

A few results in linear and mixed models with applications to complex trait analysis

Nicholas A. Furlotte

January 31, 2011

Linear models and linear mixed models have a long history in the analysis of complex traits. In my research, I have found it difficult to pin down certain results as there are many references each with a different style. I created this document as sort of a tutorial as well as a reference for certain important results. I begin by explaining ordinary least squares (OLS) and then show how these results can be extended into generalized least squares (GLS). Following these basic results, I introduce the basic variance components models often used in the analysis of complex traits in genetics.

1 Ordinary Least Squares (OLS)

1.1 A Simple Model

Let us assume that we are given phenotype measurements for n individuals and we denote each measurement by y_i . The simplest generative model for this data is given in equation 1.

$$\begin{aligned}y_i &= \mu + e_i \\ e_i &\sim N(0, \sigma_e^2)\end{aligned}\tag{1}$$

In equation 1, we assume that each phenotype measurement is simply a function of a mean value μ and a random error. We assume that the random error is normally distributed with mean 0 and variance σ_e^2 . In order to estimate the two unknown parameters we adopt a framework such as analysis of variance (ANOVA) or maximum likelihood (ML). The solutions are the same and are very well-known. The optimal estimates $\hat{\mu}$ and $\hat{\sigma}_e^2$ for μ and σ_e^2 are given as follows. These solutions should be recognizable as the sample mean and sample standard deviation introduced in basic statistics classes.

$$\hat{\mu} = \sum_i \frac{y_i}{n}\tag{2}$$

$$\hat{\sigma}_e^2 = \sum_i \frac{(y_i - \hat{\mu})^2}{n-1}\tag{3}$$

1.2 A Less Simple Model

It is reasonable to assume that we might want to include other observable variables in the model for our phenotype. For example, sex or age very often effect phenotypes and it will do us well to include them our generative models. The model in equation (4), shows a general view of a model of this type. In this model, the phenotype y_i is assumed to be a function of global mean (μ) and a set of $q - 1$ other known covariates, along with some error. The value of the j th covariate is denoted by x_j and its coefficient is denoted as β_j .

$$y_i = \mu + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{q-1} x_{q-1} + e\tag{4}$$

Going forward it will be beneficial for us to adopt matrix notation as the results are more easily obtained and more easily understood in this format. Equation (5) gives the general form of an OLS model, where \mathbf{X} is an $n \times q$ matrix encoding the global mean and $q - 1$ fixed effects, β is a coefficient vector of size q and \mathbf{e} is a random variable assumed to follow a multivariate normal distribution with mean zero and variance $\sigma_e^2 \mathbf{I}$.

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \mathbf{e} \\ \mathbf{e} &\sim N(0, \sigma_e^2 \mathbf{I}) \end{aligned} \tag{5}$$

In order to find the vector of coefficients, we define a simple objective function that we would like to minimize. The β that maximizes the likelihood of the observed data is that beta that minimizes the deviations of the predictions made given the vector to that of the actual data. To do this we find the β that minimizes the following equation.

$$Q(\beta) = \frac{1}{n}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \tag{6}$$

Differentiating this equation we get the following.

$$\frac{\partial Q}{\partial \beta} = -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta)/n = 0 \tag{7}$$

$$\tag{8}$$

Setting equal to zero and solving, we arrive at the well-known maximum likelihood solution for β , $\hat{\beta}$.

$$\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{y} \tag{9}$$

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{10}$$

This $\hat{\beta}$ is our estimate for the true β . We also want to estimate the variance σ_e^2 . To do this we can simply sum of squared residuals and divide by the degrees of freedom.

$$\hat{\sigma}_e^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})}{n - q} \tag{11}$$

These are the basic OLS results. Now let us look at some properties of these results.

1.3 Best Linear Unbiased Estimates (BLUEs)

Given the above results, we can look at some interesting properties which will help us to derive some statistics later. Most of this section is coming from [1].

Theorem 1.1 (Guass-Markov) $\hat{\beta}$ is a best linear unbiased estimator for β if the following conditions hold.

1. \mathbf{X} is non-stochastic
2. \mathbf{y} is a random vector such that $E(\mathbf{y}) = \mathbf{X}\beta$ and $\text{var}(\mathbf{y}) = \sigma_e^2 \mathbf{I}$ for some $\sigma_e^2 > 0$.

Further, if we assume normality, that is, if we assume that $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma_e^2 \mathbf{I})$, then we can also state the following without proof.

