R^2 STATISTICS WITH APPLICATION TO ASSOCIATION MAPPING

by

GUANNAN SUN

B.S., CHINA UNIVERSITY OF GEOSCIENCES (BEIJING), 2006

A REPORT

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts And Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2008

Approved by:

Major Professor
Yang, Shie-Shien

Copyright

Abstract

In fitting linear models, R^2 statistic has been wildly used as one of the measures to assess the goodness-of-fit and prediction power of the model. Unlike fixed linear models, at this time there is no single universally accepted measure for assessing goodness-of-fit and prediction power of a linear mixed model. In this report, we reviewed seven different approaches proposed to define a measure analogous to the usual R^2 statistic for assessing mixed models. One of seven statistics, R_C, has both conditional and marginal versions. Association mapping is an efficient way to link the genotype data with the phenotype diversity. When applying the R^2 statistic to the association mapping application, it can determine how well genetic polymorphisms, which are the explanatory variables in the mixed models, explain the phenotypic variation, which is the dependent variation. A linear mixed model method recently has been developed to control the spurious associations due to population structure and relative kinship among individuals of an association mapping. We assess seven definitions of R^2 statistic for the linear mixed model using data from two empirical association mapping samples: a sample with 277 diverse maize inbred lines and a global sample of 95 Arabidopsis thaliana accessions using the new method. R^2_{LR} statistic derived from the log-likelihood principle follows all the criterions of R^2 statistic and can be used to understand the overlap between population structure and relative kinship in controlling for sample relatedness. From our results, R^2_{LR} statistic is an appropriate R^2 statistic for comparing models with different fixed and random variables. Therefore, we recommend using R^2_{LR} statistic for linear mixed models in association mapping.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Thank for the help and support from Dr. Yang, Shie-Shien who is my major advisor, Dr. Yu, Jianming who is my supervisor in Sorghum Genetics Laboratory and my committee member, and Dr. Song, weixing who is my committee member to my research.

I am very grateful to Dr. Yu, Jianming's research group in Sorghum Genetics Laboratory in the department of Agronomy. This project is supported by the National Research Initiative (NRI) Plant Genome Program of the USDA-CSREES.

I appreciate the department of Statistics and the department of Agronomy.

My parents Banjun Sun and Xiaoqun Dong, and my boyfriend Guorong Chen are always giving me supports and encouragements. And thank all my teachers and friend

CHAPTER 1 - Introduction

Linear regression models are wildly used in the field of statistics and by researchers in many disciplines. The R^2 statistic, the coefficient of determination is one of the most wildly used measure of prediction power and goodness-of-fit of linear regression models. Recently, many R^2-like statistics were proposed in the statistical literature to measure the prediction power and goodness-of-fit of nonstandard linear regression models, such as generalized linear models and mixed effect models.

In this report, we are mainly interested in ways to measure the prediction power and goodness-of-fit of linear mixed effect models. Unlike R^2 statistic in linear regression models, there are no single universally accepted R^2-like statistic to measure the prediction power and goodness-of-fit of linear mixed effect models. This report will review seven R^2-like statistics proposed in the statistical literatures in the context of mixed effect models.

Recently, mixed effect models have been used in genetic research. Association mapping studies the association between a particular gene and disease susceptibility based on populations. It provides a powerful complement to the previous linkage analysis for figuring out the genetic basis of the complex traits. Yu et al. (2006) proposed a new improved association mapping method accounting for both population structure and relative kinship to complement the current available methods for association mapping. This new approach involves comparing several nested mixed effect models based on the -2log-likelihood and the Bayesian Information Criterion (BIC). These two criterions are generally for model selection and identification but not for measuring the goodness-of-fit or prediction power of the model. Therefore, R^2-like statistics for mixed models may be more appropriate for the study of association mappings. In addition, it would be desirable to many researchers if R^2 statistic can be developed for the mixed model to assess the effect of a gene as the amount of phenotypic variation explained.

This report applied the seven R^2-like statistics to the association mapping procedure in two large empirical data sets. The construction of these seven R^2-like statistics was based on different assumptions and principles. Therefore the results obtained from the analysis of these

two data sets would be different. We used the results to illustrate and compare the utility of these seven R^2-like statistics.

The remaining chapters of the report are organized as follow. In Chapter 2 of this report, we reviewed the traditional R^2 statistic for the standard linear regression model with fixed effects and seven R^2-like statistics for nonstandard linear models with different assumptions, and relations between them were also showed. In Chapter 3, we first reviewed the basic background knowledge about the association mapping and linked it to the applied area of statistics. Second we described the mixed-effect models which we used in the empirical analysis on association mapping. Last, we applied the seven R^2-like statistics to two data sets and we assessed the performance of these seven statistics to find the best one for application to association mapping. We also showed the overlap between population structure and relative kinship in the way of Venn Diagrams. In Chapter 4, we discussed more about the different R^2-like statistics including the simulation studies done by Xu (2003) and Orelien (2008). We pointed out a mistake found in Orelien's (2008) paper as well.

CHAPTER 2 - The R^2-like Statistics

§2.1 R^2 for Fixed Linear Models

For the linear model with only fixed effects:

$$y = X\beta + u \tag{1}$$

where y is a $n \times 1$ vector, X is a $n \times k$ matrix, β is a $k \times 1$ vector of unknown regression coefficients, and u is a $n \times 1$ vector consisting of i.i.d. normal variables with mean 0 and variance σ^2. Then the usual R^2 statistic is defined as

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

where $SSR = (\hat{Y} - \overline{Y})'(\hat{Y} - \overline{Y})$, $SSE = (Y - \hat{Y})'(Y - \hat{Y})$, and $SSTO = (Y - \overline{Y})'(Y - \overline{Y})$.

Since $0 \le SSE \le SSTO$, it follows that:

$$0 \le R^2 \le 1.$$

We may interpret R^2 as the proportion of total variation due to the regression model with explanatory variables X. Thus, the larger the R^2 the larger the proportion of the total variation of Y is explained by the explanatory variables X. R^2 is scale invariant and remains unchanged when the units of Y and X change. It also provides us a simple statistic to summarize the effects of covariates on the response. In the other words, R^2 gives us an easy-to-understand way to assess how well the model fits the data.

For model (1), there are several alternative R^2 statistics. Kvalseth (1985) listed eight of them which are presented below.

$R_1^2 = 1 - \sum (y - \hat{y})^2 / \sum (y - \bar{y})^2$

$R_2^2 = \sum (\hat{y} - \bar{y})^2 / \sum (y - \bar{y})^2$

$R_3^2 = \sum (\hat{y} - \bar{\hat{y}})^2 / \sum (y - \bar{y})^2$

$R_4^2 = 1 - \sum (e - \bar{e})^2 / \sum (y - \bar{y})^2$

R_5^2 = squared multiple correlation coefficient between the regressand and the regressors

R_6^2 = squared correlation coefficient between y and $\hat{y}$

$R_7^2 = 1 - \sum (y - \hat{y})^2 / \sum y^2$

$R_8^2 = \sum \hat{y}^2 / \sum y^2$

It is well known that R^2 is a biased estimator of the population multiple correlation coefficient , ρ^2 . An unbiased estimator of ρ^2 when ρ^2 =0 is

$$R_{adj}^2 = 1 - \frac{MSE}{MST} = 1 - (1 - R^2)(\frac{n-1}{n-k-1}).$$

The adjusted R^2 takes the number of independent variables used in the model into account.

Kvalseth (1985) proposed eight requirements for R^2 Statistics:

1. R^2 must have utility of a goodness of fit measure with reasonable interpretation.

2. R^2 should be dimensionless and independent of the units of measurement.

3. The potential endpoints corresponding to perfect and total lack of fit should be well defined.

4. R^2 should be applicable to any model independent of the statistical properties of the model.

5. R^2 should not be restricted to any specific model-fitting technique.

6. R^2 values for different models fitting the same data should be comparable directly.

7. R^2 is generally compatible with other acceptable measures of fit.

8. Positive and negative residuals should be weighted equally.

Cameron and Windmeijer (1996) proposed four additional properties:

1. R^2 does not decrease as regressors are added.

2. R^2 based on residual sum of squares coincides with R^2 based on explained sum of squares.

3. There is a correspondence between R^2 and a significance test on all slope parameters and between changes in R^2 and significance tests as regressors are added.

4. R^2 has an interpretation in terms of information content of the data.

Under weak conditions $R^2=\frac{\beta' S_x \beta}{\beta' S_x \beta+\sigma^2}$ almost surely as n tends to infinity, where

$S_x=\frac{1}{n-1}(X-1\bar{x})'(X-1\bar{x})$ is the sample covariance matrix for the explanatory variables.

However, $\frac{\beta' S_x \beta}{\beta' S_x \beta+\sigma^2}$ is decided by the researchers since S_x is given by the experimental design. Therefore, the random X matrix should be considered so that R^2 can also be random. Based on the assumption of the random X whose rows are independent with each other and also independent of errors, and each row of X has a multivariate normal distribution with expectation μ_x and covariance matrix Σ_x, we have $\frac{R^2}{1-R^2}=\frac{\chi_k^2(\lambda)}{\chi_{n-k-1}^2}$, where the numerator, a non-central χ^2 distribution with k degree of freedom is independent of the denominator, a central χ^2 distribution with (n-k-1) degree of freedom. The non-centrality parameter of the numerator is $\frac{\beta'\Sigma_x\beta}{\sigma^2}$ which distributes as $\frac{\rho^2}{1-\rho^2}\chi_{n-1}^2$ that depending a new χ^2 distribution with $\rho^2=\frac{\beta'\Sigma_x\beta}{\beta'\Sigma_x\beta+\sigma^2}$. Gurland (1968) show that for large n, $\frac{R^2}{1-R^2}$ can be approximated by $\frac{a\nu}{n-k-1}F_{\nu,n-k-1}=\frac{(n-1)t+k}{n-k-1}F_{\nu,n-k-1}$, where $a=\frac{(n-1)t(t+2)+k}{(n-1)t+k}$, $\nu=\frac{(n-1)t+k}{a}$, and $t=\frac{\rho^2}{1-\rho^2}$. A significance test of R^2 is equivalent to the usual F-test of the significance of the regression model.

§2.2 Mixed Models

The linear model with both fixed effects and random effects is

$$Y=X\beta+Z\gamma+u \tag{2}$$

where Y is a $n\times 1$ observation vector; X is a $n\times k$ design matrix linked to the fixed-effect β, a $k\times 1$ vector of unknown regression coefficients of fixed effects; Z is a $n\times p$ design matrix linked to the random-effects γ, a random $p\times 1$ vector of random effects with zero means and

variance-covariance matrix G; u is a $n \times 1$ random vector with zero means and variance covariance matrix R.

§2.3 R^2 Statistics for Linear Mixed Models

§2.3.1 The R^2_{LR} *Statistic*

Cox and Snell (1989), and Magee (1990) independently proposed a R^2 based on LR statistics:

$$R^2_{LR} = 1 - \exp(-\frac{2}{n}(\log L_M - \log L_0))$$

where $\log L_M$ is the maximum log-likelihood of the model of interest, $\log L_0$ is the maximum log-likelihood of the intercept-only model, n is the number of observations. Maddala (1983) also suggested this statistic for binary response models. LR Statistics can be written as $LR = 2\log(L_M/L_0)$ which asymptotically follows a χ^2 distribution. We have the relationship between R^2_{LR} and LR as $R^2_{LR} = 1 - \exp(\frac{-LR}{n})$.

R^2_{LR} is an appropriate statistic when the concept of residual variance cannot be easily defined and the maximum likelihood is the criterion of fitting the model of interest. It has seven properties as pointed out by Nagelkerke (1991):

1. It's consistent with the traditional R^2 when applied to the linear regression.
2. The maximum likelihood estimates of parameters also maximize R^2_{LR}.
3. R^2_{LR} and n, the sample size are asymptotically independent.
4. R^2_{LR} could be interpreted as the proportion of explained variation and viewed as a measure of the extent to which a distribution is not degenerate.
5. It does not have dimension.
6. Replacing 2/n by k/n in the definition may produce a generalization of the explained proportion of the k^{th} central moment of the model.
7. R^2_{LR} is the square of the Pearson correlation between the fixed effects and the efficient score of the model based on the first order Taylor expansion approximation.

§2.3.2 The R_W^2 *Statistic*

If we reformulate model (2) as a new linear model in the following form

$$Y = X\beta + u \tag{3}$$

where u is an $n \times 1$ vector of disturbances with mean 0 and variance covariance matrix V. In model (3), u is just the combination of the random effects and errors of model (2). Thus the variance covariance matrix of Y is $V = Z'GZ + R$.

Buse (1973) derived a modified R^2 based on model (3) as

$$R_W^2 = 1 - \frac{\hat{\mathbf{u}}'\mathbf{V}^{-1}\hat{\mathbf{u}}}{(\mathbf{Y} - \bar{\mathbf{Y}})'\mathbf{V}^{-1}(\mathbf{Y} - \bar{\mathbf{Y}})}$$

where $\hat{\mathbf{u}} = \mathbf{Y} - \hat{\mathbf{Y}}$ with $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ being the best predictor of $\mathbf{Y}$, and $\bar{\mathbf{Y}} = \frac{\mathbf{e}'\mathbf{V}^{-1}\mathbf{Y}}{\mathbf{e}'\mathbf{V}^{-1}\mathbf{e}}$ with $\mathbf{e}' = (1, \ldots, 1)$.

