Finite-sample genome-wide regression p-values (GWRPV) with a non-normally distributed phenotype

Gregory Connor
Department of Economics, Finance and Accounting
Maynooth University

Michael O'Neill
School of Business
University College, Dublin

November 8, 2017

Abstract

This paper derives the exact finite-sample p-value for univariate regression of a quantitative phenotype on individual genome markers, relying on a mixture distribution for the dependent variable. The p-value estimator conventionally used in existing genome-wide association study (GWAS) regressions assumes a normally-distributed dependent variable, or relies on a central limit theorem based approximation. The central limit theorem approximation is unreliable for GWAS regression p-values, and measured phenotypes often have markedly non-normal distributions. A normal mixture distribution better fits observed phenotypic variables, and we provide exact small-sample p-values for univariate GWAS regressions under this flexible distributional assumption. We illustrate the adjustment using a years-of-education phenotypic variable.

1 Introduction

This paper provides an alternative estimator of the coefficient p-values in genome-wide univariate regressions of a phenotypic variable on a single-nucleotide polymorphism (SNP). The formula is easy to apply, and can provide substantially more accurate p-values if the phenotypic variable under consideration is non-normally distributed and the number of observations is not extremely large. For a normally distributed phenotypic variable, or with an extremely large sample, the adjustment is not necessary. The magnitude of the p-value adjustment depends upon the size of the sample, the non-normal features of the phenotypic variable including skewness and kurtosis, and the minor allele frequency of the SNP.

Regression based Genome-Wide Association Studies (GWAS) are a key exploratory tool in genetic research on the heritability mechanisms of phenotypic traits, with the goal of identifying individual SNPs associated with a trait. GWAS involves a million or more individual regressions (one per SNP), in the search for SNPs with a significant relationship to an observed phenotypic variable. To account for the multiple comparisons problem, analysts use Bonferonni-corrected p-values, so that an adjusted 5% p-value with one million independent regressions requires an uncorrected univariate regression coefficient p-value (for a two-sided test) of $0.025x10^{-6}$.

In the estimation of Bonferonni-corrected p-values, analysts rely on the assumption that the estimated regression coefficient is normally distributed. This holds exactly if the phenotypic variable has a normal distribution, and approximately (for sufficiently large samples) if the phenotypic variable has any reasonably well-behaved distribution, by the central limit theorem. The quality of the central limit theorem based approximation depends upon the size of the sample, the distributional characteristics both of the observed phenotypic variable and the SNP, and (crucially in this application) on the magnitude of the p-value.

The central limit theorem guarantees uniform convergence of the true cumulative distribution to the normal distribution (see White (1984) for a review). An approximate p-value in the region of 0.025, accurate to within ± 0.0001, can be entirely adequate; an approximate p-value in the region of $0.025x10^{-6} = 0.000000025$ which is similarly accurate to within ± 0.0001 is effectively worthless. GWAS involve very large-number multiple tests and therefore extremely low p-value thresholds, with the conventional critical value set at $0.025x10^{-6}$. Invocation of the central limit theorem is problematic in this context.

In this paper, we develop an alternative approach based on fitting a Bernoulli-normal mixture distribution to the phenotypic variable. As we show, since a genome-wide regression has a trinomial explanatory variable (the three states of the SNP) and the Bernoulli-normal mixture is a combination of a binomial and a normal, the resulting regression coefficient p-value is a multinomial-based linear combination of independent normals, with a closed-form expression in terms of the standard normal distribution. Empirically, the p-value adjustment can be quite large, and can increase or decrease estimated p-values relative to the conventional approach. We illustrate the approach by comparing conventional and adjusted p-value for years-of-education, a common phenotypic variable which has a notably non-normal distribution.

Assuming that the mixture distribution correctly describes the phenotypic variable, our finite-sample adjusted p-value eliminates the reliance on the large-sample-dependent central limit theorem for significance tests in GWAS. This allows for the use of datasets with modest sample sizes and for including SNPs with very low minor allele frequencies. We provide an implementation of the method in the R programming language, as a standard R package from CRAN, along with user documentation.

2 Exact finite-sample p-values for GWAS univariate regression under a Bernoulli-normal mixture distribution

This section presents the new methodological result. We derive the exact p-values of a univariate GWAS regression under an assumed Bernoulli-normal mixture distribution. This is a reasonably straightforward exercise, combining the Bernoulli-normal mixture for the dependent variable with the three-valued explanatory variable of the regression, and then rearranging, manipulating, and simplifying the expressions.

2.1 The GWAS regression framework

The analyst has observations on $i = 1, n$ individuals with the data consisting of a phenotypic variable (such as income, years of education, life satisfaction rating, etc.) and a very long (we assume 10^6 for notational simplicity) string of genetic markers. We assume that the phenotype variable y has known mean and variance; to simplify notation (without loss of generality) we assume that y is standardized and so has zero mean and unit variance. We assume that the phenotype variable y has been pre-whitened with respect to any other confounding explanatory variables by a preliminary regression step, so that the GWAS regressions are univariate. The genetic markers are single nucleotide polymorphisms which have three potential states: major allele homozygote, minor allele homozygote, and heterozygote. Let x_{ij} be the trinomial explanatory variable, set equal to 0 if individual i is a major allele homozygote for the j^{th} genetic marker, 1 if he/she is a heterozygote, and 2 if he/she is a minor allele homozygote.

The research objective is to identify individual SNPs that have an influence on the phenotypic variable. This is done through a set of 10^6 univariate regressions, each using one genetic marker. As null hypothesis, we assume that none of the individual SNPs have an influence on the phenotype, so that:

$$E[y_i|x_{ij}] = 0; \;\; i = 1, n; \;\; j = 1, 10^6. \tag{1}$$

This null hypothesis is tested separately for each j. Let m_x and s_x^2 denote the sample average and mean-square deviation of the explanatory variable x_j. For each individual j we test the statistical significance of the ordinary least squares regression coefficient:

$$\widehat{\beta}_j = \frac{1}{ns_x^2}\sum_{i=1}^{n} y_i(x_{ij} - m_x), \tag{2}$$

where it follows from (1) that $E[\widehat{\beta}_j] = 0$ for all j.

In performing the multiple tests $H_0 : E[\widehat{\beta}_j] = 0$ with $j = 1, .., 10^6$ each tested separately, it is crucial to apply a Bonferonni correction to the individual test critical values. With 10^6 independent tests, and choosing a 95% confidence level, the two-tailed critical values for Bonferonni-corrected multiple-test significance of each coefficient uses a cumulative probability of 0.025×10^{-6} for testing a negative estimated coefficient and $1 - .025 \times 10^{-6}$ for a positive estimated coefficient.

2.2 Fitting a Bernoulli-normal mixture distribution to a phenotypic variable

As we will demonstrate later, the central limit theorem does not always provide a reliable approximation for the extremely small p-values under examination in the large-number multiple test environment of GWAS. We need an alternative estimator of regression p-values to apply in the case of a non-normally distributed phenotypic variable. We need a reasonable distributional assumption on y that fits the phenotypic variable and allows for the feasible computation of exact p-values that do not rely on the central limit theorem approximation. In this section we propose a Bernoulli-normal mixture distribution.

The Bernoulli-normal mixture distribution is a flexible family of distributions with good fit in many applications, and convenient analytical properties in our model.

Let $z_a \sim N(\mu_a, \sigma_a^2)$, $z_b \sim N(\mu_b, \sigma_b^2)$, and λ a Bernoulli distributed random variable with $\lambda = 1$ with probability p; all three random variables assumed independent. The mixed normal y is the random variable:

$$y = \lambda z_a + (1 - \lambda) z_b, \tag{3}$$

which has five parameters: $\mu_a, \sigma_a, \mu_b, \sigma_b, p$. The first two moments are:

$$E[y] = p\mu_a + (1-p)\mu_b \tag{4}$$

$$Var[y] = p(\sigma_a^2 + \mu_a^2) + (1-p)(\sigma_b^2 + \mu_b^2) - E[y]^2 \tag{5}$$

For notational simplicity, consider the case in which $E[y] = 0$ and $Var[y] = 1$, then the third and fourth moments are:

$$E[y^3] = p\mu_a^3 + (1-p)\mu_b^3 + 3(p\mu_a\sigma_a^2 + (1-p)\mu_b\sigma_b^2) \tag{6}$$

$$E[y^4] = p\mu_a^4 + (1-p)\mu_b^4 + 6(p\mu_a^2\sigma_a^2 + (1-p)\mu_b^2\sigma_b^2) + 3(p\sigma_a^4 + (1-p)\sigma_b^4) \tag{7}$$

The distribution can be fitted via EM-maximum likelihood; see McLachlan and Peel (2000) for an overview of mixture distributions and estimation methods.

