ENHANCING POWER OF SCORE TESTS FOR REGRESSION MODELS VIA FISHER TRANSFORMATION

Masao Ueki*

ABSTRACT

A simple method is presented to enhance statistical power of score tests for regression models via Fisher transformation (or Fisher’s z-transformation) by exploiting a relationship with the partial correlation coefficient. Simulation studies mimicking marginal association and gene-environment interaction analyses for genome-wide association studies (GWASs) under case-control design demonstrate that the Fisher transformation enhances power of the score tests while maintaining type I error asymptotically. The smaller the sample size is, the more the enhancement is pronounced, at the expense of inflated type I error due to invalidating asymptotic approximation. Accordingly, the proposed method may be applied when sample size is enough for valid asymptotic approximation. An illustration with real GWAS data is also presented.

1. Fisher-transformation of score tests for regression models

Suppose that \( n \) response variables \( y = (y_1, \ldots, y_n)^T \) and an \( n \times p \) design matrix \( X = (x_1, \ldots, x_n)^T \) are observed, where \( x_i \) is a \( p \)-dimensional column vector of explanatory variables for subject \( i \in \{1, \ldots, n\} \). Let \( f(y_i \mid x_i) \) denote the probability distribution of \( y_i \) conditional on \( x_i \) for each \( i \). Here, the probability density function of a continuous random variable or the probability mass function of a discrete random variable is referred to as a probability distribution (Dobson, 2002). Assume that a transformed conditional expectation of \( y_i \) through some differentiable monotone function (i.e. the link function) is written as \( x_i^T \beta \), in which \( \beta \) is a vector of corresponding \( p \) regression coefficients. Then, denote the loglikelihood by \( \ell(x_i^T \beta) = \log f(y_i \mid x_i) \) for the \( i \)th sample. Throughout, it is assumed that each \( y_i \) is independently distributed given \( x_i \). The above regression framework includes the generalized linear models (McCullagh and Nelder, 1989; Dobson, 2002) and regression with heavy-tailed error distribution (Lange and Sinsheimer, 1993). Suppose that \( X \) is partitioned into two parts as \( (X_1, X_2) \), where \( X_1 \) is a collection of \( q \) \( (q < p) \) explanatory variables to be tested for association with \( y \) and \( X_2 \) is a set of \( p - q \) covariates to be adjusted for. Correspondingly, let \( \beta = (\beta_1^T, \beta_2^T)^T \) and \( x_i = (x_{1,i}^T, x_{2,i}^T)^T \). In this article, \( X \) is assumed to be of full column rank.

1.1. Fisher-transformed score test: single parameter case

This subsection considers the case of \( q = 1 \), and hence the corresponding regression coefficient is written as \( \beta_1 \) with a non-bold letter. In genome-wide association study (GWAS)
applications (e.g. WTCCC, 2007), $X_1$ corresponds to a user-defined variable for the genotype at a single nucleotide polymorphism (SNP) including additive (0,1,2 coding for minor allele count), dominant and recessive models. For testing the null hypothesis $H_0 : \beta_1 = 0$ via the score test (Rao, 1948), the following argument considers the test statistic having the form of

$$t_1 = \frac{u_1}{v_{11}},$$

where $u_1$ is the score function corresponding to $\beta_1$, i.e. the first derivative of the log-likelihood function $L(\beta) = \sum_{i=1}^{n} \ell(x_i^T \beta)$ with respect to $\beta_1$, and $v_{11}$ is the asymptotic variance of $u_1$. Also, $u_1 = X_1^T \ell$, where $\ell = (\ell_1, \ldots, \ell_n)^T$, $v_{11} = 1/\{(X^T W X)^{-1}\}_{11}$, $W = \text{diag}(\hat{\ell}_1, \ldots, \hat{\ell}_n)$, $\hat{\ell}_i = \ell(x_i^T \hat{\beta}_2)$ and $\hat{\ell}_i = \ell(x_i^T \beta_2)$. $\ell(.)$ and $\hat{\ell}(.)$ are the first and second derivatives of $\ell(\cdot)$ with respect to $\cdot$, respectively. $\hat{\ell}_i$ is assumed to be positive for all $i$. $\hat{\beta}_2$ is the maximum likelihood estimator of $\beta_2$ under $H_0$. Let $Z_1 = W^{1/2} X_1$, $Z_2 = W^{1/2} X_2$, $Z = (Z_1, Z_2)$ and $Q Z_2 = I - P Z_2$, where $\text{diag}(\cdot)$ denotes a diagonal matrix with $\cdot$ in its diagonal elements, $I$ is the identity matrix and $P Z_2 = Z_2 (Z_2^T Z_2)^{-1} Z_2^T$. Noting that $v_{11} = ||Q Z_2 Z_1||^2$ (e.g. Ueki and Kawasaki, 2013) and that $X_2^T \ell = 0$, it holds that

$$t_1 = \frac{X_1^T \ell}{||Q Z_2 Z_1||} = \frac{(Q Z_2 Z_1)^T \hat{\ell}^*}{||Q Z_2 Z_1||} = \rho_1 ||\hat{\ell}^*||,$$

