Software for genome-wide association studies having multivariate responses: Introducing MAGWAS

Chad C. Brown[1] and Alison A. Motsinger-Reif[1,2]

[1]Department of Statistics, [2]Bioinformatics Research Center
North Carolina State University, Raleigh NC 27695

February 16, 2012
North Carolina State University Department of Statistics Technical Reports # 2641

*Address for Correspondence:

Chad C. Brown
Department of Statistics
North Carolina State University,
Raleigh NC, 27695-7566, USA
EMAIL: ccbrown2@ncsu.edu

Alison A. Motsinger-Reif, Ph.D.
Bioinformatics Research Center
North Carolina State University
Campus Box 7566
Raleigh, NC 27695-7566
EMAIL: motsinger@stat.ncsu.edu

Introduction

Continuing advances in genotyping technology have dramatically increased the amount of data available for genome-wide association studies (GWAS). In many of these studies, multiple distinct responses are observed for each individual. These responses could be measurements over time or space, or could represent related but non-identical outcomes (such as responses to various drug concentrations). One widely used method for analyzing such data are linear models adapted for multivariate responses, such as multivariate analysis of covariance. Currently, most developed statistical software packages offer analysis tools for linear models with multivariate responses. However, there is still a need for a software package that is free, fast, simple to use and accepts industry standard file formats for GWASs. MAGWAS was developed to address these needs.

MAGWAS stands for Multivariate Ancova Genome-Wide Association Software. Its purpose, as the name indicates, is to provide analysis tools for assocation studies of single nucleotide polymorphisms (SNP) in genetic data having multivariate responses and, possibly, multiple covariates. The software is quick to download, requires no installation, is easy to use, makes use of standard file types, and is computationally fast. MAGWAS tests for the the significance of each locus in a GWAS, where each individual has a vector of related outcomes. For example, each individual could have a response to a medication at three fixed time points, or each tumor cell line could give a response to a medication at four different concentrations. MAGWAS models the vector of responses for each individual jointly using a well-established multivariate analysis of covariance design.

MAGWAS can be downloaded at:
`http://www4.stat.ncsu.edu/~motsinger/Lab_Website/Software.html`

Implementation

In multivariate regression with a multivariate (vector) response, we have the model [Timm, 2002]:

$$\mathbf{Y} \quad = \quad \mathbf{X}\boldsymbol{\beta} + \mathbf{E}, \tag{1}$$

where $\mathbf{Y}$ is an $n \times p$ matrix of responses, $\mathbf{X}$ is an $n \times q$ matrix of covariates, $\boldsymbol{\beta}$ is a $q \times p$ matrix of regression parameters and $\mathbf{E}$ is an $n \times p$ matrix of residuals, where,

$$\mathbf{E} \sim \mathrm{N}_{n,p}(0, \Sigma, \mathrm{I}_n).$$

Here, $\mathrm{N}_{n,p}(0, \Sigma, \mathrm{I}_n)$ is a matrix normal distribution with covariance parameters Σ and I_n. Assuming $\mathbf{X}$ is full rank, it can be shown [Timm, 2002], that:

$$\widehat{\boldsymbol{\beta}}_{ML} \quad = \quad (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y},$$

meaning that the MLE is the same as the OLS parameter estimates from regression on each univariate response independently.

In an analysis of variance model, the total sums of squares within-group is compared to the

total sums of squares between groups. In GWASs, groups are often taken to be membership of each individual to one of three genotypes at each SNP. Mathematically, identical test statistics can also be created by comparing the residual variance of the full model to the residual variance for a reduced model that does not contain genetic (group) effects. Test statistics are computed by comparison of full and reduced models in order to increase computational efficiency.