3 The GWAS univariate regression coefficient as a linear combination of independent normals

We now use the assumption that y has a Bernoulli-normal mixture distribution to derive the exact finite-sample p-values of $\widehat{\beta}_j$. Since we now look at one particular j only, we simplify notation and drop the j subscript. To implement our technique, the analyst counts the number of major allele observations, heterozygote observations and minor allele observations in each regression sample. Let $\{n_0, n_1, n_2\}$ denote these three integer values, with $n_0 + n_1 + n_2 = n$. The sample average and mean-square deviation of the explanatory variable have simple forms, since x_i only takes the three values $0, 1, 2$:

$$\begin{aligned} m_x &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{n_1 + 2n_2}{n} \\ s_x^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)^2 = \frac{n_0 m_x^2 + n_1(1-m_x)^2 + n_2(2-m_x)^2}{n}. \end{aligned}$$

The cumulative distribution of an estimate $\widehat{\beta}$ is the probability under the null hypothesis of a random realization $\widetilde{\beta}$ having that value or less:

$$\Pr[\widetilde{\beta} \leq \widehat{\beta}] = \Pr[\frac{1}{ns_x^2}\sum_{i=1}^{n} y_i(x_i - m_x) \leq \widehat{\beta}]. \tag{8}$$

For notational convenience, we order the observations index so that the major allele observations are listed first, then the heterozygote observations, and then the minor allele observations.

$$\begin{aligned} x_i &= 0;\ i = 1, n_0 \\ x_i &= 1;\ i = n_0 + 1, n_0 + n_1 \\ x_i &= 2;\ i = n_0 + n_1 + 1, n \end{aligned}$$

Writing out the coefficient formula (8) using the observed values n_0, n_1, n_2:

$$\Pr[\widetilde{\beta} \leq \widehat{\beta}] = \Pr[\frac{1}{s_x^2}(\sum_{i=1}^{n_0}(-m_x)y_i + \sum_{i=n_0+1}^{n_0+n_1}(1-m_x)y_i + \sum_{i=n_0+n_1+1}^{n}(2-m_x)y_i) \leq \widehat{\beta}].$$

Under our distributional assumption on y, each of the three integers n_0, n_1 and n_2 in turn decomposes into two (unobserved) integers: the number of realizations of the dependent variable y_{i^*} with conditional mean and standard deviation either $\{\mu_a, \sigma_a\}$ or $\{\mu_b, \sigma_b\}$. We denote these integer realizations with a double subscript, $\{n_{0a}, n_{1a}, n_{2a}\}_h$,and $\{n_{0b}, n_{1b,} n_{2b}\}_h$, $h = 1, m$ where the first subscript refers to the minor allele frequency, $0, 1$, or 2, and the second subscript refers to the (unobserved) number of realizations of $\{\mu_a, \sigma_a\}$ or $\{\mu_b, \sigma_b\}$. Since the second set $\{n_{0b}, n_{1b,} n_{2b}\}$ consists of the remainders $n_{0b} = n_0 - n_{0a}$, $n_{1b} = n_1 - n_{1a}$ and $n_{2b} = n_2 - n_{2a}$, it is sufficient to index with $\{n_{0a}, n_{1a}, n_{2a}\}$.

Each of the integers $n_{0a}, n_{1a,} n_{2a}$ has an independent binomial distribution, since each depends upon the number of $\lambda = 1$ outcomes of the Bernoulli random variable λ within one of the three allele categories. It is easy to see that $m = (n_0 + 1)(n_1 + 1)(n_2 + 1)$. The probabilities of each of the potential outcomes $\{n_{0a}, n_{1a,} n_{2a}\}_h, h = 1, m$ can be found from the binomial formula (for three independent binomials):

$$\begin{aligned} P_h = \Pr[\{n_{0a}, n_{1a,} n_{2a}\}_h] = \\ (\frac{n_0!}{n_{0a}!n_{0b}!})(\frac{n_1!}{n_{1a}!n_{1b}!})(\frac{n_2!}{n_{2a}!n_{2b}!})\times \\ p^{(n_{0a}+n_{1a}+n_{2a})}(1-p)^{(n_{0b}+n_{1b}+n_{2b})}. \end{aligned} \tag{9}$$

The cumulative probability of $\widehat{\beta}$ is the sum of the conditional cumulative probability given each of the potential outcomes $h = 1, m$ times the probability of each outcome:

$$\Pr[\widetilde{\beta} \leq \widehat{\beta}] = \sum_{n_{01}=0}^{n_0} \sum_{n_{11}=0}^{n_1} \sum_{n_{21}=0}^{n_2} (\Pr[\widetilde{\beta} \leq \widehat{\beta}|h]) \times P_h. \tag{10}$$

The last step is to calculate

$$\Pr[\widetilde{\beta} \leq \widehat{\beta}|h] = \Pr[\frac{1}{ns_x^2}\sum_{i=1}^{n} y_i x_i \leq \widehat{\beta}|h]; \tag{11}$$

which is the sum of n independent normals, consisting of n_{0a} draws with mean $-m_x\mu_a$ and standard deviation of $m_x\sigma_a$, plus n_{0b} draws with mean $-m_x\mu_b$ and standard deviation of $m_x\sigma_b$, plus n_{1a} draws with mean $(1 - m_x)\mu_a$ and standard deviation of $(1 - m_x)\sigma_a$, plus n_{1b} draws with mean $(1 - m_x)\mu_b$ and standard deviation of $(1-m_x)\sigma_b$, plus n_{2a} draws with mean $(2-m_x)\mu_a$ and standard deviation of $(2-m_x)\sigma_a$, plus n_{2b} draws with mean $(2-m_x)\mu_b$ and standard deviation of $(2-m_x)\sigma_b$. A sum of independent normals has a normal distribution with mean equal to the sum of the means and variance equal to the sum of the variances. Applying this to (11):

$$\begin{aligned}\Pr[\frac{1}{ns_x^2}\sum_{i=1}^{n} y_i x_i \le \widehat{\beta}|h] = \Pr[&\frac{1}{ns_x^2}(-m_x(\mu_a n_{0a} + \mu_b n_{0b}) \\ &+ (1-m_x)(\mu_a n_{1a} + \mu_b n_{1b}) + (2-m_x)(\mu_a n_{2a} + \mu_b n_{2b}) + \\ &(m_x^2(n_{0a}\sigma_a^2 + n_{0b}\sigma_b^2) + (1-m_x)^2(n_{1a}\sigma_a^2 + n_{1b}\sigma_b^2) \\ &+ (2-m_x)^2(n_{2a}\sigma_a^2 + n_{2b}\sigma_b^2))^{1/2} z) \\ &\le \widehat{\beta}] \end{aligned} \tag{12}$$

where z denotes a standard normal random variable. The right-hand side of equation (12) is simply the p-value of a normally distributed random variable. Therefore, the p-value of the regression coefficient (10) is a binomial-weighted linear combination of conventional normal distribution p-values. There is no need to invoke the (sometimes unreliable) central limit theorem approximation to find the regression p-value. Given the trinomial set of values for the independent variable and a Bernoulli-normal mixture distribution for the dependent variable, the exact finite-sample p-value of the GWA regression coefficient is a directly computable linear combination of normal distribution p-values; we simply combine (9), (10) and (12).

As a tangential benefit, the model also provides exact formulas for the skewness and kurtosis of the estimated regression coefficient (under the null hypothesis that $E[\widehat{\beta}] = 0$). Since y is standardized, it also follows from (2) that $E[\widehat{\beta}_j^2] = 1/(ns_x^2)$. Simply inserting (2) and (3) into the definitions of skewness and kurtosis, and evaluating using the skewness and kurtosis of the mixture distribution, (6) and (7), give closed form expressions that are easy to evaluate, in particular:

$$\begin{aligned} skewness &= E[\widehat{\beta}^3]/(E[\widehat{\beta}^2]^{\frac{3}{2}}) \\ &= (\frac{1}{\sqrt{n}})(\frac{1}{\sqrt{s_x^2}})^3(\frac{n_0}{n}(-m_x)^3 + \frac{n_1}{n}(1-m_x)^3 + \frac{n_2}{n}(2-m_x)^3)E[y^3] \end{aligned} \tag{13}$$

$$\begin{aligned} kurtosis &= E[\widehat{\beta}^4]/(E[\widehat{\beta}^2]^2) \\ &= 3 + (\frac{1}{n})\frac{(E[y^4]-3)}{(s_x^2)^2}(\frac{n_0}{n}m_x^4 + \frac{n_1}{n}(1-m_x)^4 + \frac{n_2}{n}(2-m_x)^4); \end{aligned} \tag{14}$$

see the Technical Appendix for details. This allows the analyst to have a sense of how close to normality is the finite-sample distribution of the estimated regression coefficient in a particular application.

3.1 Computationally efficient implementation of the estimator

The p-value formula (10) requires a sum over the set of outcomes from three independent binomials with n_0, n_1 and n_2 draws, giving a total of $m = (n_0 + 1)(n_1+1)(n_2+1)$ terms. For large n, the number of terms m can be extremely large, but using an efficient computation algorithm, the vast majority of the terms can be dropped from the calculation without any discernible effect on the quality of the estimate. Each of the three random outcomes, n_{ja}, $j = 0, 1, 2$, has an independent binomial distribution with parameter p and draw size n_j. The three univariate binomial distributions for n_{ja}, $j = 0, 1, 2$ have their peaks at pn_j, $j = 0, 1, 2$. The probability of a particular value n_{ja} diminishes exponentially towards zero as n_{ja} moves further away from pn_j. For large n_j, one or both of the "tails" of the univariate probability distribution of n_{ja} can be deleted, when one or both make approximately zero contribution to the sum (10). See the Technical Appendix for details of the computation algorithm used in the GWRPV estimation code.

4 Illustration of the magnitude of the p-value adjustment using years of education

This section examines the magnitude of the adjustment arising from our finite sample p-values compared to using large-sample approximate p-values based on the central limit theorem. We illustrate the adjustment with a commonly used phenotypic variable: years of education, which is a social trait with a strongly non-normal distribution.

Given the parameters of the mixture distribution, our p-value formula (10) is exact; it does not require any simulation. The only inputs needed are the number of major allele homozygote, heterozygote, and minor allele homozygote observations in the regression sample, (n_0, n_1, n_2), the estimated regression coefficient, $\widehat{\beta}$, and the five parameters of the mixture distribution, $(p, \mu_a, \mu_b, \sigma_a, \sigma_b)$.