where $\hat{\ell}^* = W^{-1/2} \ell$ and $\rho_1 = (Q Z_2 Z_1)^T \ell^* / \{||Q Z_2 Z_1|| ||\hat{\ell}^*||\}$. The Cauchy–Schwartz inequality deduces that $|\rho_1| \leq 1$, and provides an upper bound for $|t_1|$ as $||\hat{\ell}^*|| = (-\sum_{i=1}^{n} \hat{\ell}_i^{-1} \hat{\ell}_i^2)^{1/2}$. Since $\ell^* = Q Z \hat{\ell}^*$, $\rho_1$ can be interpreted as the sample partial correlation coefficient between $Z_1$ and $\hat{\ell}$ given $Z_2$. The score test exploits the fact that the null distribution of $t_1$ is asymptotically standard normal under standard regularity conditions (e.g. van der Vaart, 1998). Another interpretation from the expression (1) is that the null distribution of $\rho_1$ is asymptotically a normal distribution with mean zero and variance $1/||\hat{\ell}^*||^2$. (Here $\hat{\ell}^*$ is evaluated at the true parameter rather than at $\hat{\beta}_2$.) For testing the null hypothesis that the correlation is zero, the Fisher transformation, $F(r) = \text{atanh}(r)$, is popularly used (Fisher, 1915, 1921). According to the delta method (e.g. van der Vaart, 1998), the null distribution of Fisher-transformed $\rho_1$, $F(\rho_1)$, is asymptotically normal with mean zero and variance $F'(0)^2/||\hat{\ell}^*||^2 = 1/||\hat{\ell}^*||^2$ since $F'(r) = 1/(1-r^2)$. The above argument naturally leads to a new hypothesis test using the Fisher-transformed score statistic,

$$t_F^2 = F(\rho_1)||\hat{\ell}^*|| = F(t_1/||\hat{\ell}^*||)||\hat{\ell}^*||.$$

Its null distribution is asymptotically standard normal. As a function of $r$, $F(r) \approx r$ for $|r| < 0.5$ and $|F(r)| > |r|$ otherwise. Furthermore, by letting $h(r) = \text{atanh}(r) - r$, since $dh(r)/dr = 1/(1-r^2) - 1 = r^2/(1-r^2) > 0$ for $r \in (0,1]$ and $h(0) = 0$, it holds that $\text{atanh}(|r|) \geq |r|$ for the entire interval $r \in (-1,1)$. Consequently, $t_F^2 \geq t_1$, that is, the test with the Fisher-transformed score statistic $t_F^2$ is more powerful than the test with the original score statistic $t_1$, because the same asymptotic null distribution is used for both tests.

The quantity $||\hat{\ell}^*|| = \left(-\sum_{i=1}^{n} \hat{\ell}_i^{-1} \hat{\ell}_i^2\right)^{1/2}$ is a key to the extent of improvement. If $||\hat{\ell}^*||$ is large, $t_1/||\hat{\ell}^*||$ is shrunken toward zero, and $F(t_1/||\hat{\ell}^*||) \approx t_1/||\hat{\ell}^*||$, i.e. less different from the original score statistic. The quantity $||\hat{\ell}^*||$ can be interpreted as a total residual
Fisher-transformed Score Test

of the null model. Actually, in a logistic regression model, \( f(y_i | \mathbf{x}_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \) with \( \pi_i = \pi(\mathbf{x}_i^T \beta) = 1/(1 + \exp(-\mathbf{x}_i^T \beta)) \), it can be written as

\[
||\hat{\ell}^*||^2 = \sum_{i=1}^{n} (y_i - \hat{\pi}_i)^2/\hat{w}_i,
\]

where \( \hat{\pi}_i \) is the null estimate of conditional probability and \( \hat{w}_i = \hat{\pi}_i(1 - \hat{\pi}_i) \).

Remark 1. For \( \mathbf{X}_2 = \mathbf{1} \), where \( \mathbf{1} \) is an \( n \)-dimensional vector of ones, the above-mentioned score test for testing the effect of \( \mathbf{X}_1 \) coincides with the Cochran–Armitage trend test. Let \( n_k = \# \{i \in \{1, \ldots, n\} : y_i = k \} \) for \( k = 0, 1 \). In this case, \( \hat{\beta}_2 \equiv \hat{\pi} = \sum_{i=1}^{n} y_i/n = n_1/n, \) \( \ell_i = y_i - \hat{\pi}, \ell_i = -\hat{\pi}(1 - \hat{\pi}), \mathbf{Z}_1 = \sqrt{\hat{\pi}(1 - \hat{\pi})}\mathbf{X}_1, \mathbf{Z}_2 = \sqrt{\hat{\pi}(1 - \hat{\pi})}\mathbf{1}, \mathbf{Q}_{\mathbf{Z}_2}\mathbf{Z}_1 = (I - 11^T/n)\sqrt{\hat{\pi}(1 - \hat{\pi})}\mathbf{X}_1, \mathbf{u}_1 = \mathbf{X}_1^T \ell, \mathbf{u}_1 = \mathbf{X}_1^T (\mathbf{y} - \hat{\pi}\mathbf{1}), \mathbf{v}_{11} = ||\mathbf{Q}_{\mathbf{Z}_2}\mathbf{Z}_1||^2 = \hat{\pi}(1 - \hat{\pi})\sum_{i=1}^{n}(X_{1i} - X_1)^2, \) and \( ||\hat{\ell}^*||^2 = \sum_{i=1}^{n} (y_i - \hat{\pi})^2/(\hat{\pi}(1 - \hat{\pi})) = \{n_1(1 - \hat{\pi})^2 + n_0\hat{\pi}^2\}/\{\hat{\pi}(1 - \hat{\pi})\} = n_1(1 - \hat{\pi})/\hat{\pi} + n_0\hat{\pi}/(1 - \hat{\pi}) = n \). Thus, for the Cochran–Armitage trend statistic,

