On 'Reverse' Regression for Robust Genetic Association Studies and Allele Frequency Estimation with Related Individuals

Lin Zhang[1] and Lei Sun[1,2]

[1]Department of Statistical Sciences, University of Toronto, 100 St. George Street, Toronto, Ontario M5S 3G3, Canada, linzhang@utstat.toronto.edu

[2]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, 155 College Street, Toronto, Ontario M5T 3M7, Canada, sun@utstat.toronto.edu

November 14, 2018

Abstract

Linear mixed-effect models are widely used for conducting family-based genetic association tests. In this note we first show that, contrary to popular belief, this standard approach is sensitive to departure from Hardy–Weinberg equilibrium at the causal SNP(s). We then consider a robust 'reverse' regression model with genotype as the response variable and phenotype as a predictor, and we provide the corresponding score test statistic. Desirably, for the simple case of Hardy–Weinberg equilibrium with or without covariates, the proposed test unifies a number of existing ones, including the generalized quasi-likelihood score test. The flexibility of the proposed framework is further demonstrated by its inherent ability to provide the best linear unbiased estimator for population allele frequency in a sample of related individuals, while adjusting for covariate effects and Hardy–Weinberg disequilibrium.

Keywords: Association tests; Dependent Samples; Hardy–Weinberg equilibrium; Heritability; Quasi-likelihood score tests; Robustness.

1 Introduction

Genotype-based genetic association tests treat a complex trait as the response variable and the genotype of a SNP as a predictor. When individuals in a sample are genetically related with each other, linear mixed-effect models have become the most popular approach for association testing (Eu-Ahsunthornwattana et al. (2014)). The variance-covariance matrix of the phenotype is partitioned into a weighted sum of correlation structure due to genetic relatedness and shared environmental effects, where the weight is usually referred to as 'heritability' (Visscher et al. (2008)). The genetic relatedness is typically represented by the kinship coefficient matrix, or estimated based on the available genome-wide genetic data if the pedigree information was not collected (Sun and Dimitromanolakis (2012); Yang et al. (2011)).

Allele-based genetic association tests or allelic tests are locally most powerful (Sasieni (1997)), but require Hardy–Weinberg equilibrium to hold in the study population. Hardy–Weinberg equilibrium states that the two alleles in a genotype are independent draws from the same distribution, or, equivalently, genotype frequencies depend solely on the allele frequencies. For a bi-allelic SNP with two possible alleles A and a, let p and $1-p$ be the respective allele frequencies. Under Hardy–Weinberg equilibrium, $p_{aa} = (1-p)^2$, $p_{Aa} = 2p(1-p)$, and $p_{AA} = p^2$, where p_{aa}, p_{Aa} and p_{AA} are the genotype frequencies of genotypes aa, Aa and AA, respectively. To measure the departure from Hardy–Weinberg equilibrium or the amount of Hardy–Weinberg disequilibrium,

$$\delta = p_{AA} - p^2 \tag{1}$$

is a widely used quantity, and $\delta = 0$ corresponds to Hardy–Weinberg equilibrium (Weir (1996)).

It is commonly assumed that genotype-based association tests are robust to departure from Hardy–Weinberg equilibrium. With a sample of independent individuals, both theoretical and empirical results support this (Sasieni (1997); Schaid and Jacobsen (1999)). However, in the presence of sample dependency, little has been discussed. We demonstrate with a simple sib-pair design that

i) when the true heritability is known, the empirical type 1 error rate of the standard linear mixed-effect model using kinship is inflated when $\delta > 0$ (or deflated if $\delta < 0$), and ii) when heritability is estimated from the data, the empirical type 1 error rate is well-controlled, but the estimated heritability is then biased. Hence, current genotype-based association tests with related individuals are not robust to Hardy–Weinberg disequilibrium.

One consideration is to reverse the roles of Y and G in the regression model. MultiPhen (O'Reilly et al. (2012)) proposed to treat the genotype G of a SNP as the response variable, and phenotype values Ys of multiple traits as predictors. However, MultiPhen relies on ordinal logistic regression and analyzes only independent samples. MultiPhen does not require the assumption of Hardy–Weinberg equilibrium, but insights to its robustness was not provided (O'Reilly et al. (2012)).