Theorem 1.2 Stuff that is useful about our estimates

1. $\hat{\beta} \sim N(\beta, \sigma_e^2(\mathbf{X}'\mathbf{X})^{-1})$
2. $(n - q)\hat{\sigma}_e^2/\sigma_e^2 \sim \chi^2(n - q)$
3. $E(\hat{\sigma}_e^2) = \sigma_e^2$ and $\text{var}(\hat{\sigma}_e^2) = 2\sigma_e^4/(n - q)$.

1.4 Hypothesis Testing

Given the estimates for β and σ_e^2 , we most likely will want to test a hypothesis about our estimates. For example, we might write the model for our phenotype to include both mean and some SNP effect. In this case, we will want to test the significance of our SNP effect. In order to define our test statistics, we first need to consider some properties of our estimates. In particular, let us consider the estimate for β , $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. If the inverse of $\mathbf{X}'\mathbf{X}$ does not exist, then we might use a pseudo-inverse. However, our estimate of β is very sensitive to our choice of pseudo-inverse and there are many. The value of $\mathbf{X}\hat{\beta}$ does not vary with choice of pseudo-inverse however, so we choose to test hypothesis on this product rather than just $\hat{\beta}$.

With this in mind, let us consider how to test the hypothesis that $\mathbf{R}\mathbf{X}\beta = \mathbf{r}$, where \mathbf{R} is a $q \times n$ matrix, \mathbf{X} is the $n \times q$, β is our $q \times 1$ vector of coefficients and \mathbf{r} is a $q \times 1$ vector of hypothetical coefficient values. The reason for this complicated test is of course because we need to use $\mathbf{X}\beta$ in order to guarantee invariance to choice of pseudo-inverse. We choose \mathbf{R} in such a way that we can obtain the values of our hypothetical vector.

Example 1.1 Choosing \mathbf{R}

For simplicity of notation, we will assume that \mathbf{R} has the form $\mathbf{M}\mathbf{X}$, so that it meets the requirements set forth previously and will subsequently test the hypothesis that $\mathbf{R}\beta = \mathbf{r}$. Assuming that $\hat{\beta}$ is normally distributed, we know that $\mathbf{R}\hat{\beta}$ is also normally distributed, as normality is preserved under linear transformations. This leads us to the following conclusions.

$$\hat{\beta} \sim N(\beta, \sigma_e^2(\mathbf{X}'\mathbf{X})^{-1}) \quad (12)$$

$$\mathbf{R}\hat{\beta} \sim N(\mathbf{R}\beta, \sigma_e^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}') \quad (13)$$

Given these we can then show that when testing the hypothesis $\mathbf{R}\beta = \mathbf{r}$, we can define the following statistic.

$$(\sigma_e^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-\frac{1}{2}}(\mathbf{R}\hat{\beta} - \mathbf{r}) \sim N(0, I) \quad (14)$$

And subsequently

$$(\mathbf{R}\hat{\beta} - \mathbf{r})'(\sigma_e^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\hat{\beta} - \mathbf{r}) \sim \chi^2(q) \quad (15)$$

Of course, we cannot use this statistic unless we know the actual value of σ_e^2 . We can use the estimate, but in order to do so we must account for the uncertainty there. We stated earlier that $(n-q)\hat{\sigma}_e^2/\sigma_e^2 \sim \chi^2(n-q)$, so we can use this to derive an F-statistic. As a reminder, an F-statistic is the ratio of two χ^2 -statistics that have been divided by their respective degrees of freedom.

$$\phi = \frac{(\mathbf{R}\hat{\beta} - \mathbf{r})'(\sigma_e^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\hat{\beta} - \mathbf{r})/q}{(n-q)\hat{\sigma}_e^2/\sigma_e^2(n-q)} \quad (16)$$

$$= \frac{(\mathbf{R}\hat{\beta} - \mathbf{r})'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\hat{\beta} - \mathbf{r})}{q\hat{\sigma}_e^2} \quad (17)$$

With this we can then claim that $\phi \sim \mathcal{F}(q, n-q)$. Under the null hypothesis this statistic follows the prescribed \mathcal{F} -distribution, so given a level of significance we can assess the significance of our SNP or other covariate of interest.