\[
t_1 = \frac{u_1}{\sqrt{v_{11}}} = \frac{X_1^T (\mathbf{y} - \hat{\pi}\mathbf{1})}{\sqrt{\hat{\pi}(1 - \hat{\pi}) \sum_{i=1}^{n}(X_{1i} - X_1)^2}},
\]

it holds that \( \rho_1 = t_1/||\hat{\ell}^*|| = t_1/\sqrt{n} \in (-1, 1) \). The Fisher-transformed Cochran–Armitage trend statistic is given by

\[
t_F^1 = \text{atanh}(\rho_1)||\hat{\ell}^*|| = \text{atanh} \left\{ \frac{X_1^T (\mathbf{y} - \hat{\pi}\mathbf{1})}{\sqrt{\hat{\pi}(1 - \hat{\pi}) \sum_{i=1}^{n}(X_{1i} - X_1)^2}} \right\} \sqrt{n},
\]

whose null distribution is asymptotically standard normal. A relationship between the Cochran–Armitage trend statistic and correlation coefficient is pointed out by Agresti (2002).

Remark 2. In the setting of Remark 1, suppose further that each component of \( \mathbf{X}_1 \) is binary, i.e. \( x_{1i} \in \{0, 1\} \) for \( i = 1, \ldots, n \). Let \( n_{jk} = \# \{i \in \{1, \ldots, n\} : x_{1i} = j, y_i = k \} \) for \( j \in \{0, 1\} \) and \( k \in \{0, 1\} \). Also, let \( n_{j+} = n_{j0} + n_{j1} \) for \( j \in \{0, 1\} \). In this case,

\[
t_1 = \frac{u_1}{\sqrt{v_{11}}} = \frac{n_{11} - n_{1+n}n_1}{\sqrt{n_1n_0n_{1+n}n_{0+n}/n}},
\]

and \( \rho_1 = t_1/\sqrt{n} \). In the context of genetics, if \( \mathbf{y} \) and \( \mathbf{X}_1 \) denote indicator variables of binary alleles at two loci, \( \rho_1 \) coincides with the Pearson’s \( r^2 \) measure of linkage disequilibrium between two loci. Wellek and Ziegler (2009) proposed a Fisher transformation of their \( r^2 \)-like measure for inference of linkage disequilibrium.

1.2. Fisher-transformed score test: multiple parameter case

This subsection presents an extension of the above Fisher transformation to the score statistics for testing \( q \)-dimensional parameters, namely, when \( \mathbf{X}_1 \) is \( n \times q \) matrix for \( 1 \leq q < p \). The corresponding null hypothesis is \( H_0 : \beta_1 = \mathbf{0} \), where \( \beta_1 \) is the \( q \)-dimensional vector of regression coefficients of \( \mathbf{X}_1 \). Without loss of generality, the tested variables are in the first \( q \) elements in \( \beta \). The test statistics considered have the form

\[
T_1 = \mathbf{u}_1^T \mathbf{V}_{11}^{-1} \mathbf{u}_1,
\]

where \( \mathbf{u}_1 = \mathbf{X}_1^T \ell \) and \( \mathbf{V}_{11} = [\{(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}\}_{11}]^{-1} \). The null distribution of \( T_1 \) is asymptotically a \( \chi^2_q \)-distribution. A generalized argument of the above single-parameter case for \( \{(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}\}_{11} \) leads to

---

39---
\[ V_{11} = \{(X^TWX)^{-1}\}_{11}^{-1} = \{(Z^TZ)^{-1}\}_{11}^{-1} = (QZ_2Z_1)^T(QZ_2Z_1). \]

By noting that \[ X_2^T\hat{e} = 0, u_1 = (QZ_2Z_1)^T\hat{e}. \] Then, \[ T_1 \] can be written as
\[ T_1 = \{(QZ_2Z_1)^T\hat{e}\}^T \{(QZ_2Z_1)^T(QZ_2Z_1)\}^{-1}\{(QZ_2Z_1)^T\hat{e}\}. \]

Put \[ \Sigma_{11} = \text{Diag}\{(QZ_2Z_1)^T(QZ_2Z_1)\}, \] where \text{Diag} is the operator which makes the off-diagonal entries of the matrix zero. By letting \[ \rho_1 = \Sigma_{11}^{-1/2}(QZ_2Z_1)^T\hat{e}/||\hat{e}||, \] the following expression is obtained:
\[ T_1 = \rho_1^{T}\{\Sigma_{11}^{-1/2}(QZ_2Z_1)^T(QZ_2Z_1)\Sigma_{11}^{-1/2}/||\hat{e}||^2\}^{-1}\rho_1. \]