In generalized least square estimation, we have the following partition of total weighted sum of squares and the normal equation for the best linear unbiased estimator $\hat{\beta}$ of β

$$Y'V^{-1}Y = \hat{Y}'V^{-1}\hat{Y} + \hat{u}'V^{-1}\hat{u}$$

$$(X'V^{-1}X)\hat{\beta} = X'V^{-1}Y.$$

By partitioning matrix X into K parts as $X = (X_1 : X_2 : \cdots : X_k)$, the normal equation can be laid out accordingly as

$$\begin{pmatrix} X_1'V^{-1}X_1 & X_1'V^{-1}X_2 & \cdots & X_1'V^{-1}X_k \\ X_2'V^{-1}X_1 & X_2'V^{-1}X_2 & \cdots & X_2'V^{-1}X_k \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ X_k'V^{-1}X_1 & X_k'V^{-1}X_2 & \cdots & X_k'V^{-1}X_k \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = \begin{pmatrix} X_1'V^{-1}Y \\ X_2'V^{-1}Y \\ \vdots \\ X_k'V^{-1}Y \end{pmatrix}.$$

The j^{th} row of the above layout gives

$$(X_j'V^{-1}X_1)\hat{\beta}_1 + (X_j'V^{-1}X_2)\hat{\beta}_2 + \ldots + (X_j'V^{-1}X_k)\hat{\beta}_k = X_j'V^{-1}Y.$$

Pre-multiplying $\hat{Y} = X\hat{\beta}$ by $X_j'V^{-1}$ yields

$$(X_j'V^{-1}X_1)\hat{\beta}_1 + (X_j'V^{-1}X_2)\hat{\beta}_2 + \ldots + (X_j'V^{-1}X_k)\hat{\beta}_k = X_j'V^{-1}\hat{Y},$$

and thus $X_j'V^{-1}(Y - \hat{Y}) = X_j'V^{-1}\hat{u} = 0$. Then we could define

$$\frac{e'V^{-1}\hat{Y}}{e'V^{-1}e} = \frac{e'V^{-1}Y}{e'V^{-1}e} = \bar{Y}$$

which states that the weighted mean of the predicted Y is equal to the weighted mean of the observed Y and the weighted sum of residuals is zero.

If we use the deviations from the weighted means as new variables in the model, then the following equations hold:

$$(\hat{Y}-\overline{Y}e)'V^{-1}(\hat{Y}-\overline{Y}e)=\hat{Y}'V^{-1}\hat{Y}-\frac{(e'V^{-1}Y)^2}{e'V^{-1}e}$$

$$(Y-\overline{Y}e)'V^{-1}(Y-\overline{Y}e)=Y'V^{-1}Y-\frac{(e'V^{-1}Y)^2}{e'V^{-1}e}.$$

Then we can rewritten $Y'V^{-1}Y=\hat{Y}'V^{-1}\hat{Y}+\hat{u}'V^{-1}\hat{u}$ as

$$(Y-\overline{Y}e)'V^{-1}(Y-\overline{Y}e)=(\hat{Y}-\overline{Y}e)'V^{-1}(\hat{Y}-\overline{Y}e)+\hat{u}'V^{-1}\hat{u}.$$

When define SSTO as $(Y-\overline{Y}e)'V^{-1}(Y-\overline{Y}e)$, SSR as $(\hat{Y}-\overline{Y}e)'V^{-1}(\hat{Y}-\overline{Y}e)$, and SSE as $\hat{u}'V^{-1}\hat{u}$ in terms of the weighted sum of squares, the above equation can be rewritten as

$$SSTO=SSR+SSE.$$

Based on the rewritten model, R^2 statistic would be defined as

$$R_W^2=\frac{(\hat{Y}-\overline{Y}e)'V^{-1}(\hat{Y}-\overline{Y}e)}{(Y-\overline{Y}e)V^{-1}(Y-\overline{Y}e)}=1-\frac{\hat{u}'V^{-1}\hat{u}}{(Y-\overline{Y}e)V^{-1}(Y-\overline{Y}e)}.$$

Note that V is estimated by ML of REML.

From another perspective developed by Magee (1990), the F statistic for testing the hypothesis of nonzero k-1 non-intercept parameters is

$$F=\frac{(SSTO-SSE)/(k-1)}{SSE/(n-k)}=\frac{R_W^2/(k-1)}{(1-R_W^2)/(n-k)}$$

It is related to the Wald statistic as

$$W=\frac{n(SSTO-SSE)}{SSE}=(k-1)(1+\frac{k}{n-k})F.$$

Then we can write $R_W^2=\frac{W}{n+W}$.

If the components of the random effect are also independent identically distributed according to a normal distribution or the model contains only the fixed effects, then we have $R_{LR}^2=R_W^2$=traditional R^2. In this case, let the variance covariance matrix of Y be $\sum=\sigma_y^2 I$. We can prove this equality in two ways:

1. The minus two times the maximum log-likelihood for model (3) with $V = \sum$, we have

$$
\begin{aligned}
-2\log L_M &= \log\left|\sum\right| + u'\sum{}^{-1} u + n\log(2\pi) \\
&= \log\left|\frac{\left\|(Y - X\hat{\beta})\right\|^2}{n} I\right| + (Y - X\hat{\beta})'(\frac{\left\|(Y - X\hat{\beta})\right\|^2}{n} I)^{-1}(Y - X\hat{\beta}) + n\log(2\pi) \\
&= n\log(\frac{\left\|(Y - X\hat{\beta})\right\|^2}{n}) + n + n\log(2\pi)
\end{aligned}
$$

where $u = Y - X\hat{\beta}$ and $\hat{\beta}$ is the MLE and also the BLUE of β. Similarly, for model (3) without the covariate, we have

$$-2\log L_0 = n\log(\frac{\left\|(Y - \bar{Y})\right\|^2}{n}) + n + n\log(2\pi).$$

Then the R^2_{LR} statistic can be written as

$$
\begin{aligned}
R^2_{LR} &= 1 - \exp(\log\frac{\left\|Y - X\hat{\beta}\right\|^2}{n} - \log\frac{\left\|Y - \bar{Y}\right\|^2}{n}) \\
&= 1 - \frac{\left\|Y - X\hat{\beta}\right\|^2}{\left\|Y - \bar{Y}\right\|^2} = R^2
\end{aligned}
$$

At the same time, under the same assumptions we have

$$
\begin{aligned}
R^2_W &= 1 - \frac{(Y - X\hat{\beta})'\sum(Y - X\hat{\beta})}{(Y - \bar{Y})'\sum(Y - \bar{Y})} \\
&= 1 - \frac{(Y - X\hat{\beta})'\sigma_y^2 I(Y - X\hat{\beta})}{(Y - \bar{Y})'\sigma_y^2 I(Y - \bar{Y})} \\
&= 1 - \frac{\left\|Y - X\hat{\beta}\right\|^2}{\left\|Y - \bar{Y}\right\|^2} \\
&= R^2_{LR}
\end{aligned}
$$

2. Since LR and W are related as $LR = n\log(1 + W/n)$, so

$$
\begin{aligned}
R^2_{LR} &= 1 - \exp(-LR/n) \\
&= 1 - \exp(-\log(1 + W/n)) \\
&= \frac{W}{n + W} = R^2_W
\end{aligned}
$$

In the context of the R^2 statistic without the distribution assumption, R^2_{LR} cannot be computed whereas R^2_W can be computed and has meaningful interpretation. R^2_{LR} measures how well the model with the given distribution fits the data. R^2_W only measures how well the means of the model predict the data without fully specified the actual form of the distribution of Y.

§2.3.3 The R_C *Statistic*

Motivated by the concordance correlation coefficient

$$\rho_c = 1 - \frac{E[(Y_1 - Y_2)^2]}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} = \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}$$

Vonesh et al. (1996) proposed

$$R_c = 1 - \frac{(\boldsymbol{Y} - \hat{\boldsymbol{Y}})'(\boldsymbol{Y} - \hat{\boldsymbol{Y}})}{(\boldsymbol{Y} - \bar{y}\mathbf{1}_n)'(\boldsymbol{Y} - \bar{y}\mathbf{1}_n) + (\hat{\boldsymbol{Y}} - \hat{y}\mathbf{1}_n)'(\hat{\boldsymbol{Y}} - \hat{y}\mathbf{1}_n) + n(\bar{y} - \hat{y})^2}$$

as a goodness of fit measure for generalized nonlinear mixed effect model, where n is the number of observations; $\boldsymbol{Y}$ is the vector of observed values; $\hat{\mathbf{Y}}$ is a predictor of $\mathbf{Y}$; $\bar{y}$ is the mean of the elements of $\boldsymbol{Y}$; and $\hat{y}$ is the mean of the elements of $\hat{\mathbf{Y}}$.

R_c can be interpreted as a measure of the degree of the agreement between the observed values and the predicted values as ρ_c measures agreement between Y_1 and Y_2.

§2.3.4 The P_{rand} *Statistic*

With a multivariate normal random effect, Zheng (2000) proposed

$$P_{rand} = 1 - \frac{-PQL_M}{-PQL_N} = 1 - \frac{(1/2\hat{\sigma})(\boldsymbol{Y} - \hat{\boldsymbol{Y}})'(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) + \hat{\boldsymbol{\gamma}}'\hat{\boldsymbol{G}}^{-1}\hat{\boldsymbol{\gamma}}/2}{(1/2\hat{\sigma})(\boldsymbol{Y} - \bar{y}\mathbf{1}_n)'(\boldsymbol{Y} - \bar{y}\mathbf{1}_n)}$$

to be a measure of the proportional reduction in penalized quasi-likelihood function, where PQL_M denotes penalized quasi-likelihood function for the model of interest; PQL_N denotes penalized quasi-likelihood function for the null model which is the model only contains the intercept; $\hat{\boldsymbol{\gamma}}$ is the estimated best linear unbiased predictor (BLUP) of $\boldsymbol{\gamma}$; $\hat{Y} = X\hat{\beta} + Z\hat{\gamma}$ is the estimated BLUP of Y; $\hat{\boldsymbol{G}}$ is the maximum likelihood estimate (MLE) of G, the variance

covariance matrix of Y; and $\hat{\sigma}$ is the MLE of σ. G and σ can also be estimated by REML. Note that when the model only has the fixed effect, P_{rand} is reduced to the traditional R^2 statistic.

Let $l(\mu,\sigma;Y)$ be the conditional log-likelihood function given mean μ and variance component σ for the model of interest and $l(Y,\sigma;Y)$ be the maximum log-likelihood value corresponding to the perfect prediction. The deviance is defined as

$$d(Y,\mu)/\sigma = -2(l(\mu,\sigma;Y) - l(Y,\sigma;Y)).$$

Thus the sample deviances for the model of interest and the fixed intercept model (null model) are $\sum_{i=1}^{n} d_i(Y_i,\hat{Y}_i)$ and $\sum_{i=1}^{n} d_i(Y_i,\bar{Y})$, respectively. Based on the above definitions, the negative penalized quasi-likelihood (-PQL) is defined as

$$-PQL = \frac{1}{2\sigma}\sum_{i=1}^{n} d_i(Y_i,\mu_i) + \frac{1}{2}\gamma' G^{-1}\gamma.$$

Hence –PQL's for the model (3) and the null model (model (3) without the covariate) are

$$-PQL_M = \frac{1}{2\hat{\sigma}}\sum_{i=1}^{n} d_i(Y_i,\hat{Y}_i) + \frac{1}{2}\hat{\gamma}'\hat{G}^{-1}\hat{\gamma}$$

and

$$-PQL_N = \frac{1}{2\hat{\sigma}}\sum_{i=1}^{n} d_i(Y_i,\bar{Y}_i).$$

Under the normality assumption,

$$-PQL_M = \frac{1}{2\hat{\sigma}}(Y-\hat{Y})'(Y-\hat{Y}) + \frac{1}{2}\hat{\gamma}'\hat{G}^{-1}\hat{\gamma}$$

and

$$-PQL_N = \frac{1}{2\hat{\sigma}}(Y-\bar{Y})'(Y-\bar{Y}).$$

It can be shown approximately that $P_{rand} = 1 - \dfrac{L_M + \frac{n}{2}\log(2\pi)}{L_N + \frac{n}{2}\log(2\pi)}$, where L_M is the maximum log-likelihood value of the model of interest and L_N is the maximum log-likelihood value of the null model. Based on the approximation, we could interpret P_{rand} as a measure of the proportional reduction of the log-likelihood comparing the model of interest with the fixed intercept only model.

The range of the statistic P_{rand} is between 0 and 1 under the above model assumptions. The larger P_{rand}, the better prediction and the smaller random effect. The penalty for the random effects in P_{rand} is analogous to the Akaike's Information Criterion (AIC) and the Schwarz's Bayesian Information Criterion (BIC).

