_j = 1$),**

$$\mathrm{E}[\delta_j | Y_j = 0] = P(\delta_j = 1 | Y_j = 0) = \frac{(1-p)e^{-\lambda_j}}{p + (1-p)e^{-\lambda_j}}$$

and

$$\mathrm{E}[1 - \delta_j | Y_j = 0] = P(\delta_j = 0 | Y_j = 0) = \frac{p}{p + (1-p)e^{-\lambda_j}}$$

(v) Use your answers in part (iii) and (iv) to show how the EM-algorithm can be used to estimate β and p (you need to state the criterion that needs to be maximise and the steps of the algorithm).

Splitting the sum $\sum_{j=1}^{n}$ into $\sum_{Y_j>0}$ and $\sum_{Y_j=0}$, and taking expectations of $\mathcal{L}_F$ with respect to $\underline{Y}$ gives

$$\begin{aligned} Q(\theta; \theta^*) &= \sum_{Y_j>0} [Y_j \log \lambda_j - \lambda_j + \log(1-p)] \\ &\quad + \sum_{Y_j=0} \left(\frac{(1-p^*)e^{-\lambda_j^*}}{p^* + (1-p^*)e^{-\lambda_j^*}} \right) [\lambda_j + \log(1-p)] \\ &\quad + \sum_{Y_j=0} \left(\frac{p^*}{p^* + (1-p^*)e^{-\lambda_j^*}} \right) \log p \\ &= Q_1(\beta; \theta^*) + Q_2(p; \theta^*), \end{aligned}$$

where $\lambda_j^* = \exp(\beta^* x_j)$, $\theta = (p, \beta)$, $\theta^* = (p^*, \beta^*)$,

$$\begin{aligned} Q_1(\beta; \theta^*) &= \sum_{Y_j>0} [Y_j \log \lambda_j - \lambda_j] + \sum_{Y_j=0} (1-\pi^*)\lambda_j \\ Q_2(p; \theta^*) &= \sum_{Y_j>0} \log(1-p) + \sum_{Y_j=0} \left[(1-\pi_j^*) \log(1-p) + \pi_j^* \log p \right] \end{aligned}$$

and

$$\pi_j^* = \frac{p^*}{p^* + (1-p^*)e^{-\lambda_j^*}}.$$

Using $Q(\theta;\theta^*)$ we can then implement the EM-algorithm:

1. **Let $p^* = \hat{p}$ and $\beta^* = \hat{\beta}$. Then evaluate $Q_1(\beta;\theta^*)$ and $Q_2(p;\theta^*)$.**
2. **Differentiate $Q_1(\beta;\theta^*)$ wrt β and $Q_2(p;\theta^*)$ wrt p (keeping θ^* fixed) and solve for p and θ (needs to be done numerically). Set the solution $\theta^* = \hat{\theta}$.**
3. **Evaluate $Q_1(\beta;\theta^*)$ and $Q_2(p;\theta^*)$ with respect to the new θ^* and go back to (2).**
4. **Keep iterating until convergence.**

(vi) Explain why in the EM-algorithm it is important to use good initial values.
The EM algorithm is an iterative scheme which successively maximises the likelihood. However, if it climbs to a local maximum it will stay at that point. By using initial values, which are consistent, thus relatively close to the global maximum we can be reasonably sure that the EM-algorithm converged to a global maximum (rather than a local one).