# Joint Multivariate Modelling and Prediction for Genetic and Biomedical Data

Mahfuzur Rahman Khokan

*BSc., University of Dhaka. MSc., University of Dhaka, University of Nottingham*

School of Mathematics & Statistics

Faculty of Science, Technology, Engineering, and Mathematics (STEM)

The Open University, United Kingdom


This dissertation is submitted for the degree of

Doctor of Philosophy

September 2023

# Abstract

In the area of statistical genetics, classical genome-wide association studies (GWAS) assess the association between a biological characteristic and genetic variants, working with one variant at a time in a regression model, and reporting the most significant associations. These studies test genetic markers individually, even though the data may exhibit multivariate structure due to the way genes are transmitted together from the parents to the offspring. Despite considering covariates like age and sex in the model, the classical GWAS does not account for the joint effects of genetic variants. Moreover, when multiple genetic variants within a gene have small effects on a phenotype, testing them individually can lack statistical power, but testing them together in a joint model can be more useful in pooling together all the evidence. In this thesis, I reviewed different multivariate testing procedures in joint multivariate model settings, explored their properties, and demonstrated them in further real-life database applications, such as enhancing statistical power by conditioning on major variants.

I studied the mathematical properties of various multivariate test procedures, particularly within the context of multiple linear regression. Considering the theoretical aspect as well as their availability in literature, I adapt various multivariate test procedures for canonical correlation in multiple regression settings. These procedures have been demonstrated to asymptotically follow the chi-square distribution. Importantly, these test procedures exhibit asymptotic equivalence among themselves and with the Wald test statistic. This indicates that the Wald test statistic may be sufficient for future studies, given its equivalence to the multivariate test procedures.

In many cases, there are known databases of major genetic variants that have a substantial effect on the trait. In such situations, it makes sense statistically to condition on these major variants to improve power in detecting associations with new variants,

but this is not a common practice in GWAS applications. In this study, we also showed theoretically and computationally how conducting a joint analysis of the genetic variants in a multiple regression model, where the estimated effect of a new variant is conditioned upon some major variants, can improve the performance of the model in terms of reducing the standard error and improving the power. The amount of gain of power will depend on the correlation between the response and the covariates, as well as the correlation between the covariates. I further show that conditional results can sometimes be obtained from publicly available summary statistics reported for univariate associations in published GWAS studies, even when the individual-level data are unavailable. A prominent example of such a trait is skin color, for which there are many studies consistently identifying a handful of major genes. I looked into a dataset of over 6,500 mixed-ethnicity Latin Americans to see how the conditioning process can improve the detection power of GWAS studies and identify new genetic variants in such a situation.

In practical applications, the statistical models I worked with for association testing can be carried forward for predictive purposes in new datasets. In this thesis, I have also demonstrated mathematical formulations of prediction errors in different linear models, including simple linear regression models, as well as shrinkage methods like ridge regression and lasso regression. These expressions for prediction errors show the inherent trade-off between bias and variance at both individual data points and across a set of observations. Moreover, these formulations have found the connections between prediction errors and genetic heritability that can enhance prediction performance in genetic association studies. Additionally, I reviewed various statistical and machine learning predictive models. Based on a dental morphology dataset, I compared their performance using classification metrics such as average error rate and maximum classification error rate per specimens.

# Dedication

*This thesis is dedicated to my beloved parents, wife, and two smart kids for their endless love, support, and ecouragement.*

*I also extend my heartfelt dedication to my cherished country, Bangladesh, a nation distinguished by its abundant heritage and vibrant culture.*

*I am grateful to the esteemed institution, the University of Dhaka, which has played an instrumental role in enriching my academic journey and fostering a spirit of knowledge and excellence.*

# Acknowledgement

Above all, I express my deepest gratitude to Almighty ALLAH for His enduring mercy and for granting me the strength and resilience necessary to overcome the challenges I encountered on this journey. It is through His blessings that I have been able to thrive despite the adversities faced. I would like to extend my profound gratitude to my lead supervisor, Dr. Kaustubh Adhikari, for his outstanding guidance and mentorship throughout this research journey. His unwavering dedication, enthusiastic support, and invaluable feedback have not only influenced the development of this thesis but have also expanded my research network through collaborations, participation in engaging public engagement events, writing research grants, and delivering talks on various platforms. I am grateful to my internal advisor, Dr. Fadlalla Elfadaly, for his constructive discussions, continuous support, and valuable feedback on my work.

Moreover, I extend my gratitude to the University of Dhaka, Bangladesh, for granting me study leave to pursue my doctoral degree. I am also thankful for the financial support provided through the prestigious Bangabandhu Overseas Scholarship, which has been instrumental in enabling me to pursue my research endeavors. I would also like to extend my sincere thanks to the Open University for providing an excellent and stimulating research environment. Many thanks to Sagnik Palmal and Soumya Paria for their valuable contributions and insightful discussions on various topics in genetics. Their constructive engagement has significantly enhanced my understanding of the fundamentals of genetics and has propelled me forward on my research journey. I am truly grateful for their support and the knowledge they have shared with me. Being in Milton Keynes has been a wonderful experience, and I consider myself fortunate and proud to be part of such a nice community. I sincerely grateful to all of my friends, especially Sagor bhai, Naznin Apu, Tanvir bhai, Shipa Apu, Mohabbat bhai, Sumaya

Apu, Riasat, priota, and Shisir for the fantastic moments we shared together.

Heartfelt thanks to my parents, brother, and sisters for their faithful wishes and prayers throughout my whole academic journey. I am truly and heartily grateful to my wife, Farhana Akond Pramy, for her incredible support and patience over the years. I consider myself truly fortunate to have such an exceptional life partner by my side. Lastly, I would like to thank Almighty Allah for blessing us with the most precious gifts, Muhammad Muhtadin Mahran and Muhammad Zaid Mahnaf. Their presence has been a constant source of inspiration throughout every stage and every word of this thesis. I am forever grateful for their love, support, and the joy they bring to my life.

# Contents

# 1 Introduction

In recent years, there has been an explosion of projects producing data from thousands of people over millions of genetic markers, for example, 500,000 volunteers in the UK BioBank [Thompson and Willeit (2015)], 500,000 Finnish individuals in FinnGen [Kurki et al. (2023)], 54,000 US participants in TOPMed [Taliun et al. (2021)], and 200,000 volunteers in BioBank Japan [Nagai et al. (2017)]. Consequently, the need for advanced statistical methods to deal with high-dimensional data became more prominent. With this explosion in genetic studies came an explosion in proposed statistical tools too, but most of the proposed methods such as GCTA [Yang et al. (2011)] do not deal with joint multivariate analysis of the input data.

Genome-wide association studies (GWAS) require two main types of input data on sampled individuals: millions of genetic markers (genotypes) and physical characteristics (phenotypes). The GWAS study aims to explore the presence of genetic variants linked to specific physical traits, such as skin color, eye color, height, or diseases. Genetic variation refers to the changes in gene sequences occurring at distinct locations in DNA across individuals within a population. These variations are carefully considered and analyzed separately to identify potential associations with the studied traits, for example, skin color.

**Figure: Skin Color Variation, a Biological Trait of Human Beings.** (https://www.earth.com/news/color-human-skin-complicated/)

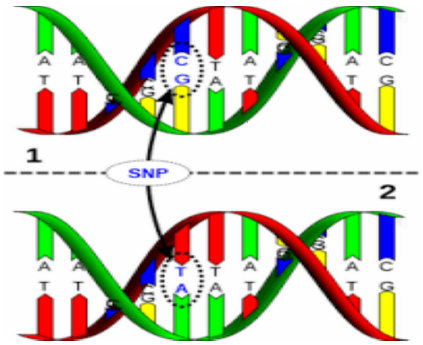



**Figure: Single Nucleotide Polymorphism (SNP) in DNA.** (http://en.wikipedia.org/wiki/Single-nucleotide_polymorphism)

Traditional analysis only performed univariate linear regression analysis by assessing one genetic marker against one phenotype at a time [Burton et al. (2007)]. An increasing number of proposed statistical methods propose more sophisticated models such as Bayesian methods [Lloyd-Jones et al. (2019); Speed et al. (2012)], but most methods do not leverage the multivariate structure in the data, for example, the correlated nature of the variables.

Biological characteristics (traits) such as skin color or height are influenced by many genes at the same time – tens, hundreds or even thousands of genetic variants contributing to a single characteristics have been discovered. And due to the way genes are transmitted together from the parents to the offspring, genetic variants that are close by on a chromosome are also correlated, a phenomenon known as LD (linkage disequilibrium). This implies that a joint analysis of all the predictors (genetic variants) in an unified model is the best way of analyzing a physical characteristic [Adhikari et al. (2015)].

In this PhD project, I have therefore worked with multivariate statistical models for joint analysis of the genotype-phenotype data. The research involved a combination of theoretical and computational approaches. Such methods proved valuable not only in identifying novel gene associations with new phenotypes but also in enhancing the prediction of physical characteristics from genetic data, thereby contributing to forensic reconstructions [Adhikari et al. (2016a)].

The project mainly has two main, interconnected parts along with some collaborative research works. The first part focuses on the identification of genetic variants that contribute to a trait. This is done with genome-wide association stud-

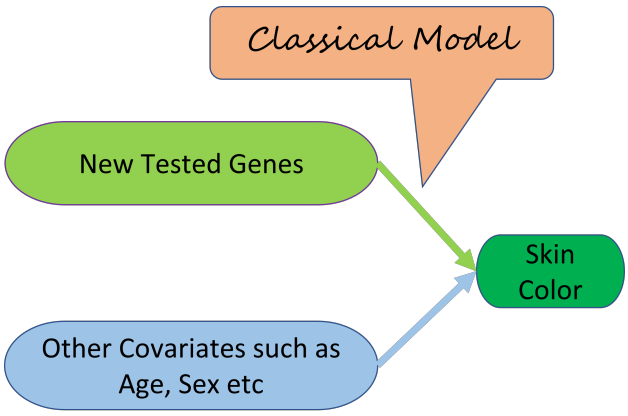





**Diagram: Classical GWAS Model for Genetic Association**

ies (GWAS), which in its simplest form do association tests between a single trait and a single genetic variant, while adjusting for the potential effects of other covariates, such as the age of the participants.

To explore the association testing procedures, I initiated the research work by conducting a comprehensive literature review of existing test procedures. Among the available options, I explored the multivariate test procedures of canonical correlation, considering their prevalence in the literature. In this context, I studied their mathematical properties, particularly within the framework of multiple linear regression, and demonstrated their asymptotic convergence to the chi-square distribution. Moreover, I showed them as an asymptotic equivalence among themselves and with the Wald Test. Consequently, I recommend the Wald Test for further studies, serving as an equivalent choice for all tests, particularly canonical correlation in multiple regression scenarios.

After obtaining the proposed test procedure, I worked with a different avenue. For example, I explored various multiple regression settings to investigate the impact of conditioning on statistical power. The theoretical investigation aimed to determine whether, in addition to conditioning on common covariates like age and sex of the participants, conditioning on known genetic variants with significant effects on the trait can enhance the power in detecting new, smaller-effect genetic variants

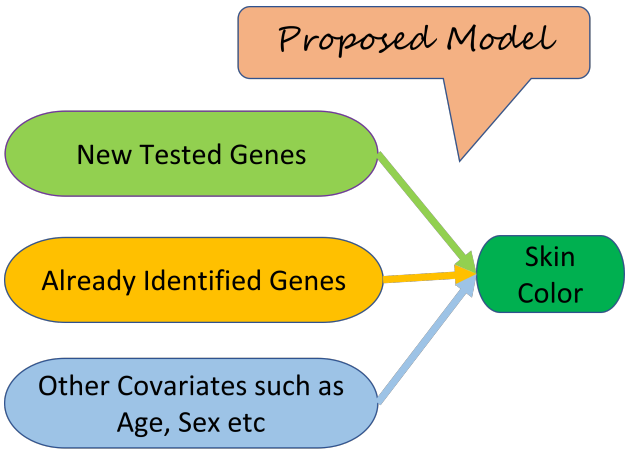



**Diagram: Proposed Conditional Model for Genetic Association**

[Yang et al. (2012)].

I explained the nature of power gain considering the behaviour of genetic variants within chromosomes, including aspects like LD structures and proxy variants. Theoretical developments were demonstrated to highlight the potential improvement in statistical power, and real-life databases such as the CANDELA cohort and UK Biobank were used to validate these findings. Moreover, I showed that these conditional results can be approximated using publicly available summary statistics from GWAS databases, even in scenarios where individual-level data is not available.

A prediction problem can arise in a real-life situation in a few different ways. For example, a geneticist may be interested in predicting a trait using already identified genetic variants, which were published in the literature via GWASes. Alternatively, the geneticist may want to first conduct a GWAS themself and proceed with the identified variants. Each of these approaches reduces the number of variants used in the prediction to a more manageable number, such as tens, hundreds, or thousands. Alternatively, other prediction methods are interested in analyzing the entire genome with millions of variants at once and let the statistical approach pick the best set of variants simultaneously while predicting, e.g. through a method like LASSO.

In prediction, the primary goal is to maximize prediction accuracy rather than testing and statistical power. However, the statistical models used in testing, such as linear models and shrinkage methods, can also be carried out for predictive purposes in new datasets and the predictive ability can be calculated with prediction accuracy or errors. A theoretical derivation of prediction accuracy is often more difficult for any particular method than the theoretical derivation of its test statistic or power while testing. Nonetheless, by exploring the pros and cons of various linear and shrinkage methods, which provide a trade-off between bias and variance, we can improve predictive performances during prediction, which is the second part of this thesis.

This PhD thesis is organized into seven chapters. After the introductory Chapter, Chapter 2 provides a brief review of linear models commonly applied in the field of statistical genetics. Following the relationship between linear mixed models with the ordinary least square regression and shrinkage methods demonstrated by Bates et al. (2014); Hoffman (2013), I expressed these relationships in a more simplified manner to illustrate how the estimates of these models are linked.

Chapter 3 of this thesis delves into the discussion of various multivariate testing procedures, with a particular focus on procedures for canonical correlation due to their prevalence in the existing literature. By adapting these test procedures to multiple regression settings, I analyzed their approximate distributions and established their equivalence to each other and the Wald test statistic. Consequently, I recommended a unified test procedure for association testing in future studies.

Chapter 4 represents a significant contribution to this thesis, where I explored the concept of power gain resulting from conditioning on major genes, both theoretically and computationally. The theoretical developments of statistical power are primarily based on multiple linear regression settings, incorporating various design matrix forms.

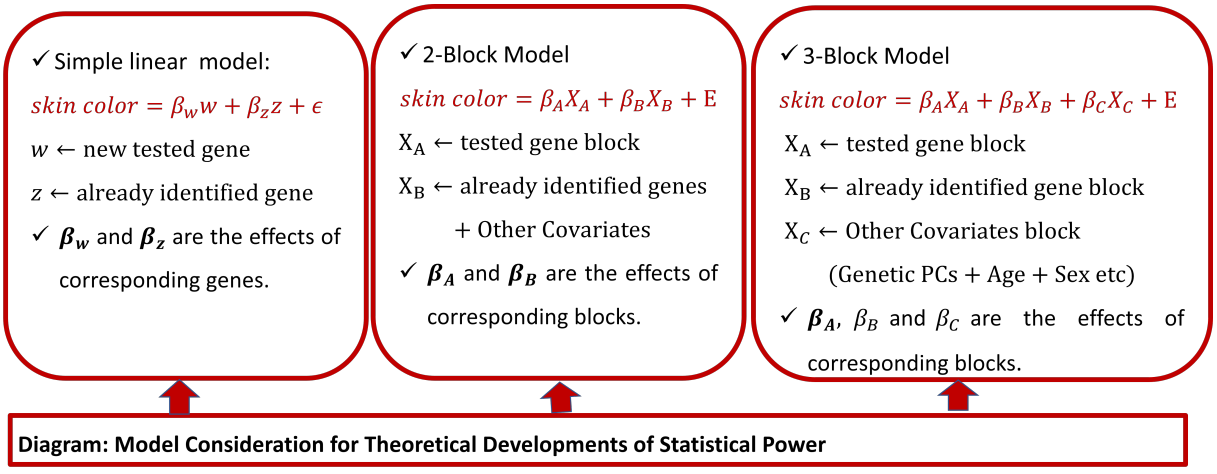



**Diagram: Model Consideration for Theoretical Developments of Statistical Power**

Moreover, I discussed the nature of power gain, considering gene structures such as LD (Linkage Disequilibrium) and the sample size in the database. To demonstrate the

concept, I conducted analyses using two genetic databases: the CANDELA cohort and UKBiobank. In particular, I derived mathematical expressions for statistical power when the design matrix is a 3-Block matrix and explained the extent to which power improves through conditioning on one block. Acknowledging that the statistical power depends on the significance of conditional genes, I also presented a mathematical approach for computing conditional results using summary statistics, even when individual-level datasets are not available.

In Chapter 5, I provided the mathematical expressions of prediction errors for various methods, such as the simple linear regression model and ridge regression. These expressions demonstrate how prediction errors can be related to genetic association studies, allowing us to understand genetic heritability based on the prediction error of a model.

Chapter 6 and Chapter 7 focus on collaborative research works related to 'Dental Morphology Data' and 'Facial Morphology Data,' respectively. In each chapter, I have provided a detailed account of the research work, outlining my specific contributions to the studies.

# 2 Literature Review on Linear Regression Models

## 2.1 Review of Linear Regression

Suppose, $y_i$ denotes a response variable which is a linear function of a set of $p$ covariates $x_1, x_2, \cdots, x_p$, and $n$ is the number of observations. Then the linear model can be modelled as [Johnson and Wichern (2007)]

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i \tag{2.1}$$

Here, $\beta' s$ are the regression coefficients, and $\epsilon_i$ indicates the error term and follows a normal distribution with mean zero and variance $\sigma_\epsilon^2$ and to be independent across the samples that is $\text{cov}(\epsilon_i, \epsilon_j) = 0$.

If the response variable $\mathbf{y}$, the input matrix $\mathbf{X}$, the regression coefficient vector $\beta$ and the error vector $\boldsymbol{\epsilon}$ are defined as follows

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \text{ and } \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Then the linear model in (2.1) can be written in matrix form as

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon} \text{ with } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{I_n}) \tag{2.2}$$

where, $\sigma_\epsilon^2$ is the error variance covariance matrix and $\boldsymbol{I_n}$ denotes the $n \times n$ identity matrix. Note that the statistical models discussed in Chapter 4, concerning the power gain through conditioning, have been considered as a mean standardized, implying the

exclusion of intercept terms within the model.

The sample variance-covariance matrix and sample correlation matrix can be defined in multiple regression setups as

$$S = \begin{pmatrix} S_y^2 & S_{yx}^T \\ S_{xy} & S_{xx} \end{pmatrix} \text{ and } R = \begin{pmatrix} 1 & r_{yx}^T \\ r_{xy} & R_{xx} \end{pmatrix}$$

The variance-covariance matrix of the response variables $Y$, in multivariate case, is expressed as $S_{yy} = \frac{1}{n}\sum_{k=1}^{n}(Y_k - \bar{Y})(Y_k - \bar{Y})$. For the single response variable case, $S_{yy}$ is equivalent to $S_y^2$. The vector $r_{yx}^T$ indicates sample correlation between $y$ and $X_i;\ i = 1, 2, \cdots, p$. Following Johnson and Wichern (2007), it is noted that the sample correlation coefficient $R_{ij}$ can be obtained in terms of the covariance $S_{ij}$ and variances $S_{ii}$ and $S_{jj}$ as follows

$$R_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}}\sqrt{S_{jj}}}$$

### 2.1.1 Maximum Likelihood Estimation (MLE)

Using the maximum likelihood method, the estimates of the unknown parameter $\beta$ and $\sigma_\epsilon^2$ can be obtained by maximizing the log-likelihood function, $\log L(\beta, \sigma_\epsilon^2|y, X)$ depending on the given dataset $X$ and $y$. Suppose $\hat{\beta}$ and $\hat{\sigma}_\epsilon^2$ are the MLE of $\beta$ and $\sigma_\epsilon^2$ respectively, then mathematically it can be written that

$$\hat{\beta}, \hat{\sigma}_\epsilon^2 = \underset{\beta, \sigma_\epsilon^2}{\operatorname{argmax}} \log L(\beta, \sigma_\epsilon^2|y, X) \tag{2.3}$$

and the solution of the maximum likelihood estimator can easily be obtained as

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T y \\ \hat{\sigma}_\epsilon^2 &= \frac{1}{n}(y - \hat{y})^T (y - \hat{y}) \end{aligned} \tag{2.4}$$

$$= \frac{1}{n}\Big(y - X\hat{\beta}\Big)^T\Big(y - X\hat{\beta}\Big) \tag{2.5}$$

here, $n$ is the number of responses, $\hat{\beta}$ is the ordinary least squares (OLS) estimator of $\beta$, $\hat{\sigma}^2_\epsilon$ will be computed using the value of $\hat{\beta}$ and $\hat{y} = X\hat{\beta}$ is the fitted value.

In Genome-wide association studies (GWAS), the linear model is a widely used approach to investigate the relationship between genetic variants (genotypes) and phenotypic traits. Traditional GWAS typically tests the association between one genetic marker and one characteristic at a time, while considering other covariates such as age, sex, and genetic principal components in the model.

However, the linear model can also be used to perform a joint analysis of multiple variants, enabling us to explore how the phenotypic trait is influenced by the combined effects of multiple genetic variants.

GWAS aims to identify genetic variants that show significant associations with the phenotypic trait under the study and to achieve this, various statistical tests are used to assess the significance of the regression coefficients $(\beta_1, \beta_2, \cdots, \beta_p)$, representing the effect of each genetic variant on the phenotypic trait.

In Chapter 3 of this thesis, I have revisited the properties of various statistical tests and a detailed comparison has also been made among them. Chapter 4 focuses on the concept of power gain resulting from conditioning. The key mathematical developments are centered around linear regression models, elucidating their various properties and relevance for statistical power calculation. Once the concept is validated in linear model settings, I further extended the exploration of power gain to multiple linear models with diverse forms of the design matrix.

## 2.2 Review of Linear Mixed Model (LMM)

In statistical genetics, the genome-wide association studies (GWAS) is the most widely used technique to analyze the thousands of millions of genetic data and traditionally it performs association testing considering a single genetic marker with a single phenotype at a time [Burton et al. (2007)]. But in practice, phenotypic characteristics may be influenced by many genetic markers simultaneously and the genotype-phenotype association may be misled by the effect of confounding factors such as population structure [Patterson et al. (2006)]. For example, skin color may be jointly influenced by many genetic markers as well as may be influenced by some hidden factors such as environment, geographical region, age, sex, and other contextual variables [Kang et al. (2008)]. To incorporate the joint effect of multiple markers as well as to deal with different confounding factors, the linear mixed model (LMM) is another attractive tool in statistical genetics [Kang et al. (2010, 2008)]. The widely used linear mixed model (LMM) is given by

$$y = X\beta + Zu + \epsilon \tag{2.6}$$

where, $y_{n\times 1}$ is a vector of responses, $X_{n\times p}$ is a design matrix for the fixed effects, $\beta_{p\times 1}$ is a vector of fixed effect parameters, $Z_{n\times q}$ is a design matrix for the random effects, $u_{q\times 1}$ is a vector of random effect which follows $u \sim N(0, G\sigma^2)$ and $\epsilon_{n\times 1}$ is a vector of residuals which follows $\epsilon \sim N(0, R\sigma^2)$.

Following Patterson and Thompson (1971), the variance-covariance matrix of the data, $y$ can be written as

$$\text{var(y)} = \sigma^2(ZGZ^T + R) = \sigma^2 H$$

where,

$$H = ZGZ^T + R$$

The matrix $H$ consists of two components that are used to model heteroscedasticity and correlation: a random effects component $ZGZ^T$ and a within-group component $R$.

The estimates of $\beta$ and $u$ can be obtained by solving the score equations obtained from the log-joint distribution of $(y, u)$ and these equations are called the mixed model equation (MMEs) as proposed by Henderson et al. (1959) and the equations can be written in matrix form as

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \tilde{u} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix}$$

Now the estimation of variance parameters can be obtained by using the maximum likelihood method and residual maximum likelihood (REML) method. The maximum likelihood estimators of the variance parameters are biased downwards, especially in small samples, because they do not take into account the degrees of freedom lost in the estimation of the fixed effects [Swallow and Monahan (1984)] and the parameter inference in linear mixed model is usually done by using residual maximum likelihood (REML) [Anderson and Bancroft (1952); Patterson and Thompson (1971)].

The R statistical package "lme4" is a versatile tool that facilitates the fitting and analysis of linear mixed models, generalized linear models, and nonlinear models. Bates et al. [Bates et al. (2014)] extensively discussed the computational methods implemented in this package and how it enables the determination of parameter estimates in a mixed model [detailed discussion in Appendix A.1].

Gumedze and Dunne (2011) extensively discussed parameter estimation and inference procedures for various components of the linear mixed model. They focused on the joint estimation of fixed and random effects, as well as the parameter estimation methods for variance, such as the maximum likelihood method and the residual maximum likelihood method. The detailed discussions has been provided in Appendix A.2.

## 2.3 Review of GWAS Methods

The previous sections provided a brief description of linear models including the parameter estimation steps of linear mixed model (LMM). In this section, I reviewed and explored the application of various multivariate statistical models such as multiple linear regression, linear mixed models etc. in genetic and other research areas. Specifically, I investigated how statistical joint modelling can effectively aggregate the joint effect of multiple genetic markers and account for hidden factors that might confound genotype-phenotype associations.

### 2.3.1 Evaluation of Contribution of Individual Variables

To portray the genotype-phenotype relationship precisely, it is sometimes essential to assess the impact of individual genetic effects on phenotype variability as this can help to choose which genetic markers are the most relevant and which are not. The selection of significant variants plays a crucial role in building a conditioning model and is also essential for prediction, which are the main focal points of this thesis. In this section, I reviewed some of the methods and journal papers regarding the evaluation of the contribution of individual variables.

Following Shabuz and Garthwaite (2019), there are two most obvious methods of evaluating the relative importance of regressors, which are the beta weight method, which looks at the beta coefficients of variables (once the variables have been variance-standardized), and the zero-order correlation method, which looks at the correlation between individual variables and the response. There are other individual contribution evaluation methods, such as Product measures [Thomas et al. (1998)], Usefulness [Darlington (1968)], Structure Coefficients [Courville and Thompson (2001)], Dominance Analysis [Budescu (1993)], Orthogonal Counterparts [Gibson (1962)], Relative Weight analysis [Johnson (2000)], Shapley value regression [Lipovetsky and Conklin (2015)],

Random forest [Liakhovitski et al. (2010)], etc.

When the regressors are uncorrelated, all these measures yield the same results, and they collectively measure the individual contributions of the regressor variables, which sum to the coefficient of determination ($R^2$). The $R^2$ value indicates the proportion of variation in the dependent variable that can be explained by the regressor variables. This metric plays a crucial role in statistics and genetic studies as it helps quantify the contribution of specific factors or genetic variants to the overall variability in a dataset or phenotypic trait.

In Chapter 4, I have provided mathematical derivations for these metrics, expressing them in terms of correlations and partial correlations between the trait and genetic variants. Sometimes, when we have GWAS summary statistics for variants (SNPs) such as SNP ID, SNP location, allele information, effect size, test statistic, and P-value, we may need to calculate the correlation coefficient between an SNP and the trait of interest. In such cases, the $R^2$ can be calculated from the test statistic, and I have demonstrated this mathematical expression in Appendix A.3.

But if the data contain collinearity among regressors then these methods become meaningless since the high collinearity can inflate the values of regression coefficients in comparison with pair correlations between the regressors and response [Lipovetsky and Conklin (2015)]. In this situation, to choose a good measure of evaluating relative importance, it should be based on concrete logic behind their development, their properties, shortcomings, and the sensibility of the results they produce [Johnson and Lebreton (2016)].

Partitioning the quadratic form into contributions from individual variables, Garthwaite and Koch (2016) argued a way of measuring the contribution of an individual named corr-max transformation that maximizes the sum of the correlations between individual variables and transformed variable. This transformation is closely related to

another transformation named as cos-max transformation proposed by Garthwaite et al. (2012). The only difference is that the cos-max transformation is designed to transform a data matrix while the corr-max transforms a random vector. Recently, Shabuz and Garthwaite (2019) developed three new measures of relative importance and compared them with well-regarded alternatives through examples and by examining theoretical properties. According to them, the new measures are much in common with the orthogonal counterpart measure and the relative weights measure but the main difference is that the new measure uses the values of both the regressor and the response in determining the transformation while the others ignore the response.

### 2.3.2 Correcting for population structure and kinship using the LMM: Theory and Extensions

Population structure and kinship influence the genetic covariance among individuals and are considered confounding factors in genome-wide association studies (GWAS) [Price et al. (2010)]. Typically, genetic studies address this by using genome-wide SNP data to exclude problematic individuals or incorporate these effects in association tests. Principal component analysis (PCA) is a widely used technique for detecting population structure, capturing genetic ancestry, often included as fixed effects in regression models [Price et al. (2006)]. In recent times, linear mixed models (LMMs) have caught attention for modelling the dependence structure of GWAS datasets by considering the genome-wide similarity among all pairs of individuals [Kang et al. (2008); Lippert et al. (2011); Segura et al. (2012)]

Failure to account for hidden genetic relatedness among individuals in the genetic database can lead to misleading association results. This can result in decreased statistical power and an increase in false positive findings [Price et al. (2010)]. Hence, LMMs or the inclusion of principal components as fixed effects in the model are commonly used approaches to address dependence [Price et al. (2010)]. While fixed effect mod-

els involve a few principal components ($i \ll n$), LMMs incorporate the genome-wide similarity among individuals to address population structure [Kang et al. (2008)].

Hoffman (2013) introduced a unified framework connecting fixed vs random effects, showcasing that the effects share the same underlying regression model. The distinction lies in their capacity to handle population structure, inference methods, and the number of included principal components in the model [Kenny et al. (2011); Price et al. (2010); Wu et al. (2011)].

Moreover, Efron (2004) noted that the effective degrees of freedom is equal to the number of parameters in the model, but increasing the number of parameters increases the covariance between the observed and fitted response due to overfitting without an increase in actual explanatory power [Kutner (2005)]. Hoffman (2013) also introduced a summary statistic, effective degrees of freedom based on the Lippert algorithm [Lippert et al. (2011)], to measure the overall model complexity and the influence of each principal component on the fit of the LMM.

In this chapter, in Section (2.5), I have demonstrated a mathematical connection between the linear mixed model and ordinary least square regression models, as suggested by Hoffman (2013). This mathematical link indicates that the estimate of the random effect model can be related to the estimate of the principal component-based ordinary least square regression model. This insight establishes a valuable connection between these two commonly used statistical approaches.

### 2.3.3 GCTA: A Tool for Genome-wide Complex Trait Analysis

Following Yang et al. (2011) genome-wide complex trait analysis (GCTA) is a versatile tool to estimate genetic relationships from genome-wide SNPs. GCTA has developed to perform 5 functions such as Data management, Estimation of the genetic relationships from SNPs, Mixed linear model analysis of variance explained by the SNPs, Estimation of the linkage disequilibrium structure, and GWAS simulation. It estimates variance

explained by all SNPs for a phenotype rather than testing the association for any particular SNP to the phenotypic characteristics. The basic difference with Single SNP based association analysis is that it takes into account all SNPs effects as random effects by fitting a mixed effect linear model (LMM). Estimation of genetic relationships considering the genetic relationship matrix (GRM) between individuals from the genetic structures is one of the important functions of GCTA [Hayes et al. (2009); VanRaden (2008)]. During this estimation process, GCTA implements the restricted maximum likelihood (REML) approach to estimate variance components relying on the GRM estimated from all the SNPs and it provides the best linear unbiased prediction (BLUP) of the genetic effect using REML [Patterson and Thompson (1971)]. BLUP is widely used by plant and animal breeders to quantify the breeding value of individuals in artificial selection programs [Henderson (1975)].

Chapter 4 is a significant part of this thesis, where I have demonstrated that conditioning on major genetic variants can lead to the discovery of new variants with improved statistical power. Through mathematical derivations in multiple regression settings, I have illustrated how the conditional results can be approximated using publicly available summary statistics from GWAS databases, even without having access to individual-level data.

The 'GCTA-COJO' command in the GCTA software played a significant role in my research, as it allowed for multi-SNP-based conditional analysis using GWAS summary statistics. By comparing the conditional results obtained from the GCTA software with those derived from my mathematical developments, I was able to validate the accuracy of the proposed approach.

### 2.3.4 A comparison of principal component regression and genomic REML for genomic prediction across populations

During the analysis of high-dimensional genetic data, one of the major problems is multicollinearity among genetic markers which misleads the least square estimates and another problem is that a large number of regressors ($p$) may be larger than the number of observation ($n$). To deal with this kind of problem, principal component analysis (PCA) is the possible way of getting rid of it which fits the model by taking the principal components instead of the original regressors. In genetic studies, PCA has been mainly used for correcting population structures and stratification during the association studies and capturing the joint effect of genetic variation [McVean (2009); Patterson et al. (2006); Price et al. (2006); Reich et al. (2008)]. The first application of PCA in population genetics was applied by Menozzi et al. (1978) to produce maps of human genetic variation across mainland regions.

Many research studies have used principal components to capture the variability present in the original variable $X$ (SNPs), concentrating primarily on selecting those principal components that maximize variance among the regressors. However, solely incorporating principal components with the highest variance might not guarantee the best prediction in the data. This is because a principal component that explains a small amount of variance in $X$ can still be significant for predicting the response variable.

To address this, some authors suggest selecting principal components not only based on the variance decomposition of covariates but also considering their contribution to the regression sum of squares. Dadousis et al. (2014) discussed various approaches for selecting principal components in the context of PCR modeling. They compared these approaches by evaluating their prediction accuracies in terms of minimum mean squared error (MSE).

One approach involved performing PCA solely on the SNP matrix of the reference

dataset. Principal components were then selected by ranking them according to decreasing eigenvalues and their contribution to the sum of squares of the regression. Another approach conducted PCA on the matrix with all SNP genotypes, including reference and test datasets. Optimal principal components for PCR modelling were chosen using a cross-validation (CV) approach within the reference dataset. These selected components were subsequently employed for predicting the test dataset.

In this thesis, I introduced a novel mathematical approach for computing conditional results in multiple regression settings, with a particular emphasis on the role of genetic principal components (PCs). The design matrix was divided into three distinct blocks: the tested genetic variant block, the conditional variants block, and the covariates block, which might include variables such as age, sex, and genetic PCs.

This approach allows for the calculation of conditional results using GWAS summary statistics, even in situations where individual-level data are not available. Since these summary statistics are typically adjusted for population structures using genetic principal components (PCs), it is essential to apply the same adjustment to the tested and conditioned genetic variants. By incorporating PCs adjustment, the proposed method ensures the accuracy of the conditional results and provides a reliable means of conducting conditional analyses in GWAS.

Furthermore, in one of the applied research works presented in Chapter 6, I extensively utilized various dimension reduction techniques for prediction. The aim was to determine the optimal number of dimensions required to build an effective predictive model. Among the approaches used, principal component analysis (PCA) played a significant role, and the reduced dimensions were selected based on the scree plot analysis. Various predictive models were then employed to evaluate the prediction accuracy, providing valuable insights into the performance of the predictive model based on the reduced dimensions.

## 2.4 Shrinkage Methods

In this section, I have conducted a comprehensive review of various shrinkage techniques. These techniques hold the potential to identify the genetic markers that contribute to human phenotypic variation, thereby leading to enhanced prediction accuracy for high-dimensional genetic data. Usually, the number of genetic markers is huge in number and possibly much larger than the number of observations, so these features can be reduced by regularization [Hastie et al. (2009)]. In this situation, there are some subset selection methods that retain a subset of the regressors and discard the rest, and doing so often exhibits high variance during the construction of the model as well as increases the prediction error. However the shrinkage methods are more continuous and trade-off between bias and variance, these methods reduce the prediction error and provide prediction accuracy during prediction.

In Chapter 6, I also provided detailed mathematical expressions for prediction errors in both the linear model and different shrinkage methods, emphasizing their significance in genetic association studies.

### 2.4.1 Principal Component Regression (PCR)

The singular value decomposition (SVD) of the centered matrix $X$ is another way of expressing the principal components of the variables in $X$. The sample covariance matrix is given by [Hastie et al. (2009)]

$$X^T X = V D^2 V^T$$

which is the eigen-decomposition of $X^T X$. The eigenvectors $v_j$ (columns of $V$) are also called the principal components directions of $X$. The first principal component direction $v_1$ has the property that $z_1 = X v_1$ has the largest sample variance amongst all

normalized linear combinations of the columns of $X$.

When the design matrix contains a large number of correlated inputs, then this method produces a small number of linear combinations $Z_m, m = 1, 2, \cdots, M$ of the original inputs $X_j$, and the $Z_m$ are then used in place of the $X_j$ as inputs in the regression. Principal component regression forms the derived input columns $z_m = Xv_m$, and then regresses $y$ on $z_1, z_2, \cdots, z_M$ for some $M \leq p$. Since the $z_m$ are orthogonal, this regression is just a sum of univariate regressions:

$$\hat{y}^{pcr}_{(M)} = \bar{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m z_m$$

where, $\theta_m = \langle z_m, y\rangle / \langle z_m, z_m\rangle$. Since the $z_m$ are each linear combinations of the original $x_j$ , the above equation can be expressed in terms of coefficients of the $x_j$.

$$\begin{aligned}
\hat{y}^{pcr}_{(M)} &= \bar{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m z_m \\
&= \bar{y}\mathbf{1} + X \sum_{m=1}^{M} \hat{\theta}_m v_m \\
&= \begin{bmatrix} 1 & X \end{bmatrix} \begin{bmatrix} \bar{y} \\ \sum_{m=1}^{M} \hat{\theta}_m v_m \end{bmatrix}
\end{aligned}$$

It can also be written as the matrix $\begin{bmatrix} 1 & X \end{bmatrix}$ times a vector $\hat{\beta}^{pcr}_{(M)}$ if the later vector taken as

$$\hat{\beta}^{pcr}_{(M)} = \begin{bmatrix} \bar{y} \\ \sum_{m=1}^{M} \hat{\theta}_m v_m \end{bmatrix}$$

This is the same equation as $\hat{\beta}^{pcr}_{(M)} = \sum_{m=1}^{M} \hat{\theta}_m v_m$, when restrict to just the last $p$ elements of $\hat{\beta}^{pcr}_{(M)}$. The principal components regression is very similar to ridge regression. Both

of them operate via the principal components of the input matrix. The ridge regression shrinks the coefficients of the principal components, shrinking more depending on the size of the corresponding eigenvalue but the principal components regression discards the $(p - M)$ smallest eigenvalue components.

### 2.4.2 Partial Least Squares (PLS)

The partial least squares (PLS) also constructs a set of linear combinations of the inputs for regression, but unlike principal components regression, it uses $y$ (in addition to $X$) for this construction [Hastie et al. (2009)]. Similar to principal component regression, PLS is not scale-invariant, assuming that each $x_j$ is standardized to have a mean of 0 and a variance of 1.

The PLS algorithm begins by computing $\hat{\varphi}_{1j} = \langle x_j, y \rangle$ for each $j$. From this, the derived input constructs as $z_1 = \sum_j \hat{\varphi}_{1j} x_j$, which represents the first partial least squares direction. In the construction of each $z_m$, the inputs are weighted by the strength of their univariate effect on $y$. The outcome $y$ is regressed on $z_1$, yielding the coefficient $\theta_1$. The input variables $x_1, \cdots, x_p$ are then orthogonalized with respect to $z_1$, and this process continues until $M \leq p$ directions have been obtained.

In this manner, partial least squares produce a sequence of derived, orthogonal inputs or directions $z_1, z_2, \cdots, z_M$. If $M = p$, the solution will be equivalent to the usual least squares estimates, while using $M < p$ directions results in a reduced regression. The general algorithm of partial least square:

1. Standardize each $x_j$ to have mean zero and variance one. Set $\hat{y}^{(0)} = \bar{y}\mathbf{1}$ and $x_j^{(0)} = x_j; j = 1, 2, \cdots, p$

2. For $m = 1, 2, \cdots, p$

   - $z_m = \sum_j^p \hat{\varphi}_{mj} x_j^{(m-1)}$, where, $\hat{\varphi}_{mj} = \langle x_j^{(m-1)}, y \rangle$.

- $\hat{\theta}_m = \langle z_m, y\rangle / \langle z_m, z_m\rangle$.
- $\hat{y}^{(m)} = \hat{y}^{(m-1)} + z_m \hat{\theta}_m$.
- Orthogonalization each $x_j^{(m-1)}$ with respect to $z_m$ as

$$x_j^{(m)} = x_j^{(m-1)} - [\langle z_m, x_j^{(m-1)}\rangle / \langle z_m, z_m\rangle] z_m; j = 1, 2, \cdots, p$$

3. Output the sequence of fitted vectors $[\hat{y}^{(m)}]_1^p$. Since the $[z_l]_1^p$ are linear in the original $x_j$, so is $\hat{y}^{(m)} = X\hat{\beta}_{(m)}^{pls}$. These linear coefficients can be recovered from the sequence of PLS transformations.

**Note:** It can show that in the orthogonal case, PLS stops after $m = 1$ steps, because subsequent $\hat{\varphi}_{mj}$ in step 2 in the above Algorithm are zero [Hastie et al. (2009)].

### 2.4.3 Ridge Regression

Ridge regression reduces overfitting and multicollinearity by applying a penalty on the size of regression coefficients [Hastie et al. (2009)]. The ridge estimator achieves this by minimizing the ridge loss function, which can be expressed as follows:

$$\begin{aligned} L_{ridge}(\beta, \lambda) = RSS(\lambda) &= ||y - X\beta||_2^2 + \lambda ||\beta||^2 \\ &= (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta \end{aligned}$$

Mininizing the above loss function with respect to $\beta$, the ridge regression solutions are easily seen to be

$$\beta^{ridge} = (X^T X + \lambda I_{pp})^{-1} X^T y$$

Note that with the choice of quadratic penalty, the ridge regression solution is a linear function of $y$. The solution adds a positive constant to the diagonal of $X^T X$ before inversion and this makes the problem non-singular, even if $X^T X$ is not of full rank,

and that was the main motivation for ridge regression when it was first introduced in statistics [Hoerl and Kennard (1970)].

When a linear regression model contains correlated variables, multicollinearity can lead to poorly determined coefficients exhibiting high variance. A wildly large positive coefficient on one variable can be canceled by a similarly large negative coefficient on its correlated cousin. By imposing a size constraint on the coefficients, as in the above loss function, this problem is alleviated. As the coefficients are penalized, their bias increases, but the trade-off leads to reduced variance, which can help mitigate the multicollinearity problem.

**Case-I: If the design matrix is Orthogonal**

In the case of orthonormal inputs, the ridge estimates are just scaled versions of the least squares estimates. That is, when $X^TX = I_{pp} = (X^TX)^{-1}$, then

$$\begin{aligned}
\hat{\beta}^{ridge} &= (X^TX + \lambda I_{pp})^{-1}X^Ty \\
&= (I_{pp} + \lambda I_{pp})^{-1}X^Ty \\
&= (1+\lambda)^{-1}I_{pp}X^Ty \\
&= (1+\lambda)^{-1}(X^TX)^{-1}X^Ty \\
&= (1+\lambda)^{-1}\hat{\beta}_{ols}
\end{aligned}$$

Hence, the ridge estimator scales the OLS estimator by a factor $(1+\lambda)^{-1}$.

**Case-II: Bayesian Approach**

Ridge estimator can also be derived as the mean or mode of a posterior distribution, with a suitably chosen prior distribution. Suppose, $y \sim N(X\beta, \sigma^2)$, and the parameter,

$\beta \sim N(0, \tau^2)$. According to Bayes rule,

$$\begin{aligned} p(\beta|y) &\propto p(y|\beta)p(\beta) \\ &= N(X\beta, \sigma^2 I)N(0, \tau^2 I) \end{aligned}$$

So, the log-posterior density of $\beta$, with $\tau^2$ and $\sigma^2$ assumed known can be expressed as

$$logp(\beta|y) = c - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) - \frac{\beta^T\beta}{2\tau^2}$$

The mean and mode of this posterior distribution can be obtained by minimizing the above expression as

$$\hat{\beta} = (X^T X + \frac{\sigma^2}{\tau^2})^{-1} X^T y$$

If $\lambda = \frac{\sigma^2}{\tau^2}$, then the expression is same as the previous one.

**Case-III: Using the Singular Value Decomposition (SVD) of the Design Matrix**

The singular value decomposition (SVD) of the centered input matrix $X$ provides a further understanding of ridge regression. The solutions of the ridge and ordinary least squares (OLS) estimates can be formulated using the singular values and vectors derived from the SVD of the design matrix $X$, showcasing the effect of regularization on the coefficients.

The singular value decomposition (SVD) of the design matrix $X$ is given by

$$X = UDV^T$$

where, $U_{n\times p}$ and $V_{p\times p}$ are orthogonal matrices, with the column of $U$ spanning the column space of $X$, and the column of $V$ spanning the row space, and $D_{p\times p}$ is a diagonal

matrix with diagonal entries $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ are called the singular values of $X$. If one or more values of $d_j = 0$, then $X$ is singular.

Using SVD, the OLS estimates can be written as

$$\begin{aligned}
\hat{y} = X\hat{\beta}_{ols} &= UDV^T[(VD^2V^T)^{-1}VDU^Ty] \\
&= UDV^T(VD^{-2}V^T)VDU^Ty \\
&= UU^Ty \\
&= \sum_{j=1}^{p} u_j(u_j^Ty)
\end{aligned}$$

Note that $U^Ty$ are the corordinates of $y$ with respect to the orthonormal basis $U$. The above expression shows that the ordinary least square (OLS) estimate does not incorporate the regularization term and relies solely on the original design matrix to estimate the coefficients.

In the context of ridge regression, the SVD can be used to express the estimates as follows:

$$\begin{aligned}
\hat{y} = X\hat{\beta}_{ridge} &= X(X^TX + \lambda I)^{-1}X^Ty \\
&= UDV^T[(VD^2V^T + \lambda VV^T)^{-1}VDU^Ty] \\
&= UDV^T[V(D^2 + \lambda I)V^T]^{-1}VDU^Ty \\
&= UDV^TV(D^2 + \lambda I)^{-1}V^TVDU^Ty \\
&= UD(D^2 + \lambda I)^{-1}DU^Ty \\
&= \sum_{j=1}^{p} u_j(\frac{d_j^2}{d_j^2 + \lambda})(u_j^Ty)
\end{aligned}$$

The solution of ridge estimate presented above includes a regularization term $\lambda$ to counter overfitting by shrinking the coefficient estimates toward zero. This regularization term, controlled by the tuning parameter $\lambda$, is added to the diagonal elements of

the SVD-derived matrix. This adjustment in ridge regression helps to address multicollinearity and enhances the stability of the inverse matrix, yielding a more reliable solution compared to OLS. It's noteworthy that when $\lambda$ is set to 0, ridge regression simplifies to the ordinary least squares solution.

**Effective Degrees of Freedom**

The effective degrees of freedom of the ridge regression estimator is defined as

$$\begin{aligned} df(\lambda) = tr(H) &= tr[X(X^T X + \lambda I)^{-1} X^T] \\ &= tr[UD(D^2 + \lambda I)^{-1} DU^T] \\ &= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda} \end{aligned}$$

The effective degrees of freedom of the ridge regression fit is a decreasing function of $\lambda$. In a linear model fit with $p$ variables, the degree of freedom is $p$, representing the number of free parameters. With ridge regression, all $p$ coefficients are non-zero, but they are constrained by $\lambda$. When $\lambda = 0$, $df(\lambda) = p$, and as $\lambda \to \infty$, $df(\lambda) \to 0$.