It implies that the null distribution of \[ \rho_1 \] is asymptotically multivariate normal with the mean a vector of zeros and variance-covariance matrix \[ \Sigma_{11}^{-1/2}(QZ_2Z_1)^T(QZ_2Z_1)\Sigma_{11}^{-1/2}/||\hat{e}||^2. \]

Since the \[ j \]-th diagonal element of \[ \Sigma_{11} \] is \[ \Sigma_{jj} = ||QZ_2Z_j||^2 \], for \[ j = 1, \ldots, q \],
\[ ||(QZ_2Z_j)^T\hat{e}|| \leq ||\hat{e}||/\sqrt{\Sigma_{jj}} \]
by the Cauchy–Schwartz inequality. Thus, each component of \[ \rho_1 \] is between \([-1, 1] \). Since \[ \hat{e} = QZ_2\hat{e} \], the \[ j \]-th component of \[ \rho_1 \],
\[ \frac{(QZ_2Z_j)^T\hat{e}}{||QZ_2Z_j||||\hat{e}||} = \frac{(QZ_2Z_j)^T(QZ_2\hat{e})}{||QZ_2Z_j||||QZ_2\hat{e}||} \]
can be interpreted as the partial correlation between \[ \hat{e} \] and \[ Z_j \] given \[ Z_2 \]. As in the single parameter case, application of the Fisher transformation is proposed for each component of \[ \rho_1 \]. Under the null hypothesis \[ \rho_1 = 0 \], the component-wise Fisher-transformed \[ \rho_1, F(\rho_1) \] say, follows asymptotically a multivariate normal distribution with the means a vector of zeros and variance-covariance matrix \[ \nabla F(0)^T\Sigma_{11}^{-1/2}(QZ_2Z_1)^T(QZ_2Z_1)\Sigma_{11}^{-1/2}\nabla F(0)/||\hat{e}||^2 \]
by the delta method. Here, \[ \nabla F(r) \] is the \[ q \times q \] matrix whose \((j,k)\)-component is given by
\[ \partial F(r_{jj})/\partial r_{kk} = 1/(1-r_{jj}^2)1_{\{j=k\}}, \]
and hence \[ \nabla F(0) = I. \] Here \( 1_A \) is the indicator function taking a value of 1 if \( A \) is true and 0 otherwise. Thus, the null distribution of \[ F(\rho_1) \] is asymptotically the same as that of \[ \rho_1 \].

The proposed Fisher-transformed test statistic for the multiple parameter case is as follows:
\[ T_1^F = F(\rho_1)^T\{\Sigma_{11}^{-1/2}(QZ_2Z_1)^T(QZ_2Z_1)\Sigma_{11}^{-1/2}/||\hat{e}||^2\}^{-1}F(\rho_1), \]

whose null distribution is asymptotically a \( \chi^2 \)-distribution. When \( q = 1 \), the above test statistic reduces to the Fisher-transformed score statistic presented in the previous subsection.

To see that the Fisher-transformed score statistic is no smaller than the score statistic, let \( S = \text{diag}\{F(\rho_1)\}\text{diag}(\rho_1)^{-1} \) and \( R = \{\Sigma_{11}^{-1/2}(QZ_2Z_1)^T(QZ_2Z_1)\Sigma_{11}^{-1/2}/||\hat{e}||^2\}^{-1} \).

Then, \( S-I \) is non-negative definite. Since \[ 0 \leq \{F(\rho_1)-\rho_1\}^TR\{F(\rho_1)-\rho_1\} = F(\rho_1)^TRF(\rho_1) + \rho_1^TR\rho_1 - 2F(\rho_1)^TR\rho_1 = F(\rho_1)^TRF(\rho_1) - \rho_1^TR\rho_1 - \rho_1^TR(\rho-1)R\rho_1, \] it holds that \[ F(\rho_1)^TRF(\rho_1) - \rho_1^TR\rho_1 \geq 2\rho_1^TR(\rho-1)R\rho_1. \]

Consequently, it follows from the non-negative definiteness of both \( S-I \) and \( R \) that
\[ T_1^F = F(\rho_1)^TRF(\rho_1) \geq \rho_1^TR\rho_1 = T_1. \]

Therefore, the test with Fisher-transformed score statistic \( T_1^F \) is more powerful than the test with the original score statistic \( T_1 \), because the same asymptotic null distribution is used for both tests.
2. Simulation studies

In this section, finite sample performance of the proposed Fisher-transformed score test is examined through simulation studies mimicking GWASs under case-control design (WTCCC, 2007). In typical GWASs, a million genetic variants such as SNPs are tested one at a time using a stringent \( p \)-value threshold, e.g. \( 5 \times 10^{-8} \) (Risch and Merikangas, 1996). Currently, the number of tested variants is increasing (e.g. whole-genome sequencing studies). Since testing a large number of variants is quite time consuming, computationally efficient methods are favorable (Lee et al., 2013). It is also often the case that there are covariates to be adjusted for, such as age or/and sex. The score tests provide an adjustment for covariates in a computationally efficient way compared with the likelihood ratio and Wald tests.