Thornton and McPeek (2007) focused on case-control studies of binary traits with related individuals. Their test was then generalized by Feng et al. (2011) and Feng (2014) to a quasi-likelihood score test for either binary or continuous traits. However, none of these methods can directly incorporate covariates. Jakobsdottir and McPeek (2013) later proposed a 'retrospective' approach, MASTOR, to study the association between G and one (approximately) normally distributed trait Y, while accommodating covariates in related individuals. All methods in this category, however, implicitly assumed Hardy–Weinberg equilibrium.

In this note, we propose a robust and flexible 'reverse' regression framework with G as the (numerical) response variable and Y as a predictor in related individuals. The proposed 'reverse' regression association test can accommodate multiple continuous or categorical phenotypes and adjust for covariate effects. It explicitly includes δ as a correction factor in the variance-covariance matrix to adjust for potential departure from Hardy–Weinberg equilibrium. Further, we show that the proposed framework implicitly unifies previous methods. In the simple cases of an independent sample, no covariates, or Hardy–Weinberg equilibrium, the proposed approach yields similar results as compared to the aforementioned tests. Lastly, we show that the proposed 'reverse' re-

gression framework can also be used to estimate allele frequency in complex pedigrees, while adjusting for covariate effects and potential departure from Hardy–Weinberg equilibrium. Interestingly, we reveal that for the simple case of no covariates and Hardy–Weinberg equilibrium, the proposed estimator is the best linear unbiased estimator of McPeek et al. (2004).

2 Genotype-based Association Test

2.1 Independent Samples

Consider a bi-allelic SNP with two alleles a and A. Define $G = 0, 1$ and 2 for genotypes aa, Aa and AA, respectively, and let Y be an (approximately) normally distributed trait. With a sample of n unrelated individuals, the traditional genotype-based association test assumes that,

$$y = \alpha^* 1 + \beta^* g + \gamma^* z + \varepsilon^*, \quad \varepsilon^* \sim N(0, \sigma^{*2} I), \tag{2}$$

where $y = (y_1, y_2, \cdots, y_n)$ is a $n \times 1$ vector for the phenotypic values, $g = (g_1, g_2, \cdots, g_n)$ is a $n \times 1$ vector for the genotypes of a SNP, ε^* is the error term with variance σ^{*2}, 1 is a $n \times 1$ vector of 1s, and I is the identity matrix. For notation simplicity but without loss of generality, we assume that there is one covariate denoted by z.

To evaluation the phenotype-genotype association, score tests are often used (Derkach et al. (2015)). In this case, the score statistic of testing $H_0 : \beta^* = 0$ can be easily derived as

$$T_{\text{indep}} = n \cdot \frac{\{(g-\bar{g}1)^T(y-\bar{y}1) - \frac{(g-\bar{g}1)^T(z-\bar{z}1)(y-\bar{y}1)^T(z-\bar{z}1)}{(z-\bar{z}1)^T(z-\bar{z}1)}\}^2}{[(g-\bar{g}1)^T(g-\bar{g}1) - \frac{\{(g-\bar{g}1)^T(z-\bar{z}1)\}^2}{(z-\bar{z}1)^T(z-\bar{z}1)}][(y-\bar{y}1)^T(y-\bar{y}1) - \frac{\{(y-\bar{y}1)^T(z-\bar{z}1)\}^2}{(z-\bar{z}1)^T(z-\bar{z}1)}]}. \tag{3}$$

After some simple algebraic manipulations, one can show that

$$\frac{1}{n}(g-\bar{g}1)^T(g-\bar{g}1) = \widehat{\text{var}}(G) = 2(\hat{p}(1-\hat{p}) + \hat{p}_{AA} - \hat{p}^2) = 2(\hat{p}(1-\hat{p}) + \hat{\delta}),$$

where $\hat{p} = (2n_{AA} + n_{Aa})/(2n)$ and $\hat{p}_{AA} = n_{AA}/n$ are, respectively, sample estimates of allele frequency of A and genotype frequency of AA. Because $\hat{\delta} = \hat{p}_{AA} - \hat{p}^2$ measures the amount of Hardy–Weinberg disequilibrium present in the data (Weir (1996)), T_{indep} inherently adjusts for departure from Hardy–Weinberg equilibrium through $\widehat{\text{var}}(G) = 2(\hat{p}(1-\hat{p}) + \hat{\delta})$. As a result, the traditional genotype-based association test is robust to Hardy–Weinberg disequilibrium in independent samples.

