Mathematics of the Linear Model and Linear Mixed Model

Brian Zhang

February 2020

Contents

1 Fundamentals **3**
1.1 Linear Algebra . . . 3
1.2 Probability . . . 3
1.2.1 Multivariate Normal . . . 4
1.2.2 Derived Distributions: Definitions . . . 4
1.2.3 Derived Distributions: Properties . . . 5
1.3 Inference for a Normal Sample . . . 5
1.3.1 Sampling from the Univariate Normal . . . 5
1.3.2 Sampling from the Multivariate Normal . . . 6

2 Standard Linear Model: Fixed Effects **6**
2.1 Tests of Significance . . . 7
2.1.1 Example: Redundant Features . . . 8
2.2 Confidence Intervals . . . 8
2.3 Approximate Testing with Wald / LRT . . . 8

3 Linear Mixed Model Part 1: Model Setup **8**

4 REML Interlude **9**
4.1 Sample Variance . . . 9
4.1.1 Going more general . . . 11
4.2 Linear Model . . . 11
4.3 Mixed Model . . . 12
4.3.1 A Lemma (Searle Casella McCulloch M.4f) . . . 13
4.3.2 REML likelihood . . . 13
4.3.3 Estimation . . . 14
4.4 An Application to Covariates . . . 14

5 Linear Mixed Model Part 2: Association and Prediction **14**
5.1 Association . . . 15
5.2 Prediction . . . 15
5.2.1 SNP or Basis Perspective . . . 16
5.2.2 Sample or Kernel Perspective . . . 16
5.2.3 Connection with Association . . . 17
5.2.4 Alternate Derivation . . . 18
5.2.5 Small N, large P limit . . . 19

6 Notes on Modern Genetics **19**
6.1 LDpred 19
6.2 Expected Heritability and MAF 19
6.3 Non-infinitesimal Priors 19
6.4 LD Score Regression and Cousins 19

7 References **19**

Thanks especially to Jonathan Marchini and Pier Palamara for supervising my DPhil and introducing me to this research background. So that I can refer to them by shorthand, textbooks that have helped me in my learning are:

- "Casella / Berger": *Statistical Inference* by George Casella and Roger L. Berger [1]
- "Wasserman": *All of Statistics* by Larry Wasserman [2]
- "ESL": *The Elements of Statistical Learning*, Second Edition by Trevor Hastie, Robert Tibshirani, and Jerome Friedman [3]
- "Bishop": *Pattern Recognition and Machine Learning* by Christopher M. Bishop [4]
- "Murphy": *Machine Learning* by Kevin P. Murphy [5]

Note 1: All likelihoods / log-likelihoods are correct up to a constant. Note 2: All references are clickable, and most PDF viewers have shortcuts to navigate back from links (`Command + [` on Preview).

1 Fundamentals

1.1 Linear Algebra

Trace trick, idempotent / projection operators, pseudo-inverse, Schur complement. Derivatives of matrix expressions. Positive (semi-)definite matrices, Cholesky decomposition, matrix square root.

For vectors u and v, we define

$$u \cdot v = u^T v,$$
$$|u| = \sqrt{u \cdot u} \geq 0.$$

Then

$$u \cdot v = |u||v| \cos(\theta), \tag{1}$$

where θ is the angle between vectors u and v.

For any real matrix X,

$$\text{rank}(X^T X) = \text{rank}(X X^T) = \text{rank}(X) = \text{rank}(X^T). \tag{2}$$

Matrix inversion lemma / Sherman-Morrison-Woodbury formula. For A n by n, C k by k, U n by k, and V k by n, all of maximum rank,

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}. \tag{3}$$

Note that the first complicated inverse is of an n by n matrix, whereas the second complicated inverse is of a k by k matrix.

TODO: add Bishop 2.76, the matrix inversion formula, which also introduces the Schur complement (Bishop 2.77). See `https://en.wikipedia.org/wiki/Block_matrix#Block_matrix_inversion` for how the matrix inversion formula / lemma relate.

1.2 Probability

$$E(X) = E(E(X|Y)) \tag{4}$$
$$\text{cov}(Y) = E(\text{cov}(Y|X)) + \text{cov}(E(Y|X)) \tag{5}$$

1.2.1 Multivariate Normal

Density:

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{6}$$

Affine transformation. If $x \sim \mathcal{N}(\mu,\Sigma)$, then

$$Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T). \tag{7}$$

Marginal and conditional. Let

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad \Lambda = \Sigma^{-1} = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix},$$

then (Bishop 2.81-2.82, 2.91, 2.96-2.98)

$$\begin{aligned} p(x_1) &= \mathcal{N}(x_1|\mu_1,\Sigma_{11}) && (8)\\ &= \mathcal{N}(x_1|\mu_1,(\Lambda_{11}-\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21})^{-1}), && (9)\\ p(x_1|x_2) &= \mathcal{N}(x_1|\mu_1+\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2),\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) && (10)\\ &= \mathcal{N}(x_1|\mu_1-\Lambda_{11}^{-1}\Lambda_{12}(x_2-\mu_2),\Lambda_{11}^{-1}). && (11) \end{aligned}$$

Bayesian updates. If

$$\begin{aligned} p(x) &= \mathcal{N}(x|\mu,\Sigma), \\ p(y|x) &= \mathcal{N}(y|Ax+b,\Pi), \end{aligned}$$

then (Bishop 2.113-2.117)

$$\begin{aligned} p(y) &= \mathcal{N}(y|A\mu+b,\Pi+A\Sigma A^T), && (12)\\ p(x|y) &= \mathcal{N}(x|\Xi\{A^T\Pi^{-1}(y-b)+\Sigma^{-1}\mu\},\Xi), && (13) \end{aligned}$$

where

$$\Xi = (\Sigma^{-1}+A^T\Pi^{-1}A)^{-1}. \tag{14}$$

Uncorrelated $\Leftrightarrow$ independent. If x_1 and x_2 are univariate normal random variables, then $\text{cov}(x_1,x_2)=0 \Leftrightarrow x_1, x_2$ independent.

1.2.2 Derived Distributions: Definitions

If $X_1,\ldots,X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, then:

$$\sum_{i=1}^{n} X_i^2 \sim \chi_n^2. \tag{15}$$

If $U \sim \mathcal{N}(0,1)$ and $V \sim \chi_p^2$ are independent, then

$$\frac{U}{\sqrt{V/p}} \sim t_p. \tag{16}$$

If $U \sim \chi_p^2$ and $V \sim \chi_q^2$ are independent, then

$$\frac{U/p}{V/q} \sim F_{p,q}. \tag{17}$$

1.2.3 Derived Distributions: Properties

1.2.3.1 χ^2 distribution

The χ^2_n distribution has mean n and variance $2n$. Also, $\chi^2_n = \text{Gamma}(n/2, 2)$ and $\chi^2_2 = \text{Gamma}(1, 2) = \text{Expo}(1/2)$.

Note that the definition of χ^2_n in (15) can be equivalently stated as follows. If $X \sim \mathcal{N}_n(0, I_n)$, then $X^T X \sim \chi^2_n$. This is because for normal random variables, uncorrelated $\Leftrightarrow$ independent. In fact, we have the following more general result. If $X \sim \mathcal{N}_n(\mu, \Sigma)$, and Σ is positive definite, then let $LL^T = \Sigma$ be the Cholesky decomposition of Σ. We can apply a transformation to X to get identity covariance, using

$$Y = L^{-1}(X - \mu) \sim \mathcal{N}_n(0, L^{-1}\Sigma L^{-T}) = \mathcal{N}_n(0, I_n),$$

where we have used (7). Thus

$$(X - \mu)^T \Sigma^{-1}(X - \mu) = (X - \mu)^T L^{-T} L^{-1}(X - \mu) = Y^T Y \sim \chi^2_n. \tag{18}$$

There is also the special case where $X \sim \mathcal{N}_n(\mu, \Sigma)$ but Σ is positive semi-definite and not of full rank. Then $X - \mu$ will always lie in a strict subspace of $\mathbb{R}^n$; this is called the degenerate case of the multivariate normal. If the rank of Σ is $k < n$, then

$$(X - \mu)^T \Sigma^{+}(X - \mu) \sim \chi^2_k, \tag{19}$$

where Σ^+ is the pseudo-inverse. The proof is omitted for now.