The following experiments consider the minor allele count of a SNP, \( x_1 \in \{0,1,2\} \), as a tested variable, for which the Hardy–Weinberg equilibrium (HWE) is assumed in a general population. Let the minor allele frequency (MAF) of \( x_1 \) be denoted by \( P \in (0,0.5] \). Therefore, \( x_1 \) follows a Binomial\((P,2)\)-distribution independently in the general population. The following considers two cases of \( P = 0.2 \) and 0.4. The experiments also consider a covariate \( x_2 \) to be adjusted for, which follows a Bernoulli distribution with parameter 0.5 independently of \( x_1 \). Development of disease in a general population is modeled by a logistic regression model with additive effect of \( x_1 \): \( P(y=1 \mid x_1, x_2) = 1/\{1 + \exp(-\beta_0 - \beta_1 x_1 - \beta_2 x_2)\} \) with regression coefficients \( \beta = (\beta_0, \beta_1, \beta_2)^T \), where the intercept \( \beta_0 \) is set so that the prevalence, \( P(y=1) \), equals to 0.0005 (i.e. rare disease) while \( \exp(\beta_2) = 3 \) is considered. The null hypothesis is \( H_0 : \beta_1 = 0 \) (marginal association test in SNP-GWAS). Given the above logistic regression model in a general population, \( N/N \) case/control samples are generated as in Ueki (2014), for \( N \in \{25,50,100,250\} \). Likelihood ratio and Wald tests are set as additional competitors.

First, behavior of the statistics, \( t^F_1 \) and \( t_1 \), under \( H_0 \) (i.e. \( \exp(\beta_1) = 1 \)) is examined through \( 10^5 \) replications, where a two-sided test based on the \( \chi^2_1 \)-distribution is considered. Test statistics for Wald and likelihood ratio tests are also evaluated. Quantile-quantile plots are given in the left-most panels in Figures 1–4, and empirical type I error rates at nominal levels 0.05, 0.005 and 0.0005 are given in Table 1. In general, the larger the sample size is, the more the \( \chi^2_1 \)-approximation is accurate for all tests. The type I error rates of the Fisher-transformed score test are well controlled for \( N \geq 50 \), while slight inflation is observed when \( N = 25 \). The likelihood ratio test also shows inflation though it is modest. The score test and Wald test behave conservatively. Second, power is evaluated through \( 10^3 \) replicates at each value of odds ratio, \( \exp(\beta_1) \), increased from 1.1, in which a two-sided test based on the \( \chi^2_1 \)-distribution is used. The results for two nominal levels, \( 5 \times 10^{-6} \) and \( 5 \times 10^{-8} \), are respectively given in right two panels in Figures 1–4. It can be seen that the Fisher-transformed score test improves on the power of the standard score test, and shows comparable with or sometimes higher power than the likelihood ratio test. All tests exhibit higher power than the Wald test.

Next, an interaction effect is incorporated in the logistic regression model as \( P(y=1 \mid x_1, x_2) = 1/\{1 + \exp(-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_{12} x_1 x_2)\} \). Regarding \( x_1 \) as a binary environmental variable, a two degrees of freedom (2df) gene-environment interaction test proposed by Kraft et al. (2007) was considered. Specifically, the null hypothesis is \( H_0 : \beta_1 = \beta_{12} = 0 \), i.e. test for multiple parameters. Throughout simulations, \( \exp(\beta_1) = 1 \) and \( \exp(\beta_2) = 3 \) are used. In this simulation, MAF of \( x_1 \), \( P = 0.2 \) and 0.4 were considered, while the \( x_2 \) is the same as above. Given the above logistic regression model in a general population, \( N/N \) case/control
Fig. 1: Simulations for marginal association testing in SNP-GWAS for 25/25 case/control samples. Top panels, MAF $P = 0.2$; Bottom panels, $P = 0.4$. Left panels, quantile-quantile plots of $p$-values of the Fisher-transformed score statistic (FT-Score), the score statistic (Score), Wald test statistic (Wald) and likelihood ratio test statistic (LRT); Power at nominal levels $5 \times 10^{-6}$ (middle panels) and $5 \times 10^{-8}$ (right panels).
Fig. 2: Simulations for marginal association testing in SNP-GWAS for 50/50 case/control samples. Top panels, MAF $P = 0.2$; Bottom panels, $P = 0.4$. Left panels, quantile-quantile plots of $p$-values of the Fisher-transformed score statistic (FT-Score), the score statistic (Score), Wald test statistic (Wald) and likelihood ratio test statistic (LRT); Power at nominal levels $5 \times 10^{-6}$ (middle panels) and $5 \times 10^{-8}$ (right panels).
Fig. 3: Simulations for marginal association testing in SNP-GWAS for 100/100 case/control samples. Top panels, MAF $P = 0.2$; Bottom panels, $P = 0.4$. Left panels, quantile-quantile plots of $p$-values of the Fisher-transformed score statistic (FT-Score), the score statistic (Score), Wald test statistic (Wald) and likelihood ratio test statistic (LRT); Power at nominal levels $5 \times 10^{-6}$ (middle panels) and $5 \times 10^{-8}$ (right panels).
Fig. 4: Simulations for marginal association testing in SNP-GWAS for 250/250 case/control samples. Top panels, MAF $P = 0.2$; Bottom panels, $P = 0.4$. Left panels, quantile-quantile plots of $p$-values of the Fisher-transformed score statistic (FT-Score), the score statistic (Score), Wald test statistic (Wald) and likelihood ratio test statistic (LRT); Power at nominal levels $5 \times 10^{-6}$ (middle panels) and $5 \times 10^{-8}$ (right panels).
UEKI

