# Project: Regression Model

Genome-wide association studies (GWAS)

Baldini Alessio
Mele Antonella

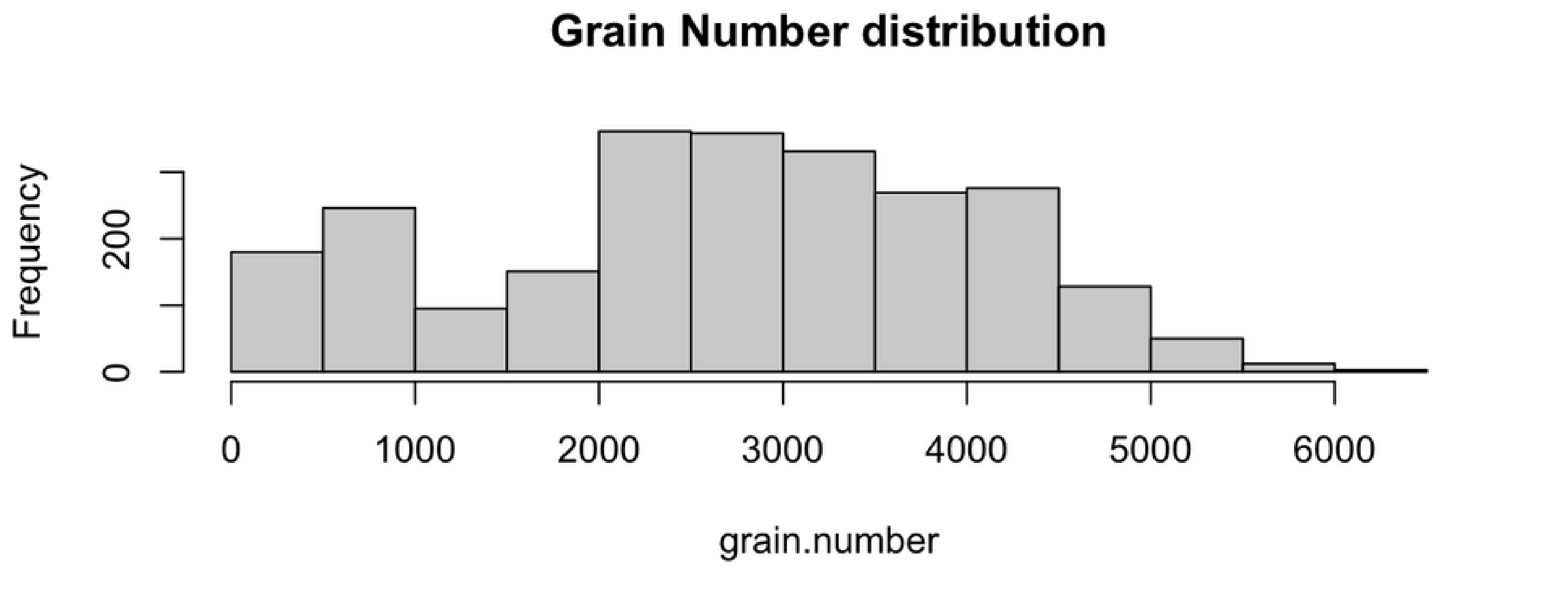
Date:
16 December 2024

# Index

**The aim: u**nderstand the possible correlation between the grain number and genetic and/or environmental factors.

# Our starting point

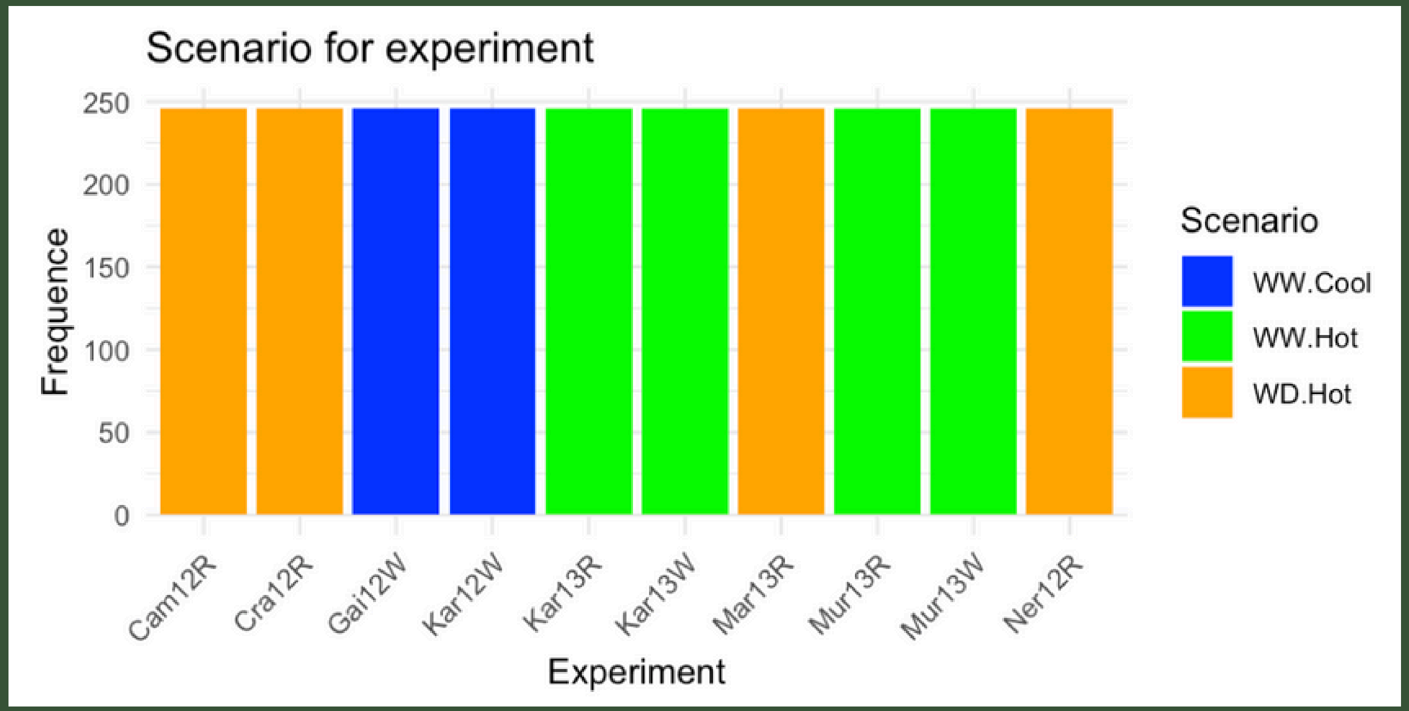
- The histogram shows the distribution of grain numbers, with the highest frequencies concentrated between 2000 and 4000. The data appears roughly symmetric, with fewer occurrences above 5000, suggesting potential outliers.