Likewise, when more than one response is observed for each individual, the test statistics are created by comparing the covariance of the residuals for a full model, and a reduced model that does not contain genetic effects. Specifically, test statistics are created from the eigenvalues of the sums of squares for the residuals (proportional to covariance) for full and reduced models:

$$eigenvalues((\widehat{E}_{Red}^T\widehat{E}_{Red} - \widehat{E}_{Full}^T\widehat{E}_{Full})(\widehat{E}_{Full}^T\widehat{E}_{Full})^{-1}),$$

where $\widehat{E}_{Full}$ and $\widehat{E}_{Red}$ represent the error terms from fitting full and reduced models. Four standard test statistics are commonly reported for multivariate regression models: Wilks' lambda, Pillai's Trace, Hotelling-Lawley Trace and Roy's Maximum Root. The last of these often gives infated type I errors, and is not reported by MAGWAS. Approximations have been made for the first three test statistics and are described in [SAS Institute Inc., 2004]. The significance of these test statistics are given in the output under the column headings "Pillai" "Wilks" and "HotellingLawley", respectively.

MAGWAS is written completely in Java and will soon be registered as open source software under a GNU public license. Matrix manipulations are accomplished using the Java package JAMA [Hicklin et al., 2011] and distributional calculations are made using the Java package DistLib [Steinmetz et al., 2007].

Advantages of multivariate analysis

When each individual has a vector of related responses, one approach would be to analyze each response across individuals one at a time. This procedure is time consuming, may give contradictory inference at different response levels, and may not capture the full array of potential difference between genotypes for some loci. Looking at the joint distribution of responses allows for the analyst to detect differences between groups that might not otherwise be possible. For example, suppose that each individual had two responses (say R1 and R2) that had high positive correlation. Looking at the differences between R1 or R2 for each genotype group individually (see Figures 1 and 2) may not be as informative as looking at the joint response of R1 and R2 jointly (see Figure 3). These plots were created by simulating (R1, R2) from a multivariate normal distribution using the "mvtnorm" package in R [R Development Core Team, 2011].

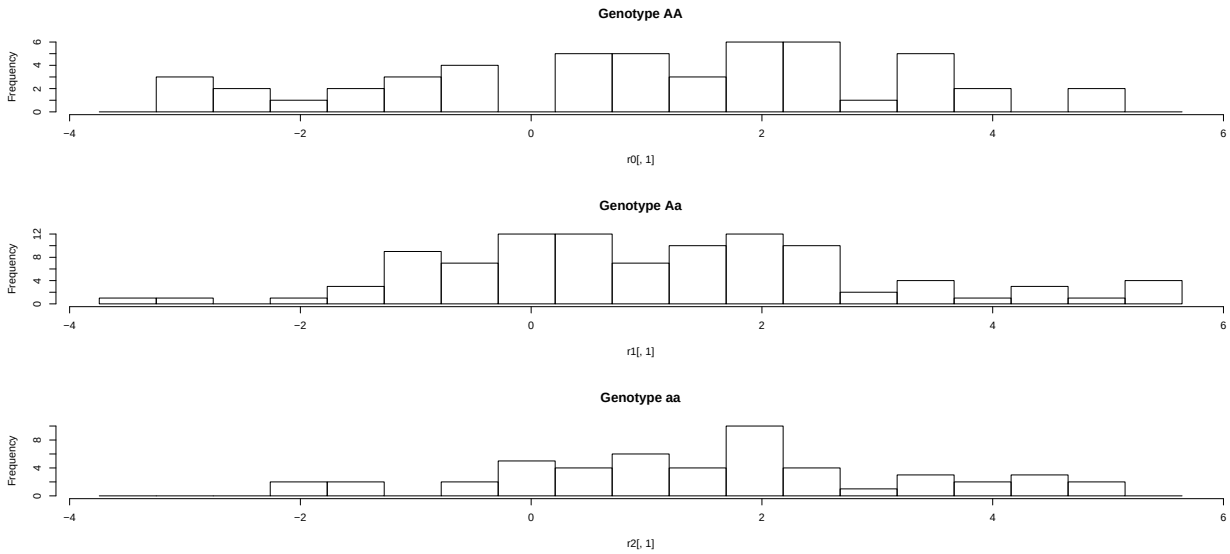

Figure 1: Overlapping histograms for response R1 for each genotype. No significant association between genotype and R1 (p-value = 0.26).