Table 1: Type I error rates from simulations for marginal association testing in SNP-GWAS. $N$, number of case or control samples; $P$, MAF; Nominal, nominal level; FT-Score, Fisher-transformed score test; Score, score test; Wald, Wald test; LRT, likelihood ratio test.

| $N$ | $P$ | Nominal | FT-Score | Score | Wald | LRT |
|-----|-----|---------|----------|-------|------|-----|
| 25  | 0.2 | 0.05    | 0.060    | 0.054 | 0.044| 0.059|
|     |     | 0.005   | 0.0076   | 0.0048| 0.0016| 0.0070|
|     |     | 0.0005  | 0.00092  | 0.00033| 0.00000| 0.00073|
|     | 0.4 | 0.05    | 0.060    | 0.054 | 0.045| 0.058|
|     |     | 0.005   | 0.0079   | 0.00501| 0.0020| 0.0065|
|     |     | 0.0005  | 0.00098  | 0.00044| 0.00002| 0.00076|
| 50  | 0.2 | 0.05    | 0.055    | 0.051 | 0.047| 0.054|
|     |     | 0.005   | 0.0062   | 0.0050 | 0.0033| 0.0059|
|     |     | 0.0005  | 0.00064  | 0.00033| 0.00014| 0.00054|
|     | 0.4 | 0.05    | 0.055    | 0.052 | 0.048| 0.054|
|     |     | 0.005   | 0.0060   | 0.0046 | 0.0034| 0.0054|
|     |     | 0.0005  | 0.00069  | 0.00041| 0.00021| 0.00057|
| 100 | 0.2 | 0.05    | 0.053    | 0.051 | 0.049| 0.053|
|     |     | 0.005   | 0.0055   | 0.0049 | 0.0038| 0.0053|
|     |     | 0.0005  | 0.00048  | 0.00032| 0.00019| 0.00046|
|     | 0.4 | 0.05    | 0.053    | 0.052 | 0.050| 0.053|
|     |     | 0.005   | 0.0053   | 0.0047 | 0.0041| 0.0051|
|     |     | 0.0005  | 0.00054  | 0.00044| 0.00030| 0.00050|
| 250 | 0.2 | 0.05    | 0.050    | 0.050 | 0.049| 0.050|
|     |     | 0.005   | 0.0054   | 0.00517| 0.0049| 0.0054|
|     |     | 0.0005  | 0.00060  | 0.00052| 0.00041| 0.00059|
|     | 0.4 | 0.05    | 0.051    | 0.051 | 0.050| 0.051|
|     |     | 0.005   | 0.0051   | 0.0049 | 0.0046| 0.0050|
|     |     | 0.0005  | 0.00050  | 0.00047| 0.00039| 0.00050|

samples are generated as in Ueki (2014), for $N \in \{25, 50, 100, 250\}$.

To check the type I error rate, behavior of the statistics, $T_1^F$ and $T_1$, under $H_0$ is examined through $10^5$ replications. Wald and likelihood ratio test statistics are also examined. Quantile-quantile plots are given in the left-most panels in Figures 5–8, and empirical type I error rates at nominal levels $0.05$, $0.005$ and $0.0005$ are given in Table 2. As in the previous simulation, the larger the sample size is, the more the $\chi^2$-approximation is accurate for all tests. The type I error rates of the Fisher-transformed score test and likelihood ratio test are well controlled for $N \geq 50$, while slight inflation is observed when $N = 25$. The score test appears to be slightly conservative, and the Wald test appears to be more conservative. By increasing $\exp(\beta_{12})$ from 1.1, power is evaluated through $10^3$ replicates. The results for two nominal levels, $5 \times 10^{-6}$ and $5 \times 10^{-8}$, are respectively given in the right two panels in Figures 5–8. It can be seen that the Fisher-transformed score test improves the power of the standard score test, and also shows comparable with or sometimes higher power than the likelihood ratio test. All tests exhibit higher power than the Wald test as in the marginal association test simulations.
Fig. 5: Simulations for gene-environment 2df test of Kraft et al. (2007) for 25/25 case/control samples. Top panels, MAF $P = 0.2$; Bottom panels, $P = 0.4$. Left panels, quantile-quantile plots of $p$-values of the Fisher-transformed score statistic (FT-Score), the score statistic (Score), Wald test statistic (Wald) and likelihood ratio test statistic (LRT); Power at nominal levels $5 \times 10^{-6}$ (middle panels) and $5 \times 10^{-8}$ (right panels).
Fig. 6: Simulations for gene-environment 2df test of Kraft et al. (2007) for 50/50 case/control samples. Top panels, MAF $P = 0.2$; Bottom panels, $P = 0.4$. Left panels, quantile-quantile plots of $p$-values of the Fisher-transformed score statistic (FT-Score), the score statistic (Score), Wald test statistic (Wald) and likelihood ratio test statistic (LRT); Power at nominal levels $5 \times 10^{-6}$ (middle panels) and $5 \times 10^{-8}$ (right panels).
Fig. 7: Simulations for gene-environment 2df test of Kraft et al. (2007) for 100/100 case/control samples. Top panels, MAF $P = 0.2$; Bottom panels, $P = 0.4$. Left panels, quantile-quantile plots of p-values of the Fisher-transformed score statistic (FT-Score), the score statistic (Score), Wald test statistic (Wald) and likelihood ratio test statistic (LRT); Power at nominal levels $5 \times 10^{-6}$ (middle panels) and $5 \times 10^{-8}$ (right panels).
Fig. 8: Simulations for gene-environment 2df test of Kraft et al. (2007) for 250/250 case/control samples. Top panels, MAF $P = 0.2$; Bottom panels, $P = 0.4$. Left panels, quantile-quantile plots of $p$-values of the Fisher-transformed score statistic (FT-Score), the score statistic (Score), Wald test statistic (Wald) and likelihood ratio test statistic (LRT); Power at nominal levels $5 \times 10^{-6}$ (middle panels) and $5 \times 10^{-8}$ (right panels).
Table 2: Type I error rates from simulations for gene-environment 2df test of Kraft et al. (2007). $N$, number of case or control samples; $P$, MAF; Nominal, nominal level; FT-Score, Fisher-transformed score test; Score, score test; Wald, Wald test; LRT, likelihood ratio test.