1.2.3.2 t distribution

The t_p distribution has mean 0 for $p > 1$ and variance $p/(p-2)$ for $p > 2$; otherwise these moments are undefined. $t_1 = \text{Cauchy}(0, 1)$, and as $p \to \infty$, $t_p \to \mathcal{N}(0, 1)$.

1.2.3.3 F distribution

The $F_{p,q}$ distribution has mean $q/(q-2)$ for $q > 2$ and variance $2q^2(p + q - 2)/(p(q-2)^2(q-4))$ for $q > 4$; otherwise these moments are undefined. Casella / Berger 5.3.8 gives:

1. If $X \sim F_{p,q}$, then $1/X \sim F_{q,p}$.
2. If $X \sim t_q$, then $X^2 \sim F_{1,q}$.
3. If $X \sim F_{p,q}$, then $\frac{(p/q)X}{1+(p/q)X} \sim \text{Beta}(p/2, q/2)$.

The first two properties are easy to prove based on the definitions of t_q and $F_{p,q}$.

1.3 Inference for a Normal Sample

1.3.1 Sampling from the Univariate Normal

Casella / Berger 5.2.2, 5.2.3, 5.2.6. For a random (i.e. i.i.d.) sample $X_1, \ldots, X_n$ from a population with mean μ and variance $\sigma^2 < \infty$, define

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad S = \sqrt{S^2}, \tag{20}$$

to be the sample mean, sample variance, and sample standard deviation. Then $E(\bar{X}) = \mu$, $\text{Var}(\bar{X}) = \sigma^2/n$, and $E(S^2) = \sigma^2$.

Casella / Berger 5.3.1. (See Section 4.1 for a proof.) If furthermore the sample is i.i.d. from $\mathcal{N}(\mu, \sigma^2)$, then

1. $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$.

2. $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

3. $\bar{X}$ and S^2 are independent.

Casella / Berger 5.3.4. If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1), \tag{21}$$

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} \sim t_{n-1}, \tag{22}$$

where we have used Casella / Berger 5.3.1 and (16).

Casella / Berger 5.3.6. If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2)$ independently, then

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{n-1,m-1}, \tag{23}$$

where we have used Casella / Berger 5.3.1 and (17).

1.3.2 Sampling from the Multivariate Normal

Now let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu, \Sigma)$ be vectors in $\mathbb{R}^p$, where $n > p$. Define

$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i, \qquad S = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T, \tag{24}$$

to be the sample mean and sample covariance. Then:

$$\bar{X} \sim \mathcal{N}_p(\mu, \Sigma/n), \tag{25}$$

$$E(S) = \Sigma, \tag{26}$$

$$(\bar{X} - \mu)^T(\Sigma/n)^{-1}(\bar{X} - \mu) \sim \chi^2_p, \tag{27}$$

$$(\bar{X} - \mu)^T(S/n)^{-1}(\bar{X} - \mu) \sim T^2_{p,n-1} = \frac{p(n-1)}{n-p} F_{p,n-p}. \tag{28}$$

The distribution $T^2_{p,n-1}$ is known as Hotelling's T-squared. More can be found on Wikipedia. There is also a Wilks' lambda distribution that I think does two-sample covariance comparisons. Notice that (27) follows from (25) using (18).

2 Standard Linear Model: Fixed Effects

Throughout, we consider N samples and M features in our model, possibly including the constant 1 as a feature. We seek to model a single output y_n for each sample, based on a feature vector x_n of size M. We can aggregate this data into a column vector Y and an N by M matrix X. For now we assume no standardization, though in GWAS it is common to standardize Y and the columns of X to have mean 0 and variance 1.

The linear regression model assumes effects β contained in a column vector of size M, with

$$Y = X\beta + \epsilon \tag{29}$$

We introduce assumptions on the errors ϵ: $\epsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, or in other words, $\epsilon \sim \mathcal{N}(0, \sigma^2 I_N)$. Then the log-likelihood given X and Y becomes:

$$l(\beta, \sigma^2 | X, Y) = -\frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta) - \frac{N}{2}\ln\sigma^2 \tag{30}$$

To perform point estimation on β, we use maximum likelihood to get

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{31}$$

so long as $X^T X$ has full rank. Define $\hat{\epsilon} = Y - X\hat{\beta}$. Then maximum likelihood of σ^2 yields

$$\hat{\sigma}^2_{MLE} = \frac{\hat{\epsilon}^T \hat{\epsilon}}{N} = \frac{RSS}{N}.$$

Using restricted maximum likelihood instead yields (details in Section 4.2)

$$\hat{\sigma}^2 = \frac{RSS}{N - M}. \tag{32}$$

This is an unbiased estimator, and will be used instead from now on.

2.1 Tests of Significance

Since $\epsilon \sim \mathcal{N}(0, \sigma^2 I_N)$, and

$$\begin{aligned}\hat{\beta} &= (X^T X)^{-1} X^T Y \\ &= (X^T X)^{-1} X^T (X\beta + \epsilon) \\ &= \beta + (X^T X)^{-1} X^T \epsilon,\end{aligned}$$

we have by (7) that

$$\hat{\beta} \sim \mathcal{N}(\beta, (X^T X)^{-1} X^T \sigma^2 I_N ((X^T X)^{-1} X^T)^T) \tag{33}$$

$$= \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1}). \tag{34}$$

Let A^{-1}_{ij} denote the ij-th entry of the inverse of matrix A, rather than the inverse of the ij-th entry. Then marginally,

$$\hat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2 (X^T X)^{-1}_{ii}), \tag{35}$$

so $\hat{\beta}_i$ is an unbiased estimator with standard error

$$se(\hat{\beta}_i) = \sqrt{\sigma^2 (X^T X)^{-1}_{ii}}. \tag{36}$$

Now we need some magic telling us the distribution of $\hat{\sigma}^2$ and saying it is independent of $\hat{\beta}$. Enter a generalized version of Casella / Berger 11.3.3! This says:

$$\frac{(N - M)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{N-M} \tag{37}$$

and $\hat{\sigma}^2$ is independent of $\hat{\beta}$. (A proof is in Section 4.2.) From this, we define

$$\hat{se}(\hat{\beta}_i) = \sqrt{\hat{\sigma}^2 (X^T X)^{-1}_{ii}} \tag{38}$$

and derive a t-statistic as:

$$\frac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)} \sim \mathcal{N}(0, 1), \tag{39}$$

$$\frac{\hat{\beta}_i - \beta_i}{\hat{se}(\hat{\beta}_i)} = \frac{(\hat{\beta}_i - \beta_i)/se(\hat{\beta}_i)}{\sqrt{\hat{\sigma}^2/\sigma^2}} \sim t_{N-M}, \tag{40}$$

where we have used Casella / Berger 11.3.3 and (16). From here we obtain the t test for testing whether the coefficient β_i is significant. When $N - M > 30$, we can consider making a normal approximation.

2.1.1 Example: Redundant Features

If we have two features in our data matrix that are redundant, we should expect some problems of identifiability. The jth feature is given by the column vector $X_{\cdot,j}$. If $X_{\cdot,j} = X_{\cdot,k}$ with $j \neq k$, then $X^T X$ will not be full rank and it is not possible to form $\hat{\beta}$ according to (31). Similarly, if $M > N$, we will suffer from an undeterdetermined system and $X^T X$ will also not be full rank. ((2) is useful.)