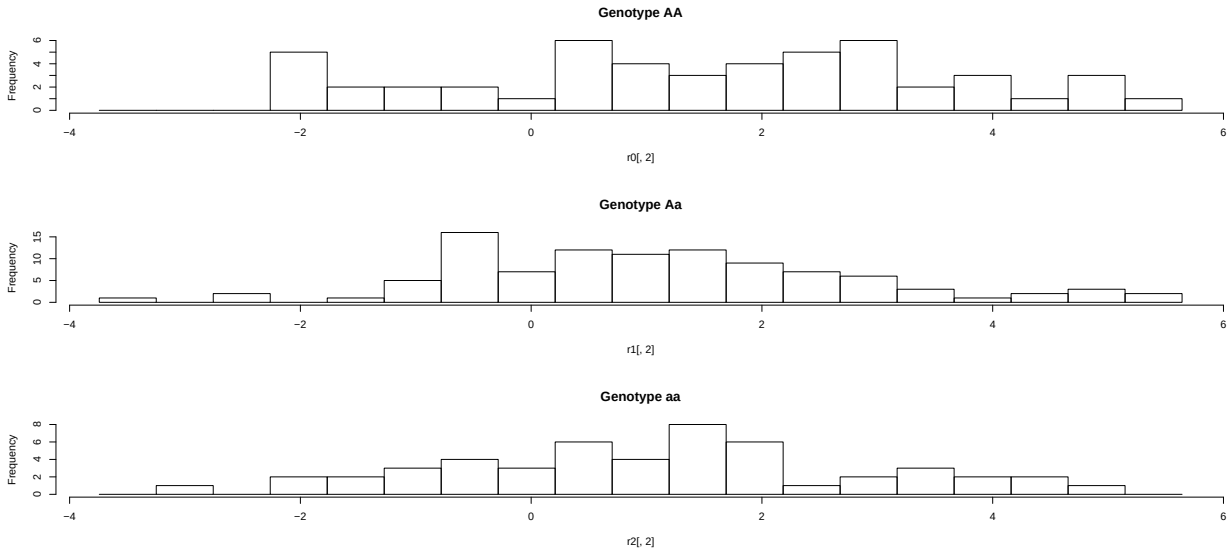

Figure 2: Overlapping histograms for response R2 for each genotype. No significant association between genotype and R2 (p-value = 0.4).

Data Formats

Genetic data must be in PLINK format, as "tfam" and "tped" files [Purcell et al., 2007; Purcell, 2009]. Responses and covariates must be contained in a separate space delimited (no tabs, semicolons, etc.) file. This data file must have a first "header" line, giving variable names (no spaces, tabs, etc in names). The first column of the data file must be a column of IDs that must also appear in tfam file as "Individual ID" (second column in .tfam file). However, not every .tfam individual ID needs to appear in the data file. The tped file must end in ".tped", but the extensions for the other files are not important.