1.5 Statistical Power

Statistical power is defined as the probability of obtaining a statistic at a particular significance level, or alternatively rejecting the null when it is in fact false. For example, if we know that we can calculate an F-statistic on our $\hat{\beta}$ and we know that when we observe a statistic that is greater than 3.14 we will reject the null hypothesis, then the power is defined as the probability of observing a statistic greater than 3.14. This is different than a p-value, as a p-value is the probability of observing a statistic as extreme under the null. For power, we must consider the alternative distribution. That is, we must be able to integrate over the alternative distribution from the value 3.14 to the end of the distribution (∞).

One motivation, which will be developed later, for computing statistical power is as follows. We may want to know what our power to discover some causal SNP is under some particular experimental setup. In order to compute such a number, we must make a number of assumptions, but given that this has been done, we simply need to figure out what our alternative distribution for the β is. That is, we need to determine what distribution our computed β statistic follows when the SNP we are testing is in fact causal. The simplest thing that we can do is to assume that our statistic still follows the distribution that we assume it to follow under the null, except that the mean has been shifted by a non-centrality parameter (NCP). Let us consider how we might derive this alternative distribution.

Let us consider that as before we are testing the hypothesis that $\mathbf{R}\beta = \mathbf{r}$, while in actuality $\mathbf{R}\beta = \mathbf{r} + \delta$. For this case, let us consider the following result.

$$(\sigma_e^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}')^{-\frac{1}{2}} (\mathbf{R}\hat{\beta} - \mathbf{r}) \quad (18)$$

$$= (\sigma_e^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}')^{-\frac{1}{2}} (\mathbf{R}(\hat{\beta} - \beta) + \delta) \quad (19)$$

$$(20)$$

Given that $\hat{\beta}$ is normally distributed, we know that

$$(\sigma_e^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}')^{-\frac{1}{2}} (\mathbf{R}(\hat{\beta} - \beta)) \sim N(0, I) \quad (21)$$

Subsequently, by adding the δ term, we shift the mean, so that

$$(\sigma_e^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}')^{-\frac{1}{2}} (\mathbf{R}\hat{\beta} - \mathbf{r}) \sim N(D^{-1/2}\delta, I) \quad (22)$$

$$(\mathbf{R}\hat{\beta} - \mathbf{r})' (\sigma_e^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}')^{-1} (\mathbf{R}\hat{\beta} - \mathbf{r}) \sim \chi^2(q; \delta' D^{-1} \delta) \quad (23)$$

where $D = \sigma_e^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'$ and the χ^2 parameters are degrees of freedom and non-centrality parameter, respectively. Using the non-centrality parameter we determined we can calculate the power under a particular hypothesis and assumption of alternative value.

(Note: Most of this is directly from [1])

2 Generalized Least Squares (GLS)

For ordinary least squares we made one very important assumption, which was that the error terms or residuals in the phenotype measurements were uncorrelated. In practice, this is not a very reasonable assumption. For example, if we are measuring cholesterol over some genetically similar strains of mice, it is reasonable to assume that part of the "measurement error" between the strains is correlated. This correlation is likely due to the "error" not actually being error, but being some unmodelled effects, such as genetic effects. We will address this issue in a later section. The important thing to note here is that for GLS we will no longer assume non-correlated error terms.

2.1 The Model

In the OLS model, we assumed that $var(\mathbf{y}) = \sigma_e^2 I$. For GLS we will no longer assume this and will instead define $var(\mathbf{y})$ as Σ , which is a positive semi-definite matrix. If we assume that our vector \mathbf{y} is of size n and each value represents the measurement of a particular phenotype in a particular strain, then the diagonal terms of Σ represent the variance of the phenotype within strain and the off diagonal terms represent the covariance of the phenotype between strains. In order to derive the parameter estimates as we did previously, let us consider a possible model.

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} + \mathbf{e} \quad (24)$$

Regardless of the distribution of \mathbf{u} and \mathbf{e} , we can derive the estimate of β while assuming that $var(\mathbf{y}) = var(\mathbf{u} + \mathbf{e}) = \Sigma$ and $E(\mathbf{y}) = \mathbf{X}\beta$. We could derive the estimates of β by assuming the normal distribution of \mathbf{y} and using maximum likelihood or restricted maximum likelihood in order to derive these values. An easier way is to transform our current distribution into the OLS version. In order to do this, let us consider an important property of the normal distribution.