But for now, let's assume $M \leq N$, and that we have two features $X_{\cdot,j}$ and $X_{\cdot,k}$ that are not identical, but very close. Now let's see what happens to the matrix $(X^T X)^{-1}$. To really simplify things, let's write:

$$X^T X = \begin{bmatrix} 1 & 1-\epsilon & 0 \\ 1-\epsilon & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{41}$$

Here the redundant features are 1 and 2. All three components have norm 1, but features 1 and 2 are both orthogonal to feature 3, while features 1 and 2 have a very small angle of offset that creates the $1-\epsilon$ term. (Remember that the dot product can be written using a cosine, as in (1).) Explicitly, we might have:

$$X = \begin{bmatrix} \frac{180}{181} & \frac{180}{181} & 0 \\ 0 & 0 & 1 \\ \frac{19}{181} & -\frac{19}{181} & 0 \end{bmatrix}, \qquad X^T X = \begin{bmatrix} 1 & 0.978 & 0 \\ 0.978 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{42}$$

Using CoCalc, I get a symbolic inverse of

$$(X^T X)^{-1} = \begin{bmatrix} \frac{1}{1-(1-\epsilon)^2} & \frac{\epsilon-1}{1-(1-\epsilon)^2} & 0 \\ \frac{\epsilon-1}{1-(1-\epsilon)^2} & \frac{1}{1-(1-\epsilon)^2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \approx \begin{bmatrix} (2\epsilon)^{-1} & -(2\epsilon)^{-1} & 0 \\ -(2\epsilon)^{-1} & (2\epsilon)^{-1} & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Numerically evaluating the inverse gives

$$(X^T X)^{-1} = \begin{bmatrix} 22.9 & -22.4 & 0 \\ -22.4 & 22.9 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

We can deduce first of all that components 1 and 2 of $\hat{\beta}$ will have a high standard error, based on the term $(X^T X)^{-1}_{ii}$ of (38) being large. Second, the terms $\hat{\beta}_1$ and $\hat{\beta}_2$ have a correlation of almost -1 (using (34)), which makes sense since they are redundant features.

Of course, redundancy doesn't need to just come from a set of two features, it can also be from a wider set. For instance, you constructed an extra feature for your linear regression which is the average of some other features. Uh-oh!

2.2 Confidence Intervals

2.3 Approximate Testing with Wald / LRT

3 Linear Mixed Model Part 1: Model Setup

From the fixed effects model (29), we can add in random effects which correspond to the other SNPs used to build the kinship matrix or genomic relatedness matrix (GRM). Let the SNPs be given by an N by P matrix Z. The random effects vector b can then be a column vector of size P, and we have

$$Y = X\beta + Zb + \epsilon \tag{43}$$

We let $\epsilon \sim \mathcal{N}(0, \sigma_e^2 I_N)$, but before we place a prior on b, the random effects, we choose to standardize Z so that each column has:

$$E(Z_{\cdot,i}) = 0, \qquad Z_{\cdot,i}^T Z_{\cdot,i} = P.$$

Alternatively we could keep Z as-is and rewrite our prior on b, but that is more complicated. After this standardization, the common prior on b is $b \sim \mathcal{N}(0, (\sigma_g^2/P) I_P)$ with b and ϵ independent.

Two notes: first, we can in general have an arbitrary covariance for b, but in genetics we typically assume a diagonal covariance. Second, this standardization / identity covariance implicitly makes a modelling assumption that per-SNP heritability does not depend on MAF. There is theory to relax this assumption, which is explored in Section 6.2.

If we "integrate out" b, we have

$$Zb|\sigma_g^2 \sim \mathcal{N}(0, (\sigma_g^2/P)ZZ^T), \tag{44}$$

$$Zb + \epsilon|\sigma_g^2, \sigma_e^2 \sim \mathcal{N}(0, (\sigma_g^2/P)ZZ^T + \sigma_e^2 I_N)), \tag{45}$$

$$Y|\beta, \sigma_g^2, \sigma_e^2 \sim \mathcal{N}(X\beta, (\sigma_g^2/P)ZZ^T + \sigma_e^2 I_N)), \tag{46}$$

where we have used (7). Define

$$V(\sigma_g^2, \sigma_e^2) = (\sigma_g^2/P)ZZ^T + \sigma_e^2 I_N. \tag{47}$$

Then the log-likelihood of the parameters is

$$l(\beta, \sigma_g^2, \sigma_e^2|X, Y, Z) = -\frac{1}{2}(Y - X\beta)^T V^{-1}(\sigma_g^2, \sigma_e^2)(Y - X\beta) - \frac{1}{2}\ln|V(\sigma_g^2, \sigma_e^2)|. \tag{48}$$

A Bayesian approach would place priors on σ_g^2, σ_e^2, and possibly β. Without priors, we need to rely on point estimates. Maximum likelihood for β gives:

$$\hat{\beta} = (X^T V^{-1}(\hat{\sigma}_g^2, \hat{\sigma}_e^2)X)^{-1} X^T V^{-1}(\hat{\sigma}_g^2, \hat{\sigma}_e^2)Y, \tag{49}$$

where $\hat{\sigma}_g^2, \hat{\sigma}_e^2$ are desired point estimates. Compare with the top of p. 1712 of [6]. $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$ can also be obtained by numerical optimization of the log-likelihood, but the results are biased; thus REML is used instead.

4 REML Interlude

TODO: move this material to an appendix?

To proceed further, it is important to actually understand REML. We present results using REML on three examples. Each time, we revisit a linear algebra paradigm and increase its generality. This presentation is not for the faint-hearted, but once understood, it covers REML in the most general fashion.

4.1 Sample Variance

We wish to derive the distribution of the sample variance and show it is independent of the sample mean. Our model is

$$X \sim \mathcal{N}(\mu 1_N, \sigma^2 I_N),$$

where X and 1_N are column vectors of size N, and μ and σ^2 are unknown parameters. As an example, let's take $N = 5!$[1] Let

$$\hat{\mu} = \frac{x_1 + \ldots + x_5}{5},$$

$$\hat{\sigma}^2 = \frac{(x_1 - \hat{\mu})^2 + \ldots + (x_5 - \hat{\mu})^2}{4}.$$

It's quite easy to show that $\hat{\mu} \sim \mathcal{N}(\mu, \sigma^2/N)$. But massaging the terms of $\hat{\sigma}^2$ to get what we want is a notational challenge, and it turns out to be clearer to introduce matrices.[2] Specifically, we introduce the

[1] If this worries you, note that estimating the sample mean is the same as a linear model with only an intercept, and use the more general results derived in the next section.

[2] Casella and Berger resort to a proof by induction.

Helmert matrix, which is an orthonormal matrix with first column pointing in the direction $(1, \ldots, 1)^T$. For $N = 5$, the unnormalized Helmert matrix is

$$G_5 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & 1 \\ 1 & 0 & -2 & 1 & 1 \\ 1 & 0 & 0 & -3 & 1 \\ 1 & 0 & 0 & 0 & -4 \end{bmatrix}.$$

It's easy to see that any two columns are orthogonal, and it should also be easy to see how to continue the pattern for larger N. We can then normalize the columns to each have norm 1, obtaining:

$$H_5 = \begin{bmatrix} 1/\sqrt{5} & 1/\sqrt{2} & 1/\sqrt{6} & 1/\sqrt{12} & 1/\sqrt{20} \\ 1/\sqrt{5} & -1/\sqrt{2} & 1/\sqrt{6} & 1/\sqrt{12} & 1/\sqrt{20} \\ 1/\sqrt{5} & 0 & -2/\sqrt{6} & 1/\sqrt{12} & 1/\sqrt{20} \\ 1/\sqrt{5} & 0 & 0 & -3/\sqrt{12} & 1/\sqrt{20} \\ 1/\sqrt{5} & 0 & 0 & 0 & -4/\sqrt{20} \end{bmatrix}.$$