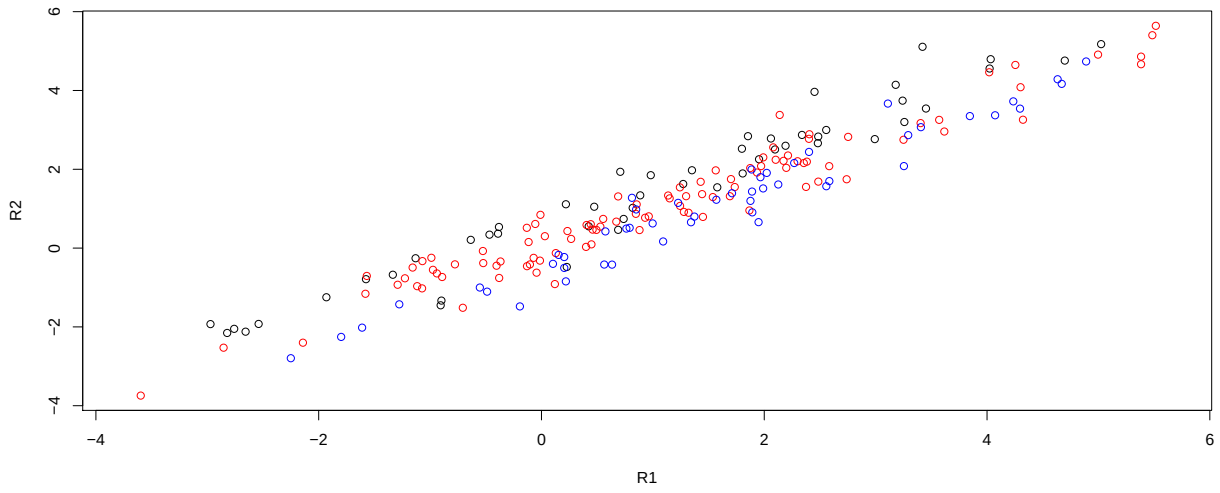

Figure 3: Scatterplot of the joint (R1, R2) response for each genotype. Black is genotype AA, red is genotype Aa and blue is genotype aa. Highly significant association between genotype and the joint response (p-value = 4.9e-17).

All variables names to be coded as the response must begin with an "R", and response variables must appear as unseparated columns immediately after the first ID column. All subsequent columns must have names that do not begin with an "R" and will be used as covariates. Entries in the ID column may contain any alpha numeric characters (no spaces) and each entry is expected to exactly match one entry in the second ("Individual ID") column in the referenced .tfam file.

All entries other than those in the ID column in the data file must contain only numbers (using "." as a decimal). Missing values, or nominal variables are not accepted at this time. However, several software packages exist to impute missing values and nominal variables can be recoded using "dummy" indicator variables (for example 1.0 for male and 0.0 for female, or vice versa).

A column of "ones" is automatically added to the covariates to account for an intercept. Therefore, the data file should not contain an intercept column. In fact, the rank of the covariates must equal the number of covariates, which in turn, must be less than the number of individuals in the data file.

Tables 1 and 2 below demonstrate what is not acceptable for a data file and how to make corrections.

	ID	First Response	Second Response	Age	Race	Treat1	Treat2
	1	1.53	-1.13	41	CHB	1	0
	2	-1.88	-0.968	25	CHD	1	0
	3	-0.747	-0.875	34	CHD	0	1
	4	0.154	0.132	19	CHB	0	1
	5	0.433	0.198	45	CHB	0	1

Table 1: Example of an unacceptable data file format. Neither responses begin with an "R" and both contain spaces. The covariate "Race" begins with an "R", and its entries contain letters. Finally, the covariates "Treat1" and "Treat2", when added together make an intercept term (giving a non-full rank covariate matrix after an intercept is added).

	ID	Response1	Response2	Age	Ancestry	Treat1
	1	1.53	-1.13	41	0	1
	2	-1.88	-0.968	25	1	1
	3	-0.747	-0.875	34	1	0
	4	0.154	0.132	19	0	0
	5	0.433	0.198	45	0	0

Table 2: The same data as above, made into an acceptable form. Notice that "Treat2" is eliminated, giving a full-rank covariate matrix.

How to use MAGWAS

To run MAGWAS, first download "MAGWAS.zip" into a new directory, and unzip to generate a set of files that include (among others) "MAGWAS.jar", "DistLib.jar" and "Jama-1.0.2.jar". These files should now be in the same directory. Navigate to this directory using a command line tool, such as Terminal on a Mac or MS-DOS in Windows. To run MAGWAS , type "java -jar MAGWAS.jar" followed by four space delimited arguments specifying, in order, the full paths, filenames and extensions for the .tfam, .tped, covariate file and file to which results will be written. For example, the command:

```
java -jar MAGWAS.jar /Research/tfamA.tfam /Research/tpedA.tped /Research/data.txt
/Research/results.txt
```

will read in genetic data from the files "tfamA.tfam" and "tpedA.tped", as well as the data file "data.txt", all located in the folder "/Research/". Results will be written to "/Research/results.txt". Do not use shortcuts for the file paths (such as "./" in Terminal).

