TESTING FOR FAMILIAL AGGREGATION WHEN THE POPULATION SIZE IS KNOWN

Yixin Fang and Daniel Rabinowitz

Georgia State University and Columbia University

Supplementary Material

This note contains the proof that there is no complete sufficient statistic under design in which all affected families from a population of a known size are obtained, calculations of the asymptotic variances and the asymptotic relative efficiencies of those six test statistics, and algorithms to obtain the estimates of β in those three test statistics for the local alternative in which the prevalence of the latent factor with a substantial effect tends to zero.

S1. Proof of completeness

For design in which all affected families from a population of a known size are obtained, under the null hypothesis, the nuisance parameters are $\{p_0, \mathbf{M}\}$. If we define a function $g(D, \mathbf{M}^a)$ as following:

$$\binom{N-1}{D-1} g(D, \mathbf{M}^a)|_{m^a_{N-1}=1, m^a_1=1} + \binom{N-1}{D} g(D, \mathbf{M}^a)|_{m^a_{N-1}=1, m^a_1=0} = 0,$$

and $g = 0$ otherwise. It implies $E\{g(D, \mathbf{M}^a)\} = 0$ for any $\{p_0, \mathbf{M}\}$. Therefore, by the definition of completeness, $\{D, \mathbf{M}^a\}$ is not complete for nuisance parameter $\{p_0, \mathbf{M}\}$. On the other hand, it is minimal sufficient statistic. Hence, there is no complete sufficient statistic for the null hypothesis.

S2. Variances of test statistics

Define $S_D(p) = \sum_{i=1}^{I}[d_i^2 - 2n_i p d_i - (n_i p - n_i p^2 - n_i^2 p^2)]$. Noting that $S_D = S_D(\hat{p}) = S_D(p_0) + (\hat{p} - p_0) E_0(\partial S_D(p_0)/\partial p) + o_p(\sqrt{N})$, it can be verified that

$$Var_0(S_D) = 2\sum_{i=1}^{I} n_i(n_i - 1) p_0^2 (1 - p_0)^2.$$

Similarly, by Taylor expansion, we obtain

$$\begin{aligned} Var_0(S_{\mathbf{M}^a}) &= \sum_{i=1}^{I}[2n_i(n_i-1)p_0^2(1-p_0)^2 + \frac{A_i(p_0)B_i(p_0)(n_ip_0-3n_ip_0^2+n_i^2p_0^2)}{C_i(p_0)}] \\ &\quad -[\sum_{i=1}^{I}\frac{n_ip_0A_i(p_0)B_i(p_0)}{C_i(p_0)}]^2/Np_0(1-p_0), \end{aligned}$$

where $A_i(p) = n_ip - n_ip^2 - n_i^2p^2, B_i(p) = (1-p)^{n_i}$ and $C_i(p) = 1-(1-p)^{n_i}$. One estimate of $Var_0(S_{\mathbf{M}^a})$ is

$$\begin{aligned} \widehat{V}ar_0(S_{\mathbf{M}^a}) &= \sum_{i=1}^{I^a}[\frac{2n_i(n_i-1)\hat{p}^2(1-\hat{p})^2}{C_i(\hat{p})} + \frac{A_i(\hat{p})B_i(\hat{p})(n_i\hat{p}-3n_i\hat{p}^2+n_i^2\hat{p}^2)}{C_i^2(\hat{p})}] \\ &\quad -[\sum_{i=1}^{I^a}\frac{n_i\hat{p}A_i(\hat{p})B_i(\hat{p})}{C_i^2(\hat{p})}]^2/N\hat{p}(1-\hat{p}). \end{aligned}$$

Again by Taylor expansion and

$$p^* - p_0 = \sum_{i=1}^{I^a}(d_i - n_ip_0/C_i(p_0))/\sum_{i=1}^{I}[n_i - n_i^2p_0(1-p_0)^{n_i-1}/C_i(p_0)] + o_p(1),$$

we obtain

$$\begin{aligned} Var_0(\hat{S}_{D,\mathbf{M}^a}) &= \sum_{i=1}^{I}[2n_i(n_i-1)p_0^2(1-p_0)^2 - \frac{A_i(p_0)^2B_i(p_0)}{C_i(p_0)}] \\ &\quad -\frac{\{\sum_{i=1}^{I}[n_ip_0A_i(p_0)B_i(p_0)/C_i(p_0) - n_ip_0(1-p_0)(1-2p_0)]\}^2}{\sum_{i=1}^{I}n_ip_0[1-p_0-B_i(p_0)(1-p_0+n_ip_0)]/C_i(p_0)}. \end{aligned}$$