H_5 is an orthonormal matrix, because the columns were originally orthogonal and have now been normalized. This gives us the property that $H_5 H_5^T = I_5$, where I_5 is the identity matrix. So in particular,

$$X^T X = X^T (H_5 H_5^T) X = (H_5^T X)^T (H_5^T X). \tag{50}$$

Just to make things more explicit, here is $H_5^T X$:

$$H_5^T X = \begin{bmatrix} 1/\sqrt{5} & 1/\sqrt{5} & 1/\sqrt{5} & 1/\sqrt{5} & 1/\sqrt{5} \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 & 0 & 0 \\ 1/\sqrt{6} & 1/\sqrt{6} & -2/\sqrt{6} & 0 & 0 \\ 1/\sqrt{12} & 1/\sqrt{12} & 1/\sqrt{12} & -3/\sqrt{12} & 0 \\ 1/\sqrt{20} & 1/\sqrt{20} & 1/\sqrt{20} & 1/\sqrt{20} & -4/\sqrt{20} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} (x_1 + x_2 + x_3 + x_4 + x_5)/\sqrt{5} \\ (x_1 - x_2)/\sqrt{2} \\ (x_1 + x_2 - 2x_3)/\sqrt{6} \\ (x_1 + x_2 + x_3 - 3x_4)/\sqrt{12} \\ (x_1 + x_2 + x_3 + x_4 - 4x_5)/\sqrt{20} \end{bmatrix}.$$

So (50), written out explicitly, becomes

$$\sum_{i=1}^{5} x_i^2 = 5\hat{\mu}^2 + \frac{1}{2}(x_1 - x_2)^2 + \frac{1}{6}(x_1 + x_2 - 2x_3)^2 + \frac{1}{12}(x_1 + x_2 + x_3 - 3x_4)^2 + \frac{1}{20}(x_1 + x_2 + x_3 + x_4 - 4x_5)^2.$$

Notice that the first entry of $H_5 X$ is $\sqrt{5}\hat{\mu}$; in general it will be $\sqrt{N}\hat{\mu}$. The sum of squares of the remaining terms is

$$\begin{aligned} \left(\sum_{i=1}^{N} x_i^2\right) - N\hat{\mu}^2 &= \left(\sum_{i=1}^{N} (x_i - \hat{\mu})^2 + 2x_i\hat{\mu} - \hat{\mu}^2\right) - N\hat{\mu}^2 \\ &= \left(\sum_{i=1}^{N} (x_i - \hat{\mu})^2\right) + 2\left(\sum_{i=1}^{N} x_i\right)\hat{\mu} - 2N\hat{\mu}^2 \\ &= \sum_{i=1}^{N} (x_i - \hat{\mu})^2 \\ &= (N-1)\hat{\sigma}^2. \end{aligned}$$

So in our case of $N = 5$,

$$4\hat{\sigma}^2 = \frac{1}{2}(x_1 - x_2)^2 + \frac{1}{6}(x_1 + x_2 - 2x_3)^2 + \frac{1}{12}(x_1 + x_2 + x_3 - 3x_4)^2 + \frac{1}{20}(x_1 + x_2 + x_3 + x_4 - 4x_5)^2.$$

Given that the x_i are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, it can be seen that each of these four terms is the square of a $\mathcal{N}(0, \sigma^2)$ random variable. Now this is where things get interesting. Our initial vector X was an isotropic (spherically

symmetric) Gaussian centered at $(\mu, \ldots, \mu)$, and multiplying by the orthonormal matrix H_N^T corresponds to a rotation around the origin. Thus, the new vector $H_N^T X$ will also be an isotropic Gaussian, with the same covariance $\sigma^2 I_N$ and centered this time at $(\sqrt{N}\hat{\mu}, 0, \ldots, 0)$. This means that each of the entries of $H_N^T X$ is a $\mathcal{N}(0, \sigma^2)$ variable and independent from all the others (uncorrelated $\Leftrightarrow$ independent for normals). Therefore, we can combine the squares of the independent $\mathcal{N}(0, \sigma^2)$ random variables to get that

$$\frac{(N-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{N-1},$$

and additionally, we see that $\hat{\sigma}^2$ is independent from the first entry of $H_N^T X$, and hence is independent from $\hat{\mu}$.

4.1.1 Going more general

At the moment, our derivation relies on Helmert matrices, which may seem a bit unmotivated. In fact, the same argument holds for a general class of $N \times N$ matrices that include H_N. Let 1_N be the column vector of all ones, so that the first column of H_N is $1_N/\sqrt{N} = v_1$. Let $v_2, \ldots, v_n$ be additional vectors such that $v_1, \ldots, v_N$ form an orthonormal basis of $\mathbb{R}^N$. Then the matrix V with ith column v_i (this generalizes H_N) is orthonormal, so $V^T V = VV^T = I_N$. We also have the identity,

$$I_N = VV^T = \begin{bmatrix} v_1 & V_{2:N} \end{bmatrix} \begin{bmatrix} v_1^T \\ V_{2:N}^T \end{bmatrix} = v_1 v_1^T + v_1 V_{2:N}^T + V_{2:N} v_1^T + V_{2:N} V_{2:N}^T = v_1 v_1^T + V_{2:N} V_{2:N}^T. \tag{51}$$

Using this, our earlier decomposition generalizes to

$$\begin{aligned} \sum_{i=1}^N x_i^2 = X^T X &= X^T (v_1 v_1^T + V_{2:N} V_{2:N}^T) X = (v_1^T X)^T (v_1^T X) + (V_{2:N}^T X)^T (V_{2:N}^T X) \\ &= \hat{\mu}^2/N + (V_{2:N}^T X)^T (V_{2:N}^T X). \end{aligned}$$

Since $X \sim \mathcal{N}(\mu 1_N, \sigma^2 I_N)$, by (7),

$$V_{2:N}^T X \sim \mathcal{N}(\mu V_{2:N}^T 1_N, V_{2:N}^T (\sigma^2 I_N) V_{2:N}) = \mathcal{N}(0_{N-1}, \sigma^2 I_{N-1}),$$

by the orthogonality of $v_1 = 1_N/\sqrt{N}$ and the other v_i. So $\hat{\mu}$ is a rescaled version of $||v_1^T X||_2^2$, $\hat{\sigma}^2$ is a rescaled version of $||V_{2:N}^T X||_2^2$ whose distribution can be worked out as earlier, and the two estimators are independent.

4.2 Linear Model

Consider the linear model from earlier:

$$\begin{aligned} Y &= X\beta + \epsilon, \\ \epsilon &\sim \mathcal{N}(0, \sigma^2 I_N), \end{aligned}$$

where X $(N \times M)$ is fixed and known, Y $(N \times 1)$ is known, and we have unknown parameters β $(M \times 1)$ and σ^2. Assume $M < N$ and X has rank M.

Let $v_1, v_2, \ldots, v_{N-M}$ be a set of orthonormal vectors that live in the subspace $\text{col}(X)^\perp = \text{null}(X^T) \subset \mathbb{R}^N$; these can be constructed using Gram-Schmidt orthogonalization. Let V be the $N \times (N-M)$ matrix with ith column v_i. Consider the identity:

$$I_N = X(X^T X)^{-1} X^T + VV^T. \tag{52}$$

The first term on the right is a projection operator onto the subspace $\text{col}(X)$, and the second term is a projection operator onto $\text{col}(X)^\perp = \text{col}(V)$ (note that it has the same form as the first term, but $V^T V =$

I_{N-M} so we can simplify). This proves (52), which is a generalization of (51). Using it, we can multiply both sides of $Y = X\beta + \epsilon$ by VV^T to get:

$$\begin{aligned} VV^TY &= VV^TX\beta + VV^T\epsilon, \\ (I_N - X(X^TX)^{-1}X^T)Y &= 0 + VV^T\epsilon, \\ Y - X\hat{\beta} &= VV^T\epsilon. \end{aligned}$$