2.2 Dependent Samples

When individuals in a sample are related to each other as in pedigree data, a common practice is to replace $\text{var}(\varepsilon^*) = \sigma^{*2} I$ in (2) with $\sigma_y^2 \Sigma_y$ to reflect the sample dependence, and use the following linear mixed-effect model,

$$y = \alpha^* 1 + \beta^* g + \gamma^* z + \varepsilon^*, \quad \varepsilon^* \sim N(0, \sigma_y^2 \Sigma_y), \quad \Sigma_y = h^2 \Sigma_\Phi + (1-h^2) I, \tag{4}$$

where h^2 is the 'heritability', and Σ_Φ is the kinship coefficient matrix. Among the total variance of Y, σ_y^2, $\sigma_a^2 = h^2 \sigma_y^2$ can be interpreted as the variance of Y due to additive genetic variation, while $\sigma_e^2 = (1-h^2)\sigma_y^2$ as the variance due to environmental variation.

For the Σ_Φ matrix, $\Sigma_\Phi(i,j) = 2\phi_{i,j}$ where $\phi_{i,j}$ is the kinship coefficient between individual i and individual j. When the pedigree information is unknown, it is a common practice to estimate $\Sigma_\Phi(i,j)$ with the averaged sample correlation across K autosomal SNPs (Yang et al. (2011)),

$$\hat{\Sigma}_\Phi(i,j) = \frac{1}{K} \sum_{k=1}^{K} \frac{(g_{ik} - 2\hat{p}_k)(g_{jk} - 2\hat{p}_k)}{2\hat{p}_k(1-\hat{p}_k)}, \tag{5}$$

where g_{ik} and g_{jk} are the genotypes of SNP k for individual i and j respectively, and $\hat{p}_k$ is the estimated allele frequency. Without the assumption of Hardy–Weinberg equilibrium, $\text{var}(G_k) = 2(p_k(1-p_k) + \delta_k)$, so (5) implies $\delta_k = 0$. Thus, the traditional kinship coefficient estimation

implicitly assumes that Hardy–Weinberg equilibrium holds at each and every SNP k, $k = 1, \cdots, K$. Consequently, the linear mixed model (4) can be sensitive to departure from Hardy–Weinberg equilibrium.

In the presence of Hardy–Weinberg disequilibrium, we now demonstrate with a simple sib-pair design that the type 1 error rate of linear mixed-effect model (4) inflates if $\delta > 0$, or deflates if $\delta < 0$, assuming h^2 is known (for example, based on previous heritability studies). Consider a continuous trait Y with $h^2 = 0.6$ and influenced by one causal SNP, G_{causal}, for which the Hardy–Weinberg disequilibrium factor is δ_{causal}. We conducted association testing between Y and a non-associated SNP, G_{tested} (with its own δ_{tested}) using simulated data from 5,000 sib-pairs. Figure 1(a) plots the empirical type 1 error rate of the test against δ_{causal} for a nominal level of 0.05. The trend of type 1 error inflation is clear as δ_{causal} increases. Here we set $\delta_{\text{tested}} = 0.08$, but we note that the root cause of the type 1 error issue is δ_{causal} when using (4). That is, even if the tested SNP is in Hardy–Weinberg equilibrium ($\delta_{\text{tested}} = 0$), the problem remains, albeit less severe, as long as $\delta_{\text{causal}} \neq 0$.

In many practical implementations of the linear mixed-effect model (4), h^2 is often treated as a nuisance parameter, and no type 1 error issue has been reported. However, in this situation, the impact of Hardy–Weinberg disequilibrium is now on the estimation of h^2. Specifically, we applied the linear mixed-effect model of (4) to the same simulated data as above but assumed h^2 as unknown. Figure 1(b) clearly shows that $\hat{h}^2$ is downward biased when $\delta_{\text{causal}} < 0$, and upward biased if $\delta_{\text{causal}} > 0$. The bias can be substantial. For example, when $\delta_{\text{causal}} = 0.10$, the estimated heritability $\hat{h}^2$ is centred at 0.97 as compared to the true value of 0.6.