**Ordinary least squares to implement ridge regression using augmented data set**

The ridge regression estimates can be obtained through ordinary least squares regression on an augmented data set. The augmentation is applied to the centered matrix $X$ with $p$ additional rows $\sqrt{\lambda}I$, and the response vector $y$ is augmented with $p$ zeros. By introducing artificial data with response values of zero, the fitting procedure is compelled to shrink the coefficients toward zero. This concept is related to the idea of hints due by [Abu-Mostafa (1995)], where model constraints are implemented by adding artificial data examples that satisfy them.

Consider the input centered data matrix $X_{p\times p}$ and the output data vector $Y$ both

appended to produce the new variable $\hat{X}$ and $\hat{Y}$ as follows

$$\hat{X} = \begin{bmatrix} X \\ \sqrt{\lambda} I_{p\times p} \end{bmatrix}$$

and

$$\hat{Y} = \begin{bmatrix} Y \\ 0_{p\times 1} \end{bmatrix}$$

with $I_{p\times p}$ and $0_{p\times 1}$ identity and zero column respectively. The least squares solution to this new problem is given by

$$\hat{\beta}_{ols} = (\hat{X}^T \hat{X})^{-1} \hat{X}^T \hat{Y}$$

Performing the block matrix multiplications required by the above expressio, it can be calculated that

$$\hat{X}^T \hat{X} = \begin{bmatrix} X^T & \sqrt{\lambda} I_{p\times p} \end{bmatrix} \begin{bmatrix} X \\ \sqrt{\lambda} I_{p\times p} \end{bmatrix} = X^T X + \lambda I_{p\times p}$$

and

$$\hat{X}^T Y = \begin{bmatrix} X^T & \sqrt{\lambda} I_{p\times p} \end{bmatrix} \begin{bmatrix} Y \\ 0_{p\times 1} \end{bmatrix} = X^T Y$$

Thus it can be written that

$$\hat{\beta}_{ols} = (X^T X + \lambda I_{p\times p})^{-1} X^T Y$$

which is the solution of the regularized least squrare estimate i.e, ridge regression estimate.

### 2.4.4 LASSO Regression

The least absolute shrinkage and selection operator (LASSO) is a shrinkage method like ridge regression with important differences in imposing the constraint [Hastie et al. (2009)]. The LASSO estimate is defined as

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \Big( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \Big)^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t,$$

The above constraint makes the solution nonlinear in $y$ and there is no closed-form expression as in ridge regression. Because of the nature of the constraint, making $t$ sufficiently small will cause some of the coefficients to be exactly zero and thus the LASSO does a kind of continuous subset selection. If $t$ is chosen larger than $t_0 = \sum_{j=1}^{p} |\hat{\beta}_j|$ (where $\hat{\beta}_j$ is the least square estimates), then the LASSO estimates are the $\hat{\beta}_j' s$. But if it is taken as $t = t_0/2, say$, then the least squares coefficients are shrunk by about 50% on average. To get the best subset variable in subset selection, the penalty parameter ($t$) should be chosen to minimize an estimate of the expected prediction error.

Zou and Hastie (2005) proposed an alternative penalty term, known as elastic net, which compromises between ridge and lasso regression. The elastic net penalty has the following form

$$\sum_{j=1}^{p} (\alpha|\beta_j| + (1-\alpha)\beta_j^2)$$

The elastic net penalty combines two important components to achieve its purpose. The first term encourages sparsity in the coefficients of correlated features, promoting a more compact and interpretable model. The second term, on the other hand, encourages correlated features to be averaged together, promoting stability and reducing multicollinearity issues. As a result, the elastic net penalty can be effectively applied in various linear models, including regression and classification tasks.

### 2.4.5 Eigenvalue Shrinkage

Hastie et al. (2009) showed that the singular value decomposition of the $n \times p$ dimensional design matrix, $X$, can be written as

$$X = U_x D_x V_x^T$$

where, $D_x$ is a $n \times n$ dimensional diagonal matrix with diagonal entries $d_1 \geq d_2 \geq \cdots \geq d_n \geq 0$ is called the singular values of $X$, $U_x$ is a $n \times n$ dimensional matrix with columns containing the left singular vector, and $V_x$ is $p \times n$ dimensional matrix with columns containing the right singular vectors. The coloumns of $U_x$ and $V_x$ are orthogonal that is $V_x^T V_x = U_x^T U_x = I_{nn}$. The OLS estimator can be written in terms of the SVD matrices as

$$\begin{aligned}
\hat{\beta}_{ols} &= (X^T X)^{-1} X^T y \\
&= (V_x D_x^2 V_x^T)^{-1} V_x D_x U_x^T y \\
&= V_x D_x^{-2} D_x U_x^T y
\end{aligned}$$

The ridge estimator can be written as

$$\begin{aligned}
\hat{\beta}_{ridge} &= (X^T X + \lambda I_{pp})^{-1} X^T y \\
&= (V_x D_x^2 V_x^T + \lambda V_x V_x^T)^{-1} V_x D_x U_x^T y \\
&= V_x (D_x^2 + \lambda I_{nn})^{-1} V_x^T V_x D_x U_x^T y \\
&= V_x (D_x^2 + \lambda I_{nn})^{-1} D_x U_x^T y
\end{aligned}$$

Combining the OLS and ridge results, it can be compared that

$$\begin{aligned} & d_{x,jj}^{-2} \geq (d_{x,jj}^{-2} + \lambda)^{-1} \\ & \Rightarrow d_{x,jj}^{-1} \geq d_{x,jj}(d_{x,jj}^{2} + \lambda)^{-1} \qquad ; \forall \lambda > 0 \end{aligned}$$

Thus, the ridge penalty shrinks the singular values and the fitted value can be expressed as

$$\begin{aligned} \hat{y}_{ridge} &= X\hat{\beta}_{ridge} \\ &= U_x D_x V_x^T V_x (D_x^2 + \lambda I_{nn})^{-1} D_x U_x^T y \\ &= U_x D_x (D_x^2 + \lambda I_{nn})^{-1} D_x U_x^T y \\ &= \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^2 y \end{aligned}$$

In principal component regression, the least square estimate of $\theta$ is

$$\begin{aligned} \hat{\theta} &= (z_m^T z_m)^{-1} z_m^T y \\ &= (V_m^T X^T X V_m)^{-1} V_m^T X^T y \\ &= (V_m^T V_x D_x U_x^T U_x D_x V_x V_m)^{-1} V_m^T V_x D_x U_x^T y \\ &= (I_{mn} D_x^2 I_{nm})^{-1} I_{mn} D_x U_x^T y \end{aligned}$$

It is observed that the PCR threshholds the singular values of $X$, but the ridge regression shrinks them depending on their size. Now, the PCR estimator of $\beta$ is

$$\hat{\beta}_{pcr} = V_x \hat{\theta} = V_x (I_{mn} D_x^2 I_{nm})^{-1} D_x U_x^T y$$

and the fitted value can be expressed as

$$\begin{aligned}\hat{y}_{pcr} &= X\hat{\beta}_{pcr} \\ &= U_x D_x V_x^T V_x (I_{mn} D_x^2 I_{nm})^{-1} D_x U_x^T y \\ &= U_x D_x (I_{mn} D_x^2 I_{nm})^{-1} D_x U_x^T y \\ &= \sum_{j=1}^{m} u_j \frac{d_j^2}{I_{mn} d_j^2 I_{nm}} u_j^2 y\end{aligned}$$

Here, if $m = p$, then the fitted value equal to the fitted value obtained by OLS.

The random effect estimator can be expressed as

$$\begin{aligned}\hat{u}_{re} &= (Z^T Z + \delta)^{-1} Z^T y \qquad [\delta = \frac{\sigma_u^2}{\sigma_\epsilon^2}] \\ &= (V_z D_z U_z^T U_z D_z V_z^T + \delta)^{-1} V_z D_z U_z^T y \\ &= V_z (D_z^2 + \delta)^{-1} V_z^T V_z D_z U_z^T y \\ &= V_z (D_z^2 + \delta)^{-1} D_z U_z^T y\end{aligned}$$

and the fitted value can be expressed as

$$\begin{aligned}\hat{y}_{re} &= X\hat{\beta}_{re} \\ &= U_x D_x V_x^T V_x (D_x^2 + \delta)^{-1} D_x U_x^T y \\ &= U_x D_x (D_x^2 + \delta)^{-1} D_x U_x^T y \\ &= \sum_{j=1}^{q} u_j \frac{d_j^2}{d_j^2 + \delta} u_j^2 y\end{aligned}$$

The estimate from the random effect model resembles the ridge estimate and can be viewed as a specific instance within ridge regression. When the value of $\delta = \frac{\sigma_u^2}{\sigma_\epsilon^2}$ matches $\lambda$, i.e., the selected penalty parameter in ridge regression equals to the ratio of the variance components in the corresponding LMM, they become mathematically the

same [Shen et al. (2013)].

## 2.5 Linking Shrinkage Methods' with Linear Mixed Models

### 2.5.1 Linear Mixed Models link with OLS Regression [Own Work based on Literature Review]

In section (2.3.2), I reviewed the research work of Hoffman (2013), where he explored the connection between modelling principal component effects as fixed versus random. He highlighted that these effects share the same underlying regression model. In this context, he considered the genotype data matrix as $X_{n\times p}$, representing $n$ individuals and $p$ genetic markers, with each entry indicating the number of minor allele copies. The singular value decomposition of the design matrix underlying principal component analysis is given by

$$X = USV^T$$

where the first $i$ principal components are the initial $i$ columns of $U_{n\times n}$, $S_{n\times n}$ is a diagonal matrix containing singular values for each principal component, and $V_{p\times n}$ represents loadings on each marker. Incorporating the first $i$ principal components as fixed effects in a linear model can be expressed as:

$$y = \mu + x_j\beta + U_{1:i}\omega + \epsilon \text{ with } \epsilon \sim N(0, \sigma_e^2 I) \tag{2.7}$$

where $y_{n\times 1}$ is a vector of phenotype values, $\mu$ is the scalar mean term, $x_j$ is the $j^{th}$ marker with scalar regression coefficient $\beta$, $U_{1:i}$ are the first $i$ principal components with coefficient vector $\omega_{i\times 1}$, and $\epsilon$ is the normally distributed residual error term with variance $\sigma_e^2$.

Now, a linear mixed model (LMM) can be expressed as

$$y = \mu + x_j\beta + \alpha + \epsilon$$

where, $\alpha_{n\times 1}$ is a random effect vector with a multivariate Gaussian prior follows a distribution as $\alpha \sim N(0, K\sigma_a^2)$ , $K_{n\times n}$ is the genetic similarities matrix between all pairs of individuals so that $\mathbf{K}_{k,l}$ represents the similarity between individuals $k$ and $l$, $\sigma_a^2$ is the additive genetic variance and $\epsilon \sim N(0, \sigma_e^2 I)$. Here, the population structure is treated as a random effect, and fitting the model involves integrating over the vector $\alpha$ with respect to the Gaussian prior so that the likelihood is maximized w.r.to $\sigma_a^2, \sigma_e^2, \mu, \beta$.

As per Patterson et al. (2006), the genetic similarity matrix $K$ can be regarded as a function of observed genotypes and factorized through the singular value decomposition as:

$$K = XX^T = USV^TVSU^T = (US)(US)^T = RR^T$$

here the columns of $R_{n\times n}$ represent the principal components weighted by their respective singular values. It is important to note that each principal component $U_t$ has a singular value $s_t$ and an eigenvalue $s_t^2$. By utilizing the property of a multivariate Gaussian distribution, where $\phi \sim N(m, \Sigma)$, it follows that $B\phi \sim N(Bm, B\Sigma B^T)$. Following the aforementioned decomposition, it can be deduced that $\gamma \sim N(0, \sigma_a^2)$, $R\gamma \sim N(0, K\sigma_a^2)$, and the Linear Mixed Model (LMM) can be equivalently reformulated as:

$$y = \mu + x_j\beta + R\gamma + \epsilon \tag{2.8}$$

here, $\gamma \sim N(0, \sigma_a^2)$ and $\epsilon \sim N(0, \sigma_e^2 I)$. Examining the relationship between equations (2.7) and (2.8), it becomes evident that modelling principal components as fixed or random effects share the same underlying regression model. Notably, even though equation (2.8) represents a Linear Mixed Model (LMM), the parameters to be estimated, $\hat{\sigma}_a^2$ and

$\hat{\sigma}^2$, as well as $\hat{\gamma_1}, \hat{\gamma_2}, \cdots, \hat{\gamma_k}, \hat{\sigma}^2$, can be obtained using ordinary least squares (OLS) estimates, showcasing a connection between them.

Subsequently, I calculated the estimate for the random effect $\gamma$ using the estimating equation outlined by Robinson (1991), as shown below

$$\begin{pmatrix} \gamma \\ y - R\gamma \end{pmatrix}^T \begin{pmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_e^2 \end{pmatrix}^{-1} \begin{pmatrix} \gamma \\ y - R\gamma \end{pmatrix} = \gamma^T \sigma_a^{-2} \gamma + (y - R\gamma)^T \sigma_e^{-2} (y - R\gamma)$$

By maximizing the above function with respect to $\gamma$ and setting the derivative to zero, I obtained the estimate for $\gamma$ as follows:

$$\begin{aligned} & 2\sigma_a^{-2}\gamma^T - 2R^T\sigma_e^{-2}y + 2\gamma^T R^T \sigma_e^{-2} R = 0 \\ & \implies [\sigma_a^{-2} + K\sigma_e^{-2}]\hat{\gamma} = R^T \sigma_e^{-2} y \\ & \implies [\frac{\sigma_a^{-2}}{\sigma_e^{-2}} + K]\hat{\gamma} = R^T y \\ & So, \hat{\gamma} = [K + \delta]^{-1} R^T y = [K + \delta]^{-1} K^{1/2} y \text{ [letting, } \delta = \frac{\sigma_a^{-2}}{\sigma_e^{-2}}] \end{aligned}$$

The estimated fitted response values based on the random effect are

$$\hat{y} = R\hat{\gamma} = [K + \delta]^{-1} K y = H y$$

where $\delta = \frac{\sigma_e^2}{\sigma_a^2}$ and $H$ is the projection matrix for the random effect model.

To validate the assertion by Hoffman (2013) that modeling principal components as either fixed or random effects shares the same underlying regression model, I examined a linear model utilizing the principal component as

$$Y = Wu + \epsilon$$

here, the singular decomposition of the genetic similarities matrix ($K$) can be written as $K = W^TW = R^TR$ and the ordinary least square estimate of this model is $\hat{\beta} = (W^TW)^{-1}W^Ty$.

From the estimates obtained using the two modeling approaches described above, I demonstrated a mathematical pathway to calculate the estimate of the random effect. This pathway involves relating the estimate of the random effect to the estimate of the principal component-based linear model, as follows:

$$\begin{aligned}\hat{\gamma} &= [K+\delta]^{-1}R^Ty = [K+\delta]^{-1}W^Ty = [K+\delta]^{-1}KK^{-1}W^Ty \\ &= [K+\delta]^{-1}K(W^TW)^{-1}W^Ty \text{ [here } K = W^TW] = H\hat{\beta} \\ &= \text{Projection Matrix of Random Effect Model} \times \text{OLS estimate based on PCA}\end{aligned}$$

The above expression indicates that the random effect solution can be obtained as multiplying the projection matrix of the random effect model by the principal component-based linear model solution.

### 2.5.2 Linear Mixed Models link with Ridge Regression

Consider a random effect model which is defined as

$$y_{n\times 1} = Z_{n\times q}u_{q\times 1} + \epsilon_{n\times 1}$$

where, $y_{n\times 1}$ is a vector of responses, $Z_{n\times q}$ is a design matrix for the random effects, $u_{q\times 1} \sim N(0, \sigma_u^2)$ is a vector of random effect and $\epsilon \sim N(0, \sigma_\epsilon^2)$. Considering $u$ as known, the conditional distribution of $y|u$ can be written as $y|u \sim N(Zu, \sigma_\epsilon^2)$. Following Bates et al. (2014), the estimate of $u$, can be obtained through methods like penalized likelihood where the penalty term is added to the log-likelihood function as follows

$$\hat{u} = \underset{u \in R^q}{\text{agrmin}} \, ||y - Zu||_2^2 + u^T u \sigma_u^2$$
$$= \underbrace{(y - Zu)^T \sigma_\epsilon^2 (y - Zu)}_{\text{Residual Sum of Square}} + \underbrace{u^T u \sigma_u^2}_{\text{Penalty term}}$$

Now, minimizing the above loss function with respect to $u$ and the estimate of the random effect can be obtained as

$$\hat{u} = [Z^T Z + \delta]^{-1} Z^T y \text{ [letting, } \delta = \frac{\sigma_u^2}{\sigma_\epsilon^2}]$$

It is observed that if the term $\delta$ is ignored, the predictor reduces to a least square estimator. But with the term, $\delta$, the predictor is actually of the shrinkage type as is the ridge estimator. In LMM, the shrinkage estimator, represented by $\delta$, indicates the variance component estimated using the REML with $\delta = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_\epsilon^2}$, whereas for the ridge regression $\delta$ is computed using the generalized cross-validation (GCV) function $GCV(\delta) = \frac{\epsilon^T \epsilon}{(n - df_e)}$, where, $df_e$ is the effective degrees of freedom [Hastie et al. (2009)].

### 2.5.3 Linear Mixed Models link with LASSO (LMM-LASSO)

The analysis of the genetic data can be misleaded and can produce false positive findings with the presence of some hidden factors such as population structure. One of the major sources of these hidden effects can be understood as deviation from the idealized assumption that the samples in the study population are unrelated. So it is really difficult to ignore the population structure in the sample and even in the stratified sample, the extent of hidden structure can not be ignored [Newman et al. (2001)]. There are some modelling approach that accounts for the presence of such structure and has been shown it greatly reduce the impact of this confounding source of variability. For example, EIGENSTRAT builds on the idea of extracting the major axes of population

differentiation using a principal component analysis decomposing of the genotype data [Price et al. (2006)], and subsequently including them into the model as additional covariates. Another statistical approach that controls the effect of such population structures through correcting of family structure and cryptic relatedness is the linear mixed model [Kang et al. (2010, 2008); Lippert et al. (2011); Yu et al. (2006); Zhang et al. (2010)].

There are a few existing approaches that are capable of addressing both the correction of population structure and the joint mapping of multiple effects jointly. In line with EIGENSTRAT, Hoggart et al. (2008) and Li et al. (2011) add principal components to the model to correct for population structure. But Rakitsch et al. (2013) proposed the LMM-LASSO modelling approach considering the effect of multi-marker genetic effects by summing the individual effects as well as taking them as random confounding factors. There are also some works that considered joint modelling of multi-marker genetic effects in a mixed model using a stepwise regression or forward selection approach [Segura et al. (2012); Yang et al. (2012)]. However, the inclusion of the SNPs in the model following an order is also a problem. As an alternative, the LMM-LASSO approach proposed by Rakitsch et al. (2013) carries out joint inference in the model and assesses all SNPs at the same time while accounting for their interdependencies and without making any assumptions on their ordering. To allow for application to genome-wide SNP data, a Laplacian shrinkage prior has been placed over the fixed effects, assigning a zero-effect size to the majority of SNPs as done in the classical LASSO [Tibshirani (1996)]. In pure LASSO, it is not clear which markers reflect merely the hidden confounders but LMM-LASSO explains confounding explicitly as random effects and these help to resolve the ambiguity between individual genetic effects and phenotype variability because of population structure.

## 2.6 Statistical View of Different Regression Methods

There are some statistical techniques such as principal component regression, partial least square regression, canonical correlation analysis, etc. that have been used in different fields of research for the reduction of high dimensional data. During the construction of a set of linear combinations, the principal component analysis considers only regressors while the partial least square regression considers the response variable in addition to the regressors. The canonical correlation analysis (CCA) is also an important technique to identify the strength of association between two sets of variables [Johnson and Wichern (2007)]. It focuses on the correlation between a linear combination of the variables in one set and a linear combination of the variables in another set. Each set of variables gets reduced to a single variable which is known as a canonical variable and finds its correlation which is known as canonical correlation. In addition to that the multiple correlation is a special case of canonical correlation as it can be obtained by considering one set of observations equal to one. The construction of the PLS algorithm has been discussed by different authors in the literature some of which are Garthwaite (1994); Hastie et al. (2009); Helland (1988) and Rosipal and Krämer (2006) e.t.c.

Frank and Friedman (1993) discussed two common methods (partial least square and principal component regression) used in chemometrics for predictive modelling and compared them with other statistical methods (ordinary least squares, variable selection, and ridge regression) to understand their apparent success as they all attempting to achieve the same operational goal but in slightly in different ways and in what situation they can be expected to work better.

Consider a regression model that finds the predictive relationship among a set of $q$ response variables ($y$) on a set of $p$ predictor variables ($x$) given a set of $N$ observations

and the structural form of this predictive relationship can be taken to be linear as

$$y_j = a_j^T x;\ j = 1, 2, \cdots, q \tag{2.9}$$

The coefficients $a_j$ are estimated using the training data. This model serves both as descriptive statistics for interpreting the data and as a prediction rule to estimate response variable values when only predictor variable values are available.

The purpose is to show how the coefficients vector $\mathbf{a}$ in the equation (2.9) shrink away from the Ordinary Least Squares (OLS) estimates. This movement is toward directions in the predictor variable space where the samples contain a larger spread, or away from the directions where sample predictor variables exhibit minimal spread, i.e.,

$$var(a^T x/|a|) = ave(a^T x/|a|)^2 = \text{small}$$

where the average is over the training sample. The comparisons among different regression methods consist of regarding the regression procedures as a two-step process as in variable subset selection Stone and Brooks (1990); first a $K$ dimensional subspace of the regression is performed under the restriction that the coefficient vector $\mathbf{a}$ lies in that subspace

$$\mathbf{a} = \sum_{k=1}^{K} a_k c_k$$

where the unit vectors $[c_k]_1^K$ span the prescribed subspace with $c_k^T c_k = 1$. The regression procedures can be compared by the way in which they define the subspace $[c_k]_1^K$ and the manner in which the constraint in regression is performed.

According to Frank and Friedman (1993), the comparison among ridge regression (RR), principal component regression (PCR), and partial least square (PLS) has been

made based on criteria

$$
\begin{aligned}
c_{OLS} &= \underset{c^T c=1}{\operatorname{argmax}}\, corr^2(y, c^T x) \\
c_{RR} &= \underset{c^T c=1}{\operatorname{argmax}}\, corr^2(y, c^T x)\frac{var(c^T x)}{var(c^T x)+\lambda} \\
c_k(PCR) &= \underset{[c^T V c_l]_1^{k-1} and\ c^T c=1}{\operatorname{argmax}} var(c^T x) \\
c_k(PLS) &= \underset{[c^T V c_l]_1^{k-1} and\ c^T c=1}{\operatorname{argmax}} corr^2(y, c^T x) var(c^T x)
\end{aligned}
$$

The above criteria indicate that RR, PCR, and PLS are applying a penalty to the OLS criterion, where the penalty increases as var($c^T x$) decreases. Then the question is under what circumstances should this lead to improved performance over OLS? According to James and Stein (1992) that OLS is inadmissible in that one can always achieve a lower mean squared estimation error with biased estimates. Again the question arises when can these estimators substantially improve performance and which one can do it best.

Considering a highly idealized situation, these questions can be answered, and in reality

$$y = \alpha^T x + \epsilon$$

for some coefficient vector $\alpha$ and considereing $\epsilon$ is an additive (i.i.d) homoscedastic error, with zero expectation and variance $\sigma^2$. Considering the coordinate system in which the predictor variables are uncorrelated that is $V = diag(e_1^2, e_2^2, \cdots, e_p^2)$.

Let $a$ be an estimate of $\alpha$ that is $\hat{y}(x) = a^T x$ for a given point $x$ in the predictor space and considering training sample predictor covariance matrix $V$ has the eigenvalues. The mean squared error (MSE) of prediction at $x$ is

$$MSE(\hat{y}(x)) = E_\epsilon[\alpha^T x - a^T x]^2$$

Since $\alpha$ (the truth) is unknown, the MSE (at $x$) for any particular estimator is also unknown.

One can, however, consider various (prior) probability distributions $\pi(\alpha)$ on $\alpha$ and compare the properties of different estimators when the relative probabilities of encountering situations for which a particular $\alpha$ occurs is given by that distribution. For a given $\pi(\alpha)$, the mean squared prediction error averaged over the situations it represents is

$$MSE(\hat{y}(x)) = E_\alpha E_\epsilon [\alpha^T x - a^T x]^2$$

Considering all coefficient vector direction $\frac{\alpha}{|\alpha|}$ equally likely that is the prior distribution depends only in its norm $|\alpha|^2 = \alpha^T \alpha$ that is $\pi(\alpha) = \pi(\alpha^T \alpha)$.

Considering the simple linear shrinkage estimates of the form $a_j = f_j \hat{\alpha}_j; j = 1, 2, \cdots, p$ where, $\hat{\alpha}_j$ is the OLS estimate and in this case, the MSE becomes

$$MSE(\hat{y}(x)) = E_\alpha E_\epsilon \Big[ \sum_{j=1}^{p} (\alpha_j - f_j \hat{\alpha}_j) x_j \Big]^2$$

Averaging over $\alpha$ using the probability distribution given by $\pi\alpha$ yields

$$MSE(\hat{y}(x)) = \sum_{j=1}^{p} \Big[ (1 - f_j)^2 E_\alpha |\alpha|^2 / p + f_j^2 \sigma^2 / (N e_j^2) \Big] x_j^2$$

here, $E_\alpha |\alpha|^2$ is the expected value of the length of the coefficient vector $\alpha$ under the prior $\pi(\alpha)$, $p$ is the number of predictor variables, $\sigma^2$ is the error variance. The two terms within the bracket in the above equation contribute to the MSE at $x$ having separate interpretations. The first term represents the bias of the estimate and the second is the variance. estimate. Setting $[f_j = 1]_1^p$ yields the least squares estimates, which are unbiased but have variance given by the second term in the above equation. Reducing any of the $[f_j]_1^p$ to a value less than 1 causes an increase in bias but decreases the variance. This usual bias-variance trade-off is encountered in nearly all estimation settings. It is

also noticeable that taking any value greater than one for this shrinkage factor causes an increase in both bias squared and variance. This above equation for the MSE illustrates the important fact that justifies the qualitative behavior of RR, PCR, and PLS discussed previously, namely, the shrinking of the solution coefficient vector away from directions of low variance in the predictor-variable space. One sees from the second term in the equation that the contribution to the variance of the model estimate from a given (eigen) direction ($x_j$) is inversely proportional to the sample predictor variance $e_j^2$ associated with that direction. Directions with a small spread in the predictor variables give rise to high variance in the model estimate. The value of $[f_j]_1^p$ that minimizes the MSE are

$$f_j^* = \frac{e_j^2}{e_j^2 + \lambda}; j = 1, 2, \cdots, p$$

with

$$\lambda = p(\sigma^2 / E_\alpha |\alpha|^2)/N$$

The quantity of $\lambda$ is the number of variables times the square of the noise-to-signal ratio, divided by the training sample size. The optimal linear shrinkage esimates can be obtained as

$$a_j = \hat{\alpha}_j \frac{e_j^2}{e_j^2 + \lambda}; j = 1, 2, \cdots, p$$

One sees that the unbiased OLS estimates $[\hat{\alpha}_j]_1^p$ are differentially shrunk with the relative amount of shrinkage increasing with decreasing predictor variable spread $e_j$. The amount of differential shrinkage is controlled by the quantity $\lambda$. The larger the value of $\lambda$, the more differential shrinkage, as well as more overall global shrinkage. The value of $\lambda$ in turn is given by the inverse product of the signal/ noise squared and the training-sample size. It is important to note that this high relative shrinkage in directions of small spread in the (sample) predictor-design distribution enters only to control the variance and not because of any prior belief that the true coefficient vector $\alpha$ is likely to align with the

high spread directions of predictor design. The prior distribution on $\alpha$, $\pi(\alpha)$, that leads to this result of $a_j$ equal mass on all directions $\alpha/|\alpha|$.

Therefore, one can at least qualitatively conclude that the common property of RR, PCR, and PLS of shrinking their solutions away from low spread directions mainly serves to reduce the variance of their estimates, and this is what gives them generally superior performance to OLS. The above results indicate that their degree of improvement (over OLS) will increase with decreasing signal-to-noise ratio and training-sample size and increasing collinearity as reflected by the disparity in the eigenvalues of the predictor variable covariance matrix.

It is well known that $a_j$ is just RR as expressed in the coordinate system defined by the eigenvectors of the sample predictor variable covariance matrix. Thus these results show (again well known) that RR is a linear shrinkage estimator that is optimal (in the sense of MSE) among all linear shrinkage estimators for the prior $\pi(\alpha)$ assumed here and the expression $\lambda$ known. PCR is also a linear shrinkage estimator

$$a_j(PCR) = \hat{\alpha}_j . I(e_j^2 - e_k^2)$$

where $k$ is the number of components used and the second factor $I(.)$ takes the value 1 for nonnegative argument values and 0 otherwise. Thus RR dominates PCR for an equidirection prior. PLS is not a linear shrinkage estimator, so RR cannot be shown to dominate PLS through this argument.

In Chapter 5, I presented an expression for the prediction error under various shrinkage factors, highlighting the trade-off between bias and variance based on this factor. Following the discussion on the shrinkage factor by Frank and Friedman (1993), I followed the connections between the qualitative behavior of ridge regression (RR), principal component regression (PCR), and partial least square (PLS) regression.

# 3 Multivariate Test Procedures (Mostly Literature Review with a little Novel Contribution)

In real-life genetic studies, multiple characteristics can share the same genetic signal, e.g. several facial measurements can be influenced by the same gene. However, the effect of the genes on any single measurement can be small, as is usually the case with most genes. In such a scenario, doing a multivariate test that jointly tests all characteristics at the same time through a multiple regression model can provide more power. Such an approach was implemented in Adhikari et al. (2016a), specifically as a Wald test.

This Chapter focuses on the review and exploration of the mathematical properties of different multivariate test procedures, especially in the context of multiple linear regression. After carefully considering their theoretical aspects and availability in the literature, I adapted various multivariate test procedures for canonical correlation in multiple regression settings. Notably, I demonstrated that these procedures asymptotically follow the chi-square distribution and, more importantly, exhibit asymptotic equivalence among themselves and with the Wald test statistic. These findings provide valuable insights into the statistical power and efficiency of these test procedures for genetic association studies.

## 3.1 Review of Large Sample Test Procedures in Multiple Regression

In multiple regression, the usual interest is to test the hypothesis of linear restrictions on $\beta$. To test this linear restriction, some of the large sample test procedures are Wald Test (W), Likelihood Ratio Test (LRT), and Lagrange Multiplier Test (LM). The key points of the three large sample testing approach are the Wald test starts at the alternative

and consider movements toward the null hypothesis, the LM approach starts at the null hypothesis and ask whether movement toward the alternative would be an improvement and the LR test compares the two hypotheses directly on an equal basis.

In a multiple regression setup, these tests can be applied to challenge the hypothesis that certain $\beta$ coefficients are equal to zero. When there's insufficient evidence in the data to reject this hypothesis, those parameters are set to zero and treated as constants. Chandra and Joshi (1983) noted that the mentioned tests assume the same null hypothesis and are asymptotically equivalent, but their test statistics construction varies significantly.

Quaglio et al. (2020) showed a graphical representation of these test statistics for testing the hypothesis in model identification frameworks based on maximum likelihood inference and described these test statistics based on the literature.

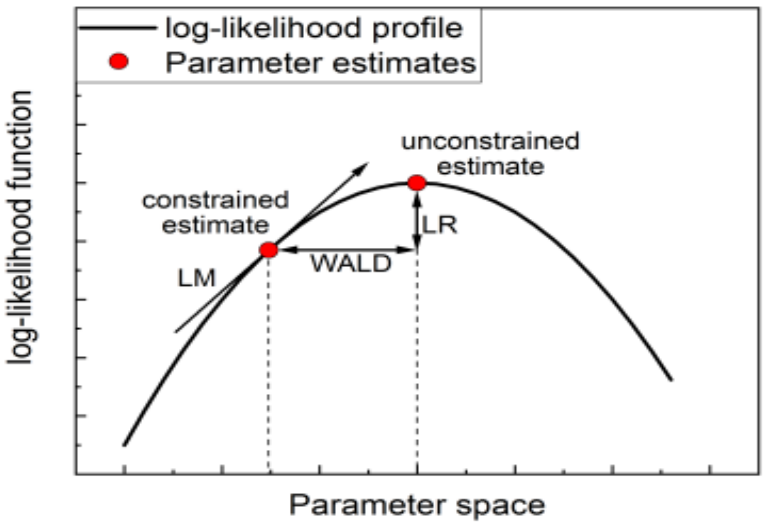



**Figure: A Visual Comparison of Likelihood Ratio (LR), Wald, and Lagrange Multiplier (LM) Statistics**

The Wald statistic quantifies the difference between the unrestricted and restricted maximum likelihood estimates in the parameter space [Wald (1943)]. The Lagrange multipliers statistic is computed from the log-likelihood gradient at the restricted esti-

mate [Rao (1948); Silvey (1959)]. The likelihood ratio test statistic measures the discrepancy between the restricted and unrestricted estimates using log-likelihood values [Wilks (1938)].

Depending on the specific scenario, one of the statistics may offer greater computational convenience compared to the others. A unified development of the three principles has been achieved by establishing the properties of the tests and interconnecting their relationships [Engle (1984)]. Vandaele (1981) indicated that Wald, Lagrange, and Likelihood Ratio tests can be derived as functions of the small sample $F$ test in the context of a multiple regression model. This derivation allows the establishment of well-known sample inequalities among these tests. Following the work by Berndt and Savin (1977), the subsequent relationship among the large sample tests has been deduced as

$$W \geq LR \geq LM$$

.

### 3.1.1 Likelihood Ratio Test (LRT)

The Likelihood Ratio Test (LRT) statistic can be obtained by taking the ratio of likelihoods with and without restrictions of the null hypothesis. Considering the multiple regression model in the equation (2.2), The Likelihood Ratio statistic for testing the null hypothesis, $H_0 : \beta = 0$ is defined as $LR = -2log\lambda$, where $\lambda$ can be obtained as

$$\lambda = \frac{\max_{\beta=0,\sigma^2} L(\beta, \sigma^2)}{\max_{\beta,\sigma^2} L(\beta, \sigma^2)} = \left[\frac{\hat{\sigma}_u^2}{\tilde{\sigma}_R^2}\right]^{n/2} \tag{3.1}$$

here,

$\hat{u} = y - \hat{y} = y - X\hat{\beta}$ and $\tilde{u} = y - X\tilde{\beta} = y$ (under the restriction)

$\hat{\sigma}_u^2 = \dfrac{\hat{u}^T\hat{u}}{n}$ = maximum likelihood estimate of the error variance in the unrestricted model

$\tilde{\sigma}_R^2 = \dfrac{\tilde{u}^T\tilde{u}}{n}$ = maximum likelihood estimate of the error variance in the restricted model

$\hat{\beta} = (X^TX)^{-1}X^Ty$ and $\tilde{\beta}$ can be expressed as a function of the unrestricted estimator as

$\tilde{\beta} = \hat{\beta} - (X^TX)^{-1}(X^TX)\hat{\beta} = 0$

$\tilde{u} = \hat{u} + X(X^TX)^{-1}(X^TX)\hat{\beta} = y$

which allows to write the restricted sum of squares as

$$\tilde{u}^T\tilde{u} = \hat{u}^T\hat{u} + \hat{\beta}^T(X^TX)\hat{\beta} \tag{3.2}$$

Using the equation (3.2), the likelihood ratio $\lambda$ can be expressed as

$$\lambda = \left[\frac{1}{1 + \frac{\hat{\beta}^T(X^TX)\hat{\beta}}{n\hat{\sigma}_u^2}}\right]^{n/2} \tag{3.3}$$

### 3.1.2 Wald and *F*-Test

The Likelihood Ratio Test (LRT) can be written in terms of Wald statistic as

$$\lambda = \left[\frac{1}{1 + \frac{W}{n}}\right]^{n/2} \tag{3.4}$$

where $W$ indicates the Wald test statistic which is constructed from the unrestricted estimates of the parameters and their estimated $\hat{\sigma}_u^2$ and defined as can be defined as

$$\begin{aligned} W =& \frac{\hat{y}^T\hat{y}}{\hat{\sigma}_u^2} \\ =& \frac{\hat{\beta}^T(X^TX)\hat{\beta}}{\hat{\sigma}_u^2} \\ =& \hat{\beta}^T[\hat{\sigma}_u^2(X^TX)^{-1}]^{-1}\hat{\beta} \end{aligned}$$

$$
\begin{aligned}
&= [(X^TX)^{-1}X^TY]^T \frac{(X^TX)}{\hat{\sigma}_u^2}(X^TX)^{-1}X^TY \\
&= Y^TX(X^TX)^{-1}\frac{X^TY}{\hat{\sigma}_u^2}
\end{aligned}
$$

Now defining the sample variances and covariances between $X$ and $Y$ as

$$
\begin{aligned}
S_{YY} &= \frac{Y^TY}{n} & \qquad S_{XY} &= \frac{X^TY}{n} \\
S_{XX} &= \frac{X^TX}{n} & \qquad S_{YX} &= \frac{Y^TX}{n}
\end{aligned}
$$

The Wald test statistic can be expressed as

$$
\frac{W}{n} = S_{YY}^{-1}S_{YX}S_{XX}^{-1}S_{XY} \tag{3.5}
$$

Alternatively it can be written that,

$$
\lambda = \left[\frac{1}{1+\frac{qF_{q,v}}{v}}\right]^{n/2} \tag{3.6}
$$

where, $F$ indicates an $F$-test statistic which is defined as

$$
\begin{aligned}
F_{q,v} &= \frac{\hat{\beta}^T[\hat{\sigma}_u^2(X^TX)^{-1}]^{-1}\hat{\beta}/q}{\frac{ns^2}{\hat{\sigma}_u^2}/(n-k)} \sim \ F_{q,n-k} \\
\Rightarrow F_{q,v} &= \frac{\hat{\beta}^T(X^TX)\hat{\beta}}{n\hat{\sigma}_u^2}\frac{(n-k)}{q} \\
\therefore \frac{q}{v}F_{q,v} &= \frac{\hat{\beta}^T(X^TX)\hat{\beta}}{n\hat{\sigma}_u^2}
\end{aligned} \tag{3.7}
$$

where, $s^2 = \hat{u}^T\hat{u}/v$ with $v = n-k$, which is unbiased estimator of $\hat{\sigma}_u^2$. Comparing (3.3) and (3.6), it can be written that

$$
\frac{W}{n} = \frac{qF_{q,v}}{v}
$$

$$\Rightarrow \frac{(n-k)}{n} W = qF_{q,v} \tag{3.8}$$

Considering the limiting distribution of $qF_{q,v}$ which is $\chi^2_q$ when $v \to \infty$, it can also be shown that the limiting distribution of $(\frac{n-k}{n})W$ will also follow $\chi^2_q$ when $v \to \infty$.

### 3.1.3 Lagrange Multiplier (LM) and *F*-Test

The Lagrange Multiplier (LM) test statistic is defined as

$$LM = \frac{\hat{\beta}^T (X^T X) \hat{\beta}}{\tilde{\sigma}^2_R} \text{ ; where, } \tilde{\sigma}^2_R = \frac{y^T y}{n} = \frac{\tilde{u}^T \tilde{u}}{n} \tag{3.9}$$

Note that, LM uses restricted estimates. Using the definition of the Wald test statistic, LM can be written as

$$LM = \left( \frac{\hat{\sigma}^2_u}{\tilde{\sigma}^2_R} \right) W \tag{3.10}$$

Relaying on (3.1)&(3.4), it can be expressed as

$$LM = \left[ \frac{W}{1 + \frac{W}{n}} \right] \tag{3.11}$$

Finally substituting equation for $W$, the LM can be expressed as,

$$LM = \frac{(v+k) qF_{q,v}}{v + qF_{q,v}} \tag{3.12}$$

Again, when $n \to \infty$, then

$$LM = qF_{q,\infty}$$

and $qF_{q,v} \sim \chi^2_q$ for large $n$.

### 3.1.4 Test Inequalities

Following Vandaele (1981), the likelihood ratio test statistic can be written as

$$LR = -2\log\lambda = n\log\left(1+\frac{W}{n}\right) \tag{3.13}$$

Using the inequality, $\frac{W}{n} \geq \log\left(1+\frac{W}{n}\right)$ the above equation (3.13) directly follows that

$$W \geq LR$$

Comparing (3.11)&(3.12), it can be written that

$$LR \geq LM$$

because, $\log\left(1+\frac{W}{n}\right) \geq \frac{\frac{W}{n}}{1+\frac{W}{n}}$. Finally, it can be said that for the multiple regression model, there is no need to use any of these large sample results as the test of the hypothesis of linear restrictions on $\beta$ can be derived as an exact $F$ test.

Buse (1982) also demonstrated an important relationship along with two examples among these large sample test procedures which is that if the log-likelihood function is quadratic then the LRT, Wald, and LM tests are numerically identical that is $W = LRT = LM$, and follow $\chi^2$ distribution for all sample sizes under the null hypothesis.

Buse (1982) demonstrated that if the log-likelihood function is quadratic then the LRT, Wald, and LM tests are numerically identical and have $\chi^2$ distribution for all sample sizes under the null hypothesis. He showed the equivalence among these test procedures demonstrating an example discussed as follows:

**Example:** Let $y_{n\times 1} = X_{n\times k}\beta_{k\times 1} + u_{n\times 1}$ with $u \sim N(0, \sigma^2 I)$ and test the hypothesis

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

The log-likelihood, a quadratic in the vector $\beta$, is given by

$$\begin{aligned} logL(\beta) &= -(n/2)(log2\pi) - 1/2log(\sigma^2) - (1/2\sigma^2)(y - X\beta)^T(y - X\beta) \\ &= c - (1/2\sigma^2)(y - X\beta)^T(y - X\beta) \end{aligned}$$

$$dlogL/d\beta = \frac{X^TY}{\sigma^2} - \frac{(X^TX)\beta}{\sigma^2}$$

and

$$d^2logL/d\beta d\beta^T = -\frac{(X^TX)}{\sigma^2}$$

In order to construct the LR statistic we need both the unrestricted and the restricted estimates. The unrestricted estimate of $\beta$ can be written as

$$\hat{\beta} = (X^TX)^{-1}X^TY$$

and under the restriction, $\tilde{\beta} = 0$.

Define $\hat{u} = y - X\hat{\beta}$ and $\tilde{u} = y - X\tilde{\beta} = y$. The unrestricted log-likelihood can be written as

$$logL(\hat{\beta}) = c - (1/2\sigma^2)\hat{u}^T\hat{u}$$

and the restricted log-likelihood as

$$logL(\tilde{\beta}) = c - (1/2\sigma^2)\tilde{u}^T\tilde{u}$$

Based on the relation that $\tilde{u}^T\tilde{u} = y^Ty = \hat{u}^T\hat{u} + \hat{\beta}^T(X^TX)\hat{\beta}$, since $\hat{\beta}^TX^T(y - X\hat{\beta}) = 0$, the above equation can be expressed as

$$logL(\tilde{\beta}) = c - (1/2\sigma^2)\hat{u}^T\hat{u} - (1/2\sigma^2)\hat{\beta}^T(X^TX)\hat{\beta}$$

The LR statistic is now given as

$$LR = (1/\sigma^2)\hat{\beta}^T(X^TX)\hat{\beta} = S_{YY}^{-1}S_{YX}S_{XX}^{-1}S_{XY}$$

The Wald statistic can be expressed as

$$W = \hat{\beta}^T[\sigma^2(X^TX)^{-1}]^{-1}\hat{\beta} = S_{YY}^{-1}S_{YX}S_{XX}^{-1}S_{XY}$$

Buse (1982) discussed the Lagrange multiplier test statistic as

$$LM = S(\tilde{\beta})^TC^{-1}S(\tilde{\beta})$$

and showed a relationship that $S(\tilde{\beta}) = \frac{(X^TX)\hat{\beta}}{\sigma^2}$. Following this relation, the LM test statistic can be expressed as

$$LM = S(\tilde{\beta})^TC^{-1}S(\tilde{\beta}) = \hat{\beta}^T[\sigma^2(X^TX)^{-1}]^{-1}\hat{\beta} = S_{YY}^{-1}S_{YX}S_{XX}^{-1}S_{XY}$$

So in the case of multiple regression, it is also found that $LR = W = LM$.

### 3.1.5 Assessing the Performance of Partial Least Squares Regression

Following Wakeling and Morris (1993), Partial Least Squares (PLS) regression is a widely used multivariate statistical method, particularly suitable when the number of factors exceeds the number of observations. It generates a set of uncorrelated score vectors,

which are linear combinations of the original variables.

However, two important issues arise in PLS: determining the number of dimensions to include and evaluating their statistical significance. Including more dimensions may lead to overfitting, although it may summarize the data well, resulting in poorer performance. Hence, it is essential to select only those dimensions that enhance predictive performance.

To assess a good model, testing for significance poses another challenge. Since there is no distributional theory available for testing PLS, model validation becomes crucial to evaluate the prediction accuracy. This can be measured using the Prediction Residual Error Sum of Squares (PRESS). A scree plot, based on PRESS values, can help determine the optimal number of dimensions to include in the model, but it is a data-specific process. For a more quantitative approach, using statistical tests is preferable.

In their work, Wakeling and Morris (1993) discussed two test statistics based on PRESS and attempted to identify a parsimonious model without a significant increase in PRESS.

In this paper, another significant test procedure based on cross-validated $R^2$ is also explored, which is analogous to $R^2$ in ordinary regression but can take negative values if the PRESS value is larger than the residual sum of squares (RSS). All the methods discussed in this paper are arbitrary in their choice of either degrees of freedom or some threshold, providing a rough idea of how many dimensions should be included.

To assess the predictive ability of the PLS model, critical values have been calculated using Markov Chain Monte Carlo (MCMC) simulation for different sample sizes. These critical values will be compared with the cross-validated $R^2$ obtained from the original data, allowing for the formulation of a statistical test statistic without resorting to a randomized test.

PLS poses another challenge regarding the determination of the appropriate number of degrees of freedom associated with a component. For instance, when performing prin-

cipal components first and then applying PLS to these principal components, the first PLS component provides the OLS solution. In this case, the first PLS component should have $r$ degrees of freedom associated with it, where $r$ is the number of explanatory variables, as that is the number of degrees of freedom typically associated with explanatory variables in an OLS model.

However, if there are high correlations between the $r$ explanatory variables, then the degrees of freedom to associate with any component should be much less than $r$. The lack of clarity on the appropriate number of degrees of freedom for a component gives rise to testing problems. Hence, cross-validation is employed to select the number of components to use in a PLS regression and to evaluate the model's performance [Wakeling and Morris (1993)].

As per Eldén (2015), the Partial Least Squares procedure is recursive, involving the computation of basis vectors (latent components) for both the explanatory variables and the solution vectors. Since the vectors of regression coefficients and predictions are nonlinear functions of the right-hand side, an algorithm for calculating Frechet derivatives of these functions is developed to address this issue. Utilizing the Frechet derivatives of the prediction vector, one can compute the number of degrees of freedom, which serves as a stopping criterion for the recursion process.

In Chapter 6, I applied various dimension reduction techniques for prediction, which included principal component analysis, between-group principal components (bgPCs), and leave-one-out cross-validated group PCs (cv-bgPCs). Each method yielded a different number of selected dimensions, and these dimensions were utilized to assess the performance of different predictive models. The valuable insights derived from Wakeling and Morris (1993)'s work were instrumental in effectively implementing these dimension reduction techniques and assessing the predictive ability of various models for dental morphology data.