§2.3.5 The r^2 *Statistic*

Xu (2003) also proposed three other kinds of R^2-like measures: r^2, R_2^2 and ρ^2 to assess the goodness of fit of the model. If we write β as $(\beta_0, \beta_1')'$ and γ as $(\gamma_0, \gamma_1')'$ where β_0 and γ_0 are the fixed and random intercepts, then two kinds of null models are possible in this case:

$$H_0: \ Y_j = \beta_0^* + \gamma_0^* + u_j^* \tag{4}$$

and

$$H_0: \ Y_j = \beta_{00}^* + u_{0j}^* \tag{5}$$

Note that null model (4) has random intercept but null model (5) does not and that model (4) and (5) are in fact nested. Denote $\sigma_0^2 = \text{var}(Y|\gamma_0^*) = \text{var}(u^*)$ for model (4) and $\sigma_{00}^2 = \text{var}(u_0^*) = \text{var}(Y)$ for model (5). Then the proportion of variation in Y explained by X is

$$\Omega^2 = 1 - \frac{\text{var}(Y|X,\gamma)}{\text{var}(Y|\gamma_0^*)} = 1 - \frac{\sigma^2}{\sigma_0^2}$$

for model (4) and

$$\Omega^2 = 1 - \frac{\text{var}(Y|X,\gamma)}{\text{var}(Y)} = 1 - \frac{\sigma^2}{\sigma_{00}^2}$$

for model (5).

We use the maximum likelihood (ML) method to estimate Ω^2 and we would have the following measures for model (4) and (5) respectively:

$$r^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}$$

$$r^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_{00}^2}$$

where $\hat{\sigma}^2$, $\hat{\sigma}_0^2$ and $\hat{\sigma}_{00}^2$ are the ML estimates of σ^2, ${\sigma_0}^2$ and σ_{00}^2. Note when the model only has the fixed effect, we only have model (5) and r^2 is equal to the traditional R^2 statistic.

§2.3.6 The R_2^2 Statistic

Xu (2003) also defined a related R^2-like statistic based on the residuals obtained from fitting model (2). Let $\hat{Y} = X\hat{\beta} + Z\hat{\gamma}$ be the linear predictor in model (2), then the residual under the fitted model (2) is $\hat{u} = Y - \hat{Y}$. Similarly, let $\hat{\beta}_0^*$ and $\hat{\gamma}_0^*$ be the predictors of β_0^* and γ_0^* under model (4), and $\hat{\beta}_{00}^*$ be the predictor of β_{00}^* under model (5). Then the residual under the fitted model (4) and (5) are $\hat{u}^* = Y - \hat{Y}^*$ and $\hat{u}_0^* = Y - \hat{Y}_0^* = Y - \overline{Y}$ where $\hat{Y}^* = \hat{\beta}_0^* + \hat{\gamma}_0^*$, $\hat{Y}_0^* = \hat{\beta}_{00}^* = \overline{Y}$ and $\overline{Y}$ is the average of observed values.

R_2^2 statistics under model (4) and model (5) are respectively

$$R_2^2 = 1 - \frac{\hat{u}'\hat{u}}{\hat{u}^{*\prime}\hat{u}^*} = 1 - \frac{RSS}{RSS_0}$$

and

$$R_2^2 = 1 - \frac{\hat{u}'\hat{u}}{\hat{u}_0^{*\prime}\hat{u}_0^*} = 1 - \frac{RSS}{RSS_{00}},$$

where RSS, RSS_0, and RSS_{00} are the residual sums of squares under model (2), (4) and (5). Notice that R_2^2 for model (5) is just the traditional R^2 statistic which is not preferred since it ignores the random components.

Since RSS/n estimates the residual variance σ^2 of model (2), RSS_0/n estimates the residual variance ${\sigma_0}^2$ of model (4), and RSS_{00}/n estimates the residual variances σ_{00}^2, R_2^2 is also an estimator of Ω^2.

§2.3.7 The ρ^2 Statistic

The explained randomness was first proposed by Kent (1983). Xu (2003) defined a R^2 statistic with the use of the conditional likelihood of the observed data given the predicted random effects.

Under model (2), define the residual randomness as $D(Y|X,\gamma) = \exp(-2E(\log l(Y|X,\gamma)))$, under model (3), and the total randomness of Y given only the random effect as

$D(Y|\gamma_0^*) = \exp(-2E(\log l(Y|\gamma_0^*)))$, where $\log l(Y|X,\gamma)$ is the log-likelihood function for model (2) and $\log l(Y|\gamma_0^*)$ is the log-likelihood function for model (4). Then the proportion of explained randomness could be defined as

$$1 - \frac{D(Y|X,\gamma)}{D(Y|\gamma_0^*)} = 1 - \exp(-\tau)$$

where $\tau = 2(E(\log l(Y|X,\gamma)) - E(\log l(Y|\gamma_0^*)))$ is twice the Kullback-Leibler information gain.

Denote the vector of unknown parameters under model (2) and (4) by θ and θ_0, then

$$n\hat{\tau} = 2\log(L(\hat{\theta})/L(\hat{\theta}_0)) = n\log(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}) - \frac{RSS}{\hat{\sigma}^2} + \frac{RSS_0}{\hat{\sigma}_0^2}$$

where $L(\theta) = \prod_{j=1}^{n} l(Y_j|\gamma)$ is the conditional likelihood of the observed data given the random effects under model (2).

A measure of explained randomness is defined as

$$\rho^2 = 1 - \exp(-\hat{\tau}) = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\exp(\frac{RSS}{n\hat{\sigma}^2} - \frac{RSS_0}{n\hat{\sigma}_0^2}).$$

Notice that when there is no random effect, the ρ^2 measure is equal to the traditional R^2 for the linear regression model.

The coefficient ρ_0^2 for model (5) can be defined as $\rho_0^2 = 1 - \exp(-\hat{\tau}_0)$, where $\hat{\tau}_0 = \log\frac{\hat{\sigma}_{00}^2}{\hat{\sigma}^2} - \frac{RSS}{n\hat{\sigma}^2} + 1$.

Again, because RSS/n estimates the residual variance σ^2 of model (2), RSS_0/n estimates the residual variance σ_0^2 of model (4), and RSS_{00}/n estimates the residual variance σ_{00}^2, ρ^2 is also an estimator of Ω^2 and it should be closed to r^2 and R_2^2. Based on a first-order Taylor approximation, we have $\rho^2 \approx 1 - \frac{RSS}{RSS_0} = R_2^2$. Then we could say that r^2 takes into account the different degree of freedom under the full and null models, however, R_2^2 and ρ^2 do not, since

$\hat{\sigma}^2 \approx \frac{RSS}{n-df}$, $\hat{\sigma}_0^{\,2} \approx \frac{RSS_0}{n-df_0}$ and $\hat{\sigma}_{00}^2 \approx \frac{RSS_{00}}{n-df_{00}}$, where df, df_0, and df_{00} are the degrees of freedom of the residual variances under model (2), (4), and (5) respectively.

§2.4 Conditional and Marginal R^2 Statistics

Vonesh et al (1996) and Vonesh and Chinchilli (1997) invited the concepts of conditional and marginal R^2 statistic. For the conditional version of R^2 statistic, the fitted value $\hat{Y}=X\hat{\beta}+Z\hat{\gamma}$ is used to compute the R^2-like statistics according to the formula introduced in Section 2.3. However, for the marginal version of the R^2 statistic, the fitted value $\hat{Y}=X\hat{\beta}$ is used to compute the R^2-like statistics in their formula. The conditional version accounts for both the fixed and random effects to measure the overall goodness of fit and prediction power, while the marginal version only measures the fixed-effect part, the mean of the model. R_C statistic has both versions of R^2 statistic.

However, the other R^2 statistics such as R_{LR}^2 R_W^2 P_{rand} r^2, R_2^2 and ρ^2 do not have both conditional and marginal versions. For R_{LR}^2 and r^2, this is because that the definitions of these two statistics do not relate to the estimated values of Y. Instead, they only concern about the log-likelihood values and the estimators for the residual variances of the models. For R_W^2 statistic, we could view the mixed model in the way of combining the random effects with the error term together as new noise with the variance covariance matrix $V = Z'GZ + R$ of a fixed-effect model (3). In this case, the predicted value can only account for the fixed effects. Therefore, R_W^2 statistic only has the marginal version of R^2 statistic. For P_{rand} statistic, when we only use the fixed effect terms to be the predicted values, we no longer has the random effect in model (3) and we could not use P_{rand} which involves the random effect as a penalty in the numerator of the definition anymore. Therefore, we have to use $\hat{Y}=X\hat{\beta}+Z\hat{\gamma}$ as the predicted values of the mixed model. Hence, P_{rand} only has the conditional version of the R^2 statistic. And for R_2^2 and ρ^2, since the RSS and RSS_0 are only defined based on the fitted value of Y as $\hat{Y}=X\hat{\beta}+Z\hat{\gamma}$, they only have

the conditional version too. In addition, if we use $X\hat{\beta}$ as $\hat{Y}$ in R_2^2 and ρ^2, it would not be appropriate of using RSS/n or $\frac{RSS}{n-df}$ to estimate σ^2 anymore, so do ${\sigma_0}^2$ and ${\sigma_{00}}^2$.

CHAPTER 3 - Association Mapping Application

§3.1 Association Mapping

A phenotype is any observed quality of an organism and the genotype is the genetic constitution and the specific allele makeup of an individual. A single nucleotide polymorphism (SNP), is a DNA sequence variation occurring when a single nucleotide - A, T, C, or G in the genome (or other shared sequence) differs between members of a species (or between paired chromosomes in an individual) and they could be used as the markers to study the complex traits. For example, two sequenced DNA fragments from different individuals, AC**C**T to AC**T**T, are differed by a single nucleotide. In this case we say that there are two alleles: C and T. Almost all common SNPs have only two alleles. During the statistical analysis of the SNPs markers, we usually code the two alleles as 0 and 1 as categorical variable. As we know, genetic factors affect the corresponding quantitative traits, and the occurrence of disease.

Association mapping holds substantial promise for unraveling the genetic basis of the interested complex traits for human and other species. Association analysis is a method to identify the relationship between molecular markers or candidate genes and the interested traits based on a given collected population. It is different from quantitative trait locus (QTL) mapping since QTL mapping needs the family-based population while association mapping does not, although alternative methods may use the family-based controls to avoid the potential problem of population stratification. Association mapping can address the targeted genes faster and more efficiently and provide much more information to the candidate genes to verify the function of candidate genes.

In candidate gene association mapping, genes are selected based on their location in a region of linkage or other evidence showing that the selected genes may impact the interested quantitative traits. However, Candidate gene study relies on the precision of the selection which is based on the biological hypotheses.

Genome-wide association study provides a powerful approach for us to understand the complex traits better than in the past. It is defined as an approach that surveys most of the genome for causal genetic variants.

Statistically, we treat the interested quantitative trait as the response and the markers information as the explanatory variables of the model. After fitting the model, we could base on the specific tests and corresponding p-values or the R^2 values to assess the effects of the functional markers or genes on the diseases.

It is well known that allele frequency differs between cases and the controls due to different systematic ancestry. This kind of population stratification can cause spurious associations in association mapping if not accounted for in the test. Since we commonly classify the individual in a sample into populations, researchers want to understand the population besides the samples with individuals. Pritchard (2000) proposed a cluster method to assign the individual into subpopulations based on multilocus genotype data to describe the population structure. If we assume there are k subpopulations, and all of them are set with the allele frequencies at each locus, then the individual could be assigned into one of these k subpopulations, or more than one admixed subpopulations. After the cluster analysis, the probability results of each individual could be arranged into a matrix denoted by Q with the size of the number of individuals by the number of subpopulations. They also designed computer software named "STRUCTURE" to calculate the Q matrix using the genotype data information.

Besides the cluster analysis using "STRUCTURE" to assign individuals to the subpopulations, Price (2006) proposed a method named "EIGENSTRAT" to detect and correct for population stratification. They applied the principal components analysis to the genotype data and used the first several eigenvectors as P matrix instead of Q matrix to account for the population structure.

The genetic relatedness of random molecular markers is also an important aspect that should be considered in association mapping. Many areas have studied this genetic relatedness very well and many methods are proposed to estimate them. Relative kinship estimates (Loiselle et al. 1995; Ritland 1996) provide both inter-coancestry and intra-individual estimates in a symmetric matrix, say K matrix with the size of the number of individuals by the number of individuals, and are used to account for the relatedness in diverse association mapping panels.

The Kinship coefficients are also called coancestry coefficients. They are computed based on the probability of identity of alleles for two homologous genes sampled in some particular way. Software named "SPAGeDi" could compute the kinship coefficients for us in two ways:

1. They are computed as a correlation coefficient between allelic states (Loiselle et al. 1995).
2. They are estimated by Ritland (1996) in the way of giving more weight to the rare alleles and having lower sampling variance than the first method.