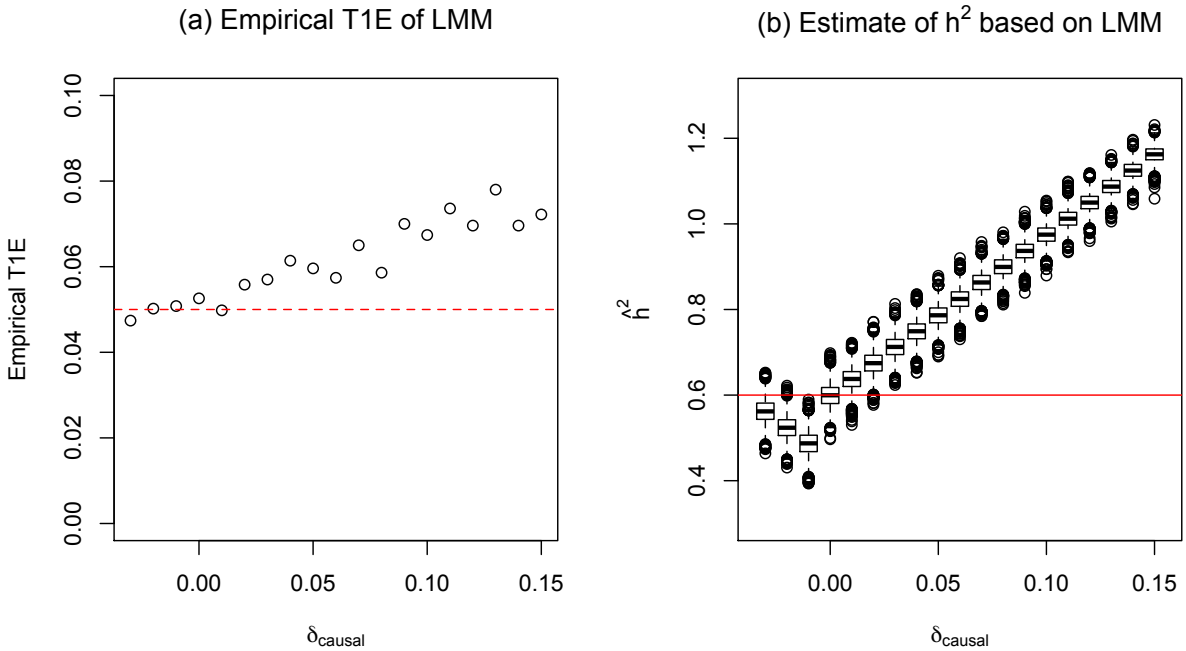

Figure 1: (a): **Empirical type 1 error rate of the linear mixed-effect model (4) against δ_{causal}, assuming the true heritability h^2 is known**; (b): **Box-plots of $\hat{h}^2$, estimated from the linear mixed-effect model (4) against δ_{causal}**. $5,000$ sib-pairs, $h^2 = 0.6$, $p = 0.2$, with $5,000$ simulated replications for each δ_{causal} value.

3 The Proposed 'Reverse' Regression Model and Association Test, T_{reverse}

To account for the potential departure from Hardy–Weinberg equilibrium, one useful direction is to reverse the roles of Y and G in the regression model. We define the 'reverse' regression model as,

$$g = \alpha 1 + \beta y + \gamma z + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 \Sigma_g), \quad \sigma^2 \Sigma_g = \sigma^2 \Sigma_\Phi + \Sigma_\delta, \tag{6}$$

where Σ_Φ is the kinship coefficient matrix as defined earlier, and Σ_δ is a function of δ, that explicitly adjusts for Hardy–Weinberg disequilibrium.

The response variable G, in the proposed 'reverse' regression (6), is a discrete random variable with three possible values 0, 1 and 2. Chen (1983) showed that under common regularity conditions, the score test statistic takes identical form for exponential family with independent samples and no covariates. In light of this result, we modelled G using a linear model. In section 4, we show that the score statistic derived from (6) shares similar properties to previous methods that treat G as discrete, in the simple case of Hardy–Weinberg equilibrium with or without covariates.