## 3.2 Review of Canonical Correlation Analysis (CCA) and its Test Procedures

Canonical correlation analysis (CCA) is a multivariate statistical technique used to explore the relationships between two sets of variables [Johnson and Wichern (2007)]. It aims to find linear combinations of variables from each set that have the highest correlation with each other. The goal is to identify the maximum correlation between the two sets, referred to as canonical variates or canonical factors. The general theoretical steps of canonical correlation can be outlined as follows:

- It seeks to identify and quantify the relationship between two sets of variables
- In multiple correlations, it takes a linear combination of $X_1, X_2, \cdots, X_p$ and find its correlation with $Y$
- In CCA, it finds the relationship between two sets of variables $X_1, X_2, \cdots, X_p$ and $X_1^*, X_2^*, \cdots, X_s^*$
- Each set of variables be reduced to a single variable and finding their correlation
- The variables obtained by this linear combination are known as canonical variables. Suppose, $X^{(1)} = X_1, X_2, \cdots, X_r$ and $X^{(2)} = X_{(r+1)}, X_{(r+2)}, \cdots, X_{(r+s)}$. Assume that $r \leq s$ and

$$E(X) = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

and

$$Cov(X) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

- Consider, $r$ and $s$ dimensional co-efficient vector $a$ and $b$ such that

$$U = a^T X^{(1)}$$
$$V = b^T X^{(2)}$$
$$Cor(U, V) = \frac{a^T \Sigma_{12} b}{(a^T \Sigma_{11} a b^T \Sigma_{22} b)^{1/2}}$$

  Now, the question is what should be the value of $a$ and $b$?

- 1st pair $(U_1, V_1)$ are chosen as to maximize $Cov(U, V)$ subject to $var(u_1) = var(v_1) = 1$.

- 2nd pair $(U_2, V_2)$ are chosen as to maximize $Cov(U, V)$ subject to their combinations being orthogonal to the 1st choice.

- This can be done till $r$. 1st canonical correlation is given by

$$cor(u_1, v_1) = \rho = \sqrt{\lambda}_1$$

  with canonical variables

$$u_1 = a^T X^{(1)} = p_1^T \Sigma_{11}^{-1/2} X^{(1)}$$

  and

$$v_1 = b^T X^{(2)} = q_1^T \Sigma_{22}^{-1/2} X^{(2)}$$

  where, $\lambda_1, \lambda_2, \cdots, \lambda_r$ are the eigen values or characteristics roots of

$$[\Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1/2}]_{r \times r}$$

  with corresponding eigen vectors $p_1, p_2, \cdots, p_r$.

- In fact $\lambda_1, \lambda_2, \cdots, \lambda_r$ are also the largest $r$ eigen values of

$$[\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}]_{s\times s}$$

In general,

$$u_k = a^T X^{(1)} = p_k^T \Sigma_{11}^{-1/2} X^{(1)}$$

and

$$v_k = b^T X^{(2)} = q_k^T \Sigma_{22}^{-1/2} X^{(2)}$$

with

$$cor(u_k, v_k) = \rho_k^* = \sqrt{\lambda}_k; k = 1, 2, \cdots, r.$$

- Result: Let $\rho_1^{*2} \geq \rho_2^{*2} \geq \cdots \geq \rho_p^{*2}$ be the $p$ ordered eigenvalues of $S_{11}^{-1/2}S_{12}S_{22}^{-1}S_{21}S_{11}^{-1/2}$ with corresponding eigen vectors $\hat{e_1}, \hat{e_2}, \cdots, \hat{e_p}$ and $p \leq q$. Let $\hat{f}_1, \hat{f}_2, \cdots, \hat{f}_q$ be the eigenvectors of $S_{22}^{-1/2}S_{21}S_{11}^{-1}S_{12}S_{22}^{-1/2}$, where the first $p$ $\hat{f}'s$ may be obtained from $\hat{f}_k = (1/\rho_k^*)S_{22}^{-1/2}S_{21}S_{11}^{-1/2}\hat{e_k}; k = 1, 2, \cdots, p.$ The $k$th sample canonical variate pair is

$$\hat{U_k} = \underbrace{\hat{e_k}^T S_{11}^{-1/2}}_{\hat{a_k^T}} x^{(1)}$$
$$\hat{V_k} = \underbrace{\hat{f_k}^T S_{22}^{-1/2}}_{\hat{b_k^T}} x^{(2)}$$

The first sample canonical variate pair has the maximum sample correlation

$$r_{\hat{U_1},\hat{V_1}} = \hat{\rho_1^*}$$

- To ease the mathematical burden, many people prefer to get the canonical corre-

lations from the eigenvalue equation

$$|\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \rho^{*2}I| = 0$$

The coefficient vectors $a$ and $b$ follow directly from the eigenvector equations

$$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}a = \rho^{*2}a$$
$$\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}b = \rho^{*2}b$$

### 3.2.1 Canonical Correlation Link with Multiple Correlation Coefficients

Multiple correlation coefficient, $\rho_{1,x^{(2)}}$ is a special case of a canonical correlation when $x^{(2)}$ has a single element *i.e.*, $p = 1$ [Johnson and Wichern (2007)]. So,

$$\rho_{1,x^{(2)}} = \max_b cor(x^{(1)}, b^T x^{(2)}) = \rho_1^*$$

When $p > 1$, $\rho_1^*$ is larger than each of the multiple correlation of $x_i^{(1)}$ with $x^{(2)}$ or the multiple correlation of $x_i^{(2)}$ with $x^{(1)}$. That is,

$$\rho_{u_k,x^{(2)}} = \max_b cor(u_k, b^T x^{(2)}) = cor(u_k, v_k) = \rho_k^*; k = 1, 2, \cdots, p.$$

Similarly,

$$\rho_{v_k,x^{(1)}} = \max_a cor(v_k, b^T x^{(1)}) = cor(v_k, u_k) = \rho_k^*; k = 1, 2, \cdots, p.$$

That is, the canonical correlations are also the multiple correlation coefficients of $U_k$ with $X^{(2)}$ or the multiple correlation coefficients of $V_k$ with $X^{(1)}$. Because of its multiple correlation coefficient interpretation, the $k^{th}$ squared canonical correlation, $\rho_k^{*2}$, is the proportion of the variance of the canonical variate, $U_k$ explained by the set $X^{(2)}$. It is

also the proportion of the variance of canonical variate $V_k$ explained by the set $X^{(1)}$. Therefore, $\rho_k^{*2}$ is often called the shared variance between the two sets $X^{(1)}$ and $X^{(2)}$.

### 3.2.2 A general parametric significance-testing System for the Canonical Correlation Analysis

If there is one set of $p$ variables and another set of $q$ variables (where, $q \leq p$), then the principle objective of canonical correlation analysis is to find a linear combination of the $p$-variables that correlate maximally with linear combinations of the $q$ variables and for sample data, to test statistical significance of that correlation [Knapp (1978)]. The weight for the $q$ variable in the second set is obtained by finding the elements of the eigenvector $v_1$ associated with the largest eigenvalue $\lambda_1$ of the matrix

$$M = R_{yy}^{-1} R_{yx} R_{xx}^{-1} R_{xy}$$

where, $R_{yy}^{-1}$ is the inverse of $q \times q$ matrix,$R_{yx}$ is the $q \times p$ matrix of cross-correlations between two sets, $R_{xx}^{-1}$ is the inverse of $p \times p$ matrix. The weights for $p$ variables in the first set are obtained by finding the elements of the vector

$$v_2 = \lambda_1^{-1/2} R_{xx}^{-1} R_{xy} v_1$$

The maximal canonical correlation $r_o$ is the square root of $\lambda_1$. Its significance is tested by referring to a table of the $F$ sampling distribution, the following statistic for $pq$ and $(ms - pq/2 + 1)$ degrees of freedom.

$$F = \frac{[1 - \Lambda^{1/s}]/pq}{\Lambda^{1/s}/(ms - pq/2 + 1)}$$

where,

$$\Lambda = \prod_{i=1}^{q}(1-\lambda_i); i = 1, 2, \cdots, q$$

$$m = N - 3//2 - (p+q)/2$$

$$s = [(p^2q^2 - 4)\boldsymbol{\nabla} \cdot \left(p^2 + q^2 - 5\right)]^{1/2}$$

The test [Rao (1952)] is exact if either $p$ or $q$ is less than or equal to two and is approximate otherwise.

#### 3.2.2.1 Link with Multiple Regression Analysis

The major difference between multiple regression analysis and canonical analysis is that the former employs just one variable in the second set, that is $q = 1$. Therefore, $R_{yy}^{-1}R_{yx}R_{xx}^{-1}R_{xy}$ reduces to $R_{yx}R_{xx}^{-1}R_{xy}$, which is recognizable as $R_{yx}b$, where, $b$ is the columns vector of beta weights (standardized partial regression coefficients) and which in turn reduces to the scalar $r^2_{y.x_1,x_2,\cdots,x_p}$ (since, $R_{yx}$ is row vector of the correlations of each of the variables in the first set with the variable in the second set that is, the square of the multiple correlation coefficient. The largest eigenvalue of $r^2$ is again $r^2$ itself. The $F$ formula reduces to

$$F = \frac{r^2/p}{(1-r^2)/(N-p-1)}$$

which is equal to the traditional formula for testing the significance of a multiple $r$, since, $q = 1$, $s = 1$, $\Lambda^{1/s} = \Lambda = 1 - \lambda_1 = 1 - r^2$ and $m = N - 3/2 - (p+1)/2$. The case $p = 2$ and $q = 1$ represents a special difficulty, since $p^2 + q^2 - 5 = 0$ and $s$ is undefined. $F$ still the same multiple regression, $F$, however.

There is a test of the significance of a canonical correlation coefficient due to Bartlett

(1941) that is not subject to this constraint

$$\chi^2 = -[N - \frac{(p+q+1)}{2}]log_e\Lambda \sim \chi_{pq}(\alpha)$$

### 3.2.3 Tests for Determining the Number of Nonzero Canonical Correlations

In canonical correlation analysis, the usual interest to test the association between two sets of variates and the number of non-zero population canonical correlation may be called as dimensionality. Caliński et al. (2006) compared the tests for the dimensionality in the canonical correlation analysis with regard to the relative frequencies of underestimation, correct estimation, and overestimation of the true dimensionality. Suppose $\underbrace{y_1, y_2, \cdots, y_p}_{\text{p}}$ and $\underbrace{X_1, X_2, \cdots, X_q}_{\text{q}}$ are two sets of observations which have the variance-covariance matrix as follow

$$\text{Cov}(Y, X) = \Sigma = \begin{pmatrix} \underbrace{\Sigma_{11}}_{p\times p} & \underbrace{\Sigma_{12}}_{p\times q} \\ \underbrace{\Sigma_{21}}_{q\times p} & \underbrace{\Sigma_{22}}_{q\times q} \end{pmatrix}$$

The likelihood ration test of $H_0 : \Sigma_{12} = 0$ vs $H_1 : \Sigma_{12} \neq 0$ rejects $H_0$ for large value of

$$\begin{aligned} -2\log\Lambda =& n\log\frac{|S_{11}||S_{22}|}{|S|} \\ =& -n\log\frac{|S|}{|S_{11}||S_{22}|} \\ =& -n\log\frac{|S_{22}||S_{11} - S_{12}S_{22}^{-1}S_{21}|}{|S_{11}||S_{22}|} \\ =& -n\log\big(|I - S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}|\big) \\ =& -n\log\prod_{i=1}^{min(p,q)}(1-\hat{\rho}_i^2) \end{aligned} \tag{3.14}$$

where, $\rho_1, \rho_2, \cdots \rho_{min(p,q)}$ are the square root of the eigenvalue of $S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}$. The likelihood ratio statistic compares the sample generalized variance under $H_0$

$$\begin{pmatrix} S_{11} & 0 \\ 0 & S_{22} \end{pmatrix} = |S_{11}||S_{22}|$$

with the unrestricted generalized variance $|S|$. Bartlett (1941) suggests replacing the multipliplicative factor $n$ in LRT with the factor $(n - 1 - 1/2(p + q + 1))$ to improve the $\chi^2$ approximation to the sampling distribution of $-2\log\Lambda$. Thus for large $n$, Reject $H_0 : \Sigma_{12} = 0(\rho_1^* = \rho_2^* = \cdots = \rho_p^* = 0)$ at significance level $\alpha$ if

$$-(n - 1 - 1/2(p + q + 1))\log \prod_{i=1}^{min(p,q)} (1 - \hat{\rho}_i^2) > \chi_{pq}^2(\alpha)$$

where, $\chi_{pq}^2(\alpha)$ is the upper $100\alpha$th percentile of a chi-square distribution with $pq$ degrees of freedom. If the $H_0 : \Sigma_{12} = 0$ is rejected, it is natural to examine the significance of the individual canonical correlations. Because the coefficients are ordered from largest to smallest, it can start by assuming the first canonical correlation is nonzero and the remaining $(p - 1)$ are zero. If it is rejected, assume the first two canonical correlations are nonzero but the remaining $(p - 2)$ are zero, and so forth. That is

$$H_d : \rho_{d+1} = \rho_{d+2} = \cdots = \rho_p = 0 \text{ for } d = 1, 2, \cdots, (p - 1)$$

Bartlett (1941) has argued that the above situation can be tested by the LR criterion. Specifically, reject null hypothesis if

$$-(n - 1 - 1/2(p + q + 1))\log \prod_{i=d+1}^{min(p,q)} (1 - \hat{\rho}_i^2) > \chi_{(p-d)(q-d)}^2(\alpha) \tag{3.15}$$

It is noticed that the test statistic involves $\prod_{i=d+1}^{min(p,q)}(1 - \hat{\rho}_i^2)$, the residual after the

first $d$ sample canonical correlations have been removed from the total criterion

$$\Lambda^{2/n} = \prod_{i=1}^{min(p,q)} (1 - \hat{\rho}_i^2)$$

$$\therefore -2 \log \Lambda = -n \log \prod_{i=1}^{min(p,q)} (1 - \hat{\rho}_i^2)$$

For testing $H_d$, the following three test statistic analogus to the statistics proposed for MANOVA can be considered

(1) Lawley-Hotelling trace statistic

$$T_d^2 = \sum_{i=d+1}^{min(p,q)} \frac{\hat{\rho}_i^{*2}}{1 - \hat{\rho}_i^{*2}}$$

(2) Wilk's statistic (Likelihood ratio statistic)

$$\Lambda_d = \prod_{i=d+1}^{min(p,q)} (1 - \hat{\rho}_i^2)$$

(3) Bartlett-Nanda-Pillai trace statistic

$$V_d = \sum_{i=d+1}^{min(p,q)} \hat{\rho}_i^2$$

Under $H_d$, Bartlett $\chi^2$-approximation can be applied to $T_d^2, \Lambda_d$ and $V_d$ when transformed to

$$S_B(T_d^2) = (n - p - q - 2)T_d^2$$

$$S_B(\Lambda_d) = -[n - 1 - 1/2(p + q + 1)] \log \Lambda_d$$

$$S_B(V_d) = (n - 1)V_d$$

Their distribution under $H_d$ is approximately $\chi^2$ with $(p-d)(q-d)$ degrees of freedom.

Olson (1976) suggested that multivariate analysis of variance must be chosen such that test statistics can be expressed as a generalization of the usual univariate $F$ statistic. He also reviewed the above-described tests which leads to the recommendation that Pillai-Bartlett statistic for general use. He discussed that the Hotelling-Lawley statistic makes use of all the sample eigenvalues by summing them but the eigenvalue can take any value from 0 to $\infty$ which will result that a single deviant eigenvalue may have a severe effect on the test statistic. In Wilks's likelihood ratio statistic, the eigenvalues are multiplied which also affects severely if a deviant of eigenvalue occurs. On the other hand, the Pillai-Bartlette statistic is defined in such a way that it is relatively well protected from the deviant of eigenvalues as it uses summing rather than multiplying them.

## 3.3 Optimality of Tests

When conducting association tests to detect genetic signals, the primary objective is to identify the most powerful test, whenever possible. This allows us to maximize the chances of discovering a larger number of relevant genes associated with the trait under investigation. Discovering more genes also means their subsequent use in prediction, making the prediction results more accurate. However, there is no broad discussion available about the power and optimality of the various statistical tests employed in the statistical genetics literature. It is therefore of interest to discuss such properties of the various tests encountered in this literature review and find out if any of them are better than the others, at least in certain scenarios.

In practice, sometimes uniformly most powerful test does not exist because the class of level $\alpha$ tests is so large that no one test dominates all the others in terms of power [Casella and Berger (2002)]. In such cases, a common method of continuing the search

for a good test is to consider some subset of the class of level $\alpha$ tests and attempt to find a UMP test in this subset. When no UMP level $\alpha$ test exists within the class of all tests, it might be tried to find a UMP level $\alpha$ test within the class of unbiased tests.Young and Smith (2005) also discussed that sometimes characterizing UMPU tests for two-sided problems is a much harder task than characterizing UMP tests for one-sided hypotheses, but for one specific but important example, that of a one-parameter exponential family, it ables to find UMPU tests. The extension to multiparameter exponential families involves the notion of conditional tests. Here it have been discussed two situations where conditional tests naturally arise, one when there are ancillary statistics and the other where conditional procedures construct similar tests. The basic idea behind an ancillary statistic is that of a quantity with distribution not depending on the parameter of interest and eliminating the dependency on nuisance parameters.

Edelman demonstrated that if one accepts a restriction, and attention is limited to the class of tests $\zeta$ that have this property, then uniformly most powerful tests do indeed exist for the two-sided testing problem, and they are equal to the ones usually proposed for the two-sided testing problem. It is well known that the usual one-sided test for the mean of a normal distribution (with variance known) is uniformly most powerful but within the class of all possible tests, there is no uniformly most powerful two-sided test for the mean of a normal distribution. He explained that in many two-sided testing situations, considering the symmetry suggests that it is reasonable to restrict attention to tests that have a critical region that is symmetric about the null mean $\boldsymbol{\mu_0} = (\mu_0, \mu_0, \cdots, \mu_0)$; that is, if an observation $\boldsymbol{x} = (x_1, \cdots, x_n)$ is to lead to rejection, then an observation of $(2\boldsymbol{\mu_0} - \boldsymbol{x})$ should also lead to rejections. The previous argument may be extended to include certain testing problems in which there exists a sufficient statistic for the parameter of interest and when certain symmetry properties are present.

## 3.4 Novel Contribution

In the previous sections of this Chapter, I conducted a review of various statistical test procedures in both multiple regression and canonical correlation settings. Based on the characteristics of these test procedures, my first step was to compare the test procedures in the multiple regression setting and assess their suitability as uniformly most powerful tests or based on specific observations.

Next, I revisited several multivariate test procedures from the canonical correlation setting and adapted them for use in the multiple regression setting. I then calculated their approximate distributions to facilitate a comprehensive comparison of all the test procedures, including the previously recommended one. This comparison allowed me to identify a unified test procedure for further association studies.

### 3.4.1 Comparing Test Procedures in Multiple Regression Settings: A Literature Review-Based Analysis

Classical genome-wide association studies (GWAS) model a single phenotype against one genetic marker at a time. So even though a large number of markers are tested, they are tested separately. And the model employed is usually a multiple regression model, since in addition to the genetic marker, other covariates like age and sex are usually included.

However, other studies have employed multiple regression models more directly, where multiple genetic markers against a phenotype, or multiple phenotypes against a genetic marker [Adhikari et al. (2015)], have been tested. Such uses have particular advantages. For example, when multiple genetic variants within a gene have small effects on a phenotype, testing them individually can be underpowered to their weak effects, but testing them together in a joint model can be useful in pooling together all the evidence. A joint model also allows us to take into consideration the correlation

structure between the variables.

In the first part of my PhD project, I am looking into multivariate statistical procedures, particularly in the context of multiple linear regression. My aim is to assess various testing procedures to look for more powerful tests and see if the most powerful test exists in any particular scenario. For example, in studying two-sided tests for one-parameter exponential families, we have seen that UMP tests don't exist in this scenario, but a UMPU test exists [Young and Smith (2005)]. However such discussions do not exist for multi-parameter scenarios, such as the multiple linear regression problem studied here.

But even if explicit discussion of the power of tests is not available, some observations can be made by looking at the tests themselves. In the multiple linear regression context, Berndt and Savin (1977) showed an inequality between the Wald test, likelihood ratio test, and Lagrange multiplier test:

$$W \geq LR \geq LM$$

As Vandaele (1981) showed that they all are tested in small samples by the same F-test, it means that the test statistic with the largest numerical value, i.e. $W$, will have the highest power. However, Buse (1982) demonstrated that if the log-likelihood function is quadratic then the three test statistics are identical, i.e. $W = LRT = LM$.

These observations inform us that from this class of tests, it suffices to focus on only the Wald test for the rest of the analysis. Building on the same premise, I compared other test procedures applicable to multiple linear regression with the Wald test. The next section focuses on the adaptation of various test procedures for canonical correlation to multiple regression settings, enabling a thorough comparison among these procedures as well as the Wald test.

### 3.4.2 Equivalence of Multivariate Test Procedures in Canonical Correlation for Multiple Regression

In the literature review, only the canonical correlation method was found to have a theoretical testing procedure. Hence, I modified the multivariate test procedures to suit multiple regression settings and provided a comparison of its testing procedures to the Wald test.

I first revisit the Wald test for multiple linear regression. Consider a multiple regression model discussed in equation (2.2) and suppose, we would like to test the null hypothesis that the entire coefficient vector is zero, $H_0 : \beta = 0$. It has shown that the Wald test statistic can be obtained as a classical $F$ test statistic:

$$\begin{aligned} F_{q,v} =& \frac{\hat{\beta}^T[\hat{\sigma}_u^2(X^TX)^{-1}]^{-1}\hat{\beta}/q}{\frac{ns^2}{\hat{\sigma}_u^2}/(n-k)} \sim F_{q,n-k} \\ \Rightarrow F_{q,v} =& \frac{\hat{\beta}^T(X^TX)\hat{\beta}}{n\hat{\sigma}_u^2}\frac{(n-k)}{q} \quad \text{[Using the equations (3.7) \& (3.8)]} \\ \Rightarrow qF_{q,v} =& \Big(\frac{n-k}{n}\Big)W \end{aligned}$$

Since the limiting distribution of $qF_{q,v}$ is $\chi_q^2$ when $n-k = v \to \infty$, so the limiting distribution of $(\frac{n-k}{n})W$ will also follow $\chi_q^2$ when $n \to \infty$. Put in a simpler way, for large samples, $\Big(\frac{n-k}{n}\Big)W \approx W$ will follow a $\chi_q^2$ distribution asymptotically.

Next, I considered the method of canonical correlations. It generally considers the relationship between two multivariate data sets $Y$ and $X$. However, when $Y$ is univariate, it reduces to the same problem conceptually as multiple linear regression. In this particular context, I have considered the different multivariate test procedures for canonical correlation as described in section (3.2). The widely used testing procedures are the Lawley-Hotelling trace statistic, Wilks's lambda statistic, and Bartlett-Nanda-Pillai trace statistic. In canonical correlation, it is attempted to test whether there are

any correlations among two sets of variates e.g., $H_0 : \Sigma_{12} = 0$ vs $H_1 : \Sigma_{12} \neq 0$. Initially, the purpose was to make comparisons among various testing procedures in multiple regression setups, so I attempted to simplify these multivariate test procedures to multiple regression setups first and then make a comparison among them. At this point, I tried to simplify all the test procedures in terms of correlation-covariance and the reason behind this is that it will help to form all the test statistics considering the covariance structure of the data. In addition to that comparing the test procedures in this way and referring to a powerful test, the impact of joint effects can be evaluated and eventually, the most relevant factors can be chosen.

Firstly, I simplified the Wald test statistic in terms of covariance structures and the simplified form has been shown in the equation (3.5) as $W = nS_{yy}^{-1}S_{yX}S_{XX}^{-1}S_{Xy}$. In multivariate canonical test procedures, if we constraint one set of covariate equal to one then the term $R_{yy}^{-1}R_{yx}R_{xx}^{-1}R_{xy}$ will be reduced to $R_{yx}R_{xx}^{-1}R_{xy}$ which is also recognizable as $R_{yx}\beta$ where $\beta$ is the column vector of beta weights, and eventually it will reduce to the scalar $R^2_{y.x_1,x_2,\cdots,x_p}$ that is the square of the multiple correlation coefficient [Knapp (1978)]. With this reference, I simplified the reduced term in terms of covariance as

$$\begin{aligned} R_{yx}R_{xx}^{-1}R_{xy} =& (S_{yy}^{-1/2}S_{yx}S_{xx}^{-1/2})(S_{xx}^{-1/2}S_{xx}S_{xx}^{-1/2})(S_{xx}^{-1/2}S_{xy}S_{yy}^{-1/2}) \\ =& S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{xy} \end{aligned}$$

Now, the squared multiple correlation between $y$ and $X_1, X_2, \cdots, X_p$ is

$$R^2 = r_{yx}^T R_{xx}^{-1}R_{xy} = S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{xy} = \frac{W}{n}$$

It is noticeable that the multiple correlations ($R^2$) and $\frac{W}{n}$ are equivalent if the test statistics are expressed in terms of covariance structure.

Considering different multivariate test procedures for canonical correlation setup in

section (3.2), I simplified these test procedures to multiple regression setup. As described earlier, by letting one set of covariates equal to one, the canonical correlation setup can be linked with multiple regression by expressing the correlations in terms of $R^2$. The values of $\hat{\rho}_i^*$ are called canonical correlations. As $p = 1$, the largest canonical correlation $\hat{\rho}_1^*$, is used as a measure of association of the two sets of variables and $\hat{\rho}_1^{*2}$ is the maximum squared correlation between a linear combination of the $y$ variable and a linear combination of the $x$ variables. It is also noticeable that the first eigenvalue of $R^2$ is again $R^2$, that is $\hat{\rho}_1^* = R^2$. So, to test the hypothesis $H_0 : \rho_1^* = 0 (d = 0)$, the test statistics for canonical correlations can be expressed as

(1) Lawley-Hotelling trace statistic

$$T_0^2 = \frac{\hat{\rho}_1^{*2}}{1 - \hat{\rho}_1^{*2}} = \frac{R^2}{1 - R^2} = \frac{W/n}{1 - W/n} \tag{3.16}$$

(2) Wilk's statistic

$$\Lambda_0 = (1 - R^2) = (1 - W/n) \tag{3.17}$$

(3) Bartlett- Nanda-Pillai trac statistic

$$V_0 = R^2 = W/n \tag{3.18}$$

Under $H_0$, Bartlett $\chi^2$ approximation can be applied to transform as

$$\begin{aligned} S_B(T_0^2) =& (n - q - 3) T_0^2 \\ =& (n - q - 3)(\frac{R^2}{1 - R^2}) \\ =& (n - q - 3)(\frac{W/n}{1 - W/n}) \sim \chi_q^2(\alpha) \\ S_B(\Lambda_0) =& - [n - 1 - 1/2(q + 2)] \ \log \Lambda_0 \\ =& - [n - 1 - 1/2(q + 2)] \ \log\big(1 - R^2\big) \end{aligned} \tag{3.19}$$

$$= -\left[n-1-1/2(q+2)\right]\ \log(1-W/n) \sim \chi^2_q(\alpha) \tag{3.20}$$

$$S_B(V_0) = (n-1)V_0 = (n-1)R^2 = (n-1)\frac{W}{n} \sim \chi^2_q(\alpha) \tag{3.21}$$

Just like the comparison among conventional test procedures in multiple regression settings ($W \geq LR \geq LM$), a similar comparison can be made for the test procedures mentioned above. To facilitate the comparison, equations (3.19), (3.20), and (3.21) can be expanded as follows:

$$\begin{aligned}
S_B(T_0^2) &= (n-q-3)\frac{W}{n}\left(1-\frac{W}{n}\right)^{-1} \\
&= \left(\frac{n-q-3}{n}\right) W\left[1+\frac{W}{n}+\frac{W^2}{n^2}+\frac{W^3}{n^3}+\cdots\right] \qquad (3.22) \\
&\quad [\text{Using } (1-x)^{-1} = 1+x+x^2+x^3+\cdots] \\
S_B(\Lambda_0) &= -[n-1-\frac{1}{2}(q+2)]\ \log(1-W/n) \\
&= -\left(n-1-\frac{1}{2}(q+2)\right)\left[-\frac{W}{n}-\frac{W^2}{2n^2}-\frac{W^3}{3n^3}-\cdots\right] \\
&\quad [\text{Using } log(1-x) = -x-\frac{x^2}{2}-\frac{x^3}{3}-\cdots] \\
&= \left(n-1-\frac{1}{2}(q+2)\right)\frac{W}{n}\left[1+\frac{W}{2n}+\frac{W^2}{3n^2}+\cdots\right] \\
&= \left(\frac{n-1-\frac{1}{2}(q+2)}{n}\right) W\left[1+\frac{W}{2n}+\frac{W^2}{3n^2}+\cdots\right] \qquad (3.23) \\
S_B(V_0) &= (n-1)\frac{W}{n} \\
&= \left(\frac{n-1}{n}\right) W \qquad (3.24)
\end{aligned}$$

The expansions of the test statistics in equations (3.22), (3.23), and (3.24) reveal that they can be arranged in order of magnitude, which in turn affects their respective power. The sequence of test procedures can be expressed as $S_B(T_0^2) > S_B(\Lambda_0) > S_B(V_0)$. In other words, Lawley-Hotelling Trace Statistic > Wilk's Statistic > Bartlett- Nanda-Pillai

Trac Statistic.

By using the simplified test procedures in multiple regression scenarios, an asymptotic comparison has been conducted to illustrate their behavior for large samples (n) and various values of the Wald test statistic (W). When examining the asymptotic distribution of $W$ as $n$ approaches infinity, it becomes apparent that not only is $W/n = R^2 \leq 1$ but also under the null hypothesis, $W$ follows a $\chi^2_q$ distribution. This implies that $W/n$ converges to 0 in probability, or in other words, $W/n$ is of the order $o_p(1)$. Leveraging these insights and considering large sample sizes, the aforementioned test procedures, described in equations (3.22), (3.23), and (3.24), can be further simplified as:

$$S_B(T_0^2) \approx W \sim \chi^2_q(\alpha)$$
$$S_B(\Lambda_0) \approx W \sim \chi^2_q(\alpha)$$
$$S_B(V_0) \approx W \sim \chi^2_q(\alpha)$$

That is, under the null when the sample size goes to infinity, the three test procedures are asymptotically equal and follow $\chi^2_q$ distribution asymptotically, and they are also asymptotically equivalent to the Wald test statistic itself.

The simplification of canonical correlation to multiple regression and the equivalence between $R^2$ and $(W/n)$ suggest that the Wald test statistics is a suitable choice for testing the coefficient vector in the multiple regression setup. Furthermore, it is asymptotically equivalent to test procedures used in the area of canonical correlation. Therefore, in further studies, it will suffice to investigate only the Wald test statistic in further studies.

# 4 Gain of Power by Conditioning

This chapter commenced with an exploration of genetic data in section (4.1), explaining aspects like the genetic structures among SNPs and individuals, implications of relatedness, and population structures. Section (4.2) then describes the outlines of GWAS models, encompassing both traditional and conditional models. It includes a motivational example showcasing how conditioning on major genes can enhance association power for the 'melanin' trait. This example involves selecting a list of major conditional SNPs based on unconditional models and identifying the major genes through a Manhattan plot. This plot is drawn by using the P-values for the genetic variations and their respective chromosomes.

Theoretical aspects including standard errors and statistical power are thoroughly explored in sections (4.4), (4.5), and (4.6), across simple linear and multiple regression scenarios. These theoretical developments consider various modelling setups and elucidate how conditioning on major variants improves the discovery power of new genetic variants. The viability of enhancing power through conditioning on major variants is extensively scrutinized using mathematical expressions, while also considering genetic variant characteristics within chromosomes, such as LD structure, proxy variants, and the population size of the database. Subsequently, these mathematical advances are demonstrated in two genetic databases—the CANDELA cohort and UK BioBank—in section (4.11), showcasing detailed power gain results through power graphs.

Moreover, mathematical expressions demonstrating how conditional results can be derived from publicly available summary statistics of univariate associations in GWAS studies, even without access to individual-level data, are discussed. This begins with a mathematical linkage between a single SNP model and joint SNP models. Sections (4.8) and (4.9) extend the regression modelling setup by considering 2-Block and 3-

Block matrix design matrices, illustrating the calculation of conditional results using publicly available summary statistics even when individual-level data are inaccessible. A comparison between the conditional coefficient results from the 3-Block approach and those derived from the GCTA software is made, demonstrating that the 3-Block approach's conditional coefficients closely align with true conditional results for all tested SNPs, surpassing those from the GCTA software. Finally, the chapter concludes with a comprehensive conclusion in section (4.12), summarizing all mathematical derivations regarding power gain and their real-world implications in databases.

## 4.1 Overview of Genetic Data

Suppose genetic data contains $n$ individuals, $y$ is the output variable indicating the phenotypes of the individuals, which can take single or multiple traits (e.g., height, pigmentations, diseases, etc.), $X$ denotes the information regarding genotypes of $n$ individuals which can be three cases:

- Single SNP case, where a single genetic marker will look for a genetic variation on the traits, and the performance will be evaluated by a single P-value and plotting Manhattan plot. In this situation, the dimension of the genetic variable, $X$, is $n \times 1$.

- Multi-SNP case, where a set of SNPs will be taken as genetic variants and multiple regression analysis can be applied to evaluate the potential genetic variants. A penalized regression method, 'LASSO', can be used to select a subset of variants, say, $q$, and then the dimension of the genotype matrix will be $n \times q$.

- All SNPs cases, where all possible SNPs will be included. Considering all SNPs (say, $m$) in the model, the genotype matrix will be $n \times m$ dimensional, where $m$ can be one million or even more and much larger than $n$. The product $XX^T$

(dimension $n \times n$) will represent population structure among $n$ individuals, also known as the kinship matrix.

The population structure indicates the genetic covariance/correlation matrix among the individuals derived from the genetic data, i.e., SNPs. Usually, the covariance matrix is analyzed with a linear mixed model (LMM) or some such methods; for example, the genome-wide complex trait analysis (GCTA) tool uses an LMM to estimate the heritability of a trait from genetic data [Yang et al. (2011)]. In GWAS, it's common to incorporate a small number of principal components that capture the most variability in the trait to address kinship effects.

Suppose $U$ denotes the information regarding covariates which can be entered in the model as:

- Basic covariates such as age, sex, etc. Sometimes $Y$ (and $X$) is pre-adjusted to remove effects of $U$, i.e., regressing $Y$ on $U$, and using residuals for further analysis such as GCTA heritability.

- Genetic-derived variables such as genetic principal components (PCs) are calculated by decomposing the eigenvalue of the kinship matrix (classical GWAS). Here, the number of principal components (PCs) is determined empirically and used as a fixed effect in an OLS. In such a case, the whole kinship matrix is not used in an LMM anymore. Often software implementations pre-adjust $Y$ with $U$, and take residuals, to simplify the analysis of $Y$ with $X$.

### 4.1.1 Genetic Structure among SNPs

Linkage disequilibrium (LD) indicates the correlation among SNPs ($X$ variables) measured as $r^2$ or $D'$. The correlation matrix obtained from $X^TX$ (dimension: $m \times m$) represents $r^2$ between pairwise SNPs. Linkage disequilibrium (LD) between SNPs arises

due to the transmission of genetic materials through generations, and it can be expressed as LD $\propto r^2$.

SNPs on different chromosomes will not be in LD, as chromosomes are inherited independently. The kinship matrix has correlation blocks due to the genetic recombination process being non-uniform throughout the genome. Closely situated SNPs on a genome are highly correlated; therefore, including them in a regression model can create collinearity. Therefore, it is essential to consider no LD (i.e., uncorrelated SNPs) when working with multiple $X$ variables simultaneously.

### 4.1.2 Genetic Structure among People

Genetic similarity is a measure of genetic relatedness among individuals that can arise due to

- Recent relatedness, also termed as familial kinship.
- Historical relatedness, also known as demographic history, population structure, or admixture.

These two sources of relatedness can have very different effects on the overall kinship matrix $XX^T$, especially on genetic PCs. Recent kinship may produce a small block of related people, which can inflate the variance of all PCs even when a single family is included [Hoffman (2013)]. Then it becomes necessary to use the entire $XX^T$ matrix through an LMM.

On the other hand, historical population structure may affect everyone and appear as a gradient instead of blocks. Then these gradients are well represented by the top few PCs of $XX^T$.

### 4.1.3 Relatedness Consequence on $Y$

Familial relatedness or demographic history affects not only the genetic structure but the shared environment, too, causing correlation structure in $X$. For example, a similar diet ($U$) can affect the weight phenotype ($Y$). So, $Y$ can have familial similarity through $U$ without any effect of genes $X$, which may be called confounded effects.

### 4.1.4 Adjustment of Population Structure

Genome-wide association studies (GWAS) usually aim to identify genetic associations among individuals that are ancestrally similar but differ phenotypically. However, different ethnicities and familial relatedness may be included in the same study and cause population structure in the dataset due to having similar genetic signals of associations. As a consequence, the power, as well as the efficiency of genetic association can be severely jeopardized.

GWAS and all other genetic studies carefully consider the impact of ancestry and relatedness, mainly when the participants of a data set come from diverse backgrounds, to avoid false positive or negative genetic signals [Marchini et al. (2004)]. GWAS usually adjusts these population structures by calculating principal components from the genotype of all individuals and including them as covariates in subsequent GWAS regression models [Price et al. (2006)].

The classical GWAS models take the genetic principal components (PCs) along with other demographic factors, such as age and sex, as a fixed or random effect in the model but do not check their empirical performance explicitly. While a fixed effect model includes relatively a few principal components ($i \ll n$) as a fixed effect, an LMM uses the genome-wide similarity between all pairs of individuals to account for population structure but it requires higher computational cost due to its complexity in handling both fixed and random effects [Kang et al. (2008)].

Hoffman (2013) suggested that under a transformation, it can be shown that there is a relationship between modelling the effect of principal components as a fixed vs. random, and the effects share the same underlying regression model. Only the difference is in their ability to account for population structure, inference method, and number of principal components included in the model [Kenny et al. (2011); Price et al. (2010); Wu et al. (2011)].

## 4.2 Overview of GWAS Models

### 4.2.1 Notation

Consider a GWAS model which consists of $n$ individuals. Let $X_j$ be the genotype value of a particular SNP $j$ ($j = 1, 2, \cdots, m$), $y$ be the trait of interest ( say, skin color), and $U$ be a set of covariates such as age, sex, genetic PCs etc. Traditional GWAS uses a simple regression model to assess the genetic association between an SNP and a phenotypic trait of interest with the adjustment of covariates such as sex, genetic PCs as

$$response = SNP_j + \underbrace{age + sex + \text{genetic PCs}}_{covariates} + \epsilon \; ; \quad j = 1, 2, \cdots, m$$

$$or, \; y = x_j b_j + U + \epsilon \tag{4.1}$$

here, $b_j$ indicates the regression coefficients of the classical GWAS model which represents the strength and direction of the association between a genetic variant and a phenotypic trait. Sometimes, the power of detecting the genetic association can be improved considering additional SNPs along with the effect SNP, i.e., conditioning on some

other SNPs, and the conditional GWAS model can be modelled as

$$response = SNP_j + \underbrace{SNP_{m+1} + SNP_{m+2} + \cdots + SNP_{m+k}}_{major\ SNPs} + U + \ \epsilon \ ; \quad j = 1, 2, \cdots, m \tag{4.2}$$

The above conditional GWAS model is conditioned on $(m - k)$ major SNPs along with some non-genetic covariates. Choosing a set of major SNPs typically involves methods like Genome-Wide Association Studies (GWAS), which assess the association between SNPs and a phenotype of interest. Selection criteria vary but often include statistical significance (such as p-values), effect size, linkage disequilibrium, and biological relevance, for example, validation in wet lab or animal model experiments.

Determining the sufficiency of major SNPs involves considering factors like effect sizes, significance thresholds, multiple testing control, and the explained genetic variance. There's no fixed rule, but researchers often stop when the increase in predictive power becomes negligible.

## 4.3 Motivating Examples

The CANDELA Cohort Database is a valuable resource for genetic research and studies related to human populations. It comprises a diverse dataset collected from individuals of Latin American ancestry, particularly from Mexico, Colombia, Peru, Chile, and Brazil. The database contains comprehensive genetic, phenotypic, and environmental data, making it an excellent platform for studying various health-related traits and conditions. [Adhikari et al. (2016a, 2015); Ruiz-Linares et al. (2014)].

The classical GWAS model generates summary statistics based on equation (4.1) to summarize the results of association analysis between SNPs and the phenotypic trait of interest. These statistics typically include SNP identifiers, effect sizes (regression coefficients), standard errors, P-values, and allele frequencies. Summary statistics are often

publicly available in large-scale GWAS efforts, facilitating the utilization and integration of results across multiple studies.

The P-value is commonly used in GWAS to assess statistical significance. It quantifies the probability of observing the association between a SNP and the phenotype by chance alone. Manhattan plots are frequently employed in GWAS as a visual tool to depict the statistical significance of associations between genetic variants and a phenotypic trait across the entire genome. These plots utilize the negative logarithm of p-values (-log p) on the y-axis and allow for the identification of significant associations through peaks above the significance threshold. Manhattan plots are valuable for detecting the presence of multiple associated genetic variants and providing an overview of the genomic landscape of associations in GWAS studies [Paria et al. (2022)].

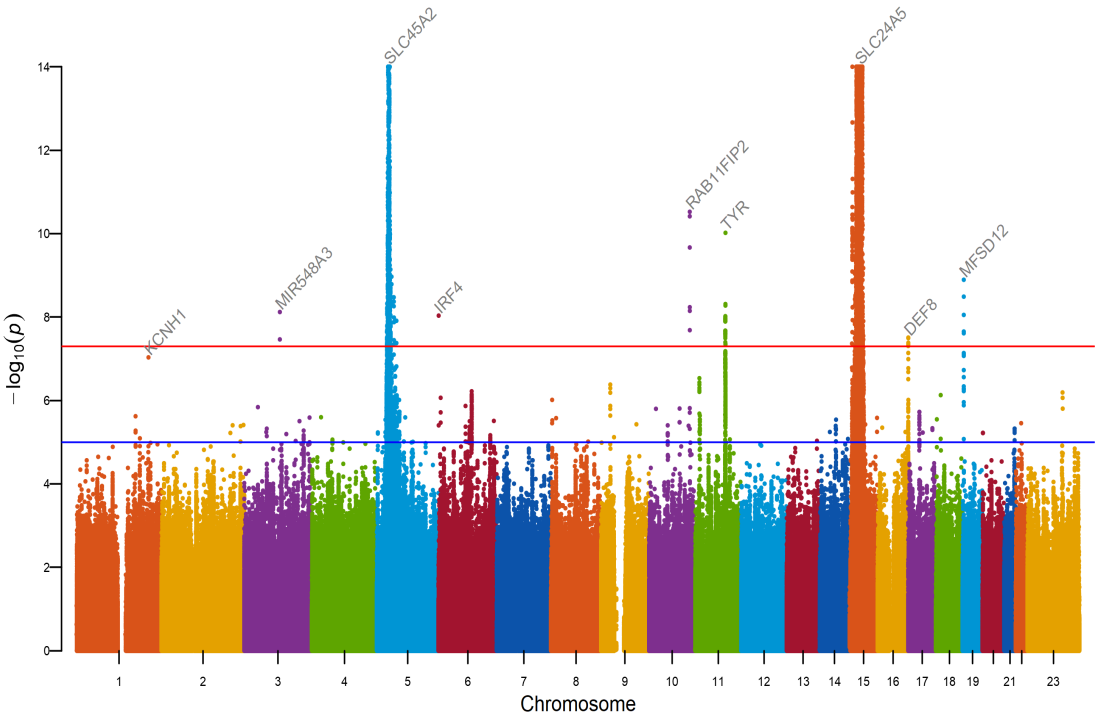



**Figure | Manhattan Plot: Displaying the significantly associated genes related to skin color through classical GWAS or unconditional models. Genome-wide suggestive:** $10^{-}5$**, blue line; Genome-wide significance:** $5 \times 10^{-}8$**, red line**

The above Manhattan plot indicates the P-values for corresponding SNPs which

were calculated by using the classical or unconditional GWAS model. Blue and red lines represent the genome-wide suggestive and genome-wide significance levels respectively. The P-values above these thresholds (genome-wide suggestive: $10^{-}5$, blue line; genome-wide significance: $5 \times 10^{-}8$, red line) suggest significant SNP association with skin color. The names of the candidate genes, which are closest to each association peak, are shown in the figure. Notably, among all the other associated genes, two genes such as $SLC45A2$ and $SLAC24A5$ exhibit the strongest association with skin color.

Using the CANDELA Cohort Database and equation (4.2), we can investigate the efficacy of conditioning techniques and assess the improvement in statistical power when identifying associations between genetic variants and traits like melanin. To demonstrate the concept of power gain through conditioning, I have extracted two sets of summary-level data from this cohort. The first dataset comprises unconditional "P-values" for approximately 9 million genetic variants. The second dataset includes "P-values" for genetic variants conditioned on known larger-effect variants, such as those present in genes $SLC45A2$ and $SLAC24A5$. These datasets provide valuable context and motivation for understanding the concept of power gain achieved through conditioning.

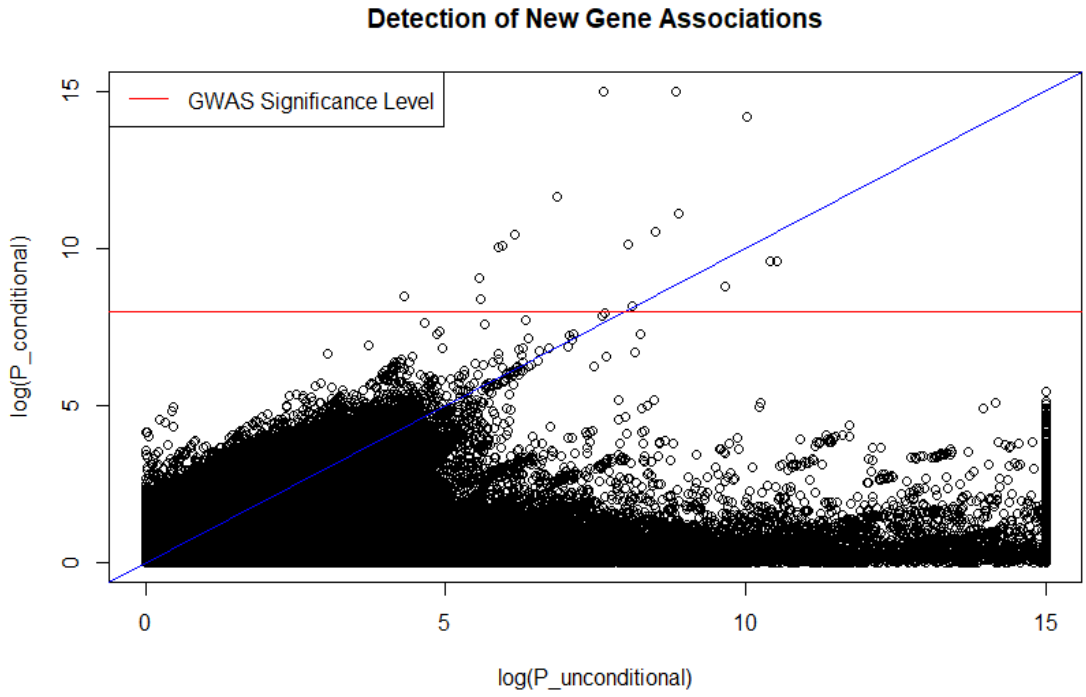



**Figure | Proof of Concept: Detection of new gene association imposing a condition on some major effect SNPs**

The conditional P-values for the genetic variants are represented by the points located above the blue diagonal line, indicating that certain SNPs exhibit increased significance after conditioning compared to their corresponding unconditional p-values. This obser-

vation implies that some genetic variants, which were initially less significant, become stronger through conditioning, as demonstrated by the points positioned above both the red and blue lines. Consequently, conditioning on well-known major genetic variants enables the identification of smaller-effect genetic variants with enhanced power. Furthermore, biological verification ensured that these newly associated genes were relevant for skin color [Adhikari et al. (2016a)]

## 4.4 Mathematical Derivations for 2-Variable LM

Let's consider a linear regression model as follows

$$y = \beta_w w + \epsilon \tag{4.3}$$

where $y$ is the response variable, $w$ corresponds to a genetic variant of interest extracted from $X$, and $\epsilon$ denotes the error term. For asymptotic distributions, assume that $\epsilon$ is normally distributed *i.e.*, $\epsilon \sim N(0, \sigma^2)$. Additionally, all variables in the model are mean standardized, implying that there are no intercept terms included.

