Applications of Resampling-based Simultaneous Confidence Interval Method to Over-dispersed Binomial Responses

Bo Li *

Abstract

In this article, we briefly overview the simultaneous confidence interval method in pairwise comparisons based on quasi-likelihood estimation of regression coefficients in generalized linear models. The simultaneous confidence intervals for the associated odd ratios are obtained by direct end-point transformation. A real example is used for illustration.

Key Words: Simultaneous Confidence Intervals; Quasi-likelihood Estimation; Percentile-t Bootstrap; Over-dispersion; Odds Ratio

1. Introduction

Dichotomous responses, obtained from surveys or experiments in scientific investigations, are often fit to generalized linear models to inference the regression coefficients as well as the unknown responding probabilities. Researchers often compare the odds for a given pair values of a predictor variable to examine the sensitivity of the responding probability to level changes. In multiple comparisons for a family of odds, a natural consideration is to control the family-wise error rate, or equivalently, to attain the prescribed joint confidence level.

Large-sample approximation methods have been widely used in simultaneous inference for a collection of responding probabilities, Kutner et al (2004). Simulation studies, including Simonoff and Tsai (1988) and Li (2020), have provided a caveat of using large-sample approximation method when data show over-dispersion. It is noted in Li (2020) that when over-dispersion presents, simultaneous confidence interval method based on normal theory provides liberal interval estimation of linear contrasts of regression coefficients; percentile-t bootstrap method is proposed as a resampling-based alternative. When functions of responding probabilities are of interest in a study, for instance, all pairwise odds ratios among treatment groups in case-control studies in Breslow and Day (1980), we construct simultaneous confidence intervals for odds ratios based on bootstrap method in Li (2020).

2. Simultaneous Confidence Intervals for Odds Ratios in Oneway Layouts

Let Y_{ij} be the $j-th$ Binomial observation from the $i-th$ treatment group, $i = 1, \cdots, a,\ j = 1, \cdots, n_i$ such that the variance and the mean follow the structure

*Department of Mathematical Sciences, The Citadel, The Military College of South Carolina, Charleston, SC, 29409

$$Var(Y_{ij}) = \phi n_i p_i(1-p_i), \tag{2.1}$$

where ϕ is the scale parameter and p_i is the common responding probability for each binomial response in group i. Note that ϕ is often used to capture over-dispersion in count data analysis, Wedderburn (1974), Hoef and Boveng (2007), Auer and Doerge (2010), and Li (2020^a). We assume that logit function of p_i is fit to the mean structure of Oneway ANOVA model that

$$log(\frac{p_i}{1-p_i}) = \mu + \tau_i, \tag{2.2}$$

where μ is the overall mean and τ_i denotes the treatment i effect, $\sum_{i=1}^{a} \tau_i = 0$. For details of quasi-binomial models, see section 5 of Li (2020).

To proceed, let C be the $k \times (a+1)$ Tukey-Type contrast matrix such that

$$C\boldsymbol{\beta} = [\tau_1 - \tau_2, \tau_1 - \tau_3, \cdots, \tau_{a-1} - \tau_a]' \tag{2.3}$$

where $k = \binom{a}{2}$ and $\boldsymbol{\beta} = [\mu, \tau_1, \cdots, \tau_a]'$.

Let $\widehat{\boldsymbol{\beta}}_Q = [\widehat{\mu}_Q, \widehat{\tau}_{1,Q}, \cdots, \widehat{\tau}_{a,Q}]'$ be the quasi-likelihood estimation of $\boldsymbol{\beta}$ in (2.2), Wedderburn (1974). We have that $(1-\alpha)100\%$ simultaneous confidence intervals of $C\boldsymbol{\beta}$ are given by

$$C\widehat{\boldsymbol{\beta}}_Q \pm q_{1-\alpha}[\widehat{\phi}\widehat{\Lambda}\mathbf{1}]^{1/2} \tag{2.4}$$