If we extend our vectors $v_1, \ldots, v_{N-M-1}$ to a full orthonormal basis $v_1, \ldots, v_N$ of $\mathbb{R}^N$, we can write

$$\epsilon = \delta_1 v_1 + \ldots + \delta_N v_N,$$

where each $\delta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. Then

$$VV^T\epsilon = \delta_1 v_1 + \ldots + \delta_{N-M} v_{N-M}.$$

Hence

$$||Y - X\hat{\beta}||_2^2 = ||VV^T\epsilon||_2^2 = \delta_1^2 + \ldots + \delta_{N-M}^2.$$

Denoting the left hand side as RSS, the residual sum of squares, and taking expectations, we have

$$E(RSS) = (N-M)\sigma^2.$$

Hence our earlier estimator of σ^2 from (32) is unbiased:

$$E\left(\hat{\sigma}^2\right) = E\left(\frac{RSS}{N-M}\right) = \frac{(N-M)\sigma^2}{N-M} = \sigma^2.$$

More generally,

$$\hat{\sigma}^2 = \frac{RSS}{N-M} = \frac{\sigma^2(\delta_1^2 + \ldots + \delta_{N-M}^2)}{N-M},$$

so

$$(N-M)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{N-M}.$$

This is (37) and forms the first part of Casella / Berger 11.3.3, which we needed to derive the linear model t-test. The second part is to show that $\hat{\sigma}^2$ is independent of $\hat{\beta}$. Note that $\hat{\sigma}^2$ is a deterministic function of the stochastic variables $\delta_1, \ldots, \delta_{N-M}$, so it is independent of any deterministic function of the limited set of variables $\delta_{N-M+1}, \ldots, \delta_N$, since the δ_i are i.i.d. From writing

$$\begin{aligned} \hat{\beta} &= (X^TX)^{-1}X^T(X\beta + \epsilon) \\ &= \beta + (X^TX)^{-1}X^T\epsilon \\ &= \beta + (X^TX)^{-1}(0 + X^T\delta_{N-M+1}v_{N-M+1} + \ldots + X^T\delta_N v_N), \end{aligned}$$

we have this desired property.

4.3 Mixed Model

From Sections 4.1 and 4.2, we have encountered the key ideas of REML. In each case, we were trying to estimate an unknown variance, and by multiplying our data by an orthonormal matrix, we could decompose our variance estimator into a function of independent normal random variables. Perhaps because the modifications of REML in these cases – changing a division by N to $N-1$ for sample variance and $N-M$ for the linear model – are simple and "intuitive"[3], the details of the REML derivation are frequently skipped in textbooks.

In contrast, the results of REML for the mixed model do not follow from a quick modification from the maximum likelihood case. In this context, where the results are less intuitive, having the entire machinery we have developed proves essential.

[3]The intuition usually given is that of limiting degrees of freedom.

4.3.1 A Lemma (Searle Casella McCulloch M.4f)

4.3.2 REML likelihood

Similar to the case of the linear model, we assume that X is full rank (has rank M) and multiply both sides of the linear mixed model equation by the projection operator $I_N - X(X^TX)^{-1}X^T$:

$$Y = X\beta + Zb + \epsilon,$$
$$\left(I_N - X(X^TX)^{-1}X^T\right)Y = 0 + \left(I_N - X(X^TX)^{-1}X^T\right)(Zb + \epsilon).$$

Recall that

$$Zb + \epsilon|\sigma_g^2, \sigma_e^2 \sim \mathcal{N}(0, (\sigma_g^2/P)ZZ^T + \sigma_e^2 I_N)) = \mathcal{N}(0, V(\sigma_g^2, \sigma_e^2)).$$

So

$$\left(I_N - X(X^TX)^{-1}X^T\right)Y \sim \mathcal{N}(0, \left(I_N - X(X^TX)^{-1}X^T\right)V(\sigma_g^2, \sigma_e^2)\left(I_N - X(X^TX)^{-1}X^T\right)^T),$$

where the normal distribution is a degenerate one. Equivalently,

$$KK^TY \sim \mathcal{N}(0, KK^TV(\sigma_g^2, \sigma_e^2)KK^T),$$

where K is $N \times N - M$ consisting of orthonormal columns with $K^TX = 0$. Now, because V is positive definite, it admits a symmetric matrix square root $V^{1/2}$. Notice that $(V^{1/2}K)^T(V^{-1/2}X) = 0$, so (skipping many steps and referring to Searle M.4f), we have:

$$KK^T(KK^TVKK^T)^+KK^T = V^{-1} - V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}.$$

([7] calls the right hand side P.) Recall that

$$\hat{\beta} = (X^TV^{-1}(\hat{\sigma}_g^2, \hat{\sigma}_e^2)X)^{-1}X^TV^{-1}(\hat{\sigma}_g^2, \hat{\sigma}_e^2)Y,$$

so our new log likelihood is

$$\begin{aligned} l_{REML}(\sigma_g^2, \sigma_e^2|X, Y, Z) &= -\frac{1}{2}(KK^TY)^T(KK^TVKK^T)^+KK^TY - \frac{1}{2}\ln|KK^TVKK^T|_+ \\ &= -\frac{1}{2}Y^TV^{-1}Y + \frac{1}{2}Y^TV^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}Y - \frac{1}{2}\ln|KK^TVKK^T|_+ \\ &= -\frac{1}{2}Y^TV^{-1}(Y - X\hat{\beta}) - \frac{1}{2}\ln|KK^TVKK^T|_+, \end{aligned}$$

where $|A|_+$ refers to the pseudo-determinant. Note:

$$\begin{aligned} (X\hat{\beta})^TV^{-1}(Y - X\hat{\beta}) &= Y^TV^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}Y - Y^TV^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}Y \\ &= Y^TV^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}Y - Y^TV^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}Y \\ &= 0 \end{aligned}$$

Hence,

$$\begin{aligned} l_{REML}(\sigma_g^2, \sigma_e^2|X, Y, Z) &= -\frac{1}{2}Y^TV^{-1}(Y - X\hat{\beta}) - \frac{1}{2}\ln|KK^TVKK^T|_+ \\ &= -\frac{1}{2}(Y - X\hat{\beta})^TV^{-1}(Y - X\hat{\beta}) - \frac{1}{2}\ln|KK^TVKK^T|_+. \end{aligned}$$

Finally, it remains to show that

$$\ln|KK^TVKK^T|_+ = \ln|V| + \ln|X^TV^{-1}X| - \ln|X^TX|.$$

How? V is $N \times N$, X^TX and $X^TV^{-1}X$ are $M \times M$, KK^TVKK^T is $N \times N$ but of rank $N - M$. Since we specified K to have orthonormal columns, I think we can write

$$\ln|KK^TVKK^T|_+ = \ln|K^TVK|.$$

Aha! Consulting Mick O'Neill's notes ([8], `https://www.stats.net.au/Maths_REML_manual.pdf`), we want to use the block matrix determinant formulas found here: `https://en.wikipedia.org/wiki/Determinant#Block_matrices`. We express this in a basis consisting of the columns of K and the columns of $\widetilde{X}$, where $\widetilde{X}$ is an orthonormal basis of the column space of X. The block matrix determinant can be written in this basis:

$$\begin{aligned}
|V| &= |I_N V I_N| = |(KK^T + \widetilde{X}\widetilde{X}^T)V(KK^T + \widetilde{X}\widetilde{X}^T)| \\
&= |(KK^T V KK^T + \widetilde{X}\widetilde{X}^T V KK^T + \widetilde{X}\widetilde{X}^T) \\
&\quad (KK^T + (KK^T V KK^T)^+ KK^T V\widetilde{X}\widetilde{X}^T + \widetilde{X}\widetilde{X}^T V\widetilde{X}\widetilde{X}^T - \widetilde{X}\widetilde{X}^T V KK^T (KK^T V KK^T)^+ KK^T V\widetilde{X}\widetilde{X}^T)| \\
&= |KK^T V KK^T|_+ \cdot |\widetilde{X}\widetilde{X}^T V\widetilde{X}\widetilde{X}^T - \widetilde{X}\widetilde{X}^T V KK^T (KK^T V KK^T)^+ KK^T V\widetilde{X}\widetilde{X}^T|_+ \\
&= |K^T VK| \cdot |\widetilde{X}^T V\widetilde{X} - \widetilde{X}^T V KK^T (KK^T V KK^T)^+ KK^T V\widetilde{X}| \\
&= |K^T VK| \cdot |\widetilde{X}^T V\widetilde{X} - \widetilde{X}^T V(V^{-1} - V^{-1}X(X^T V^{-1}X)^{-1}X^T V^{-1})V\widetilde{X}| \\
&= |K^T VK| \cdot |\widetilde{X}^T X(X^T V^{-1}X)^{-1}X^T\widetilde{X}| \\
&= |K^T VK| \cdot |(X^T V^{-1}X)^{-1}X^T\widetilde{X}\widetilde{X}^T X| \\
&= |K^T VK| \cdot |(X^T V^{-1}X)^{-1}X^T X| \\
&= |K^T VK| \cdot |(X^T V^{-1}X)^{-1}| \cdot |X^T X|.
\end{aligned}$$