For a pair of siblings,

$$\sigma^2 \Sigma_g = \sigma^2 \Sigma_\Phi + \Sigma_\delta = \sigma^2 \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 0.5\delta \\ 0.5\delta & 0 \end{pmatrix}. \tag{7}$$

Under the null hypothesis of $H_0 : \beta = 0$, the corresponding score test statistic is

$$\begin{gathered} T_{\text{reverse}} = n \cdot \frac{(V_{gy}V_{zz} - V_{gz}V_{yz})^2}{(V_{gg}V_{zz} - V_{gz}^2)(V_{yy}V_{zz} - V_{yz}^2)}, \\ \text{where } V_{uv} = u^T \hat{\Sigma}_g^{-1} v - u^T \hat{\Sigma}_g^{-1} 1 (1^T \hat{\Sigma}_g^{-1} 1)^{-1} 1^T \hat{\Sigma}_g^{-1} v, \text{ for } u, v \in \{g, z, y\}, \end{gathered} \tag{8}$$

and T_{reverse} is asymptotically χ_1^2 distributed under the null of no association.

In the absence of sample correlation, $\Sigma_g = I$, and the proposed test statistic in (8) is then simplified to T_{indep} in (3); T_{reverse} is symmetric with respect to G and Y, provided the same Σ is used. Thus, for independent samples, T_{reverse} adjusts for Hardy–Weinberg disequilibrium through $\text{var}(G) = 2(p(1-p)+\delta)$ as in the traditional association test. However, differences emerge in the presence of sample correlation. In that case, the traditional linear mixed-effect model uses the averaged kinship coefficient matrix Σ_Φ to model Σ_y, while the proposed method adjusts for sample correlation through Σ_g that includes both Σ_Φ for relatedness and a correction factor δ for potential Hardy–Weinberg disequilibrium at the tested SNP.

4 Connections with Quasi-likelihood Association Tests without Covariates, $T_{\text{QL-allelic}}$ and T_{GQLS}

For case-control studies with related samples, Thornton and McPeek (2007) proposed an association test that compares the sample allele frequency estimates between the case and control groups, while adjusting for relatedness between individuals. The test statistic is defined as,

$$\begin{aligned} T_{\text{QL-allelic}} &= \frac{(\hat{p}_r - \hat{p}_c)^2}{\hat{\text{var}}(\hat{p}_r - \hat{p}_c)}, \quad \text{where} \quad \hat{p}_r = \frac{1_r^T \Sigma_\Phi^{-1} x}{1_r^T \Sigma_\Phi^{-1} 1}, \quad \hat{p}_c = \frac{1^T \Sigma_\Phi^{-1} x}{1^T \Sigma_\Phi^{-1} 1}, \quad x = g/2, \\ \hat{\text{var}}(\hat{p}_r - \hat{p}_c) &= \frac{1}{2}\hat{p}_c(1-\hat{p}_c)[(1_r^T \Sigma_\Phi^{-1} 1)^{-2}(1_r^T \Sigma_\Phi^{-1} 1_r) - (1^T \Sigma_\Phi^{-1} 1)^{-1}], \end{aligned} \tag{9}$$

and 1_r is a $n \times 1$ vector with the ith observation to be 1 if individual i is from the case group, and 0 otherwise. Thus, $\hat{p}_r$ is a sample estimate of the allele frequency in cases, while $\hat{p}_c$ is the pooled estimate using the case-control combined sample, both adjusting for sample correlation through Σ_Φ.

Feng et al. (2011) and Feng (2014) reformulated $T_{\text{QL-allelic}}$ as a generalized quasi-likelihood score test that handles both continuous and binary traits but does not allow for covariates. The

quasi-likelihood based regression framework assumes that,

$$E(x_i|y_i) = \mu_i, \quad \log(\frac{\mu_i}{1-\mu_i}) = \beta_0 + \beta_1 y_i, \quad \text{var}(x|y) = \Lambda^{1/2}\Sigma_\Phi\Lambda^{1/2}, \tag{10}$$

where Λ is an $n \times n$ diagonal matrix, and $\{\text{diag}\{\Lambda\}\} = \{\mu_1(1-\mu_1), \mu_2(1-\mu_2), \cdots, \mu_n(1-\mu_n)\}$.

