Overdispersion

For binomial or Poisson distribution, the variance is determined if the expected value is known. Sometimes in real application, we observe a deviance of a Pearson goodness of fit much larger than the expected if we assume the binomial or Poisson model. This can be explained by an overdispersion model. Overdispersion model describes the case when the observed variances are proportionally enlarged to the expected variance under the binomial or Poisson assumptions.

Binomial Data. Suppose $y_i \sim Bin(n_i, p_i)$. Then, $E(y_i|p_i) = n_i p_i$ and $V(y_i|p_i) = n_i p_i(1-p_i)$. In addition, suppose p_i is also a random variable with expected value p_{i0} and τ_i^2. Then, we can calculate the marginal expected value and variance of Y_i by

$$E(y_i) = E[E(y_i|p_i)] = E(n_i p_i) = n_i p_{i0}$$

and

$$\begin{aligned} V(y_i) =& V[E(y_i|p_i)] + E[V(y_i|p_i)] \\ =& V(n_i p_i) + E[n_i p_i(1-p_i)] \\ =& n_i^2\tau_i^2 + n_i p_{i0} - n_i(p_{i0}^2 + \tau_i^2) \\ =& (n_i^2 - n_i)\tau_i^2 + n_i p_{i0}(1-p_{i0}). \end{aligned}$$

In the overdispersion model, we need to choose τ_i^2 satisfying

$$\frac{V(y_i)}{n_i p_{i0}(1-p_{i0})} = \phi$$

where ϕ is a constant not depending on n_i and p_{i0}. Thus, we need to set up τ_i^2 as

$$\frac{(n_i-1)\tau_i^2}{p_{i0}(1-p_{i0})} = \phi - 1 \Rightarrow \tau_i^2 = \frac{(\phi-1)p_{i0}(1-p_{i0})}{n_i - 1}.$$

In this case, the observed variance is ϕ times as large as the expected variance.

Poisson Data. Suppose $y_i \sim Poisson(\lambda_i)$. Then, $E(y_i|\lambda_i) = V(y_i|\lambda_i) = \lambda_i$. For the same reason, assume λ_i is also aa random variable with expected value λ_{i0} and variance τ_i^2. Then, we have

$$E(y_i) = E[E(y_i|\lambda_i)] = E(\lambda_i) = \lambda_{i0}$$

and

$$V(y_i) = E[V(y_i|\lambda_i)] + V[E(y_i|\lambda_i)] = \lambda_{i0} + \tau_i^2.$$

Similarly, if we choose $\tau_i^2 = (\phi-1)\lambda_{i0}$, then, we have $V(y_i) = \phi E(y_i)$. Also in this case, the observed variance is ϕ times as large as the expected variance.

Models for overdispersion. When the overdispersion effect is significant, we recommend to estimate (McCullagh(1983)) ϕ by

$$\hat{\phi} = \frac{X^2}{df_{residual}}$$

where X^2 is the Pearson goodness of fit statistic. In this case, we need to adjust the likelihood ratio goodness of fit statistic by G^2/ϕ and the standard error of parameter estimate by the standard

error from the fitted model multiplied by $\sqrt{\hat{\phi}}$. For example, suppose there are 10 rows in the logistic model

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x.$$

Then, $df_{residual} = 8$. Suppose the Pearson goodness of fit is $X^2 = 40$, $\hat{\beta}_0 = -4$, $\hat{\beta}_1 = 0.5$ with standard error 0.5 and 0.1 respectively. Then, the z-values are -8 for the intercept term and the 5 for the slope term. When the overdispersion model is considered, we have

$$\hat{\phi} = \frac{40}{8} = 5,$$

and the adjusted standard error of intercept term is $0.5\sqrt{5} = 1.118$ and the adjusted standard error of the slope term is $0.1\sqrt{5} = 0.2236$. Thus, the adjusted z values become 3.578 and 2.236 respectively.

Other Models. There are quite a few models which can not described by the overdispersion model. Look at the following example.

Suppose in a disease study, we observe disease count y_i and at risk population ξ_i at m units. Suppose x_i is the corresponding independent variable. A Poisson regression model can be proposed as

$$\log(\lambda_i) = \log(\xi_i) + \beta_0 + \beta_1 x_i.$$

Then, we have

$$E(y_i|\lambda_i) = V(y_i|\lambda_i) = \xi_i e^{\beta_0+\beta_1 x_i}.$$

Suppose the true model is not the exactly logliner model but a loglinear mixed effect model as

$$\log(\lambda_i) = \log(\xi_i) + \beta_0 + \beta_1 x_i + N(0, \sigma^2).$$

Note that if $log(U) \sim N(\mu, \sigma^2)$, we have $E(U) = e^{\mu+\sigma^2/2}$ and $V(U) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$. Then, we have

$$E(\lambda_i) = \xi_i e^{\beta_0+\beta_1 x_i} e^{\sigma^2/2}$$

and

$$V(\lambda_i) = \xi_i^2 e^{2(\beta_0+\beta_1 x_i)}(e^{2\sigma^2} - e^{\sigma^2}).$$

Then, we have

$$E(y_i) = \xi_i e^{\beta_0+\beta_1 x_i} e^{\sigma^2/2}$$

and

$$V(y_i) = \xi_i e^{\beta_0+\beta_1 x_i} e^{\sigma^2/2} + \xi_i^2 e^{2(\beta_0+\beta_1 x_i)}(e^{2\sigma^2} - e^{\sigma^2}).$$

Since

$$\frac{V(y_i)}{E(y_i)} = 1 + \xi_i e^{\beta_0+\beta_1 x_i}(e^{3\sigma^2/2} - e^{\sigma^2/2})$$

is not a constant, this model is not an overdispersion model.

To make an overdispersion model, we need to choose

$$\log(\lambda_i) = \log(\xi_i) + \beta_0 + \beta_1 x_i + N(-\frac{\sigma_i^2}{2}, \sigma_i^2).$$

Then, we have

$$E(\lambda_i) = \xi_i e^{\beta_0+\beta_1 x_i}$$

and

$$V(\lambda_i) = \xi_i^2 e^{2(\beta_0+\beta_1 x_i)}(e^{\sigma_i^2} - 1).$$

If we choose

$$e^{\sigma_i^2} = 1 + \frac{(\phi - 1)}{\xi_i e^{\beta_0+\beta_1 x_i}} \Rightarrow e^{\sigma_i^2} - 1 = \frac{(\phi - 1)}{\xi_i e^{\beta_0+\beta_1 x_i}},$$

then, we have

$$\frac{V(y_i)}{E(y_i)} = \phi$$

and so this model is an overdispersion model.