The inclusion of additional covariates in the linear regression framework leads to a reduction in the mean square error (MSE) and standard error, consequently enhancing the precision of estimating and testing the regression coefficients of other covariates [Rao (2002)]. In line with the steps described by Rao (2002) and Chatterjee (2012), we introduce an additional variable, denoted as $z$, into equation (4.3) to demonstrate the increased power gained through conditioning on this additional covariate. The conditional linear regression model can be represented as follows:

$$y = \beta_w w + \beta_z z + \epsilon \tag{4.4}$$

Here, $\beta_w$ and $\beta_z$ represent the regression coefficients for variables $w$ and $z$, respec-

tively, while $\epsilon$ denotes the error term.

## 4.5 Derivation of Regression Coefficients

Let $\sigma_w^2$ and $\sigma_z^2$ represent the variances of $w$ and $z$, respectively. The covariance of $w$ and $z$ can be denoted as $\sigma_{wz}$, while their correlation is denoted as $r_{wz}$. Equation (4.4) can be equivalently expressed in matrix form as:

$$y = X\beta + \epsilon \tag{4.5}$$

where,

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X^T = \begin{bmatrix} w_1 & w_2 & \cdots & x_n \\ z_1 & z_2 & \cdots & z_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_w \\ \beta_z \end{bmatrix} \text{ and } \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The estimate of the regression coefficient, denoted as $\hat{\beta}$, represents an unbiased estimate of $\beta$ and can be obtained by performing the following calculation:

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T y \\ &= \begin{pmatrix} \sum w^2 & \sum wz \\ \sum zw & \sum z^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum wy \\ \sum zy \end{pmatrix} \\ &= \begin{pmatrix} \sigma_w^2 & \sigma_{wz} \\ \sigma_{zw} & \sigma_z^2 \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{yw} \\ \sigma_{yz} \end{pmatrix} \\ &= \frac{1}{\sigma_w^2 \sigma_z^2 - \sigma_{wz}^2} \begin{pmatrix} \sigma_z^2 \sigma_{yw} - \sigma_{wz}\sigma_{yz} \\ \sigma_w^2 \sigma_{yz} - \sigma_{wz}\sigma_{yw} \end{pmatrix} \end{aligned}$$

$$= \begin{pmatrix} \hat{\beta}_w \\ \hat{\beta}_z \end{pmatrix}$$

### 4.5.1 Derivation of $R^2$

The coefficient of determination ($R^2_{y.x_1,x_2,\cdots,x_k}$) represents the proportion of variation in the dependent variable $y$ that can be explained by the regressors $X$. It can be mathematically expressed as follows:

$$\begin{aligned}
R^2_{y.X} &= \frac{SSR}{SST} = \frac{y^T X\hat{\beta}}{y^T y} = \frac{y^T X(X^T X)^{-1}X^T y}{y^T y} = \frac{S_{yX}S_{XX}^{-1}S_{Xy}}{S_{yy}} \\
&= \frac{S_{yX}}{S_{yy}^{1/2}S_{XX}^{1/2}}\left(\frac{S_{XX}}{S_{XX}^{1/2}S_{XX}^{1/2}}\right)^{-1}\frac{S_{Xy}}{S_{XX}^{1/2}S_{yy}^{1/2}} = R_{yX}R_{XX}^{-1}R_{Xy} \\
&= \begin{pmatrix} r_{yw} & r_{yz} \end{pmatrix}\begin{pmatrix} 1 & r_{wz} \\ r_{zw} & 1 \end{pmatrix}^{-1}\begin{pmatrix} r_{yw} \\ r_{yz} \end{pmatrix} \\
&= \frac{r_{yw}^2 + r_{yz}^2 - 2r_{yw}r_{yz}r_{wz}}{(1-r_{wz}^2)} \\
&= \frac{r_{yw}^2 - r_{yw}^2 r_{wz}^2 + r_{yw}^2 r_{wz}^2 + r_{yz}^2 - 2r_{yw}r_{yz}r_{wz}}{(1-r_{wz}^2)} \\
&= r_{yw}^2 + \frac{(r_{yz} - r_{yw}r_{wz})^2}{(1-r_{wz}^2)} \\
&= r_{yw}^2 + \frac{(r_{yz} - r_{yw}r_{wz})^2}{(1-r_{wz}^2)(1-r_{yw}^2)}.(1-r_{yw}^2) \\
&= r_{yw}^2 + r_{yz.w}^2(1-r_{yw}^2) \text{ since } r_{yz.w} = \frac{r_{yz} - r_{yw}r_{wz}}{\sqrt{(1-r_{yw}^2)(1-r_{wz}^2)}}
\end{aligned}$$

Moreover, $R^2$ can also be expressed in terms of the test statistics $t^2$, specifically in the case of univariate linear regression. Considering the linear regression equation (4.3), the expression for $R^2_{y.w}$ is derived as follows [Proof in the appendix A.3]:

$$R^2_{y.w} = \frac{t^2}{t^2 + n}$$

Here, the numerator represents the squared test statistic $t^2$, while the denominator incorporates both the squared test statistic $t^2$ and the sample size $n$.

### 4.5.2 Derivation of MSE

The sum square of error (SSE) can also be simplified as:

$$
\begin{aligned}
SSE &= y^T y - y^T X (X^T X)^{-1} X^T y \\
&= n\sigma_y^2 - n \begin{pmatrix} \sigma_{wy} & \sigma_{zy} \end{pmatrix} \begin{pmatrix} \sigma_w^2 & \sigma_{wz} \\ \sigma_{zw} & \sigma_z^2 \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{yw} \\ \sigma_{yz} \end{pmatrix} \\
&= n\sigma_y^2 - \left( \frac{n}{\sigma_w^2 \sigma_z^2 - \sigma_{wz}^2} \right) \begin{pmatrix} \sigma_{wy} & \sigma_{zy} \end{pmatrix} \begin{pmatrix} \sigma_z^2 & -\sigma_{wz} \\ -\sigma_{zw} & \sigma_w^2 \end{pmatrix} \begin{pmatrix} \sigma_{yw} \\ \sigma_{yz} \end{pmatrix} \\
&= n\sigma_y^2 - \left( \frac{n}{\sigma_w^2 \sigma_z^2 - \sigma_{wz}^2} \right) \left[ \sigma_z^2 \sigma_{yw}^2 + \sigma_w^2 \sigma_{yz}^2 - 2\sigma_{yw}\sigma_{yz}\sigma_{wz} \right] \\
&= n\sigma_y^2 - n \left[ \frac{\sigma_z^2 \left( \frac{\sigma_{yw}}{\sigma_y \sigma_w} \right)^2 \sigma_w^2 \sigma_y^2 + \sigma_w^2 \left( \frac{\sigma_{yz}}{\sigma_y \sigma_z} \right)^2 \sigma_z^2 \sigma_y^2 - 2 \frac{\sigma_{yw}}{\sigma_w \sigma_y} \frac{\sigma_{yz}}{\sigma_z \sigma_y} \frac{\sigma_{wz}}{\sigma_w \sigma_z} \sigma_y^2 \sigma_w^2 \sigma_z^2}{\sigma_w^2 \sigma_z^2 \left( 1 - \left( \frac{\sigma_{wz}}{\sigma_w \sigma_z} \right)^2 \right)} \right] \\
&= n\sigma_y^2 - n \left[ \frac{\sigma_y^2 \sigma_w^2 \sigma_z^2 \left( r_{yw}^2 + r_{yz}^2 - 2 r_{yw} r_{yz} r_{wz} \right)}{\sigma_w^2 \sigma_z^2 (1 - r_{wz}^2)} \right] \\
&= n\sigma_y^2 \left[ 1 - \left( \frac{r_{yw}^2 + r_{yz}^2 - 2 r_{yw} r_{yz} r_{wz}}{1 - r_{wz}^2} \right) \right]
\end{aligned}
$$

So, the mean squared error (MSE) can be written as:

$$
\begin{aligned}
MSE = \hat{\sigma}^2 &= \frac{SSE}{n - k - 1} \\
&= \frac{n}{n - k - 1} \sigma_y^2 \left[ 1 - \left( \frac{r_{yw}^2 + r_{yz}^2 - 2 r_{yw} r_{yz} r_{wz}}{1 - r_{wz}^2} \right) \right] \\
&\cong \sigma_y^2 \left[ 1 - \left( \frac{r_{yw}^2 + r_{yz}^2 - 2 r_{yw} r_{yz} r_{wz}}{1 - r_{wz}^2} \right) \right] \\
&= \sigma_y^2 \left[ 1 - \left( \frac{r_{yw}^2 - r_{yw}^2 r_{wz}^2 + r_{yw}^2 r_{wz}^2 + r_{yz}^2 - 2 r_{yw} 2 r_{yz} r_{wz}}{(1 - r_{wz}^2)} \right) \right]
\end{aligned}
$$

$$
\begin{aligned}
&= \sigma_y^2 \left[1 - \left(r_{yw}^2 + \frac{(r_{yz} - r_{yw}r_{wz})^2}{(1 - r_{wz}^2)}\right)\right] \\
&= \sigma_y^2 \left[1 - \left(r_{yw}^2 + \frac{(r_{yz} - r_{yw}r_{wz})^2}{(1 - r_{wz}^2)(1 - r_{yw}^2)}.(1 - r_{yw}^2)\right)\right] \\
&= \sigma_y^2 \left[1 - \left(r_{yw}^2 + r_{yz.w}^2(1 - r_{yw}^2)\right)\right] \text{ since } r_{yz.w} = \frac{r_{yz} - r_{yw}r_{wz}}{\sqrt{(1 - r_{yw}^2)(1 - r_{wz}^2)}} \qquad (4.6) \\
&\therefore \text{MSE} \cong \sigma_y^2 \left[1 - \left(r_{yw}^2 + r_{yz.w}^2(1 - r_{yw}^2)\right)\right] = \sigma_y^2 \left[1 - R_{y.X}^2\right]
\end{aligned}
$$

### 4.5.3 Derivation of SE

The calculation for obtaining the asymptotic variance-covariance matrix of the regression coefficient ($\hat{\beta}$) is as follows:

$$
\begin{aligned}
Var(\hat{\beta}) &= \hat{\sigma}^2 (X^T X)^{-1} \\
&= \frac{\hat{\sigma}^2}{n(\sigma_w^2 \sigma_z^2 - \sigma_{wz}^2)} \begin{pmatrix} \sigma_z^2 & -\sigma_{wz} \\ -\sigma_{zw} & \sigma_w^2 \end{pmatrix} \quad \left[\text{here, } \hat{\sigma}^2 \cong \sigma_y^2 \left(1 - r_{yw}^2 - r_{yz.w}^2(1 - r_{yw}^2)\right)\right]
\end{aligned}
$$

So, the asymptotic distribution of the estimate of the regression coefficients can be written as

$$
\hat{\beta}_w \sim N\left(\frac{\sigma_z^2 \sigma_{yw} - \sigma_{wz}\sigma_{yz}}{\sigma_w^2 \sigma_z^2 - \sigma_{wz}^2}, \frac{\sigma_z^2}{n(\sigma_w^2 \sigma_z^2 - \sigma_{wz}^2)} \sigma_y^2 \left(1 - r_{yw}^2 - r_{yz.w}^2(1 - r_{yw}^2)\right)\right)
$$

$$
\hat{\beta}_z \sim N\left(\frac{\sigma_w^2 \sigma_{yz} - \sigma_{wz}\sigma_{yw}}{\sigma_w^2 \sigma_z^2 - \sigma_{wz}^2}, \frac{\sigma_w^2}{n(\sigma_w^2 \sigma_z^2 - \sigma_{wz}^2)} \sigma_y^2 \left(1 - r_{yz}^2 - r_{yw.z}^2(1 - r_{yz}^2)\right)\right)
$$

### 4.5.4 MSE when Covariates are Uncorrelated

Considering the regression model (4.4) under the assumption that the covariates $w$ and $z$ are uncorrelated, denoted by $\sigma_{wz} = 0$ or $r_{wz} = 0$, the mean square error (MSE) takes

on a specific form as derived in equation (4.6) as follow:

$$MSE = \hat{\sigma}^2 \cong \sigma_y^2 \left[1 - r_{yw}^2 - r_{yz}^2\right] \cong \sigma_y^2 \left[1 - R_{y.X}^2\right]$$

The asymptotic distribution of the estimate of the regression coefficients will be

$$\begin{aligned} \hat{\beta}_w &\sim N\left(\frac{\sigma_{wy}}{\sigma_w^2}, \frac{\sigma_y^2}{n\sigma_w^2}[1 - r_{wy}^2 - r_{zy}^2]\right) \\ \hat{\beta}_z &\sim N\left(\frac{\sigma_{zy}}{\sigma_z^2}, \frac{\sigma_y^2}{n\sigma_z^2}[1 - r_{zy}^2 - r_{wy}^2]\right) \end{aligned} \tag{4.7}$$

### 4.5.5 Proof of Concept: Gain of Power

Let's consider the scenario where we want to assess the significance of a specific SNP, denoted as $w$. We are interested in testing the hypothesis $H_0 : \beta_w = 0$ versus $H_a : \beta_w \neq 0$, but without loss of generality, we assume that $\beta_{alt}$ is positive.

The asymptotic variance of the estimated effect size, $\hat{\beta}_w$, is represented by $V$. Under the null hypothesis, we can express the probability as follows: $P\Big(|\frac{\hat{\beta}_w - 0}{\sqrt{V}}| > t_\alpha\Big) = \alpha$. Under the alternative hypothesis, the power of the test can be obtained as

$$\begin{aligned} Q &= P\Big(|\frac{\hat{\beta}_w - \beta_{alt}}{\sqrt{V}}| > t_\alpha\Big) \\ &= P\Big(\frac{\hat{\beta}_w - \beta_{alt}}{\sqrt{V}} < -t_\alpha\Big) + P\Big(\frac{\hat{\beta}_w - \beta_{alt}}{\sqrt{V}} > t_\alpha\Big) \\ &= P\Big(\frac{\hat{\beta}_w}{\sqrt{V}} < \frac{\beta_{alt}}{\sqrt{V}} - t_\alpha\Big) + P\Big(\frac{\hat{\beta}_w}{\sqrt{V}} > \frac{\beta_{alt}}{\sqrt{V}} + t_\alpha\Big) \\ &= \Phi\Big(\frac{\beta_{alt}}{\sqrt{V}} - t_\alpha\Big) + 1 - \Phi\Big(\frac{\beta_{alt}}{\sqrt{V}} + t_\alpha\Big) \\ &= \Phi\Big(\frac{\beta_{alt}}{\sqrt{V}} - t_\alpha\Big) + \Phi\Big(-\frac{\beta_{alt}}{\sqrt{V}} - t_\alpha\Big) \\ &\cong \Phi\Big(\frac{\beta_{alt}}{\sqrt{V}} - t_\alpha\Big) \end{aligned} \tag{4.8}$$

here, $\Phi$ indicates the cumulative distribution function (CDF) of the normal distribution

which is a monotonically increasing function. As the asymptotic variance $V$ decreases when an additional covariate is included in the model, the entire term within the $\Phi$ function becomes larger. Consequently, the corresponding term $Q$ increases in magnitude, resulting in a higher power of the statistical test.

Considering the unconditional regression model represented by equation (4.3), the Mean Squared Error (MSE) can be approximated as; $\hat{\sigma}^2 \cong \sigma_y^2 \left(1 - r_{wy}^2\right)$. Additionally, the asymptotic distribution of $\hat{\beta}_w$ follows a normal distribution given by

$$\hat{\beta}_w \sim N\left(\frac{\sigma_{wy}}{\sigma_w^2}, \frac{\sigma_y^2(1 - r_{wy}^2)}{n\sigma_w^2}\right) \tag{4.9}$$

Here, $\sigma_{wy}$ represents the covariance between the dependent variable $y$ and the independent variable $w$, $\sigma_w^2$ denotes the variance of variable $w$, $r_{wy}^2$ signifies the squared correlation coefficient between $w$ and $y$, and $n$ represents the sample size.

Equations (4.7) and (4.9) demonstrate that incorporating an additional covariate ($z$) along with the tested SNP leads to a smaller mean square error (MSE) for $\hat{\beta}_w$. The reduction in MSE is quantified as $r_{zy}^2$, which represents a fraction of the trait variance ($\sigma_y^2$). Consequently, the standard error of $\hat{\beta}_w$ for the model (4.7) becomes smaller, thereby enhancing the statistical power to test the effect of a new genetic variant.

These findings imply that incorporating additional covariates, even when uncorrelated, in a regression model, can enhance the precision of regression coefficients and increase statistical power by explaining a portion of the original trait variance. This mathematical justification supports the practice of conditioning on major single nucleotide polymorphisms (SNPs) in genome-wide association studies (GWAS). Notably, the estimation and testing of $\beta_w$ remain valid under conditioning, as $\hat{\beta}_w$ remains an unbiased and asymptotically normal estimator of $\beta_w$ in both scenarios.

Though GWAS assumes the simplifying assumption of uncorrelated covariates due to the index SNPs residing in separate linkage disequilibrium (LD) blocks, similar results

hold true for correlated covariates. In such cases, the gain in power also depends on the partial correlation between the covariates ($r^2_{yw.z}$) [Rao (2002)]. These findings highlight the importance of considering the effects of correlated covariates in GWAS and support the use of conditioning techniques to improve the accuracy and reliability of genetic association analyses.

The identification of new genetic variants with increased power is a desirable goal, but its feasibility depends on various factors such as the nature of the variants (e.g., effect sizes), the correlation among the variants, and the sample size. To understand the nature of changes in power when identifying a new variant, it is necessary to consider the following different modeling setups, particularly discussed in the context of two variant situations.

#### 4.5.5.1 Expression of Gain of Power when Fitting a Simple Regression Model but the True Model is a 2-Variable Model

Let's consider a scenario where two genetic variants, denoted as $w$ and $z$, are responsible for the variation in a phenotypic trait of interest, $y$. However, we mistakenly fit a null model that includes only one of these variants. The null model can be expressed as:

$$y_i = \beta^*_w w_i + \epsilon^*_i \tag{4.10}$$

In this equation, $\epsilon^*$ represents a random error term following a normal distribution, i.e., $\epsilon^* \sim N(0, \sigma^2)$. For simplicity, we assume a mean-standardized model without an intercept term. However, in reality, the phenotypic trait depends on both $w$ and $z$, i.e., $y_i \sim N(\beta_w w + \beta_z z, \sigma^2)$. By considering the wrong null model, the estimated effect of the tested genetic variant, denoted as $\hat{\beta}^*_w$, may be influenced by neighboring variants. We can express the effect size of a model in terms of the correlation coefficient as follows: $r_{yw} = \frac{\sigma_{yw}}{\sigma_y \sigma_w} = \beta_w \frac{\sigma_w}{\sigma_y}$ and the estimate of the considered regression model (4.10) can be

obtained as

$$\begin{aligned}
\hat{\beta}_w^* &= \frac{\sum w_i y_i}{\sum w_i^2} = \frac{\sum w_i(\beta_w w_i + \beta_z z_i + \epsilon_i)}{\sum w_i^2} \\
&= \beta_w + \beta_z \frac{\sum w_i z_i}{\sum w_i^2} + \frac{\sum w_i \epsilon_i}{\sum w_i^2} \\
\therefore E(\hat{\beta}_w^*) &= \beta_w + \beta_z \frac{\sum w_i z_i}{\sum w_i^2} = \beta_w + \beta_z \frac{\sigma_{wz}}{\sigma_w^2} \\
&= \beta_w + \beta_z \left(\frac{\sigma_{wz}}{\sigma_w \sigma_z}\right)\left(\frac{\sigma_z}{\sigma_w}\right) \\
&= \beta_w + \beta_z r_{wz} \left(\frac{\sigma_z}{\sigma_w}\right) \\
&= \beta_w + r_{wz}^2
\end{aligned}$$

That implies that the estimated effect size of the considered regression model is biased due to ignoring the effect of the related genetic variant ($z$). This will be unbiased only if the two variants are uncorrelated, i.e., $r_{wz}^2 = 0$. The variance of the estimated effect size can be obtained as $V(\hat{\beta}_w^*) = \frac{\hat{\sigma}^2}{\sum w_i^2}$; where $\hat{\sigma}^2 = MSE = \frac{SSE}{n-1}$. The sum square error (SSE) can be obtained as:

$$\begin{aligned}
SSE &= \sum (y_i - \hat{y})^2 = \sum (y_i - \hat{\beta}_w^* w)^2 \\
&= \sum (y_i^2 - 2\hat{\beta}_w^* y_i w_i + \hat{\beta}_{w^*}^2 w_i^2) \\
&= \sum y_i^2 - 2\hat{\beta}_w^* \sum w_i y_i + \hat{\beta}_{w^*}^2 \sum w_i^2 \\
&= \sum y_i^2 - 2\frac{\sum w_i y_i}{\sum w_i^2} \sum w_i y_i + \left(\frac{\sum w_i y_i}{\sum w_i^2}\right)^2 \sum w_i^2 \\
&= \sum y_i^2 - \frac{\left(\sum w_i y_i\right)^2}{\sum w_i^2} \\
&= n\sigma_y^2 - \frac{n^2 \sigma_{wy}^2}{n\sigma_w^2} \\
&= n\sigma_y^2 - n\left(\frac{\sigma_{wy}}{\sigma_w \sigma_y}\right)^2 \sigma_y^2 \\
&= n\sigma_y^2 - n r_{wy}^2 \sigma_y^2
\end{aligned}$$

$$
\begin{aligned}
&= n\sigma_y^2(1 - r_{wy}^2) \\
\therefore \hat{\sigma}^2 &= \frac{SSE}{(n-1)} = \frac{n\sigma_y^2(1 - r_{wy}^2)}{(n-1)} \approx \sigma_y^2(1 - r_{wy}^2) \\
\text{and } V(\hat{\beta}_w^*) &= \frac{\hat{\sigma}^2}{\sum w_i^2} = \frac{\sigma_y^2(1 - r_{wy}^2)}{n\sigma_w^2}
\end{aligned}
$$

That is, the asymptotic distribution of $\hat{\beta}_w^*$ follows as $\hat{\beta}_w^* \sim N\left(\beta_w + r_{wz}^2, \frac{\sigma_y^2(1-r_{wy}^2)}{n\sigma_w^2}\right)$. It indicates that the asymptotic mean of the effect size is biased, which is coming through the LD with the neighborhood variant (z), i.e., $r_{wz}^2$, and the asymptotic variance becomes as same as the single variable model. However, only the difference is that the variation of the response variable ($\sigma_y^2$) depends on both of the true effects ($w\&z$).

Neglecting the presence of another true effect (represented by z) in a statistical model can have a substantial impact on the statistical power of a hypothesis test. The reason is that the ignored SNP may have a substantial impact on the trait of interest, and by ignoring it, we eliminate crucial information that could assist in the detection of the SNP being tested. Assuming that the hypothesis is being tested for the effect described in the model (4.10) and the true $\beta$ is positive, then the power of the test can be expressed as follows:

$$
\begin{aligned}
\text{Power } Q^* &= \Phi\Big(\frac{\beta_{alt}}{\sqrt{V}^*} - t_\alpha\Big) \\
&= \Phi\left(\beta_{alt}\sqrt{\frac{n\sigma_w^2}{\sigma_y^2(1 - r_{wy}^2)}} - t_\alpha\right)
\end{aligned} \tag{4.11}
$$

Consider again the scenario where we are testing the hypothesis for a specific variant while simultaneously including another uncorrelated true variant in the model. Assuming that the true $\beta$ is positive, we can determine the power of the test by utilizing the

asymptotic distribution of $\hat{\beta}_w$ as discussed in equation (4.7) as follows:

$$\begin{aligned} \text{Power } Q &= \Phi\Big(\frac{\beta_{alt}}{\sqrt{V}} - t_\alpha\Big) \\ &= \Phi\left(\beta_{alt}\sqrt{\frac{n\sigma_w^2}{\sigma_y^2(1-r_{wy}^2-r_{yz}^2)}} - t_\alpha\right) \end{aligned} \tag{4.12}$$

The quantification of power gain can be determined by calculating the difference between two expressions of power using equations (4.11) and (4.12), as follows:

$$Q - Q^* = \Phi\left(\beta_{alt}\sqrt{\frac{n\sigma_w^2}{\sigma_y^2(1-r_{wy}^2-r_{yz}^2)}} - t_\alpha\right) - \Phi\left(\beta_{alt}\sqrt{\frac{n\sigma_w^2}{\sigma_y^2(1-r_{wy}^2)}} - t_\alpha\right) \tag{4.13}$$

When no LD situation holds, equation (4.13) demonstrates that the gain in power is influenced by various factors, including the correlation between the trait of interest and the genetic variants ($r_{yw}^2$ and $r_{yz}^2$), the variance of the trait ($\sigma_y^2$), the effect size of the variant, and the sample size. Notably, the correlation between the trait of interest and the conditional variant ($r_{yz}^2$) plays a crucial role in determining the magnitude of power gain. A higher correlation indicates a greater potential for power gain through conditioning.

However, for both cases described in equations (4.11) and (4.12), the power primarily depends on the effect size of the variant. In other words, variants with smaller effect sizes exhibit lower power. Eventually, as the effect size decreases, the power gain achieved through conditioning becomes negligible, requiring a substantial sample size to observe any significant power gain.

To further understand the nature of power gain resulting from conditioning, real-life genetic databases such as the CANDELA Cohort and UK Biobank have been used to provide empirical evidence and insights into the power gain phenomenon in the demonstration section.

#### 4.5.5.2 Validataion of the Model under the Null Scenario i.e., $\beta_w = 0, \beta_z \neq 0, r^2_{wz} \neq 0$

Suppose we are fitting a regression model with two genetic variants (w&z), where the variant $z$ is causal but $w$ is not. The regression model can be written as $y_i = \beta_w w_i + \beta_z z_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$.The estimate of the effect size $\beta_w$ can be obtained as

$$\begin{aligned}
\hat{\beta}_w &= \frac{\sum z_i^2 \sum w_i y_i - \sum w_i z_i \sum z_i y_i}{\sum w_i^2 \sum z_i^2 - (\sum w_i z_i)^2} \\
\text{and } E(\hat{\beta}_w) &= \frac{\sum z_i^2 \sum w_i(\beta_w w_i + \beta_z z_i) - \sum w_i z_i \sum z_i(\beta_w w_i + \beta_z z_i)}{\sum w_i^2 \sum z_i^2 - (\sum w_i z_i)^2} \\
&= \frac{\beta_w \left[\sum z_i^2 \sum w_i^2 - (\sum w_i z_i)^2\right]}{\sum w_i^2 \sum z_i^2 - (\sum w_i z_i)^2} \\
&= \beta_w \\
&= 0 \quad \text{[Under the Null i.e., } \beta_w = 0\text{]}
\end{aligned}$$

This implies that, under the null model ($H_0 : \beta_w = 0$), the mean of the estimated effect size ($\hat{\beta}_w$) is zero, even when the causal variant ($z$) is accounted for in the model. However, as we discussed in the previous section, if a model is incorrectly treated as a null model, the mean of the variant effect size ($\hat{\beta}^*_w$) becomes biased i.e., $E(\hat{\beta}^*_w) = \beta_w + r^2_{wz} \neq 0$, where $r^2_{wz}$ represents the linkage disequilibrium (LD) between the two variants.

In the scenario where the two variants are uncorrelated (i.e., $r^2_{wz} = 0$), the null model suggests that the mean effect size of the variant will be zero, representing an ideal situation.

However, if the two variants exhibit linkage disequilibrium (LD) with each other, indicated by a non-zero squared LD coefficient ($r^2_{wz} \neq 0$), the null model will no longer have a mean effect size of zero ($E(\hat{\beta}_w) \neq 0$). This is because there is a possibility of an LD effect originating from the neighboring variant. Consequently, if we have a

proxy variant located near the causal variant, which itself does not directly impact the trait, conditioning on the causal variant will reveal that the proxy variant is merely a result of LD and not causal itself. This scenario frequently occurs in genome-wide association studies (GWAS), where a causal variant may be mistakenly ungenotyped, and the association picks that show significant associations may be entirely due to a proxy variant.

When a fine-mapping approach investigates variant causality, it is necessary to incorporate the functional annotations of the genetic variant as well as understand the biological mechanisms underlying the association signals identified in GWAS [Wang and Huang (2022)]. So, in the case of evaluating a genetic variant's causality with a trait, the aforementioned mathematical model is not practically useful. However, if the objective is to assess the effect of a specific variant ($w$) by conditioning on a potential causal variant ($z$), our mathematical derivations align with the standard conventions of genome-wide association studies (GWAS).

#### 4.5.5.3 Validation of the Model when the Conditional Variant has no Effect (i.e., $\beta_z = 0$)

Let us consider a regression model with two genetic variants where the effect size of the conditional genetic variant $z$ has no impact on the phenotypic trait $y$ i.e., $\beta_z = 0$ or $r_{yz} = 0$ (since $\beta_z = r_{yz}\frac{\sigma_y}{\sigma_z}$). Referring to the asymptotic distribution of effect sizes for a regression model with two genetic variants discussed in section (4.5.3), we can determine the asymptotic mean and variance of the other variant $w$ as follows:

$$\begin{aligned}\hat{\beta}_w &= \frac{\sigma_z^2\sigma_{yw} - \sigma_{wz}\sigma_{yz}}{\sigma_w^2\sigma_z^2 - \sigma_{wz}^2} \\ &= \frac{\sigma_{yw} - \sigma_{wz}\beta_z}{\sigma_w^2 - \sigma_{wz}^2\frac{\beta_z}{\sigma_{yz}}}\end{aligned}$$

$$
\begin{aligned}
&= \frac{\sigma_{yw}}{\sigma_w^2} \\
\text{and } V(\hat{\beta}_w) &= \frac{\sigma_z^2 \sigma_y^2}{n(\sigma_w^2 \sigma_z^2 - \sigma_{wz}^2)} \left[1 - \left(\frac{r_{yw}^2 + r_{yz}^2 - 2r_{yw}r_{yz}r_{wz}}{1 - r_{wz}^2}\right)\right] \\
&= \frac{\sigma_y^2}{n\sigma_w^2(1 - r_{wz}^2)} \left[1 - \left(\frac{r_{yw}^2 + r_{yz}^2 - 2r_{yw}r_{yz}r_{wz}}{1 - r_{wz}^2}\right)\right] \\
&= \frac{\sigma_y^2}{n\sigma_w^2(1 - r_{wz}^2)} \left[1 - \frac{r_{yw}^2}{(1 - r_{wz}^2)}\right] \\
&= \frac{\sigma_y^2}{n\sigma_w^2} \left[\frac{1 - r_{yw}^2 - r_{wz}^2}{(1 - r_{wz}^2)^2}\right]
\end{aligned}
$$

Setting $\beta_z = 0$ in the 2-variable regression model leads to specific simplifications. Notably, the asymptotic mean of the effect size, denoted as $\hat{\beta}_w$, becomes the same as that of the single variable model. However, there is a difference in the asymptotic variance, as both the numerator and the denominator have reduced amounts of non-negative fractions

When there is no linkage disequilibrium (LD) or correlation between the two variables, i.e., $r_{wz}^2 = 0$, the asymptotic variance of $\hat{\beta}_w$ becomes exactly the same as in the single variable model, that is, $v(\hat{\beta}_w) = \frac{\sigma_y^2(1-r_{yw}^2)}{n\sigma_w^2}$.

However, when the variants are in linkage disequilibrium (LD) or are correlated, the extent of change in the asymptotic variance can be expressed as follows:

$$
\begin{aligned}
V(\hat{\beta}_w) - V(\hat{b}_w) &= \frac{\sigma_y^2}{n\sigma_w^2} \left[\frac{1 - r_{yw}^2 - r_{wz}^2}{(1 - r_{wz}^2)^2} - (1 - r_{yw}^2)\right] \\
&= \frac{\sigma_y^2}{n\sigma_w^2((1 - r_{wz}^2)^2)} \left(1 - r_{yw}^2 - r_{wz}^2 - 1 + r_{yw}^2 + 2r_{wz}^2 - 2r_{wz}^2 r_{yw}^2 - r_{wz}^4 + r_{wz}^4 r_{yw}^2\right) \\
&= \frac{\sigma_y^2}{n\sigma_w^2((1 - r_{wz}^2)^2)} \left(r_{wz}^2 - r_{wz}^4 - 2r_{wz}^2 r_{yw}^2 + r_{yw}^2 r_{wz}^4\right)
\end{aligned}
$$

So, even if the conditioning variant (SNP) is not associated with the trait of interest ($\beta_z = 0$), it may still exhibit linkage disequilibrium (LD) with the target SNP, influencing the asymptotic variance. Considering that the asymptotic variance plays a crucial role

in determining statistical power, caution should be made when interpreting the results if the variants are correlated. The presence of LD between the variants can impact the precision and reliability of the effect size estimates, which in turn affects the ability to detect true associations between genetic variants and the trait.

## 4.6 Mathematical Derivations for Multiple Regression Model

A joint multi-SNP model refers to a statistical framework where a quantitative trait is influenced by multiple genetic variants. Considering the collective impact of multiple SNPs (Single Nucleotide Polymorphisms), a joint multi-SNP model can be defined as

$$y = X\beta + \epsilon \tag{4.14}$$

where, $y$ is an $(n \times 1)$ vector of phenotypes of size $n$, $X$ is an $(n \times m)$ genotype matrix, and $\beta$ is the $(m \times 1)$ vector of joint SNP effects. The joint estimates of this model can be obtained as $\hat{\beta} = (X^T X)^{-1} X^T y$ and the variance-covariance matrix will be $Var(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1}$, where, $\hat{\sigma}^2$ is the mean square error (MSE) which can be obtained as

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{SSE}{n-m-1} \\
&= \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{n-m-1} \\
&= \frac{y^T y - y^T X\hat{\beta} - \hat{\beta}^T X^T y + \hat{\beta}^T X^T X\hat{\beta}}{n-m-1} \\
&= \frac{y^T y - y^T X\hat{\beta} - \hat{\beta}^T X^T y + \hat{\beta}^T X^T y}{n-m-1} \\
&= \frac{y^T y - y^T X(X^T X)^{-1} X^T y}{n-m-1} \\
&= \frac{y^T (I - H) y}{n-m-1} \quad \text{[here, H=}X(X^T X)^{-1} X^T\text{, is a hat matrix ]}
\end{aligned}$$

For the joint model, the multiple correlation coefficient ($R^2_{y.x_1,x_2,\cdots,x_m}$) can be ex-

pressed as

$$\begin{aligned} R^2_{y.x_1,x_2,\cdots,x_m} &= \frac{SSR}{SST} \\ &= \frac{y^T X\hat{\beta}}{y^T y} \\ &= \frac{y^T X(X^T X)^{-1} X^T y}{y^T y} \\ &= \frac{S_{yX} S_{XX}^{-1} S_{Xy}}{S_{yy}} \\ &= \frac{S_{yX}}{S_{yy}^{1/2} S_{XX}^{1/2}} \left( \frac{S_{XX}}{S_{XX}^{1/2} S_{XX}^{1/2}} \right)^{-1} \frac{S_{Xy}}{S_{XX}^{1/2} S_{yy}^{1/2}} \\ &= R_{yX} R_{XX}^{-1} R_{Xy} \end{aligned}$$

## 4.7 Single-SNP Model vs. Joint-SNP Model

Genome-wide association studies (GWAS) usually test the association between phenotypic traits and genetic variants, taking each SNP separately based on a single-SNP model as

$$y = x_j b_j + \epsilon \; ; \; j = 1, 2, \cdots, m. \tag{4.15}$$

where, $x_j$ is the $j$th genetic variant and $b_j$ is the marginal effect for the SNP $j$. The marginal effects for all the genetic variants from a single SNP-based model can be expressed in a matrix form as

$$\begin{aligned} \hat{b} &= D^{-1} X^T y \\ &= \left( \frac{D}{n} \right)^{-1} \frac{1}{n} X^T y \\ &= V_d^{-1} \frac{1}{n} X^T y \\ \therefore V_d \hat{b} &= \frac{1}{n} X^T y \end{aligned}$$

where, $\hat{b}$ is a $m \times 1$ vector of marginal SNP effects and $D$ is the diagonal matrix of $X^T X$. Suppose, $V$ is a covariance matrix of $X$ i.e., $V = \frac{1}{n} X^T X$ and $V_d$ is a diagonal variance matrix which is defined as $V_d = \frac{D}{n}$, then it can be shown that the covariance matrix can be interchanged with the diagonal matrix as $V = \sqrt{V_d} R \sqrt{V_d}$. Based on this expressions, the joint SNP effect $\hat{\beta}$ can be obtained from the marginal SNP effects ($\hat{b}$) as

$$\begin{aligned}
\hat{\beta} = (X^T X)^{-1} X^T y &= \left(\frac{1}{n} X^T X\right)^{-1} \left(\frac{1}{n} X^T Y\right) = V^{-1} \frac{1}{n} X^T Y \\
&= V_d^{-1/2} R^{-1} V_d^{-1/2} \left(\frac{1}{n} X^T Y\right) = V_d^{-1/2} R^{-1} V_d^{-1/2} V_d \hat{b} \\
&= V_d^{-1/2} R^{-1} V_d^{1/2} \hat{b}
\end{aligned}$$

Yang et al. (2012) expressed the estimate of the single SNP model as $D\hat{b} = X^T y$ and demonstrated the connection with the multi-SNP estimate as $\hat{\beta} = (X^T X)^{-1} D\hat{b}$.

## 4.8 2-Block LM

In genome-wide association studies (GWAS), numerous SNPs have been found to be associated with variations in different phenotypes. However, certain phenotypes, such as human pigmentation traits, are primarily influenced by a few key SNPs that explain a significant proportion of the phenotypic variation. To enhance the power of identifying new SNPs associated with these traits, it is possible to consider these major SNPs as a block and condition on them.

In this section, I approached the design matrix of a multiple regression model as a two-block matrix. The first block consists of the tested SNPs, while the second block includes additional covariates such as conditioning SNPs, genetic principal compo-

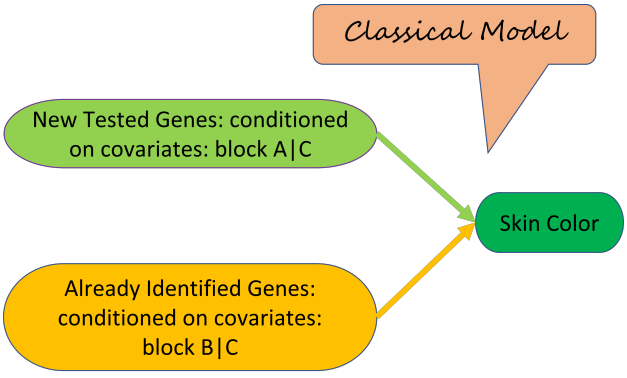





**Diagram: Classical GWAS Model for Genetic Association**

nents (PCs), or non-genetic covariates like age and sex. In classical GWAS, the genetic data is typically adjusted using genetic PCs before performing univariate regression on each SNP individually. Consequently, the major SNPs in the conditioning block have already been adjusted.

However, when conducting joint or conditional analysis, it becomes crucial to adjust the tested SNPs to maintain uniformity with the conditioning block. Interestingly, in their demonstration of conditional analysis for two-block matrix scenarios, Yang et al. (2012) neglected this adjustment despite using PC-adjusted summary statistics to identify conditional effects.

### 4.8.1 Expression of Regression Coefficeints

Suppose, a multiple linear regression setup, where the design matrix, $X$ consists of 2 blocks such as

$$X = (\underbrace{X_1, X_2, \cdots, X_k}_{X_A}, \underbrace{X_{k+1}, \cdots, X_m}_{X_B} = (X_A, X_B)$$

$$\Sigma_{Xy} = \frac{1}{n}\begin{bmatrix} X_A^T y \\ X_B^T y \end{bmatrix} = \begin{bmatrix} \Sigma_{Ay} \\ \Sigma_{By} \end{bmatrix}$$

$$\Sigma_{XX} = \frac{1}{n}\begin{bmatrix} X_A^T X_A & X_A^T X_B \\ X_B^T X_A & X_B^T X_B \end{bmatrix} = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}$$

here, $X_A$ is a block of tested SNPs and $X_B$ is a set of any other covariates such as conditioning SNPs, genetic PCs, or other covariates say, age, sex, etc. $\Sigma_{AA}$ and $\Sigma_{BB}$ are invertible, and $\Sigma_{AB}^T = \Sigma_{BA}$. The inverse of the covariance matrix, when the design matrix has two blocks, i.e., $\Sigma_{XX}^{-1}$ can be found with the help of the following formula

[Lu and Shiou (2002)]

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & 0 \\ 0 & (D - CA^{-1}B)^{-1} \end{bmatrix} \begin{bmatrix} I & -BD^{-1} \\ -CA^{-1} & I \end{bmatrix}$$

$$
\begin{aligned}
\Sigma_{XX}^{-1} &= \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}^{-1} \\
&= \begin{bmatrix} (\Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA})^{-1} & 0 \\ 0 & (\Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB})^{-1} \end{bmatrix} \begin{bmatrix} I & -\Sigma_{AB}\Sigma_{BB}^{-1} \\ -\Sigma_{BA}\Sigma_{AA}^{-1} & I \end{bmatrix} \\
&= n \begin{bmatrix} (X_A^T X_A - X_A^T X_B (X_B^T X_B)^{-1} X_B^T X_A)^{-1} & 0 \\ 0 & (X_B^T X_B - X_B^T X_A (X_A^T X_A)^{-1} X_A^T X_B)^{-1} \end{bmatrix} \begin{bmatrix} I & -X_A^T X_B (X_B^T X_B)^{-1} \\ -X_B^T X_A (X_A^T X_A)^{-1} & I \end{bmatrix} \\
&= n \begin{bmatrix} (X_A^T X_A - X_A^T H_B X_A)^{-1} & 0 \\ 0 & (X_B^T X_B - X_B^T H_A X_B)^{-1} \end{bmatrix} \begin{bmatrix} I & -X_A^T X_B (X_B^T X_B)^{-1} \\ -X_B^T X_A (X_A^T X_A)^{-1} & I \end{bmatrix} \text{ [where, } H_B = X_B (X_B^T X_B)^{-1} X_B^T \text{]} \\
&= n \begin{bmatrix} (X_A^T (I - H_B) X_A)^{-1} & 0 \\ 0 & (X_B^T (I - H_A) X_B)^{-1} \end{bmatrix} \begin{bmatrix} I & -X_A^T X_B (X_B^T X_B)^{-1} \\ -X_B^T X_A (X_A^T X_A)^{-1} & I \end{bmatrix} \\
&= n \begin{bmatrix} \left(X_A^T (I - H_B) X_A\right)^{-1} & -\left(X_A^T (I - H_B) X_A\right)^{-1} X_A^T X_B (X_B^T X_B)^{-1} \\ -\left(X_B^T (I - H_A) X_B\right)^{-1} X_B^T X_A (X_A^T X_A)^{-1} & \left(X_B^T (I - H_A) X_B\right)^{-1} \end{bmatrix}
\end{aligned}
$$

$$
\hat{\beta} = \Sigma_{XX}^{-1} \Sigma_{Xy}
$$

$$
\begin{aligned}
&= \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{Ay} \\ \Sigma_{By} \end{bmatrix} \\
&= n \begin{bmatrix} \left(X_A^T(I-H_B)X_A\right)^{-1} & -\left(X_A^T(I-H_B)X_A\right)^{-1} X_A^T X_B (X_B^T X_B)^{-1} \\ -\left(X_B^T(I-H_A)X_B\right)^{-1} X_B^T X_A (X_A^T X_A)^{-1} & \left(X_B^T(I-H_A)X_B\right)^{-1} \end{bmatrix} . \frac{1}{n} \begin{bmatrix} X_A^T y \\ X_B^T y \end{bmatrix} \\
&= \begin{bmatrix} \left(X_A^T(I-H_B)X_A\right)^{-1} X_A^T y - \left(X_A^T(I-H_B)X_A\right)^{-1} X_A^T X_B (X_B^T X_B)^{-1} X_B^T y \\ \left(X_B^T(I-H_A)X_B\right)^{-1} X_B^T y - \left(X_B^T(I-H_A)X_B\right)^{-1} X_B^T X_A (X_A^T X_A)^{-1} X_A^T y \end{bmatrix} \\
&= \begin{bmatrix} \left(X_A^T(I-H_B)X_A\right)^{-1} X_A^T y - \left(X_A^T(I-H_B)X_A\right)^{-1} X_A^T H_B y \\ \left(X_B^T(I-H_A)X_B\right)^{-1} X_B^T y - \left(X_B^T(I-H_A)X_B\right)^{-1} X_B^T H_A y \end{bmatrix} \\
&= \begin{bmatrix} \left(X_A^T(I-H_B)X_A\right)^{-1} X_A^T (I-H_B) y \\ \left(X_B^T(I-H_A)X_B\right)^{-1} X_B^T (I-H_A) y \end{bmatrix} \\
&= \begin{bmatrix} \hat{\beta}_A^T \\ \hat{\beta}_B^T \end{bmatrix}
\end{aligned}
$$

Note that, in the 2-block case, the estimate for the tested SNPs block, $\hat{\beta}_A^T = \left(X_A^T(I-H_B)X_A\right)^{-1} X_A^T(I-H_B)y$ is obtained by conditioning on the second block which contains already explored major SNPs.

### 4.8.2 Expression of Regression Coefficients in terms of Residuals

The effects of the conditioning model can also be examined by analyzing the residuals of regression models. In the scenario where the response variable ($y$) and the tested SNPs block ($X_A$) are regressed on the conditioning block ($X_B$), the outcomes of the tested SNPs block in the 2-block case can be considered as linear models using the residuals.