Under the null of no association, the framework also assumes that $E(x_i) = p_c$, and $\text{var}(x) = p_c(1-p_c)\Sigma_\Phi$. Thus, the generalized quasi-score statistic of testing $H_0 : \beta_1 = 0$ is,

$$T_{\text{GQLS}} = \frac{1}{\frac{1}{2}\hat{p}_c(1-\hat{p}_c)} \cdot \frac{[(y-\bar{y}1)^T\Sigma_\Phi^{-1}(x-\hat{p}_c 1)]^2}{(y-\bar{y}1)^T\Sigma_\Phi^{-1}(y-\bar{y}1)}, \tag{11}$$

where $\bar{y} = (1^T\Sigma_\Phi^{-1}1)^{-1}1^T\Sigma_\Phi^{-1}y$, and $\hat{p}_c$ is identical to that in $T_{\text{QL-allelic}}$ of (9). When Y is binary, $y = 1_r$. Substituting y with 1_r in (11), it is easy to show that

$$T_{\text{GQLS,binary}} = T_{\text{QL-allelic}}.$$

When individuals are independent of each other in a sample, $\Sigma_\Phi = I$, then $\hat{p}_c = \hat{p}$. Thus, $\hat{p}_c(1-\hat{p}_c)$ in the denominator of (11) assumes $\delta = 0$ in the estimate of $\text{var}(G/2)$, i.e. Hardy–Weinberg equilibrium. In contrast,

$$T_{\text{reverse,indep,no covariate}} = T_{\text{GQLS, indep}} \cdot \frac{1}{1+\frac{\hat{\delta}}{\hat{p}(1-\hat{p})}}. \tag{12}$$

5 Connections with MASTOR, an Alternative Mixed-Model Association Test with Covariates, T_{MASTOR}

For continuous traits with covariates, Jakobsdottir and McPeek (2013) proposed MASTOR, a mixed-effect model-based association test. The MASTOR test statistic of no association is de-

fined as,

$$T_{\text{MASTOR}} = \frac{(x^T P_y^* y)^2}{\hat{\text{var}}(x^T P_y^* y | W, y)} = \frac{(x^T P_y^* y)^2}{\hat{\sigma}_x^2 y^T P_y^{*T} \Sigma_\Phi P_y y}, \tag{13}$$

where

$$P_y^* = V_{xy}^* - V_{xz}^* V_{yz}^* / V_{zz}^*, \ \hat{\sigma}_x^2 = x^T \{\Sigma_\Phi^{-1} - \Sigma_\Phi^{-1} 1 (1^T \Sigma_\Phi^{-1} 1)^{-1} 1^T \Sigma_\Phi^{-1}\} x/(n-1),$$
$$V_{uv}^* = u^T \hat{\Sigma}_y^{-1} v - u^T \hat{\Sigma}_y^{-1} 1 (1^T \hat{\Sigma}_y^{-1} 1)^{-1} 1^T \hat{\Sigma}_y^{-1} v, \text{ for } u, v \in \{x, z, y\}, \ x = g/2,$$

and Σ_y is the same as that in the linear mixed-effect model of (4), and $\hat{\Sigma}_y$ is the maximum likelihood estimate under the null of no association.

To compare T_{reverse} and T_{MASTOR}, we rewrite T_{reverse} in (8) as

$$T_{\text{reverse}} = \frac{(g^T P_g y)^2}{\hat{\sigma}_g^2 y^T P_g y},$$

where $P_g = V_{gy} - V_{gz} V_{yz} / V_{zz}$, and $\hat{\sigma}_g^2 = g^T P_g g / n$. Note that although V_{uv} in (8) has the same form as V_{uv}^* above, V_{uv} uses Σ_g that contains both Σ_Φ and Σ_δ as illustrated in (7) for sib-pairs. Indeed, although both T_{MASTOR} and T_{reverse} measure the correlation between G and Y while adjusting for covariate effects, the two approaches are distinct from each other. T_{reverse} is directly derived from a regression framework with G as the response variable and Y as a covariate, explicitly adjusting for Hardy–Weinberg disequilibrium through $\sigma^2 \Sigma_g = \sigma^2 \Sigma_\Phi + \Sigma_\delta$ for the tested SNP as in (6). In contrast, T_{MASTOR} in (13) is a hybrid score statistic, where the score function is derived from the linear mixed-effect model with Y as the response variable, while the variance of the score function is estimated from a 'retrospective' approach that models $X = G/2$ as the response variable. When estimating σ_x^2 and modelling Σ_y, T_{MASTOR} considers Σ_Φ alone to account for the genetic relationship between related individuals. Hence, T_{MASTOR} implicitly assumes that Hardy–Weinberg equilibrium holds in the study population.