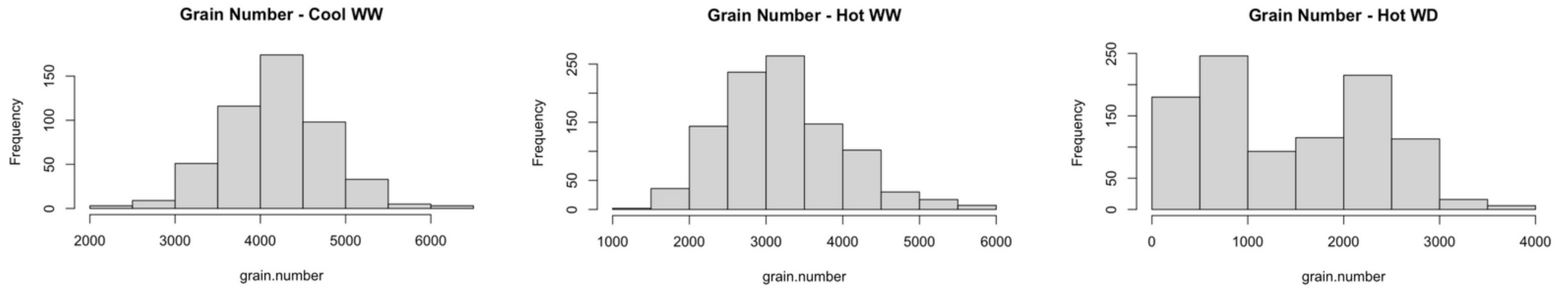

- The sample is made of 2460 observations, for which we know phenotypes and environmental situations in which the experiment took place (in pheno) and the SPN with the genotypes with respect to the reference allele (in df1).

- Then, for each SPN we have some characteristics with respect to their position and genetic variability (in geno_map) and for each genotype the allelic status (homozigote or heterozigote) for each SNP (geno).