The normal distribution is actually one of a class of spherically symmetric distributions [2]. This set of distributions has the property that under orthogonal transformations the likelihood function remains the same. That is, $\mathcal{L}(\Gamma y) = \mathcal{L}(y)$, where Γ is an $n \times n$ orthogonal matrix and \mathcal{L} represents the likelihood function of y . This property has the more familiar and not-so-obvious consequence that when we apply a linear transformation to a normally distributed random variable, we obtain another normally distributed random variable. This rule or property can be summarized as follows.

$$\mathbf{y} \sim N(\mu, \Sigma) \quad (25)$$

$$\mathbf{A}\mathbf{y} + \mathbf{b} \sim N(\mathbf{A}\mu + \mathbf{b}, \mathbf{A}\Sigma\mathbf{A}') \quad (26)$$

We can use this property to transform our model in such a way that we can obtain a model 24 for which we can use OLS. Consider this transformation.

$$\mathbf{G}\mathbf{y} = \mathbf{G}\mathbf{X}\beta + \mathbf{G}\mathbf{u} + \mathbf{G}\mathbf{e} \quad (27)$$

$$\mathbf{G}\mathbf{y} \sim N(\mathbf{G}\mathbf{X}\beta, \mathbf{G}\Sigma\mathbf{G}') \quad (28)$$

If we can select \mathbf{G} in such a way so that $\mathbf{G}\Sigma\mathbf{G}' = I$, then we can effectively return our model to the assumptions necessary for estimating β using OLS. When Σ has certain properties, like being positive semi-definite, we can show that $\Sigma^{1/2}\Sigma^{-1/2} = I$. This means that if we choose $\mathbf{G} = \Sigma^{-1/2}$, then we will transform our distribution appropriately. Let's consider the estimate for β in this case. Previously, we showed that $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. By substituting \mathbf{X} with $\mathbf{G}\mathbf{X}$ and \mathbf{y} with $\mathbf{G}\mathbf{y}$ we obtain the following.

$$\hat{\beta}_{GLS} = ((\Sigma^{-1/2}\mathbf{X})' \Sigma^{-1/2}\mathbf{X})^{-1} (\Sigma^{-1/2}\mathbf{X})' \Sigma^{-1/2}\mathbf{y} \quad (29)$$

$$= (\mathbf{X}'\Sigma^{-1/2'}\Sigma^{-1/2}\mathbf{X})^{-1} \mathbf{X}'\Sigma^{-1/2'}\Sigma^{-1/2}\mathbf{y} \quad (30)$$

$$= (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1} \mathbf{X}'\Sigma^{-1}\mathbf{y} \quad (31)$$

2.2 Best Linear Unbiased Estimates (BLUEs)

The way in which we derived the estimate for β using GLS was done in such a way to guarantee that $\hat{\beta}_{GLS}$ is a BLUE.

Theorem 2.1 (Aiken) *Given that*

1. \mathbf{X} is non-stochastic

2. \mathbf{y} is a random vector such that $E(\mathbf{y}) = \mathbf{X}\beta$ and $\text{var}(\mathbf{y}) = \Sigma$.

Then $\hat{\beta}_{GLS}$ is a BLUE for β and $\hat{\beta}_{GLS} \sim N(\beta, (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1})$.

2.3 Hypothesis Testing

Given theorem 2.1 and our previous derivation of test statistics in section 1.4 for OLS, we can make a similar statement about the statistic necessary to test the hypothesis that $\mathbf{R}\beta = \mathbf{r}$.

$$(\mathbf{R}\hat{\beta}_{GLS} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta}_{GLS} - \mathbf{r}) \sim \chi^2(q) \quad (32)$$

However, we find that this is problematic, as the true value of Σ is in practice unknown. That is, we have most likely estimated Σ from our data in a domain specific instance. We can get around this by making the assumption that we know Σ up to a scalar multiple. Consider our transformation ($G\mathbf{y} = \Sigma^{-1/2}\mathbf{y}$). We chose \mathbf{G} in such a way that we guaranteed that our variance would be I . If we consider that the true variance structure of \mathbf{y} is $\mathbf{\Sigma}$ and that what we used was actually an estimate such that $\Sigma = \sigma_c^2\hat{\Sigma}$, then our transformed \mathbf{y} will have variance $\sigma_c^2\mathbf{I}$. With this we see that

$$(\mathbf{R}\hat{\beta}_{GLS} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta}_{GLS} - \mathbf{r}) \quad (33)$$

$$= (\mathbf{R}\hat{\beta}_{GLS} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'(\hat{\Sigma}\sigma_c^2)^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta}_{GLS} - \mathbf{r}) \quad (34)$$

$$= \frac{(\mathbf{R}\hat{\beta}_{GLS} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta}_{GLS} - \mathbf{r})}{\sigma_c^2} \sim \chi^2(q) \quad (35)$$