For the purposes of this comparison, we use five sample sizes, $n = 500$, 1000, 5000, 50000, and 100000. For each sample size we fit n_0, n_1, n_2 from the range of values typically encountered in genome-wide regression tests. Let MAF denote the minor allele frequency of the SNP; we chose four representative values: $MAF = 0.1\%$, 0.5%, 1%, 5%. To choose the observation numbers n_0, $n_{1,}$ n_2 we assume that the SNP is in Hardy-Weinberg equilibrium, which implies that $n_0 = n(1 - MAF)^2$, $n_1 = 2nMAF(1 - MAF)$ and $n_2 = nMAF^2$. The numbers of observations n_0, n_1, n_2 must be integers, so for fractional values we stick the "leftover" one or two observation(s) in the heterozygote category.

Note that the relative numbers of explanatory variables across the three categories, n_0, n_1, n_2, can affect the quality of the central limit theorem approximation. For example, with $MAF = 0.1\%$, only 0.0001% of SNP observations take the value $x = 2$ (zero observations in most small samples); 0.1998% take the value 1, and for 99.8% of the regression sample, $x = 0$. This unbalanced distribution impacts the speed at which the central limit theorem acts upon the probability distribution of the coefficient estimate, and the asymmetry (right-tail probability versus left-tail probability) of its finite-sample distribution, unless the dependent variable is exactly normal. Our p-value adjustment is particularly valuable for genome-wide regressions with low minor allele frequency since it provides a finite-sample test statistic when the large-sample approximation is unreliable. This will become clear in the tables and figures below.

To calibrate the parameters of the mixture distribution, $(p, \mu_a, \mu_b, \sigma_a, \sigma_b)$, we run EM-maximum likelihood on the phenotypic variable; see below for details.

4.1 Application to a non-normal phenotype: Years-of-education

In this subsection we calibrate the Bernoulli-normal mixture using data on years of education from the U.S. Census Bureau Current Population Survey of Educational Attainment, 2015. See Rietveld, et al. (2013, 2015), Okbay et al. (2016), and references therein for details on the considerable number of

GWA regression studies with years-of-education as the phenotypic variable.

Figure 1 shows a frequency distribution of the years-of-education data, along with fitted normal and Bernoulli-normal mixture distributions. See the Technical Appendix for description of the U.S. census data. The mixture distribution picks up the high-peakedness and asymmetry in the data distribution, associated with the 76% frequency of data observations in the range $12-16$ years, and the secondary clump of observations in the $0-6$ years range with frequency 3.04%. These features are missed by the fitted normal. The data has skewness of -0.676781 and kurtosis of 5.126954, which both differ significantly from the normal distribution values (zero and three, respectively) with 99% confidence. The Jarque-Bera test rejects normality with 99% confidence.

The 81,913 years-of-education data observations are fitted to a Bernoulli-normal mixture distribution using the *normalmixEM* command in the *mixtools* library of programming language *R;* see Benaglia et al. (2009) for details on the estimation routine. The estimated parameter values are $p = 0.9654$, $\mu_a = 13.872$, $\mu_b = 4.628$, $\sigma_a = 2.588$, $\sigma_b = 2.518$.

For comparative purposes, in the tables below we assume $\widehat{\beta}$ values at the normal-distribution critical values for 95%, and 99% confidence, and for Bonferonni-adjusted multiple-test 95% confidence with 10^6 independent tests. These $\widehat{\beta}$ values are easily derived from the t-statistics, using $Pr(\frac{\widehat{\beta}}{\sqrt{ns_x^2}} \leq -1.96) = 2.5\%$; $Pr(\frac{\widehat{\beta}}{\sqrt{ns_x^2}} \leq -2.58) = 0.5\%$; and $Pr(\frac{\widehat{\beta}}{\sqrt{ns_x^2}} \leq -5.45) = 2.5\% \times 10^{-6}$; the upper-tail tests are analogous, with a positive sign. We take each of the normal-distribution critical values of $\widehat{\beta}$ and find their p-values under the mixture distribution, which we can then compare to the normality-based p-values, $2.5\%, 0.5\%$, and $2.5\% \times 10^{-6}$.

Table 1 Panel A considers a single-hypothesis, two-sided test with a 95% confidence limit. The table shows exact p-values under the mixture distributions for estimated regression coefficients with normality-based approximate p-values (using the central limit theorem) of 2.5%. The central limit theorem approximation gives quite accurate p-values in almost all cases, even with small sample sizes and low minor allele frequencies. The approximation error from invoking the central limit theorem to estimate p-values is never severe.

Panel B of the table repeats the exercise for a 99% confidence test, so that the p-value under normality is 0.5%. The central limit theorem approximation continues to work reasonably well, with the exception of smallest sample size (500 observations) with minor allele frequencies of 1% or less.

For 10^6 multiple test Bonferonni-corrected p-values with 95% confidence, shown in Table 2, the approximation error from relying on the central limit theorem is severe. Convergence of the p-value toward its normality-derived value is much slower, and the small-sample asymmetry in the approximation error is more notable. For small to medium sample sizes, the true p-value for a negative-tail test is very substantially above $2.5\% \times 10^{-6}$, the p-value for the positive-tail test is substantially below $2.5\% \times 10^{-6}$, and the sum of the two tail probabilities (which should be $5\% \times 10^{-6}$) is substantially higher. The central limit theorem approximation only works reasonably accurately for sample sizes of ten thousand or more, and only with relatively high minor allele frequency. In the other cases considered in Table 2, the finite-sample adjustment is critically important.

Table 3 shows the skewness and kurtosis of the estimated coefficients, using (13) and (14), for the twenty sample specifications considered in Tables 1 and 2. The results in this table explain the pattern of findings in Tables 1 and 2. The estimated coefficient, viewed as a random variable under the null hypothesis, has strong negative skewness, inherited in turn from the negative skewness in the years-of-education variable. The negative skewness in the distribution of the years-of-education variable interacts with the asymmetry in the spread of SNP values (vastly more major allele observations than heterozygote or minor allele) to produce biases in the conventional p-values. For negative estimated betas, the conventional p-value understates Type I error, rejecting the null hypothesis when it should be accepted. For positive estimated betas, the conventional p-value sacrifices power by accepting the null hypothesis when it could be rejected with confidence.

5 Summary

This paper provides a new approach to estimating Bonferonni-corrected multiple-test p-values for regressions of individual genetic markers on a phenotypic variable. The current standard approach in computing coefficient p-values is to assume a normal distribution for the phenotypic variable, or to invoke the central limit theorem to justify the approximate normality of the coefficient estimate. Many phenotypic variables, particularly for social traits like income and education levels, have distributions which are far from normality. The central limit theorem, as we show, does not give reliable p-values for the types of sample sizes (and multiple-test numbers) used in genome-wide association studies (GWAS) with non-normally distributed phenotypic variables.

We suggest a new approach, based on fitting a Bernoulli-normal mixture distribution to the phenotypic variable (prewhitened with respect to any other confounding explanatory variables) before running univariate GWAS regressions. We derive the exact finite-sample distribution of GWAS regression coefficients p-values under this more flexible distributional assumption. We illustrate the magnitude of the p-value adjustment from our approach (relative to the conventional approach) with sample data on a commonly-used phenotypic variable with a notably non-normal distribution: years of education. The derived p-values differ markedly from the conventional, normality-based, p-values. This new approach allows for smaller samples sizes and lower minor allele frequencies since it does not rely on a large-sample central limit theorem approximation. We provide an implementation of the method in the R programming language, which is available as a standard R package from CRAN, along with user documentation.

Acknowledgments

We would like to thank Philipp Koellinger, Robert Korajczyk, Richard Linner, Oliver Linton, Brian O'Kelly and Donal O'Neill for helpful comments. We wish to acknowledge support from the Science Foundation of Ireland.

References

[1] Benaglia T., D. Chauveau, DR. Hunter, D. Young (2009). "Mixtools: An R Package for Analyzing Finite Mixture Models," *Journal of Statistical Software*, 32(6):1-29.

[2] McLachlan, G., and D. Peel (2000). *Finite Mixture Models*, Wiley Series in Probability and Statistics, New York.

[3] Okbay, A., JP. Beauchamp, MA. Fontana, JJ. Lee, TH. Pers, CA. Rietveld, P. Turley, G. Chen, V. Emilsson, SFW. Meddens, S. Oskarsson, JK. Pickrell, K. Thom, P. Timshel, R. de Vlaming, A. Abdellaoui, TS. Ahluwalia, J. Bacelis, C. Baumbach, G. Bjornsdottir, JH. Brandsma, MP. Concas, J. Derringer, NA. Furlotte, TE. Galesloot, et al. (2016). "Genome-wide association study identifies 74 loci associated with educational attainment," *Nature,* 533: 539–542.

[4] Rietveld CA, , SE. Medland, J. Derringer, J. Yang, T. Esko, NW. Martin, HJ. Westra, et al. (2013). "GWAS of 126,559 individuals identifies genetic variants associated with educational attainment," *Science,* 340(6139): 1467–1471.

[5] Rietveld, CA., T. Esko, G. Davies, TH. Pers, P. Turley, B. Benyamin, CF. Chabris, V. Emilsson, AD. Johnson, J J. Lee, C. de Leeuw, RE. Marioni, SE. Medland, MB. Miller, O. Rostapshova, SJ. van der Lee, AAE. Vinkhuyzen, N. Amin, D. Conley, J. Derringer, CM. van Duijn, R. Fehrmann, L. Franke, EL. Glaeser, NK. Hansell, C. Hayward, WG. Iacono, C. Ibrahim-Verbaas, V. Jaddoe, J. Karjalainen, D. Laibson, P. Lichtenstein, DC. Liewald, PKE. Magnusson, NG. Martin, M. McGue, G. McMahon, NL. Pedersen, S. Pinker, DJ. Porteous, D. Posthuma, F. Rivadeneira, BH. Smith, JM. Starr, H. Tiemeier, NJ. Timpson, M. Trzaskowski, AG. Uitterlinden, FC. Verhulst, ME. Ward, MJ. Wright, GD. Smith, IJ. Deary, M. Johannesson, R. Plomin, PM. Visscher, DJ. Benjamin, D. Cesarini, and PD. Koellinger (2015). "Common genetic variants associated with cognitive performance identified using the proxy-phenotype method," *Publications of the National Association of Science, Psychological and Cognitive Sciences, Genetics*, January, 12(15): 13790-13794.