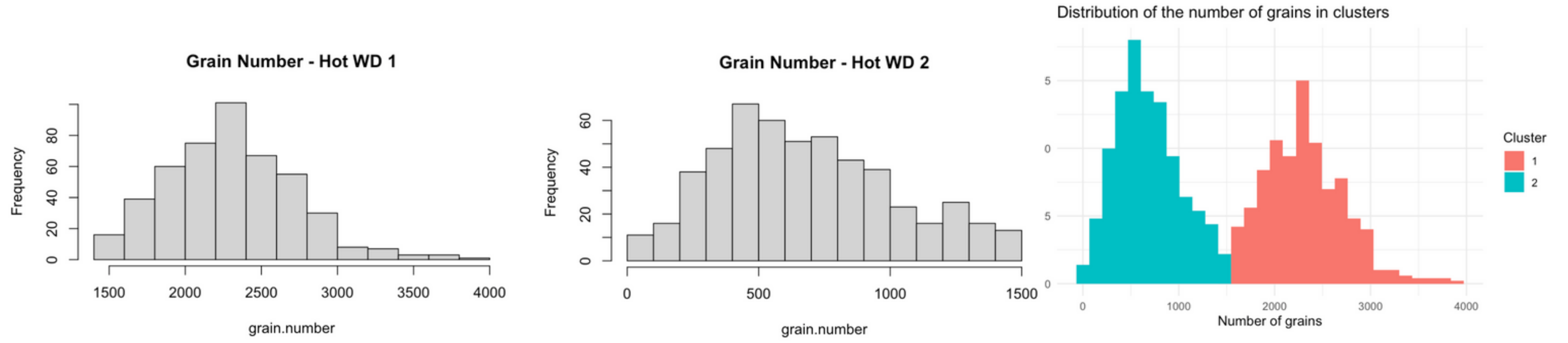
# Scenario frequencies across experiments




## Managing dimensions




## Clustering in HotWD



# Variable Selection (SIS)

**Data preparation:** For each scenario we extract the relevant columns

**Standardization:** The target variable (grain.number) is standardized

**Correlation:** We calculate the correlation with the std target variable and SNPs

**Selecting Threshold:** 3 * number of rows / log(number of rows)

**Top SNP selection:** SNPs are sorted by their abs correlation values and then selected

**Environment and SNPs Integration:** environmental variables are combined with the SNPs and the target variable to create a final dataset

**Function Application:** The select_snps function is applied to each dataset

# Analysis' Models

To execute the analysis, the **linear regression model** has been performed

Definition of parsimonious model:

- Backward regression
- Stepward regression
- Forward regression

Definition of penalied model:

- Lasso penalization
- Elasticnet penalization

Summaries made using formulas **robust to heteroskedasticity**

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} = \overset{\text{SNPs}}{\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}}_{n\times p} \times \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}_{p\times 1} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1}$$

Linear model and variables

Theoretical Model: $Y = X\beta + \epsilon$

Fitted Values: $\hat{y} = X\hat{\beta}$

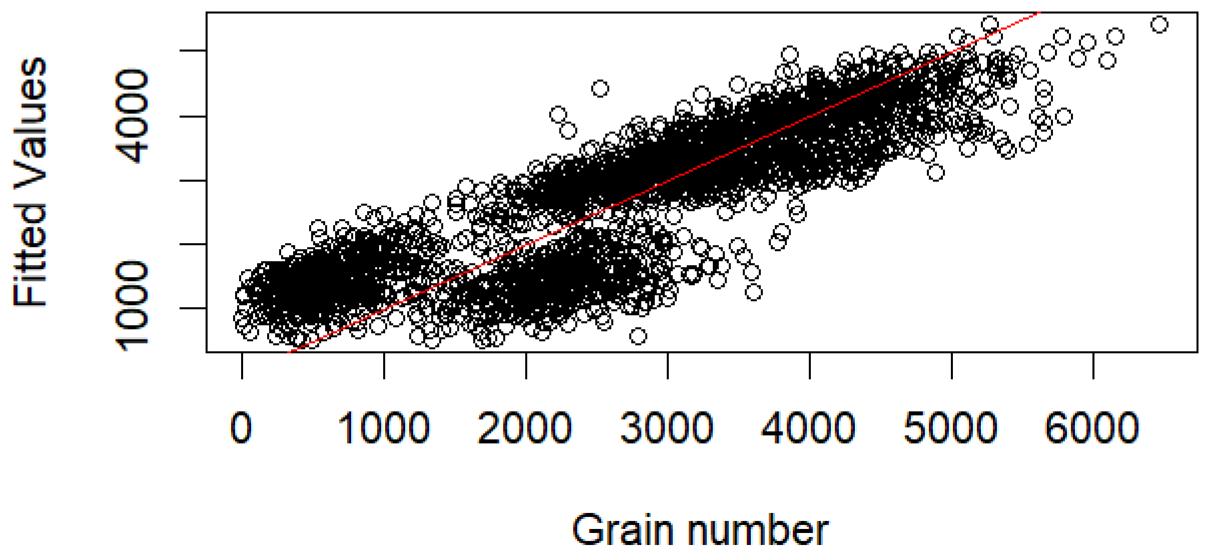
Residuals: $\hat{\epsilon} = y - \hat{y}$