Yu et al. (2006) proposed a new method named unified mixed-model association mapping that takes into account both the population structure and the familial relatedness. While the previous method of genomic control adjusts the test statistics obtained from a model that does not consider the population structure or kinship, the new proposed method adjusts the test statistic internally by considering the multiple levels of relatedness. They showed that this new method could control the type Ⅰ and type ⅠⅠ errors much better than the other methods used in the association mapping. The mixed effect model they proposed is illustrated in Section 3.2.

§3.2 The Mixed Model in Association Mapping

The mixed model for Q+K method is

$$\mathbf{y} = \mu + \mathrm{Q}\mathbf{v} + Z\mathbf{u} + \mathbf{e} \tag{6}$$

where $\mathbf{y}$ is a vector of phenotype observation, μ is a vector of intercepts, $\mathbf{v}$ is a $k \times 1$ vector of population effects, $\mathbf{u}$ is a $n \times 1$ vector of random polygene background effects, $\mathbf{e}$ is a vector of random experimental errors with mean 0 and covariance matrix Var($\mathbf{e}$), Q is an $n \times k$ matrix defining the subgroup membership, Z is an incidence matrix relating $\mathbf{y}$ to $\mathbf{u}$. In our case, since we do not have the replication for each subject, the Z design matrix of the model is the identity matrix with the size of number of observations. We have Var($\mathbf{u}$)=2KV_g, where K is a known $n \times n$ matrix of kinship coefficients, V_g is the unknown genetic variance which is a scalar. Var($\mathbf{e}$)=RV_R, where R is an $n \times n$ matrix, and V_R is the unknown residual variance which is a scalar too.

We examined four different models: Q, K, P and P+K model to compare the results (Table 3.1). The definitions of Q and K were the same as in previous sections. In both P and P+K model, the P matrix consists of first several principal components that are eigenvectors calculated by the principal component analysis (PCA) of SNPs data. In our empirical analysis presented in next section, we followed analyses we chose the first three PCAs for maize data and first eight PCAs for Arabidopsis data to be consistent with analyses in previous publications (Yu et al. 2006; Zhao et al. 2007).

Table 3.1 Models Used in the Application

Model Name	Model form
The Q+K model	$\mathbf{y} = \mu + \mathrm{Q}\mathbf{v} + Z\mathbf{u} + \mathbf{e}$
The Q model	$\mathbf{y} = \mu + \mathrm{Q}\mathbf{v} + \mathbf{e}$
The K model	$\mathbf{y} = \mu + Z\mathbf{u} + \mathbf{e}$
The P+K model	$\mathbf{y} = \mu + P\mathbf{v} + Z\mathbf{u} + \mathbf{e}$
The P model	$\mathbf{y} = \mu + P\mathbf{v} + \mathbf{e}$

§3.3 Empirical Analysis

§3.3.1 Maize Data Example

We used the data collected from maize (Yu et al. 2006 Nat Genet 38.) with 277 inbred lines. There are three quantitative traits: flowering time, ear height and ear diameter, which we will use as the response variables in the models. The Q matrix and the K matrix are derived using 553 Single Nucleotide Polymorphism (SNPs) markers in STRUCTURE and SPAGeDi software, respectively.

The intercept only model ($\mathbf{y} = \mu + \mathbf{e}$) and five other models were fitted with each of the three traits as the response. The R^2_{LR} values were calculated between each of the five models and the intercept-only model (Table 3.2).

Table 3.2 R^2_{LR} Values for Three Traits in Different Models

R^2_{LR}	Flowering Time	Ear Height	Ear Diameter
Q+K	0.42	0.25	0.22
Q	0.35	0.16	0.05
K	0.35	0.21	0.21
P+K	0.41	0.25	0.22
P	0.31	0.15	0.05

It is illustrated in Table 3.2 that the R^2_{LR} value of the Q+K model is less than the sum of R^2_{LR} values of the Q model and the K model across three traits and that the same can be said for the P+K model, the P model and the K model. This is due to the fact that the Q matrix representing the population structure, the P matrix characterizing the main information of the interested population and the K matrix showing the relatedness of 277 inbred lines are all derived from the common source, the same genotype data but in different ways. The overlap portion of what Q matrix explains and what K matrix explains, denoted by "Q • K" is the sum of the R^2_{LR} values of the Q model and the K model minus that of the Q+K model. Similarly, the overlap portion of what P matrix explains and K matrix explains, denoted by "P • K" is the sum of the R^2_{LR} value of the P model and the K model minus that of the P+K model. Based on our maize data set, we have the following Table3.3:

Table 3.3 Overlaps for Three Traits in Different Models

R^2_{LR}	Flowering Time	Ear Height	Ear Diameter
Q • K	0.28	0.12	0.04
P • K	0.25	0.11	0.04

We could also show this fact in the way of Venn Diagrams.

Figure 3.1 Flowering Time Trait Involving Q and K Models

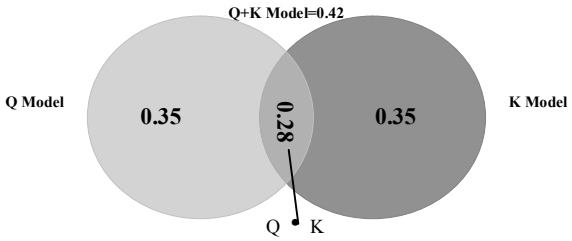


Figure 3.2 Ear Height Trait Involving Q and K Models

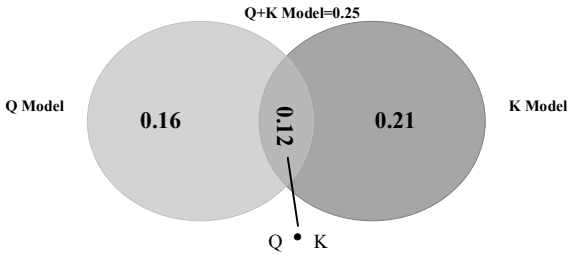


Figure 3.3 Ear Diameter Trait Involving Q and K Models

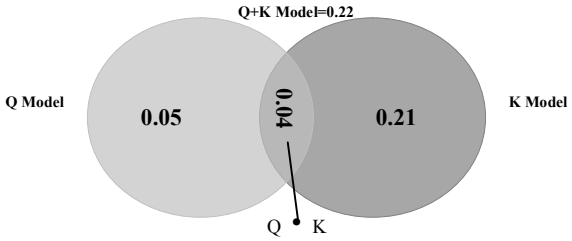


Figure 3.4 Flowering Time Trait Involving P and K Models

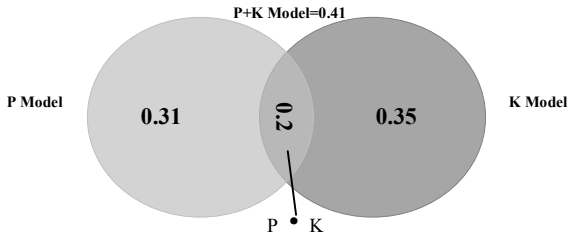


Figure 3.5 Ear Height Trait Involving P and K Models

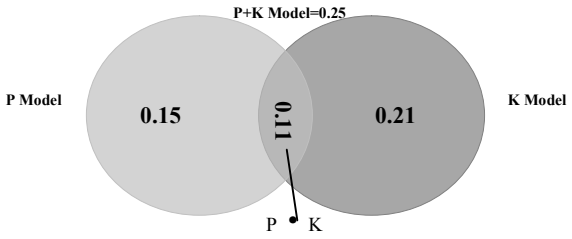


Figure 3.6 Ear Diameter Trait Involving P and K Models

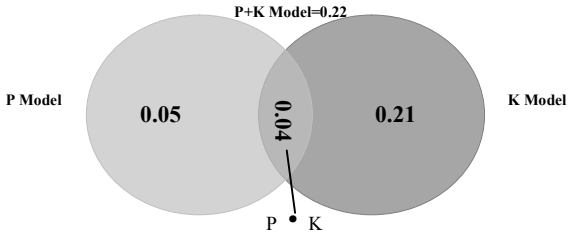


Table 3.4 R^2_W Values for Three Traits in Different Models

R^2_W	Flowering Time	Ear Height	Ear Diameter
Q+K	0.11	0.05	0.01
Q	0.35	0.16	0.05
K	0.00003	0.00003	0.00004
P+K	0.09	0.05	0.01
P	0.31	0.15	0.05

From Table 3.4, we will find that K model's R^2_W values are very near to zero. Because based on the definition of R^2_W statistic under model (3), when there is only intercept and random effects in the model, $\hat{Y}$ is just the estimated intercept and $(Y-\hat{Y})'V^{-1}(Y-\hat{Y})$ and $(Y-\overline{Y})'V^{-1}(Y-\overline{Y})$ are the same theoretically in this case, which will lead the R^2_W value to be 0. Therefore, R^2_W cannot report the correlation between random effects and the responses, which means that R^2_W is not a useful statistic in this situation.

When we compare the R^2_W values of the mixed models and the models only consist of the fixed effects, we always see that the previous values are less than the later values. It seems like there is a contradiction with the criterion 1 proposed by Cameron and Widmeijer (1996) in Section 2.1. In fact, the criterion is not applicable in this case because we are comparing two models with different assumptions of the errors. For the mixed model the variance covariance matrix is assumed to be V and for the models only have the fixed effects, the variance covariance matrix is assumed to be σ^2I. We just add the random effect into the error term instead of adding more regressors to the model.

From Table 3.3 and 3.4, we see that the R^2_W and R^2_{LR} for the Q model that only contains the fixed effects are the same, which we have proved in section 2.3. However, two R^2 statistics of model Q+K are different obviously. The reason for this result is that for the mixed-model we have $-2l(G,R) = \log|\mathrm{V}| + u'V^{-1}u + n\log(2\pi)$ in the ML method where $u = Y - X\hat{\beta}$, so R^2_{LR} takes into accounts the determinant of the variance covariance matrix V of Y, $u'V^{-1}u$ and also $(Y-\overline{Y})'R^{-1}(Y-\overline{Y})$ that are the generalized sums of squares with respect to the full model and the null model. While R^2_W only involves $u'V^{-1}u$, the same as what R^2_{LR} does, and $(Y-\overline{Y})'V^{-1}(Y-\overline{Y})$, called weighted sum of square, which is different from $(Y-\overline{Y})'R^{-1}(Y-\overline{Y})$, where R is the variance covariance matrix of Y of the model that only includes the intercept.

Table 3.5 Components of Maximum Log-Likelihood Values Using ML Method (Mixed-Model)

Q+K model	Flowering Time	Ear Height	Ear Diameter
$\log\lvert V\rvert$	773.19	1507.40	605.51
$u'V^{-1}u$	274.00058	276.00122	247.00071
$(Y-\bar{Y})'R^{-1}(Y-\bar{Y})$	274	276	247
$(Y-\bar{Y})'V^{-1}(Y-\bar{Y})$	308.49	289.29	249.50

Table 3.5 involves the computation of $\log\lvert V\rvert$. When we use SAS or Matlab to compute $\log\lvert V\rvert$, the software can not compute the result directly since the value is too large. So we do the following steps to resolve this problem:

First, we compute $\lvert V\rvert$. The value of $\lvert V\rvert$ is also too big to output by the software, so we use one of the properties of the determinant of matrix which is

$$\begin{vmatrix} e_{11} & e_{12} & \dots & e_{1n} \\ e_{21} & e_{22} & \dots & e_{2n} \\ \dots & \dots & \dots & \dots \\ e_{n1} & e_{n2} & \dots & e_{nn} \end{vmatrix} = k \begin{vmatrix} e_{11} & e_{12} & \dots & e_{1n} \\ e_{21}/k & e_{22}/k & \dots & e_{2n}/k \\ \dots & \dots & \dots & \dots \\ e_{n1} & e_{n2} & \dots & e_{nn} \end{vmatrix}.$$

For example, $\lvert V_1\rvert$ is the determinant of V_1 that is the variance covariance matrix of the response variable corresponding to ear height trait. We first divided every row of the matrix V by 100, and then we calculate the determinant of the restructured matrix, say $\lvert V_{1p}\rvert$, with much smaller elements. Finally, $\lvert V_1\rvert$ is just that 100^{264} times $\lvert V_{1p}\rvert$, and $\log\lvert V\rvert = \log\lvert V_{1p}\rvert + \log 100^{264}$.

Notably, in Table 3.5 the values of the row named $(Y-\bar{Y})'R^{-1}(Y-\bar{Y})$ always equal to the n, the number of observations. This is because:

$$r'V^{-1}r = (Y-\hat{Y})'V^{-1}(Y-\hat{Y})$$
$$= (Y-\hat{Y})'R^{-1}(Y-\hat{Y})$$

Where $R = \hat{\sigma}^2 I = \dfrac{\left\|Y-\hat{Y}\right\|^2}{n} I$

$$r'V^{-1}r = \left(\frac{\left\|Y-\hat{Y}\right\|^2}{n}\right)^{-1} (Y-\hat{Y})'I(Y-\hat{Y}) = n$$

Table 3.6 Marginal R_C Values for Three Traits in Different Models

R_C	Flowering Time	Ear Height	Ear Diameter
Q+K	0.53	0.28	0.09
Q	0.51	0.27	0.10
K	0	0	0
P+K	0.48	0.28	0.09
P	0.47	0.25	0.10

From Table 3.6, all R_C values for K model are equal to zero for the same reason as for R^2_W statistic. Therefore, R_C is not a preferred R^2 statistic.