One estimate of it is

$$\begin{aligned} \widehat{V}ar_0(\hat{S}_{D,\mathbf{M}^a}) &= \sum_{i=1}^{I^a}[\frac{2n_i(n_i-1)p^{*2}(1-p^*)^2}{C_i(p^*)} - \frac{A_i(p^*)^2B_i(p^*)}{C_i^2(p^*)}] \\ &\quad -\frac{\{\sum_{i=1}^{I^a}[n_ip^*A_i(p^*)B_i(p^*)/C_i^2(p^*) - n_ip^*(1-p^*)(1-2p^*)/C_i(p^*)]\}^2}{\sum_{i=1}^{I^a}n_ip^*[1-p^*-B_i(p^*)(1-p^*+n_ip^*)]/C_i^2(p^*)}. \end{aligned}$$

Similarly,

$$Var_0(\hat{T}_D) = \sum_{i=1}^{I}\{[1+e^{\alpha}\frac{(e^{\beta}-1)^2}{(1+e^{\alpha+\beta})^2}]^{n_i} - 1 - n_ie^{\alpha}\frac{(e^{\beta}-1)^2}{(1+e^{\alpha+\beta})^2}\},$$

$$
\begin{aligned}
Var_0(T_{\mathbf{M}^a}) &= \sum_{i=1}^{I}\{[1+e^{\alpha}\frac{(e^{\beta}-1)^2}{(1+e^{\alpha+\beta})^2}]^{n_i}-1-B_i(p_0)[1-(\frac{1+e^{\alpha}}{1+e^{\alpha+\beta}})^{n_i}]^2/C_i(p_0)\} \\
&\quad -\{\sum_{i=1}^{I}\frac{n_ip_0}{C_i(p_0)}[\frac{e^{\beta}-1}{1+e^{\alpha+\beta}}-\frac{B_i(p_0)e^{\beta}}{1+e^{\alpha+\beta}}+(1-p_1)^{n_i}]\}^2/Np_0(1-p_0),
\end{aligned}
$$

and

$$
\begin{aligned}
Var_0(\hat{T}_{D,\mathbf{M}^a}) &= \sum_{i=1}^{I}\{[1+e^{\alpha}\frac{(e^{\beta}-1)^2}{(1+e^{\alpha+\beta})^2}]^{n_i}-1-B_i(p_0)[1-(\frac{1+e^{\alpha}}{1+e^{\alpha+\beta}})^{n_i}]^2/C_i(p_0)\} \\
&\quad -\frac{\sum_{i=1}^{I}\frac{n_ip_0}{C_i(p_0)}[\frac{e^{\beta}-1}{1+e^{\alpha+\beta}}-\frac{B_i(p_0)e^{\beta}}{1+e^{\alpha+\beta}}+(1-p_1)^{n_i}]^2}{\sum_{i=1}^{I}n_ip_0(1-p_0-B_i(p_0)(1-p_0+n_ip_0))/C_i(p_0)}.
\end{aligned}
$$

S3. Asymptotic relative efficiency

Start with the tests for the first locally alternative. Let $\bar{d}$ be $\{d_1,\cdots,d_k\}$. Denote probability distribution of $\bar{d}$ by $f(\bar{d};\theta,p,F)$, probability distribution of D by f_D, and conditional probability distribution of $\bar{d}$ given D by f_C. Because $E_0\{S_D\frac{\partial \log f_D}{\partial\theta}|_{\theta=0}\}=0$ and the derivative of the conditional log-likelihood $\frac{\partial \log f_C}{\partial\theta}|_{\theta=0}$ is zero, we have

$$
\frac{\partial E_\theta S_D}{\partial\theta}|_{\theta=0}=E_0\{S_D\frac{\partial \log f}{\partial\theta}|_{\theta=0}\}-E_0\{S_D\frac{\partial \log f_D}{\partial\theta}|_{\theta=0}\}=E_0\{S_D\frac{\partial \log f_C}{\partial\theta}|_{\theta=0}\}=0.
$$