Hence

$$\ln|KK^T V KK^T|_+ = \ln|K^T VK| = \ln|V| + \ln|X^T V^{-1}X| - \ln|X^T X|,$$

and therefore, our final REML expression is (see [6, 9])

$$\begin{aligned}
l_{REML}(\beta, \sigma_g^2, \sigma_e^2 | X, Y, Z) &= -\frac{1}{2}\Big((Y - X\hat{\beta})^T V^{-1}(\sigma_g^2, \sigma_e^2)(Y - X\hat{\beta}) + \ln|V(\sigma_g^2, \sigma_e^2)| \\
&\qquad - \ln|X^T X| + \ln|X^T V^{-1}(\sigma_g^2, \sigma_e^2)X|\Big) \\
&= l(\hat{\beta}, \sigma_g^2, \sigma_e^2 | X, Y, Z) + \frac{1}{2}\left(\ln|X^T X| - \ln|X^T V^{-1}(\sigma_g^2, \sigma_e^2)X|\right).
\end{aligned}$$

with $\hat{\beta}$ from (49).

4.3.3 Estimation

Maximizing the REML likelihood is a 2D optimization problem, in σ_g^2 and σ_e^2. Many software packages exist, such as GCTA [7], and a variety of methods can be used. The three that are implemented in GCTA are average information (AI), Fisher scoring, and expectation maximization (EM). See also FaST-LMM [9], which uses Brent's method and presented an influential matrix decomposition approach to speed up calculations.

4.4 An Application to Covariates

We can derive results for regressing out covariates, aka the Frisch-Waugh-Lovell theorem of econometrics, using the same machinery.

5 Linear Mixed Model Part 2: Association and Prediction

This part proceeds using empirical Bayes / type 2 maximum likelihood after we have estimated the variance parameters using REML. Let $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$ be the resulting estimates. Define the shorthand

$$\hat{V} = V(\hat{\sigma}_g^2, \hat{\sigma}_e^2) = (\hat{\sigma}_g^2/P)ZZ^T + \hat{\sigma}_e^2 I_N,$$

and let V be the true underlying covariance. Then we have from earlier:

$$Y \sim \mathcal{N}(X\beta, V), \tag{53}$$

$$\hat{\beta} = (X^T\hat{V}^{-1}X)^{-1}X^T\hat{V}^{-1}Y. \tag{54}$$

5.1 Association

We examine (54) as describing an affine transformation on Y, and use (53) and (7) to get:

$$\begin{aligned}\hat{\beta} &\sim \mathcal{N}((X^T\hat{V}^{-1}X)^{-1}X^T\hat{V}^{-1}(X\beta), (X^T\hat{V}^{-1}X)^{-1}X^T\hat{V}^{-1}V((X^T\hat{V}^{-1}X)^{-1}X^T\hat{V}^{-1})^T) \\ &= \mathcal{N}(\beta, (X^T\hat{V}^{-1}X)^{-1}X^T\hat{V}^{-1}V\hat{V}^{-1}X(X^T\hat{V}^{-1}X)^{-1}).\end{aligned}$$

We note that the true covariance V of Y is unobserved. Therefore we make an approximation by substituting $\hat{V}$ for V, hoping that this gives correct asymptotic results. The expression for $\hat{\beta}$ dramatically simplifies:

$$\hat{\beta} \approx\sim \mathcal{N}(\beta, (X^T\hat{V}^{-1}X)^{-1}).$$

In the particular case where we are testing a single SNP, X is a column vector of size N, and we have

$$\frac{\hat{\beta}-\beta}{\hat{se}(\hat{\beta})} = (\hat{\beta}-\beta)\sqrt{X^T\hat{V}^{-1}X} \approx\sim \mathcal{N}(0,1).$$

To test the null hypothesis $\beta = 0$, we can substitute the expression for $\hat{\beta}$ to get

$$\frac{X^T\hat{V}^{-1}Y}{X^T\hat{V}^{-1}X}\sqrt{X^T\hat{V}^{-1}X} \approx\sim \mathcal{N}(0,1),$$

and squaring yields:

$$\frac{\left(X^T\hat{V}^{-1}Y\right)^2}{X^T\hat{V}^{-1}X} \approx\sim \chi_1^2. \tag{55}$$

Compare with Eqs. (5) and (7) of [10] or Eq. (2) of [11], which both cite [12].

5.2 Prediction

Consider two identical sets of Eq. (43):

$$\begin{aligned}Y_1 &= X_1\beta + Z_1 b + \epsilon_1, \\ Y_2 &= X_2\beta + Z_2 b + \epsilon_2.\end{aligned}$$

The interpretation is as follows: the index 1 denotes our training set, and the index 2 denotes our test set. For set 1, we observe Y_1, X_1, and Z_1: these are used to establish inferences on β and b. For set 2, we observe X_2 and Z_2, and want to predict Y_2 by extrapolating from the training set. Note that the residuals ϵ_1 and ϵ_2 are always unobserved. The number of columns of our variables are the same as before, and there are N_1 rows for variables in the training set, and N_2 rows for variables in the test set.

The best inference for β is the maximum-likelihood estimate given earlier:

$$\hat{\beta} = (X_1^T\hat{V}_1^{-1}X_1)^{-1}X_1^T\hat{V}^{-1}Y_1.$$

In the context of prediction, $\hat{\beta}$ is called the best linear unbiased estimate (BLUE).

For inference on b, we actually get more than a point estimate. Since we have priors $P(b) = \mathcal{N}(b|0, (\sigma_g^2/P)I_P)$ and a restricted likelihood $P(Y_1|X_1, Z_1, b)$, we can get out a posterior:

$$P(b|Y_1, X_1, Z_1) = \frac{P(b)P(Y_1|X_1, Z_1, b)}{\int P(b')P(Y_1|X_1, Z_1, b')db'}.$$

The mean of this posterior, $\hat{b} = E(b|Y_1, X_1, Z_1)$, is then the best linear unbiased predictor (BLUP).

Putting these two parts together, our final prediction for Y_2 becomes:

$$\hat{Y_2} = X_2\hat{\beta} + Z_2\hat{b}.$$

Because we have already explored the maximum-likelihood estimate or BLUE, we now focus on the BLUP. To ease presentation, we consider the case when there are only random effects, so that the restricted likelihood becomes just a likelihood:

$$P(Y_1|Z_1, b) = \mathcal{N}(Y_1|Z_1 b, \sigma_e^2 I_{N_1}).$$

For simplicity, we will also occassionally drop the subscript 1 when only the training set is being considered.