I did not plug in $\hat{\Sigma}$ in the β s above, but if we do, the σ_c s will cancel. We can assume that the β s will be calculated with the estimates and not write it about for brevity. We now observe that we can use our estimate for Σ , but we have a new parameter σ_c that is unknown. As we did before, we can use this statistic to create an \mathcal{F} -statistic, so that this parameter goes away. Without the full derivation we will have an equation similar to the OLS statistic.

$$\phi = \frac{(\mathbf{R}\hat{\beta}_{GLS} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta}_{GLS} - \mathbf{r})}{\hat{\sigma}_c^2 q} \quad (36)$$

$$\phi \sim \mathcal{F}(q, n - q) \quad (37)$$

The estimate of σ_c comes from the basic sample variance estimate, but we use the transformed data, $\hat{\Sigma}^{-1/2}\mathbf{y}$. If we plug this in and do some manipulation we will find that

$$\hat{\sigma}_c^2 = \frac{\mathbf{y}'[\hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Sigma}^{-1}]\mathbf{y}}{n - q} \quad (38)$$

One way around these extra calculations is to normalize the data. I'm not sure what the implications are, but it shouldn't make a difference. In that case, we know that each variance component will have to be less than one and that the total variance should add to one. Variances will just have to be scaled and we should not lose any information... I think.

2.4 Statistical Power

We can derive the statistics and non-centrality parameters to estimate statistical power for GLS the same way as for OLS, shown in section 1.5. We first consider that $\hat{\beta}_{GLS} \sim N(\beta, (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1})$. In the previous section, we showed how we can obtain a χ^2 statistic for the hypothesis test $\mathbf{R}\beta = \mathbf{r}$. We can derive a non-centrality parameter for this statistic by considering the case when $\mathbf{R}\beta = \mathbf{r} + \delta$ ($\mathbf{r} = \mathbf{R}\beta - \delta$).

$$\begin{aligned}
& [\mathbf{R}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1/2}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \\
= & [\mathbf{R}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1/2}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{R}\boldsymbol{\beta} + \boldsymbol{\delta}) \\
= & [\mathbf{R}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1/2}(\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \boldsymbol{\delta})
\end{aligned}$$

Because we know the distribution of $\hat{\boldsymbol{\beta}}$ and therefore $\mathbf{R}\hat{\boldsymbol{\beta}}$, we can say that

$$\begin{aligned}
& [\mathbf{R}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1/2}(\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})) \sim N(0, I) \\
& [\mathbf{R}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1/2}(\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \boldsymbol{\delta}) \sim N([\mathbf{R}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1/2}\boldsymbol{\delta}, I)
\end{aligned}$$

We can then square this statistic in order to obtain a χ^2 statistic. Let $\mathbf{W} = [\mathbf{R}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']$.

$$\begin{aligned}
& (\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \boldsymbol{\delta})'\mathbf{W}^{-1}(\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \boldsymbol{\delta}) \\
= & (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'\mathbf{W}^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \sim \chi^2(q; \boldsymbol{\delta}'\mathbf{W}^{-1}\boldsymbol{\delta})
\end{aligned} \tag{39}$$

The final solution is a χ^2 statistic with q degrees of freedom and a non-centrality parameter of $\boldsymbol{\delta}'\mathbf{W}^{-1}\boldsymbol{\delta}$.

3 Modeling Phenotypes

Note: This part is not very well written and needs a lot of work...

It is of interest to us to use mixed models and GLS in order to model phenotypes and test for associations between variables of interest. It seems to have become rather generally accepted by many people in the field that phenotypes can be modeled using a multivariate normal framework. The observation that phenotypes follow a normal distribution was made by Galton in the late 1800's and an explanation and statistical model for this phenomenon was suggested by Fisher in 1918 [6]. It is questionable as to under what conditions the assumption of normality holds and this was discussed at length by [4]. In our case, we will assume that the phenotypes of interest follow a multivariate normal distribution and we will develop a model for our phenotypes for which we can use the results from GLS in order to test hypotheses and estimate variables of interest.

