Asymptotic standard errors for the logs of variance–inflated Poisson–variance means and of their ratios

Roger B. Newson

December 6, 2016

1 Formulas

We consider Poisson–variance variables Y_i (typically event counts), with positive means μ_i and variance inflation (or deflation) factors ϕ_i, such that the variance of Y_i is given by $\phi_i\mu_i$. Models with such variables were among those considered by Wedderburn (1974)[2]. The variance inflation (or deflation) may be imagined to be caused by clustering and/or weighting and/or adjustment for covariates. We focus on the case where there are two such variables Y_0 and Y_1, having expectations μ_0 and μ_1, with sum $\mu_+ = \mu_0 + \mu_1$, proportions $p_i = \mu_i/\mu_+$, and ratio $R = \mu_1/\mu_0 = p_1/p_0 = p_1/(1-p_1) = \text{odds}(p_1)$. Then we find, using Taylor polynomials, that the variances of the logs of the Y_i converge in ratio, as the μ_i become large, to

$$\text{Var}\left[\ln(Y_i)\right] = \mu_i^{-2}\phi_i\mu_i = \phi_i/\mu_i. \tag{1}$$

If we assume an equal–dispersion model in which the variances of Y_0 and Y_1 have a common variance inflation factor ϕ, then the variance of their log ratio, assuming independence, converges in ratio to

$$\begin{aligned} \text{Var}\left[\ln(Y_1/Y_0)\right] = &\ \phi\left(\tfrac{1}{\mu_0} + \tfrac{1}{\mu_1}\right) \\ = &\ \tfrac{\phi}{\mu_+}\left(\tfrac{1}{1-p_1} + \tfrac{1}{p_1}\right) \\ = &\ \tfrac{\phi}{\mu_+}\left(\tfrac{R+1}{1} + \tfrac{R+1}{R}\right) \\ = &\ \tfrac{\phi}{\mu_+}\left(\tfrac{R^2+2R+1}{R}\right) \\ = &\ \tfrac{\phi}{\mu_+}\tfrac{(R+1)^2}{R}. \end{aligned} \tag{2}$$

It follows that, if $\ln(Y_1/Y_0)$ is the estimator for $\ln(R)$, then its asymptotic sampling standard error is given by

$$\text{SE}\left[\ln(R)\right] = \sqrt{\frac{\phi}{\mu_+}\frac{(R+1)^2}{R}}. \tag{3}$$

This formula can be useful in power calculations for $\ln(R)$, if we think we have a good prior guess for the values of μ_+ and ϕ. Note that the standard error formula still applies if we are testing null hypotheses other than $R = 1$, as will be the case when Y_0 and Y_1 are event counts, associated with different exposure–time totals having a known exposure–time ratio.

The factor $(R+1)^2/R$ in (3) has the feature that

$$\begin{aligned} \tfrac{d}{dR}\left[\tfrac{(R+1)^2}{R}\right] = &\ \left[R\tfrac{d}{dR}(R+1)^2 - (R+1)^2\tfrac{dR}{dR}\right]/R^2 \\ = &\ \left[2(R+1)R - (R+1)^2\right]/R^2 \\ = &\ \left(2R^2 + 2R - R^2 - 2R - 1\right)/R^2 \\ = &\ (R^2-1)/R^2 \\ = &\ 1 - \tfrac{1}{R^2}. \end{aligned} \tag{4}$$

This expression is positive for $R > 1$, negative for $R < 1$, and zero for $R = 1$. So, for each value of ϕ/μ_+, the standard error of $\ln(R)$ is a smile-shaped function of R, with a minimum at $R = 1$. The log transformation therefore does not stabilize variances perfectly.

1.1 Applications to power calculations

Power calculations for Poisson rate ratios fall into the general theory summarized in Newson (2004)[1]. In that theory, the 5 quantities that may be calculated are power ($\gamma = 1 - \beta$ where β is the probability of Type

2 error), α (the probability of Type 1 error), δ (the detectable difference between parameter values under 2 hypotheses), σ (the standard deviation of the influence function of the parameter being estimated), and n (the necessary number of sampling units). Any one of these quantities can be calculated from the other 4 quantities. However, for our purposes, we will assume that $n = 1$, because the sum of several independent count variables is a count variable, and, if the variables being summed are Poisson, then so is their sum. So, we will assume that there is a single primary sampling unit, composed of 2 variance–inflated Poisson variables, with a common dispersion parameter ϕ, measuring overdispersion or underdispersion, which might be caused by clustering and/or sampling–probability weighting, or which might be caused by the component count variables themselves being overdispersed or underdispersed.

We assume that we are measuring the log ratio of the population means of the 2 count variables. And we assume that, under the null hypothesis being tested, the ratio R has the value of R_0. This value may be 1, if we are testing a null hypothesis of equal mean counts. Alternatively, R_0 may be a ratio between exposure times at risk, if we are testing a null hypothesis of equal event rates per unit time at risk. It is important to note that, if R_0 is an exposure–time ratio, then the ratio R between mean counts, under an alternative hypothesis, will be equal to the product of 2 ratios, namely the exposure–time ratio R_0 and the ratio of event rates per unit time at risk under that alternative hypothesis. Therefore, if we are doing power calculations for the detection of a ratio of event rates per unit exposure, then this rate ratio must be multiplied by R_0 to derive R for input into a power calculation. (Or, if the detectable ratio of mean counts R is output from a power calculation, then R must be divided by R_0 in order to derive a detectable rate ratio per unit exposure.)

Under these assumptions, once we have defined μ_+, ϕ, R and R_0, we have $n = 1$, $\delta = \ln(R/R_0) = \ln(R) - \ln(R_0)$, and the standard deviation σ of the influence function is given by (3) as a function of R, ϕ and μ_+. The quantities n, δ and σ appear in the equations in Newson (2004)[1], together with the values of γ and α under the appropriate scenarios. Given that n is fixed at 1, we can use these equations to derive any one of the 4 quantities δ, σ, α and γ from the other 3. This can be done using the `powercal` package in Stata, described in Newson (2004)[1], to do the power calculations for multiple scenarios, in a dataset with 1 observation per scenario. Each scenario will be defined by values of the 6 parameters γ, α, R, R_0, μ_+ and ϕ.

References

[1] Newson R. Generalized power calculations for generalized linear models and more. *The Stata Journal* 2004; **4(4)**: 379-401.

[2] Wedderburn RWM. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* 1974; **61(3)**: 439-447.