5.2.1 SNP or Basis Perspective

Taking Z as always known, we have:

$$P(b) = \mathcal{N}(b|0, (\sigma_g^2/P)I_P), \tag{56}$$

$$P(Y|b) = \mathcal{N}(Y|Zb, \sigma_e^2 I_N). \tag{57}$$

Thus, we can apply (13) to get

$$p(b|Y) = \mathcal{N}(b|\Xi\{A^T\Pi^{-1}(Y-0) + \Sigma^{-1}\mu\}, \Xi),$$

where we fill in:

$$\begin{aligned} A &= Z, \\ \mu &= 0, \\ \Sigma &= (\sigma_g^2/P)I_P, \\ \Pi &= \sigma_e^2 I_N, \\ \Xi = (\Sigma^{-1} + A^T\Pi^{-1}A)^{-1} &= \left\{(P/\sigma_g^2)I_P + (1/\sigma_e^2)Z^TZ\right\}^{-1}. \end{aligned}$$

Performing the substitutions yields:

$$p(b|Y) = \mathcal{N}\left(b|\left\{(P/\sigma_g^2)I_P + (1/\sigma_e^2)Z^TZ\right\}^{-1}(Z^TY/\sigma_e^2), \left\{(P/\sigma_g^2)I_P + (1/\sigma_e^2)Z^TZ\right\}^{-1}\right), \tag{58}$$

And thus, substituting in $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$ from REML, our BLUP is:

$$\hat{b} = E(b|Y) = \left(Z^TZ + \frac{P\hat{\sigma}_e^2}{\hat{\sigma}_g^2}I_P\right)^{-1}Z^TY \tag{59}$$

$$= \left(\frac{1}{N}Z^TZ + \frac{P\hat{\sigma}_e^2}{N\hat{\sigma}_g^2}I_P\right)^{-1}\frac{Z^TY}{N}. \tag{60}$$

There are at least two ways to interpret the above expression. The first form, (59), highlights that what we are doing is identical to ridge regression, with the ridge parameter being $\lambda = P\hat{\sigma}_e^2/\hat{\sigma}_g^2$. This makes sense since our prior on b is equivalent to L2 regularization, and because the posterior mean and posterior mode for a Gaussian are the same.

In the second form, (60), the expression Z^TY/N corresponds to the summary statistics obtained by doing univariate linear regression of genome-wide SNPs against the phenotype. We then multiply these summary statistics by a shrinkage matrix to get our optimal predictor. Within this shrinkage matrix, Z^TZ/N is exactly the LD matrix or covariance matrix between SNPs. This is the presentation followed in the LDPred paper [13].

5.2.2 Sample or Kernel Perspective

Now let us go back to considering a training and test set. Once we have learned $\hat{b}$ from the training set, our prediction on the test set is:

$$\hat{Y}_2 = Z_2\hat{b} = Z_2\left(Z_1^TZ_1 + \frac{P\hat{\sigma}_e^2}{\hat{\sigma}_g^2}I_P\right)^{-1}Z_1^TY_1.$$

Define

$$k = \frac{\hat{\sigma}_g^2}{P\hat{\sigma}_e^2}. \tag{61}$$

We rewrite the matrix inverse using the matrix inversion lemma (3):

$$\begin{aligned}\left(Z_1^T Z_1 + \frac{I_P}{k}\right)^{-1} &= \left(\frac{I_P}{k}\right)^{-1} - \left(\frac{I_P}{k}\right)^{-1} Z_1^T \left(I_{N_1} + Z_1 \left(\frac{I_P}{k}\right)^{-1} Z_1^T\right)^{-1} Z_1 \left(\frac{I_P}{k}\right)^{-1} \\ &= kI_P - k^2 Z_1^T \left(I_{N_1} + kZ_1 Z_1^T\right)^{-1} Z_1.\end{aligned}$$

So

$$\begin{aligned}\hat{Y}_2 &= Z_2 \left\{kI_P - k^2 Z_1^T \left(I_{N_1} + kZ_1 Z_1^T\right)^{-1} Z_1\right\} Z_1^T Y_1 \\ &= \left\{kZ_2 Z_1^T - kZ_2 Z_1^T \left(I_{N_1} + kZ_1 Z_1^T\right)^{-1} kZ_1 Z_1^T\right\} Y_1 \\ &= \left\{kZ_2 Z_1^T - kZ_2 Z_1^T \left(I_{N_1} + kZ_1 Z_1^T\right)^{-1} \left(I_{N_1} + kZ_1 Z_1^T - I_{N_1}\right)\right\} Y_1 \\ &= \left\{kZ_2 Z_1^T - kZ_2 Z_1^T + kZ_2 Z_1^T \left(I_{N_1} + kZ_1 Z_1^T\right)^{-1}\right\} Y_1 \\ &= kZ_2 Z_1^T \left(I_{N_1} + kZ_1 Z_1^T\right)^{-1} Y_1. \qquad (62)\end{aligned}$$

What is surprising is that the matrices Z_1 and Z_2 only enter into the expression via the forms $Z_1 Z_1^T$ and $Z_2 Z_1^T$. These matrices, N_1 by N_1 and N_2 by N_1 respectively, occur in sample space, not in SNP space. In fact, they represent a SNP-based similarity among samples. If we form $Z_1 Z_1^T / P$, then we have exactly the kinship matrix or genome relatedness matrix (GRM).

So if we had Y_1 and wanted to predict Y_2, it turns out that the matrices Z_1 and Z_2 themselves are not needed. Once we have estimates $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$, all we need is a kernel function $\mathcal{K}(z, z')$ describing how similar two samples are. In this case, we are using a linear kernel, $\mathcal{K}(z, z') = z^T z'/P$, where z and z' are vectors of size P containing genome-wide SNP data for the two samples. This kernel-based perspective is also how Gaussian processes are motivated from Bayesian linear regression (see Bishop Sections 3.3.3 and 6.4), and often more complex kernels are then used.

5.2.3 Connection with Association

To connect our results on mixed-model prediction with association, let us first consider prediction on the training set. We can substitute Z_1 for Z_2 in (62) to get:

$$\hat{Y}_1 = kZ_1 Z_1^T \left(I_{N_1} + kZ_1 Z_1^T\right)^{-1} Y_1,$$

and work from there. Alternatively, here is another approach I came up with, which uses the Taylor expansion for $(I + A)^{-1}$. We have:

$$\begin{aligned}\hat{Y}_1 = Z_1 \hat{b} &= Z_1 \left(Z_1^T Z_1 + \frac{I_P}{k}\right)^{-1} Z_1^T Y_1 \\ &= kZ_1 \left(I_P + kZ_1^T Z_1\right)^{-1} Z_1^T Y_1 \\ &= kZ_1 \left(I_P - kZ_1^T Z_1 + (kZ_1^T Z_1)^2 - \cdots\right) Z_1^T Y_1 \\ &= \left(kZ_1 Z_1^T - (kZ_1 Z_1^T)^2 + (kZ_1 Z_1^T)^3 - \cdots\right) Y_1 \\ &= \left(I_{N_1} - \left(I_{N_1} - kZ_1 Z_1^T + (kZ_1 Z_1^T)^2 - \cdots\right)\right) Y_1 \\ &= \left(I_{N_1} - \left(I_{N_1} + kZ_1 Z_1^T\right)^{-1}\right) Y_1.\end{aligned}$$

Now, recall that we had defined

$$\begin{aligned}\hat{V} &= (\hat{\sigma}_g^2/P) ZZ^T + \hat{\sigma}_e^2 I_N \\ &= \hat{\sigma}_e^2 \left(I_{N_1} + \frac{\hat{\sigma}_g^2}{P\hat{\sigma}_e^2} Z_1 Z_1^T\right) \\ &= \hat{\sigma}_e^2 \left(I_{N_1} + kZ_1 Z_1^T\right).\end{aligned}$$

Thus,

$$\begin{aligned}\hat{Y}_1 &= \left(I_{N_1} - \left(I_{N_1} + kZ_1Z_1^T\right)^{-1}\right) Y_1 \\ &= \left(I_{N_1} - (\hat{V}/\hat{\sigma}_e^2)^{-1}\right) Y_1 \\ &= Y_1 - \hat{\sigma}_e^2 \hat{V}^{-1} Y_1.\end{aligned}$$