6 Allele Frequency Estimation with the Proposed 'Reverse' Regression

Another feature of the proposed 'reverse' regression framework is its inherent ability to estimate allele frequencies in dependent samples while adjusting for covariate effects. When no phenotype Y is included in (6), the 'reverse' regression is simplified to,

$$g = \alpha 1 + \gamma Z + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 \Sigma_g), \tag{14}$$

where Z is the matrix of covariates. The intercept α can be interpreted as twice the population allele frequency, p, for the SNP of interest. The maximum likelihood estimator of α is

$$\hat{\alpha}_z = \frac{1^T \hat{\Sigma}_g^{-1} [I - Z(Z^T \hat{\Sigma}_g^{-1} Z)^{-1} Z^T \hat{\Sigma}_g^{-1}] g}{1^T \hat{\Sigma}_g^{-1} [I - Z(Z^T \hat{\Sigma}_g^{-1} Z)^{-1} Z^T \hat{\Sigma}_g^{-1}] 1}. \tag{15}$$

When no covariate is included in (14) and Σ_g is replaced with the kinship coefficient matrix Σ_Φ, $\hat{\alpha}_z$ is then reduced to $\hat{\alpha} = (1^T \Sigma_\Phi^{-1} 1)^{-1} 1^T \Sigma_\Phi^{-1} g$. Interestingly, $\hat{\alpha}/2$ is the best linear unbiased estimator of p as studied in McPeek et al. (2004), while $\hat{\alpha}_z/2$ is a generalized estimator that can adjust for covariate effects and potential departure from Hardy–Weinberg equilibrium.

Table 1: Comparison of genotype-based association tests

Methods	Mixed-effect Model	Quasi-likelihood Score Test	MASTOR	MultiPhen	Proposed 'Reverse' Regression
Dependent Sample	Yes	Yes	Yes	No	Yes
Covariates	Yes	No	Yes	Yes	Yes
Binary Phenotypes	No	Yes	No	Yes	Yes
Continuous Phenotypes	Yes	Yes	Yes	Yes	Yes
Multiple Phenotypes	No	Yes	No	Yes	Yes
Robust to Hardy–Weinberg disequilibrium	No	No	No	Yes	Yes

7 Discussion

In this note we first showed that, contrary to popular belief, the existing genetic association tests based on linear mixed-effect models are not robust to departure from Hardy–Weinberg equilibrium at the causal SNP(s). We then proposed a 'reverse' regression model with genotype G as the response variable and phenotype Y as a covariate. At a tested SNP, because the proposed 'reverse' framework is conditional on Y, the variance-covariance matrix only concerns G_{tested}, i.e. Σ_g. The modelling and estimation of Σ_g can account for potential departure from Hardy–Weinberg equilibrium (Σ_δ), in addition to genetic correlation as captured by the kinship coefficient matrix (Σ_Φ), resulting in a more robust association test for related individuals. Desirably, for the simple cases of independent samples, no covariates, or Hardy–Weinberg equilibrium, the proposed test numerically coincides with a number of existing methods. Table 1 provides a summary of the method comparisons.

Departure from Hardy–Weinberg equilibrium generally falls into two categories: disequilibrium at the causal SNP(s) and disequilibrium at the tested SNP. If Hardy–Weinberg equilibrium does not hold at the causal SNP(s), models that rely on Σ_y (e.g. the linear mixed-effect model and MASTOR) may have inflated type 1 error rate if $\delta_{\text{causal}} > 0$, or deflated type 1 error rate if $\delta_{\text{causal}} < 0$, assuming the true heritability h^2 is known; if h^2 is estimated based on the data, the estimate is then biased. If Hardy–Weinberg disequilibrium occurs at the tested SNP, models that assumes $\text{var}(g_{\text{tested}}) = \sigma_g^2 \Sigma_\Phi$ (e.g. quasi-likelihood score test and MASTOR) may be sensitive to $\delta_{\text{tested}} \neq 0$. In practice, SNPs out of Hardy–Weinberg equilibrium are typically not analyzed due to concerns for low genotyping quality. However, the issue identified here remains relevant: Hardy–Weinberg equilibrium quality control screening uses a p-value threshold in the range of 10^{-8} (Consortium et al. (2007)), and this practice itself can be called into question because a truly associated SNP is likely to be out of Hardy–Weinberg equilibrium (Turner et al. (2011); Ryckman and Williams (2008)). MultiPhen (O'Reilly et al. (2012)) treats G as a categorical response vari-

able and is robust to Hardy–Weinberg disequilibrium, but its application is limited to independent samples.