Suppose, $\tilde{y}_{|B} = (I-H_B)y$ and $\tilde{X}_{A|B} = (I-H_B)X_A$ are the residuals, if the response ($y$) and tested SNPs block ($X_A$) are regressed on conditioning block ($X_B$); where $H_B = X_B(X_B^TX_B)^{-1}X_B^T$ and it is idempotent. The estimates of regression coefficients, in the 2-block case, can be expressed in terms of residuals as

$$
\begin{aligned}
\hat{\beta} &= \Sigma_{XX}^{-1}\Sigma_{Xy} \\
&= \begin{bmatrix} \left(X_A^T(I-H_B)X_A\right)^{-1} X_A^T(I-H_B)y \\ \left(X_B^T(I-H_A)X_B\right)^{-1} X_B^T(I-H_A)y \end{bmatrix} \quad [\text{Using, } (I-H_B)^T(I-H_B) = (I-H_B)] \\
&= \begin{bmatrix} \left(X_A^T(I-H_B)^T(I-H_B)X_A\right)^{-1} X_A^T(I-H_B)^T(I-H_B)y \\ \left(X_B^T(I-H_A)^T(I-H_A)X_B\right)^{-1} X_B^T(I-H_A)^T(I-H_A)y \end{bmatrix} \\
&= \begin{bmatrix} \left(((I-H_B)X_A)^T(I-H_B)X_A)\right)^{-1} ((I-H_B)X_A)^T(I-H_B)y \\ \left(((I-H_A)X_B)^T(I-H_A)X_B)\right)^{-1} ((I-H_A)X_B)^T(I-H_A)y \end{bmatrix} \\
&= \begin{bmatrix} (\tilde{X}_{A|B}^T\tilde{X}_{A|B})^{-1}\tilde{X}_{A|B}^T\tilde{y}_{|B} \\ (\tilde{X}_{B|A}^T\tilde{X}_{B|A})^{-1}\tilde{X}_{B|A}^T\tilde{y}_{|A} \end{bmatrix}
\end{aligned}
$$

### 4.8.3 Derivation of Estimates arising from Yang et al. (2012)

Yang et al. (2012) derived the conditional estimates, considering a design matrix with having two sets of covariates such as $X_A$ and $X_B$, but not as a blocked matrix. The normal equations for this setup of multivariate regression can be written as

$$X_A^T X_A \hat{\beta}_A + X_A^T X_B \hat{\beta}_B = X_A^T y$$
$$X_B^T X_A \hat{\beta}_A + X_B^T X_B \hat{\beta}_B = X_B^T y$$

They showed that the estimate for the first block ($X_A$) can be obtained by conditioning on the second block ($X_B$) as follows

$$\hat{\beta}_{A|B} = (X_A^T X_A)^{-1} X_A^T y - (X_A^T X_A)^{-1} X_A^T X_B \hat{\beta}_B \tag{4.16}$$

Yang et al. (2012) also expressed the above conditional estimate in terms of marginal SNP effects as

$$\begin{aligned}
\hat{\beta}_{A|B} &= (X_A^T X_A)^{-1} X_A^T y - (X_A^T X_A)^{-1} X_A^T X_B \hat{\beta}_B \\
&= (X_A^T X_A)^{-1} X_A^T y - (X_A^T X_A)^{-1} X_A^T X_B (X_B^T X_B)^{-1} X_B^T y \\
&= (X_A^T X_A)^{-1} X_A^T y - (X_A^T X_A)^{-1} X_A^T X_B (X_B^T X_B)^{-1} D_B \hat{b}_B \\
&= (X_A^T X_A)^{-1} D_A \hat{b}_A - (X_A^T X_A)^{-1} X_A^T X_B (X_B^T X_B)^{-1} D_B \hat{b}_B
\end{aligned} \tag{4.17}$$

where, both $\hat{\beta}_A$ and $\hat{\beta}_B$ are the estimates of the joint multivariate regression model and $\hat{b}_B$ is the univariate regression estimates for the set of covariates on which condition is applied. If the first set, $X_A$ contains just one variable $(say, X_1)$, then the equation (4.17) can be written as

$$\hat{\beta}_{A|B} = (X_1^T X_1)^{-1} X_1^T y - (X_1^T X_1)^{-1} X_1^T X_B (X_B^T X_B)^{-1} D_B \hat{b}_B$$

$$= \hat{b}_1 - (X_1^T X_1)^{-1} X_1^T X_B (X_B^T X_B)^{-1} D_B \hat{b}_B \tag{4.18}$$

### 4.8.4 MSE in Multiple Regression Model

Suppose the two blocks of the design matrix contain $m_1$ and $m_2$ number of covariates in each block, respectively and it has been shown in the previous section that the coefficient of regression for the first block conditioning on the second block is $\hat{\beta}_{A|B} = \left(\tilde{X}_{A|B}^T \tilde{X}_{A|B}^T\right)^{-1} \tilde{X}_{A|B}^T \tilde{y}_{|B}$ and the variance of this estimate will be $\text{Var}(\hat{\beta}_{A|B}) = \hat{\tilde{\sigma}}^2 \left(\tilde{X}_{A|B}^T \tilde{X}_{A|B}^T\right)^{-1}$, where, $\hat{\tilde{\sigma}}^2$ is the mean square error (MSE) which can be obtained as

$$MSE = \hat{\tilde{\sigma}}^2 = \frac{SSE}{n - m_1 - m_2} = \frac{y^T y - y^T H y}{n - m_1 - m_2}$$

The term $y^T H y$ can be found as

$$
\begin{aligned}
H &= X(X^TX)^{-1}X^T \\
&= \frac{1}{n}X(\frac{1}{n}X^TX)^{-1}X^T \\
&= \frac{1}{n}X\Sigma_{XX}^{-1}X^T \\
&= \frac{1}{n}\begin{bmatrix} X_A & X_B \end{bmatrix} n \begin{bmatrix} \left(X_A^T(I-H_B)X_A\right)^{-1} & -\left(X_A^T(I-H_B)X_A\right)^{-1}X_A^TX_B(X_B^TX_B)^{-1} \\ -\left(X_B^T(I-H_A)X_B\right)^{-1}X_B^TX_A(X_A^TX_A)^{-1} & \left(X_B^T(I-H_A)X_B\right)^{-1} \end{bmatrix} \begin{bmatrix} X_A^T \\ X_B^T \end{bmatrix} \\
&= X_A\left(X_A^T(I-H_B)X_A\right)^{-1}X_A^T - X_B\left(X_B^T(I-H_A)X_B\right)^{-1}X_B^TX_A(X_A^TX_A)^{-1}X_A^T \\
&+ X_B\left(X_B^T(I-H_A)X_B\right)^{-1}X_B^T - X_A\left(X_A^T(I-H_B)X_A\right)^{-1}X_A^TX_B(X_B^TX_B)^{-1}X_B^T \\
&= X_A\left(X_A^T(I-H_B)X_A\right)^{-1}X_A^T - X_B\left(X_B^T(I-H_A)X_B\right)^{-1}X_B^TH_A \\
&+ X_B\left(X_B^T(I-H_A)X_B\right)^{-1}X_B^T - X_A\left(X_A^T(I-H_B)X_A\right)^{-1}X_A^TH_B \\
&= X_A\left(X_A^T(I-H_B)X_A\right)^{-1}X_A^T(I-H_B) + X_B\left(X_B^T(I-H_A)X_B\right)^{-1}X_B^T(I-H_A) \\
&= X_A\left(X_A^T(I-H_B)^T(I-H_B)X_A\right)^{-1}X_A^T(I-H_B)^T(I-H_B) + X_B\left(X_B^T(I-H_A)^T(I-H_A)X_B\right)^{-1}X_B^T(I-H_A)^T(I-H_A) \\
&= X_A\left(\tilde{X}_{A|B}^T\tilde{X}_{A|B}\right)^{-1}\tilde{X}_{A|B}^T(I-H_B) + X_B\left(\tilde{X}_{B|A}^T\tilde{X}_{B|A}\right)^{-1}\tilde{X}_{B|A}^T(I-H_A)
\end{aligned}
$$

$$
\begin{aligned}
y^T H y &= y^T X_A \left(\tilde{X}_{A|B}^T \tilde{X}_{A|B}\right)^{-1} \tilde{X}_{A|B}^T (I - H_B) y + y^T X_B \left(\tilde{X}_{B|A}^T \tilde{X}_{B|A}\right)^{-1} \tilde{X}_{B|A}^T (I - H_A) y \\
&= y^T X_A \left(\tilde{X}_{A|B}^T \tilde{X}_{A|B}\right)^{-1} \tilde{X}_{A|B}^T \tilde{y}_{|B} + y^T X_B \left(\tilde{X}_{B|A}^T \tilde{X}_{B|A}\right)^{-1} \tilde{X}_{B|A}^T \tilde{y}_{|A} \\
&= y^T X_A \hat{\beta}_A + y^T X_B \hat{\beta}_B \\
&= y^T X_A \left[(X_A^T X_A)^{-1} X_A^T y - (X_A^T X_A)^{-1} X_A^T X_B \hat{\beta}_B\right] + y^T X_B \hat{\beta}_B \text{ [Using equation (4.16)]} \\
&= y^T X_A (X_A^T X_A)^{-1} X_A^T y - y^T X_A (X_A^T X_A)^{-1} X_A^T X_B \hat{\beta}_B + y^T X_B \hat{\beta}_B \\
&= y^T H_A y - y^T H_A X_B \hat{\beta}_B + y^T X_B \hat{\beta}_B \\
&= y^T H_A y + y^T (I - H_A) X_B \hat{\beta}_B \\
&= y^T H_A y + y^T (I - H_A)^T (I - H_A) X_B \left(\tilde{X}_{B|A}^T \tilde{X}_{B|A}\right)^{-1} \tilde{X}_{B|A}^T \tilde{y}_{|A} \\
&= y^T H_A y + \tilde{y}_{|A}^T \tilde{X}_{B|A} \left(\tilde{X}_{B|A}^T \tilde{X}_{B|A}\right)^{-1} \tilde{X}_{B|A}^T \tilde{y}_{|A} \\
&= y^T H_A y + \tilde{y}_{|A}^T \tilde{H}_{B|A} \tilde{y}_{|A} \qquad (4.19)
\end{aligned}
$$

In GWAS, it is common practice to exclude individuals who exhibit close genetic relationships with all autosomal single nucleotide polymorphisms (SNPs). This is done to focus on causal variants that are not in linkage disequilibrium (LD) with other genetic variations [Yang et al. (2011)].

When there is no LD, the individuals are considered genetically pairwise unrelated, and this allows for simplifications in the mathematical expressions. Specifically, in the case where the blocks of the design matrix are uncorrelated, denoted as $X_A^T X_B = 0$, the expression for $y^T H y$ can be simplified as follows.

$$
\begin{aligned}
y^T H y &= y^T H_A y + \tilde{y}_{|A}^T \tilde{H}_{B|A} \tilde{y}_{|A} \\
&= y^T H_A y + y^T (I - H_A)^T (I - H_A) X_B \left(X_B^T (I - H_A) X_B\right)^{-1} X_B^T (I - H_A)^T (I - H_A) y \\
&= y^T H_A y + y^T (I - H_A) X_B \left(X_B^T X_B - X_B^T H_A X_B\right)^{-1} X_B^T (I - H_A) y
\end{aligned}
$$

$$= y^T H_A y + y^T X_B \left(X_B^T X_B\right)^{-1} X_B^T y = y^T H_A y + y^T H_B y \tag{4.20}$$

### 4.8.5 Gain of Power in Multiple Regression Model Setting

Considering a multivariate regression case, where the design matrix has only one block say, $X_A$, then the regression coefficient of the model will be $\hat{\beta}_A = (X_A^T X_A)^{-1} X_A^T y$ and the variance of the estimate, $\text{Var}(\hat{\beta}_A) = \hat{\sigma}^2 (X_A^T X_A)^{-1}$, where, $\hat{\sigma}^2 = MSE = \frac{SSE(A)}{n - m_1 - 1}$ and $SSE(A) = y^T y - y^T H_A y = y^T y - y^T X_A (X_A^T X_A)^{-1} X_A^T y$.

If the design matrix contains an additional block of a matrix, say, $X_B$, then the regression coefficient of $X_A$ can be obtained by conditioning on the other block $X_B$, which has been discussed in subsection (3.8.2) as$\hat{\beta}_{A|B} = \left(\tilde{X}_{A|B}^T \tilde{X}_{A|B}^T\right)^{-1} \tilde{X}_{A|B}^T \tilde{y}_{|B}$ and the the variance can be obtained as, $\text{Var}(\hat{\beta}_{A|B}) = \hat{\tilde{\sigma}}^2 \left(\tilde{X}_{A|B}^T \tilde{X}_{A|B}^T\right)^{-1}$, where, $\tilde{X}_{A|B}^T \tilde{X}_{A|B}^T = X_A^T (I - H_B) X_A$ and $\hat{\tilde{\sigma}}^2$ is the mean square error (MSE) of the conditioned model, and the sum square error of the conditoned model, $SSE(A|B)$, can be shown as

$$\begin{aligned}
SSE(A|B) &= y^T y - y^T H y \\
&= y^T y - y^T H_A y - \tilde{y}_{|A}^T \tilde{H}_{B|A} \tilde{y}_{|A} \quad \text{[Using the equation (4.19)]} \\
&= SSE(A) - \tilde{y}_{|A}^T \tilde{H}_{B|A} \tilde{y}_{|A} \\
&= SSE(A) - \text{a non-negative term} \\
\therefore\ SSE(A|B) &\leq SSE(A)
\end{aligned}$$

Eventually, $\hat{\tilde{\sigma}}^2 \leq \hat{\sigma}^2$

The above mathematical simplification indicates that the error sum of square gets a reduction by the amount denoted as $\tilde{y_{|A}^T} \tilde{H_{B|A}} \tilde{y}_{|A}$, which is a non-negative term. This reduction is achieved by conditioning on an additional set of covariates, represented by $X_B$. As the standard error or mean square error (MSE) has an inverse relationship with statistical power and precision, the decrease in MSE leads to an improvement in the

power to detect new SNPs. Therefore, this reduction in MSE enhances the ability to discover and identify novel genetic variants with greater power.

When the two blocks of the design matrix are not correlated, i.e., $X_A^T X_B = 0$, then the above simplifications can be shown as $\hat{\beta}_{A|B} = \left(\tilde{X}_{A|B}^T \tilde{X}_{A|B}^T\right)^{-1} \tilde{X}_{A|B}^T \tilde{y}_{|B} = \hat{\beta}_A$ and the the variance, $\text{Var}(\hat{\beta}_{A|B}) = \hat{\tilde{\sigma}}^2 \left(\tilde{X}_{A|B}^T \tilde{X}_{A|B}^T\right)^{-1} = \hat{\tilde{\sigma}}^2 (X_A^T X_A)^{-1}$, and the sum square error, $SSE(A|B)$, can be shown as

$$
\begin{aligned}
SSE(A|B) &= y^T y - y^T H y \\
&= y^T y - y^T H_A y - \tilde{y}_{|A}^T \tilde{H}_{B|A} \tilde{y}_{|A} \\
&= y^T y - y^T H_A y - y^T H_B y \ \ \text{[Using the equation (4.20)]} \\
&= SSE(A) - \text{a non-negative term} \\
\therefore \ SSE(A|B) &\leq SSE(A)
\end{aligned}
$$

Eventually, $\hat{\tilde{\sigma}}^2 \leq \hat{\sigma}^2$

The inequality presented above demonstrates that even if the two blocks of the design matrix are uncorrelated, incorporating an extra block of covariates or conditioning on an additional set of covariates in the model can result in a reduction in the mean square error. This reduction, denoted as $y^T H_B y$, is always a non-negative term, indicating that it will inevitably enhance statistical power.

## 4.9 3-Block LM

In the classical Genome-Wide Association Study (GWAS), it is common practice to adjust the genotype data by incorporating genetic principal components (PCs) and other non-genetic covariates like age and sex. This adjustment helps to account for population structure and reduce variations that could arise due to such factors. Following the adjustment, a single-SNP modeling approach is typically employed to generate GWAS

summary statistics.

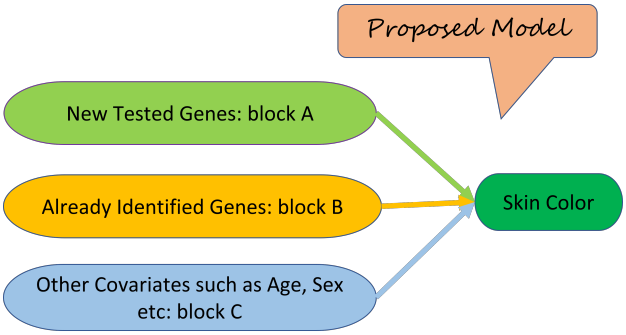



**Diagram: Proposed Conditional Model for Genetic Association**

To address the impact of PCs adjustment in conditional models, I presented a mathematical simplification for a multiple regression model. I split the design matrix of a multiple regression model into three blocks: the tested SNPs block, a set of major SNPs on which conditioning will be performed, and a covariates block that may include genetic PCs and non-genetic covariates like age and sex.

To compute the joint or conditional effects from the derived mathematical expressions, GWAS summary statistics are necessary, which are usually calculated using genetic databases that are covariates adjusted. Additionally, a reference sample with individual-level genotype data is required to calculate the correlations, and this dataset also needs to be appropriately adjusted with the covariates.

Suppose, the design matrix is partitioned into three blocks such as $X_A, X_B$, and $X_C$, and the 3-block joint linear model may be expressed as

$$y = X_A\beta_A + X_B\beta_B + X_C\beta_C + \epsilon \tag{4.21}$$

The summary statistics published in GWAS typically involve calculations for each SNP individually, while conditioning on genetic principal components (PCs). To account for these PCs, we can express the adjusted model for the two blocks of genetic variants

as follows:

$$y = X_A b_A + X_C b_C + \epsilon$$

$$y = X_B b_B + X_C b_C + \epsilon$$

The above two models can be thought of as a 2-block linear model, and the regression coefficient of $X_A$ conditioning on the covariates block, $X_C$, can be expressed as

$$\begin{aligned} \hat{b}_{A|C} &= \left(\tilde{X}_{A|C}^T \tilde{X}_{A|C}\right)^{-1} \tilde{X}_{A|C}^T \tilde{y}_{|C} \ ; \quad \text{where,} \quad \tilde{X}_{A|C} = (I - H_C) X_A \\ &= D_{A|C}^{-1} \tilde{X}_{A|C}^T \tilde{y}_{|C} \qquad\qquad\qquad\qquad H_C = X_C (X_C^T X_C)^{-1} X_C^T \\ \therefore D_{A|C} \hat{b}_{A|C} &= \tilde{X}_{A|C}^T \tilde{y}_{|C} \qquad\qquad\qquad\qquad D_{A|C} = \tilde{X}_{A|C}^T \tilde{X}_{A|C} = X_A^T (I - H_C) X_A \end{aligned}$$

Similarly, the regression coefficient of $X_B$ conditioning on the covariates block, $X_C$ can be expressed as

$$\begin{aligned} \hat{b}_{B|C} &= \left(\tilde{X}_{A|C}^T \tilde{X}_{A|C}\right)^{-1} \tilde{X}_{A|C}^T \tilde{y}_{|C} \\ &= D_{B|C}^{-1} \tilde{X}_{A|C}^T \tilde{y}_{|C} \\ \therefore D_{B|C} \hat{b}_{B|C} &= \tilde{X}_{B|C}^T \tilde{y}_{|C} \end{aligned}$$

The above regression coefficients, $\hat{b}_{A|C}$ and $\hat{b}_{B|C}$, are the univariate regression coefficients conditioning on the covariates block, and they are usually available in the GWAS published results, $D_{A|C}$ and $D_{B|C}$ are the diagonal of $(\tilde{X}_{A|C}^T \tilde{X}_{A|C})$ and $(\tilde{X}_{B|C}^T \tilde{X}_{B|C})$ respectively and $\tilde{y}_{|C}$ is the residual of the regression model after regressing the response $y$ on $X_C$.

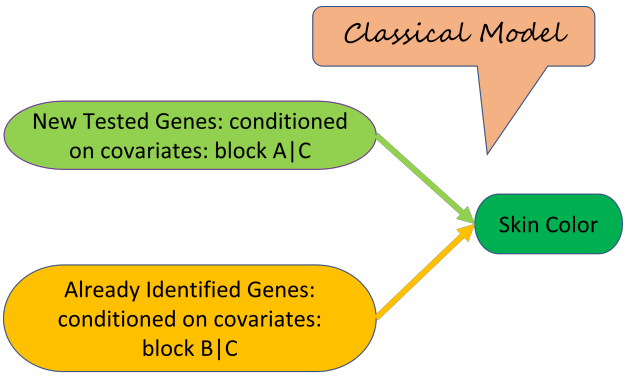



**Diagram: Converting a 3-Block Linear Model to 2-Block by adjusting for the Covariates**

### 4.9.1 Expression of Model Estimates through Residuals

As the genotyped data is usually adjusted with PCs first, so the joint regression model (4.13) is equivalent to the expression of the following model

$$\tilde{y}_{|C} = \tilde{X}_{A|C}\beta_{A|C} + \tilde{X}_{B|C}\beta_{B|C} + \epsilon \tag{4.22}$$

where, $\tilde{X}_{A|C}$ and $\tilde{X}_{B|C}$ are the residuals of the models when both $X_A$ and $X_B$ are regressed on $X_C$.

Now, the regression equation (4.22) can be considered as a two-block situation, and the regression coefficient $\hat{\beta}_{A|C}$, can be obtained as equivalently as follow

$$\hat{\beta}_{A|BC} = \left(\tilde{X}^T_{A|BC}\tilde{X}_{A|BC}\right)^{-1}\tilde{X}^T_{A|BC}\tilde{y}_{|BC}$$

here, $\tilde{X}_{A|BC} = (I - H_{B|C})\tilde{X}_{A|C}$ and $\tilde{y}_{|BC} = (I - H_{B|C})\tilde{y}_{|C}$ are the residuals of the models after regressing both $\tilde{X}_{A|C}$ and $\tilde{y}_{|C}$ on $\tilde{X}_{B|C}$ respectively, and $H_{B|C} = \tilde{X}_{B|C}\left(\tilde{X}^T_{B|C}\tilde{X}_{B|C}\right)^{-1}\tilde{X}^T_{B|C}$.

The expression of $\hat{\beta}_{A|BC}$ can be simplified as

$$\therefore \hat{\beta}_{A|BC} = \left(\tilde{X}^T_{A|BC}\tilde{X}_{A|BC}\right)^{-1}\tilde{X}^T_{A|BC}\tilde{y}_{|BC}$$

$$
\begin{aligned}
&= \left(\tilde{X}_{A|C}^T (I - H_{B|C}) \tilde{X}_{A|C}\right)^{-1} \tilde{X}_{A|C}^T (I - H_{B|C}) \tilde{y}_{|C} \\
&= M^{-1} \tilde{X}_{A|C}^T (I - H_{B|C}) \tilde{y}_{|C} \\
&= M^{-1} \tilde{X}_{A|C}^T \tilde{y}_{|C} - M^{-1} \tilde{X}_{A|C}^T H_{B|C} \tilde{y}_{|C} \\
&= M^{-1} \underbrace{\tilde{X}_{A|C}^T \tilde{y}_{|C}}_{D_{A|C}\hat{b}_{A|C}} - M^{-1} \tilde{X}_{A|C}^T \tilde{X}_{B|C} \left(\tilde{X}_{B|C}^T \tilde{X}_{B|C}\right)^{-1} \underbrace{\tilde{X}_{B|C}^T \tilde{y}_{|C}}_{D_{B|C}\hat{b}_{B|C}} \\
&= M^{-1} D_{A|C} \hat{b}_{A|C} - M^{-1} \tilde{X}_{A|C}^T \tilde{X}_{B|C} \left(\tilde{X}_{B|C}^T \tilde{X}_{B|C}\right)^{-1} D_{B|C} \hat{b}_{B|C} \qquad (4.23)
\end{aligned}
$$

where, $M = \tilde{X}_{A|C}^T (I - H_{B|C}) \tilde{X}_{A|C}$

Yang et al. (2012) demonstrated a conditional analysis that involves a design matrix with two sets of covariates. The first set consists of test SNPs, while the second set comprises conditioning SNPs. The conditional estimate is expressed in terms of univariate GWAS summary statistics, which was discussed in section (4.8.3) as:

$$\hat{\beta}_{A|B} = (X_A^T X_A)^{-1} D_A \hat{b}_A - (X_A^T X_A)^{-1} X_A^T X_B (X_B^T X_B)^{-1} D_B \hat{b}_B$$

Here, $\hat{\beta}_{A|B}$ represents the conditional estimate of the joint multiple regression model. The summary statistics $\hat{b}_A$ and $\hat{b}_B$ are obtained from corresponding univariate regression models and can be found in the online Gentic databases.

It is worth highlighting that, Yang et al. (2012) expressed the joint estimates or conditional estimates using univariate GWAS summary statistics, typically adjusted with genetic principal components (PCs). However, they did not account for the influence of genetic PCs on the set of tested SNPs.

### 4.9.2 Conditional Coefficients with Summary Statistics: Comparison between 3-Block Approach and Yang's GCTA Approach

In this thesis, I have calculated the conditional beta coefficients and assessed joint associations based on the proposed 3-block approach and the GCTA software (introduced by Yang et al. (2011)). Both methods relied on summary-level statistics from genome-wide association studies (GWAS) and estimates of linkage disequilibrium (LD) from a reference sample with individual-level genotype data. I compared the performance of calculating these coefficients against true values obtained from the original raw data or individual-level genotype data.

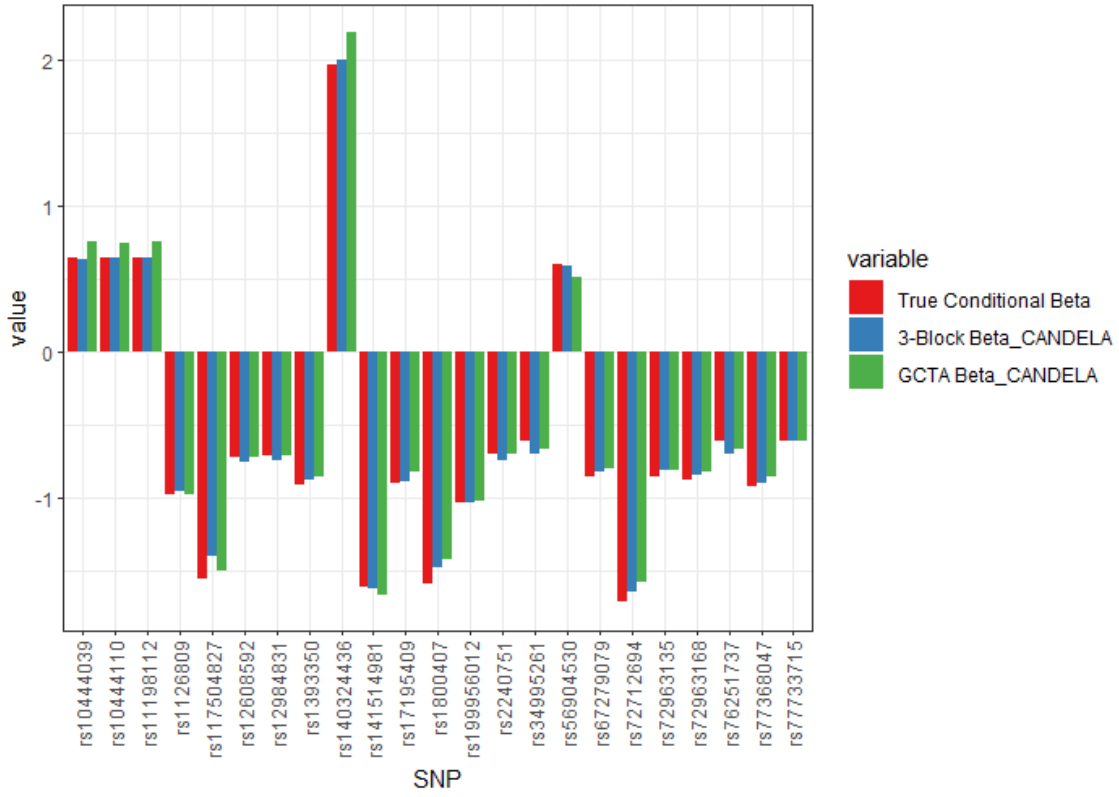



**Figure: Conditional beta coefficients were calculated by the 3-Block approach and GCTA software, and compared with the true conditional coefficients.**

The reference sample was obtained from the CANDELA cohort database, consisting of information on 30 single nucleotide polymorphisms (SNPs). Among these, 7 SNPs

were used as conditional SNPs, while the remaining 23 SNPs were tested. To validate the results, we also employed the PLINK software to generate unconditional beta coefficients and conditional coefficients based on the same set of 7 SNPs. The entire analysis was conducted using the CANDELA Cohort database.

Using the reference sample SNP information and the unconditional beta coefficients (summary-level statistics), I calculated the conditional beta coefficients using both the 3-block approach and the GCTA software. The accompanying picture illustrates the true conditional beta coefficients alongside the values obtained through the 3-block approach and GCTA software for each of the 23 tested SNPs. The image clearly demonstrates that the conditional coefficients obtained through the 3-block approach closely align with the true coefficients for all tested SNPs, outperforming the coefficients derived from the GCTA software.

## 4.10 Implementation

In order to address the potential confounding effects of population structure, genetic studies including GWAS often incorporate principal components (PCs) derived from the genotype covariance matrix, as well as other non-genetic factors such as age and sex, as fixed effects in their models. The results of these analyses are then published as summary-level statistics for GWAS.

To calculate conditional effects using the available summary-level statistics, it is most realistic to consider the genetic data matrix as a 3-block structure. This entails dividing the data into three blocks: the tested SNP block, the conditioning SNP block, and the block containing other covariates such as genetic PCs, age, and sex. By considering this 3-block structure, we can appropriately account for the interplay between the tested SNPs, the conditioning SNPs, and the relevant covariates, enabling the estimation of conditional effects.

The mathematical derivation for conditional effects, based on the 3-block multiple regression model, has been presented in equation (4.23). To proceed with this derivation, information is required from the available GWAS unconditional summary level statistics regarding the effects of major affected SNPs on which the conditioning will be performed. Since individual-level genotype data are typically not accessible for the entire cohort due to privacy concerns, additional information such as genetic variant correlations, represented by $R^2$ matrices, can be computed from a reference cohort that is publicly available and exhibits similar ethnicity characteristics.

If the raw data or individual-level genotype data are accessible, conditional GWAS can be performed using the PLINK software [Purcell et al. (2007)]. Additionally, GCTA-COJO [Yang et al. (2012)] can be employed to calculate conditional GWAS by utilizing PC-adjusted summary-level statistics from GWAS, a list of major conditioning SNPs, and a reference cohort with comparable ethnic background. These tools facilitate the analysis of genetic associations while accounting for the effects of specific conditioning SNPs on the traits of interest.

It is important to note that the GCTA-COJO software utilizes PC-adjusted summary statistics, but it does not incorporate adjustment of the reference cohort using genetic principal components (PCs). As a result, the conditional GWAS results obtained from GCTA-COJO may differ from the true conditioned GWAS results, particularly in cases where individual-level data are accessible.

To bridge this gap, the concept of the 3-block regression model, which considers PC-adjusted summary level statistics and a PC-adjusted reference cohort, can be employed. By utilizing the derived formulas for conditional effects in the 3-block regression model, it becomes possible to enhance the accuracy of the conditional GWAS results and address the limitation posed by GCTA-COJO's lack of adjustment for genetic PCs in the reference cohort.

## 4.11 Demonstration

The statistical power of testing the effect of a genetic variant typically increases with larger sample sizes. This increase in power is due to the reduction in standard error, improved precision of estimates, and reduced impact of random variation on the results [William G. Cochran (1977)]. However, when examining the power gain resulting from conditioning on major variants, equation (4.13) indicates that the gain may initially increase but stops after reaching a peak. Subsequently, it will decrease and eventually diminish to zero for a specific sample size. The sample size at which this occurs largely depends on various factors, such as the effect size and correlation coefficients of both the target SNP and the conditioned SNP with the trait of interest.

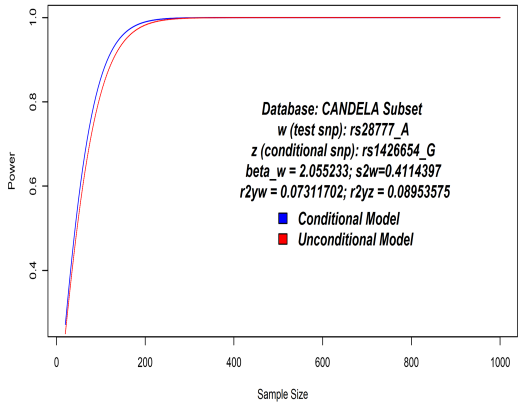
In this study, I have examined the statistical power of a specific SNP in two scenarios: when it is included uniformly in the regression model and when it is incorporated into the model while conditioning on one or more major SNPs. I have analyzed the power curve and the power difference resulting from conditioning. To simplify the analysis, we have considered three SNPs adjusted with principal components (PCs): $rs28777\_A$, $rs7118677\_T$, and $rs2240751\_G$. Additionally, we have selected $rs1426654\_G$ as the conditional SNP due to its stronger association with the trait of interest, namely skin color. These SNPs were chosen from different chromosomes (CHR: 5, 11, 19, and 15, respectively) to ensure no linkage disequilibrium (LD) among them. The analysis draws upon real-life genetic databases, including the CANDELA Cohort and UK Biobank, to provide empirical insights into the nature of power changes resulting from conditioning.

### 4.11.1 Analysis with CANDELA Cohort Database

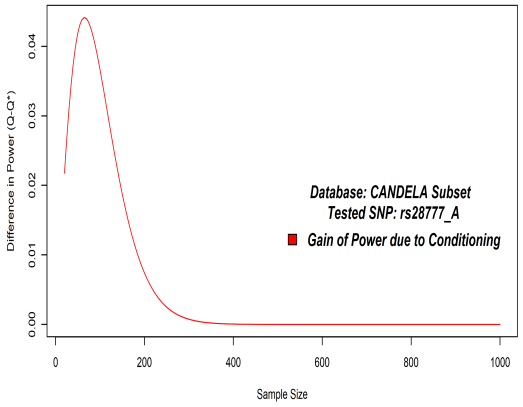
In the CANDELA cohort subset, the correlation coefficients ($r^2$) between the melanin trait (the trait of interest) and the SNPs $rs28777\_A$, $rs7118677\_T$, $rs2240751\_G$, and $rs1426654\_G$ are 0.073, 0.0081, 0.007, and 0.089, respectively. Among these SNPs,

$rs1426654\_G$ has the highest correlation with melanin and is selected as the conditional SNP. Multiple conditional regression models have been performed separately for each of the three SNPs, resulting in respective effect sizes of 2.03, 0.633, and 0.798.

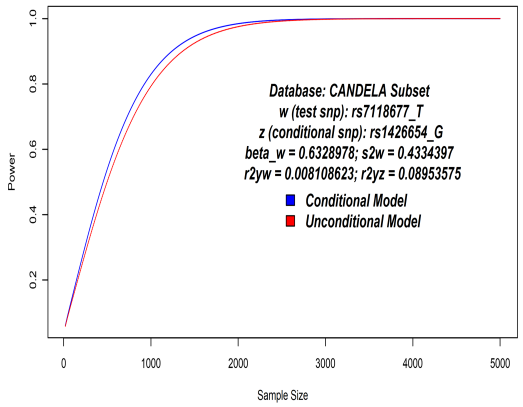
The power curve for each of the three variants, along with the power gain due to conditioning (based on equation 4.13), has shown in figures (a) to (f). These figures illustrate that the power gain initially reaches its peak and then diminishes as the effect of conditioning varies with smaller sample sizes. The effect sizes of the variants play a role in determining the sample size required to observe changes in power gain. For instance, the SNP $rs28777\_A$ with a larger effect size requires approximately 300 sample sizes to diminish the effect of conditioning. On the other hand, the other two SNPs require around 4000 sample sizes due to their smaller effect sizes and a weaker association with melanin.



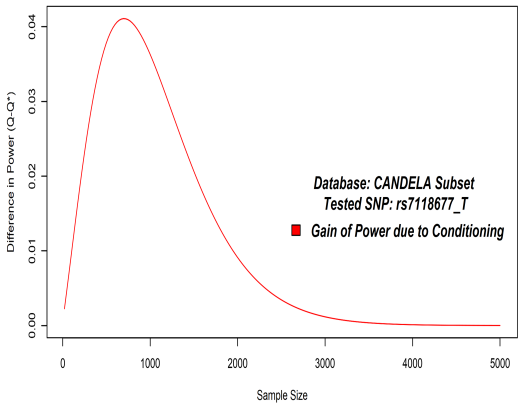

(a) Power Comparision due to conditioning; Database: CANDELA Subset; Tested SNP: rs28777_A



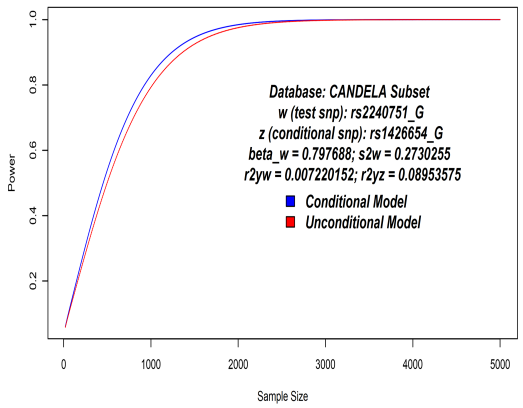

(b) Nature of Gain of Power with different sample sizes



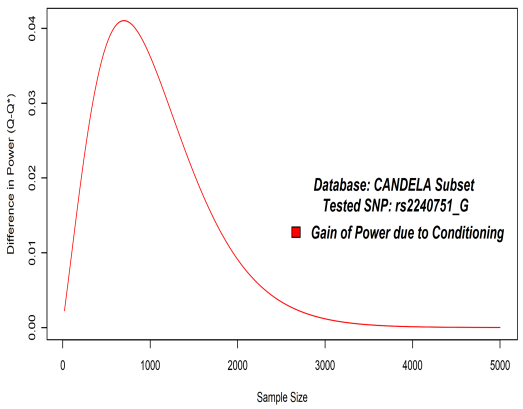

(c) Power Comparision due to conditioning; Database: CANDELA Subset; Tested SNP: rs7118677_T




(d) Nature of Gain of Power with different sample sizes




(e) Power Comparision due to conditioning; Database: CANDELA Subset; Tested SNP: rs2240751_G




(f) Nature of Gain of Power with different sample sizes

### 4.11.2 Analysis with UK Biobank Database

The UK Biobank is one of the world's largest biobanks where various phenotypic information and biological samples have been collected for each of the approximately 500,000 individuals from across the United Kingdom, aged between 40 and 69 at recruitment [Bycroft et al. (2018)]. Various genetic studies have been conducted based on these databases to explore the genetic architecture of many complex traits. For example, human skin and hair color are visible traits that can vary dramatically within and across ethnic populations. Many genetic variants have already been identified by analyzing the UK Biobank databases.

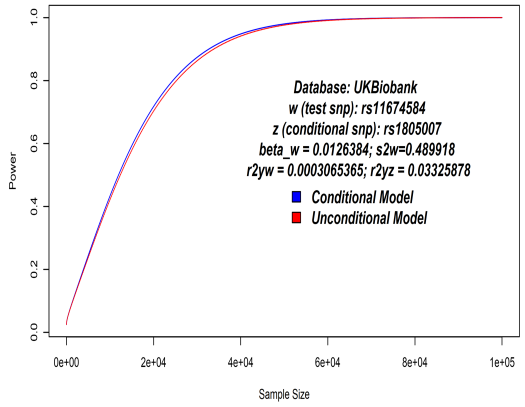
To validate the concept of enhancing the power to detect newly associated genetic variants by conditioning on major SNPs, a specific subset of individuals from the UK Biobank databases was utilized. The conditional analysis was performed using classical genome-wide association study (GWAS) models implemented in PLINK [Purcell et al. (2007)] software.

I considered three tested SNPs: rs11674584, rs4689317, and rs79844047, originating from chromosomes 2, 4, and 18, respectively. The variant rs1805007 (CHR:16) was chosen as a conditional SNP due to its strongest association ($r^2_{yz}$) with the trait of interest, skin color. The correlation coefficients between the tested SNPs and the melanin trait are 0.0003065365, 0.00007807342, and 0.00007920819, respectively, while the conditional variant exhibits a coefficient of 0.03325878, justifying its selection as a major variant. The effect sizes of the tested SNPs are observed as 0.0126384, 0.0125183, and 0.0113362, respectively. Simultaneously, the variances of these variants are 0.489918, 0.150552, and 0.15535407 respectively. Moreover, the phenotype (skin color) in the UK Biobank is a categorical variable with three categories and the variance (var(y)) is calculated from the frequency distribution, which is 0.2520505.

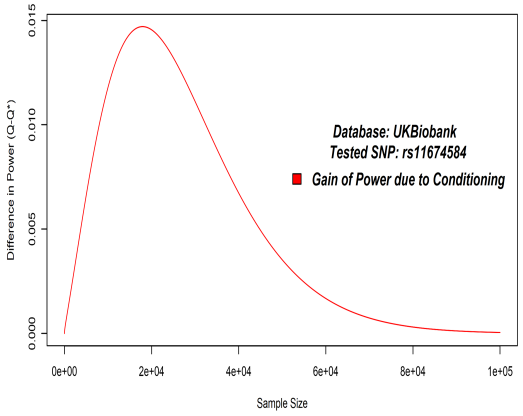
The power curve for these variants is presented in figures (g) to (l). In reviewing

the summary statistics of the tested SNPs, there are negligible differences observed in terms of correlation coefficients and effect sizes. However, a notable distinction arises in their variance, with rs11674584 demonstrating the highest variability in comparison to the other two SNPs.

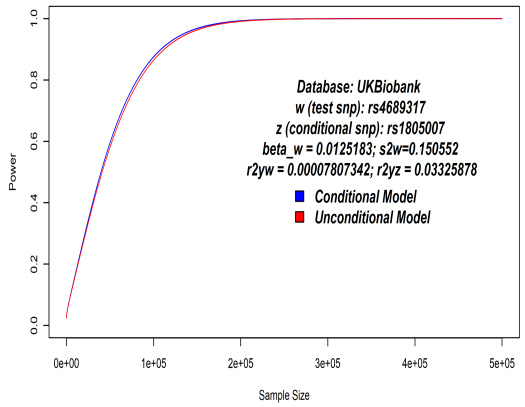
The power curves for these tested SNPs reveal that larger sample sizes are necessary to detect the highest difference in power resulting from conditioning. Notably, SNP rs11674584 showcases the most substantial power gain at a sample size of 20,000, while the other two SNPs necessitate almost 5 times larger sample sizes (100,000) to observe a similarly impactful difference in power.



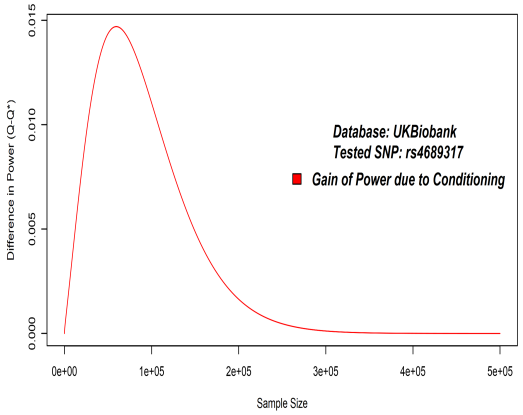

(g) Power Comparision due to conditioning; Database: UK Biobank Subset; Tested SNP: rs11674584



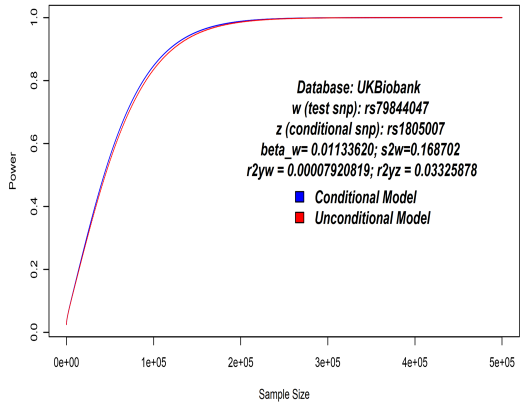

(h) Nature of Gain of Power with different sample sizes



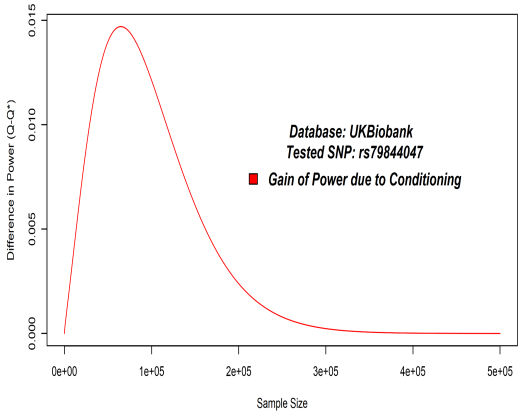

(i) Power Comparision due to conditioning; Database: UK Biobank Subset; Tested SNP: rs4689317




(j) Nature of Gain of Power with different sample sizes




(k) Power Comparision due to conditioning; Database: UK Biobank Subset; Tested SNP: rs79844047




(l) Nature of Gain of Power with different sample sizes

## 4.12 Conclusions and Discussion

The significance of a genetic association between a specific genetic variant (SNP) and a phenotypic trait relies on determining its statistical significance. The mean square error (MSE) plays a crucial role in computing the asymptotic variance of a regression coefficient, which in turn affects the denominator during the calculation of statistical power. It is worth noting that the MSE and the standard error of a regression coefficient are inversely related to statistical power. The main objective of this study is to investigate the behavior of the MSE in different modelling setups, aiming to understand its impact on statistical power.

The mathematical derivation of the mean square error (MSE) has been conducted for two different scenarios within the regression modelling framework to elucidate how conditioning on an additional or a set of known major single nucleotide polymorphisms (SNPs) can enhance the power to discover new genetic variants. Firstly, a linear regression model with two genetic variants ($w$, a test SNP, and $z$, a major effect SNP) was considered, leading to the simplification of the MSE and standard error. Inclusion of the significant SNP, e.g., ($z$) in the model shows a reduction in MSE, with the extent of reduction determined by the partial correlation between the covariates, denoted as $r^2_{yz.w}$ and eventually, increase the power of detecting the effect of tested SNP, $w$.

Assuming that the two covariates ($w$ & $z$) are uncorrelated, which is often the case in genome-wide association studies (GWAS) where the SNPs are located in separate linkage disequilibrium blocks, the expression of MSE of the tested SNP ($w$) contains a reduction term quantifying as $r^2_{yz}$, which represents a fraction of the trait variance ($\sigma^2_y$). Notably, the reduction in MSE also occurs if the covariates are correlated. This implies that incorporating an additional SNP in the model improves the performance of the new SNP by increasing the precision of the regression coefficient and enhancing the statistical power, especially if the additional SNP has a major effect.

An expression for regression coefficients has been derived in a multiple regression setup, where the design matrix is split into two blocks. One block comprises the tested Single Nucleotide Polymorphisms (SNPs), while the other block contains a list of major or conditioning SNPs, genetic principal components (PCs), and non-genetic factors like age and sex. Additionally, the expression for Mean Squared Error (MSE) has also been derived. This demonstrates how much the MSE decreases when conditioning on an additional block of covariates in the model. Mathematically, this reduction in MSE can be quantified by the term $\tilde{y}_A^T \tilde{H}_B \tilde{y}_A$, which is non-negative and due to the conditioning on the additional block of covariates. As MSE inversely affects power calculations, this reduction in MSE leads to increased power in discovering a new set of test SNPs when conditioning is applied. Moreover, the number of samples in the dataset plays an important role in calculating power and the gain of power is optimal for intermediate sample size, as for very large size power is 100% anyway.

Although conditioning can enhance the power to discover new genetic associations in certain cases, it may not always be effective, particularly for phenotypic traits that depend on a combination of numerous smaller affected Single Nucleotide Polymorphisms (SNPs). For instance, in the context of human skin pigmentation, major SNPs related to skin color and hair color have been identified, and conditioning on these SNPs may enable the exploration of new associations. However, for complex traits like height and BMI, hundreds of associated variants with having smaller effects, have been identified, and there is limited linkage disequilibrium (LD) among them. Consequently, the effect sizes obtained from a joint analysis are likely to be only marginally different from those obtained through single-SNP analysis.

The objective of this study is to explore how the ability to detect new genetic associations in a genomic region can be enhanced by conditioning on major or index SNPs, rather than focusing on establishing the causal relationship of the SNPs. Additionally, it

is reasonable to assume that the SNPs are uncorrelated, as the index SNPs are situated in distinct linkage disequilibrium (LD) blocks. However, it is worth noting that the same findings would also apply to correlated SNPs.