Hence if we form the residuals from BLUP on the training set, and call these $\widetilde{Y}$ (we now get rid of the subscript 1), we have

$$\widetilde{Y} = Y - \hat{Y} = \hat{\sigma}_e^2 \hat{V}^{-1} Y. \tag{63}$$

Earlier we derived a chi-squared test (55) for association of a fixed effect β for a single SNP X:

$$\frac{(X^T \hat{V}^{-1} Y)^2}{X^T \hat{V}^{-1} X} \approx\sim \chi_1^2.$$

If we substitute our formula (63) for the BLUP residuals into this expression, we have:

$$\frac{\left(X^T(\widetilde{Y}/\hat{\sigma}_e^2)\right)^2}{X^T \hat{V}^{-1} X} = \frac{\left(X^T \widetilde{Y}/N\right)^2}{X^T \hat{V}^{-1} X (\hat{\sigma}_e^2/N)^2} \approx\sim \chi_1^2.$$

The numerator of this expression is precisely the square of a "summary statistic" between a standardized SNP X and the residual phenotype $\widetilde{Y}$ from BLUP. Indeed, the BOLT-LMM paper [10] observes that "the $\chi^2_{BOLT-LMM-inf}$ statistic is equivalent to computing (and then calibrating) the squared correlations between SNPs x_{test} and BLUP residuals" (Online Methods p. 1). BOLT-LMM additionally notes that "in human genetics applications, the denominator ... $x'_{test} V^{-1} x_{test}$, is nearly indepedent of the SNP x_{test} being tested" (Online Methods p. 1, with a citation to [11]). This simplifies the computation from performing a matrix-vector product per SNP to estimating a constant, which they do by sampling 30 pseudorandom SNPs. The resulting statistic is

$$\chi^2_{BOLT-LMM-inf} = \frac{1}{c_{inf}} \frac{N\left(X^T \widetilde{Y}\right)^2}{(X^T X)(\widetilde{Y}^T \widetilde{Y})}.$$

According to values in the BOLT-LMM supplement (Supplementary Tables 14 and 15), the estimated values of c_{inf} are often very close to but slightly less than 1. If we set $c_{inf} = 1$, we get the GRAMMAR test statistic [14], which [15] observes is slightly underpowered. The idea to estimate c_{inf} was introduced in the GRAMMAR-Gamma paper [11].

5.2.4 Alternate Derivation

Here we offer a shorter derivation of (62) that does not require ever entering into SNP / basis space. Recalling that $b \sim \mathcal{N}(0, (\sigma_g^2/P) I_P)$, we have the properties:

$$\begin{aligned}E(Y_1) &= E(Z_1 b + \epsilon_1) = 0, \\ E(Y_2) &= E(Z_2 b + \epsilon_2) = 0, \\ \mathrm{Var}(Y_1) &= \mathrm{Var}(Z_1 b + \epsilon_1) = \mathrm{Var}(Z_1 b) + \mathrm{Var}(\epsilon_1) = \frac{\sigma_g^2}{P} Z_1 Z_1^T + \sigma_e^2 I_{N_1}, \\ \mathrm{Var}(Y_2) &= \mathrm{Var}(Z_2 b + \epsilon_2) = \mathrm{Var}(Z_2 b) + \mathrm{Var}(\epsilon_2) = \frac{\sigma_g^2}{P} Z_2 Z_2^T + \sigma_e^2 I_{N_2}, \\ \mathrm{cov}(Y_1, Y_2) &= \mathrm{cov}(Z_1 b + \epsilon_1, Z_2 b + \epsilon_2) = \mathrm{cov}(Z_1 b, Z_2 b) = Z_1 \mathrm{Var}(b) Z_2^T = \frac{\sigma_g^2}{P} Z_1 Z_2^T.\end{aligned}$$

Thus, we can model the joint distribution of Y_1 and Y_2 as a multivariate normal satisfying

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{\sigma_g^2}{P} Z_1 Z_1^T + \sigma_e^2 I_{N_1} & \frac{\sigma_g^2}{P} Z_1 Z_2^T \\ \frac{\sigma_g^2}{P} Z_2 Z_1^T & \frac{\sigma_g^2}{P} Z_2 Z_2^T + \sigma_e^2 I_{N_2} \end{bmatrix} \right).$$

From (10), we then have

$$p(Y_2|Y_1) = \mathcal{N}(Y_2|\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(Y_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}),$$

for which our mean is

$$\begin{aligned} E(Y_2|Y_1) &= \Sigma_{21}\Sigma_{11}^{-1} Y_1 \\ &= \frac{\sigma_g^2}{P} Z_2 Z_1^T \left(\frac{\sigma_g^2}{P} Z_1 Z_1^T + \sigma_e^2 I_{N_1} \right)^{-1} Y_1, \end{aligned}$$

which matches (62), except that that expression has used inferences for $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$ in place of their true values.

This is more similar to the presentation seen in Bishop Chs. 3 and 6. Note that to derive (10), Bishop uses a matrix identity (2.76) which they call the "matrix inversion formula" on p. 91. Proving that turns out to be of comparable difficulty to proving the matrix inversion lemma (3), which we used in our earlier derivation.

5.2.5 Small N, large P limit

The BLUP on the training set is:

$$\hat{Y} = \frac{\hat{\sigma}_g^2}{P} Z Z^T \left(\frac{\hat{\sigma}_g^2}{P} Z Z^T + \hat{\sigma}_e^2 I_N \right)^{-1} Y.$$

If N is small and individuals are randomly sampled from the population, then we expect no close relatives. In this case, especially as we take P to be large, we expect $ZZ^T \approx P I_N$. Hence

$$\begin{aligned} \hat{Y} &\approx \hat{\sigma}_g^2 I_N \left(\hat{\sigma}_g^2 I_N + \hat{\sigma}_e^2 I_N \right)^{-1} Y \\ &= \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_e^2} Y \\ &= \hat{h^2} Y. \end{aligned}$$

In other words, regardless of the phenotype Y, we obtain a BLUP that is a rescaled version of Y.

6 Notes on Modern Genetics

6.1 LDpred

6.2 Expected Heritability and MAF

Negative selection, LDAK α parameter

6.3 Non-infinitesimal Priors

6.4 LD Score Regression and Cousins

7 References

References

[1] George Casella and Roger L. Berger. *Statistical Inference*. Duxbury Resource Center, 2nd edition, 2002.

[2] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Science & Business Media, 2004.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2nd edition, February 2009.

[4] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer Science & Business Media, 2006.

[5] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[6] Hyun Min Kang, Noah A Zaitlen, Claire M Wade, et al. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, 2008.

[7] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *American journal of human genetics*, 88(1):76–82, 01 2011.

[8] Mick O'Neill. The mathematics of reml, December 2013.

[9] Christoph Lippert, Jennifer Listgarten, Ying Liu, et al. Fast linear mixed models for genome-wide association studies. *Nature Methods*, 8:833 EP –, 09 2011.

[10] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47:284 EP –, 02 2015.

[11] Gulnara R Svishcheva, Tatiana I Axenovich, Nadezhda M Belonogova, Cornelia M van Duijn, and Yurii S Aulchenko. Rapid variance components–based method for whole-genome association analysis. *Nature Genetics*, 44:1166 EP –, 09 2012.

[12] Wei-Min Chen and Goncalo R Abecasis. Family-based association tests for genomewide association scans. *American journal of human genetics*, 81(5):913–926, 11 2007.

[13] Bjarni J Vilhjálmsson, Jian Yang, Hilary K Finucane, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *American journal of human genetics*, 97(4):576–592, 10 2015.

[14] Yurii S Aulchenko, Dirk-Jan de Koning, and Chris Haley. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, 177(1):577–585, 09 2007.

[15] William Astle and David J. Balding. Population structure and cryptic relatedness in genetic association studies. *Statist. Sci.*, 24(4):451–471, 2009.