| $N$ | $P$ | Nominal | FT-Score | Score | Wald   | LRT   |
|-----|-----|---------|----------|-------|--------|-------|
| 25  | 0.2 | 0.05    | 0.060    | 0.051 | 0.016  | 0.074 |
|     |     | 0.005   | 0.0069   | 0.0036 | 0.00019 | 0.0088 |
|     |     | 0.0005  | 0.00078  | 0.00018 | 0.00000 | 0.00090 |
|     | 0.4 | 0.05    | 0.064    | 0.053 | 0.021  | 0.071 |
|     |     | 0.005   | 0.0070   | 0.0035 | 0.00021 | 0.0098 |
|     |     | 0.0005  | 0.00094  | 0.00016 | 0.00000 | 0.0013 |
| 50  | 0.2 | 0.05    | 0.055    | 0.051 | 0.047  | 0.054 |
|     |     | 0.005   | 0.0062   | 0.0050 | 0.0033  | 0.0060 |
|     |     | 0.0005  | 0.00064  | 0.00033 | 0.00014 | 0.00054 |
|     | 0.4 | 0.05    | 0.055    | 0.052 | 0.048  | 0.054 |
|     |     | 0.005   | 0.0060   | 0.0046 | 0.0034  | 0.0054 |
|     |     | 0.0005  | 0.00069  | 0.00041 | 0.00021 | 0.00057 |
| 100 | 0.2 | 0.05    | 0.052    | 0.050 | 0.041  | 0.054 |
|     |     | 0.005   | 0.0056   | 0.0048 | 0.0029  | 0.0063 |
|     |     | 0.0005  | 0.00062  | 0.00050 | 0.00018 | 0.00072 |
|     | 0.4 | 0.05    | 0.053    | 0.051 | 0.045  | 0.054 |
|     |     | 0.005   | 0.0053   | 0.0046 | 0.0030  | 0.0055 |
|     |     | 0.0005  | 0.00055  | 0.00038 | 0.00015 | 0.00057 |
| 250 | 0.2 | 0.05    | 0.051    | 0.050 | 0.046  | 0.052 |
|     |     | 0.005   | 0.0051   | 0.0048 | 0.0038  | 0.0052 |
|     |     | 0.0005  | 0.00044  | 0.00036 | 0.00025 | 0.00043 |
|     | 0.4 | 0.05    | 0.051    | 0.050 | 0.048  | 0.051 |
|     |     | 0.005   | 0.0050   | 0.0047 | 0.0040  | 0.0050 |
|     |     | 0.0005  | 0.00050  | 0.00045 | 0.00030 | 0.00058 |

3. Application to real GWAS data

This section provides an illustration through real data taken from Tanaka et al. (2009) who report a genome-wide association to null virological response (NVR) in the recommended treatment for patients with hepatitis C virus (HCV) genotype 1 within a Japanese population, pegylated interferon-alpha (PEG-IFN-alpha) plus ribavirin (RBV), compared with patients with sustained virologic response (SVR). The authors found two SNPs near the gene IL28B on chromosome 19 to be strongly associated with NVR. The authors replicated these associations in an independent cohort. One of the reported SNPs, rs8099917, is used for illustration. The genotype frequency is given in Table 3. Using additive coding for the minor allele count instead of the dominant coding used in Tanaka et al. (2009), Table 3 provides $p$-values from Fisher-transformed score, likelihood ratio and Wald tests. As observed in simulation studies, the Fisher-transformation increases the significance of the standard score test. In this application, the Fisher-transformed score test gave the highest significance among the other tests.
Table 3: GWAS data at SNP rs8099917 (T/G alleles) taken from Tanaka et al. (2009) (Supplementary Table 2). NVR, TVR and SVR represent null virologic response, transient virologic response and sustained virologic response, respectively. GWAS, discovery samples; replication, replication samples.