The mathematical derivation of power gain through conditioning has been extensively demonstrated in various regression models where the response or phenotype variable is continuous. It is worth noting that most morphological variations, such as human dentition and pigmentations, exhibit a continuous scale of variation [Carayon et al. (2019)]. However, in anthropological assessment schemes, these variations are often simplified and categorized into discrete categories for ease of representation. For example, the amount of melanin pigmentation in the eye, despite being a continuous quantity, is traditionally analyzed as blue versus brown to indicate the presence or absence of melanin. Initially, it was believed to follow Mendelian inheritance until quantitative analysis revealed its complex polygenic nature.

In the context of dental traits, Scott et al. (1997) noted that certain nonmetric traits, like incisor shoveling, exhibit a "quasi-continuous" nature. This implies that these traits can be considered as ordinal or dichotomous, derived from an underlying continuous quantity. Incisor shoveling, for instance, can be dichotomized into the presence or absence of curvature based on a specific threshold. In this context, the continuous underlying variable is referred to as a "latent variable" that corresponds to the assessed categorical variable.

The work of McCullagh (1980) provides a comprehensive exploration of regression models tailored to ordinal data, taking into account the existence of an underlying continuous latent variable. It emphasizes the interpretability of models based on this scale. However, it acknowledges that the concept of conditioning can also be useful in generalized linear models (GLMs) settings when a clear latent variable is not present, though we don't provide the mathematical details in the context of GLMs.

In genome-wide association studies (GWAS), case-control study designs are commonly employed to investigate the association between genetic variants and a binary response variable within a single cohort. Logistic regression is the prevailing method, comparing allele or genotype frequencies between cases and controls while considering potential confounding factors. However, conditioning in case-control studies presents challenges due to the complex impact of disease prevalence and allelic effect size on the power of conditioned GWAS [Pirinen et al. (2012); Zaitlen et al. (2012)]. Consequently, conditional analyses have been infrequently utilized in GWAS, primarily focused on disease phenotypes in case-control samples. Nevertheless, there are instances where conditioned GWAS has been successfully applied within this framework. For instance, a recent study addressed the inherent complexities of the case-control scenario, conditioning a GWAS for Alzheimer's disease on two well-established disease variants and identifying novel associations [Mez et al. (2017)]. While the concept of conditioning is commonly used in the literature, exploring the change in statistical power due to conditioning in case-control studies remains a promising area for future research.

# 5 Prediction Error (PE)

Throughout the preceding chapters, I explored various statistical models, including linear models and shrinkage methods, for different purposes in genetic association studies, such as association testing and improving statistical power. These models can be extended for predictive purposes in new datasets. Thus, it becomes imperative to discuss performance measures, like prediction error, and understand how these models trade-off between bias and variance, resulting in improved predictive performance.

In this chapter, I showed a brief illustration of the bias-variance trade-off for linear models and shrinkage methods, both at a single observation level and collectively. This demonstration helps predict the same trait from genetic data in new individuals and effectively utilize these models for accurate predictions and gain valuable insights into genetic associations.

For example, face GWASes provide insights into the genetic basis of various physical appearance traits, such as facial features, hair color, and eye color. Genetic associations from these studies can be used to develop prediction models to estimate an individual's appearance using their genetic information. These models can be valuable in forensic studies, where experts can potentially use them to reconstruct the physical appearance of unidentified individuals using their DNA remains.

## 5.1 Prediction Error at a single point (say, $x_0$)

Consider a simple linear regression model as

$$y = x\beta + \epsilon \ ; \text{ where}, \epsilon \sim N(0, \sigma^2) \tag{5.1}$$

The estimate of the parameter $\beta$ using ordinary least squares (OLS) can be obtained as follows:

$$\hat{\beta} = \frac{\sum xy}{\sum x^2} = \frac{\text{cov(xy)}}{\text{var(x)}}$$

The mean and variance of the estimate $\beta$ can be shown as follows:

$$E(\hat{\beta}) = \frac{\sum xy}{\sum x^2} = \frac{\sum x(x\beta + \epsilon)}{\sum x^2} = \beta$$

$$\begin{aligned}
\text{Var}(\hat{\beta}) &= \left(\frac{1}{\sum x^2}\right)^2 v\left[\sum_{i=1}^{n} x(x\beta + \epsilon)\right] \\
&= \left(\frac{1}{\sum x^2}\right)^2 v\left[\beta \sum x^2 + \sum x\epsilon\right] \\
&= \left(\frac{1}{\sum x^2}\right)^2 \sum_{i=1}^{n} x^2 v(\epsilon) \\
&= \frac{\sigma^2}{\sum x^2} = \frac{1}{n}\left[\frac{\sigma^2}{\text{Var(x)}}\right]
\end{aligned}$$

That is,

$$\hat{\beta} \sim N\left(\beta, \frac{1}{n}\left[\frac{\sigma^2}{\text{Var(x)}}\right]\right)$$

Suppose the regression model at a specific value of $x_0$ is $y_0 = f(x_0) + \epsilon_0$, where $f(x_0) = x_0\beta$ and the predictive model at $x_0$ is $\hat{f}(x_0) = x_0\hat{\beta}$. It can be shown that $\hat{f}(x_0)$ is an unbiased estimate of $f(x_0)$ that is,

$$E[\hat{f}(x_0)] = E[x_0\hat{\beta}] = x_0 E[\hat{\beta}] = x_0\beta = f(x_0)$$

and

$$Var[\hat{f}(x_0)] = Var[x_0\hat{\beta}] = x_0^2 Var[\hat{\beta}] = \frac{x_0^2\sigma^2}{nVar(x)}$$

The prediction error of the estimate $\hat{f}(x_0)$ can be obtained as follows

$$\begin{aligned}
\text{Prediction Error (PE)} &= E[y_0 - \hat{f}(x_0)]^2 \\
&= E[x_0\beta + \epsilon_0 - x_0\hat{\beta}]^2 \\
&= E[\epsilon_0 - x_0(\hat{\beta} - \beta)]^2 \\
&= E(\epsilon_0^2) + x_0^2 E(\hat{\beta} - \beta)^2 \\
&= \sigma^2 + x_0^2 \text{Var}(\hat{\beta}) \\
&= \sigma^2 \left[1 + \frac{1}{n}\frac{x_0^2}{\text{Var(x)}}\right] \qquad (5.2)
\end{aligned}$$

The expression of the prediction error in equation (5.2) indicates that as the sample size ($n$) approaches infinity, the prediction error (PE) will converge to the lower bound of $\sigma^2$. This lower bound represents the inherent error variance that contributes to the variability of $y$. In other words, no matter how large the sample size becomes, there will always be a residual error due to the inherent variability of the data.

Mathematically, as $n$ tends to infinity:

$$PE \to \sigma^2$$

This result shows that even with an infinitely large sample, it is impossible to completely eliminate the prediction error, as there will always be some level of inherent variability that cannot be explained or predicted by the model.

### 5.1.1 Conecting Prediction Accuracy to Heritability

In genetic studies, heritability is a measure of the proportion of phenotypic variation in a trait that can be attributed to additive genetic variation. It represents the extent to which the genetic makeup of individuals contributes to the observed variability in the trait of interest. Heritability is often denoted as $h^2$.

In the context of the additive model described in equation (5.1), the heritability can be defined as follows:

$$h^2 = \frac{\sigma_g^2}{\sigma_y^2} = 1 - \frac{\sigma_\epsilon^2}{\sigma_y^2}$$

where, $\sigma_g^2$ is the additive genetic variance, $\sigma_y^2$ is the variation in a trait of interest, and $\sigma_\epsilon^2$ is the variation due to error.

The heritability value ranges from 0 to 1, where 0 indicates that the trait's variation is entirely explained by environmental factors, and 1 indicates that the variation is solely attributed to genetic factors. A heritability value between 0 and 1 suggests that both genetic and environmental factors contribute to the trait's variability. Using the training data, we can estimate the variations in the trait ($\sigma_y^2$) and the error ($\sigma_\epsilon^2$), also known as the mean square error (MSE). With these estimates, we can then calculate the estimate of heritability ($\hat{h}^2$).

Based on the prediction error obtained in the equation (5.2), the prediction accuracy (PA), which will be discussed in detail later, can be expressed as

$$\begin{aligned} PA &= 1 - \frac{PE}{Var(y)} \\ &= 1 - \frac{\sigma^2 + x_0^2 \mathrm{Var}(\hat{\beta})}{Var(y)} \\ &= 1 - \frac{\sigma^2}{Var(y)} - \frac{x_0^2 \mathrm{Var}(\hat{\beta})}{Var(y)} \\ &= h^2 - \text{a positive term} \end{aligned}$$

From the above expression, it is noticeable that for the linear model in the case of OLS situation, the highest possible value of the prediction accuracy (PA) is the heritability. In other words, when the prediction accuracy is equal to the heritability, the model is performing at its best in capturing the genetic influence on the trait of interest.

## 5.2 Prediction Error at a set of (say, $m$) observations

### 5.2.1 Simple Linear Model Case

Suppose, we would like to predict the model (5.1) for a set of $m$ observations of the predictor variable $x_{01}, x_{02}, \cdots, x_{0m}$ and the corresponding predicted values are $\hat{y}_{01}, \hat{y}_{02}, \cdots, \hat{y}_{0m}$ whereas the true values of the response values are $y_{01}, y_{02}, \cdots, y_{0m}$. For the set of $m$ observations of the predictor variable, the prediction error (PE) can be obtained as

$$\begin{aligned} PE &= \frac{1}{m}\sum_{i=1}^{m} E\left[(y_{0i} - \hat{y}_{0i})^T (y_{0i} - \hat{y}_{0i})\right] \\ &= \frac{1}{m}\sum_{i=1}^{m} E[\epsilon_0^T \epsilon_0] + \frac{1}{m}\sum_{i=1}^{m} \sigma^2 trace\left[x_{0i}(X^T X)^{-1} x_{0i}^T\right] \\ &= \sigma^2 + \frac{\sigma^2}{mnVar(X)}\sum_{i=1}^{m} x_{0i}^2 \\ &= \sigma^2 + \frac{\sigma^2}{nVar(X)} Var(x_o) \end{aligned}$$

This expression represents the mean squared error (MSE) between the true response values $y_{0i}$ and the corresponding predicted values $\hat{y}_{0i}$ for each observation in the set. The prediction error quantifies the accuracy of the predictive model in estimating the true response values for the given predictor variable observations.

### 5.2.2 Multiple Linear Model Case

Consider a regression model as $y = X\beta + \epsilon$, where, $X$ is the matrix independent predictor variables and $\beta$ is the vector of regression coefficients. Then the prediction error (PE) for the new set of observations at $X_p$ can be obtained as

$$\begin{aligned} PE &= E\left[(y_p - X_p\hat{\beta})^T (y_p - X_p\hat{\beta})\right] \\ &= E\left[(X_p\beta + \epsilon - X_p\hat{\beta})^T (X_p\beta + \epsilon - X_p\hat{\beta})\right] \end{aligned}$$

$$
\begin{aligned}
&= E\left[\Big(\epsilon - X_p(\hat{\beta}-\beta)\Big)^T\Big(\epsilon - X_p(\hat{\beta}-\beta)\Big)\right] \\
&= E\left[\epsilon^T\epsilon + (\hat{\beta}-\beta)^T X_p^T X_p(\hat{\beta}-\beta)\right] \\
&= E[\epsilon^T\epsilon] + E\left[(\hat{\beta}-\beta)^T X_p^T X_p(\hat{\beta}-\beta)\right] \\
&= \sigma^2 + \sigma^2 trace\left[X_p(X^TX)^{-1}X_p^T\right] \\
&= \sigma^2 + \frac{\sigma^2}{n} trace\left[X_p\left(Var(X)\right)^{-1}X_p^T\right] \qquad (5.3)
\end{aligned}
$$

## 5.3 Expression of Prediction Error considering two independent covariates (say, $w, z$)

Following the general expression of prediction error in equation (5.3), let us now focus on a true regression model that includes two independent covariates ($w$ and $z$), with corresponding regression coefficients ($\beta_w$ and $\beta_z$). The fitted model can be represented as follows:

$$\hat{y}_0 = f_w w_0 \hat{\beta}_w + f_z z_0 \hat{\beta}_z$$

where $f_w, f_z$ are shrinkage factors corresponding to the two covariates, $w_0, z_0$ are the values of the covariates at which the prediction will be made. When both of the shrinkage factors consider the value one, then it will indicate the OLS situation otherwise the need for shrinkage.

Now if the response variable is predicted based on only one covariate say, $w_0$ only, then it will be a situation that the value $f_w = 1$ and $f_z = 0$ and the prediction error can be derived as

$$
\begin{aligned}
PE &= E\left[f_w w_0\beta_w + f_z z_0\beta_z + \epsilon - f_w w_0\hat{\beta}_w - f_z z_0\hat{\beta}_z\right]^2 \\
&= E\left[\epsilon - w_0(f_w\hat{\beta}_w - \beta_w) - z_0(f_z\hat{\beta}_z - \beta_z)\right]^2 \\
&= E[\epsilon^2] + w_0^2 E\left[f_w\hat{\beta}_w - f_w\beta_w + f_w\beta_w - \beta_w\right]^2 + z_0^2 E\left[f_z\hat{\beta}_z - f_z\beta_z + f_z\beta_z - \beta_z\right]^2
\end{aligned}
$$

$$= \sigma^2 + w_0^2 f_w^2 E[\hat{\beta}_w - \beta_w]^2 + w_0^2 \beta_w^2 (1 - f_w)^2 + z_0^2 f_z^2 E[\hat{\beta}_z - \beta_z]^2 + z_0^2 \beta_z^2 (1 - f_z)^2$$
$$= \sigma^2 + \frac{1}{n} \frac{\sigma^2 w_0^2 f_w^2}{Var(w)} + w_0^2 \beta_w^2 (1 - f_w)^2 + \frac{1}{n} \frac{\sigma^2 z_0^2 f_z^2}{Var(z)} + z_0^2 \beta_z^2 (1 - f_z)^2$$
$$= \sigma^2 + \frac{1}{n} \frac{\sigma^2 w_0^2}{Var(w)} + z_0^2 \beta_z^2 \text{ [Putting, } f_w = 1 \text{ and } f_z = 0] \tag{5.4}$$

## 5.4 Expression of Prediction Error considering Ridge Regression Model

For a ridge regression model, say, $y = X\beta + \epsilon$, where, the parameter of the model estimated by the ridge estimator as

$$\begin{aligned}
\hat{\beta}_\lambda =& (X^T X + \lambda I)^{-1} X^T y \\
=& \left[I + \lambda (X^T X)^{-1}\right]^{-1} (X^T X)^{-1} X^T y \\
=& \left[I + \lambda (X^T X)^{-1}\right]^{-1} \hat{\beta} \\
=& W_\lambda \hat{\beta} \text{ where, } W_\lambda = \left[I + \lambda (X^T X)^{-1}\right]^{-1}
\end{aligned}$$

It can be shown that the ridge estimate is a biased estimate of $\beta$ that is,

$$\begin{aligned}
E[\hat{\beta}_\lambda] =& E[W_\lambda \hat{\beta}] = W_\lambda \beta \\
& \text{and} \\
Var[\hat{\beta}_\lambda] =& Var[W_\lambda \hat{\beta}] \\
=& W_\lambda Var(\hat{\beta}) W_\lambda^T \\
=& W_\lambda [\sigma^2 (X^T X)^{-1}] W_\lambda^T \\
=& \sigma^2 \left[I + \lambda (X^T X)^{-1}\right]^{-1} (X^T X)^{-1} \left[\left(I + \lambda (X^T X)^{-1}\right)^{-1}\right]^T \\
=& \sigma^2 (X^T X + \lambda I)^{-1} (X^T X) \left[(X^T X + \lambda I)^{-1}\right]^T
\end{aligned}$$

Suppose a predictve model is $y_p = X_p\beta_\lambda + \epsilon$ and the corresponding fitted predictive model is $\hat{y}_p = X_p\hat{\beta}_\lambda$, where, $\hat{\beta}_\lambda$ is the ridge estimate. Then the prediction error for this predictive model can be expressed as

$$
\begin{aligned}
&PE = E[(y_p - \hat{y}_p)^T(y_p - \hat{y}_p)] \\
&= E[(X_p\beta + \epsilon - X_p\hat{\beta}_\lambda)^T(X_p\beta + \epsilon - X_p\hat{\beta}_\lambda)] \\
&= E[\epsilon^T\epsilon] + E[(X_p\hat{\beta}_\lambda - X_p\beta)^T(X_p\hat{\beta}_\lambda - X_p\beta)] \\
&= \sigma^2 + E[(X_p\hat{\beta}_\lambda - X_p\beta)^T(X_p\hat{\beta}_\lambda - X_p\beta)] \\
&= \sigma^2 + E\left[\hat{\beta}_\lambda^T X_p^T X_p\hat{\beta}_\lambda - \hat{\beta}_\lambda^T X_p^T X_p\beta - \beta^T X_p^T X_p\hat{\beta}_\lambda + \beta^T X_p^T X_p\beta\right] \text{ [here, } \hat{\beta}_\lambda = W_\lambda\hat{\beta}] \\
&= \sigma^2 + E[\hat{\beta}^T W_\lambda^T X_p^T X_p W_\lambda\hat{\beta} - \hat{\beta}^T W_\lambda^T X_p^T X_p W_\lambda\beta - \beta^T W_\lambda^T X_p^T X_p W_\lambda\hat{\beta} + \beta^T W_\lambda^T X_p^T X_p W_\lambda\beta \\
&+ \hat{\beta}^T W_\lambda^T X_p^T X_p W_\lambda\beta + \beta^T W_\lambda^T X_p^T X_p W_\lambda\hat{\beta} - \beta^T W_\lambda^T X_p^T X_p W_\lambda\beta - \hat{\beta}^T W_\lambda^T X_p^T X_p\beta \\
&- \beta^T X_p^T X_p W_\lambda\hat{\beta} + \beta^T X_p^T X_p\beta] \\
&= \sigma^2 + E\left[\left((\hat{\beta} - \beta)X_p W_\lambda\right)^T\left((\hat{\beta} - \beta)X_p W_\lambda\right)\right] \\
&+ \beta^T W_\lambda^T X_p^T X_p W_\lambda\beta - \beta^T W_\lambda^T X_p^T X_p\beta - \beta^T X_p^T X_p W_\lambda\beta + \beta^T X_p^T X_p\beta \text{ [As, } E(\hat{\beta}) = \beta] \\
&= \sigma^2 + E\left[\left((\hat{\beta} - \beta)X_p W_\lambda\right)^T\left((\hat{\beta} - \beta)X_p W_\lambda\right)\right] + \left[\beta^T W_\lambda^T X_p^T X_p - \beta^T X_p^T X_p\right](W_\lambda - I)\beta \\
&= \sigma^2 + E\left[\left((\hat{\beta} - \beta)X_p W_\lambda\right)^T\left((\hat{\beta} - \beta)X_p W_\lambda\right)\right] + \beta^T(W_\lambda - I)^T X_p^T X_p(W_\lambda - I)\beta \\
&= \sigma^2 + \sigma^2 trace\left[W_\lambda^T X_p^T(X^T X)^{-1} W_\lambda X_p\right] + \beta^T(W_\lambda - I)^T X_p^T X_p(W_\lambda - I)\beta \qquad (5.5)
\end{aligned}
$$

The above expression of prediciton error can be simplified using the singular value decomposition of the design matrix $(X)$ as

$$X = UDV^T$$

where, $U$ is a $(n \times p)$ orthogonal matrix, $V$ is a $(p \times p)$ orthogonal matrix, and $D$ is a

$(p \times p)$ diagonal matrix having diagonal elements as the singular values $d_1, d_2, \cdots, d_p$

$$\begin{aligned} W_\lambda &= \left[I + \lambda (X^T X)^{-1}\right]^{-1} \\ &= \left[V D^2 D^{-2} V^T + \lambda V D^{-2} V^T\right]^{-1} \\ &= \left[V D^{-2} (D^2 + \lambda I_p) V^T\right]^{-1} \\ &= V D^2 (D^2 + \lambda I_p)^{-1} V^T \\ \text{So, } XW_\lambda &= U D V^T V D^2 (D^2 + \lambda I_p)^{-1} V^T \\ &= U D^3 (D^2 + \lambda I_p)^{-1} V^T \end{aligned}$$

Second term of the equation (5.5),

$$\begin{aligned} \sigma^2 trace\left[XW_\lambda (X^T X)^{-1} W_\lambda^T X^T\right] &= \sigma^2 trace\left[U D^3 (D^2 + \lambda I_p)^{-1} V^T V D^{-2} V^T V D^2 (D^2 + \lambda I_p)^{-1} D U^T\right] \\ &= \sigma^2 trace\left[U D^3 (D^2 + \lambda I_p)^{-2} D U^T\right] \\ &= \sigma^2 trace\left[D^4 (D^2 + \lambda I_p)^{-2}\right] \end{aligned}$$

The $i$th element of the above term can be written as

$$\sigma^2 trace\left[D^4 (D^2 + \lambda I_p)^{-2}\right]_{ii} = \sigma^2 \left(\frac{d_i^2}{d_i^2 + \lambda}\right)^2 \text{ ; for } i = 1, 2, \cdots, p$$

Third term of the equation (5.5),

$$\begin{aligned} &\beta^T (W_\lambda - I)^T X^T X (W_\lambda - I)\beta \\ &= \beta^T \left[V D^2 (D^2 + \lambda I_p)^{-1} V^T - V V^T\right]^T V D^2 V^T \left[V D^2 (D^2 + \lambda I_p)^{-1} V^T - V V^T\right] \\ &= \beta^T V \left[D^2 (D^2 + \lambda I_p)^{-1} - I_p\right]^T V^T V D^2 V^T V \left[D^2 (D^2 + \lambda I_p)^{-1} - I_p\right] V^T \beta \\ &= \beta^T V D^2 \left[D^2 (D^2 + \lambda I_p)^{-1} - I_p\right]^2 V^T \beta \end{aligned}$$

Letting, a $(p \times 1)$ vector as $S = V^T \beta$, the $i$th element of the above term can be written

as

$$\beta^T V D^2 \left[D^2(D^2+\lambda I_p)^{-1} - I_p\right]_{ii}^2 V^T \beta = \lambda^2 \left(\frac{Sd_i}{d_i^2+\lambda}\right)^2$$

Finally, the prediction error obtained in the equation (5.5) can be written for the singular value decomposition of the design matrix $X$ as

$$PE = \sigma^2 \sum_{i=1}^{p} \left(\frac{d_i^2}{d_i^2+\lambda}\right)^2 + \lambda^2 \sum_{i=1}^{p} \left(\frac{Sd_i}{d_i^2+\lambda}\right)^2$$

## 5.5 Expression of Prediction Error considering general case

For a general linear regression model, $y = f(x) + \epsilon$, where the true mean of the model can be anything, the expected prediction error of a regression fit $\hat{f}(x)$ at a specific value $x = x_0$ can be obtained as

$$\begin{aligned}
PE[\hat{f}(x_0)] &= E[y_0 - \hat{f}(x_0)]^2 \\
&= E[f(x_0) + \epsilon_0 - \hat{f}(x_0)]^2 \\
&= E[\epsilon_0 - \left(\hat{f}(x_0) - f(x_0)\right)]^2 \\
&= E[\epsilon_0]^2 + E\left[\hat{f}(x_0) - f(x_0)\right]^2 \\
&= E[\epsilon_0]^2 + E\left[\hat{f}(x_0) - E\left(\hat{f}(x_0)\right) + E\left(\hat{f}(x_0)\right) - f(x_0)\right]^2 \\
&= E[\epsilon_0]^2 + E\left[\hat{f}(x_0) - E\left(\hat{f}(x_0)\right)\right]^2 + \left[E\left(\hat{f}(x_0)\right) - f(x_0)\right]^2 \\
&= \sigma^2 + \text{Var}\left[\hat{f}(x_0)\right] + \text{Bias}^2\left[\hat{f}(x_0)\right] \\
&= \text{Irreducible Error} + \text{Variance} + \text{Bias}^2
\end{aligned} \tag{5.6}$$

It is evident that when the model estimate is taken as a simple linear regression fit, i.e., $\hat{f}(x_0) = x_0\hat{\beta}$, the equation (5.2) is obtained.

## 5.6 Prediction Error at different Shrinkage Factors ($f$)

Considering the shrinkage factor ($f$) in the simple linear estimate, it is possible to explain the behavior of different shrinkage methods for the different values of the shrinkage factor. The predictive model for the simple linear shrinkage estimate at a particular value $x_0$ can be expressed as $\hat{f}(x_0) = x_0 f\hat{\beta}$ and the prediction error can be obtained as

$$\begin{aligned}
PE(\hat{f}(x_0)) &= E[y_0 - \hat{f}(x_0)]^2 \\
&= E[x_0\beta + \epsilon_0 - x_0 f\hat{\beta}]^2 \\
&= E[\epsilon_0 - x_0(f\hat{\beta} - \beta)]^2 \\
&= E(\epsilon^2) + x_0^2 E\left[f\hat{\beta} - \beta\right]^2 \\
&= E(\epsilon^2) + x_0^2 E\left[f\hat{\beta} - f\beta + f\beta - \beta\right]^2 \\
&= E(\epsilon^2) + x_0^2 E\left[f(\hat{\beta} - \beta) - \beta(1-f)\right]^2 \\
&= E(\epsilon^2) + x_0^2 E\left[f^2(\hat{\beta} - \beta)^2\right] + x_0^2\beta^2(1-f)^2 \\
&= \sigma^2 + x_0^2[f^2 \underbrace{\text{var}(\hat{\beta})}_{\text{a}} + \underbrace{\beta^2}_{\text{b}}(1-f)^2] \\
&= \text{Irreducible Error} + \text{Variance} + \text{Bias}^2
\end{aligned} \tag{5.7}$$

It is observed from the above expression of prediction error that if the shrinkage factor, $f$, takes the value one, then the prediction error (PE) is equivalent to the OLS case, but trade-off between the variance and bias will be encountered for the different values of $f$. Reduction of any value of $f$ less than one, it will cause an increase in the bias but decrease in the variance. It is also noticeable that both the bias and variances will be increased for the value of the shrinkage factor greater than one. Based on different values of shrinkage fractor, $f$, the bias-variance trade-off has also been discussed in Frank and Friedman (1993) and subsequently he justified the qualitative behavior of RR, PCR, and

PLS.

**Mathematical Note:**

$$
\begin{aligned}
z &= af^2 + b(1-f)^2 \\
\frac{\mathrm{d}z}{\mathrm{d}f} &= 2af - 2b(1-f) = 0 \\
\therefore \hat{f} &= \frac{b}{a+b} \\
z(\hat{f}) &= a(\frac{b}{a+b})^2 + b(\frac{a}{a+b})^2 \\
&= \frac{ab^2}{(a+b)^2} + \frac{a^2b}{(a+b)^2} \\
&= \frac{ab}{a+b}
\end{aligned}
$$

Following this note, the optimum value of the shrinkage factor that minimizes the prediction error can be found as

$$
\hat{f} = \frac{\beta^2}{\beta^2 + Var(\hat{\beta})} = \frac{\beta^2}{\beta^2 + \frac{1}{n}\frac{\sigma^2}{\mathrm{Var(x)}}} = \frac{\mathrm{Var(x)}}{\mathrm{Var(x)} + \lambda};\ \text{letting, } \lambda = \frac{\sigma^2}{n\beta^2}
$$

and the prediction error at this optimum value of the shrinkage factor $\hat{f}$ will be

$$
\begin{aligned}
PE(\hat{f}) &= \sigma^2 + x_0^2 . \frac{\frac{\sigma^2}{nVar(x)}\beta^2}{\frac{\sigma^2}{nVar(x)} + \beta^2} \\
&= \sigma^2 + \frac{\sigma^2}{n} . x_0^2 . \frac{\beta^2}{\beta^2 Var(x) + \frac{\sigma^2}{n}}
\end{aligned} \tag{5.8}
$$

It was also shown in the equation (5.2) that the prediction error of a regression model in the case of OLS situation as

$$
\begin{aligned}
PE(OLS) &= \sigma^2 \left[1 + \frac{1}{n}\frac{x_0^2}{\mathrm{Var(x)}}\right] \\
&= \sigma^2 + \frac{\sigma^2}{n} . x_0^2 . \frac{\beta^2}{\beta^2 Var(x)}
\end{aligned} \tag{5.9}
$$

Comparing the prediction error obtained in equations (5.8) and (5.9), it is observed that in the case of introducing shrinkage, the prediction error has a larger denominator than the OLS situation which is a clear indication of getting a smaller prediction error.

On the other hand, once the optimal shrinkage factor is obtained, the linear shrinkage estimate can be expressed as follows

$$f\hat{\beta} = \left[\frac{\text{Var(x)}}{\text{Var(x)} + \lambda}\right]\hat{\beta}$$

Following Frank and Friedman (1993), it is noticeable here that the OLS estimates, $\hat{\beta}$, are differentially shrunk with the increase of the relative value of the shrinkage factor and decrease of the variance of $x$. The amount of shrinkage is also controlled by the value of $\lambda$, which is defined as the error variance ($\sigma^2$) divided by the training sample size that is, the larger the value of $\lambda$, the more differential shrinkage as well as more overall global shrinkage. It is also noticeable that if the values of the sample size go to infinity then the value of $\lambda$ will also go to zero and the optimum value of $f$ will turn into one. That indicates that it is not necessary to think about the shrinkage if the sample size goes to infinity.

With the different values of $f$, the optimal value of prediction error or MSE as well as the choice of best predictors can be obtained. Since $f$ is the true parameter value and we don't usually know this value, we can calculate the estimates of $f$ from the training data. The estimate of error variance ($\sigma^2$) can be obtained by fitting the model on the training data and the $var(x)$ can also be calculated from the training data. With the help of training data, the distribution of $\hat{\beta}$ can be obtained and by plugging in the values of the estimates, $\hat{\beta}$, it is also possible to estimate the value of $\beta^2$.

## 5.7 Prediction Accuracy for different Models

Prediction accuracy is a measure of the association between the response and the predictors. It measures the proportion of variation in the response ($y$) that can be explained by the predictors and it is useful in comparing different choices of predictor variables in any given problem [Rao (2002)]. Mathematically, it can be expressed as

$$\text{Prediction accuracy}, R^2 = 1 - \frac{PE}{\text{Var(y)}}$$

Here, $PE = E[y - \hat{y}]^2$, is the unexplained mean residual variance in the test data. It is observed that $R^2$ goes to unity if the unexplained error variance goes to zero which means the predictor variables explain the model well. On the other hand, if the predictor variable fails to predict the model well that is, no significant reduction in error variance due to the use of predictors, then the prediction accuracy ($R^2$) will tend to be zero. In the situation when the predictor variables predict the response perfectly, then there will be no prediction error, and eventually, the prediction accuracy will be one.

Considering the prediction error measured for the true parameter linear regression model obtained in the equation (5.3), the prediction accuracy can be expressed as

$$\begin{aligned}\text{Prediction accuracy}, R^2 &= 1 - \frac{PE[\hat{f}(x_0)]}{\text{Var(y)}} \\ &= 1 - \left[\frac{\sigma^2 + \text{Var}\left[\hat{f}(x_0)\right] + \text{Bias}^2\left[\hat{f}(x_0)\right]}{\beta^2\text{var(x)} + \sigma^2}\right]\end{aligned}$$

In the case of the OLS situation, the prediction accuracy will be

$$\begin{aligned}R^2 &= 1 - \frac{PE[\hat{f}(x_0)]}{\text{Var(y)}} \\ &= 1 - \left[\frac{\sigma^2 + \text{Var}\left[\hat{f}(x_0)\right]\left[\hat{f}(x_0)\right]}{\beta^2\text{var(x)} + \sigma^2}\right]\end{aligned}$$

$$= 1 - \left[ \frac{\sigma^2 \left( 1 + \frac{1}{n} \frac{x_0^2}{\text{var(x)}} \right)}{\beta^2 \text{var(x)} + \sigma^2} \right]$$

and in the situation when the shrinkage factor is involved in the linear regression estimate then

$$\begin{aligned} R^2 &= 1 - \frac{PE[\hat{f}(x_0)]}{\text{Var(y)}} \\ &= 1 - \left[ \frac{\sigma^2 + \frac{1}{n} \frac{x_0^2 f^2 \sigma^2}{\text{var(x)}} + x_0^2 \beta^2 (1-f)^2}{\beta^2 \text{var(x)} + \sigma^2} \right] \end{aligned}$$

# 6 Applied Research Work: Dental Morphology Analysis

## 6.1 Introduction and Aims

This study investigates the dental shape variation for 70 specimens and 7 groups of specimens (species) such as Homo sapiens (HOMO), Homo neanderthalensis (NEA), Australopithecus africanus (AUST), Paranthropus robustus (PROB), Gorilla, Pan troglodytes (PAN), and Pongo pygmaeus (PONGO). Each of the groups consists of an equal number of specimens which indicates the data is balanced. To assess the accuracy of specimen classification, different geometric morphometric data have been analyzed separately which are based on 2D, 3D, and deformetrica approaches of data representation.

Various classical and machine learning predictive models have been used to predict the specimen classes for these datasets. Model performance was assessed using the average prediction error rate and the maximum prediction error (PE) rate within each of the predicted classes. Based on the discussion of the previous chapter, I used the prediction error metrics to classify the specimens and determine the optimal predictive model by comparing their prediction error rates.

## 6.2 Data Structure

### 2D Data

The 2D data contains 103 Procrustes coordinates (206 variables), among which 97 coordinates (194 variables) are based on semi-landmarks and 6 coordinates (12 variables) are based on main landmarks. It is also found that Centroid size plays an important role in separating the categories which has been confirmed later by different predictive modelling approaches.

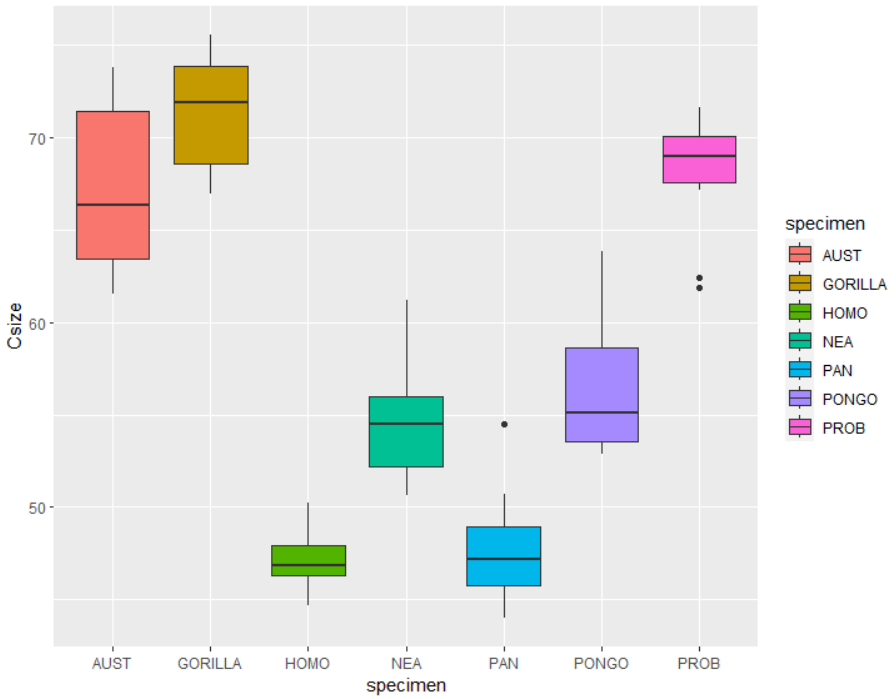



**Figure: Boxplot showing variation of Centroid Size among Specimens (species)**

The total number of covariates analyzed for 2D data is 207, among them 206 are based on Procrustes coordinates, and one is Centroid size. Centroid size, a widely used measure in geometric morphometrics, represents the size of an object. It is calculated as the square root of the sum of squared distances between all landmarks of the object and its centroid. The centroid is the center of gravity determined by averaging the x and y coordinates of all landmarks [Klingenberg (2016)].

The 2D dataset has been analyzed in two ways such as

- All Landmarks (Main and Semi-landmark) [206 variables]
- Main Landmark Only [12 variables]

### 3D Data

The 3D data contains the variables with two types of information. Firstly, the variables that contain information regarding landmarks and semi-landmarks on the

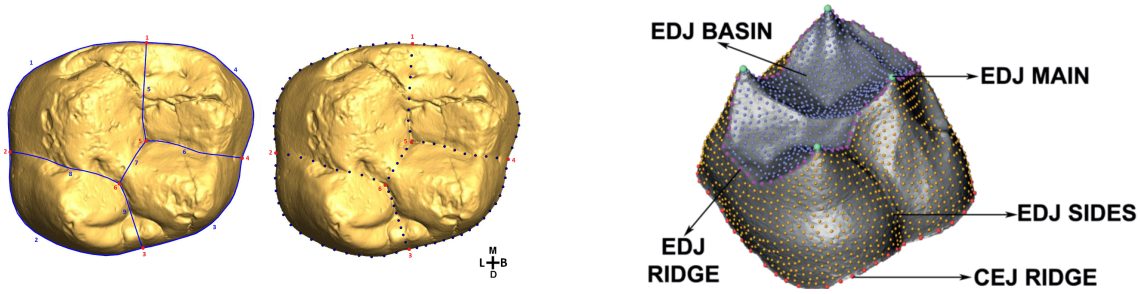



**Figure: Illustration of landmark-based methods. On the left: 2D landmarks/semi-landmarks located on the outer enamel surface (OES), including 6 main landmarks and 99 curve semi-landmarks. On the right: 3D landmarks/semi-landmarks situated at the enamel-dentin junction (EDJ), comprising 4 main landmarks and 128 curve and 1757 surface semi-landmarks [Delgado et al. (2021)]**

Curves only. Secondly, all the digitized points which include both landmarks and semi-landmarks information on the Curves as well as the Surface. So, the dataset has been analyzed separately as

- Landmark points on Curves Only [125 points = 375 variables]
- Landmark points on both Curves and Surface [5667 variables]

### Deformetrica Data

Deformetrica is a method that analyses the 3D data and identifies key locations on the surface where there is a major change of shape. The amount and direction of change at those locations are represented by two types of information, 'Momenta' and 'Velocity'. The two sets of information have been analyzed separately as

- Momenta Data which contains 5376 points on the surfaces and considered them as variables
- Velocity Data which also contains 5376 variables

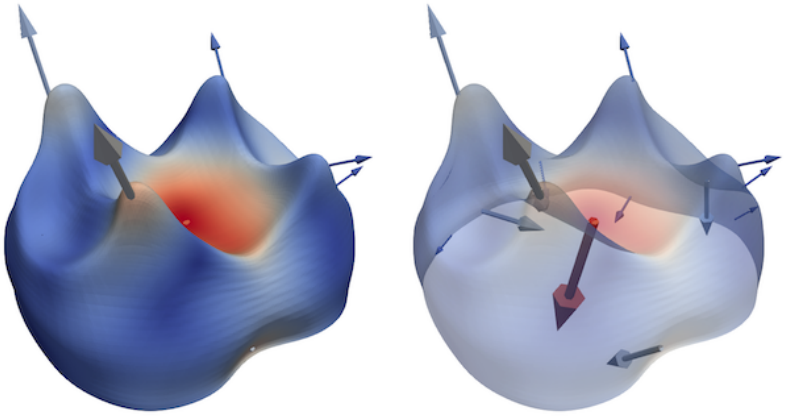


**Figure: An illustration of the Deformetrica method, shown on a dental surface. The amount and direction of shape changes at key locations are represented with arrows.** [https://gitlab.com/jeandumoncel/tools-for-deformetrica/]

## 6.3 Dimension Reduction Techniques for Prediction (My Contribution)

This chapter focuses on the comprehensive utilization of a 2D data structure to conveniently and consistently represent the results of data analysis. This approach enables us to compare the results within the same avenues, facilitating effective evaluation and comparison of different dimension reduction techniques and prediction models. The ultimate goal is to recommend a modeling technique that minimizes the error in specimen classifications.

Given that the 2D data consists of a larger number of features (206) compared to the number of observations (70), it becomes necessary to reduce the dimensionality. Various dimension reduction techniques can be applied to achieve this reduction, ensuring that the essential information is retained while reducing the complexity of the data. By implementing these strategies, we aim to get a reduced number of dimensions, which will help to compare different prediction modelling techniques and eventually, we can make recommendations for the most accurate modeling technique for specimen classifications.

### 6.3.1 Principal Component Analysis (PCA)

The principal component analysis (PCA) is a widely used dimension reduction method that transforms a high-dimensional dataset into a lower-dimensional representation while retaining most of the original data's variability [Johnson and Wichern (2007)]. The reduced data set contains a new set of uncorrelated (principal components) that captures the maximum variance in the original data. The scree plot was plotted to see how drastically the explained variations are slopping down for different numbers of PCs and a threshold of explained variation was considered at 0.5% to select top PCs. Based on this setup, different numbers of PCs were selected for different data structures.

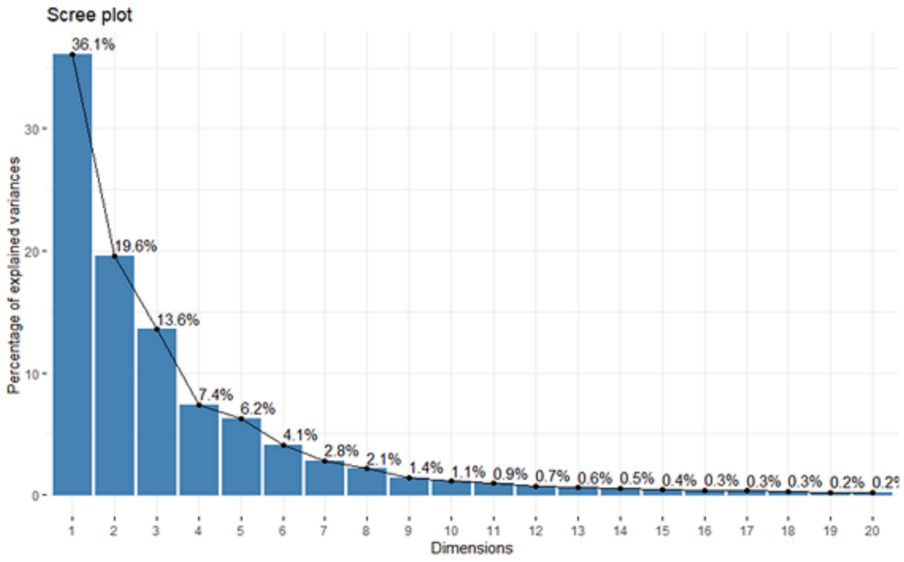



**Figure: Scree plot, which represents the proportion of variance captured by each principal component from the data.**

For example, considering the main and semi-landmark (206 variables) in the 2D data structure, the scree plot above indicates that the variance of PCs falls rapidly, and considering the threshold of 0.5% variance explained, the top 15 PCs are sufficient. In addition to top PCs, Centroid size was also used as a covariate. The scatter plot of Centroid size and PC1 indicated that Centroid size has a fair amount of information for separating the categories which is later confirmed by the different predictive modeling

approaches.

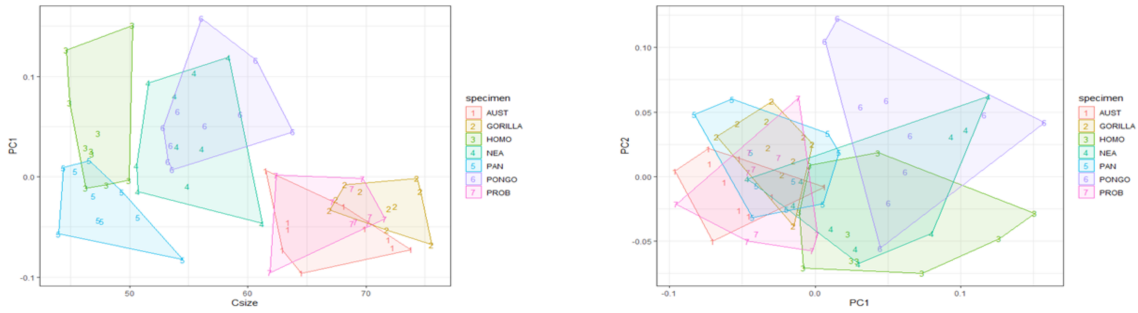



**Figure: Scatter plot of Centroid Size vs PC1 and PC1 vs PC2.**

### 6.3.2 Between Group Principal Component Analysis (bgPCA)

Between-group principal components analysis (bgPCA) is another dimension reduction technique that captures group differences in the data set and aims to find principal components that maximize the between-group variance. This method calculates the covariance matrix of the group means and then does a principal component analysis (PCA) on that matrix. Note that, this method is equivalent to a reduced-rank linear discriminant analysis (LDA).

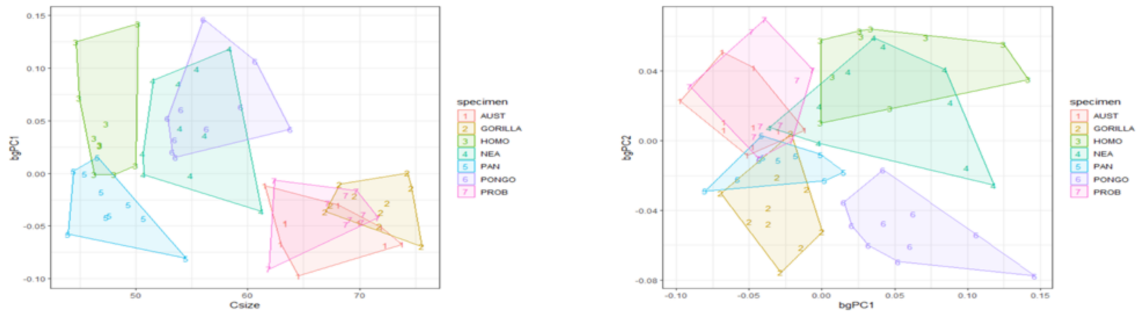



**Figure: Scatter plot of Centroid Size vs bgPC1 and bgPC1 vs bgPC2.**

The 2D Procrustes coordinates dataset consists of 7 specimens and 206 variables derived from landmark-based coordinates. In the R software, the "groupPCA" function was applied to the $(70 \times 206)$ data matrix to compute bgPCs. Initially, this function

calculates group mean matrices and performs principal component analysis (PCA) to extract a reduced number of principal components. Considering seven groups, the function identified 6 bgPCs, all of which were included in the predictive models. Visualizations displaying the relationships between Centroid size and bgPC1, as well as bgPC1 and bgPC2, are depicted in the figures above.

