Published in final edited form as:

Biometrics. 2015 June ; 71(2): 556–557. doi:10.1111/biom.12260.

Reader Reaction On the generalized Kruskal-Wallis test for genetic association studies incorporating group uncertainty

Baolin Wu* and Weihua Guan

Division of Biostatistics, School of Public Health University of Minnesota

Summary

Acar and Sun (2013, *Biometrics*, 69, 427-435) presented a generalized Kruskal-Wallis (GKW) test for genetic association studies that incorporated the genotype uncertainty and showed its robust and competitive performance compared to existing methods. We present another interesting way to derive the GKW test via a rank linear model.

Keywords

Kruskal-Wallis test; Linear model; Score test; Wald test

1 Introduction

Acar and Sun (2013) proposed a generalized Kruskal-Wallis (GKW) statistic for association test of imputed SNPs accounting for the genotype imputation uncertainty. The GKW test is based on contrasting weighted rank sum statistics across different groups. Its asymptotic chi-square null distribution is derived using the rank statistics based central limit theorem.

For this article, we presented a rank linear regression model and derived the proposed GKW statistic as a score test statistic. The linear model approach makes the derivation more straightforward and transparent, and leads to a simplified and unified approach to the general rank based multi-group comparison problem. We present a F-type statistic based on the Wald test and numerically compare its performance to the GKW test.

2 Rank based association test

Consider N samples randomly sampled from a large population consisting of $k \geq 2$ disjoint groups. Denote by G the categorical variable taking values on $\{1, 2, \dots, k\}$. We want to

compare the k groups, of sizes $n_1, \dots, n_k$ with $\sum_{i=1}^k n_i = N$, based on a continuous response variable Y . Denote the rank of Y_j among the overall samples as r_j for $j = 1, \dots, N$. Suppose available to us are not G but probabilistic data of G : $(p_{1j}, p_{2j}, \dots, p_{kj})$, where $p_{ij} = \Pr(G_j = i)$

for $i = 1, 2, \dots, k$ and $j = 1, 2, \dots, N$, with $\sum_{i=1}^k p_{ij} = 1$.

*baolin@umn.edu

2.1 The generalized Kruskal-Wallis (GKW) test

Acar and Sun (2013) assumed a location-shift model for the continuous outcome. Formally, letting the distribution function of Y be of the form $F_i(y) = F(y - \theta_i)$ for group i , we wish to test $H_0 : \theta_1 = \theta_2 = \dots = \theta_k$ against H_A : not all θ_i 's are equal.

The GKW test is based on testing the equality of the weighted rank sum $R_i^* = \sum_{j=1}^N p_{ij} r_j$ across group $i = 1, \dots, k$. Acar and Sun (2013) showed the asymptotic multivariate normality of $\{R_i^* : i=1, \dots, k\}$ and constructed the corresponding GKW chi-square statistic.

2.2 Rank linear regression model

Another interesting way to derive the GKW test is via a rank linear model

$$r_j = \mu_1 + \sum_{i=2}^k p_{ij} \mu_i + \varepsilon_j, \quad \text{Var}(\varepsilon_j) = \sigma^2, \quad j=1, \dots, N. \quad (1)$$

Here μ_1 represents the mean rank of the first group, and $\mu_i (i > 1)$ represents the mean rank difference between the i -th and first group. The null hypothesis is $H_0 : \mu_2 = \dots = \mu_k = 0$, which amounts to testing $H_0 : \theta_1 = \dots = \theta_k$ under the location-shift model of Acar and Sun (2013). For the case of genetic association test ($k = 3$), this model treats the three genotypes as categorical without imposing any genetic model. Denote $\mathbf{R} = (r_1, \dots, r_N)^T$, and $\mathbf{P}$ the design matrix for the rank linear model (1): $P_{j1} = 1$, $P_{ji} = p_{ij}$, $i = 2, \dots, k$. With some algebra,

we can show that the Score statistic is $S_0 = \frac{1}{\hat{\sigma}_0^2} (\mathbf{R} - \hat{\mu}_0)^T \{ \mathbf{P}(\mathbf{P}^T \mathbf{P})^{-1} \mathbf{P}^T \} (\mathbf{R} - \hat{\mu}_0)$, where

$$\hat{\mu}_0 = \frac{N+1}{2}, \quad \text{and} \quad \hat{\sigma}_0^2 = \frac{\sum_{j=1}^N \{j - (N+1)/2\}^2}{N-1} = \frac{N(N+1)}{12}. \quad \text{In the following we show that } S_0 \text{ equals to the GKW statistic.}$$