Table 3.7 Conditional R_C Values for Three Traits in Different Models

R_C	Flowering Time	Ear Height	Ear Diameter
Q+K	0.92	0.88	0.96
Q	0.51	0.27	0.10
K	0.94	0.91	0.96
P+K	0.92	0.88	0.96
P	0.47	0.25	0.10

Table 3.8 P_{rand} Values for Three Traits in Different Models

P_{rand}	Flowering Time	Ear Height	Ear Diameter
Q+K	0.83	0.81	0.85
Q	0.35	0.16	0.05
K	0.86	0.85	0.86
P+K	0.83	0.81	0.85
P	0.31	0.15	0.05

Table 3.9 r^2 Values for Three Traits in Different Models

r^2	Flowering Time	Ear Height	Ear Diameter
Q+K	0.74	0.66	0.79
Q	0.35	0.16	0.05
K	0.78	0.70	0.79
P+K	0.74	0.65	0.79
P	0.31	0.15	0.05

Table 3.10 R_2^2 Values for Three Traits in Different Models

R_2^2	Flowering Time	Ear Height	Ear Diameter
Q+K	0.98	0.98	0.99
Q	0.35	0.16	0.05
K	0.98	0.98	0.99
P+K	0.87	0.82	0.93
P	0.31	0.15	0.05

Table 3.11 ρ^2 Values for Three Traits in Different Models

ρ^2	Flowering Time	Ear Height	Ear Diameter
Q+K	0.84	0.78	0.89
Q	0.35	0.16	0.05
K	0.88	0.82	0.89
P+K	0.84	0.79	0.89
P	0.31	0.15	0.05

From Table 3.7 to Table 3.11, the R^2 values for the K model all are larger than the R^2 values for the mixed effect models, which contradicts the criterion 1 proposed by Cameron and Widmeijer (1996) in Section 2.1. With the same assumption of the variance covariance structure, after adding more regressors into the model, the R^2 should not decrease. The proposers of the R^2-like statistics only considered the non-decreasing property of R^2 values by adding more fixed regressors into the mixed models with the same random effects. They did not compare the

models with different random and fixed effects. Therefore, conditional R_C, P_{rand}, r^2, R_2^2, and ρ^2 are not the preferred R^2 statistics to use for the application considered here.

On the other hand, R_{LR}^2 is based on the likelihood principle where the log-likelihood function increases as the number of parameters increases. It does not have the same problem as the other R^2-like statistics. All in all, R_{LR}^2 statistic is the most useful and make the most sense R^2-like statistic relative to the others based on our results using Maize data example.

To detect the effect of SNP on the phenotype, we could add the marker data into the interested model to check if there is a big difference in the R^2 values. We used the Q+K model as the example. After adding 553 SNPs markers into the Q+K mixed models one at a time as the fixed regressors, we obtained the R_{LR}^2 values across three traits. The following figures show the important SNPs with significant peaks to certain traits.

Figure 3.7 Flowering Time Trait with SNPs Markers

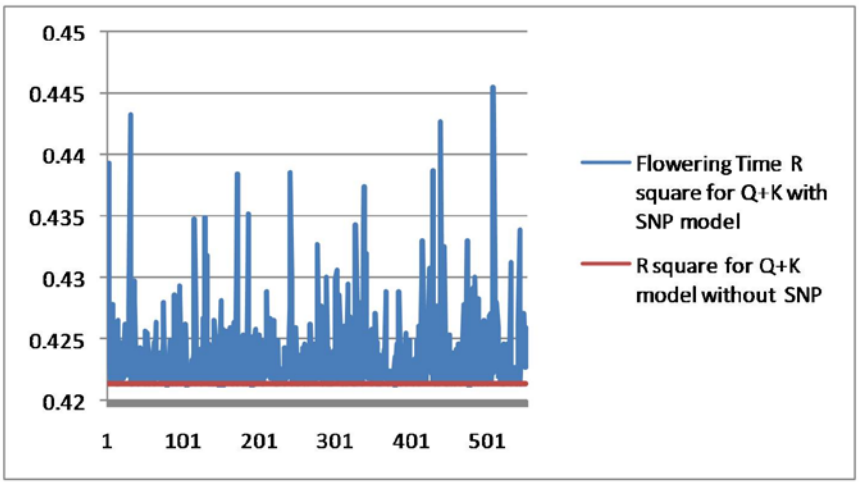


Figure 3.8 Ear Height Trait with SNPs Markers

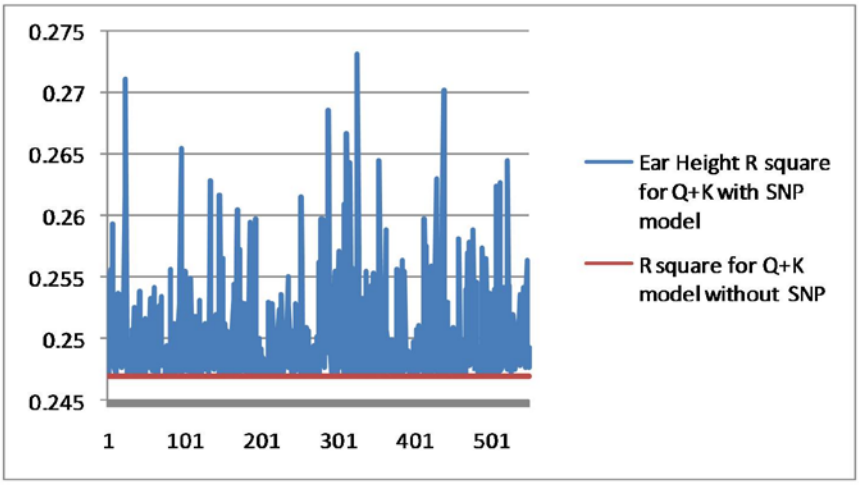
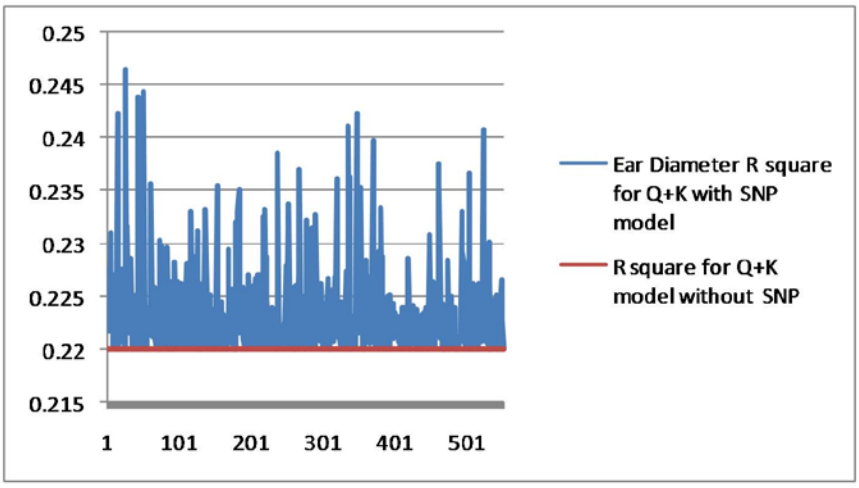


Figure 3.9 Ear Diameter Trait with SNPs Markers



§3.3.2 Arabidopsis Data Example

The data used in this example is a global sample of 95 Arabidopsis thaliana accessions (Zhao et al. 2007 PloS Genet 3.) with three phenotypes: SDV, JIC8W and FRI. These phenotypes are the mean flowering time for accessions under different experimental conditions obtained at University of Southern California (USC) and at the John Innes Centre (JIC) and also an expression level of key flowering time gene. We use the same methods to compute Q, K and P matrices as in the Maize example based on 5419 SNPs markers.

Five models were fitted with each of the three traits as the response. The R^2_{LR} values are reported in Table 3.13.

Table 3.12 Descriptions of The Phenotype

Phenotype	Description
SDV	Short days without 5-week vernalizaiton at USC
JIC8W	Long days with 8-week vernalization at JIC
FRI	FRI expression

Table 3.13 R^2_{LR} Values for Three Traits in Different Models

R^2_{LR}	SDV	JIC8W	FRI
Q+K	0.40	0.54	0.35
Q	0.37	0.53	0.31
K	0.17	0.08	0.18
P+K	0.47	0.56	0.39
P	0.47	0.56	0.44

From Table 3.13, we conclude that P+K model is better than the other models and the P matrix capture more character of the SNPs markers than the Q matrix.

Again, there exist overlaps between what Q matrix and K matrix explain about the genotype information.

Table 3.14 Overlaps for Three Traits in Different Models

R^2_{LR}	SDV	JIC8W	FRI
Q·K	0.14	0.07	0.14
P·K	0.17	0.08	0.23

We show this fact in the way of following Venn Diagram figures.

Figure 3.7 SDV Trait Involving Q and K Models

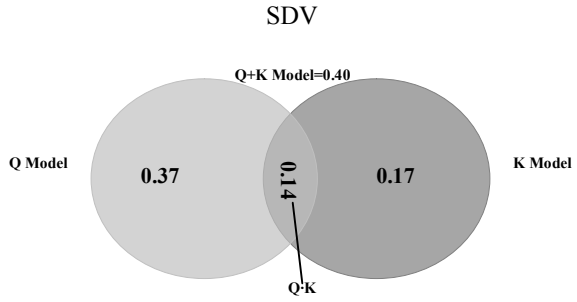


Figure 3.8 JIC8W Trait Involving Q and K Models

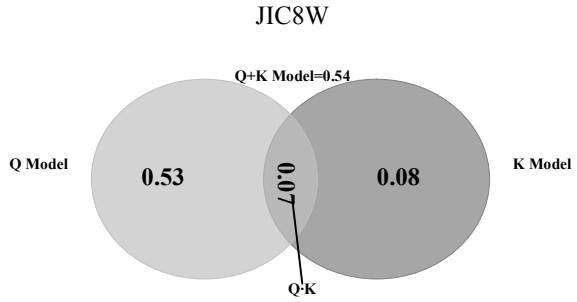


Figure 3.9 FRI Trait Involving Q and K Models

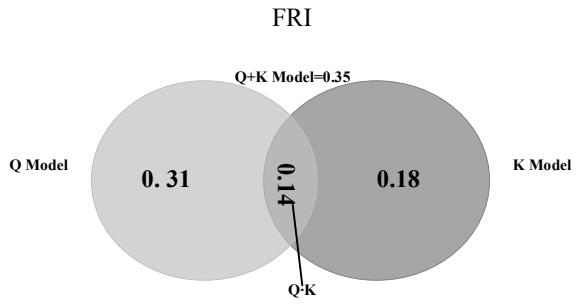


Figure 3.10 SDV Trait Involving P and K Models

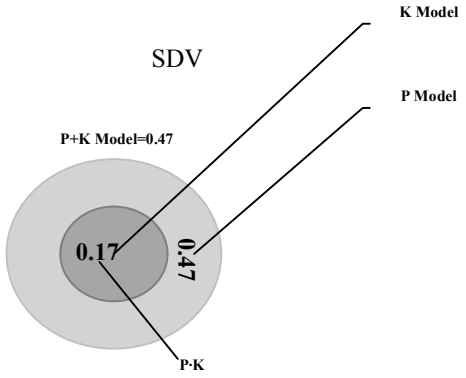


Figure 3.11 JIC8W Trait Involving P and K Models

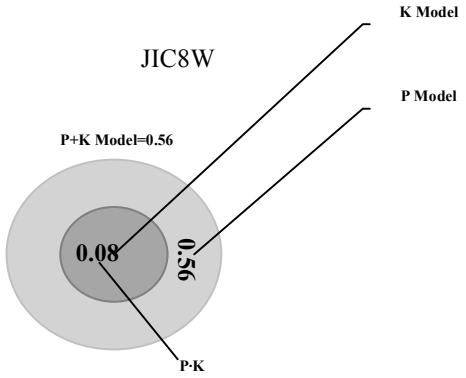


Figure 3.12 FRI Trait Involving P and K Models

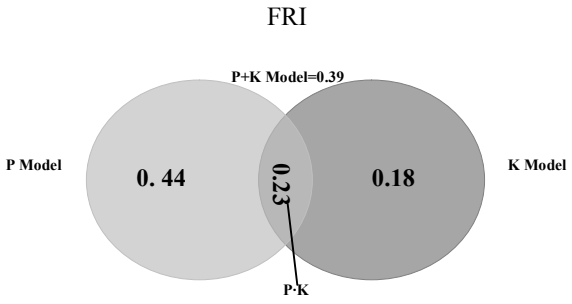


Table 3.15 R^2_W Values for Three Traits in Different Models

R^2_W	SDV	JIC8W	FRI
Q+K	0.28	0.49	0.58
Q	0.37	0.53	0.31
K	0.0000	0.00005	0.0011
P+K	0.42	0.53	0.27
P	0.47	0.56	0.44

Table 3.16 Marginal R_C Values for Three Traits in Different Models

R_C	SDV	JIC8W	FRI
Q+K	0.56	0.69	0.58
Q	0.54	0.69	0.48
K	0	0	0
P+K	0.64	0.72	0.67
P	0.64	0.72	0.60

From Table 3.16, marginal R_C statistic for K model is near to zero. Because the fitted value of Y is the mean in this case, the R_C value equals to zero based on the definition. Therefore, R_C cannot provide useful information to the researchers.