3.1 Variance Component Model

The assumption of normality and the use of variance components has been heavily employed by the genetics community since Fisher's seminal paper in 1918 (*insert large number of references*). The basic idea of a variance component model, as described by Fisher [6], is that a normally distributed phenotype has a variance V_T that can be represented as the sum of additive variances each due to some effect [8]. For example, we might have a phenotype for which the variance can be partitioned into three components. We might assume that part of the variance is explained by a genetic component (V_G), part is explained by a household component (V_H) and part by random error (V_E). In this case our total variance will just be the sum of each of these components $V_T = V_G + V_H + V_E$. Obviously, we assume that the variances are uncorrelated. We can generalize this to say that $\mathbf{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega})$, where \mathbf{y} is our phenotype of interest and $\boldsymbol{\Omega} = \sum_i V_i$, where each V_i is an individual matrix contributing to the overall multivariate normal variance-covariance matrix.

Let us consider the case when we are interested only in genetic variance and error variance. We might suggest a phenotypic model of the following form.

$$y_i = \mu_i + g_i + e_i \tag{40}$$

In this model, y_i represents an individual's phenotype for trait i , g_i represents the genetic contribution to the phenotype and e_i represents the contribution of error. An individual will have many genetic markers across

the genome that will contribute portions of the phenotype, so we let $a_l^{(i)}$ represent the additive contribution of locus l to phenotype i . We can assume that the genetic component of the phenotype can be summarized as the sum of all individual contributions from each locus. If we assume that there are no dominance effects, we can then write our model as follows.

$$y_i = \mu_i + \sum_l a_l^{(i)} + e_i \quad (41)$$

We can further decompose the a_l 's into individual contributions from each particular variant and we can calculate the covariance of this genetic effect between any two individuals. The solutions for these types of calculations are well-known and can be found in [5] for example. The final result is that $\text{var}(g_i) = 2\sigma_a^2\Phi$, where Φ is the kinship matrix and again we are assuming no dominance effects. The kinship coefficient between two individuals represents the probability of them sharing a gene identical by descent at a randomly chosen locus, so it makes sense that the covariance between individuals is proportional to this. If we assume that the number of genetic loci is very large and that the contribution of each is very small, we can represent this genetic component as a random effect in our model. If we were to see the derivation of $\text{var}(g_i)$, we would know that $E[a_l] = 0$, so that $g_i \sim N(0, 2\sigma_a^2\Phi)$.

The calculation of the kinship matrix (Φ) is laborious and it was shown by [7] that an identity by state (IBS) matrix can be used to represent the polygenic background. In this formulation of the model, we think of each of the genetic loci as being a covariate in our model.

$$y_i = \mu_i + \sum_k \beta^{(i)} X_k + e_i \quad (42)$$

As we stated earlier, if we assume that k is large and that each loci contributes a small amount, we can assume that $\sum_k \beta^{(i)} X_k$ follows a multivariate normal distribution. Kang *et al.* shows that this variance can be estimated by $\sigma_g K$, where σ_g is the additive genetic variance and K is the IBS allele sharing matrix. They also present a method for rapidly estimating this variance component and performing association tests.

With all of this said, we will adopt the model assumed by Kang *et al.*, which is fully summarized in equations below.

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{g} + \mathbf{e} \quad (43)$$

$$\mathbf{g} \sim N(0, \sigma_g^2 K) \quad (44)$$

$$\mathbf{e} \sim N(0, \sigma_e^2 I) \quad (45)$$

In this model, \mathbf{y} is a phenotype vector of size n , \mathbf{X} is an $n \times q$ matrix of covariates, β is a vector of coefficients of size q and \mathbf{u} and \mathbf{e} are random variables accounting for genetics and error, respectively. With this model, we can say that our phenotype y is $N(\mathbf{X}\beta, \Sigma)$, where $\Sigma = \sigma_g^2 K + \sigma_e^2 I$ and we can utilize our previous GLS results. In the next section, I will describe what we would like to do for hypothesis testing.

3.2 Hypothesis Testing

Given, our variance components σ_g^2 and σ_e^2 , we can use the GLS F-test described previously. In association studies, we want to test the significance that one particular SNP has on a phenotype. We do this by fitting the model described in equation 43, where \mathbf{X} is an $n \times q$ matrix of covariates, one of which is a snp of interest. That is, one covariate included in \mathbf{X} is a vector of SNP values for each strain of mouse. These values could be coded as 0, 0.5 and 1 for homozygous major allele, heterozygous and homozygous minor allele. We can fit the model and then design an appropriate \mathbf{R} matrix to test the hypothesis that our SNP has no effect. For example, let us consider the case when the \mathbf{X} matrix has one column of all ones and another column that is representative of the SNP we are testing. In this case, our β vector would be of size 2. The appropriate \mathbf{R} would then be $\begin{bmatrix} 0 & 1 \end{bmatrix}$ so that we would be testing $\mathbf{R}\beta = \mathbf{r} = 0$.