where (i) $\mathbf{1}$ is a $k \times 1$ vector of $1's$; (ii) $\widehat{\Lambda}$ is a $k \times k$ diagonal matrix whose $l-th$ diagonal element equals $\boldsymbol{c}_l'(X'\widehat{W}X)^{-1}\boldsymbol{c}_l, l = 1, \cdots, k$; (iii) $\widehat{W} = diag\{n_i e^{\widehat{\mu}_Q + \widehat{\tau}_{i,Q}}/(1 + e^{\widehat{\mu}_Q + \widehat{\tau}_{i,Q}})^2\}_{i=1,\cdots,a}$; (iv) the plug-in estimation of the scale parameter $\widehat{\phi} = \sum_{i,j}(Y_{ij} - \widehat{\mu}_{i,Q})^2/\{(N-a)[\widehat{\mu}_{i,Q}(n_i - \widehat{\mu}_{i,Q})/n_i]\}$ with $\widehat{\mu}_{i,Q} = n_i e^{\widehat{\mu}_Q + \widehat{\tau}_{i,Q}},\ i = 1, \cdots, a,\ N = \sum_{i=1}^{a} n_i$; (v) a widely used approximation method for the quantile $q_{1-\alpha}$ is based on normal theory, Kutner et al (2004) and Li (2020). In specific, Kutner et al (2004) used Bonferroni method and Li (2020) approximate $q_{1-\alpha}$ by the $(1-\alpha)-th$ quantile of the multivariate normal distribution with specifications $MVN(\mathbf{0}, \widehat{\Lambda}^{-1/2}C(X'\widehat{W}X)^{-1}C'\widehat{\Lambda}^{-1/2})$. The simulation study of Li (2020) shows that the quantiles generated by the corresponding multivariate normal distribution are below the "true" quantiles, obtained through Monte-Carlo simulation.

Li (2020) provided a resampling-based method to approximate $q_{1-\alpha}$. In brief, at the $b-th$ step of the bootstrap method, $b = 1, \cdots, B$ we draw a sample with replacement from Pearson residuals to obtain $b-th$ bootstrap copy of the pivotal quantities $T_Q^b = (\widehat{\phi}^{(b)}\widehat{\Lambda}^{(b)})^{-1/2}C(\widehat{\boldsymbol{\beta}}_Q^{(b)} - \widehat{\boldsymbol{\beta}}_Q)$ and that of the maximum modulus statistic $T_{Q,M}^b = max\{|T_Q^b|\}$. In the expression, the term with superscript (b) denotes the estimation of the corresponding parameter using the same estimation method above, based on $b-th$ bootstrap dataset. The $(1-\alpha)-th$ quantile of the sampling distribution of $T_{Q,M}^{(b)}$ gives a resampling-based approximation of $q_{1-\alpha}$ in (2.4).

Let $Odds_i = p_i/(1-p_i),\ i = 1, \cdots, 1$. The odds ratio of $(i, i') - th$ treamtent groups is given by

$$OR_{i,i'} = Odds_i/Odds_{i'} = e^{\tau_i - \tau_{i'}} \tag{2.5}$$

for $i \neq i' = 1, \cdots, a$. By end-point transformation, for example, page 562 of Kutner et al (2004), we have $(1-\alpha)100\%$ simultaneous confidence intervals for all pairwise odds ratios $\{OR_{i,i'}\}_{i \neq i'=1,\cdots,a}$ given by

$$[e^{L_{i,i'}}, e^{U_{i,i'}}] \tag{2.6}$$

where $L_{i,i'}$ is the lower bound of the simultaneous confidence interval for $\tau_i - \tau_{i'}$ in (2.4) and $U_{i,i'}$ is the upper bound of the interval.

Though we focus on all pairwise comparisons in this article, an extension of the proposed method to many-to-one comparisons is straightforward.

3. Example

To study the effect of plant nutrients on germination rate, two types of seeds ($O.aegyptiaca$ 75 and $O.aegyptiaca$ 73) were cultured under diluted root extracts from beans and cucumbers. The data is from Crowder (1978), who fit logit function of the unknown germination rates $p_i,\ i = 1, \cdots, 4$ to the mean structure of Oneway ANOVA model with 4 treatment groups: $O.aegyptiaca$ 75 with root extracts of beans, $O.aegyptiaca$ 75 with root extracts of cucumbers, $O.aegyptiaca$ 73 with root extracts of beans, $O.aegyptiaca$ 73 with root extracts of cucumbers, in order.