[6] U.S. Census Bureau (2015). *Current Population Survey Data on Educational Attainment: 2015,* available at http:// www.census.gov/ hhes/ socdemo/ education/.

[7] U.S. Center for Disease Control and Prevention (2010). *Behavioral Risk Factor Surveillance System,* available at http://www.cdc.gov/brfss/ annual_data/annual_2010.htm.

[8] White, Halbert (1984). *Asymptotic theory for econometricians*, Academic Press, Inc., London.

Figure 1

The frequency distribution of years-of-education and fitted normal and mixture distributions

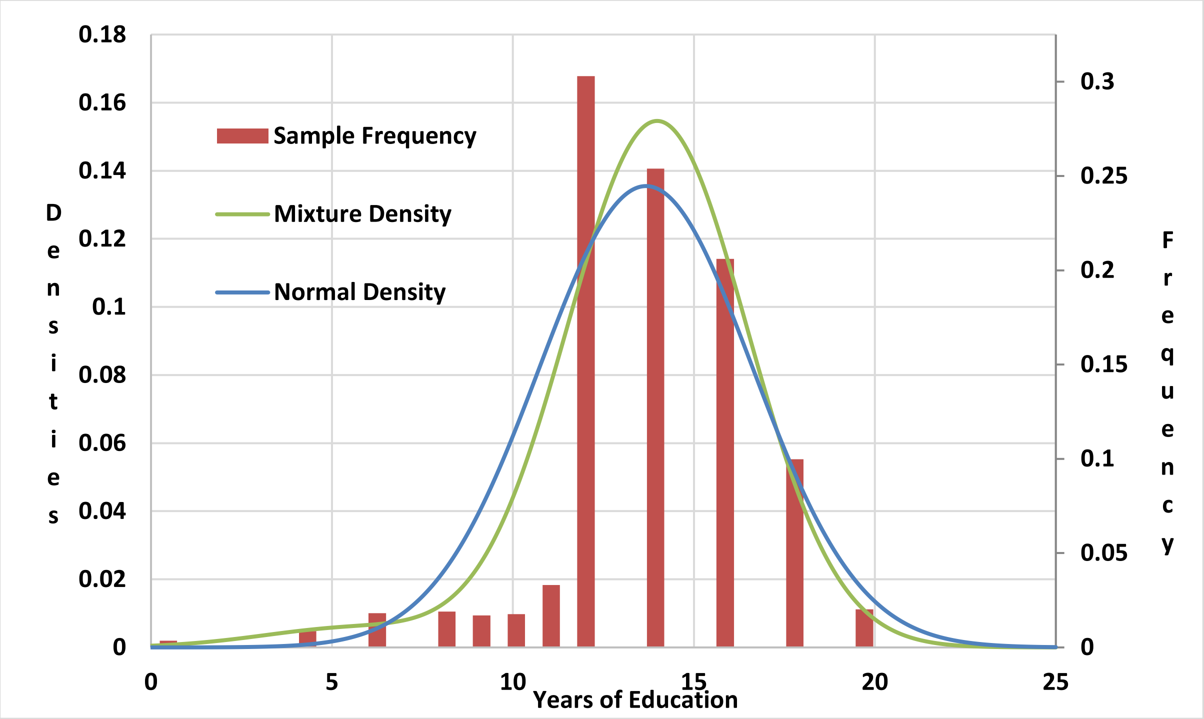

Table 1

Comparison of adjusted/unadjusted single-test p-values for GWA regression coefficients under a mixture distribution: Years of education

Panel A: 95% two-tailed confidence test (conventional p-value=0.025)

	Sample Size:	**500**	**1000**	**5000**	**50000**	**100000**
Minor allele frequency	Sign of tested coefficient	p-value				
0.1%	Positive	0.012936	0.014220	0.018346	0.022736	0.023396
	Negative	0.036828	0.040418	0.031954	0.027251	0.026596
0.5%	Positive	0.016554	0.018427	0.021854	0.023992	0.024287
	Negative	0.034740	0.031877	0.028132	0.026006	0.025712
1%	Positive	0.018530	0.020212	0.022799	0.024290	0.024503
	Negative	0.031779	0.029837	0.027191	0.025707	0.025505
5%	Positive	0.022143	0.022958	0.024067	0.024706	0.024787
	Negative	0.027909	0.027063	0.025937	0.025300	0.025212

Panel B: 99% two-tailed confidence test (conventional p-value=0.005)

	Sample Size:	**500**	**1000**	**5000**	**50000**	**100000**
Minor allele frequency	Sign of tested coefficient	p-value				
0.1%	Positive	0.001539	0.001750	0.002587	0.003988	0.004262
	Negative	0.023088	0.017138	0.009214	0.006196	0.005829
0.5%	Positive	0.002191	0.002611	0.00365	0.004525	0.004659
	Negative	0.011185	0.009165	0.006713	0.005513	0.005358
1%	Positive	0.002642	0.003103	0.004015	0.004663	0.004758
	Negative	0.009103	0.007784	0.006163	0.005357	0.005250
5%	Positive	0.003778	0.00409	0.004561	0.004856	0.004898
	Negative	0.006610	0.006104	0.005479	0.005147	0.005106

Table 2

Comparison of adjusted/unadjusted 10^6 multiple-test p-values for GWA regression coefficients under a mixture distribution: Years of education

95% two-tailed confidence test (conventional p-valuex10^6 = 0.025)

	Sample Size:	**500**	**1000**	**5000**	**50000**	**100000**
Minor allele frequency	Sign of tested coefficient	p-valuex10^6				
0.1%	Positive	0.000086	0.000115	0.000349	0.003322	0.005567
	Negative	28.85998	56.65438	4.016193	0.199445	0.112986
0.5%	Positive	0.000213	0.000365	0.001747	0.009510	0.012572
	Negative	11.83049	3.895693	0.404728	0.065809	0.049793
1%	Positive	0.000394	0.000741	0.003448	0.012650	0.015399
	Negative	3.748166	1.306868	0.190816	0.049647	0.040670
5%	Positive	0.002526	0.004296	0.010469	0.018774	0.020353
	Negative	0.374346	0.182129	0.062877	0.033458	0.030762

Table 3

Skewness and kurtosis of GWAS regression coefficients under a mixture distribution: Years of education

		500	**1000**	**5000**	**50000**	**100000**
Sample Size:						
Minor allele frequency:						
0.1%	Skewness	-0.81933	-0.57935	-0.25909	-0.08193	-0.05794
	Kurtosis	5.071641	4.035821	3.207164	3.020716	3.010358
0.5%	Skewness	-0.36198	-0.25596	-0.11447	-0.03642	-0.02575
	Kurtosis	3.407713	3.203856	3.040771	3.004160	3.00208
1%	Skewness	-0.25201	-0.1782	-0.07969	-0.02560	-0.0181
	Kurtosis	3.19976	3.09988	3.019976	3.002101	3.00105
5%	Skewness	-0.10559	-0.07466	-0.03383	-0.01073	-0.00759
	Kurtosis	3.041966	3.020983	3.004343	3.000438	3.000219

Appendix A: Technical Appendix

Gregory Connor
Department of Economics, Finance and Accounting
Maynooth University

Michael O'Neill
Graduate School of Business
University College, Dublin

November 9, 2017

1 Introduction

This technical appendix discusses the computationally efficient procedures used in the R code, provides detailed derivations of the skewness and kurtosis of the estimated regression coefficient, and describes the years-of-education data set used in the paper.

2 The regression format

The p-value formula in GWRPV assumes that the regression coefficient $\widehat{\beta}$ comes from a univariate genome-wide regression model, in which the influence of other confounding variates have been removed in a first-stage regression. The analyst begins with a raw phenotypic variable $\widetilde{y}$ which potentially includes the influence of confounding variates X_c, such as the dominant principal components of the genetic variants matrix. The raw phenotype $\widetilde{y}$ is first regressed on these confounding variates to remove their influence:

$$y = (I - X_c(X_c'X_c)^{-1}X_c')\widetilde{y}.$$

The prewhitened phenotypic variate y is then used throughout.

For a given SNP, the analyst has regressed the prewhitened phenotypic variable y on a single nucleotide polymorphism (SNP) x:

$$\widehat{\beta} = \frac{1}{ns_x^2}\sum_{i=1}^{n} y_i(x_i - m_x) \tag{A1}$$

with sample $\{y_i, x_i\}_{i=1,n}$, where $x_i = 0$ if the SNP is a major homozygote, $x_i = 1$ if the SNP is a heterozygote, $x_i = 2$ if the SNP is a minor homozygote. The inputs for computing the p-value of a given coefficient estimate are the five parameters of the mixture distribution, $\mu_a, \sigma_a, \mu_b, \sigma_b, p$, the number of each of the three values for the independent variable, n_0, n_1, n_2, and the estimated coefficient, $\widehat{\beta}$.

The model assumes that y has a Bernoulli-normal mixture distribution. For notational simplicity, the main paper and this technical note also assume that y is standardized in the regression, so that $E[y] = 0$ and $E[y^2] = 1$. However this assumption is only for notational simplicity; the analyst need not standardize y before running the regression (A1) or the GWRPV program. Under the assumed distribution, $E[y] = p\mu_a + (1-p)\mu_b$ and $Var[y] = p(\mu_a^2 + \sigma_a^2) + (1-p)(\mu_b^2 + \sigma_b^2) - E[y]^2$. The GWRPV program takes the inputted regression coefficient $\widehat{\beta}$ and, using the inputted parameters to compute $Var[y]$, scales the inputted regression estimate by $Var[y]^{-1/2}$. This has the effect of replacing the regression model with the equivalent one with standardized y. This linear transform of the regression model does not affect the p-value of the regression coefficient but simplifies the calculations.