Note that $\mathbf{H} = \mathbf{P}(\mathbf{P}^T \mathbf{P})^{-1} \mathbf{P}^T$ is a projection matrix into the column space of $\mathbf{P}$, and $\mathbf{H}$ is invariant under centering and scaling. Let $\bar{\mathbf{P}}$ be the covariate centered and scaled design

$$\text{matrix: } \bar{P}_{ji} = \frac{p_{ij} - \bar{p}_i}{\sqrt{\sum_{j=1}^N (p_{ij} - \bar{p}_i)^2}}, \quad \forall i > 1, \quad \text{where } \bar{p}_i = \sum_{j=1}^N p_{ij} / N. \quad \text{Therefore we can}$$

also write $S_0 = \frac{1}{\hat{\sigma}_0^2} (\mathbf{U}^T \mathbf{V}^{-1} \mathbf{U})$, where $\mathbf{U} = \mathbf{P}^T (\mathbf{R} - \hat{\mu}_0)$ and $\mathbf{V} = \mathbf{P}^T \bar{\mathbf{P}}$. We can check that $V_{11} = N$, $V_{1i} = V_{i1} = 0$, $\forall i > 1$, and

$$V_{il} = \frac{\sum_{j=1}^N (p_{ij} - \bar{p}_i)(p_{lj} - \bar{p}_l)}{\sqrt{\sum_{j=1}^N (p_{ij} - \bar{p}_i)^2} \sqrt{\sum_{j=1}^N (p_{lj} - \bar{p}_l)^2}}, \quad \forall i > 1, l > 1,$$

which is the sample correlation of $\mathbf{P}_i$ and $\mathbf{P}_l$, where $\mathbf{P}_i = (p_{i1}, \dots, p_{iN})^T$, and

$$U_1=0, U_i=\frac{\sum_{j=1}^N r_j p_{ij} - \frac{N+1}{2} N \bar{p}_i}{\sqrt{\sum_{j=1}^N (p_{ij} - \bar{p}_i)^2}}, \forall i > 1.$$

Note that $U_i/(\sqrt{N-1}\hat{\sigma}_0)$ is just the sample correlation of $\mathbf{R}$ and $\mathbf{P}_i$. Therefore a general form of the Score statistic (and hence the GKW) is

$$S_0=(N-1)\mathbf{\Gamma}^T\mathbf{\Sigma}^{-1}\mathbf{\Gamma}, \quad (2)$$

where $\mathbf{\Gamma}=(\gamma_2,\dots,\gamma_k)^T$ is the sample correlation vector of response rank and probabilistic covariates $\{\mathbf{P}_i : i > 1\}$, and $\mathbf{\Sigma}=(\rho_{il})$ is the sample correlation matrix of corresponding probabilistic covariates, $\rho_{il}=\text{corr}(\mathbf{P}_i, \mathbf{P}_l)$ ($i > 1, l > 1$). This general form also readily leads to the asymptotic chi-square distribution based on the standard central limit theorem argument. When $k=2$, $S_0=(N-1)\gamma_2^2$. For association testing of imputed SNPs, $k=3$, and

$S_0=\frac{N-1}{1-\rho_{23}^2}(\gamma_2^2+\gamma_3^2-2\rho_{23}\gamma_2\gamma_3)$. And they exactly equal to the GKW statistic defined in Acar and Sun (2013, Section 2.2).

The proposed rank linear model (1) also provides a simplified approach to testing more general hypotheses. For example, with a binary environment covariate and an imputed SNP, we can create a group variable with six categories. Detecting gene-environment interaction or testing the overall genetic effect by leveraging gene-environment interactions (Kraft et al., 2007) using the robust sample rank can both be formulated as testing the linear combination of regression parameters under the linear rank model, and are worthy of further study.

As we have shown the equivalence of the GKW and Score statistic. It is natural to consider the asymptotically equivalent Wald statistic, which is the commonly used F-statistic in linear model. For linear regression model, the Score and Wald statistics have the same form and differ only in their variance calculation with the Wald statistic variance computed under the full model. The Wald statistic can be checked to be

$$W_0=\frac{1}{\hat{\sigma}_1^2}(\mathbf{R}-\hat{\mu}_0)^T\{\mathbf{P}(\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T\}(\mathbf{R}-\hat{\mu}_0), \text{ where } \hat{\sigma}_1^2=\frac{1}{n-k}\hat{\sigma}_0^2(N-1-S_0).$$

Computationally, these tests can be readily and very efficiently computed from existing software for linear model regressing the sample rank on the probabilistic covariates.

3 Numerical studies

We follow the same simulation settings of Acar and Sun (2013, Section 3.1) to evaluate the performance of GKW and the rank linear model based Score and Wald statistics for SNP association test. Numerically we have checked the exact equivalence of GKW and Score statistics, and consistently observed that the Wald statistic performs slightly better than the GKW statistic (especially for relatively small sample size).

The rank linear model framework (1) and general formula for the Score/GKW statistic (2) also suggest that it helps to use normally transformed rank vector. Specifically we use

$\Phi^{-1}(\frac{R_j}{N+1})$ instead of $r_j, j = 1, \dots, N$, in the rank linear model (1). In further simulation studies (not shown) with minor allele frequency of 0.1, 0.2, and 20% group uncertainty, we found that for all of the Score statistic S_0 , GKW statistic, and the Wald statistic W_0 , using the normally transformed ranks gave slightly improved power. All statistics have the right size.

Acknowledgments

This research was supported in part by NIH grant GM083345. We are grateful to the University of Minnesota Supercomputing Institute for assistance with the computations. We would like to thank the editor and associate editor for their valuable comments, which have greatly improved the presentation of the paper.

References

- Acar EF, Sun L. A generalized Kruskal-Wallis test incorporating group uncertainty with application to genetic association studies. *Biometrics*. 2013; 69:427–435. [PubMed: 23441822]
- Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. *Human Heredity*. 2007; 63:111–119. [PubMed: 17283440]