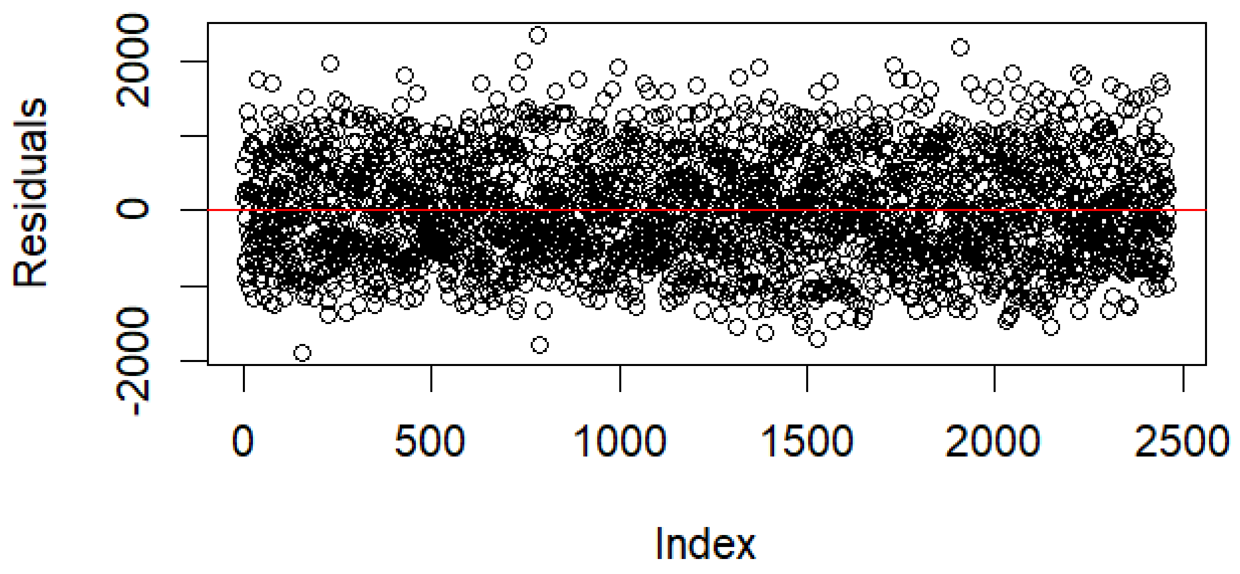
# Linear Model on the entire dataset

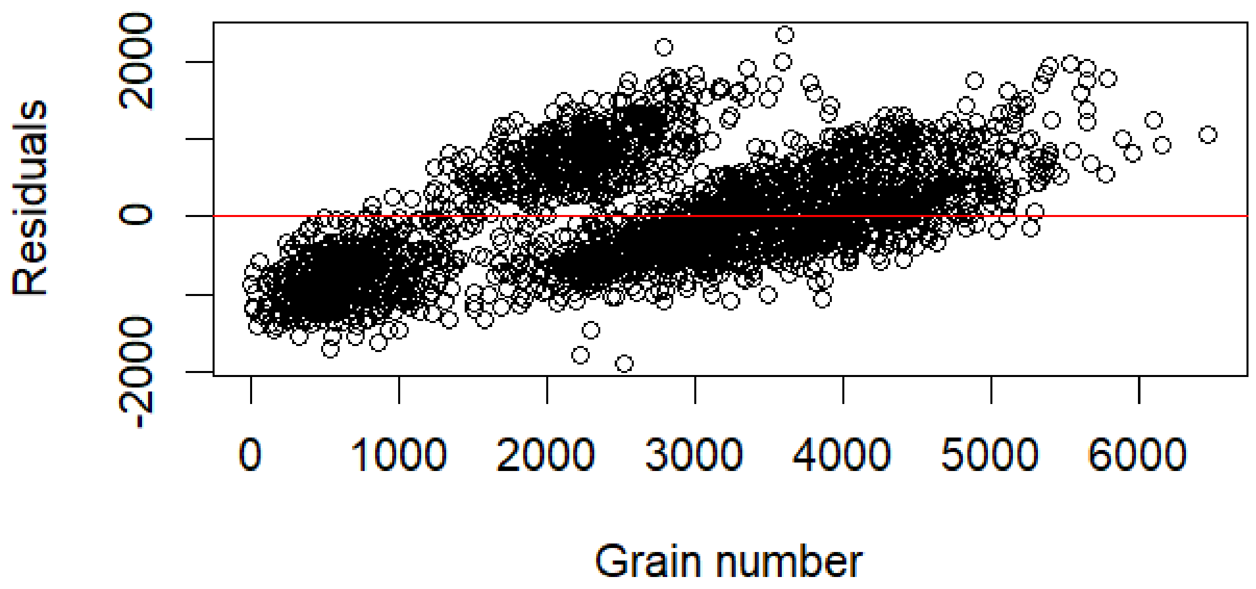
Numerical results:

- No significant regressors, highly only Temperature (-), Water (+)
- Adjusted R-squared: 0.6972

## Graphical Results




## Results and Interpretation:

Possible presence of **clusters**

Almost **centered residuals with constant variance**

Possible **linearity in residuals**

# Linear Regression CoolWW

### Standard Linear Model

Dataset: Cool_WW

Adjusted R: 0.4411

### Lasso Penalized Model

Dataset: Cool_WW

Removed variables: 161

Selected variables: 50

### Backward Regression

Dataset: Reduced Cool_WW

Adjusted R: 0.4815

Significant variables: 22

AIC: 6031.5

**Common Significant Variables** (linear model vs backward on reduced set):

**6** Common **Significant SNPs**

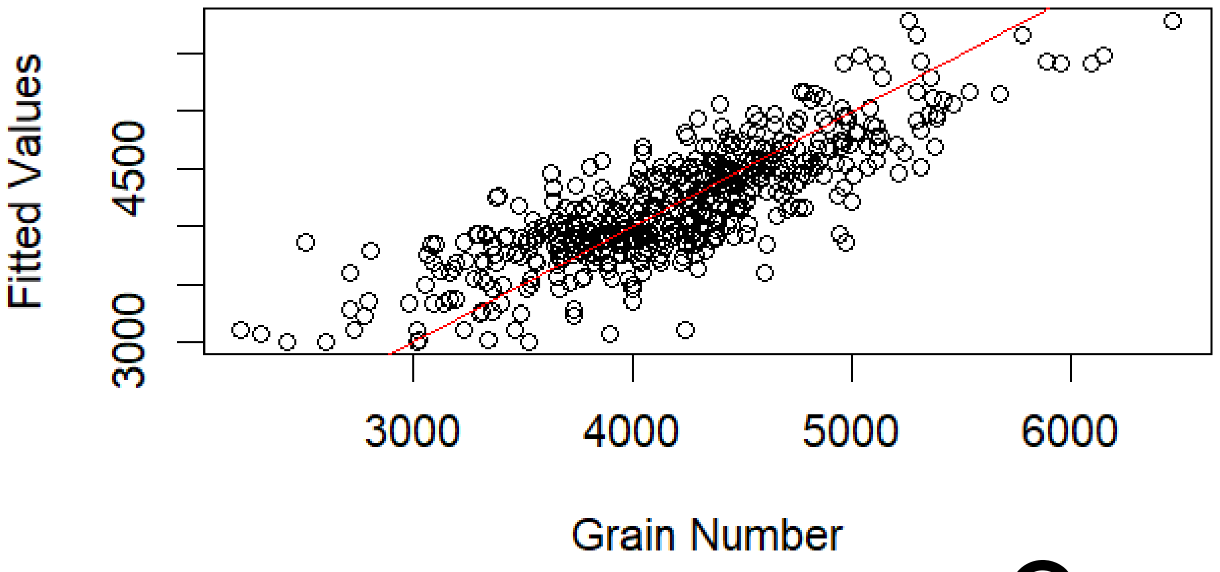
# CoolWW Graphical Comparison

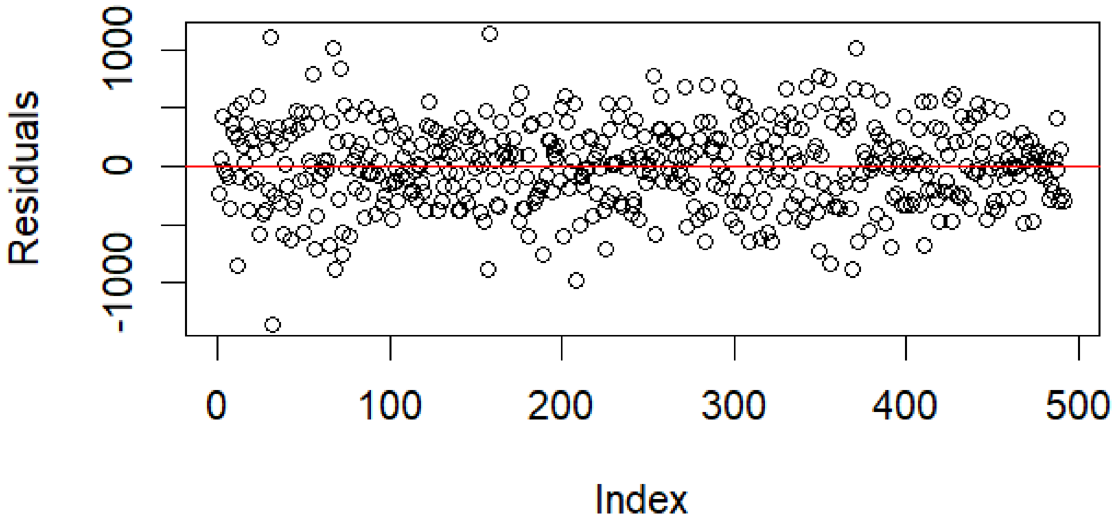
**Interpretation:**

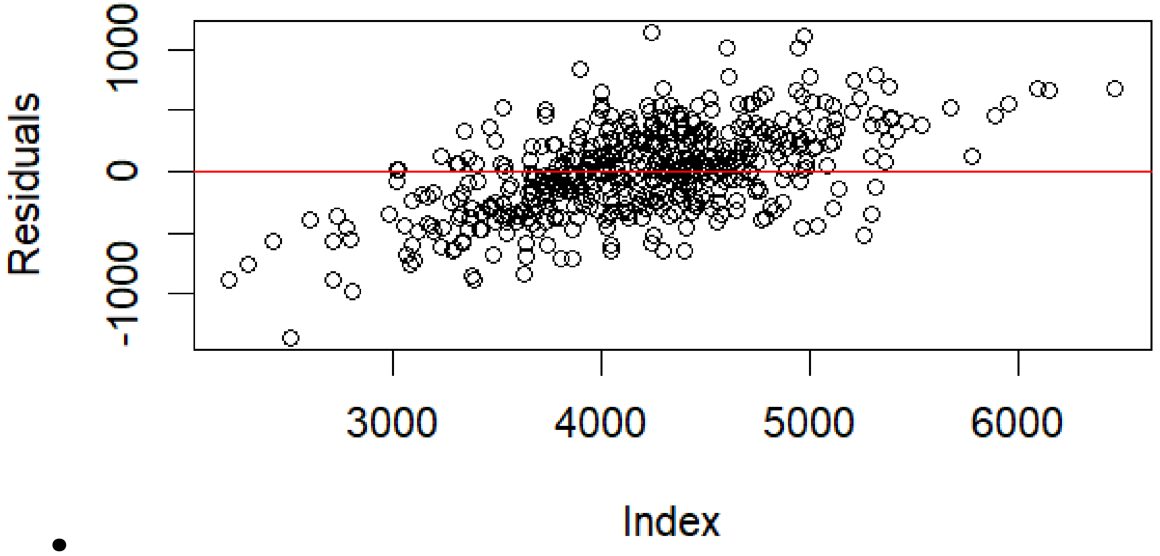
Almost **centered residuals with constant variance**

**Increasing linearity** in residuals reducing variables

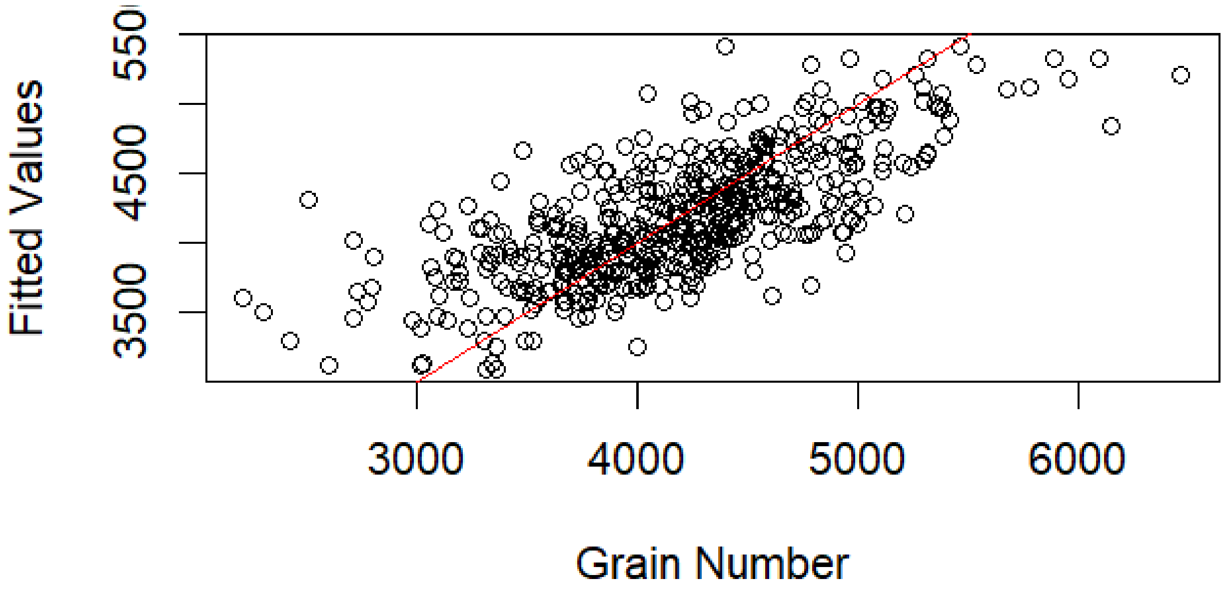
## Graphical Results: Standard Linear Model