The sample average sum of squares of x is:

$$s_x^2 = \frac{n_0 m_x^2 + n_1(1-m_x)^2 + n_2(2-m_x)^2}{n}. \tag{A2}$$

In running the regression, it is acceptable to subtract one from x and rescale so that $x = -1, 0, 1$ since this does not impact $\widehat{\beta}$ or s_x^2. For notational simplicity we will assume $x = 0, 1, 2$.

3 Computing the exact p-value

Consider the case $\widehat{\beta} \leq 0$. In order to determine the p-value, one considers this observed $\widehat{\beta}$ estimate and finds the cumulative probability of random outcomes which would give this value or less. (If $\widehat{\beta} > 0$ the p-value is computed as one minus this probability.) The cumulative distribution of an estimate $\widehat{\beta}$ under the null hypothesis $\beta = 0$ is:

$$\Pr[\frac{1}{ns_x^2}\sum_{i=1}^{n} y_i(x_i - m_x) \leq \widehat{\beta}]. \tag{A3}$$

Each of the three integers n_0, n_1 and n_2, decomposes into two (unobserved) integers: the number of realizations of the dependent variable y_i with the conditional normal distribution with mean and standard deviation $\{\mu_a, \sigma_a\}$ or $\{\mu_b, \sigma_b\}$. We denote these integer realizations with a double subscript, $\{n_{0a}, n_{1a}, n_{2a}\}_h$,and $\{n_{0b}, n_{1b}, n_{2b}\}_h$, $h = 1, m$ where the first subscript refers to the minor allele frequency, $0, 1$, or 2, and the second subscript refers to the (unobserved) number of realizations of $\{\mu_a, \sigma_a\}$ or $\{\mu_b, \sigma_b\}$, depending upon whether the Bernoulli random variable λ equals one or zero. Note that $m = (n_0+1)(n_1+1)(n_2+1)$. Since the second set $\{n_{0b}, n_{1b,} n_{2b}\}$ consists of the remainders $n_{0b} = n_0 - n_{0a}$, $n_{1b} = n_1 - n_{1a}$ and $n_{2b} = n_2 - n_{2a}$, it is sufficient to index with $\{n_{0a}, n_{1a}, n_{2a}\}$. Let $P_h, h = 1, m$ denote the probabilities of all these unobserved potential outcomes $\{n_{0a}, n_{1a}, n_{2a}\}_h$. The probability of the $n-$sum (A3) can be written as a probability-weighted $m-$sum over all the potential outcomes:

$$\Pr[\frac{1}{ns_x^2}\sum_{i=1}^{n} y_i(x_i - m_x) \leq \widehat{\beta}] =$$

$$\sum_{n_{0a}=0}^{n_0} \sum_{n_{1a}=0}^{n_1} \sum_{n_{2a}=0}^{n_2} \Pr[\frac{1}{ns_x^2}\sum_{i=1}^{n} y_i(x_i - m_x) \leq \widehat{\beta} | \{n_{0a}, n_{1a}, n_{2a}\}_h] \times P_h. \tag{A4}$$

The probabilities of each potential outcome $P_h = \Pr[\{n_{0a}, n_{1a}, n_{2a}\}_h]$, can be found from the binomial formula (for three independent binomials).

$$\begin{gathered} P_h = \Pr[\{n_{0a}, n_{1a}, n_{2a}\}_h] = \\ (\frac{n_0!}{n_{0a}!n_{0b}!})(\frac{n_1!}{n_{1a}!n_{1b}!})(\frac{n_2!}{n_{2a}!n_{2b}!})\times \\ p^{(n_{0a}+n_{1a}+n_{2a})}(1-p)^{(n_{0b}+n_{1b}+n_{2b})}. \end{gathered} \tag{A5}$$

The formula (A5) consists of a potentially very large number, $(\frac{n_0!}{n_{0a}!n_{0b}!})(\frac{n_1!}{n_{1a}!n_{1b}!})(\frac{n_2!}{n_{2a}!n_{2b}!})$, multiplied by a potentially extremely small number, $p^{(n_{0a}+n_{1a}+n_{2a})}(1-p)^{(n_{0b}+n_{1b}+n_{2b})}$. To prevent numeric overflow, the program computes the log of each of these two terms, sums the logs, and then takes the antilog.

4 Trimming the set of feasible combinations

The p-value formula (A4) requires a sum over the set of outcomes from three independent binomials with n_0, n_1 and n_2 draws, giving a total of $m = (n_0 + 1)(n_1 + 1)(n_2 + 1)$ terms. If all three values n_0, n_1, n_2 are very large (or even if only two are very large and the third is moderately large) this can be an extremely large number of terms to compute in the sum, and not computationally necessary. It is possible to shrink this computational burden substantially without sacrificing any estimation accuracy.

For each $h = 1, m$, rearrange (A4) listing the values associated with n_0, n_1, n_2 separately:

$$\begin{aligned}\Pr[\widetilde{\beta} \leq \widehat{\beta}] = \sum_{n_{0a}=0}^{n_0} \sum_{n_{1a}=0}^{n_1} \sum_{n_{2a}=0}^{n_2} & (\frac{n_0!}{n_{0a}!n_{0b}!})p^{(n_{0a})}(1-p)^{(n_{0b})} \times \\ & (\frac{n_1!}{n_{1a}!n_{1b}!})p^{(n_{1a})}(1-p)^{(n_{1b})} \times \\ & (\frac{n_2!}{n_{2a}!n_{2b}!})p^{(n_{2a})}(1-p)^{(n_{2b})} \times \\ \Pr[\frac{1}{ns_x^2}\sum_{i=1}^{n} y_i(x_i - m_x) \leq \widehat{\beta} | (n_{0a}, n_{1a}, n_{2a})]. \end{aligned} \tag{A6}$$

Each of the four multiplicative terms in (A6) is a probability and lies between zero and one, for every (n_{0a}, n_{1a}, n_{2a}). Depending upon the values of n_0, n_1, n_2, and p, the vast majority of the terms in (A6) are indistinguishable from zero.

To speed computation in the GWRPV programme, a range of beginning and/or ending terms in (A6) with extremely low cumulative probability are dropped from the sum. We divide the summation into two parts: a range of index values A which by construction covers virtually all of the total probability, and the complement set with total probability very close to zero. The separate computation interval for each of n_{0a}, n_{1a}, and n_{2a} is chosen using the univariate distributions. Let $(\overline{n}_{0a}, \overline{\overline{n}}_{0a})$ denote the upper and lower limits, with $\overline{n}_{0a} \geq 0$, $\overline{\overline{n}}_{0a} \leq n_0$ and $\overline{n}_{0a} \leq \overline{\overline{n}}_{0a}$. Let δ denote a small positive number (the GWRPV code uses $\delta = 10^{-16}$ as the default value but this can be altered by the user). The range limits are chosen such that at most δ probability lies outside the range:

$$\overline{n}_{0a} = \arg\max_{\overline{n}_{0a}} \Pr[n_{0a} < \overline{n}_{0a}] \leq (\frac{1}{6})\delta$$

where of course $\Pr[n_{0a} < 0] = 0$. Note that $\Pr[n_{0a} < \overline{n}_{0a}]$ is the cumulative probability of a univariate binomial with parameter p (a trivial computation). Similarly for the upper limit on the range:

$$\overline{\overline{n}}_{0a} = \arg\min_{\overline{\overline{n}}_{0a}} \Pr[n_{0a} > \overline{\overline{n}}_{0a}] \leq (\frac{1}{6})\delta$$

where of course $\Pr[n_{0a} > n_0] = 0$. The values for $(\overline{n}_{1a}, \overline{\overline{n}}_{1a})$ and $(\overline{n}_{2a}, \overline{\overline{n}}_{2a})$ are chosen analogously. Since each of the ommitted tail ranges have probability less than or equal to $(\frac{1}{6})\delta$, the set A has total probability greater than or equal to $1 - \delta$. The p-value is separated into the conditional probability given $(n_{0a}, n_{1a}, n_{2a}) \in A$ times the probability of $(n_{0a}, n_{1a}, n_{2a}) \in A$, plus the remaining probability:

$$\begin{aligned}
&\Pr[\widetilde{\beta} \leq \widehat{\beta}] = \Pr[\widetilde{\beta} \leq \widehat{\beta} | (n_{0a}, n_{1a}, n_{2a}) \in A] \Pr[(n_{0a}, n_{1a}, n_{2a}) \in A] + \\
&\Pr[\widetilde{\beta} \leq \widehat{\beta} | (n_{0a}, n_{1a}, n_{2a}) \notin A] \Pr[(n_{0a}, n_{1a}, n_{2a}) \notin A].
\end{aligned}$$

Using the restricted range A in (A6) in place of the full range:

$$\begin{aligned}
\Pr[\widetilde{\beta} \leq \widehat{\beta}] = \sum_{n_{0a}=\overline{n}_{0a}}^{\overline{\overline{n}}_{0a}} \sum_{n_{1a}=\overline{n}_{1a}}^{\overline{\overline{n}}_{1a}} \sum_{n_{2a}=\overline{n}_{2a}}^{\overline{\overline{n}}_{2a}} (\frac{n_0!}{n_{0a}! n_{0b}!}) p^{(n_{0a})} (1-p)^{(n_{0b})} \times \\
(\frac{n_1!}{n_{1a}! n_{1b}!}) p^{(n_{1a})} (1-p)^{(n_{1b})} \times \\
(\frac{n_2!}{n_{2a}! n_{2b}!}) p^{(n_{2a})} (1-p)^{(n_{2b})} \times \\
\Pr[\frac{1}{n s_x^2} \sum_{i=1}^{n} y_i (x_i - m_x) \leq \widehat{\beta} | (n_{0a}, n_{1a}, n_{2a})] + f[\delta]. \qquad \text{(A7)}
\end{aligned}$$

where $f[\delta] \leq \delta$.