### 6.3.3 Leave-one-out cross-validated group PCA (cv-bgPCA)

Leave-one-out cross-validated group PC (cv-bgPCAs) is another technique to calculate the reduced dimensions by using the leave-one-out cross-validated process. The data set was divided into two parts such as train and test data, where train data contains the information for all individuals except the first. Then the function "prcomp" was performed on the trained data which returned a list of components. The trained and test data were then predicted with these components and the principal component scores were extracted for the train as well as test data. The process was repeated for the whole data set within a loop in R, which eventually provided us a $(70 \times 7)$ principal component matrix. The scatter plot of Centroid size vs. cv-bgPC1 and cv-bgPC1 vs. cv-bgPC2 have been shown in the following figures:

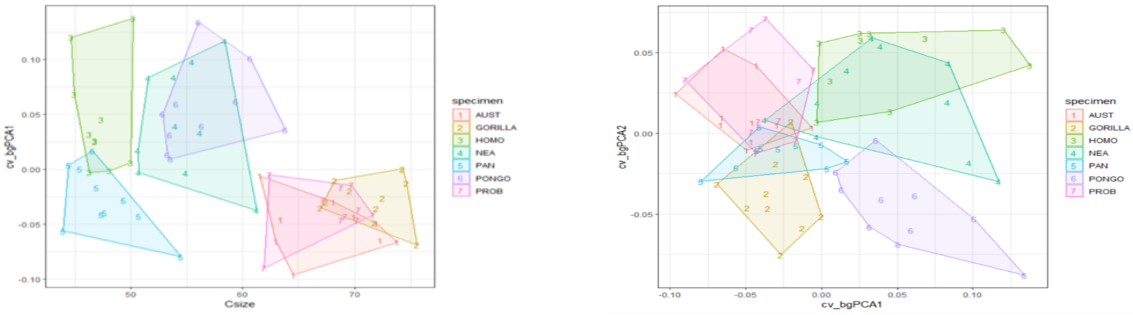



**Figure: Scatter plot of Centroid Size vs cv-bgPC1 and cv-bgPC1 vs cv-bgPC2.**

### 6.3.4 tSNE

t-SNE (t-Distributed Stochastic Neighbor Embedding) is also a dimension reduction technique that reduces the dimensionality of high-dimensional data for visualization in lower-dimensional spaces [Zhou et al. (2018)]. Unlike PCA and other linear techniques, t-SNE focuses on retaining pairwise similarities among data points instead of overall variance. The performance of t-SNE has been checked at different value levels of hyper-parameter 'Perplexity'.

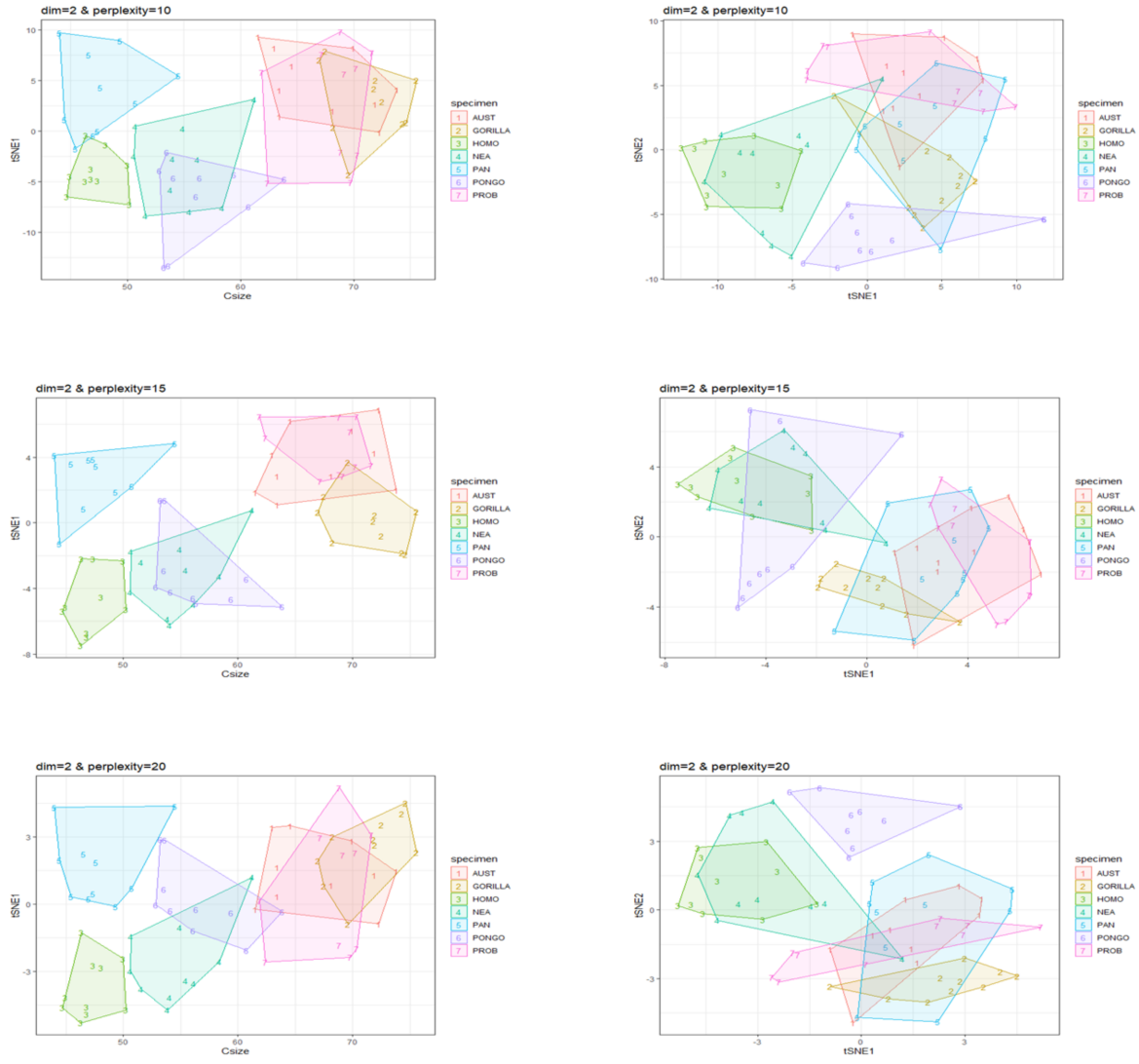



**Figure: Scatter plot of Centroid Size vs tSNE1, and tSNE1 vs tSNE2. The tSNEs were calculated at different value levels of the hyperparameter 'Perplexity'**

## 6.4 Prediction Accuracy with Different Prediction Models (My Contribution)

The performance of a predictive model depends on how accurately it can predict or classify a species, and this is evaluated by calculating the "Confusion Matrix", which provides the proportion of misclassification for each category. Since the Confusion Matrix is usually very large, the performance of a predictive model is evaluated with the "average error rate". But it may be the situation that the average error rates are similar or very close for different models, and the error rates of each class vary quite a bit. Therefore, in addition to the average error rate, another metric such as the maximum misclassification rate for any category was also calculated. In the evaluation of predictive models, I compared various predictive models based on both the average error rate as well as groupwise maximum error rate, leading to a comprehensive assessment of their performance.

The prediction models presented in this study were based on the selected principal components (PCs) and the bgPCs, as described earlier. Additionally, I included the covariate centroid size ("Csize") to gain deeper insights into species classification, as it demonstrated significance in the scatter plot. Consequently, I evaluated each set of covariates both with and without the inclusion of the "Csize" variable. The results for the prediction models were calculated using the following covariate setups:

- Model Accuracy with PCs only
- Model Accuracy with PCs + Centroid Size (Csize)
- Model Accuracy with bgPCs only
- Model Accuracy with bgPCs + Centroid Size (Csize)

The performance of the different predictive models has been evaluated with different metrics for the above covariate setup. Firstly, the predictive models considered the covariates by adding PCs or bgPCs one by one in the models and then applied for the leave-one-out cross-validation (LOOCV) to measure the performance metrics.

### 6.4.1 Random Forest (RF) Model

Random Forest (RF) approach has been performed to evaluate how correctly the species can be classified. It is a widely used classification or regression-based method that grows many classification trees randomly taking subsamples of the observations as well as subsamples of the variables and taking an average of them which eventually reduces the variance. The performance of the RF model has been evaluated with different metrics for the covariate setup described above. Since the process is random and the results may vary for executing the model each time, we have run the model at different numbers of trees and the result has been reported with the highest number of trees i.e., B= 10003, since this variation reduces when the number of trees increases.

### 6.4.2 Multinomial Logistic Regression

Since we have a multiclass classification task i.e., 7 groups of species to classify, another widely used classification approach, Multinomial logistic regression (MLR), has also been used. This model also predicts different categories with the help of different performance metrics.

### 6.4.3 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a statistical method commonly used for prediction and classification tasks. It aims to find a linear combination of features that maximally separates different classes or categories in a dataset. Considering the above

covariate setup, LDA has been performed to classify the species and measure the prediction accuracy based on different performance metrics.

### 6.4.4 K-nearest Neighbour (K-NN)

The k-nearest neighbor approach has been performed to evaluate the performance metrics and compare them with other methods as well. It is a memory-based classifier that classifies or predicts a data point based on the majority class or average of its k-nearest neighbors in the feature space [Hastie et al. (2009)]. The algorithm calculates the distance between the new data points and all existing data points in the training dataset and then selects the $K$ training points with the closest distances. The predicted class or value for the new data point is then determined by the majority vote or the average value among these $K$ nearest neighbors.

### 6.4.5 Support Vector Machine (SVM)

Support Vector Machine (SVM) algorithm is a supervised learning technique used for classification and regression problems. This method classifies the data points into different categories by finding a hyperplane that maximizes the margin between the classes [Hastie et al. (2009)]. If the data points are non-overlapping, the algorithm can effectively classify them. However, real-world datasets may not be perfectly separable. In such cases, SVM uses strategies like soft margin to allow limited misclassification while aiming to find a reasonable separation. This method has also been performed in this research work to evaluate the performance- of classification metrics and compare them with other predictive models.

### 6.4.6 Results

#### 6.4.6.1 Modelling with PCs, with and without Centroid Size

To classify species, various predictive models were utilized and evaluated for accuracy using two metrics obtained from the confusion matrix. The assessment began by gradually introducing the 15 principal components (PCs), selected based on the scree plot depicting their maximum variability coverage, in a sequential fashion. Performance metrics were derived through leave-one-out cross-validation to determine model effectiveness.

The average error rate (OOB) and maximum classification error rate per species were computed for each number of PCs used in the predictive models. These metrics were calculated both with and without the inclusion of the centroid size (Csize) variable in the models. The results were plotted to visualize the relationship between the number of PCs and the corresponding error rates.

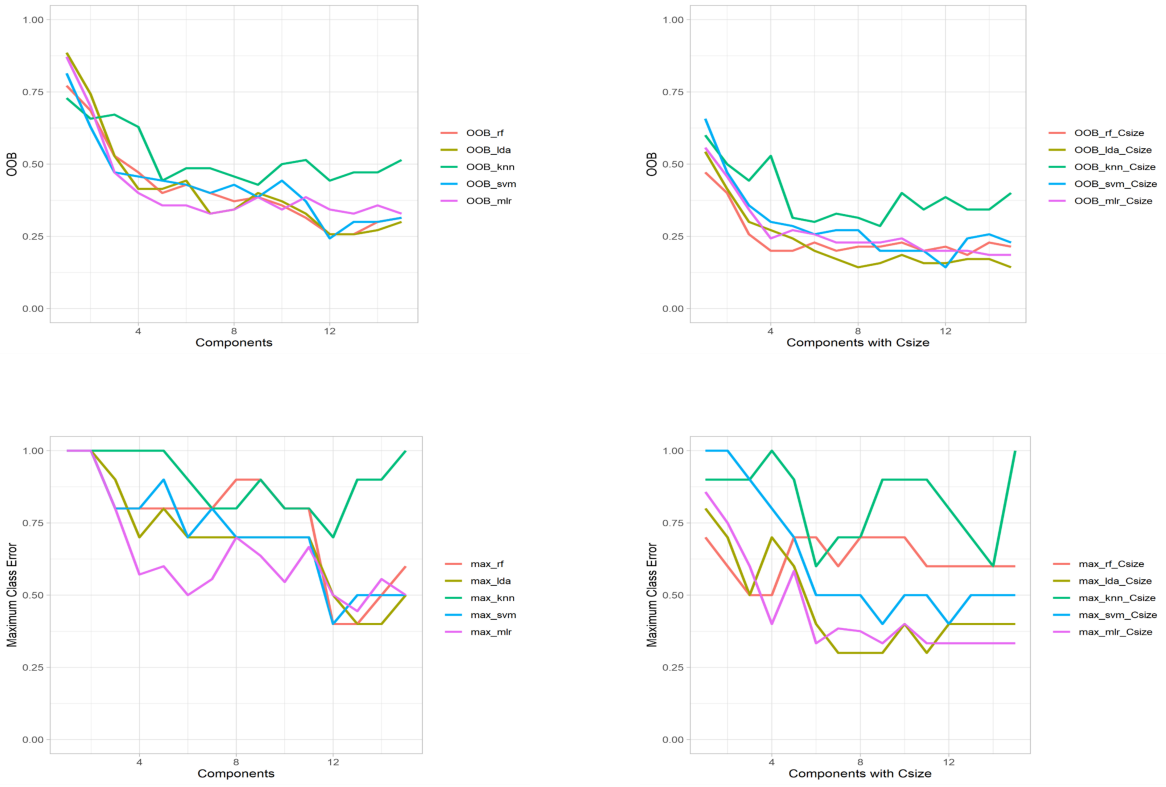



**Figure: Average error rate (OOB) and Maximum misclassification rate for any category when PCs are considered in the model adding one by one along with and without Csize**

Incorporating the centroid size variable notably decreased error rates in the models.

As a significant finding, error plots stabilized sooner when Csize was included, indicating the efficacy of fewer principal components in evaluating model performance. The observed decrease in error rates at lower PC counts is anticipated as Csize offers valuable species-distinguishing information (demonstrated in the boxplot in section 6.2), allowing similar classification performance with fewer additional PC variables.

The inclusion of Csize exhibited a consistent decrease pattern in the error rates, with minimum levels dropping from approximately 25% to 13%. This trend was observed across all the predictive models. Notably, models such as random forest, linear discriminant analysis, and support vector machines consistently achieved lower error rates, regardless of whether the centroid size variable was included in the model or not.

Upon comparing all the models, linear discriminant analysis demonstrated superior accuracy, particularly when the centroid size was considered in the model. This trend was consistent with the findings from the maximum classification error rate per species graphs.

In conclusion, the analysis demonstrated that incorporating the centroid size variable in the predictive models resulted in improved accuracy. Among the different models examined, linear discriminant analysis consistently outperformed the others, exhibiting lower error rates.

#### 6.4.6.2 Modelling with bgPCs, with and without Centroid Size

Similar to the previous scenario, we systematically added 6 between-group principal components (bgPCs) to the predictive models, considering both the inclusion and exclusion of the centroid size variable (Csize). The performance of the models was evaluated at each step, considering the metrics average error rate (OOB) and the maximum classification error rate for each species category. These accuracy metrics were plotted against the number of bgPCs used during the execution of the predictive models.

The inclusion of the centroid size variable played a vital role in reducing error rates,

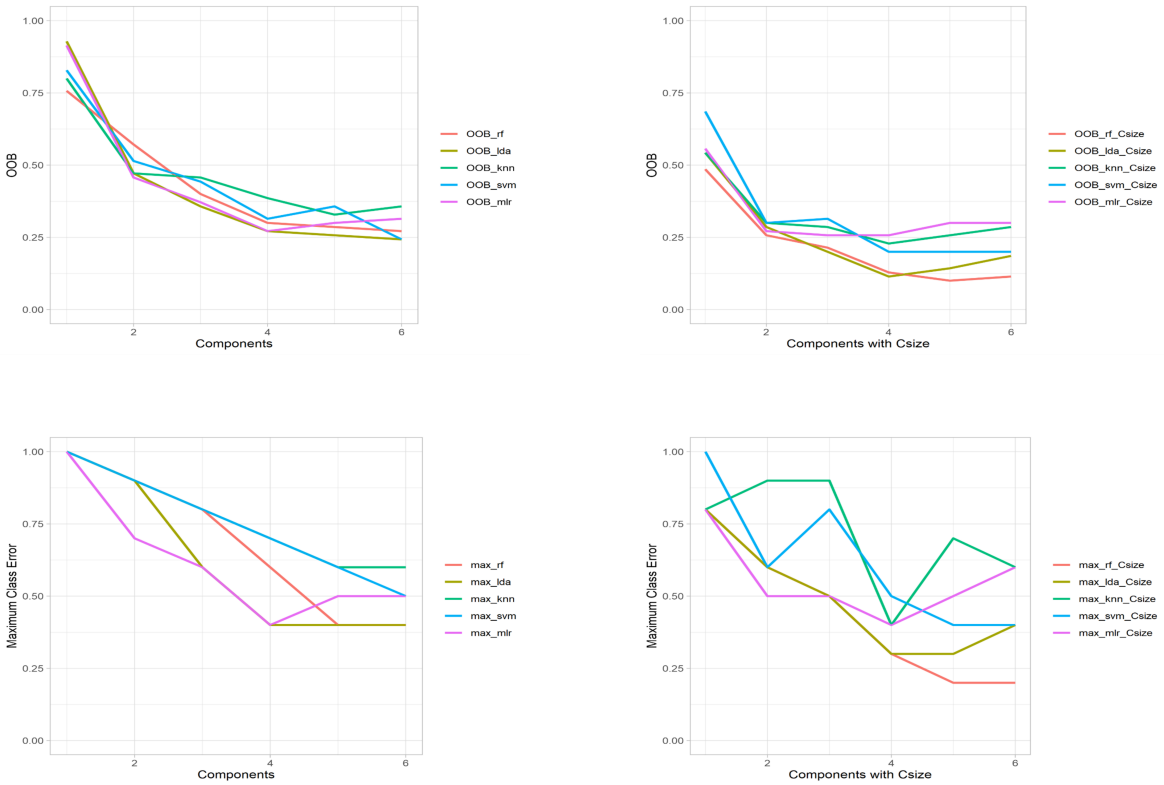



**Figure: Average error rate (OOB) and Maximum misclassification rate for any category when bgPCs are considered in the model adding one by one along with and without Csize**

resulting in a substantial decrease in the minimum error level. Specifically, the error rates decreased by approximately 25% to below 12%, representing an improvement over the models that solely considered principal components (PCs). Notably, the random forest and linear discriminant analysis models consistently outperformed the other models. In particular, the random forest model exhibited slightly superior performance, particularly when the centroid size variable was included in the analysis. This trend was consistent with the maximum classification error rate graph, which displayed lower error rates for the random forest model across the species categories.

To summarize, incorporating the centroid size variable in the predictive models significantly contributed to the reduction of error rates. The random forest model, in particular, exhibited better overall performance, particularly when the centroid size variable was included alongside the bgPCs, rather than with the principal components (PCs).

# 7 Applied Research Work: Facial Morphology Analysis

## 7.1 Introduction and Aims

Face GWAS is an area of genetic research focused on identifying genetic variants associated with facial morphology and traits. It involves analyzing large-scale genomic data from individuals to investigate the genetic basis of facial features and their variations within populations [Adhikari et al. (2016b)]. Though the initial face GWAS studies focused primarily on individuals of European descent [Liu et al. (2012)], researchers have gradually expanded their investigations to include non-European populations. This expansion has been instrumental in obtaining a more comprehensive understanding of the genetic architecture underlying facial variation in humans.

Geometric morphometrics is an area of study that analyzes the shape variation in facial features using geometric landmarks [Webster and Sheets (2010)]. Procrustes distances are a crucial component of this approach, as they quantify the differences in shape between objects or individuals after superimposing and aligning their landmarks. These distances provide a reliable measure of shape variation and have been widely used to compare and analyze morphological differences in various biological studies. In the context of genetic association studies, these Procrustes distances are usually used as phenotypic measurements and researchers investigate shape-based association analyses to identify genetic variants linked to specific shape characteristics.

Traditionally GWASs of facial morphology use different phenotyping approaches such as including qualitative assessment of morphological features on 2D photographs, measurements derived from manual landmarking of 2D photographs, and semi-automatic analyses of 3D facial images. Each approach has its own drawbacks, such as variations

in cost, informativity, ease of application, and labor intensiveness. However, fully automatic landmarking approaches can be a feasible option, which combines automatic landmarking with manual editing [Liu et al. (2012)].

This study focused on analyzing facial features obtained through a fully automatic landmarking technique applied to 2D frontal photographs of Latin Americans with diverse European, Native American, and African ancestry. The objective was to identify genetic loci that are significantly associated with various Procrustes distances, treated as phenotypic traits. This study identified 33 novel genes associated with the facial shape variation in diverse ethnicities within the CANDELA cohort and some of them overlap with previous GWAS findings. In addition to that, most of the novel signals identified here show evidence of statistical replication in other datasets such as European, East Asian, or African GWAS data.

Notably, one of the novel genes, possibly inherited from the Neanderthals among Native Americans and East Asians, has been found to contribute to increasing nasal heights and this result is consistent with the morphological differentiation between Neanderthals and modern humans. These findings have been published in the journal 'Communication Biology'. (DOI: `https://doi.org/10.1038/s42003-023-04838-7`)

## 7.2 Methods

### 7.2.1 Study Sample and Phenotyping

The sample population in this study consisted of 6,486 individuals from the CANDELA cohort, which was collected across five Latin American countries [Ruiz-Linares et al. (2014)]. These individuals were genotyped using Illumina's OmniExpress chip, which included over 700,000 SNPs. Additionally, they were assessed for various standard covariates, including age, sex, BMI, and genetic ancestry estimated from the chip data [Adhikari et al. (2016b)].

In this study, the Face++ cloud service platform (`https://www.faceplusplus.com`) was used to automatically locate 106 landmarks on frontal 2D photographs obtained from CANDELA individuals. A subset of these individuals had previously undergone manual placement of 16 landmarks, which served as a reference for evaluating the robustness of the Face++ landmarking method [Adhikari et al. (2016b)].

Interclass correlation coefficients and median Euclidean distances were calculated between the manually placed landmarks and those obtained using both the Face++ platform. These metrics allowed for a comprehensive comparison and evaluation of the accuracy and consistency of the automated landmarking methods.

Following the Procrustes superposition, inter-landmark distances (ILDs) between 34 landmarks were obtained from the Face++ landmarking method. These landmarks primarily corresponded to distinct anatomical features [Adhikari et al. (2016b)] which are shown in the following figure.

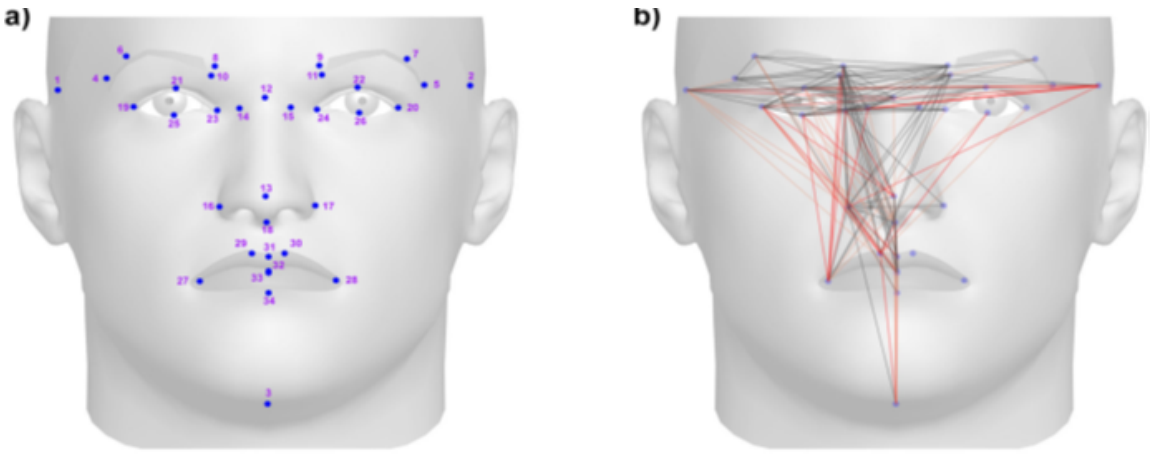



**Figure (a): The dots on the plot represent the positions of the 34 facial landmarks that were utilized for the computation of 301 inter-landmark distances; Figure (b): The lines on the plot depict the 148 inter-landmark distances (ILDs) that exhibited a significant association with at least one genomic region in the CANDELA dataset.**

Taking face symmetry into consideration, we derived a total of 301 distances. These distances exhibited notable variation and displayed an approximately normal distribution. A significant correlation was observed between several distances and three head

angles (pitch, roll, and yaw angle) estimated by Face++, indicating the influence of head pose. To account for this effect, 76 individuals were excluded with extreme head angle values and incorporated these angles as covariates in the genetic association tests.

## 7.3 Results

### 7.3.1 Overview of GWAS results

Following the application of quality control (QC) filters to the genotype and phenotype data, association analyses were conducted between inter-landmark distances (ILDs) and up to 11,532,785 SNPs across a maximum of 5,988 individuals. A significance threshold of P-value $< 5e{-}8$ was used to identify significant associations, in accordance with the convention for GWAS. In total, 42 genomic regions demonstrated significant associations with at least one ILD, and among these regions, 148 distances exhibited associations with at least one of the 42 genomic regions. Notably, nine of these regions had previously been reported in GWAS studies on facial features, including six regions that were identified in previous face GWASs conducted within the CANDELA cohort.

### 7.3.2 Follow-up of newly associated regions: Replication in independent cohorts

The replication of the newly identified genome regions was validated by examining the results from independent studies conducted on individuals with different continental ancestries, including East Asians, Europeans, and Africans. This consideration of diverse ancestries aligns with the admixed ancestry of the CANDELA individuals.

For East Asians, data was gathered from available frontal 2D photographs and genome-wide SNP data for 5078 individuals [Zhang et al. (2022)]. In the case of Europeans and Africans, associated P-values were extracted from a GWAS meta-analysis involving data from 10115 Europeans and 78 interlandmark distances (ILDs) [Xiong

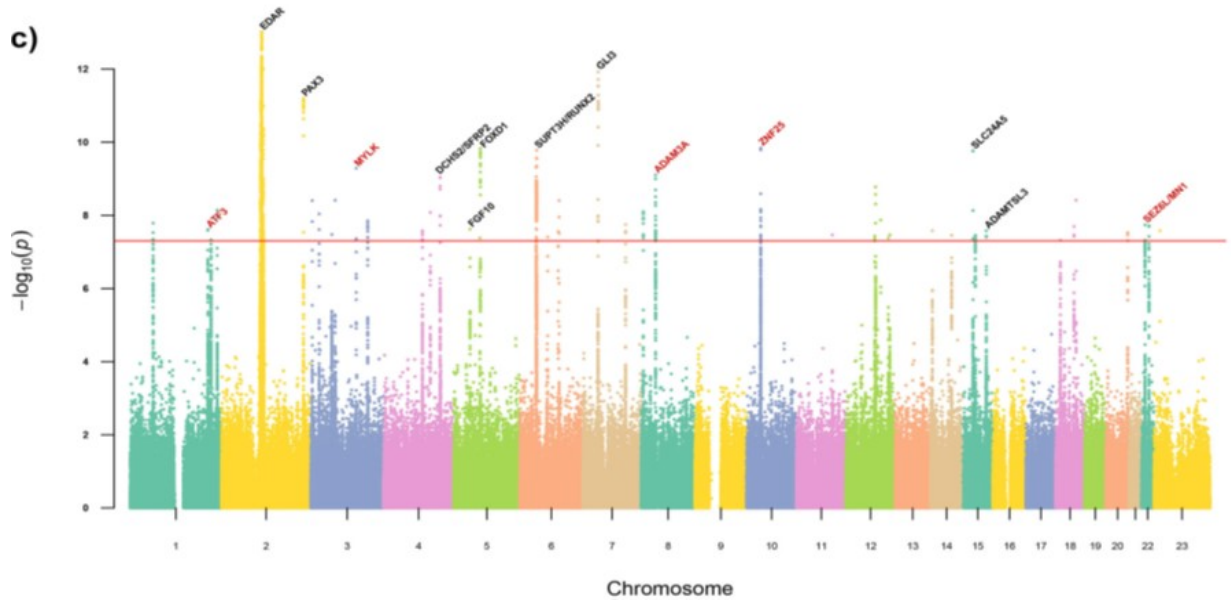



**Figure: The combined Manhattan plot displays all significant GWAS findings, with a threshold set at log(P) > 7.3 (indicated by the red line). Candidate genes identified in previous GWAS studies are represented by black labels, while the main candidate genes within the five novel regions are highlighted with red labels.**

et al. (2019)], as well as from a GWAS conducted on 3631 Africans, focusing on 34 size and shape-related facial traits [Null et al. (2022)]. To determine the significance threshold for replication, the Benjamini-Hochberg FDR procedure was employed. Among the 33 regions of interest, 26 showed associations for at least one distance in at least one of the replication datasets, providing evidence of statistical replication of these genomic regions across different populations.

### 7.3.3 Neanderthal introgression in a genomic region 1q32.3 and Nasal height comparison across the various Ethnicities

One of the novel regions identified in this study, located in chromosome region $1q32.3$, showed replication and association with various Procrustes distances, particularly affecting nasal heights. Interestingly, previous research has reported Neanderthal introgression in this same region, and the evidence of introgression was also observed in the CANDELA data. Approximately 31% of CANDELA chromosomes were found to carry

Neanderthal inherited genetic tracts in this region, which displayed significant associations with Procrustes distances, ultimately leading to an increase in nasal heights.

To gain further insights, a comparison of nasal heights was made between modern humans and Neanderthals using available data on Neanderthals for equivalent Procrustes distances. This comparison included 1190 modern human skulls from three continental populations and data from 10 Neanderthals, revealing that Neanderthals exhibited notably higher nasal heights. The modern human data was obtained from Howell's database [Oxnard (1974)], while the Neanderthal data was sourced from Weaver and Stringer [Weaver and Stringer (2015)].

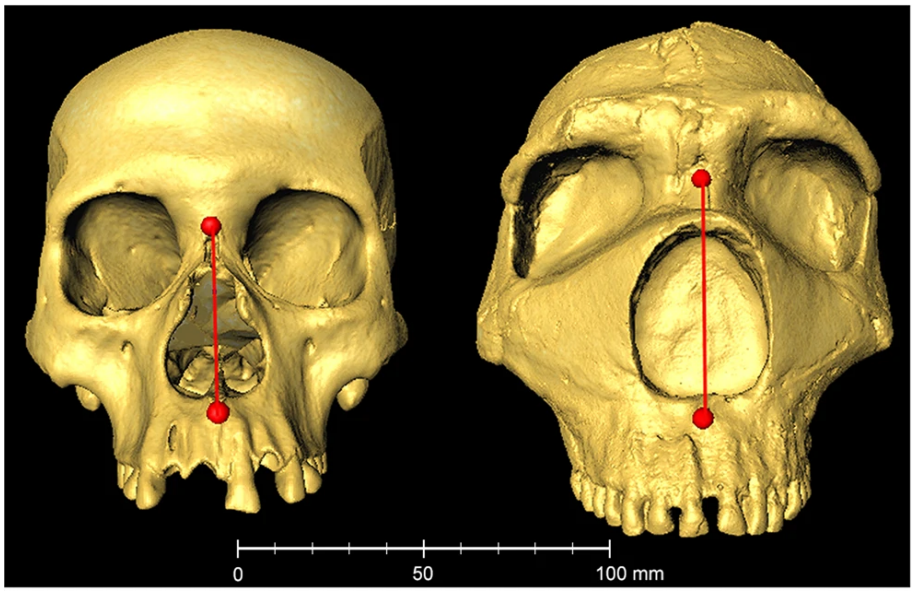



**Figure: Displayed here are 3D images of a modern human skull and a Neanderthal skull. Nasal height is represented as a red line, measuring 50.2 mm for the modern human and 63.8 mm for the Neanderthal.**

Upon comparing the genomes, it becomes evident that Neanderthal tracts are present on a Native American chromosomal background. This finding aligns with previous analyses that identified 1q32.3 introgression primarily in Native Americans [Sankararaman et al. (2016)].

## 7.4 My Contribution

Although this research mainly focused on identifying genetic variants associated with facial traits, the critical concept of power gain through conditioning, a significant chapter in my thesis, was not applied in this particular area. As part of the research collaboration, I extensively examined the CANDELA cohort dataset, exploring various variables. This involved investigating a conversion method to enable comparisons across diverse ancestries and with Neanderthals.

The CANDELA cohort dataset has two different types of variable setups. The GWAS above, which found a significant association of the Neanderthal-inherited gene, was based on Procrustes distance between landmarks. However, when comparing between different species, i.e. the humans and Neanderthals, it is important to compare the actual size (i.e. in mm), since that is the primary point of difference between the species, as illustrated above.

To enable this comparison, I developed a conversion method between the two types of measurements: size-based measurements (in mm) and shape-based measurements (obtained through the Procrustes process). For this purpose, I used the two kinds of variables that are available in CANDELA. The first dataset contained background information, such as country, head size, etc., and various distance measurements for different facial features in actual scales (e.g., in mm), for example, 'NASION_GNASION' is for nasal height, 'CH_CH_BREADTH' is for the width of noses, etc. The second dataset consisted of landmark-based Procrustes distances for various facial features.

Initially, I started with the various types of variables in the entire dataset and produced summary statistics for different countries including Brazil, Chile, Colombia, Mexico, and Peru. Next, I extracted the variable 'NASION-GNASION', which pertained to nasal height distance, from the CANDELA cohort dataset. I then examined relevant summary statistics for this variable. The background information provided details

on the number of individuals in each country and how many had information on 'NASION_GNASION'.

By doing a pairwise comparison of 'NASION_GNASION' and the equivalent landmark-based Procrustes distance of nasal height, I transformed the landmark-based Procrustes distances for CANDELA individuals into actual scales (e.g., in mm), similar to the Neanderthal individuals. I plotted the nasal distances of CANDELA individuals using box plots, categorized by continent (i.e., Africans, Europeans, and Native Americans). Additionally, I included the Neanderthal individuals in the same graph to observe which continent closely resembled the Neanderthal facial features.

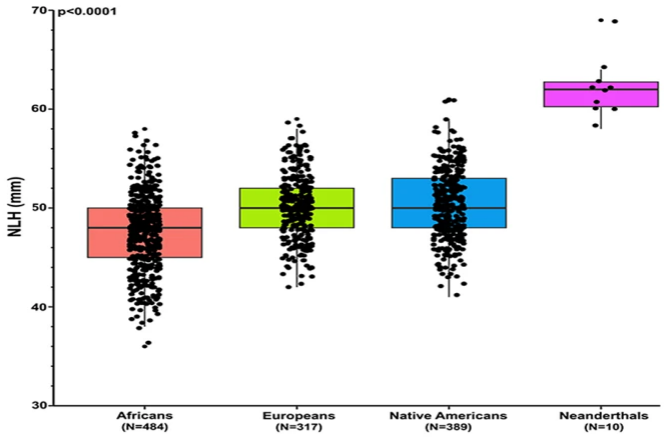



**Figure: The box plot illustrates the variation in nasal height between modern humans and Neanderthals. The P-value ($< 0.0001$) indicates a significant contrast in nasal height data between the two groups.**

In the end, to evaluate the consistency of the introgression effect, a genome-wide association study (GWAS) was conducted using the reconstructed actual size measurements, and the results showed that the genetic associations and effect directions were consistent with the Procrustes distance-based GWAS.

# 8 Overall Conclusion

In this doctoral research, I've delved into multivariate statistical models for analyzing genotype-phenotype data. This exploration encompassed both theoretical and computational methodologies. I've shown the connection between linear mixed models, ordinary least square regression, and shrinkage, simplifying their interrelations. Chapter 3 navigates through various multivariate testing techniques, preceded by an extensive literature review. I have considered the multivariate test procedures for the Canonical Correlation setup due to their availability in the literature. Here, I delved into their mathematical properties within multiple linear regression, establishing their convergence to the chi-square distribution. Moreover, I demonstrated their equivalence with each other and the Wald Test, recommending the Wald Test for future studies due to its compatibility across scenarios.

Chapter 4 contains a significant contribution to the thesis which involves the gain of power in identifying new associations from conditioning on major genes. Through various multiple regression settings, I explored the impact of conditioning on statistical power theoretically and computationally. Theoretical developments revolved around multiple linear regression models, encompassing diverse design matrix forms. I elucidated power gain considering genetic variant behavior within chromosomes, incorporating LD structures and proxy variants. To illustrate this concept, I performed analyses using two genetic databases: the CANDELA cohort and UKBiobank. Specifically, I derived mathematical expressions for statistical power when the design matrix takes the form of a 3-Block matrix, clarifying the degree of enhancement in power achieved by conditioning on a single block. Furthermore, I introduced a mathematical formula, a 3-block approach, to compute conditional results from summary statistics, even without individual-level datasets. Comparing the conditional beta coefficients obtained from

this 3-block approach with those derived from the GCTA software, I found that the former closely aligned with the true coefficients for all the tested SNPs, outperforming the GCTA-derived coefficients.

In prediction, the central aim revolves around optimizing predictive precision over focusing solely on testing or statistical power. However, the statistical models used in testing, such as linear models and shrinkage methods, can also be carried out for predictive purposes in new datasets and the predictive ability can be calculated with prediction accuracy or errors. Chapter 5 delves into the mathematical depiction of prediction errors across various methods like simple linear regression models and ridge regression. These formulations elucidate the connection between prediction errors and genetic association studies, facilitating an understanding of genetic heritability by gauging model prediction errors.

Chapter 6 and Chapter 7 concentrate on collaborative research related to "Dental Morphology Data" and "Facial Morphology Data", respectively. In Chapter 6, various predictive models such as random forest, linear discriminant analysis, multinomial logistic regression, K-nearest neighbor, and support vector machines were employed to assess species prediction accuracy based on dental morphology data. The evaluation involved calculating the "Confusion Matrix", offering misclassification proportions for each category. Since the Confusion Matrix is usually very large, the performance of a predictive model is evaluated with the "average error rate". But it may be the situation that the average error rates are similar or very close for different models, and the error rates of each class vary quite a bit. Therefore, additional metrics, like the maximum misclassification rate for any category, were also computed. The analysis demonstrated that incorporating the centroid size variable in the predictive models resulted in improved accuracy and upon comparing all the models, linear discriminant analysis demonstrated superior accuracy. Chapter 7 centered on analyzing facial traits from automatic landmarking of

2D frontal photographs from individuals with varied ancestries. The aim was to pinpoint genetic loci linked to distinct Procrustes distances treated as phenotypic traits. Extensive examination of the CANDELA cohort dataset involved exploring variables and devising a conversion method for cross-ancestral comparisons, including with Neanderthals. This exploration revealed 33 new genes associated with facial shape variations in diverse ethnicities within the CANDELA cohort. Some of these genes overlapped with previous GWAS findings, with most novel signals showing statistical replication in other datasets, such as European, East Asian, or African GWAS data. Notably, a novel gene inherited from Neanderthals among Native Americans and East Asians was found to impact nasal heights, consistent with Neanderthal and modern human morphological differences. These discoveries were published in the journal "Communication Biology".

The mathematical derivation illustrating power gain through conditioning has been extensively demonstrated in various regression models primarily applied when the response or phenotype variable is continuous. However, in practical scenarios, some non-metric traits, like incisor shoveling, exhibit a "quasi-continuous" nature. This means these traits can be seen as dichotomous, arising from an underlying continuous quantity. In such cases, this continuous underlying variable is termed a "latent variable" that corresponds to the evaluated categorical variable. In that case, the concept of conditioning can also be useful in generalized linear models (GLMs) settings when a clear latent variable is not present, though we don't provide the mathematical details in the context of GLMs.

In genome-wide association studies (GWAS), the trait of interest is often binary, indicating a categorical variable with two categories denoting the presence or absence of the outcome. Logistic regression is typically used to analyze the association between genetic variants and such categorical outcomes, considering confounding factors. However, the coefficient of regression ($\beta$) is unbiased in the presence of uncorrelated confounders

in linear regression setup but not in the logistic regression case. Consequently, conditional analyses in GWAS using logistic regression have been limited, mainly focusing on disease phenotypes in case-control studies. Binary responses are common in various study designs like case-control studies, cohort studies, clinical trials, and twin studies. Case-control studies often employ logistic regression to explore the relationship between genetic variants and binary responses within a single cohort. However, conditioning in these studies poses challenges due to the complex interplay of disease prevalence and allelic effect size on conditioned GWAS power. While conditioning is well-established, further exploration of its impact on statistical power in case-control studies remains an area for future research.

# A APPENDICES

## A.1 Computational methods for mixed models

Bates (2010) discussed the computational methods for the "lme4" package which provides R functions to fit and analyze linear mixed models, generalized linear mixed models, and nonlinear mixed models. These models are called mixed-effects models because they incorporate both fixed-effects parameters, which apply to an entire population or to certain well-defined and repeatable subsets of a population, and random effects, which apply to the particular experimental units or observational units in the study. Here, it has been discussed the general form of the mixed models that can be represented in the "lme4" package and the computational approach embodied in the package. In the mixed effect model, the $n$-dimensional response variable $y$, maybe on a continuous scale or they may be on a discrete scale, such as binary responses or responses representing a count, and the $q$-dimensional random effect vector, $\boldsymbol{B}$ is always continuous. The conditional distribution of $(y|\boldsymbol{B} = b)$ is the multivariate Gaussian distribution of the form

$$(y|\boldsymbol{B} = b) \sim N(Zb + X\beta, \sigma^2 I_n)$$

where, $I_n$ denotes the identity matrix of size, $n$. The conditional mean, $E(y|\boldsymbol{B} = b)$, depends on $b$ only through the value of the linear predictor, $Zb + X\beta$, $X$ is the $n \times p$ model matrix, $\beta$ is a $p$-dimensional coefficient vector, $Z$ is the $n \times q$ model matrix for the $q$-dimensional vector valued random effect variables, $\boldsymbol{B}$. The unconditional distribution of $\boldsymbol{B}$ is also multivariate Gaussian distribution as

$$\boldsymbol{B} \sim N(0, \Sigma)$$

As a variance-covariance matrix must be positive semidefinite. It is convenient to express the mixed model in terms of a relative covariance factor, $\Lambda_\theta$ which is a $q \times q$ matrix, depending on the variance-component parameter, $\theta$, and generating the symmetric $q \times q$ variance-covariance matrix, $\Sigma$, according to

$$\Sigma_\theta = \sigma^2 \Lambda_\theta \Lambda_\theta^T$$

A linear transformation, ($\boldsymbol{B} = \Lambda_\theta u$ )has been made on the random effect $\boldsymbol{B}$ to formulate it in terms of a spherical random effect variable $u$ such that the conditional mean or the linear predictor may be reformed as

$$\gamma(u) = E(y|\boldsymbol{B} = b) = Z\Lambda_\theta u + X\beta$$

here, $u \sim N(0, \sigma^2 I_q)$ and this spherical formulation allows to work with singular covariance matrices, which regularly arise in practice.

The estimates of the parameters in a mixed model are determined as the values that optimize an objective function either the likelihood of the parameters given the observed data, for maximum likelihood (ML) estimates, or a related objective function called the REML criterion. As this objective function must be evaluated at many different values of the model parameters during the optimization process, the evaluation of the objective function and the computational methods for maximum likelihood fitting the linear mixed model involve repeated applications of the penalized least squares (PLS) method. In particular, the PLS problem is to minimize the penalized weighted residual sum of squares,

$$r^2(\theta, \beta, u) = \rho^2(\theta, \beta, u) + ||u||^2$$

over $(\theta, \beta)^T$, where, $\rho^2(\theta, \beta, u)$ is the weighted residual sum of squares. The reason for the word "penalized" is that the term $||u||^2$, penalizes models with larger magnitude

values of $u$.

For the purpose of statistical inference, it is always interesting to deal with the conditional probability density of $u|y$, and the unnormalized conditional density can be written as

$$h(u|y,\theta,\beta,\sigma) = f_{y|u}(y|u,\theta,\beta,\sigma)f_u(u|\sigma)$$

and to obtain the conditional density, $h$ needs to be normalized by dividing by the value of the integral

$$L(\theta,\beta,\sigma|y) = \int_{R^q} h(u|y,\theta,\beta,\sigma)du$$

On the deviance scale, the unnormalized conditional density can be written as

$$\begin{aligned} -2log(h(u|y,\theta,\beta,\sigma)) &= (n+q)log(2\pi\sigma^2) + \frac{||y - Z\Lambda(\theta)u - X\beta||^2 + ||u||^2}{\sigma^2} \\ &= (n+q)log(2\pi\sigma^2) + \frac{d(u|y,\theta,\beta)}{\sigma^2} \end{aligned}$$

here, $d(u|y,\theta,\beta) = ||y - Z\Lambda(\theta)u - X\beta||^2 + ||u||^2$ is the discrepancy function and it has the form of a penalized residual sum of squares in which $||y - Z\Lambda(\theta)u - X\beta||^2$ is the residual sum of squares for $y, u, \theta$ and $\beta$ and the second term, $||u||^2$, is a penalty on the size of $u$.

In the so-called "pseudo-data" approach, the discrepancy function can be written as the squared length of a block matrix equation

$$d(u|y,\theta,\beta) = \left|\left| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} Z\Lambda(\theta) & X \\ I_q & 0 \end{bmatrix} \begin{bmatrix} u \\ \beta \end{bmatrix} \right|\right|^2$$

The term"pseudo data" reflects the fact that we have added $q$ "pseudo observations" to the observed response, $y$, and to the linear predictor, $\gamma(u) = Z\Lambda_\theta u + X\beta$, in such a way that their contribution to the overall residual sum of squares is exactly the penalty

term in the discrepancy. It is seen that the form of the discrepancy is a quadratic form in both $u$ and $\beta$. Furthermore, because we require that $X$ has full column rank, the discrepancy is a positive-definite quadratic form in $u$ and $\beta$ that is minimized at $\tilde{u}$ and $\tilde{\beta}$ satisfying

$$\begin{bmatrix} \Lambda^T(\theta)Z^TZ\Lambda(\theta) + I_q & \Lambda^T(\theta)Z^TX \\ X^TZ\Lambda(\theta) & X^TX \end{bmatrix} \begin{bmatrix} \tilde{u}(\theta) \\ \tilde{\beta}(\theta) \end{bmatrix} = \begin{bmatrix} \Lambda^T(\theta)Z^Ty \\ X^Ty \end{bmatrix}$$

## A.2 Parameter estimation and inference in the linear mixed model

Gumedze and Dunne (2011) discussed the parameter estimation for the different components of the linear mixed model and inference procedures for the fixed effects, random effects, or a combination of both, and random effects are discussed in this paper. The widely used linear mixed model (LMM) is given by

$$y = X\beta + Z\mu + \epsilon$$

where, $y_{n\times 1}$ is a vector of responses, $X_{n\times p}$ is a design matrix for the fixed effects, $\beta_{p\times 1}$ is a vector of fixed effect parameters, $Z_{n\times q}$ is a design matrix for the random effects, $u_{q\times 1}$ is a vector of random effect. It is assumed that $u$ and $\epsilon$ follow independent and multivariate Gaussian distribution such that

$$\begin{bmatrix} u \\ \epsilon \end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} G(\gamma) & 0 \\ 0 & R(\rho) \end{bmatrix} \right)$$

where,$\gamma$ and $\rho$ are $r \times 1$ and $s \times 1$ (with $s \leq n(n+1)/2$) vectors of unknown variance parameters corresponding to $u$ and $\epsilon$, respectively. If the random terms are correlated then the dimension of $\gamma$ may excced $q$, i.e. $\gamma$ may be of dimension $r \leq q(q+1)/2$. Following **?**, the variance-covariance matrix of the data, $y$ can be written as

$$var(y) = \sigma^2(ZGZ^T + R) = \sigma^2 H$$

where

$$H = ZGZ^T + R$$

The matrix $H$ consists of two components that are used to model heteroscedasticity

and correlation: a random-effects component $ZGZ^T$ and a within-group component $R$. In some applications, the within-group component $R$ is used to directly model the variance–covariance matrix of the data without incorporating random effects in the model to account for dependence among observations.