3.3 Statistical Power

In order to estimate statistical power, we need to be able to calculate a non-centrality parameter for our association statistic. For the GLS case, we showed how we can calculate a non-centrality parameter and estimate power by assuming that our \mathbf{r} is actually equal to something else. We can do the same trick here, but in order for it to be meaningful we need to make some more assumptions about how phenotypes are modeled. More specifically, it doesn't mean much to say that when $\mathbf{r} = 5$ we have a certain power. Instead, we need to determine the value of \mathbf{r} by using concepts that are familiar to geneticists and have some meaning within this domain. We consider two parameters of interest that we use in order to determine the value of \mathbf{r} and in turn calculate our non-centrality parameter. These are heritability (hg^2) and SNP effect.

3.3.1 Heritability

Heritability is at the core of genetics. The definition of genetics is in fact the study of heritability. Geneticists study heritable traits. We can think of heritability as the portion of phenotypic variance that is explained by the genetic component. Within our model this means that heritability (hg^2) is defined as follows.

$$hg^2 = \frac{\sigma_g^2}{\sigma_e^2 + \sigma_g^2} \quad (46)$$

Simply put, we estimate the heritability of a trait as the proportion of additive genetic variance to total additive variance. For a given heritability and given σ_e^2 , we can solve for the genetic variance component.

$$\sigma_g^2 = \frac{hg^2\sigma_e^2}{(1 - hg^2)} \quad (47)$$

In our case, we have to consider that we are using a multivariate framework. Our calculation of heritability has to consider the covariance between individuals when computing heritability. We define heritability as follows.

$$hg^2 = \frac{var(\mathbf{g})}{var(\mathbf{g}) + var(\mathbf{e})} \quad (48)$$

The $var(\mathbf{g})$ is not known with out instantiating \mathbf{g} . Also, the sample variance does not follow a known distribution. We can however attempt to approximate it with the expected value and this seems to work relatively well in practice.

$$var(\mathbf{g}) = \frac{1}{n-1} \mathbf{g}' S \mathbf{g} \quad (49)$$

$$E[var(\mathbf{g})] = \frac{1}{n-1} \sigma_g^2 Tr(SK) \quad (50)$$

, where $\mathbf{S} = \mathbf{I}_n - (1/n)\mathbf{J}_n$. This is given by well-known equalities for quadratic forms (see matrix cookbook or basic statistical results in a linear models book). Substituting the expected value of the sample variance of \mathbf{g} and solving for σ_g^2 we get

$$\sigma_g^2 = \frac{hg^2 var(\mathbf{e})(n-1)}{(1 - hg^2) Tr(\mathbf{SK})} \quad (51)$$

3.3.2 SNP Effect

We also want to incorporate SNP effect into the calculation of our non-centrality parameter. The idea of SNP effect is that when a SNP is causal it will have some effect on the phenotype. The magnitude of this effect is important. Specifically, the percentage of the total variance that we can account for by the addition of this SNP is important. This quantity is what we call SNP effect.

We can think of SNP effect in the following way. Let's say that we have a phenotype vector \mathbf{y} such that $\mathbf{y} \sim N(0, \Sigma)$, where Σ is our variance-covariance matrix that is representative of the genetic background. We can create a new phenotype by the addition of a SNP effect.

$$\mathbf{y}' = \mathbf{y} + \mathbf{x}\beta \quad (52)$$

The new phenotype has an effect from some SNP \mathbf{x} , which will change the base level phenotype by β when an individual is homozygous. The SNP effect (s) can then be thought of as the proportion of the total sample variance of \mathbf{y}' that is due to the effect of the SNP. We can represent SNP effect in the following way.

$$s = \frac{\beta^2 \text{var}(\mathbf{x})}{\beta^2 \text{var}(\mathbf{x}) + \text{var}(\mathbf{y})} \quad (53)$$

If we solve for β^2 we get that $\beta^2 = \frac{s \text{var}(\mathbf{y})}{(1-s)\text{var}(\mathbf{x})}$. We can consider that the \mathbf{x} is a given SNP and that s is a given SNP effect, but we are not able to solve directly for the sample variance of \mathbf{y} unless given a specific instance. We get around this by doing the following.