By fitting the model in (2.2), the scale parameter has plug-in estimation $\widehat{\phi} = 1.86(> 1)$, and Pearson statistic $(N-4)\widehat{\phi} = 31.65$ gives $p-value = 0.017$, indicating over-dispersion among observations, McCullagh and Nelder (1989) and Agresti (2007). The group averaged mean-variance plot in Figure 1 shows that the quadratic form presumed in (2.1) roughly holds for mean values less than 30 and a linear trend is observed for mean values greater than 30. The simulation results in Li (2020) show that the large-sample approximation method is sensitive to such deviation and the percentile-t bootstrap method in section 2 provides a robust alternative. The robustness of validity holds when the working variance-mean structure in (2.1) lies in a neighborhood of the "true" variance-mean structure, implied by mean-variance plot in practice. Hence, we apply the bootstrap method in section 2 to obtain 95% simultaneous confidence intervals for $OR_{i,i'},\ i \neq i' = 1, \cdots, 4$. As a comparison, we include the interval estimation based on normal theory in section 2 and Bonferroni method discussed in Kutner et al (2004). The results are summarized in Table 1. It shows that the bootstrap method in section 2 provides wider intervals than that using the large-sample approximation method in all pairwise comparisons for odds ratios. This is consistent with the results in inference linear contrasts of regression coefficients such as (2.3) in Li (2020). Moreover, Bonferroni method gives almost similar (slightly wider) interval estimation as that of the large-sample approximation method.

Table 1: Confidence Intervals of All Pairwise Comparisons for Odds Ratios - Nominal Confidence Level $1 - \alpha = 0.95$

Comparisons	$mvn^{\dagger}$	BS	Bonferroni
$OR_{1,2}$	$(0.144, 0.498)$	$(0.137, 0.524)$	$(0.141, 0.507)$
$OR_{1,3}$	$(0.396, 1.886)$	$(0.372, 2.010)$	$(0.387, 1.930)$
$OR_{1,4}$	$(0.241, 1.052)$	$(0.227, 1.117)$	$(0.236, 1.075)$
$OR_{2,3}$	$(1.483, 7.031)$	$(1.391, 7.495)$	$(1.449, 7.195)$
$OR_{2,4}$	$(0.903, 3.921)$	$(0.850, 4.165)$	$(0.884, 4.007)$
$OR_{3,4}$	$(0.243, 1.395)$	$(0.226, 1.499)$	$(0.237, 1.432)$

The bootstrap size B = 1,000,000.
† We use the package "mvtnorm" of Genz et al (2017) to generate the quantiles $q_{1-\alpha}$ in (2.4).
The user time to generate the results in Table 1 is 106.5 seconds on a desktop with the processor: Intel(R) Core(TM) i5-7600 CPU @ 3.50GHz, 3504 Mhz and Installed physical memory (RAM): 16.0 GB

REFERENCES

Agresti A. (2007), *An Itroduction to Categorical Data Analysis*, Hoboken, USA:Wiley.

Auer P. L., and Doerge R.W. (2010), "Statistical Design and Analysis of RNA Sequencing Data," *Genetics*, 185, 405–416.

Breslow N.E., and Day N.E. (1980), *Statistical Methods in Cancer Research Volume I: The Analysis of Case-Control Studies*, Lyon, International Agency for Research on Cancer (IARC Scientific Publications No. 32).

Crowder M.J. (1978), "Beta-Binomial Anova for Proportions," *Journal of Royal Statistical Society, Series C (Applied Statistics)*, 27(1), 34–37.

Genz A., Bretz F., Miwa T., Mi X., Leisch F., Scheipl F., and Hothorn T. (2017), mvtnorm: Multivariate Normal and t Distributions, R package version 1.0-6, URL http://CRAN.R-project.org/package=mvtnorm.

Hoef J.M.V., and Boveng, P.L. (2007), "Quasi-Poisson vs. Negative Binomial Regression: How Should We Model Overdispered Count Data?" *Ecology*, 88(11), 2766 – 2772.

Kutner M.H., Nachtsheim C.J., Neter J., and Li W. (2004), *Applied Linear Statistical Medels*, (5th ed.), New York: McGraw-Hill/Irwin.

Li B. (2020^{a}), "Simultaneous Inference of Differentially Expressed Isoforms for RNA Sequencing Data," *REVSTAT– Statistical Journal*, 18(2), 153 –163.