When we use ML method to compute the estimates of the variance components of K models, all three traits encountered the problems of non-convergence. So we set the parameters from Q+K model into K model to make the iteration convergent.

Table 3.17 Conditional R_C Values for Three Traits in Different Models

R_C	SDV	JIC8W	FRI
Q+K	0.74	0.80	0.60
Q	0.54	0.69	0.48
K	0.58	0.45	0.49
P+K	0.71	0.79	0.60
P	0.64	0.72	0.60

Table 3.18 P_{rand} Values for Three Traits in Different Models

P_{rand}	SDV	JIC8W	FRI
Q+K	0.61	0.68	0.66
Q	0.37	0.53	0.31
K	0.47	0.36	0.49

P+K	0.56	0.65	0.78
P	0.47	0.56	0.44

Table 3.19 r^2 **Values for Three Traits in Different Models**

r^2	SDV	JIC8W	FRI
Q+K	0.54	0.63	0.50
Q	0.37	0.53	0.31
K	0.36	0.25	0.37
P+K	0.52	0.62	0.50
P	0.47	0.56	0.44

Table 3.20 R_2^2 **Values for Three Traits in Different Models**

R_2^2	SDV	JIC8W	FRI
Q+K	0.62	0.69	0.33
Q	0.37	0.53	0.31
K	0.47	0.37	0.21
P+K	0.56	0.66	0.28
P	0.47	0.56	0.44

Table 3.21 ρ^2 **Values for Three Traits in Different Models**

ρ^2	SDV	JIC8W	FRI
Q+K	0.61	0.69	0.31
Q	0.37	0.53	0.31
K	0.46	0.36	0.19
P+K	0.56	0.66	0.22
P	0.47	0.56	0.44

From Table 3.20 and 3.21, we found that for trait FRI, the R^2 statistics values of P models are larger than that of P+K models. Because the non-convergence of the P+K model, we estimated the variance components of P+K model by those obtained from estimating from Q+K

model. However, with this kind of procedure, the variance component estimates may not be valid, which will lead to a "contradiction" with the non-decreasing criterion. Actually, it is not a contradiction and we could find a better set of parameters to make the non-decreasing criterion holds. Also, another reason is the difference of the variance covariance structures for mixed models and fixed effect models.

From Table 3.17 to Table 3.21, conditional R_C, P_{rand}, r^2, R_2^2,and ρ^2 values of the K model no longer larger than the mixed model as in Maize data example which means the violation of the non-decreasing criterion does not always happen. Although based on the Arabidopsis data example, our results did not show the contradiction with the non-decreasing criterion, we still prefer to R_{LR}^2 statistic since the other R^2-like statistics are not stable and reliable for using.

We could also plot the figures of R_{LR}^2 values in the same way as in Maize data example to address the SNPs markers to the phenotype traits.

CHAPTER 4 - Discussion

After reviewing seven R^2-like statistics and applying these statistics to two empirical data sets, we have obtained some general impression about various R^2-like statistics for measuring the goodness-of-fit and prediction power of a mixed effect linear model. Notice that R_C statistic is the only R^2-like statistic that has both the marginal and conditional version, and it is also the only R^2-like statistic that could be used both in linear and nonlinear mixed effect models.

For R_W^2, R_2^2, and ρ^2 statistics, for some traits the R^2 values of the fixed effect models are larger than those of the corresponding mixed effect models. The different covariance structures of the two models are the key reasons for this contradiction. Moreover the requirement for the fixed variance components to the nonnegative for the mixed models are also the reasons during handling the non-convergence problem in the iteration to obtain the estimates of the parameters of the mixed effect model using SAS.

For R_C, P_{rand}, r^2, R_2^2,and ρ^2 statistics, the R^2 values of the random effect models are larger than those of the corresponding mixed effect models. Although the random-effect models and the mixed effect models have the same random components, their variance components are estimated from two different models and they will be different in general, especially when the fixed effect terms have significant impact on the response y. This will cause the failure of R^2 statistics satisfying the monotone increasing property. If the variance components are known or given and the same variance components are used in the two models to compute the R^2 statistics, then the nonstandard R^2 statistics will have the monotone increasing property. Hence, we should use the variance components estimated in the mixed effect models to compute the R^2 statistics. This approach make sense because if the fixed effect term is significant, then the variance components estimated from the K model are not correct and should not be used in the calculation of R^2 statistics for the K model, and the variance components estimated from the mixed effect models should be used in both K and Q+K models.

Since R_{LR}^2 is based on the likelihood principle where the log-likelihood function increases as the number of parameters increases, it would not have the non-monotone increasing

problem as the others have. Therefore, R^2_{LR} statistic seems to be the preferred R^2-like statistic for the mixed effect models, in our study of association mapping.

Xu (2003) studied the R^2-like statistics such as r^2, R^2_2,and ρ^2 through Monte Carlo simulations. In the study of the behavior of these three R^2-like statistics, he concluded that r^2, R^2_2,and ρ^2 adequately quantify the predictability of the variables as given by the fixed and random effects. r^2 could give accurate estimates of the population Ω^2 with small or large sample sizes. R^2_2 and ρ^2 could give good estimates of the population Ω^2 with large cluster sizes, but overestimate Ω^2 with small sample sizes.

Orelien (2008) reported the performance of the R^2-like statistics such as R_C, P_{rand}, r^2, R^2_2,and ρ^2 by method of simulation. All models involved have the same random effects. The results of their simulation is that R^2-like statistics that involve the residuals are unable to adequately distinguish between the right model and the model without important fixed effects when the random effects are included to compute the fitted values. And they also demonstrated that the R^2-like statistics proposed by Xu (2003) behave poorly since the variation is little in r^2, R^2_2,and ρ^2 from the model of interest (the full model) to a reduced model.

When we did the review of seven R^2-like statistics, we found a mistake in the paper written by Orelien and Edwards (2008). They asserted that the concept of conditional and marginal R^2 could be applied to other statistics such as P_{rand} and R^2_2. However, the discussion in Section 2.4 reviews that this statement is not quite true.

Based on the results of the empirical analysis, we conclude that the R^2_{LR} statistic is the most useful R^2-like statistic for mixed effect models in association mapping. The other six statistics violate the non-decreasing criterion for R^2 statistic.

References

1. Breslow, N. E., Clayton, D. G. 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88: 9-25.
2. Buse, A. 1973. Goodness of fit in generalized least squares estimation. The American Statistician 27: 106-108.
3. Cameron, C., Windmeijer, F. A. G. 1996. R-squared measures for count data regression models with applications to health-care utilization. Journal of Business and Economic Statistics 14: 209-220.
4. Cox, D. R., Snell, E. J. 1989. Analysis of binary data (2^{nd} Edition). London: Chapman and Hall.
5. Gurland J. 1968. A relatively simple form of the distribution of the multiple correlation coefficient. Journal of the Royal Statistical Society B30: 276-283.
6. Helland, I. S. 1987. On the interpretation and use of R^2 in regression analysis. Biometrics 43: 61-69.
7. Hirschhorn, J. N., Daly, M. J. 2005. Genome-wide association studies for common diseases and complex traits. Nature Genetics 6: 95-108.
8. Kent, J. T. 1983. Information gain and a general measure of correlation. Biometrika 70 (1): 163-171.
9. Kramer, M. 2005. R^2 statistics for mixed models. Proceedings of the Conference on Applied Statistics in Agriculture: 148.e
10. Kvalseth, T. O. 1985. Cautionary note about R^2. The American Statistician 39: 279-285.
11. Maddala, G. S. 1983. Limited-Dependent and Qualitative Variables in Econometrics. Econometric Society Monographs.
12. Magee, L. 1990. R^2 measures based on Wald and likelihood ratio joint significance tests. The American Statistician 44: 250-253.
13. Nagelkerke, N. J. D. 1991. A note on a general definition of the coefficient of determination. Biometrika 78: 691-692.

14. Orelien, J. G., Edwards, L. J. 2008. Fixed-effect variable selection in linear mixed models using R^2 statistics. Computational Statistics & Data Analysis 52 (4): 1896-1907.
15. Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick N. A., Reich, D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38: 904-909.
16. Pritchard, J. K., Stephens, M., Donnelly, P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945-959.
17. Vonesh, E. F., Chinchilli, V. M., Pu, K. 1996. Goodness-of-fit in generalized nonlinear mixed-effects models. Biometrics 52: 572-587.
18. Wikipedia web-search. http://en.wikipedia.org/wiki/Main_Page
19. Xu, R. 2003. Measuring explained variation in linear mixed effects models. Statistics in Medicine 22: 3527-3541.
20. Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., Buckler, E. S. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38: 203-208.
21. Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian, C., Zheng, H., Dean, C., Marjoram, P., Nordborg, M. 2007. An Arabidopsis example of association mapping in structured samples. PloS Genetics 3: e4.
22. Zheng, B. 2000. Summarizing the goodness of fit of generalized linear models for longitudinal data. Statistics in Medicine 19: 1265-1275.

Appendix A - Computer Codes of SAS

Maize Data Example

Models involving Q matrix

```
option nodate nonumber center ls=115 ps=55 notes ;
/*Infile data sets*/
data AltMaz;
infile 'E:\files\work\read paper2 mixed model\AltMaz.prn';
input  TAXA $ Trait1 @@;
run;
proc sort data=AltMaz ;
by taxa;
run;
data DPOLL;
infile 'E:\files\work\read paper2 mixed model\DPOLL.prn';
input  TAXA $ Trait2 @@;
run;
proc sort data=DPOLL ;
by taxa;
run;
data EarDia;
infile 'E:\files\work\read paper2 mixed model\EarDia.prn';
input  TAXA $ Trait3 @@;
run;
proc sort data=EarDia;
by taxa;
run;
Data Pheno_data ;
merge AltMaz DPOLL EarDia;
by taxa;
RUN;
%let Trait_N=3 ;
Data STRUCT ;
set 'E:\files\work\Rsqure\sas files\q.sas7bdat';
RUN;
proc sort;
by taxa;
run;
proc sort data=pheno_data ;
by taxa;
run;
Data Similarity_matrix ;
set 'E:\files\work\read paper2 mixed model\kinship2.sas7bdat' ;
run;
proc sort;
by taxa;
run;
```

```
/*macro part with model fitting*/
%MACRO JY247sgn ;
 %global T I N M K ;

data results;
trait=.;
run;
%Do T=1 %to 1;

dm "output" clear;

data pheno;
set pheno_data ;
trait=trait&T ;
drop trait1--trait&Trait_N ;
run;

data all;
merge pheno STRUCT   Similarity_matrix ;
by TAXA ;
if row=. then delete ;
RUN;

data new;
set all;
drop col1--col277 ;
run;

data all_2;
set all;
if G1=.   then delete ;
if trait=. then delete ;
drop trait G1--row ;
run;

proc transpose data=all_2 out=all_2 ;
run;

data new;
merge new all_2 ;
if G1=.   then delete ;
if trait=. then delete ;
row=_n_ ;
drop _NAME_ ;
last=999 ;
run;

data sim1;
set new ;
row=_n_ ;
drop  taxa trait  G1--G3 last ;
run;

proc transpose data=sim1 out=sim1 ;
run;

data sim1 ;
```

```
set sim1 ;
if _NAME_='parm' then delete ;
if _NAME_='row'  then delete ;
temp=-999 ;
parm=1 ;
row=_n_ ;
run;

data temp;
set sim1 ;
drop col1--temp ;

data sim1 ;
merge temp sim1 ;
by parm row ;
drop _NAME_ temp ;
row=_n_ ;
run;


data maize; /*the data set we will use to analyze */
merge new pheno_data ;
by taxa;
drop parm--last ;
if G1=.   then delete ;
run;

proc sort;
by taxa ;
run;

data critical;
bic=.;
aic=.;
Neg2LogLike=.;
trait=.;
run;


/*the Q+K model*/
/*Proc mixed data=maize METHOD=ML noprofile covtest itdetails ic ;
class Taxa ;
model trait = G1 G2/ solution outp=pred ddfm=satterth outpm=pred2;
random Taxa / type=lin(1) ldata=Sim1 vi v gi solution;
repeated/ r;
ods output  infocrit=critical invV=inverseV V=v R=R invG=inverseG
solutionR=er solutionF=mysolutiontable1;
run;*/


/* the intercept only model*/
/*Proc mixed data=maize METHOD=ML noprofile covtest itdetails ic ;
class Taxa ;
model trait=/solution outp=pred ddfm=satterth ;
repeated/ r;
ods output  infocrit=critical r=r;
run;*/
```

```
/* the Q model*/
/*Proc mixed data=maize METHOD=ML noprofile covtest itdetails ic ;
class Taxa ;
model trait = G1 G2 /solution outp=pred ddfm=satterth ;
repeated/r ri;
ods output  infocrit=critical R=R invR=inverseR solutionF=mysolutiontable1;
run;*/


/* the K model*/
Proc mixed data=maize METHOD=ML noprofile covtest itdetails ic ;
class Taxa ;
model trait = / solution outp=pred ddfm=satterth outpm=pred2;
random Taxa / type=lin(1) ldata=Sim1 vi v gi solution;
*parms (6.7141)(8.1834); /*trait2 parms from Q+K*/
repeated/ r;
ods output  infocrit=critical invV=inverseV v=v R=r solutionR=er
invG=inverseG;
run;



data inverseV_dr; set inverseV; drop index row; run;
data inverseG_dr; set inverseG; drop row effect taxa; run;
data R_dr; set R; drop index row; run;

data critical; set critical; trait=&T ; drop parms aicc hqic caic; run;

data results; merge results critical; by trait ; if trait=. then delete; run;
/*result table containing -2log-likelikhood values*/

%END;
%mend JY247sgn ;
run;
%JY247sgn
run;


/*compute R2 values except RLR using iml */
proc iml;
use R_dr; read all into R;
use inverseV_dr; read all into invV;
/*use pred2; read all var {resid} into u;*/  /*fitted Y consisting the fixed
effect*/
use pred2; read all var {pred} into p2;
use maize; read all var {trait} into y;
use pred; read all var {resid} into u;  /*fitted Y consisting the fixed
effect and the random effect*/
use pred; read all var {pred} into p;
use er; read all var {estimate} into er;
use inverseG_dr; read all into invG;
use inverseR; read all var {col1} into invR;

n=nrow(y);
*invR2=invR*i(n);
```

```
*m=t(u)*invV*u;
*m=t(u)*invR2*u;
m=t(u)*u;
*m=t(r)*invV*r;
*m=(t(u)*u)/(2*sqrt(R))+t(er)*invG*er*0.5;
m=(t(u)*u*sqrt(invR))/2;


total=sum(y);
avg=total/n;
ymean=avg*j(n,1);
y2=y-ymean;
*l=t(y2)*invV*y2;
*l=t(y2)*invR2*y2;
l=t(y2)*y2;
/*avgp=sum(p)/n;
pm=avgp*j(n,1);
y3=p-pm;
l=t(y2)*y2+t(y3)*y3+n*(avg-avgp)*(avg-avgp);*/
*l=(t(y2)*y2)/(2*sqrt(R));
*l=(t(y2)*y2*sqrt(invR))/2;

w=log(397.83/R)-(t(u)*u)/(n*R)+1; /*trait1*/
*w=log(31.8156/R)-(t(u)*u)/(n*R)+1; /*trait2*/
*w=log(16.2915/R)-(t(u)*u)/(n*R)+(t(y2)*y2)/(n*16.2915); /*trait3*/
r2=1-exp(-w);


R2s=1-m/l;
print r2 /*R2s*/;
```

