BIOS 667: Longitudinal Data Analysis

Overdispersion - summary

The term *over-dispersion* refers to the case of a random variable having a larger variance than some theoretical distribution. So it is a relative term.

For example, *extra-Poisson* variation is relative to the Poisson distribution. A random variable Y is said to have extra-Poisson variation if $\text{var}(Y) > E[Y]$. A random variable that has extra-Poisson variation certainly can't have a Poisson distribution. The term is usually applied to counts, although it can be applied to any non-negative random variable.

Effect on inference: Underestimation of the variance; this affects the performance of confidence intervals and hypothesis tests.

Example: If we assume that $Y_1, \dots, Y_n$ are iid Poisson(μ), we obtain $\hat{\mu} = \bar{Y}$ and estimate $\text{var}(\bar{Y}) = \mu/n$ by $\hat{\mu}/n = \bar{Y}/n$. We would compute an approximate (for large n) 95% confidence interval for μ as $\bar{Y} \pm 1.96\sqrt{\bar{Y}/n}$. But suppose that $Y_1, \dots, Y_n$ are iid but not Poisson(μ), and $\text{var}(Y) = 4\mu$, then the above interval would be half as wide as it should be for proper 95% coverage. The coverage of $\bar{Y} \pm 1.96\sqrt{\bar{Y}/n}$ is about 67%.

Methods for handling extra-Poisson variation.

1. Replace the Poisson assumption by another parametric family such as the negative binomial. This typically introduces additional parameters that need to be estimated. The advantage is that maximum-likelihood estimation will be possible with all its advantages, provided the assumed family is correct. If it is not, then the consistency of the regression parameters β can be lost, even if the assumed model for the mean (link function and $X\beta$) is correct.

2. Use $\hat{\beta}$ from the Poisson model (i.e. likelihood), but use a robust variance estimator (RVE, this will be studied later). The RVE is a large-sample procedure that provides a valid estimate of $\text{cov}(\hat{\beta})$ provided the assumed model for the mean (link function and $X\beta$) is correct.

Extra-binomial variation refers to the situation: Y takes values in $[0, m]$, $\text{E}[Y] = \mu$ and $\text{var}(Y) > \mu(1 - \mu/m)$. If we define $p = \mu/m$, then $\mu = mp$ and $\text{var}(Y) > mp(1-p)$. Note that $\mu \in [0, m]$ and $p \in [0, 1]$. Again, this is commonly applied to counts, but in theory can be applied to continuous random variables too. Again, there are parametric families such as the beta-binomial that allow extra-binomial variation, but they have disadvantages as in the Poisson case. A good strategy is to use $\hat{\beta}$ from the Binomial model (i.e. likelihood), and a robust variance estimator.

An important case in which extra-binomial variation is mathematically impossible is: Y is Bernoulli, $Y \in \{0, 1\}$, $\text{E}[Y] = \mu = p$. The variance is $\text{var}(Y) = \mu(1 - \mu) = p(1 - p)$, (and can't be anything else).