For your convenience, a fictitious set of testing data files are included in the distribution. The file with covariates is labeled "testData.txt", and the associated .tfam and .tped files are "test.tfam" and "test.tped", respectively. Using MAGWAS to analyze this data will produce a results file with the first few lines as reading:

```
ChrNum RSID GenDis bpPos nAA nAa naa Pillai Wilks HotellingLawley
1 RS7908684599 0 643117 399 77 4 0.03750095617513627 0.03727448577359227 0.0368304471825307
1 RS4368828077 0 1008450 113 239 127 0.4531835033476319 0.4533808543510326 0.4528247186364942
1 RS8944428831 0 1988692 98 234 148 0.7635110264907106 0.7640341065373445 0.7640816962151011
1 RS4472219575 0 1998984 333 133 14 0.8560654489761544 0.8565427846356581 0.856700587854669
1 RS8052434382 0 2102803 58 222 200 0.19070332467699802 0.19021417271522456 0.18911665763324426
1 RS3406445063 0 2247755 270 179 31 0.11280860015481808 0.11329185564664557 0.11330916879662312
```

Alternatively, MAGWAS can run analysis for a batch of tped files (for example, one for each chromosome). To do this, place all the tped files in the same directory and specify this directory where the tped (individual) file would normally be placed. In this directory, all (and only) tped files must end in ".tped". Results for each tped file will be written as a separate file. These files will be placed in the directory (plus an optional the first part of the filename), specified as the last command line argument (where the individual results file normally goes). In the example above, if the files "Chr1.tped" and "Chr2.tped" were located in the directory "/Research/tpedFiles/", they could be analyzed using the command:

```
java -jar MAGWAS.jar /Research/tfamA.tfam /Research/tpedFiles/ /Research/data.txt
/Research/resultsFolder/myResults
```

The output would be written in the folder "/Research/resultsFolder/", and the file names would be "myResultsChr1.txt" and "myResultsChr2.txt".

Run times

The time to complete a MAGWAS analysis depends heavily on the number of individuals in the study, the number of SNPs evaluated, the number of responses for each individual and the number of covariates in the model (not to mention the computational resources). However, an analysis of a real data set with 486 individuals having two responses and 4 covariates each across about 2 million SNPs took approximately 7 minutes on a modest computing cluster (two Intel(R) Xeon(R) CPU E5450 processors). I estimate that the same data set on a typical newer personal machine would take 10-15 minutes. The time to run the test data set should take about a second or less on any system.

Interpretation

After running MAGWAS, results will be written to the file specified on the command line. The first line is a header, giving a brief description for each column. The first three columns give the chromosome, rs number, genetic distance and base pair position (labeled ChrNum, RSID, GenDis and bpPos respectively) from the .tped file. The next three columns contain the number of individuals having each of the three possible genotypes (labeled nAA, nAa and naa), in alphabetical order by nucleotide (for example, a GT polymorphism will have genotype frequencies in the order GG, GT and TT). The last three lines give the levels of significance (p-values) for SNP using three common test statistics used in multivariate analysis of covariance. These are described in more detail below. If any locus has only two observed genotypes (for instance a AC polymorphism with only AA and AC genotypes observed), test statistics are not calculated, but instead a flag of "-3.0" is given. Similarly, if any locus has only one observed genotype, a flag of "-2.0" will be given.

It is important to note p values are based on asymptotic distributions under the null. This means that if the genotype frequencies for any group are small, results may not be valid. This is the reason that genotype frequencies are given as part of the output. Although it depends heavily on the specific problem, a total sample size of at least 250, with at least 20 in each genotype category seems to work for the problems that I have encountered.

Acknowledgments

This work was supported by NIH NCI R01CA161608 and T32GM081057 from the National Institute of General Medical Sciences and the National Institute of Health. We would like to thank Kevin Long for testing the software.

References

Hicklin, J., Moler, C., Webb, P., Boisvert, R. F., Miller, B., Pozo, R. & Remington, K. (2011). JAMA v1.0.2. http://math.nist.gov/javanumerics/jama/.

Purcell, S. (2009). PLINK v1.07. http://pngu.mgh.harvard.edu/purcell/plink/.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., Maller, J., Sklar, P., De Bakker, P., Daly, M. et al. (2007). The American Journal of Human Genetics 81, 559–575.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria. ISBN 3-900051-07-0.

SAS Institute Inc. (2004). SAS User's Guide (Release 9.2): Statistics SAS Inst. SAS Cary, NC.

Steinmetz, P. N., Warnes, G. & Warnes, J. (2007). DistLib v0.9.1. http://statdistlib.sourceforge.net/.

Timm, N. (2002). Applied multivariate analysis. Springer Verlag.