Li, B. (2020), " Simultaneous confidence intervals of estimable functions based on quasi- likelihood in generalized linear models for over-dispersed data," *Journal of Statistical Computation and Simulation*, DOI: 10.1080/00949655.2020.1807548.

McCullagh P., and Nelder J. (1989), *Generalized Linear Models*, (2nd ed.), London: Chapman & Hall.

Simonoff J.S., and Tsai, C.L. (1988), "Jackknifing and Bootstrapping Quasi-Likelihood Estimators," *Journal of Statistical Computation and Simulation*, 30(3), 213 – 232.

Wedderburn R. W. M. (1974), "Quasi-likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method," *Biometrika*, 61(3), 439 – 447.

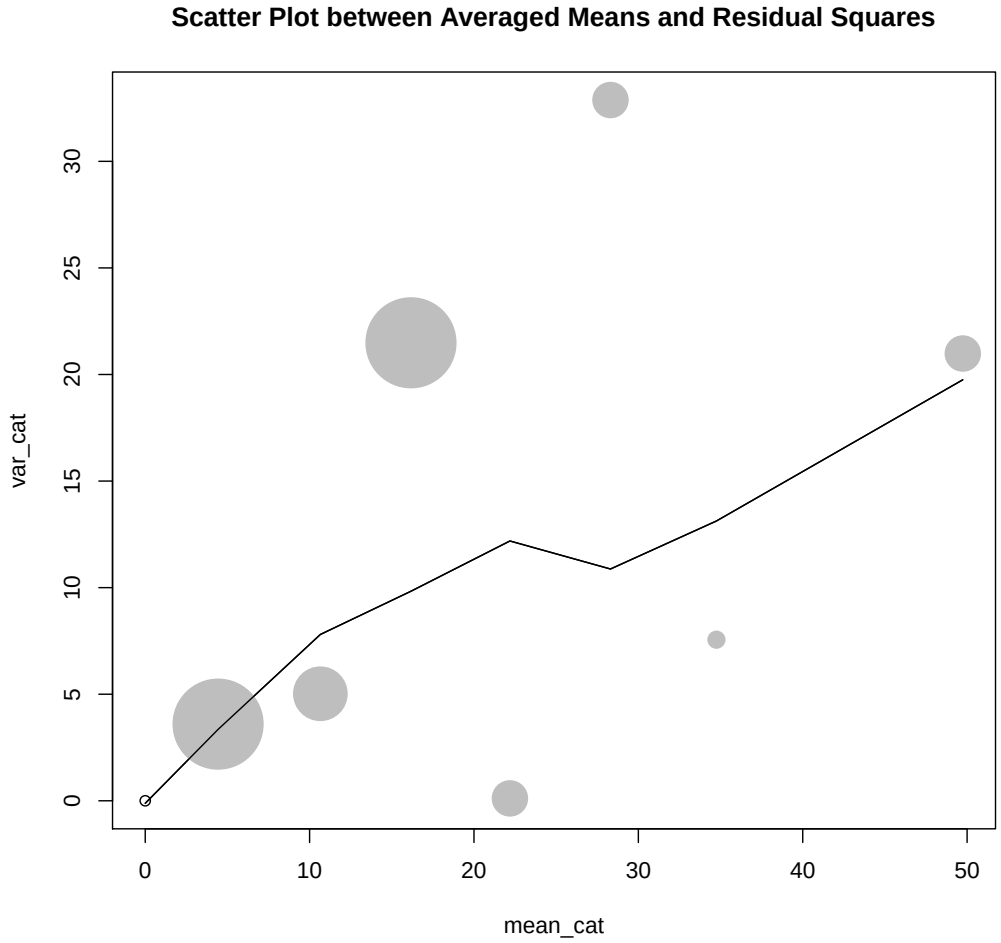

Figure 1. Scatter plot of group averaged means $\widehat{\mu}_i$ versus group averaged residual squares $(Y_i - \widehat{\mu}_i)^2$ for seed germination data, $i = 1, \cdots, 4$. The diameter of each circle is proportional to the size of the group, which has bin width 6.5. A $lowess$ (locally-weighted smoothing scatterplot) curve is included to observe the trend by using the smoother span $f = 0.75$.