Appendix C

Modelling Over-Dispersed Count Data

C.1 Extra-Poisson Variation

A key assumption of the Poisson regression model is that the variance equals the mean

$$\mathrm{var}(Y) = E(Y) = \mu.$$

However, count data often exhibit over-dispersion, with a variance larger than the mean. We now consider models that accommodate the excess residual variation.

An interesting feature of the IRSL algorithm used in generalized linear models is that it depends only on the mean and variance of the observations. Nelder and Wedderburn proposed specifying just the mean and variance and then applying the algorithm. The resulting estimates are called maximum quasi-likelihood estimates (MQLE), and have been shown to share many of the nice properties of maximum likelihood estimates (MLE) under fairly general conditions.

In the present context, suppose we were to assume that the variance is *proportional* to the mean, say

$$\mathrm{var}(Y) = \phi E(Y) = \phi\mu.$$

If $\phi = 1$ then the variance equals the mean. If $\phi > 1$, we have over-dispersion.

It turns out that applying the IRLS algorithm with this variance structure leads to exactly the same estimates as Poisson maximum likelihood. This implies that Poisson estimates are consistent when the variance is proportional (not just equal) to the mean.

However, the variance of the estimator in the more general case is

$$\text{var}(\hat{\beta}) = \phi(X'WX)^{-1}.$$

Under the Poisson assumption $\phi = 1$. Thus, Poisson standard errors tend to be conservative in the presence of over-dispersion.

If we knew ϕ we could, of course, correct the standard errors. Several authors have proposed estimating ϕ using Pearson's chi-squared statistic divided by its degrees of freedom:

$$\hat{\phi} = \frac{\chi_p^2}{n-p}.$$

A word of caution in using this approach is in order. Normally one would consider a large χ_p^2 as evidence of lack of fit. What we are doing here, put rather crudely, is relabelling lack of fit as extra-Poisson variation, and inflating our standard errors accordingly.

This suggests that one should be reasonably sure that the lack of fit is not due to poor specification of the systematic part of the model.

C.2 Negative Binomial Regression

An alternative approach to modelling over-dispersion is to start from a standard Poisson regression model and add a *random effect* θ_i to represent unobserved heterogeneity.

Suppose then, that the *conditional* distribution of the outcome Y_i given θ_i is indeed Poisson with mean (and variance) $\theta_i\mu_i$,

$$Y_i \sim P(\theta_i\mu_i).$$

The idea is that if we observed θ_i the data would be Poisson. Unfortunately, we do not observe θ_i. Instead, we assume that it has a given distribution. It turns out to be convenient to assume that θ_i has a gamma distribution with parameters $\alpha = \beta = 1/\sigma^2$, where σ^2 represents the variance of the unobservable.

With this information we can compute the *unconditional* distribution of the outcome, which happens to be a negative binomial distribution, with density

$$\Pr\{Y = y\} = \frac{\Gamma(\alpha + y)}{y!\Gamma(\alpha)} \frac{\beta^\alpha \mu^y}{(\mu + \beta)^{\alpha+y}},$$

where $\alpha = \beta = 1/\sigma^2$.

The negative binomial distribution is best known as the distribution of the number of failures before k successes in a series of Bernoulli trials with common probability of success π. The resulting density can be obtained from the expression above setting $\alpha = k$ and $\pi = \beta/(\mu + \beta)$.

The negative binomial distribution with $\alpha = \beta = 1/\sigma^2$ has mean

$$E(Y) = \mu$$

and variance

$$\text{var}(Y) = \mu(1 + \sigma^2 \mu)$$

If σ^2 is zero we obtain the Poisson variance. If $\sigma^2 > 0$ then the variance is larger than the mean. Thus, the negative binomial distribution is over-dispersed relative to the Poisson.

Interestingly, one can derive the same mean and variance without assuming that the unobservable has a gamma distribution. One can then proceed to estimate the parameters affecting μ for a fixed value of σ^2 using maximum quasi-likelihood. This strategy has been implemented in Stata's `glm` procedure. This doesn't solve the problem of estimating σ^2 itself. Breslow has proposed a strategy based on Pearson's chi-square, but we won't pursue this further.

The alternative is to use maximum likelihood, which requires assuming a gamma distribution for the unobservable, so that the outcome has a negative binomial distribution. This strategy, which makes stronger assumptions but yields estimates of both σ^2 and the parameters affecting μ, has been implemented in Stata's `nbreg` procedure.

Because the Poisson model is a special case of the negative binomial, namely the case with $\sigma^2 = 0$, one can use a standard likelihood ratio test to compare them. There is, however, one small difficulty. Because the Poisson model is in a boundary of the parameter space, the test statistic does not have the standard χ^2 distribution with one d.f. Some research suggests that the distribution is better approximated as a 50:50 mixture of zero and a chi-squared with one d.f., and this is what Stata does.

We note in closing that there are alternative formulations of the negative binomial model that lead to slightly different models, including one that leads to the over-dispersed Poisson of the previous section. The formulation given here, however, is the one in common use.