Q+K Models with SNPs markers

```
option nodate nonumber center ls=115 ps=55 notes ;
/*Infile data sets*/
data AltMaz;
infile 'E:\files\work\read paper2 mixed model\AltMaz.prn';
input  TAXA $ Trait1 @@;
run;
proc sort data=AltMaz ;
by taxa;
run;
data DPOLL;
infile 'E:\files\work\read paper2 mixed model\DPOLL.prn';
input  TAXA $ Trait2 @@;
run;
proc sort data=DPOLL ;
by taxa;
run;
data EarDia;
infile 'E:\files\work\read paper2 mixed model\EarDia.prn';
input  TAXA $ Trait3 @@;
run;
proc sort data=EarDia;
by taxa;
run;
```

```
Data Pheno_data ;
merge AltMaz DPOLL EarDia;
by taxa;
RUN;
%let Trait_N=3 ;
Data STRUCT ;
set 'E:\files\work\Rsqure\sas files\q.sas7bdat';
RUN;
proc sort;
by taxa;
run;
proc sort data=pheno_data ;
by taxa;
run;
Data Similarity_matrix ;
set 'E:\files\work\read paper2 mixed model\kinship2.sas7bdat' ;
run;
proc sort;
by taxa;
run;
%let Trait_N=3 ;
%let SNP_N=553 ;   /*number of SNP in the data set*/

Data SNP_data ;
infile 'E:\files\work\read paper2 mixed model\snp553c.txt' expandtabs lrecl=
10000;
array SNP{&SNP_N} $ SNP1-SNP&SNP_N;
input TAXA$ SNP1-SNP&SNP_N @@;
run;


/*macro part with model fitting*/
%MACRO JY247 ;
 %global T S I N M K ;

data results;
trait=.;
assay=. ;
run;

data results2;
assay=. ;
run;
%Do T=1 %to &Trait_N ;
%DO S=1 %TO *&SNP_N;

%if &S>5 %then %do ;
option nonotes ;
%end;

dm "output" clear;

data pheno;
set pheno_data ;
trait=trait&T ;
drop trait1--trait&Trait_N ;
```

```
run;

data all;
merge pheno SNP_data STRUCT Similarity_matrix ;
by TAXA ;
SNP=SNP&S ;
if row=. then delete ;
RUN;

data new;
set all;
drop col1--snp ;
run;

data all_2;
set all;
if SNP='' then delete ;
if G1=.   then delete ;
if trait=. then delete ;
drop trait snp1--row snp ;
run;

proc transpose data=all_2 out=all_2 ;
run;

data new;
merge new all_2 ;
SNP=SNP&S ;
if SNP='' then delete ;
if G1=.   then delete ;
if trait=. then delete ;
row=_n_ ;
drop _NAME_ ;
last=999 ;
run;

data SNP;
set new ;
drop trait parm--last ;
SNP=SNP&S ;
run;

data sim1;
set new ;
row=_n_ ;
drop  taxa trait snp1--G3 snp last ;
run;

proc transpose data=sim1 out=sim1 ;
run;

data sim1 ;
set sim1 ;
if _NAME_='parm' then delete ;
if _NAME_='row'  then delete ;
temp=-999 ;
parm=1 ;
```

```
row=_n_ ;
run;

data temp;
set sim1 ;
drop col1--temp ;

data sim1 ;
merge temp sim1 ;
by parm row ;
drop _NAME_ temp ;
row=_n_ ;
run;

data maize;
merge SNP pheno_data ;
by taxa ;
trait=trait&T ;
run;

data maize;    /*the data set we will use to analyze */
set maize;
SNP=SNP&S ;
if SNP='' then delete ;
if G1=.   then delete ;
run;

proc sort;
by taxa ;
run;

data mysolutiontable1 ;
assay=.;
DF=-999;
SNP='?';
run;

data critical;
assay=.;
bic=.;
Neg2LogLike=.;
trait=.;
run;

Proc mixed data=maize METHOD=ML noprofile covtest itdetails ic ;
class Taxa SNP;
model trait = SNP G1 G2 /solution outp=pred ddfm=satterth ;
random Taxa / type=lin(1) ldata=Sim1 ;
ods output solutionF=mysolutiontable1 infocrit=critical ;
run;

data mysolutiontable1;
set mysolutiontable1;
if DF=. then delete;
if SNP='' then delete;
TRAIT=&T   ;
assay=&S ;
```

```
run;

data critical;
set critical;
TRAIT=&T   ;
assay=&S ;
drop parms aic aicc hqic caic;
run;


data results;
merge results mysolutiontable1 critical;
by trait assay ;
if assay=. then delete;
run;  /*result table containing -2log-likelikhood values*/

%END;
%end;
%mend JY247 ;
run;
%JY247
run;
```

Models involving P matrix

```
option nodate nonumber center  ls=115 ps=55 notes ;

/*Infile data sets*/
data AltMaz;
infile 'E:\files\work\read paper2 mixed model\AltMaz.prn';
input  TAXA $ Trait1 @@;
run;
proc sort data=AltMaz ;
by taxa;
run;
data DPOLL;
infile 'E:\files\work\read paper2 mixed model\DPOLL.prn';
input  TAXA $ Trait2 @@;
run;
proc sort data=DPOLL ;
by taxa;
run;
data EarDia;
infile 'E:\files\work\read paper2 mixed model\EarDia.prn';
input  TAXA $ Trait3 @@;
run;
proc sort data=EarDia;
by taxa;
run;
Data Pheno_data ;
merge AltMaz DPOLL EarDia;
by taxa;
RUN;
%let Trait_N=3 ;

proc sort data=pheno_data ;
```

```
by taxa;
run;
Data Similarity_matrix ;
set 'E:\files\work\read paper2 mixed model\kinship2.sas7bdat' ;
run;
proc sort;
by taxa;
run;

filename snpc "E:\files\work\Rsqure\sas files\snp553ci.txt";
data snp553c;
infile snpc  expandtabs lrecl= 100000;
array SNP{553} SNP1-SNP553;
input TAXA$ SNP1-SNP553;
run;
proc sort;
by taxa;
run;


proc princomp data=snp553c n=3 out=eigenvector ;
var SNP1-SNP553  ;
run;
data ptable;
set eigenvector;
keep taxa prin1-prin3;
if prin1=. then  delete;
run;
proc sort;
by taxa;
run;

/*macro part with model fitting*/
%MACRO JY247sgn ;
 %global T  I N M K ;

data results;
trait=.;
run;
%Do T=3 %to 3;*&Trait_N;

dm "output" clear;

data pheno;
set pheno_data ;
trait=trait&T ;
drop trait1--trait&Trait_N ;
run;

data all;
merge pheno ptable   Similarity_matrix ;
by TAXA ;
if row=. then delete ;
RUN;

data new;
set all;
```

```
drop col1--col277 ;
run;

data all_2;
set all;
if prin1=. then delete;
if trait=. then delete ;
drop trait prin1--row ;
run;

proc transpose data=all_2 out=all_2 ;
run;

data new;
merge new all_2 ;
if prin1=. then delete;
if trait=. then delete ;
row=_n_ ;
drop _NAME_ ;
last=999 ;
run;

data sim1;
set new ;
row=_n_ ;
drop  taxa trait prin1-prin3  last ;
run;

proc transpose data=sim1 out=sim1 ;
run;

data sim1 ;
set sim1 ;
if _NAME_='parm' then delete ;
if _NAME_='row'  then delete ;
temp=-999 ;
parm=1 ;
row=_n_ ;
run;

data temp;
set sim1 ;
drop col1--temp ;

data sim1 ;
merge temp sim1 ;
by parm row ;
drop _NAME_ temp ;
row=_n_ ;
run;


data maize; /*the data set we will use to analyze */
merge new pheno_data ;
by taxa;
drop parm--last ;
if prin1=. then delete;
```

```
run;

proc sort;
by taxa ;
run;

data critical;
bic=.;
aic=.;
Neg2LogLike=.;
trait=.;
run;


/*the P+K model*/
Proc mixed data=maize METHOD=ML noprofile covtest itdetails ic ;
class Taxa ;
model trait = prin1 prin2 / solution outp=pred ddfm=satterth outpm=pred2;
random Taxa / type=lin(1) ldata=Sim1 vi v gi solution;
repeated/ r;
ods output  infocrit=critical invV=inverseV V=v R=R invG=inverseG
solutionR=er solutionF=mysolutiontable1;
run;


/* the P model*/
/*Proc mixed data=maize METHOD=ML noprofile covtest itdetails ic ;
class Taxa ;
model trait = Prin1 Prin2 /solution outp=pred ddfm=satterth ;
repeated/ r ri;
ods output  infocrit=critical R=R invR=inverseR solutionF=mysolutiontable1;
run;*/


data inverseV_dr; set inverseV; drop index row; run;
data inverseG_dr; set inverseG; drop row effect taxa; run;
data R_dr; set R; drop index row; run;

data critical; set critical; trait=&T ; drop parms aicc hqic caic; run;

data results; merge results critical; by trait ; if trait=. then delete; run;
/*result table containing -2log-likelikhood values*/

%END;
%mend JY247sgn ;
run;
%JY247sgn

run;

/*compute R2 values except RLR using iml */
proc iml;
use R_dr; read all into R;
use inverseV_dr; read all into invV;
/*use pred2; read all var {resid} into u;*/  /*fitted Y consisting the fixed
effect*/
use pred2; read all var {pred} into p2;
```

```
use maize; read all var {trait} into y;
use pred; read all var {resid} into u;  /*fitted Y consisting the fixed
effect and the random effect*/
use pred; read all var {pred} into p;
use er; read all var {estimate} into er;
use inverseG_dr; read all into invG;
use inverseR; read all var {col1} into invR;

n=nrow(y);
*invR2=invR*i(n);
*m=t(u)*invV*u;
*m=t(u)*invR2*u;
m=t(u)*u;
*m=t(r)*invV*r;
*m=(t(u)*u)/(2*sqrt(R))+t(er)*invG*er*0.5;
m=(t(u)*u*sqrt(invR))/2;


total=sum(y);
avg=total/n;
ymean=avg*j(n,1);
y2=y-ymean;
*l=t(y2)*invV*y2;
*l=t(y2)*invR2*y2;
l=t(y2)*y2;
/*avgp=sum(p)/n;
pm=avgp*j(n,1);
y3=p-pm;
l=t(y2)*y2+t(y3)*y3+n*(avg-avgp)*(avg-avgp);*/
*l=(t(y2)*y2)/(2*sqrt(R));
*l=(t(y2)*y2*sqrt(invR))/2;

w=log(397.83/R)-(t(u)*u)/(n*R)+1; /*trait1*/
*w=log(31.8156/R)-(t(u)*u)/(n*R)+1; /*trait2*/
*w=log(16.2915/R)-(t(u)*u)/(n*R)+1; /*trait3*/
r2=1-exp(-w);


R2s=1-m/l;
print r2 /*R2s*/;
```