The trimming parameter is set by the parameter *logdelta* in the optional input file trimparameters.txt. *logdelta* is in log base 10 format and has a default value of -16.

5 Using a normal approximation for the sample average phenotype of the major homozygote observations

If the sample size n is large and the minor allele frequency is low (so that the number of major allele observations n_0 is large) then it may be possible to greatly speed computation by applying a central limit theorem approximation to the sample average of the phenotype values within the subset of major allele observations.

Recall that, for notational convenience, the observations are ordered by allele type. The beta coefficient $\widehat{\beta}$ can be written as a linear combination of three random variables, the average of the phenotype over the major homozygote, heterozygote, and minor homozygote observations:

$$\widehat{\beta} = \frac{1}{s_x^2}(\frac{n_0}{n}(-m_x)m_{n0} + \frac{n_1}{n}(1-m_x)m_{n1} + \frac{n_2}{n}(2-m_x)m_{n2})$$

where

$$m_{n0} = \frac{1}{n_0}\sum_{i=1}^{n_0} y_i;\ m_{n1} = \frac{1}{n_1}\sum_{i=n_0+1}^{n_0+n_1} y_i;\ m_{n2} = \frac{1}{n_2}\sum_{i=n_0+n_1+1}^{n} y_i.$$

Since y has a Bernoulli-normal mixture distribution, if follows easily that m_{n0} is asymptotically normal for large n_0. The GWRPV program computes the skewness and kurtosis of m_{n0} and measures how far they are from their normal-based values of zero and three. See Section 7 below for the formulas for $Skew[m_{n0}]$ and $Kurt[m_{n0}]$. If $(Skew[m_{n0}])^2 + (Kurt[m_{n0}] - 3)^2 < nearnorm$, then the GWRPV program uses a normal approximation for m_{n0}. This greatly shrinks the computation time of the program, since the computation loop (A6) need only run over n_1 and n_2, which tend to be much smaller than n_0. The program has a default value of *nearnorm*$= 10^{-6}$. The analyst can alter the value of *nearnorm* by inserting a different value of *lognearnorm* (in base 10 format) into trimparameters.txt.

6 Setting an upper limit on the computation sum

The GWRPV program has a built-in computation limit, *topsum*, designed to prevent an individual regression result from causing a long computation delay. Before running the main computation sum (A7) for each regression case $(\widehat{\beta}, n_0, n_1, n_2)$, the program computes the number of terms in this computation sum, which is $(\overline{\overline{n}}_{0a} - \overline{n}_{0a} + 1)(\overline{\overline{n}}_{1a} - \overline{n}_{1a} + 1)(\overline{\overline{n}}_{2a} - \overline{n}_{2a} + 1)$. If this number is greater than *topsum*, and the skewness and kurtosis are not sufficiently close to normal values to invoke a normality-based p-value, then the program skips the p-value computation for that $\widehat{\beta}$ and procedes to the next regression case. A value of -999.9 is inserted in place of the p-value in the output file pvalues.txt to flag that the p-value computation has been skipped. The analyst can alter the value of *topsum* by inserting a different value of *logtopsum* (in base 10 format) into trimparameters.txt. The code has a default value of *topsum*$= 10^8$.

7 The skewness and kurtosis of the estimated coefficient

In addition to computing the p-value, the GWRPV programme computes the skewness and kurtosis of the estimated regression coefficient, under the assumed mixture distribution of the dependent variable and under the assumed null hypothesis that the true coefficient equals zero. That is, the programme computes:

$$\begin{aligned} skewness &= E[\widehat{\beta}^3]/(E[\widehat{\beta}^2]^{\frac{3}{2}}) \\ &= (\frac{1}{\sqrt{n}})(\frac{1}{\sqrt{s_x^2}})^3(\frac{n_0}{n}(-m_x)^3 + \frac{n_1}{n}(1-m_x)^3 + \frac{n_2}{n}(2-m_x)^3)E[y^3] \end{aligned} \tag{A8}$$

$$\begin{aligned} kurtosis &= E[\widehat{\beta}^4]/(E[\widehat{\beta}^2]^2) \\ &= 3 + (\frac{1}{n})\frac{(E[y^4]-3)}{(s_x^2)^2}(\frac{n_0}{n}m_x^4 + \frac{n_1}{n}(1-m_x)^4 + \frac{n_2}{n}(2-m_x)^4) \end{aligned} \tag{A9}$$

under the model's assumptions. The first equalities in (A8) and (A9) are definitional; the second equalities will be derived here. Note that $skewness \rightarrow 0$ and $kurtosis \rightarrow 3$ as $n \rightarrow \infty$.

Recall the definition of $\widehat{\beta}$, where we order the observations according to SNP allele type:

$$\widehat{\beta} = \frac{1}{ns_x^2}(\sum_{i^*=1}^{n_0}(-m_x)y_{i^*} + \sum_{i^*=n_0+1}^{n_0+n_1}(1-m_x)y_{i^*} + \sum_{i^*=n_0+n_1+1}^{n}(2-m_x)y_{i^*}). \quad \text{(A10)}$$

The three additive terms in (A10) will be denoted a, b and c. Since y_i is independently distributed across i and $E[y_i] = 0$, we have $E[a^p b^q c^r] = E[a^p]E[b^q]E[c^r]$ and $E[a^p b^q c^r] = 0$ if p, q or r equals one. First consider the value in the numerator of both the skewness and kurtosis formulas (A8) and (A9). Writing the square using a, b and c:

$$E[\widehat{\beta}^2] = E[(a+b+c)^2] = E[a^2] + E[b^2] + E[c^2] \quad \text{(A11)}$$

since all other terms in the product have at least one unit power of a, b, or c and so have zero expectation. Expanding out $E[a^2]$ and dropping terms which have a unit power of y_i for some i :

$$E[a^2] = E[(\frac{1}{ns_x^2}\sum_{i^*=1}^{n_0}(-m_x)y_{i^*})^2] \quad \text{(A12)}$$

$$= (\frac{-m_x}{ns_x^2})^2 \sum_{i_1=1}^{n_0}\sum_{i_2=1}^{n_0} E[y_{i_1}y_{i_2}] = (\frac{1}{ns_x^2})^2 n_0 m_x^2 E[y^2],$$

and recall that $E[y^2] = 1$ since this variable is standardized. Repeating (A12) for $E[b^2]$ and $E[c^2]$:

$$E[b^2] = (\frac{1}{ns_x^2})^2 n_1(1-m_x)^2$$

$$E[c^2] = (\frac{1}{ns_x^2})^2 n_2(2-m_x)^2$$

Inserting the expressions for $E[a^2]$, $E[b^2]$ and $E[c^2]$ into (A11):

$$E[\widehat{\beta}^2] = (\frac{1}{ns_x^2})^2(n_0 m_x^2 + n_1(1-m_x)^2 + n_2(2-m_x)^2)$$

$$= \frac{1}{ns_x^2}(\frac{n_0 m_x^2 + n_1(1-m_x)^2 + n_2(2-m_x)^2}{n})(\frac{1}{s_x^2}) = \frac{1}{ns_x^2}, \quad \text{(A13)}$$

substituting the expression (A2) for s_x^2 and cancelling from the numerator and denominator.

Next, we turn to the numerator in the definition of skewness (A8). Using (A10) gives:

$$E[\widehat{\beta}^3] = E[(a+b+c)^3] = E[a^3] + E[b^3] + E[c^3]$$
$$= (\frac{1}{ns_x^2})^3(n_0(-m_x)^3 + n_1(1-m_x)^3 + n_2(2-m_x)^3)E[y^3]. \tag{A14}$$

Taking the numerator divided by the denominator to get skewness:

$$E[\widehat{\beta}^3]/(E[\widehat{\beta}^2]^{\frac{3}{2}}) = (\frac{1}{\sqrt{n}})(\frac{1}{\sqrt{s_x^2}})^3(\frac{n_0}{n}(-m_x)^3 + \frac{n_1}{n}(1-m_x)^3 + \frac{n_2}{n}(2-m_x)^3)E[y^3].$$

Following the same procedure as above, now for the computation of the numerator in the expression for kurtosis (A9):

$$E[\widehat{\beta}^4] = E[(a+b+c)^4] = E[a^4] + E[b^4] + E[c^4] + \tag{A15}$$
$$6E[a^2]E[b^2] + 6E[a^2]E[c^2] + 6E[b^2]E[c^2].$$

Consider first the term $E[a^4]$. Expanding this out using the definition of a in (A10):

$$E[(\frac{1}{ns_x^2}(\sum_{i=1}^{n_0}(-m_x)y_i)^4] = (\frac{1}{ns_x^2})^4\sum_{i_1=1}^{n_0}\sum_{i_2=1}^{n_0}\sum_{i_3=1}^{n_0}\sum_{i_4=1}^{n_0}(-m_x)^4 y_{i_1}y_{i_2}y_{i_3}y_{i_4} \tag{A16}$$