$$\text{var}(\mathbf{y}) = \frac{\sum_i (y_i - \bar{y})^2}{n-1} \quad (54)$$

$$= \frac{1}{n-1} \mathbf{y}' \mathbf{S} \mathbf{y} \quad (55)$$

where $\mathbf{S} = \mathbf{I}_n - \frac{1}{n} \mathbf{J}_n$ (\mathbf{J}_n is an $n \times n$ matrix of 1s) and $Tr()$ is the trace of a matrix. With this we have a solution for β .

$$\beta^2 = \frac{s}{1-s} \frac{\mathbf{y}' \mathbf{S} \mathbf{y}}{\mathbf{x}' \mathbf{S} \mathbf{x}} \quad (56)$$

3.3.3 Estimating Power

Using the previous found in equation 39, we know that our null distribution is $\chi^2(q)$ and that our alternative distribution is $\chi^2(q; \delta' \mathbf{W}^{-1} \delta)$, where $\delta = \sqrt{\beta^2}$ and $\mathbf{W} = \mathbf{R}(\mathbf{X}' \Sigma^{-1} \mathbf{X})' \mathbf{R}'$. The \mathbf{W} in this equation is a function of the variance of $\hat{\beta}$ and the \mathbf{R} used in the hypothesis test. Both of these will be dependent on the experimental setup. More specifically, the \mathbf{X} will depend on the SNP we are testing and the sample variance of \mathbf{y} will depend on Σ , which will depend on the kinship matrix and the heritability. Assuming that we have a given heritability and the corresponding genetic variance component σ_g calculated by equation 51, we can setup our hypothesis test with the following set of equations.

$$\begin{aligned} \Sigma &= \sigma_g^2 \mathbf{K} + \sigma_e \mathbf{I} \\ \mathbf{R} &= [0 \quad 1]' \\ \mathbf{X} &= [\mathbf{1}_n \quad \mathbf{x}_i] \\ \mathbf{W} &= \mathbf{R}(\mathbf{X}' \Sigma^{-1} \mathbf{X})' \mathbf{R}' \\ q &= 2 \end{aligned}$$

Here \mathbf{x}_i is SNP vector i . We will assume that $\mathbf{y} \sim N(0, \Sigma)$ under the null. If we introduce a SNP effect of magnitude δ then $\mathbf{y} \sim N(\delta \mathbf{x}_i, \Sigma)$. We would like to estimate our power to detect this SNP effect. In order to do this we define our null test as $\mathbf{R}\beta = 0$, which follows $\chi^2(q)$. Our alternative alternative distribution then follows $\chi^2(q; \delta' \mathbf{W}^{-1} \delta)$. This follows from our earlier derivation of the alternative distribution (equation 39). To calculate the power we then simply integrate over the alternative distribution from v to ∞ , where v is the value for which $(1 - \alpha)\%$ of $\chi^2(q)$ values fall below.

The problem with this is the calculation of δ . We know that $\delta = \sqrt{\beta^2}$, where β is given by equation 56. Our calculation for β^2 necessitates knowledge of the distribution of the sample variance of \mathbf{y} , which we do not have. In order to get around this, we can approximate the distribution through sampling. For each sampled point, we can calculate β^2 and the corresponding non-centrality parameter and then average the power over each of these.

References

- [1] Chung-Ming Kuan. Introduction to Econometric Theory. Academia Sinica, Taipei. 2000.
- [2] Takeaki Kariya and Hiroshi Kurata. Generalized Least Squares. John Wiley and Sons, UK. 2004.
- [3] Charles E. McCulloch and Shayle R. Searle. Generalized, Linear and Mixed Models. John Wiley and Sons, NY. 2001.
- [4] Kenneth Lange Central Limit Theorems for Pedigrees. J. Math. Biology 6, 59-66 1978.
- [5] Kenneth Lange. Mathematical and Statistical Methods for Genetic Analysis. Springer 2002.
- [6] R.A. Fisher. The Correlation Between Relatives on the Supposition of Mendelian Inheritance Trans R Soc Edinburgh 52:399-433 1918.
- [7] H M Kang , N A Zaitlen , A Kirby , C M Wade , D Heckerman , M J Daly and E Eskin. Efficient Control for Population Structure in Model Organism Association Mapping. Genetics 178.3:1709-23 2008.
- [8] Mark Abney, Mary Sara McPeck and Carole Ober. Estimation of Variance Components of Quantitative Traits in Inbred Populations. Am. J. Hum. Genet. 66:629-650 2000.