In addition, $E_0\{S_D\frac{\partial \log f}{\partial\theta}\cdot\frac{\partial \log f_D}{\partial\theta}|_{\theta=0}\}=E_0\{\frac{\partial \log f_D}{\partial\theta}|_{\theta=0}E_0\{S_D\frac{\partial \log f}{\partial\theta}|D\}\}=0.$ Then we have

$$
\begin{aligned}
\frac{\partial^2 E_\theta S_D}{\partial\theta^2}|_{\theta=0} &= E_0\{S_D\frac{\partial^2 \log f}{\partial\theta^2}|_{\theta=0}\}+E_0\{S_D(\frac{\partial \log f}{\partial\theta}-\frac{\partial \log f_D}{\partial\theta})^2|_{\theta=0}\} \\
&= E_0\{S_D\frac{\partial^2 \log f_C}{\partial\theta^2}|_{\theta=0}\}=CVar_0(S_D),
\end{aligned}
$$

where $C=Var(A_1)$. Similarly, $\frac{\partial E_\theta S_{D,\mathbf{M}^a}}{\partial\theta}|_{\theta=0}=0$ and $\frac{\partial^2 E_\theta S_{D,\mathbf{M}^a}}{\partial\theta^2}|_{\theta=0}=CVar_0(S_{D,\mathbf{M}^a})$. Therefore, the calculation of $AE(S_D)$ and $AE(S_{D,\mathbf{M}^a})$ is straightforward. Now define

$$
\Delta_S=S_{\mathbf{M}^a}-S_D=\sum_{i=1}^{I}\frac{1}{1-(1-\hat{p})^{n_i}}(n_i\hat{p}-n_i\hat{p}^2-n_i^2\hat{p}^2)[1-(1-\hat{p})^{n_i}-I(d_i>0)],
$$

where $I(\cdot)$ is an indicator function. By the facts that

$$
\frac{\partial E_\theta[1-(1-\hat{p})^{n_i}-I(d_i>0)]}{\partial\theta}|_{\theta=0}=0,\quad \frac{\partial^2 E_\theta\hat{p}}{\partial\theta^2}|_{\theta=0}=\int\frac{\partial^2 p_\theta(a)}{\partial\theta^2}|_{\theta=0},\ \text{and}
$$

$$\frac{\partial^2 E_\theta I(d_i>0)}{\partial\theta^2}|_{\theta=0} = n_i(1-p_0)^{n_i-2}\{(1-p_0)\int\frac{\partial^2 p_\theta(a)}{\partial\theta^2}|_{\theta=0} - (n_i-1)\int(\frac{\partial p_\theta(a)}{\partial\theta}|_{\theta=0})^2\},$$

we have $\partial E_\theta S_{\mathbf{M}^a}/\partial\theta|_{\theta=0}=0$, and

$$\frac{\partial^2 E_\theta \Delta_S}{\partial\theta^2}|_{\theta=0} = C\sum_{i=1}^{I}\frac{(n_ip_0-n_ip_0^2-n_i^2p_0^2)(1-p_0)^{n_i}}{1-(1-p_0)^{n_i}}n_i(n_i-1)p_0^2.$$

Then the calculation of the asymptotic efficiency of $S_{\mathbf{M}^a}$ is straightforward, by the fact that

$$AE(S_{\mathbf{M}^a}) = (\frac{\partial^2 E_\theta\Delta_S}{\partial\theta^2}|_{\theta=0}+\frac{\partial^2 E_\theta S_D}{\partial\theta^2}|_{\theta=0})/NVar_0(S_{\mathbf{M}^a}).$$

Similarly, for the second local alternative, $PAE(T_D)=Var_0(T_D)/N$, $PAE(T_{D,\mathbf{M}^a}) = Var_0(T_{D,\mathbf{M}^a})/N$, and $PAE(T_{\mathbf{M}^a}) = (Var_0(\hat{T}_D)+\partial E_\theta\Delta_T/\partial\theta|_{\theta=0})^2/NVar_0(T_{\mathbf{M}^a})$, where $\Delta_T = T_{\mathbf{M}^a}-\hat{T}_D$ and $\partial E_\theta\Delta_T/\partial\theta|_{\theta=0}$ equals

$$\sum_{i=1}^{I}\frac{(1-p_0)^{n_i}}{1-(1-p_0)^{n_i}}\{n_ie^{\alpha}\frac{e^{\beta}-1}{1+e^{\alpha+\beta}}[1-(\frac{1+e^{\alpha}}{1+e^{\alpha+\beta}})^{n_i}]-[1-(\frac{1+e^{\alpha}}{1+e^{\alpha+\beta}})^{n_i}]^2\}.$$