Dropping all terms in (A16) which include a unit power of y_{i_1}, y_{i_2}, y_{i_3}, or y_{i_4} gives n_0 terms with $i_1 = i_2 = i_3 = i_4$ and $n_0(n_0-1)$ terms for each of the three cases $\{i_1 = i_2, i_3 = i_4,\ i_1 \neq i_3\}$, $\{i_1 = i_3, i_2 = i_4,\ i_1 \neq i_2\}$, $\{i_1 = i_4, i_2 = i_3,\ i_1 \neq i_2\}$. The first set of terms all have the same individual term value of $E[y^4]m_x^4$ and similarly the other three sets of terms all have individual term values of $E[y^2]^2m_x^4 = m_x^4$. Summing all sets of terms:

$$E[a^4] = (\frac{1}{ns_x^2})^4(n_0E[y^4] + 3n_0(n_0-1))m_x^4. \tag{A17}$$

Repeating exactly the same steps to generate (A17) for b and c :

$$E[b^4] = (\frac{1}{ns_x^2})^4(n_1E[y^4] + 3n_1(n_1-1))(1-m_x)^4 \tag{A18}$$

$$E[c^4] = (\frac{1}{ns_x^2})^4(n_2E[y^4] + 3n_2(n_2-1))(2-m_x)^4. \tag{A19}$$

Inserting the expressions for $E[a^2]$, $E[b^2]$, $E[c^2]$, $E[a^4]$, $E[b^4]$, and $E[c^4]$ into (A15) gives:

$$\begin{aligned} E[\widehat{\beta}^4] = (\frac{1}{ns_x^2})^4((n_0E[y^4]{+}3(n_0(n_0-1))(m_x)^4+ \\ &(n_1E[y^4]+3(n_1(n_1-1))(1-m_x)^4+ \\ &(n_2E[y^4]+3(n_2(n_2-1))(2-m_x)^4+ \\ &6n_0n_1(m_x)^2(1-m_x)^2+6n_0n_2(m_x)^2(2-m_x)^2+ \\ &6n_1n_2(1-m_x)^2(2-m_x)^2. \end{aligned} \tag{A20}$$

In order to simplify (A20) it is necessary to cancel out $(ns_x^2)^2$ from a collection of terms in the the numerator. First, rewriting $(ns_x^2)^2$ using the expression (A2) for s_x^2 :

$$\begin{aligned} (ns_x^2)^2 = n_0^2m_x^4+n_1^2(1-m_x)^4+n_2^2(2-m_x)^4+ \\ 2n_0n_1m_x^2(1-m_x)^2+2n_0n_2m_x^2(2-m_x)^2+2n_1n_2(1-m_x)^2(2-m_x)^2. \end{aligned} \tag{A21}$$

Gathering terms in (A20) that match those in (A21) inside the first curly bracket and the remainders (which all multiply $(E[y^4]-3)$) inside the second curly bracket:

$$\begin{aligned} E[\widehat{\beta}^4] = (\frac{1}{ns_x^2})^4(3\{n_0^2(m_x)^4+n_1^2(1-m_x)^4+n_2^2(2-m_x)^4+ \\ 2n_0n_1(m_x)^2(1-m_x)^2+2n_0n_2(m_x)^2(2-m_x)^2+ \\ 2n_1n_2(1-m_x)^2(2-m_x)^2\}+ \\ (E[y^4]-3)\{(n_0(m_x)^4+(n_1(1-m_x)^4+ \\ n_2(2-m_x)^4\} \end{aligned}$$

Dividing by $E[\widehat{\beta}^2]^2$ using (A13) and (A21) gives 3 plus a remainder:

$$E[\widehat{\beta}^4]/\left(E[\widehat{\beta}^2]^2\right) = 3+\frac{(E[y^4]-3)}{(ns_x^2)^2}(n_0m_x^4+n_1(1-m_x)^4+n_2(2-m_x)^4).$$

Next we derive the skewness and kurtosis of the Bernoulli-normal mixture distribution. We only consider the standardized case $E[y]=0$ and $E[y^2]=1$. Recall that $y=\lambda(\mu_a+\sigma_a z_a)+(1-\lambda)(\mu_b+\sigma_b z_b)$ where λ is a Bernoulli random variable with probability p that $\lambda=1$ and z_a, z_b are independent standard normal variates. Note that $\lambda(1-\lambda)=0$ and $\lambda^2=\lambda^3=\lambda^4=\lambda$. Taking the cube, dropping all terms which include $\lambda(1-\lambda)$, and simplifying powers of λ and $(1-\lambda)$:

$$\begin{aligned} y^3 = (\lambda(\mu_a+\sigma_a z_a) + (1-\lambda)(\mu_b+\sigma_b z_b))^3 = \\ \lambda(\mu_a+\sigma_a z_a)^3 + (1-\lambda)(\mu_b+\sigma_b z_b)^3. \end{aligned} \tag{A22}$$

The third moment of a normal variate with mean μ and standard deviation σ is $\mu^3+3\mu\sigma^2$. Taking the expectation of (A22) gives:

$$E[y^3] = p\mu_a^3 + (1-p)\mu_b^3 + 3(p\mu_1\sigma_a^2 + (1-p)\mu_b\sigma_b^2).$$

For the kurtosis of y, take the fourth power of y, drop all terms which include $\lambda(1-\lambda)$, and simplify powers of λ and $(1-\lambda)$:

$$\begin{aligned} y^4 = (\lambda(\mu_a+\sigma_a z_a) + (1-\lambda)(\mu_b+\sigma_b z_b))^4 = \\ \lambda(\mu_a+\sigma_a z_a)^4 + (1-\lambda)(\mu_b+\sigma_b z_b)^4. \end{aligned} \tag{A23}$$

The fourth moment of a normal variate with mean μ and standard deviation σ is $\mu^4+6\mu^2\sigma^2+3\sigma^4$. Taking the expectation of (A23) gives:

$$E[y^4] = p\mu_a^4 + (1-p)\mu_b^4 + 6(p\mu_a^2\sigma_a^2 + (1-p)\mu_b^2\sigma_b^2) + 3(p\sigma_a^4 + (1-p)\sigma_b^4).$$

Finally, we derive the skewness and kurtosis of m_{n0}, which is an average of n_0 independent observations of y.Finding the second, third and fourth moments:

$$\begin{aligned} E[(m_{n0})^2] &= \frac{1}{n_0}E[y^2] = \frac{1}{n_0} \\ E[(m_{n0})^3] &= (\frac{1}{n_0})^3(n_0 E[y^3]) = (\frac{1}{n_0})^2 E[y^3] \\ E[(m_{n0})^4] &= (\frac{1}{n_0})^4(n_0(n_0-1)3E[y^2] + n_0 E[y^4]). \end{aligned}$$

Taking ratios as in (A8) and (A9) to get skewness and kurtosis:

$$\begin{aligned} Skew[m_{n0}] &= (\frac{1}{n_0})^{\frac{1}{2}} E[y^3] \\ Kurt[m_{n0}] &= 3 + (\frac{1}{n_0})(E[y^4]-3). \end{aligned}$$

8 Years-of-education data description

The first two columns of Table A-1 below reproduce two rows from *Table 1: Educational Attainment of the Population 18 Years and Over, by Age, Sex, Race, and Hispanic Origin: 2015* in Current Population Survey Data on Educational Attainment (U.S. Census Bureau (2015)). We choose the subsample "U.S. white males ages 25 and greater" from that data source, which is row 25 of their *Table 1.* The [white/male/age 25 and over] subsample has 81,913 observations. Row three of Table A-1 below transforms the qualitative categories into a quantitative variable. There are a few minor subjective judgements in transforming the survey categories into quantitative years-of-education. The final column shows the frequency distribution of the data.

Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. mixtools: An R package for analyzing finite mixture models. Journal of Statistical Software, 32(6):1-29, 2009.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analyses. URL: http://pngu.mgh.harvard.edu/purcell/plink/.

Table A-1

Years of education of U.S. white males, ages 25+

Completed years of education, survey categories	Number of survey respondents in each category	Quantitative measure for years-of-education	Frequency distribution of the quantitative variable
None	293	0	0.36%
1st – 4th grade	703	4	0.86%
5th – 6th grade	1,487	6	1.82%
7th – 8th grade	1,558	8	1.90%
9th grade	1,397	9	1.71%
10th grade	1,445	10	1.76%
11th grade	2,704	11	3.30%
High school graduate	24,812	12	30.29%
Some college, no degree	13,316	13	16.26%
Associate’s degree, occupational	3,478	14	4.25%
Associate's degree, academic	4,001	15	4.88%
Bachelor’s degree	16,882	16	20.61%
Master’s degree	6,475	18	7.90%
Professional degree	1,705	19	2.08%
PhD graduate	1,657	20	2.02%

Appendix B: Guide to the gwrpv program in the R package gwrpvr

Gregory Connor
Department of Economics, Finance and Accounting
Maynooth University

Michael O'Neill
School of Business
University College, Dublin

October 17, 2017

1 Introduction

The Genome-Wide Regression P-Value (gwrpv) program computes the sample probability value (p-value) for the estimated coefficient from a standard genome-wide univariate regression. It computes the exact finite-sample p-value under the assumption that the measured phenotype (the dependent variable in the regression) has a known Bernoulli-normal mixture distribution. This appendix provides instructions for using the gwrpv program contained in the R package gwrpvr version 1.0.

2 Preliminary Research Steps

The gwrpv program only has added value if the phenotypic variable in the genome-wide regression study does not follow a normal distribution. Before running the gwrpv program, the analyst should compute the sample skewness and kurtosis of the phenotypic variable, and apply the Jarque-Bera test for normality. The Jarque-Beta test is provided in most statistical software packages; for example, *ajb.norm.test* in the normtest package in R (see ftp://

`cran.r-project.org/pub/R/web/packages/normtest/normtest.pdf`) gives a standard implementation. If normality is not rejected by the Jarque-Bera test, the `gwrpv` routine is not appropriate; conventional normality-based p-values should be used instead.

