STAT4116: Project Proposal

Jiazhe Lin (u7487349)

1 Introduction

The selective breeding of crops is an ancient practice that dates back to the emergence of human civilisation. Following the industrial revolution and the rising demand for agronomic production, modern plant breeding began to leverage statistical tools to analyse cultivars for desired traits. The abundance of data in agricultural statistics in turn gave rise to numerous fundamental statistical theories, such as the analysis of variance (ANOVA) (reference).

A key aspect of the plant breeding cycle involves comparing varieties across different trials, which makes it possible to assess plant performance under varying soil, water, and other environmental conditions (reference). The goal of this aggregated analysis is, first, to distinguish true genotypic performance from environmental influences, and second, to evaluate the stability of genotype performance across conditions (reference).

With the objective of multi-environment trials relatively straightforward, the key challenge lies in modeling the interaction between genotype and environment effects, a phenomenon referred to as the G $\times$ E effect (reference). One modern approach is to model trait responses using a linear mixed model, incorporating fixed environmental effects and random genotype, design, and interaction effects, along with auto-correlated residual errors to account for spatial correlation (reference).

In this assignment, we aim to provide a comprehensive comparison of estimation accuracy between the frequentist linear mixed model and its Bayesian counterpart, formulated as a posterior model with appropriate prior assumptions. We will begin by examining a simple case that uses phenotype data only, with genotype, environment and their interaction effect. From there, we will gradually explore the inclusion of experimental design variables such as row and column effects, auto-correlated residual errors, and, if time permits, genetic and marker information within the Bayesian data analytic framework introduced in the course.

2 Proposed Research Question

This project sets out to investigate several key research questions. First, we ask how a Bayesian framework can be applied to analyse the dataset and how its performance compares with the traditional frequentist approach. Building on this, we explore the incorporation of different variables into the Bayesian model, examining how alternative prior assumptions influence inference and prediction. Finally, if there's time we extend the analysis to consider the inclusion of gene marker data and spatial correlation structures, evaluating how these additional sources of information can be effectively modeled within a Bayesian framework to improve accuracy and interpretability.

3 Dataset Description

```
## [1]   323 17317
```

```
##       Me                 Mum                 Dad                 fgen
##  Length:9333        Length:9333        Length:9333        Min.   : 0.000
##  Class :character   Class :character   Class :character   1st Qu.: 1.000
##  Mode  :character   Mode  :character   Mode  :character   Median : 3.000
##                                                           Mean   : 2.851
##                                                           3rd Qu.: 4.000
##                                                           Max.   :10.000

##       plot           col            row            gen            env
##  1      :  36   1      : 792   1      : 432   G0008  : 126   E01    : 288
##  2      :  36   10     : 792   10     : 432   G0010  : 126   E02    : 288
##  3      :  36   11     : 792   11     : 432   G0324  : 126   E04    : 288
##  4      :  36   12     : 792   12     : 432   G0013  :  72   E05    : 288
##  5      :  36   2      : 792   13     : 432   G0009  :  63   E07    : 288
##  6      :  36   3      : 792   14     : 432   G0002  :  55   E08    : 288
##  (Other):9288   (Other):4752   (Other):6912   (Other):8936   (Other):7776
##      yield
##  Min.   :0.3134
##  1st Qu.:2.3147
##  Median :3.1950
##  Mean   :3.3570
##  3rd Qu.:4.3300
##  Max.   :7.6599
##  NA's   :31

## # A tibble: 36 x 2
##    env   n_plots
##    <fct>   <int>
##  1 E01       288
##  2 E02       288
##  3 E03       192
##  4 E04       288
##  5 E05       288
##  6 E06       192
##  7 E07       288
##  8 E08       288
##  9 E09       192
## 10 E10       288
## # i 26 more rows
```

- What data set are you using?
- What variables in the data set will you use?

4 Preliminary Analysis

- What are your initial thoughts on appropriate models/distributions?
- What questions and/or concerns do you have about the project?
- What metrics do we use for comparison?

5 Reference