**Joint estimation of fixed and random effects**

There are many methods for obtaining the joint estimates of the fixed and random effects [Searle et al. (1992)]. These methods include Henderson's mixed model equation, Goldberger's approach of predicting future observations, techniques based on two-stage regression, linearity in $y$, partitioning of $y$, and Bayes estimation. Here, Henderson's mixed model equations have been discussed because it produces sampling variances for the estimators and it has a connection with the maximum likelihood estimation of the variance parameters. Hendeson assumed $u$ and $y$ to be jointly Gaussian distributed as

$$\begin{bmatrix} u \\ y \end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ X\beta \end{bmatrix}, \sigma^2 \begin{bmatrix} G & GZ^T \\ GZ & H \end{bmatrix} \right)$$

Henderson maximized the log-joint distribution of $(y, u)$ to obtain estimators of $\beta$ and $u$. However, this logarithmic function is not a log-likelihood function as $u$ is not observed. The marginal distribution of u is

$$u \sim N(0, \sigma^2 G)$$

and the conditional distribution of $y$ given $u$is

$$y|u \sim N(X\beta + Zu, \sigma^2 R)$$

Hence, the log-joint distribution of $(y, u)$ is given by

$$\begin{aligned} logf(y,u) &= logf(y|u) + logf(u) \\ &= -\frac{1}{2}nlog\sigma^2 + logR + (y - X\beta - Zu)^T R^{-1}(y - X\beta - Zu)/\sigma^2 \\ &- \frac{1}{2}qlog\sigma^2 + logG + u^T G^{-1}u/\sigma^2 \\ &= -\frac{1}{2}(n+q)log\sigma^2 + logR + logG + (y - X\beta)^T R^{-1}(y - X\beta)/\sigma^2 \\ &- \frac{1}{2\sigma^2}u^T(ZR^{-1}Z^T + G^{-1})u - 2(y - X\beta)^T R^{-1}Zu \end{aligned}$$

Now, the estimates of $\beta$ and $u$ can be obtained by solving the score equations

$$X^T R^{-1}(y - X\hat{\beta}) - X^T R^{-1}Z\tilde{u} = 0$$

$$Z^T R^{-1}(y - X\hat{\beta}) - (Z^T R^{-1}Z + G^{-1})\tilde{u} = 0$$

These equations are called the mixed model equations (MMEs) as proposed by Henderson (1975) and the equations can be written in matrix form as

$$\begin{bmatrix} X^T R^{-1}X & X^T R^{-1}Z \\ Z^T R^{-1}X & Z^T R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \tilde{u} \end{bmatrix} = \begin{bmatrix} X^T R^{-1}y \\ Z^T R^{-1}y \end{bmatrix} \tag{A.1}$$

Gilmour et al. (1995) rewrote the mixed model equation (2.8) as

$$C\psi = W^T R^{-1}y \tag{A.2}$$

where, $W = [XZ], \psi = (\beta^T, u^T)^T$ and $C = W^T R^{-1}W + G^{*+}$ with

$$G^* = \begin{bmatrix} 0 & 0 \\ 0 & G \end{bmatrix}$$

and

$$G^{*+} = \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} \end{bmatrix}$$

where, the superscript " $+$ " denotes the Moore-Penrose inverse. Assuming $H$ is known, the fixed effects parameters $\beta$ can be estimated by generalized least square (GLS) to obtain

$$\hat{\beta} = (X^T H^{-1} X)^{-1} X^T H^{-1} y \tag{A.3}$$

which is the best linear unbiased estimator (BLUE) of $\beta$. If $X$ is not full rank, then any generalized inverse $(X^T H^{-1} X)^-$ is used instead of $(X^T H^{-1} X)^{-1}$ to obtain a solution for $\beta$. The resulting solution for $\beta$ is not unique and is no longer unbiased. However, $X\hat{\beta}$ is unique and unbiased for $X\beta$.The computational challenge of using GLS to estimate $\beta$ is that it requires the inverse of $H$ which is an $n \times n$ matrix. In contrast, the joint estimators for $\beta$ and $u$ can be obtained by solving either (A.1) or (A.2), i.e.,

$$\tilde{\psi} = C^{-1} W^T R^{-1} y$$

*Lemma 1:*

The solution for $\beta$ and $u$ from solving the MMEs, for $G$ and $R$ known, are given by

$$\hat{\beta} = (X^T H^{-1} X)^{-1} X^T H^{-1} y$$

$$\tilde{u} = G Z^T H^{-1} (y - X\hat{\beta})$$

with corresponding variance matrices

$$var(\hat{\beta}) = \sigma^2 [(X^T H^{-1} X)^{-1} X^T H H^{-1} X (X^T H^{-1} X)^{-1}] = \sigma^2 (X^T H^{-1} X)^{-1}$$

and

$$var(\tilde{u}) = \sigma^2 GZ^T PHPZG = \sigma^2 GZ^T PZG \tag{A.4}$$

respectively, where $P = H^{-1} - H^{-1}X(X^T H^{-1} X)^{-1} X^T H^{-1}$ . It can also be written that

$$var(\tilde{u} - u) = \sigma^2 G - var(\tilde{u})$$

which unlike (A.4) takes into account the variability of $u$ and can therefore be useful for constructing confidence intervals for $u$.

**Parameter Estimation for Variance**

Maximum likelihood (ML) and Residual Maximum Likelihood (REML), also known as restricted maximum likelihood, are now standard methods for estimating variance parameters for both balanced and unbalanced data. The main attraction of these methods is that they can handle a much wider class of variance models than simple variance components. ML estimators of the variance parameters are biased downwards, especially in small samples, because they do not take into account the degrees of freedom lost in the estimation of the fixed effects [Lin and McAllister (1984); Swallow and Monahan (1984)]. Hence, REML estimation of the variance parameters is preferable to ML estimation. ML estimation of the variance parameters has been discussed by several researchers (e.g.,[Hartley and Rao (1967)]. ML and REML estimation for variance parameters in linear mixed models has been discussed here:

**Maximum likelihood Method**

The marginal distribution of $y$ in the linear mixed model is given by $N(X\beta, \sigma^2 H)$ and hence the marginal log-likelihood function of $y$ is [Hartley and Rao (1967)]

$$l_{ML}(\beta, \phi; y) = -\frac{1}{2}\Big[nlog2\pi + nlog\sigma^2 + log|H| + \frac{(y - X\beta)^T H^{-1}(y - X\beta)}{\sigma^2}\Big]$$

where, $\phi = (k^T, \sigma^2)^T, k = (\gamma^T, \rho^T)^T$. Differentiating the marginal log-likelihood function with respect to $\beta, \sigma^2$ and $k_j; j = 1, \cdots, r+s$ yields the partial derivatives and setting equal to zero gives

$$X^T \hat{H}^{-1} X\hat{\beta} = X^T \hat{H}^{-1} y$$

$$n\hat{\sigma}^2 = (y - X\hat{\beta})^T \hat{H}^{-1}(y - X\hat{\beta})$$

$$tr(\hat{H}^{-1}\ \tilde{\hat{H}}_j) = \frac{1}{2}(y - X\hat{\beta})^T \hat{H}^{-1}\ \tilde{\hat{H}}_j \hat{H}^{-1}(y - X\hat{\beta}) \tag{A.5}$$

Solving the above equation yields the maximum likelihood estimators

$$\hat{\beta} = (X^T \hat{H}^{-1} X)^{-1} X^T \hat{H}^{-1} y$$

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})^T \hat{H}^{-1}(y - X\hat{\beta})$$

Solving (A.5), the solution for $k_j$ must be found which depends on $\hat{\beta}$ and $\hat{\sigma}^2$.

**Residual Maximum Likelihood (REML)**

The downward biasedness of ML estimators of the variance parameters, hidden in $H$, can be overcome by using residual maximum likelihood (REML) estimation [Anderson and Bancroft (1952); Patterson et al. (2006)]. REML maximizes the likelihood of linearly independent error contrasts, i.e. independent contrasts of linear combinations of the data $y$, orthogonal to the design matrix $X$. The linear combinations are chosen as $K^T y$ so that $K^T y$ is of maximal rank but is free of the fixed effects $\beta$. These linear combinations are the residuals obtained after fitting the fixed effects hence the name residual maximum likelihood. For $y \sim N(X\beta, \sigma^2 H)$ and $K^T X = 0$, it can be written as $K^T y \sim N(0, \sigma^2 K^T H K)$ and the residual (REML) log-likelihood function is

$$l_R(\phi; K^T y) = -\frac{1}{2}\Big[(n-p)log2\pi + (n-p)log\sigma^2 + log|K^T H^{-1} K| + \frac{1}{\sigma^2} y^T K (K^T H^{-1} K)^{-1} K^T y\Big]$$

where, $\phi = (k^T, \sigma^2)^T, k = (\gamma^T, \rho^T)^T$. Patterson and Thompson (1971) derived the probability distribution of $K^T y$ by carefully choosing $K$ as an$[n-p]\times n$ matrix whose rows are $[n-p]$ linearly independent rows of $[I - X(X^TX)^{-1}X^T]$. Since $[I - X(X^TX)^{-1}X^T]$ is symmetric, idempotent and has rank $[n-p]$, it can be expressed as $KK^T$ such that $K^TK = I$. Patterson and Thompson [1971] argued that since $E(K^Ty = 0$, $K^Ty$ lies in the error space, and hence contains no information about the fixed effects, but it does contain information about the variance parameters. Then the REML log-likelihood function (ignoring constants) for the model is

$$\begin{aligned} l_R(\phi; y) &= -\frac{1}{2}\Big[(n-p)log\sigma^2 + log|H| + log|X^TH^{-1}X| + \frac{1}{\sigma^2}(y - X\hat{\beta})^T\hat{H}^{-1}(y - X\hat{\beta})\Big] \\ &= -\frac{1}{2}\Big[(n-p)log\sigma^2 + log|H| + log|X^TH^{-1}X| + \frac{y^TPy}{\sigma^2}\Big] \end{aligned}$$

where, $\hat{\beta}$, the GLS estimate of $\beta$, and $P$ are given in Lemma 1. Differentiating the REML log-likelihood function with respect to$\sigma^2$ and $k_j; j = 1, 2, \cdots, r+s$ and setting equal to zero and solving gives a REML estimator for the error variance as

$$\hat{\sigma}^2 = \frac{y^T\hat{P}y}{n-p}$$

which should be computed iteratively since it depends on $\hat{k}$ through $P$. The REML estimate for $k$must also be found iteratively [Gilmour et al. (1995)].

**Iterative Schemes**

To calculate the ML or REML estimates of the variance of the variance parameters, related iterative procedures have been discussed and a comparison has also been made. The iterative methods are:

- Newton-Raphson (NR)
- Fisher Scoring (FS)

- Average Information (AI) algorithm

The theory of inferential procedures used for the estimated parameters in the linear mixed model has also been discussed separating into different sections as follows

- Inference for fixed effects
- Inference for variance parameters
- Inference for random effects
- Inference on a combination of fixed and random effects

## A.3 Expression of $R^2$ in terms of z-Score

Let us consider a simple linear regression model as:

$$y = x\beta + \epsilon; \quad \text{where, } \epsilon \sim N(0, \sigma^2)$$

The ordinary least square (OLS) estimate of the parameter $\beta$ can be obtained as

$$\hat{\beta} = \frac{\sum xy}{\sum x^2} = R\frac{\sigma_y}{\sigma_x}$$

We know that,

$$R^2 = 1 - \frac{MSE}{var(y)}$$
$$\therefore MSE = \hat{\sigma}_e^2 = \sigma^2[1 - R^2]$$

We also know that,

$$v(\hat{\beta}) = \frac{MSE}{nvar(x)} = \frac{\hat{\sigma}_e^2}{n\sigma_x^2} = \frac{\sigma^2[1 - R^2]}{n\sigma_x^2}$$

The test statistic:

$$\begin{aligned} z^2 &= \frac{\hat{\beta}^2}{var(\hat{\beta})} \\ &= \frac{R^2\frac{\sigma_y^2}{\sigma_x^2}}{\frac{\sigma^2[1-R^2]}{n\sigma_x^2}} \\ &= \frac{nR^2}{1 - R^2} \\ \Rightarrow \frac{z^2}{n} &= \frac{R^2}{1 - R^2} \end{aligned}$$

$$\Rightarrow 1 + \frac{z^2}{n} = 1 + \frac{R^2}{1 - R^2}$$

$$\Rightarrow 1 - R^2 = \frac{1}{1 + \frac{z^2}{n}}$$

$$\therefore R^2 = \frac{z^2}{z^2 + n}$$

# References

Abu-Mostafa, Y. S. (1995). Hints. *Neural Computation*, 7(4):639–671. [TLDR] The systematic use of hints in the learning-from-examples paradigm, which is tantamount to combining rules and data in learning, is compatible with different learning models, optimization techniques, and regularization techniques.

Adhikari, K., Fontanil, T., Cal, S., Mendoza-Revilla, J., Fuentes-Guajardo, M., Chacón-Duque, J.-C., Al-Saadi, F., Johansson, J. A., Quinto-Sanchez, M., Acuña-Alonzo, V., Jaramillo, C., Arias, W., Barquera Lozano, R., Macín Pérez, G., Gómez-Valdés, J., Villamil-Ramírez, H., Hunemeier, T., Ramallo, V., Silva de Cerqueira, C. C., Hurtado, M., Villegas, V., Granja, V., Gallo, C., Poletti, G., Schuler-Faccini, L., Salzano, F. M., Bortolini, M.-C., Canizales-Quinteros, S., Rothhammer, F., Bedoya, G., Gonzalez-José, R., Headon, D., López-Otín, C., Tobin, D. J., Balding, D., and Ruiz-Linares, A. (2016a). A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. *Nature Communications*, 7(1):10815.

Adhikari, K., Fuentes-Guajardo, M., Quinto-Sánchez, M., Mendoza-Revilla, J., Camilo Chacón-Duque, J., Acuña-Alonzo, V., Jaramillo, C., Arias, W., Lozano, R. B., Pérez, G. M., Gómez-Valdés, J., Villamil-Ramírez, H., Hunemeier, T., Ramallo, V., Silva de Cerqueira, C. C., Hurtado, M., Villegas, V., Granja, V., Gallo, C., Poletti, G., Schuler-Faccini, L., Salzano, F. M., Bortolini, M.-C., Canizales-Quinteros, S., Cheeseman, M., Rosique, J., Bedoya, G., Rothhammer, F., Headon, D., González-José, R., Balding, D., and Ruiz-Linares, A. (2016b). A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation. *Nature Communications*, 7(1):11616.

Adhikari, K., Reales, G., Smith, A. J. P., Konka, E., Palmen, J., Quinto-Sanchez,

M., Acuña-Alonzo, V., Jaramillo, C., Arias, W., Fuentes, M., Pizarro, M., Barquera Lozano, R., Macín Pérez, G., Gómez-Valdés, J., Villamil-Ramírez, H., Hunemeier, T., Ramallo, V., Silva de Cerqueira, C. C., Hurtado, M., Villegas, V., Granja, V., Gallo, C., Poletti, G., Schuler-Faccini, L., Salzano, F. M., Bortolini, M.-C., Canizales-Quinteros, S., Rothhammer, F., Bedoya, G., Calderón, R., Rosique, J., Cheeseman, M., Bhutta, M. F., Humphries, S. E., Gonzalez-José, R., Headon, D., Balding, D., and Ruiz-Linares, A. (2015). A Genome-Wide Association Study Identifies Multiple Loci for Variation in Human Ear Morphology. *Nature Communications*, 6:7500.

Anderson, R. L. and Bancroft, T. A. (1952). *Statistical Theory in Research [by] R.L. Anderson [and] T.A. Bancroft.* McGraw-Hill, New York.

Bartlett, M. S. (1941). The Statistical Significance of Canonical Correlations. *Biometrika*, 32(1):29–37.

Bates, D. (2010). Computational Methods for Mixed Models.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Berndt, E. R. and Savin, N. E. (1977). Conflict among Criteria for Testing Hypotheses in the Multivariate Linear Regression Model. *Econometrica*, 45(5):1263–1277.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, 114(3):542–551.

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., Samani, N. J., Todd, J. A., Donnelly, P., Barrett, J. C., Burton, P. R., Davison, D., Donnelly, P., Easton,

D., Evans, D., Leung, H.-T., Marchini, J. L., Morris, A. P., Spencer, C. C. A., Tobin, M. D., Cardon, L. R., Clayton, D. G., Attwood, A. P., Boorman, J. P., Cant, B., Everson, U., Hussey, J. M., Jolley, J. D., Knight, A. S., Koch, K., Meech, E., Nutland, S., Prowse, C. V., Stevens, H. E., Taylor, N. C., Walters, G. R., Walker, N. M., Watkins, N. A., Winzer, T., Todd, J. A., Ouwehand, W. H., Jones, R. W., McArdle, W. L., Ring, S. M., Strachan, D. P., Pembrey, M., Breen, G., St Clair, D., Caesar, S., Gordon-Smith, K., Jones, L., Fraser, C., Green, E. K., Grozeva, D., Hamshere, M. L., Holmans, P. A., Jones, I. R., Kirov, G., Moskvina, V., Nikolov, I., O'Donovan, M. C., Owen, M. J., Craddock, N., Collier, D. A., Elkin, A., Farmer, A., Williamson, R., McGuffin, P., Young, A. H., Ferrier, I. N., Ball, S. G., Balmforth, A. J., Barrett, J. H., Bishop, D. T., Iles, M. M., Maqbool, A., Yuldasheva, N., Hall, A. S., Braund, P. S., Burton, P. R., Dixon, R. J., Mangino, M., Stevens, S., Tobin, M. D., Thompson, J. R., Samani, N. J., Bredin, F., Tremelling, M., Parkes, M., Drummond, H., Lees, C. W., Nimmo, E. R., Satsangi, J., Fisher, S. A., Forbes, A., Lewis, C. M., Onnie, C. M., Prescott, N. J., Sanderson, J., Mathew, C. G., Barbour, J., Mohiuddin, M. K., Todhunter, C. E., Mansfield, J. C., Ahmad, T., Cummings, F. R., Jewell, D. P., Webster, J., Brown, M. J., Clayton, D. G., Lathrop, G. M., Connell, J., Dominiczak, A., Samani, N. J., Marcano, C. A. B., Burke, B., Dobson, R., Gungadoo, J., Lee, K. L., Munroe, P. B., Newhouse, S. J., Onipinla, A., Wallace, C., Xue, M., Caulfield, M., Farrall, M., Barton, A., and Genomics (BRAGGS), T. B. i. R. G., Bruce, I. N., Donovan, H., Eyre, S., Gilbert, P. D., Hider, S. L., Hinks, A. M., John, S. L., Potter, C., Silman, A. J., Symmons, D. P. M., Thomson, W., Worthington, J., Clayton, D. G., Dunger, D. B., Nutland, S., Stevens, H. E., Walker, N. M., Widmer, B., Todd, J. A., Frayling, T. M., Freathy, R. M., Lango, H., Perry, J. R. B., Shields, B. M., Weedon, M. N., Hattersley, A. T., Hitman, G. A., Walker, M., Elliott, K. S., Groves, C. J., Lindgren, C. M., Rayner, N. W., Timpson, N. J., Zeggini, E., McCarthy, M. I.,

Newport, M., Sirugo, G., Lyons, E., Vannberg, F., Hill, A. V. S., Bradbury, L. A., Farrar, C., Pointon, J. J., Wordsworth, P., Brown, M. A., Franklyn, J. A., Heward, J. M., Simmonds, M. J., Gough, S. C. L., Seal, S., Susceptibility Collaboration (UK), B. C., Stratton, M. R., Rahman, N., Ban, M., Goris, A., Sawcer, S. J., Compston, A., Conway, D., Jallow, M., Newport, M., Sirugo, G., Rockett, K. A., Kwiatkowski, D. P., Bumpstead, S. J., Chaney, A., Downes, K., Ghori, M. J. R., Gwilliam, R., Hunt, S. E., Inouye, M., Keniry, A., King, E., McGinnis, R., Potter, S., Ravindrarajah, R., Whittaker, P., Widden, C., Withers, D., Deloukas, P., Leung, H.-T., Nutland, S., Stevens, H. E., Walker, N. M., Todd, J. A., Easton, D., Clayton, D. G., Burton, P. R., Tobin, M. D., Barrett, J. C., Evans, D., Morris, A. P., Cardon, L. R., Cardin, N. J., Davison, D., Ferreira, T., Pereira-Gale, J., Hallgrimsdóttir, I. B., Howie, B. N., Marchini, J. L., Spencer, C. C. A., Su, Z., Teo, Y. Y., Vukcevic, D., Donnelly, P., Bentley, D., Brown, M. A., Cardon, L. R., Caulfield, M., Clayton, D. G., Compston, A., Craddock, N., Deloukas, P., Donnelly, P., Farrall, M., Gough, S. C. L., Hall, A. S., Hattersley, A. T., Hill, A. V. S., Kwiatkowski, D. P., Mathew, C. G., McCarthy, M. I., Ouwehand, W. H., Parkes, M., Pembrey, M., Rahman, N., Samani, N. J., Stratton, M. R., Todd, J. A., Worthington, J., The Wellcome Trust Case Control Consortium, Management Committee, Data and Analysis Committee, UK Blood Services and University of Cambridge Controls, 1958 Birth Cohort Controls, Bipolar Disorder, Coronary Artery Disease, Crohn's Disease, Hypertension, Rheumatoid Arthritis, Type 1 Diabetes, Type 2 Diabetes, Tuberculosis, Ankylosing Spondylitis, Autoimmune Thyroid Disease, Breast Cancer, Multiple Sclerosis, Gambian Controls, DNA, Data QC and Informatics, G., Statistics, and Primary Investigators (2007). Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls. *Nature*, 447(7145):661–678.

Buse, A. (1982). The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An

Expository Note. *The American Statistician*, 36(3a):153–157.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., and Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209.

Caliński, T., Krzysko, M., and Wolynski, W. (2006). A Comparison of Some Tests for Determining the Number of Nonzero Canonical Correlations.

Carayon, D., Adhikari, K., Monsarrat, P., Dumoncel, J., Braga, J., Duployer, B., Delgado, M., Fuentes-Guajardo, M., De Beer, F., Hoffman, J. W., Oettlé, A. C., Donat, R., Pan, L., Ruiz-Linares, A., Tenailleau, C., Vaysse, F., Esclassan, R., and Zanolli, C. (2019). A geometric morphometric approach to the study of variation of shovel-shaped incisors. *American Journal of Physical Anthropology*, 168(1):229–241.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Thomson Learning.

Chandra, T. K. and Joshi, S. N. (1983). Comparison of the Likelihood Ratio, Rao's and Wald's Tests and a Conjecture of C. R. Rao. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 45(2):226–246.

Chatterjee, S. (2012). *Regression Analysis by Example*. Wiley.

Courville, T. and Thompson, B. (2001). Use of Structure Coefficients in Published Multiple Regression Articles: $\beta$ is not Enough. *Educational and Psychological Measurement*, 61(2):229–248.

Dadousis, C., Veerkamp, R. F., Heringstad, B., Pszczola, M., and Calus, M. P. (2014). A comparison of principal component regression and genomic REML for genomic prediction across populations. *Genetics Selection Evolution*, 46(1):60.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, 69(3):161–182.

Delgado, M., Imbrasas, M., Davies, T., Khokan, M. R., Adhikari, K., Skinner, M., and Zanolli, C. (2021). *Pattern and Magnitude of Taxonomic Classification Accuracy of Living and Fossil Hominoid Upper Molars through Landmark-Based and Surface-Based Approaches*.

Edelman, D. A Note on Uniformly Most Powerful Two-Sided Tests. page 3.

Efron, B. (2004). The Estimation of Prediction Error. *Journal of the American Statistical Association*, 99(467):619–632.

Eldén, L. (2015). Computing Frechet derivatives in partial least squares regression. *Linear Algebra and its Applications*, 473:316–338.

Engle, R. F. (1984). Chapter 13 Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. In *Handbook of Econometrics*, volume 2, pages 775–826. Elsevier.

Frank, I. E. and Friedman, J. H. (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, 35(2):109–135.

Garthwaite, P. H. (1994). An interpretation of partial least squares. *Journal of the American Statistical Association*, 89(425):122–127.

Garthwaite, P. H., Critchley, F., Anaya-Izquierdo, K., and Mubwandarikwa, E. (2012). Orthogonalization of vectors with minimal adjustment. *Biometrika*, 99(4):787–798.

Garthwaite, P. H. and Koch, I. (2016). Evaluating the Contributions of Individual Variables to a Quadratic Form. *Australian & New Zealand Journal of Statistics*, 58(1):99–119.

Gibson, W. A. (1962). Orthogonal predictors: A possible resolution of the Hoffman-Ward controversy. *Psychological reports*, 11(1):32–34.

Gilmour, A. R., Thompson, R., and Cullis, B. R. (1995). Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics*, 51(4):1440.

Gumedze, F. N. and Dunne, T. T. (2011). Parameter estimation and inference in the linear mixed model. *Linear Algebra and its Applications*, 435(8):1920–1944.

Hartley, H. O. and Rao, J. N. K. (1967). Maximum-Likelihood Estimation for the Mixed Analysis of Variance Model. *Biometrika*, 54(1/2):93–108.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY.

Hayes, B. J., Visscher, P. M., and Goddard, M. E. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research*, 91(1):47–60.

Helland, I. S. (1988). On the structure of partial least squares regression. *Communications in statistics-Simulation and Computation*, 17(2):581–607.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, pages 423–447.

Henderson, C. R., Kempthorne, O., Searle, S. R., and von Krosigk, C. M. (1959). The Estimation of Environmental and Genetic Trends from Records Subject to Culling. *Biometrics*, 15(2):192.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67.

Hoffman, G. E. (2013). Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions. *PLoS ONE*, 8(10):e75707.

Hoggart, C. J., Whittaker, J. C., Iorio, M. D., and Balding, D. J. (2008). Simultaneous Analysis of All SNPs in Genome-Wide and Re-Sequencing Association Studies. *PLOS Genetics*, 4(7):e1000130.

James, W. and Stein, C. (1992). Estimation with quadratic loss. In *Breakthroughs in Statistics*, pages 443–460. Springer.

Johnson, J. W. (2000). A Heuristic Method for Estimating the Relative Weight of Predictor Variables in Multiple Regression. *Multivariate Behavioral Research*, 35(1):1–19.

Johnson, J. W. and Lebreton, J. M. (2016). History and Use of Relative Importance Indices in Organizational Research:. *Organizational Research Methods*.

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics*, 178(3):1709.

Kenny, E. E., Kim, M., Gusev, A., Lowe, J. K., Salit, J., Smith, J. G., Kovvali, S., Kang, H. M., Newton-Cheh, C., Daly, M. J., Stoffel, M., Altshuler, D. M., Friedman, J. M., Eskin, E., Breslow, J. L., and Pe'er, I. (2011). Increased power of mixed models

facilitates association mapping of 10 loci for metabolic traits in an isolated population. *Human Molecular Genetics*, 20(4):827.

Klingenberg, C. P. (2016). Size, shape, and form: Concepts of allometry in geometric morphometrics. *Development Genes and Evolution*, 226(3):113–137.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance-testing system. *Psychological Bulletin*, 85(2):410–416.

Kurki, M. I., Karjalainen, J., Palta, P., Sipilä, T. P., Kristiansson, K., Donner, K. M., Reeve, M. P., Laivuori, H., Aavikko, M., Kaunisto, M. A., Loukola, A., Lahtela, E., Mattsson, H., Laiho, P., Della Briotta Parolo, P., Lehisto, A. A., Kanai, M., Mars, N., Rämö, J., Kiiskinen, T., Heyne, H. O., Veerapen, K., Rüeger, S., Lemmelä, S., Zhou, W., Ruotsalainen, S., Pärn, K., Hiekkalinna, T., Koskelainen, S., Paajanen, T., Llorens, V., Gracia-Tabuenca, J., Siirtola, H., Reis, K., Elnahas, A. G., Sun, B., Foley, C. N., Aalto-Setälä, K., Alasoo, K., Arvas, M., Auro, K., Biswas, S., Bizaki-Vallaskangas, A., Carpen, O., Chen, C.-Y., Dada, O. A., Ding, Z., Ehm, M. G., Eklund, K., Färkkilä, M., Finucane, H., Ganna, A., Ghazal, A., Graham, R. R., Green, E. M., Hakanen, A., Hautalahti, M., Hedman, Å. K., Hiltunen, M., Hinttala, R., Hovatta, I., Hu, X., Huertas-Vazquez, A., Huilaja, L., Hunkapiller, J., Jacob, H., Jensen, J.-N., Joensuu, H., John, S., Julkunen, V., Jung, M., Junttila, J., Kaarniranta, K., Kähönen, M., Kajanne, R., Kallio, L., Kälviäinen, R., Kaprio, J., Kerimov, N., Kettunen, J., Kilpeläinen, E., Kilpi, T., Klinger, K., Kosma, V.-M., Kuopio, T., Kurra, V., Laisk, T., Laukkanen, J., Lawless, N., Liu, A., Longerich, S., Mägi, R., Mäkelä, J., Mäkitie, A., Malarstig, A., Mannermaa, A., Maranville, J., Matakidou, A., Meretoja, T., Mozaffari, S. V., Niemi, M. E. K., Niemi, M., Niiranen, T., O´Donnell, C. J., Obeidat, M., Okafo, G., Ollila, H. M., Palomäki, A., Palotie, T., Partanen, J., Paul, D. S., Pelkonen, M., Pendergrass, R. K., Petrovski, S., Pitkäranta, A., Platt,

A., Pulford, D., Punkka, E., Pussinen, P., Raghavan, N., Rahimov, F., Rajpal, D., Renaud, N. A., Riley-Gillis, B., Rodosthenous, R., Saarentaus, E., Salminen, A., Salminen, E., Salomaa, V., Schleutker, J., Serpi, R., Shen, H.-y., Siegel, R., Silander, K., Siltanen, S., Soini, S., Soininen, H., Sul, J. H., Tachmazidou, I., Tasanen, K., Tienari, P., Toppila-Salmi, S., Tukiainen, T., Tuomi, T., Turunen, J. A., Ulirsch, J. C., Vaura, F., Virolainen, P., Waring, J., Waterworth, D., Yang, R., Nelis, M., Reigo, A., Metspalu, A., Milani, L., Esko, T., Fox, C., Havulinna, A. S., Perola, M., Ripatti, S., Jalanko, A., Laitinen, T., Mäkelä, T. P., Plenge, R., McCarthy, M., Runz, H., Daly, M. J., and Palotie, A. (2023). FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*, 613(7944):508–518.

Kutner, M. H., editor (2005). *Applied Linear Statistical Models*. The McGraw-Hill/Irwin Series Operations and Decision Sciences. McGraw-Hill Irwin, Boston, 5th ed edition.

Li, J., Das, K., Fu, G., Li, R., and Wu, R. (2011). The Bayesian lasso for genome-wide association studies. *Bioinformatics*, 27(4):516–523.

Liakhovitski, D., Bryukhov, Y., and Conklin, M. (2010). Relative importance of predictors: Comparison of Random Forests with Johnson's Relative Weights. *Model Assisted Statistics and Applications*, 5(4):235–249.

Lin, C. and McAllister, A. (1984). Monte Carlo Comparison of Four Methods for Estimation of Genetic Parameters in the Univariate Case. *Journal of Dairy Science*, 67(10):2389–2398.

Lipovetsky, S. and Conklin, W. M. (2015). Predictor relative importance and matching regression parameters. *Journal of Applied Statistics*, 42(5):1017–1031.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman,

D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835.

Liu, F., van der Lijn, F., Schurmann, C., Zhu, G., Chakravarty, M. M., Hysi, P. G., Wollstein, A., Lao, O., de Bruijne, M., Ikram, M. A., van der Lugt, A., Rivadeneira, F., Uitterlinden, A. G., Hofman, A., Niessen, W. J., Homuth, G., de Zubicaray, G., McMahon, K. L., Thompson, P. M., Daboul, A., Puls, R., Hegenscheid, K., Bevan, L., Pausova, Z., Medland, S. E., Montgomery, G. W., Wright, M. J., Wicking, C., Boehringer, S., Spector, T. D., Paus, T., Martin, N. G., Biffar, R., and Kayser, M. (2012). A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in Europeans. *PLOS Genetics*, 8(9):e1002932.

Lloyd-Jones, L. R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K. E., Wang, H., Zheng, Z., Magi, R., Esko, T., Metspalu, A., Wray, N. R., Goddard, M. E., Yang, J., and Visscher, P. M. (2019). Improved Polygenic Prediction by Bayesian Multiple Regression on Summary Statistics. *bioRxiv*, page 522961.

Lu, T.-T. and Shiou, S.-H. (2002). Inverses of 2 $\times$ 2 block matrices. *Computers & Mathematics with Applications*, 43(1-2):119–129.

Marchini, J., Cardon, L. R., Phillips, M. S., and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature Genetics*, 36(5):512–517.

McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127.

McVean, G. (2009). A Genealogical Interpretation of Principal Components Analysis. *PLOS Genetics*, 5(10):e1000686.

Menozzi, P., Piazza, A., and Cavalli-Sforza, L. (1978). Synthetic maps of human gene frequencies in Europeans. *Science*, 201(4358):786–792.

Mez, J., Chung, J., Jun, G., Kriegel, J., Bourlas, A. P., Sherva, R., Logue, M. W., Barnes, L. L., Bennett, D. A., Buxbaum, J. D., Byrd, G. S., Crane, P. K., Ertekin-Taner, N., Evans, D., Fallin, M. D., Foroud, T., Goate, A., Graff-Radford, N. R., Hall, K. S., Kamboh, M. I., Kukull, W. A., Larson, E. B., Manly, J. J., Alzheimer's Disease Genetics Consortium, Haines, J. L., Mayeux, R., Pericak-Vance, M. A., Schellenberg, G. D., Lunetta, K. L., and Farrer, L. A. (2017). Two novel loci, *COBL* and *SLC10A2* , for Alzheimer's disease in African Americans. *Alzheimer's & Dementia*, 13(2):119–129.

Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T., Murakami, Y., Yuji, K., Furukawa, Y., Zembutsu, H., Tanaka, T., Ohnishi, Y., Nakamura, Y., Shiono, M., Misumi, K., Kaieda, R., Harada, H., Minami, S., Emi, M., Emoto, N., Daida, H., Miyauchi, K., Murakami, A., Asai, S., Moriyama, M., Takahashi, Y., Fujioka, T., Obara, W., Mori, S., Ito, H., Nagayama, S., Miki, Y., Masumoto, A., Yamada, A., Nishizawa, Y., Kodama, K., Kutsumi, H., Sugimoto, Y., Koretsune, Y., Kusuoka, H., Yanai, H., and Kubo, M. (2017). Overview of the BioBank Japan Project: Study design and profile. *Journal of Epidemiology*, 27(3, Supplement):S2–S8.

Newman, D. L., Abney, M., McPeek, M. S., Ober, C., and Cox, N. J. (2001). The Importance of Genealogy in Determining Genetic Associations with Complex Traits. *American Journal of Human Genetics*, 69(5):1146–1148.

Null, M., Yilmaz, F., Astling, D., Yu, H.-C., Cole, J. B., Hallgrímsson, B., Santorico, S. A., Spritz, R. A., Shaikh, T. H., and Hendricks, A. E. (2022). Genome-wide analysis of copy number variants and normal facial variation in a large cohort of Bantu Africans. *Human Genetics and Genomics Advances*, 3(1):100082.

Olson, C. L. (1976). On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, 83(4):579–586.

Oxnard, C. E. (1974). Cranial variation in man: A study by multivariate analysis of patterns of difference among recent human populations. By W. W. Howells. ix + 259 pp., figuress, tables, bibliography. Vol. 67, Paper of the Peabody Museum of Archaeology and Enthnology, Harvard University, Cambridge, Mass. 1973. $10.00 (paper). *American Journal of Physical Anthropology*, 41(2):349–351.

Paria, S. S., Rahman, S. R., and Adhikari, K. (2022). Fastman: A fast algorithm for visualizing GWAS results using Manhattan and Q-Q plots. [TLDR] A new R package, fastman, is developed for fast and efficient visualization of GWAS results and other genomewide scores using Manhattan and Q-Q plots, and is equipped to handle big datasets with fast plot generation.

Patterson, H. D. and Thompson, R. (1971). Recovery of Inter-Block Information when Block Sizes are Unequal. *Biometrika*, 58(3):545–554.

Patterson, N., Price, A. L., and Reich, D. (2006). Population Structure and Eigenanalysis. *PLOS Genetics*, 2(12):e190.

Pirinen, M., Donnelly, P., and Spencer, C. C. A. (2012). Including known covariates can reduce power to detect genetic effects in case-control studies. *Nature Genetics*, 44(8):848–851.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909.

Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches to pop-

ulation stratification in genome-wide association studies. *Nature Reviews. Genetics*, 11(7):459–463.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3):559–575.

Quaglio, M., Fraga, E. S., and Galvanin, F. (2020). A diagnostic procedure for improving the structure of approximated kinetic models. *Computers & Chemical Engineering*, 133.

Rakitsch, B., Lippert, C., Stegle, O., and Borgwardt, K. (2013). A Lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics*, 29(2):206–214.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 44(1):50–57.

Rao, C. R. (1952). *Advanced Statistical Methods in Biometric Research*. Advanced Statistical Methods in Biometric Research. Wiley, Oxford, England.

Rao, C. R. (2002). *Linear Statistical Inference and Its Applications*. Wiley Series in Probability and Statistics. Wiley, New York, 2. ed., paperback ed edition.

Reich, D., Price, A. L., and Patterson, N. (2008). Principal component analysis of genetic data. *Nature Genetics*, 40(5):491–492.

Robinson, G. K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science*, 6(1):15–32.

Rosipal, R. and Krämer, N. (2006). Overview and recent advances in partial least squares.

Ruiz-Linares, A., Adhikari, K., Acuña-Alonzo, V., Quinto-Sanchez, M., Jaramillo, C., Arias, W., Fuentes, M., Pizarro, M., Everardo, P., de Avila, F., Gómez-Valdés, J., León-Mimila, P., Hunemeier, T., Ramallo, V., de Cerqueira, C. C. S., Burley, M.-W., Konca, E., de Oliveira, M. Z., Veronez, M. R., Rubio-Codina, M., Attanasio, O., Gibbon, S., Ray, N., Gallo, C., Poletti, G., Rosique, J., Schuler-Faccini, L., Salzano, F. M., Bortolini, M.-C., Canizales-Quinteros, S., Rothhammer, F., Bedoya, G., Balding, D., and Gonzalez-José, R. (2014). Admixture in Latin America: Geographic Structure, Phenotypic Diversity and Self-Perception of Ancestry Based on 7,342 Individuals. *PLOS Genetics*, 10(9):e1004572.

Sankararaman, S., Mallick, S., Patterson, N., and Reich, D. (2016). The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Current biology: CB*, 26(9):1241–1247.

Scott, G. R., Turner Ii, C. G., Townsend, G. C., and Martinón-Torres, M. (1997). *The Anthropology of Modern Human Teeth: Dental Morphology and Its Variation in Recent and Fossil* Homo Sapien. Cambridge University Press, 2 edition.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). Variance Components. page 537.

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., and Nordborg, M. (2012). An efficient multi-locus mixed model approach for genome-wide association studies in structured populations. *Nature genetics*, 44(7):825–830.

Shabuz, Z. R. and Garthwaite, P. H. (2019). Contribution of individual variables to the regression sum of squares. *Model Assisted Statistics and Applications*, 14(4):281–296.

Shen, X., Alam, M., Fikse, F., and Rönnegård, L. (2013). A Novel Generalized Ridge Regression Method for Quantitative Genetics. *Genetics*, 193(4):1255–1268.

Silvey, S. D. (1959). The Lagrangian Multiplier Test. *The Annals of Mathematical Statistics*, 30(2):389–407.

Speed, D., Hemani, G., Johnson, M. R., and Balding, D. J. (2012). Improved Heritability Estimation from Genome-wide SNPs. *American Journal of Human Genetics*, 91(6):1011–1021.

Stone, M. and Brooks, R. J. (1990). Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(2):237–258.

Swallow, W. H. and Monahan, J. F. (1984). Monte Carlo Comparison of ANOVA, MIVQUE, REML, and ML Estimators of Variance Components. *Technometrics*, 26(1):47–57.

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S.-b., Tian, X., Browning, B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., Shetty, A. C., Blackwell, T. W., Smith, A. V., Wong, Q., Liu, X., Conomos, M. P., Bobo, D. M., Aguet, F., Albert, C., Alonso, A., Ardlie, K. G., Arking, D. E., Aslibekyan, S., Auer, P. L., Barnard, J., Barr, R. G., Barwick, L., Becker, L. C., Beer, R. L., Benjamin, E. J., Bielak, L. F., Blangero, J., Boehnke, M., Bowden, D. W., Brody, J. A., Burchard, E. G., Cade, B. E., Casella, J. F., Chalazan, B., Chasman, D. I., Chen, Y.-D. I., Cho, M. H., Choi, S. H., Chung, M. K., Clish, C. B., Correa, A., Curran, J. E., Custer, B., Darbar, D., Daya, M., de Andrade, M., DeMeo, D. L., Dutcher, S. K., Ellinor, P. T., Emery, L. S., Eng, C., Fatkin, D., Fingerlin, T.,

Forer, L., Fornage, M., Franceschini, N., Fuchsberger, C., Fullerton, S. M., Germer, S., Gladwin, M. T., Gottlieb, D. J., Guo, X., Hall, M. E., He, J., Heard-Costa, N. L., Heckbert, S. R., Irvin, M. R., Johnsen, J. M., Johnson, A. D., Kaplan, R., Kardia, S. L. R., Kelly, T., Kelly, S., Kenny, E. E., Kiel, D. P., Klemmer, R., Konkle, B. A., Kooperberg, C., Köttgen, A., Lange, L. A., Lasky-Su, J., Levy, D., Lin, X., Lin, K.-H., Liu, C., Loos, R. J. F., Garman, L., Gerszten, R., Lubitz, S. A., Lunetta, K. L., Mak, A. C. Y., Manichaikul, A., Manning, A. K., Mathias, R. A., McManus, D. D., McGarvey, S. T., Meigs, J. B., Meyers, D. A., Mikulla, J. L., Minear, M. A., Mitchell, B. D., Mohanty, S., Montasser, M. E., Montgomery, C., Morrison, A. C., Murabito, J. M., Natale, A., Natarajan, P., Nelson, S. C., North, K. E., O'Connell, J. R., Palmer, N. D., Pankratz, N., Peloso, G. M., Peyser, P. A., Pleiness, J., Post, W. S., Psaty, B. M., Rao, D. C., Redline, S., Reiner, A. P., Roden, D., Rotter, J. I., Ruczinski, I., Sarnowski, C., Schoenherr, S., Schwartz, D. A., Seo, J.-S., Seshadri, S., Sheehan, V. A., Sheu, W. H., Shoemaker, M. B., Smith, N. L., Smith, J. A., Sotoodehnia, N., Stilp, A. M., Tang, W., Taylor, K. D., Telen, M., Thornton, T. A., Tracy, R. P., Van Den Berg, D. J., Vasan, R. S., Viaud-Martinez, K. A., Vrieze, S., Weeks, D. E., Weir, B. S., Weiss, S. T., Weng, L.-C., Willer, C. J., Zhang, Y., Zhao, X., Arnett, D. K., Ashley-Koch, A. E., Barnes, K. C., Boerwinkle, E., Gabriel, S., Gibbs, R., Rice, K. M., Rich, S. S., Silverman, E. K., Qasba, P., Gan, W., Papanicolaou, G. J., Nickerson, D. A., Browning, S. R., Zody, M. C., Zöllner, S., Wilson, J. G., Cupples, L. A., Laurie, C. C., Jaquish, C. E., Hernandez, R. D., O'Connor, T. D., and Abecasis, G. R. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, 590(7845):290–299.

Thomas, D. R., Hughes, E., and Zumbo, B. D. (1998). On Variable Importance in Linear Regression. *Social Indicators Research*, 45(1):253–275.

Thompson, S. G. and Willeit, P. (2015). UK Biobank Comes of Age. *Lancet (London,*

*England)*, 386(9993):509–510.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Vandaele, W. (1981). Wald, likelihood ratio, and Lagrange multiplier tests as an F test. *Economics Letters*, 8(4):361–365.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of dairy science*, 91(11):4414–4423.

Wakeling, I. N. and Morris, J. J. (1993). A test of significance for partial least squares regression. *Journal of Chemometrics*, 7(4):291–304.

Wald, A. (1943). Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. *Transactions of the American Mathematical Society*, 54(3):426–482.

Wang, Q. S. and Huang, H. (2022). Methods for statistical fine-mapping and their applications to auto-immune diseases. *Seminars in Immunopathology*, 44(1):101–113.

Weaver, T. D. and Stringer, C. B. (2015). Unconstrained cranial evolution in Neandertals and modern humans compared to common chimpanzees. *Proceedings of the Royal Society B: Biological Sciences*, 282(1817):20151519.

Webster, M. and Sheets, H. D. (2010). A Practical Introduction to Landmark-Based Geometric Morphometrics. *The Paleontological Society Papers*, 16:163–188.

Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *Annals Math. Statist.*, 9(1):60–62.

William G. Cochran (1977). *Sampling Techniques (3th Edition) William G. Cochran.*

Wu, C., DeWan, A., Hoh, J., and Wang, Z. (2011). A Comparison of Association Methods Correcting for Population Stratification in Case–Control Studies. *Annals of human genetics*, 75(3):418.

Xiong, Z., Dankova, G., Howe, L. J., Lee, M. K., Hysi, P. G., de Jong, M. A., Zhu, G., Adhikari, K., Li, D., Li, Y., Pan, B., Feingold, E., Marazita, M. L., Shaffer, J. R., McAloney, K., Xu, S.-H., Jin, L., Wang, S., de Vrij, F. M., Lendemeijer, B., Richmond, S., Zhurov, A., Lewis, S., Sharp, G. C., Paternoster, L., Thompson, H., Gonzalez-Jose, R., Bortolini, M. C., Canizales-Quinteros, S., Gallo, C., Poletti, G., Bedoya, G., Rothhammer, F., Uitterlinden, A. G., Ikram, M. A., Wolvius, E., Kushner, S. A., Nijsten, T. E., Palstra, R.-J. T., Boehringer, S., Medland, S. E., Tang, K., Ruiz-Linares, A., Martin, N. G., Spector, T. D., Stergiakouli, E., Weinberg, S. M., Liu, F., and Kayser, M. (2019). Novel genetic loci affecting facial shape variation in humans. *eLife*, 8:e49898.

Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Madden, P. A. F., Heath, A. C., Martin, N. G., Montgomery, G. W., Weedon, M. N., Loos, R. J., Frayling, T. M., McCarthy, M. I., Hirschhorn, J. N., Goddard, M. E., and Visscher, P. M. (2012). Conditional and Joint Multiple-SNP Analysis of GWAS Summary Statistics Identifies Additional Variants Influencing Complex Traits. *Nature Genetics*, 44(4):369–375, S1–3.

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics*, 88(1):76–82.

Young, G. A. and Smith, R. L. (2005). *Essentials of Statistical Inference*. Cambridge

Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., and Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208.

Zaitlen, N., Pasaniuc, B., Patterson, N., Pollack, S., Voight, B., Groop, L., Altshuler, D., Henderson, B. E., Kolonel, L. N., Le Marchand, L., Waters, K., Haiman, C. A., Stranger, B. E., Dermitzakis, E. T., Kraft, P., and Price, A. L. (2012). Analysis of case-control association studies with known risk variants. *Bioinformatics (Oxford, England)*, 28(13):1729–1737.

Zhang, M., Wu, S., Du, S., Qian, W., Chen, J., Qiao, L., Yang, Y., Tan, J., Yuan, Z., Peng, Q., Liu, Y., Navarro, N., Tang, K., Ruiz-Linares, A., Wang, J., Claes, P., Jin, L., Li, J., and Wang, S. (2022). Genetic variants underlying differences in facial morphology in East Asian and European populations. *Nature Genetics*, 54(4):403–411.

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., and Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 42(4):355–360.

Zhou, H., Wang, F., and Tao, P. (2018). T-Distributed Stochastic Neighbor Embedding (t-SNE) Method with the Least Information Loss for Macromolecular Simulations. *Journal of chemical theory and computation*, 14(11):5499–5510.

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net.

*Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320.