In order to run the program, the analyst must fit a normal mixture distribution to the phenotypic variable. This can be done for example using the `normalmixEM` command in the mixtools library in R (see Benaglia et al. (2009)). The `gwrpv` program requires the six parameters of the fitted mixture distribution, $\mu_a, \sigma_a, \mu_b, \sigma_b$, p_a and p_b. The parameters μ_a, σ_a are the mean and standard deviation given that the Bernoulli random variable equals one, and μ_b, σ_b are the mean and standard deviation given that the Bernoulli random variable equals zero. The parameter p_a is the probability that the Bernoulli random variable equals one and p_b is the probability that the Bernoulli random variable equals zero.

Before running the `gwrpv` program, the analyst must run a standard set of genome-wide regressions. That is, for each single-nucleotide polymorphism (SNP) in a large data set, the analyst has regressed a phenotypic variable y, observed over n individuals, on the realized SNP values across the individuals. The univariate regression has been run separately for each SNP in the data set, resulting in a very large number (up to tens of millions) of individual regression coefficients, one per SNP. This can be done using the plink package of routines (see Purcell, et al. (2007)). The `gwrpv` program can use as input the regression coefficient outputs of plink.

The `gwrpv` program allows the analyst to find p-values for any reasonable number of regression coefficients. The number of candidate regression coefficients should be a relatively small number, roughly in the range one to one thousand. It is not appropriate to compute adjusted p-values on the entire universe of tens of millions of regression coefficients, due to the processing-time demands of computing each exact finite-sample p-value. The candidate coefficients should be preselected by the analyst, based for example on the magnitude of their t-statistics, as the ones most likely to have statistically significant p-values. A sensible cutoff is to restrict the analysis to regression coefficients with a t-statistic greater than 3.5 in magnitude. For convenience we have also provided a batch version of the program (`gwrpv_batch`), which can process the results from a set of regressions.

3 Inputs and Outputs

The gwrpv program requires a number of inputs. The first required set of inputs are the six values (μ_a, σ_a, μ_b, σ_b, p_a, p_b). The inputs μ_a and μ_b can have any real values; σ_a and σ_b must be positive; p_a and p_b both must lie between zero and one, and must sum to one.

There are also a set of trimming parameters for efficient computation of the p-values (see Section 5 below), each of these is set to default values, which can be overridden.

The second required set of inputs are four values[1]. The first is the set of candidate regression coefficient estimates for the regressions being analyzed. The second, third, and fourth values must give the number of major heterozygote, homozygote, and minor heterozygote observations from each of the corresponding regression samples. The three number-of-observations columns must all be non-negative integers; the beta estimates can have any real values.

The gwrpv program outputs the p-value of the regression coefficient based on the mixture distribution and some supplementary statistics, including the skewness and kurtosis of the coefficient estimate, based on the assumed mixture distribution for the dependent variable.

4 Procedure for Running the Program

This short section lists the steps needed to run the program. The gwrpv program is contained in the R package gwrpvr, which is available as standard from CRAN. The current version of the program is v1.0.

In your favourite R environment install the package gwrpvr.

```
> install.packages("gwrpvr")
> library(gwrpvr)
```

To see the associated help file for the gwrpv program with the required parameters run

```
> help("gwrpv")
```

[1] There is a batch version of the function gwrpv_batch, to which a list of candidate regression coefficients whose adjusted p-values are to be computed are passed.

The following is an example of how to use `gwrpv`, having initially calculated the inputs as per Sections 2 and 3.

```
> beta <- 6.05879 # candidate regression cooefficient estimate
> n0 <- 499  # number of major heterozygote observations
> n1 <- 1    # number of major homozygote observations
> n2 <- 0    # number of minor heterozygote observations
> mua <- 13.87226 # mean of the mixture distribution given that the Bernoulli random variable equals zero
> siga <- 2.58807 # stdev of the mixture distribution given that the Bernoulli random variable equals zero
> mub <- 4.62829  # mean of the mixture distribution given that the Bernoulli random variable equals one
> sigb <- 2.51803 # stdev of the mixture distribution given that the Bernoulli random variable equals one
> pa <- 0.96544  # pa is the probability that the Bernoulli random variable equals one
> pb <- 0.03456  # 1 - pa
> g <- gwrpv(beta, n0, n1, n2, mua, siga, mub, sigb, pa, pb)
> g$pvalue # display the p-value
> g # display all the output statistics
```

4.1 The skewness and kurtosis of the estimated coefficient

In addition to computing the exact p-value, the `gwrpv` programme computes the skewness and kurtosis of the estimated regression coefficient, under the assumed mixture distribution of the dependent variable, and assuming the null hypothesis that the true coefficient equals zero.

Following on from the sample input parameters in the earlier example we can retrieve the output skewness and kurtosis as follows:

```
> g <- gwrpv(beta, n0, n1, n2, mua, siga, mub, sigb, pa, pb)
> g$pvalue # display the p-value
> g$skew # display the skewness
> g$kurt # display the kurtosis
```

5 Efficient Computation Procedures

If applied naively, the p-value computation in the model requires a sum over a potentially very large number of terms. The `gwrpv` programme uses efficient computation procedures to minimize run-time, while maintaining a high degree of accuracy in the p-value computation. There are three parameter inputs which control these computation features in the programme: *logdelta*, *lognearnorm* and *logtopsum*. All three parameters are real numbers inputted in log base ten format.

The parameter *logdelta* controls the trimming of the three univariate binomial distributions, each binomial distribution corresponding to the sample of one allele type (see the Technical Appendix for analytical details). *logdelta* has a default value of -16, which means that the p-value computation is only accurate for 16 decimal places. The analyst can increase or decrease *delta* by changing the optional parameter value; a larger-magnitude negative delta will result in a more accurate computation and slower run-time. *logdelta* must be a negative number, in log base ten format, so that $10^{logdelta}$ is less than one.

The parameter *lognearnorm* is used to determine whether to use a triple computation loop over $n0$, $n1$ and $n2$ observations, or apply the central limit theorem to approximate the distribution of the sample average phenotype of the major homozygote observations. This approximation eliminates the need to loop over $n0$, giving a double loop over $n1$ and $n2$ only. The program computes the skewness and kurtosis of the sample average phenotype of the major homozygote observations. If the sum of squared difference of skewness and kurtosis from their normal distribution values is less than nearnorm, then the `gwrpv` program uses the central limit theorem to eliminate the major homozygote observations from the computation loop.

The final parameter, *logtopsum*, ensures that the `gwrpv` programme does not spend too long computing a p-value. Before running the main computation sum for each regression case $(\widehat{\beta}, n_0, n_1, n_2)$, the program computes the number of terms in this computation sum. If this number is greater than *logtopsum*, the program skips the p-value computation for that $\widehat{\beta}$ and proceeds to the next regression case, and a value of -999.9 is inserted in place of the p-value in the output to indicate that the computation has been skipped.

5.1 Inputting the control parameters for efficient computation

In the `gwrpv` program we provide parameters to faciliate efficient computation. These are the control parameters described above, the three real numbers, *logdelta*, *lognearnorm* and *logtopsum*. All three of these inputs must be provided in log base ten, so that $(-16, -5, 8)$ means that δ is set at 10^{-16} and *topsum* is 10^8.

By default we set values for these three parameters (i.e., logdelta=-16, lognearnorm=-5, and logtopsum=8). These do not need to be explicitly passed in unless you want to override them. For example, the following will result in the same output as the earlier example. Here we are explicitly setting the trimming parameters..

```
> g <- gwrpv(beta, n0, n1, n2, mua, siga, mub, sigb, pa, pb, logdelta=-12, lognearnorm=-5, logtopsum=8)
> g$pvalue
```

6 Batch mode

If one wishes to compute p-values for mutiple regressions there is a batch version of the function, gwrpv_batch. The following are examples of its use. They illustrate how the results of each regression are presented to gwrpv_batch as a list of lists.

```
# create a list of the beta's
> beta <- c(6.05879, -6.05879, 2.72055, -2.72055, 1.93347, -1.93347, 0.88288, -0.88288, 4.28421, -4.28421)

# create a list of the number of major heterozygote observations
> n0 <- c(499, 499, 495, 495, 490, 490, 451, 451, 998, 998)

# create a list of the number of major homozygote observations
> n1 <- c(1, 1, 5, 5, 10, 10, 48, 48, 2, 2)

# create a list of the number of minor heterozygote observations
> n2 <- c(0, 0, 0, 0, 0, 0, 1, 1, 0, 0)

# create the list of lists
> myregresults <- list(beta = beta, n0 = n0, n1 = n1, n2 = n2)

> g <- gwrpv_batch(myregresults,13.87226,2.58807,4.62829,2.51803,0.96544,0.03456)
```

In the second example we illustrate how to load the regression results from a comma separated file. Connor & O'Neill (2017) describe an illustrative sample data set of regressions to which the Genome-Wide Regression P-Value method is applied. The R package gwrpvr contains a folder called data/ in which this data set is provided. The data is contained in the file named regresults.csv

```
# alternatively the regression results may be contained in a .csv file
# let's call the file "regresults.csv"
# assuming four comma separated columns with a single header line
# containing columns names: beta,n0,n1,n2

# readr is a handy package to read in csv files, install this if not in your environment
> install.packages('readr')

# load the readr package into your environment
> library(readr)

# use the read_csv function from the readr package to load in the csv
> myregresults <- read_csv("data/regresults.csv")
> g <- gwrpv_batch(myregresults,13.87226,2.58807,4.62829,2.51803,0.96544,0.03456)
```

References

Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. mixtools: An R package for analyzing finite mixture models. Journal of Statistical Software, 32(6):1-29, 2009.

Connor, G., O'Neill, M. (2017). Finite-sample genome-wide regression p-values (GWRPV) with a non-normally distributed phenotype.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis/ American Journal of Human Genetics, 81 http://pngu.mgh.harvard.edu/purcell/plink/.

R Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org