S4. Estimation of β

To see the identifiability of mixture binomial model, readers are referred to Teicher (1961, p.248) or Teicher (1963, Proposition 4). Simply put, it is identifiable provided that the proportion of families with size greater or equal to three is not trivial. For estimation of β in the mixture binomial, there are many packages for the simple setting in which a simple random sample of families of same size is obtained; for example, see a review paper Haughton (1997). Here we review a method of moment proposed by Blischke (1962) for the case where family sizes are the same and greater or equal to three.e and greater or equal to three. Define the jth sample factorial moment

$$F_j = \frac{1}{I}\sum_{i=1}^{I}\frac{d_i(d_i-1)\cdots(d_i-j+1)}{n(n-1)\cdots(n-j+1)},\ \text{for } j=1,\cdots,n,$$

where n is the common family size. Because $E(F_j)=\theta p_1^j+(1-\theta)p_0^j$, by substituting F_j for $E(F_j)$, $j=1,2,3$, the moment estimates of p_0, p_1 and θ are, respectively,$\hat{p}_0 = A/2-(A^2-4AF_1+4F_2)^{\frac{1}{2}}/2, \hat{p}_1 = A/2+(A^2-4AF_1+4F_2)^{\frac{1}{2}}/2$, and $\hat{\theta}=(F_1-\hat{p}_0)/(\hat{p}_1-\hat{p}_0)$, where $A=(F_3-F_1F_2)/(F_2-F_1^2)$. If $A^2-4AF_1+4F_2\le$

0 or $(A^2-4AF_1+4F_2)^{\frac{1}{2}} \le \min(A, 2-A)$, p_0, p_1 and θ can be estimated by F_1, F_1 and 0, respectively. Blischke (1962) also analyzed the asymptotic efficiency of these estimates. This method can be generalized to the case in which family sizes are various. To this end, we can replace jth sample factorial moment F_j by

$$F_j = \sum_{i=1}^{I} \frac{d_i(d_i-1)\cdots(d_i-j+1)I(n_i \ge j)}{n_i(n_i-1)\cdots(n_i-j+1)} / \sum_{i=1}^{I} I(n_i \ge j), \text{ for } j = 1, 2, 3.$$

Furthermore, for the design in which all affected families from a population of a known size are obtained, we embed an iterative procedure into the above method of moment as following. Starting with the initial estimates $p_0^{(0)}, p_1^{(0)}$ and $\theta^{(0)}$, we estimate $P(d_i > 0)$ by $P_i^{(0)} = 1 - \theta^{(0)}(1-p_1^{(0)})^{n_i} - (1-\theta^{(0)})(1-p_0^{(0)})^{n_i}$, for $i = 1, 2, \cdots, I^a$. Then replacing $F_j, j = 1, 2, 3$, by

$$\begin{aligned} F_1 &= \sum_{i=1}^{I^a} d_i/N, \quad F_2 = \sum_{i=1}^{I^a} \frac{d_i(d_i-1)}{n_i(n_i-1)} P_i^{(0)} I(n_i \ge 2) / \sum_{i=1}^{I^a} I(n_i \ge 2), \\ F_3 &= \sum_{i=1}^{I^a} \frac{d_i(d_i-1)(d_i-2)}{n_i(n_i-1)(n_i-2)} P_i^{(0)} I(n_i \ge 3) / \sum_{i=1}^{I^a} I(n_i \ge 3), \end{aligned}$$

respectively, leads to updated moment estimates $p_0^{(1)}, p_1^{(1)}$ and $\theta^{(1)}$.

References

Blischke, W. R. (1962). Moment estimators for the parameters of a mixture of two binomial distributions. *The Annals of Mathematical Statistics* **33**, 444-454.

Haughton, D. (1997). Packages for estimating finite mixtures: a review. *The American Statistician* **51**, 194-205.

Teicher, H. (1961). Identifiability of mixtures. *The Annals of Mathematical Statistics* **32**, 244-248.

Teicher, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics* **34**, 1265-1269.