For genetic epidemiological interpretation, it may appear that the traditional Y-regressing-on-G approach is more intuitive. But, as pointed out by Schaid et al. (2013) in a different setting, the retrospective alternative that treats G as random while conditioning on Y "overcomes the problem of modelling the ascertainment process, which would be particularly challenging for highly enriched pedigrees." How to model gene-environment interaction using the proposal 'reverse' regression framework, however, remains an open question.

8 Acknowledgment

This research is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC, RGPIN-2018-04934, RGPIN-250053-2013) and the Canadian Institutes of Health Research (CIHR, 201309MOP-310732-G-CEAA-117978) to LS. LZ is a trainee of the CIHR STAGE (Strategic Training in Advanced Genetic Epidemiology) training program at the University of Toronto.

References

Chen, C.-F., 1983: Score tests for regression models. *Journal of the American Statistical Association*, **78 (381)**, 158–161.

Consortium, W. T. C. C., and Coauthors, 2007: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447 (7145)**, 661.

Derkach, A., J. F. Lawless, and L. Sun, 2015: Score tests for association under response-dependent sampling designs for expensive covariates. *Biometrika*, **102 (4)**, 988–994.

Eu-Ahsunthornwattana, J., E. N. Miller, M. Fakiola, S. M. Jeronimo, J. M. Blackwell, H. J. Cordell, W. T. C. C. C. 2, and Coauthors, 2014: Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genetics*, **10 (7)**, e1004 445.

Feng, Z., 2014: A generalized quasi-likelihood scoring approach for simultaneously testing the genetic association of multiple traits. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **63 (3)**, 483–498.

Feng, Z., W. W. Wong, X. Gao, and F. Schenkel, 2011: Generalized genetic association study with samples of related individuals. *The Annals of Applied Statistics*, 2109–2130.

Jakobsdottir, J., and M. S. McPeek, 2013: Mastor: mixed-model association mapping of quantitative traits in samples with related individuals. *The American Journal of Human Genetics*, **92 (5)**, 652–666.

McPeek, M. S., X. Wu, and C. Ober, 2004: Best linear unbiased allele-frequency estimation in complex pedigrees. *Biometrics*, **60 (2)**, 359–367.

O'Reilly, P. F., C. J. Hoggart, Y. Pomyen, F. C. Calboli, P. Elliott, M.-R. Jarvelin, and L. J. Coin, 2012: Multiphen: joint model of multiple phenotypes can increase discovery in gwas. *PloS One*, **7 (5)**, e34 861.

Ryckman, K., and S. M. Williams, 2008: Calculation and use of the hardy-weinberg model in association studies. *Current Protocols in Human Genetics*, **57 (1)**, 1–18.

Sasieni, P. D., 1997: From genotypes to genes: doubling the sample size. *Biometrics*, 1253–1261.

Schaid, D. J., and S. J. Jacobsen, 1999: Biased tests of association: comparisons of allele frequencies when departing from hardy-weinberg proportions. *American Journal of Epidemiology*, **149 (8)**, 706–711.

Schaid, D. J., S. K. McDonnell, J. P. Sinnwell, and S. N. Thibodeau, 2013: Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. *Genetic Epidemiology*, **37 (5)**, 409–418.

Sun, L., and A. Dimitromanolakis, 2012: Identifying cryptic relationships. *Statistical Human Genetics: Methods and Protocols*, 47–57.

Thornton, T., and M. S. McPeek, 2007: Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *The American Journal of Human Genetics*, **81 (2)**, 321–337.

Turner, S., and Coauthors, 2011: Quality control procedures for genome-wide association studies. *Current Protocols in Human Genetics*, **68 (1)**, 1–19.

Visscher, P. M., W. G. Hill, and N. R. Wray, 2008: Heritability in the genomics era—concepts and misconceptions. *Nature Reviews Genetics*, **9 (4)**, 255.

Weir, B. S., 1996: *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. 575.1072 W4.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher, 2011: Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, **88 (1)**, 76–82.