Arabidopsis Data Example

Models involving Q matrix

```
option nodate nonumber center  ls=115 ps=55 notes ;

/*Infile data sets*/
data phenodata;
set 'E:\files\work\Rsqure\sas files\phenodata.sas7bdat';
run;
proc sort;
by accession;
run;
%let Trait_N=4 ;
```

```
Data STRUCT ;
set 'E:\files\work\Rsqure\sas files\structsnp.sas7bdat';
RUN;

proc sort;
by accession;
run;
Data Similarity_matrix ;
set 'E:\files\work\Rsqure\sas files\similatrity_matrixsnp.sas7bdat' ;
run;

proc sort;
by accession;
run;

%MACRO JY247sgn2 ;
 %global T I N M K ;

data results;
trait=.;
run;
%Do T=1 %to 1;

dm "output" clear;

data pheno;
set phenodata ;
trait=trait&T ;
drop trait1--trait&Trait_N ;
run;

data all;
merge pheno  Struct  Similarity_matrix ;
by accession ;
if row=. then delete ;
RUN;

data new;
set all;
drop col1--col95 ;
run;

data all_2;
set all;
if Q1=.   then delete ;
if trait=. then delete ;
drop trait Q1--row ;
run;

proc transpose data=all_2 out=all_2 ;
run;

data new;
merge new all_2 ;
if Q1=.   then delete ;
if trait=. then delete ;
row=_n_ ;
```

```
drop _NAME_ ;
last=999 ;
run;

data sim1;
set new ;
row=_n_ ;
drop  accession trait Q1--Q8 last ;
run;

proc transpose data=sim1 out=sim1 ;
run;

data sim1 ;
set sim1 ;
if _NAME_='parm' then delete ;
if _NAME_='row'  then delete ;
temp=-999 ;
parm=1 ;
row=_n_ ;
run;

data temp;
set sim1 ;
drop col1--temp ;

data sim1 ;
merge temp sim1 ;
by parm row ;
drop _NAME_ temp ;
row=_n_ ;
run;


data arabidopsis; /*the data set we will use to analyze */
merge new phenodata ;
by accession;
drop parm--last ;
if Q1=.   then delete ;
run;

proc sort;
by accession ;
run;

data critical;
bic=.;
aic=.;
Neg2LogLike=.;
trait=.;
run;


/*the Q+K model*/
Proc mixed data=arabidopsis METHOD=ML noprofile covtest itdetails ic ;
class accession ;
model trait = Q1-Q7 /solution outp=pred ddfm=satterth outpm=pred2;
```

```
random accession / type=lin(1) ldata=Sim1 vi v gi solution;
repeated/ r ri;
ods output  infocrit=critical invV=inverseV V=v R=R invG=inverseG
solutionR=er;
run;


/* the intercept only model*/
/*Proc mixed data=arabidopsis METHOD=ML noprofile covtest itdetails ic ;
class accession ;
model trait=/solution outp=pred ddfm=satterth ;
ods output  infocrit=critical ;
run;*/

/* the Q model*/
/*Proc mixed data=arabidopsis METHOD=ML noprofile covtest itdetails ic ;
class accession ;
model trait = Q1-Q7 /solution outp=pred ddfm=satterth  ;
repeated / ri r;
ods output  infocrit=critical invR=inverseR R=r ;
run;*/

/* the K model*/
/*Proc mixed data=arabidopsis METHOD=ML noprofile covtest itdetails ic ;
class accession ;
model trait = /solution outp=pred ddfm=satterth outpm=pred2;
random accession / type=lin(1) ldata=Sim1 vi v gi solution;
*parms (102.87)(536.43);/*tait1*/
*parms (5.1536)(32.1574);/*trait3*/
parms (0.06714)(0.2674);/*trait4*/
repeated/ r;
ods output infocrit=critical invV=inverseV V=v R=R invG=inverseG solutionR=er;
run;*/

data inverseV_dr; set inverseV; drop index row; run;
data inverseG_dr; set inverseG; drop row effect taxa; run;
data inverseR; set inverseR; drop index row; run;
data R_dr; set R; drop index row; run;

data critical; set critical; trait=&T ; drop parms aicc hqic caic; run;

data results;
merge results   critical;
by trait ;
if trait=. then delete;
run;  /*result table containing -2log-likelikhood values*/

%END;
%mend JY247sgn2 ;
run;
%JY247sgn2

run;

/*compute R2 values except RLR using iml */
proc iml;
use R_dr; read all into R;
```

```
use inverseV_dr; read all into invV;
/*use pred2; read all var {resid} into u;*/  /*fitted Y consisting the fixed
effect*/
use pred2; read all var {pred} into p2;
use arabidopsis; read all var {trait} into y;
use pred; read all var {resid} into u;  /*fitted Y consisting the fixed
effect and the random effect*/
use pred; read all var {pred} into p;
use er; read all var {estimate} into er;
use inverseG_dr; read all into invG;
use inverseR; read all var {col1} into invR;

n=nrow(y);
*invR2=invR*i(n);
*m=t(u)*invV*u;
*m=t(u)*invR2*u;
m=t(u)*u;
*m=t(r)*invV*r;
*m=(t(u)*u)/(2*sqrt(R))+t(er)*invG*er*0.5;
*m=(t(u)*u*sqrt(invR))/2;


total=sum(y);
avg=total/n;
ymean=avg*j(n,1);
y2=y-ymean;
*l=t(y2)*invV*y2;
*l=t(y2)*invR2*y2;
l=t(y2)*y2;
/*avgp=sum(p)/n;
pm=avgp*j(n,1);
y3=p-pm;
l=t(y2)*y2+t(y3)*y3+n*(avg-avgp)*(avg-avgp);*/
*l=(t(y2)*y2)/(2*sqrt(R));
*l=(t(y2)*y2*sqrt(invR))/2;

w=log(1157.27/R)-(t(u)*u)/(n*R)+1; /*trait1*/
*w=log(87.602/R)-(t(u)*u)/(n*R)+1; /*trait3*/
*w=log(0.5381/R)-(t(u)*u)/(n*R)+1; /*trait4*/
r2=1-exp(-w);

R2s=1-m/l;
print m l  r2/* R2s */;
```

Models involving P matrix

```
option nodate nonumber center  ls=115 ps=55 notes ;

/*Infile data sets*/
data phenodata;
set 'E:\files\work\Rsqure\sas files\phenodata.sas7bdat';
run;
proc sort;
by accession;
run;
%let Trait_N=4 ;
```

```
Data Similarity_matrix ;
set 'E:\files\work\Rsqure\sas files\similatrity_matrixsnp.sas7bdat' ;
run;

proc sort;
by accession;
run;

/*data arabidopsis_snp;
infile "E:\files\work\Rsqure\sas files\snp5419c.txt" expandtabs lrecl= 100000;
array snp{5419} snp1-snp5419;
input accession $ snp1-snp5419;
run;

proc princomp data=arabidopsis_snp n=8 out=eigenvector noprint;
var SNP1-SNP5419  ;
run;
data ptable;
set eigenvector;
keep accession prin1-prin8;
if prin1=. then  delete;
run;
proc sort;
by accession;
run;*/

data ptable;
set 'E:\files\work\Rsqure\sas files\ptable.sas7bdat';
run;
proc sort;
by accession;
run;

/*macro part with model fitting*/
%MACRO JY247sgn2 ;
 %global T I N M K ;

data results;
trait=.;
run;
%Do T=4 %to 4;*&Trait_N;

dm "output" clear;

data pheno;
set phenodata ;
trait=trait&T ;
drop trait1--trait&Trait_N ;
run;

data all;
merge pheno  ptable Similarity_matrix ;
by accession ;
if row=. then delete ;
RUN;

data new;
```

```
set all;
drop col1--col95 ;
run;

data all_2;
set all;
if Prin1=.   then delete ;
*if P1=.   then delete ;
if trait=. then delete ;
drop trait Prin1--row ;
run;

proc transpose data=all_2 out=all_2 ;
run;

data new;
merge new all_2 ;
if Prin1=.   then delete ;
*if P1=.   then delete ;
if trait=. then delete ;
row=_n_ ;
drop _NAME_ ;
last=999 ;
run;

data sim1;
set new ;
row=_n_ ;
drop  accession trait /* P1-P8*/ prin1--prin8 last ;
run;

proc transpose data=sim1 out=sim1 ;
run;

data sim1 ;
set sim1 ;
if _NAME_='parm' then delete ;
if _NAME_='row'  then delete ;
temp=-999 ;
parm=1 ;
row=_n_ ;
run;

data temp;
set sim1 ;
drop col1--temp ;

data sim1 ;
merge temp sim1 ;
by parm row ;
drop _NAME_ temp ;
row=_n_ ;
run;


data arabidopsis;  /*the data set we will use to analyze */
merge new phenodata ;
```

```
by accession;
drop parm--last ;
if Prin1=.   then delete ;
*if P1=.   then delete ;
run;

proc sort;
by accession ;
run;

data critical;
bic=.;
aic=.;
Neg2LogLike=.;
trait=.;
run;


/*the P+K model*/
Proc mixed data=arabidopsis METHOD=ML noprofile covtest itdetails ic ;
class accession ;
model trait =prin1-prin7 /solution outp=pred ddfm=satterth outpm=pred2;
random accession / type=lin(1) ldata=Sim1 vi v gi solution;
parms (0.06714)(0.2674)/noiter;/*trait4*/
repeated/ r;
ods output  infocrit=critical invV=inverseV V=v R=R invG=inverseG
solutionR=er;
run;


/* the P model*/
/*Proc mixed data=arabidopsis METHOD=ML noprofile covtest itdetails ic ;
class accession ;
model trait = Prin1-Prin7 /solution outp=pred ddfm=satterth ;
repeated/ ri;
ods output  infocrit=critical invR=inverseR;
run;*/

data inverseV_dr; set inverseV; drop index row; run;
data inverseG_dr; set inverseG; drop row effect taxa; run;
data inverseR; set inverseR; drop index row; run;
data R_dr; set R; drop index row; run;

data critical; set critical; trait=&T ; drop parms aicc hqic caic; run;

data results;
merge results   critical;
by trait ;
if trait=. then delete;
run;
/*result table containing -2log-likelikhood values*/


%END;
%mend JY247sgn2 ;
run;
%JY247sgn2
```

```
run;


/*compute R2 values except RLR using iml */
proc iml;
use R_dr; read all into R;
use inverseV_dr; read all into invV;
/*use pred2; read all var {resid} into u;*/  /*fitted Y consisting the fixed
effect*/
use pred2; read all var {pred} into p2;
use arabidopsis; read all var {trait} into y;
use pred; read all var {resid} into u;  /*fitted Y consisting the fixed
effect and the random effect*/
use pred; read all var {pred} into p;
use er; read all var {estimate} into er;
use inverseG_dr; read all into invG;
use inverseR; read all var {col1} into invR;

n=nrow(y);
*invR2=invR*i(n);
*m=t(u)*invV*u;
*m=t(u)*invR2*u;
m=t(u)*u;
*m=t(r)*invV*r;
*m=(t(u)*u)/(2*sqrt(R))+t(er)*invG*er*0.5;
*m=(t(u)*u*sqrt(invR))/2;


total=sum(y);
avg=total/n;
ymean=avg*j(n,1);
y2=y-ymean;
*l=t(y2)*invV*y2;
*l=t(y2)*invR2*y2;
l=t(y2)*y2;
/*avgp=sum(p)/n;
pm=avgp*j(n,1);
y3=p-pm;
l=t(y2)*y2+t(y3)*y3+n*(avg-avgp)*(avg-avgp);*/
*l=(t(y2)*y2)/(2*sqrt(R));
*l=(t(y2)*y2*sqrt(invR))/2;

w=log(1157.27/R)-(t(u)*u)/(n*R)+1; /*trait1*/
*w=log(87.602/R)-(t(u)*u)/(n*R)+1; /*trait3*/
*w=log(0.5381/R)-(t(u)*u)/(n*R)+1; /*trait4*/
r2=1-exp(-w);

R2s=1-m/l;
print m l  r2/* R2s */;
```