| Data         | NVR+TVR | SVR     | FT-Score | Score  | LRT     | Wald   |
|--------------|---------|---------|----------|--------|---------|--------|
| GWAS         | TT/TG/GG| TT/TG/GG|          |        |         |        |
|              | 30/58/3 | 47/4/0  | 9.1 x 10^{-14} | 3.8 x 10^{-11} | 3.7 x 10^{-13} | 3.2 x 10^{-8} |
| replication  | 41/40/2 | 78/11/0 | 7.6 x 10^{-9}  | 5.7 x 10^{-8}  | 1.9 x 10^{-8}  | 4.2 x 10^{-7}  |

4. Concluding remarks

This article presents a simple method to enhance power of the score tests for regression models via Fisher-transformation. Simulation studies show that its power is comparable with or sometimes higher power than the likelihood ratio test. The likelihood ratio test is computationally expensive due to requirement of estimation under the alternative. The Wald test is less computationally expensive but Xing et al. (2012) reported that it exhibits a paradoxical behavior and results in loss of power. Lower power of the Wald test is also observed in the simulation studies in this article. The score test is fast in computation due to the closed-form expression except for the null estimation, and its improved power by the proposed method without additional computational cost would be valuable, e.g. in the situation where a large number of tests are required as in GWASs. Because $||\hat{\ell}'||^2 = - \sum_{i=1}^n \hat{r}_i ^{-1} \hat{r}_i ^2$ tends to be small when $n$ is small, the improvement appeared to be pronounced when sample sizes get smaller. However, simultaneously, the asymptotic normality of the score statistic may be violated when sample size is too small, e.g. when $N = 25$. Therefore, the proposed method could be particularly useful when the sample size is large enough to make the asymptotic approximation valid. For instance, simulation studies mimicking GWAS under balanced case-control design suggest that at least a hundred in total is necessary. Notably, however, it is unclear whether the result of needed sample size is generalizable to other scenarios such as unbalanced design. Finally, one of the reviewers pointed out an interesting property of the Fisher-transformation in improving asymptotic normality (Konishi, 1981; Taniguchi et al., 1989) and questioned how it relates to the proposed test. This point is worth investigating further in future work.

Acknowledgements

The author is grateful to two anonymous reviewers for providing insightful comments and providing directions for additional work which has resulted in this article. This work was supported by JSPS KAKENHI Grant Numbers JP16K00064, JP16K08638, JP15H01678.

REFERENCES

Agresti, A. (2002). Categorical Data Analysis, Second Edition. Wiley, New Jersey.
Dobson, A. (2002). An Introduction to Generalized Linear Models, Second Edition. Chapman & Hall, London.
Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. Biometrika 10, 507–521.
Fisher, R. A. (1921). On the ‘probable error’ of a coefficient of correlation deduced from a small sample. Metron 1, 3–32.
Fisher-transformed Score Test

Konishi, S. (1981). Normalizing transformations of some statistics in multivariate analysis. *Biometrika* **68**, 647–651.

Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J. and Gauderman, W. J. (2007). Exploiting gene-environment interaction to detect genetic associations. *Human Heredity* **63**, 111–119.

Lange, K. and Sinsheimer, J. S. (1993). Normal/Independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics* **2**, 175–198.

Lee, S. M., Karrison, T. G., Cox, N. J. and Im, H. K. (2013). Quantitative allelic test – a fast test for very large association studies. *Genetic Epidemiology* **37**, 831–839.

McCullagh, P. and Nelder, J. A., (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall, London.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society* **44**, 50–57.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517.

Tanaka, Y., Nishida, N., Sugiyama, M., Kurosaki, M., Matsuura, K., Sakamoto, N., Nakagawa, M., Korenaga, M., Hino, K., Hige, S., Ito, Y., Mita, E., Tanaka, E., Mochida, S., Murawaki, Y., Honda, M., Sakai, A., Hiasa, Y., Nishiguchi, S., Koike, A., Sakaida, I., Imamura, M., Ito, K., Yano, K., Masaki, N., Sugauchi, F., Izumi, N., Tokunaga, K. and Mizokami, M. (2009). Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C. *Nature Genetics* **41**, 1105–1109.

Taniguchi, M., Krishnaiah, P. R. and Chao, R. (1989). Normalizing transformations of some statistics of Gaussian ARMA processes. *Annals of the Institute of Statistical Mathematics* **41**, 187–197.

Ueki, M. (2014). On the choice of degrees of freedom for testing gene-gene interactions. *Statistics in Medicine* **33**, 4934–4948.

Ueki, M. and Kawasaki, Y. (2013). Multiple choice from competing regression models under multicollinearity based on standardized update. *Computational Statistics & Data Analysis* **63**, 31–41.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, New York.

Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678.

Wellek, S. and Ziegler, A. (2009). A genotype-based approach to assessing the association between single nucleotide polymorphisms. *Human Heredity* **67**, 128–139.

Xing, G., Lin, C. Y., Wooding, S. P. and Xing, C. (2012). Blindly using Wald’s test can miss rare disease-causal variants in case-control association studies. *Annals of Human Genetics* **76**, 168–177.

(Received: February 3, 2017, Accepted: May 22, 2017)