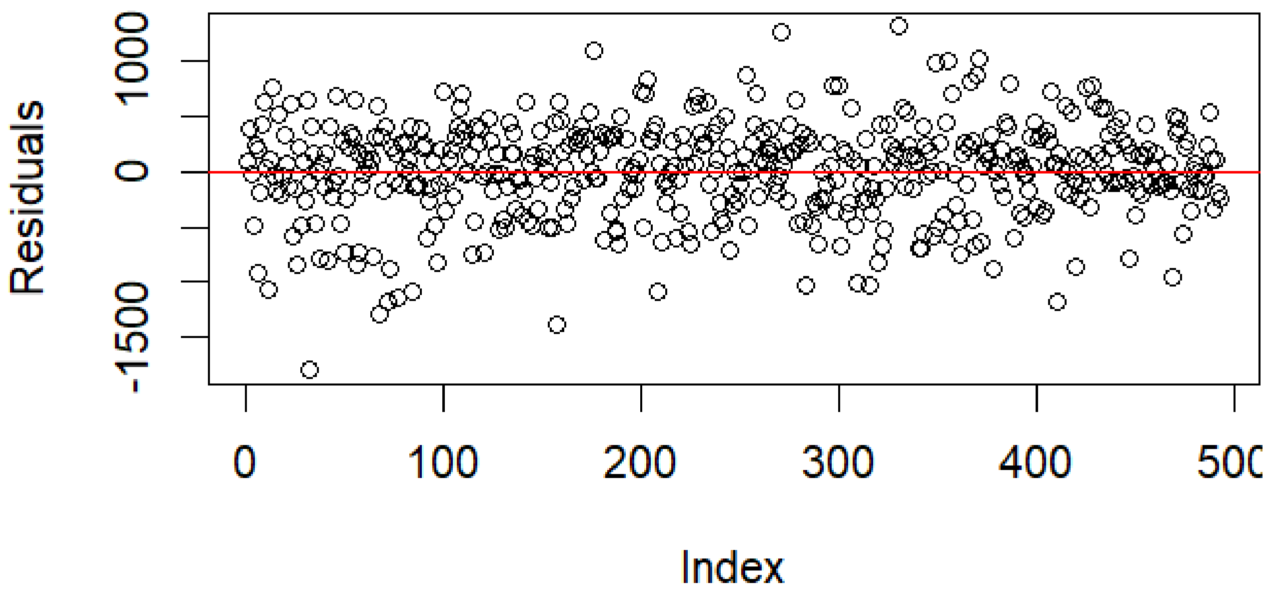


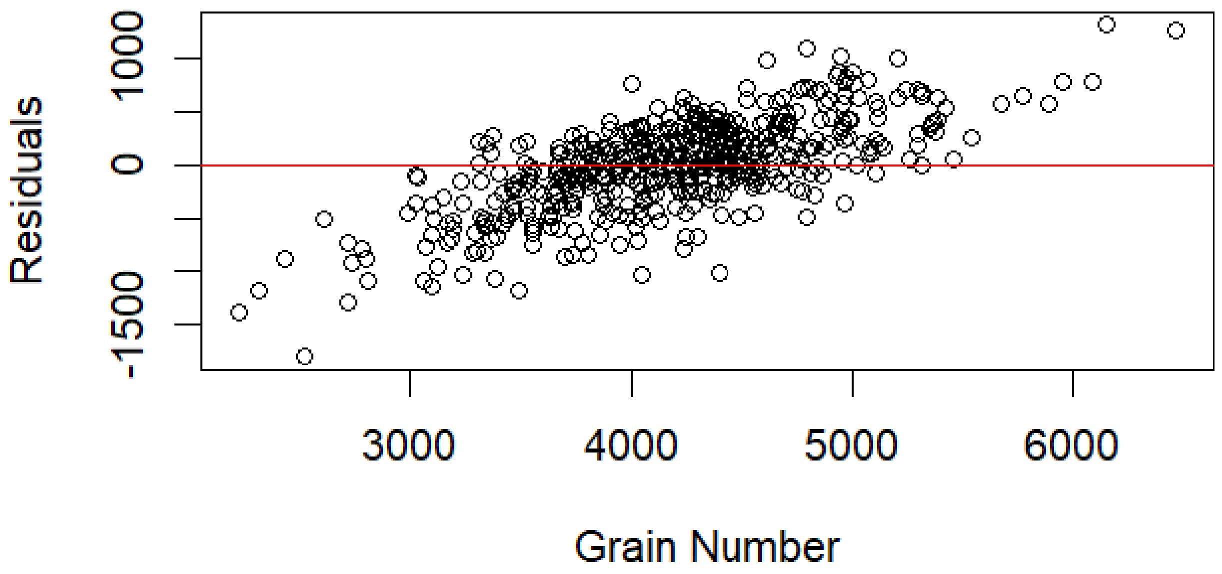




## Graphical Results: Backward Regression



# Linear Regression HotWW

### Standard Linear Model

Dataset: hot_WW

Adjusted R: 0.3601

### Lasso Penalized Model

Dataset: hot_WW

Removed variables: 325

Selected variables: 63

### Backward Regression

Dataset: Reduced hot_WW

Adjusted R: 0.397

Significant variables: 18

AIC: 12609

**Common Significant Variables** (linear model vs backward on reduced set):

**No common significant variables**

Standard Linear Model had few significant variables

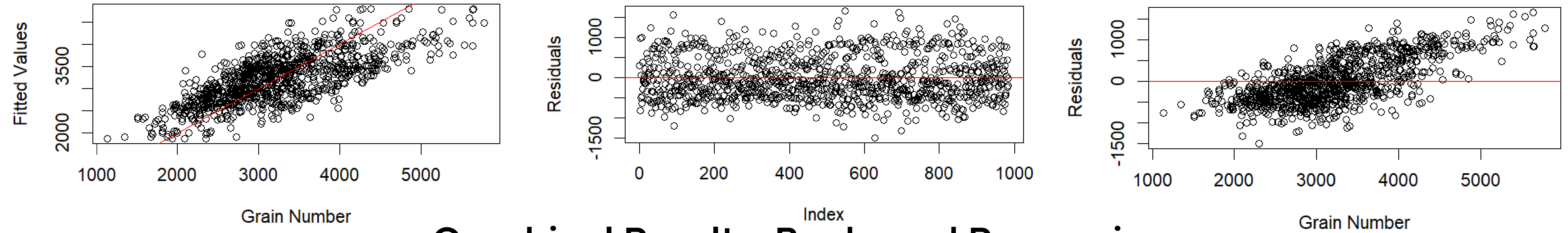
# HotWW Graphical Comparison

### Interpretation:

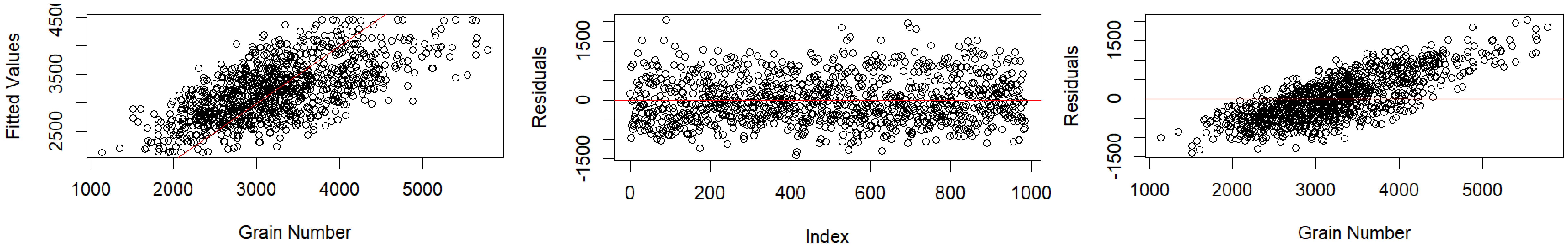
Barely de**centered residuals with constant variability**

Reducing variables slightly r**educes linearity of the model**

## Graphical Results: Standard Linear Model




## Graphical Results: Backward Regression



# Linear Regression HotWD_1

### Standard Linear Model

Dataset: hot_WD_1

Adjusted R: 0.002768

### Lasso Penalized Model

Dataset: hot_WD_1

Removed variables: 149

Selected variables: 56

### Backward Regression

Dataset: Reduced hot_WD_1

Adjusted R: 0.211

Significant variables: 7

AIC: 5535.9

**Common Significant Variables** (linear model vs backward on reduced set):

**No common significant variables**

Standard Linear Model had few significant variables

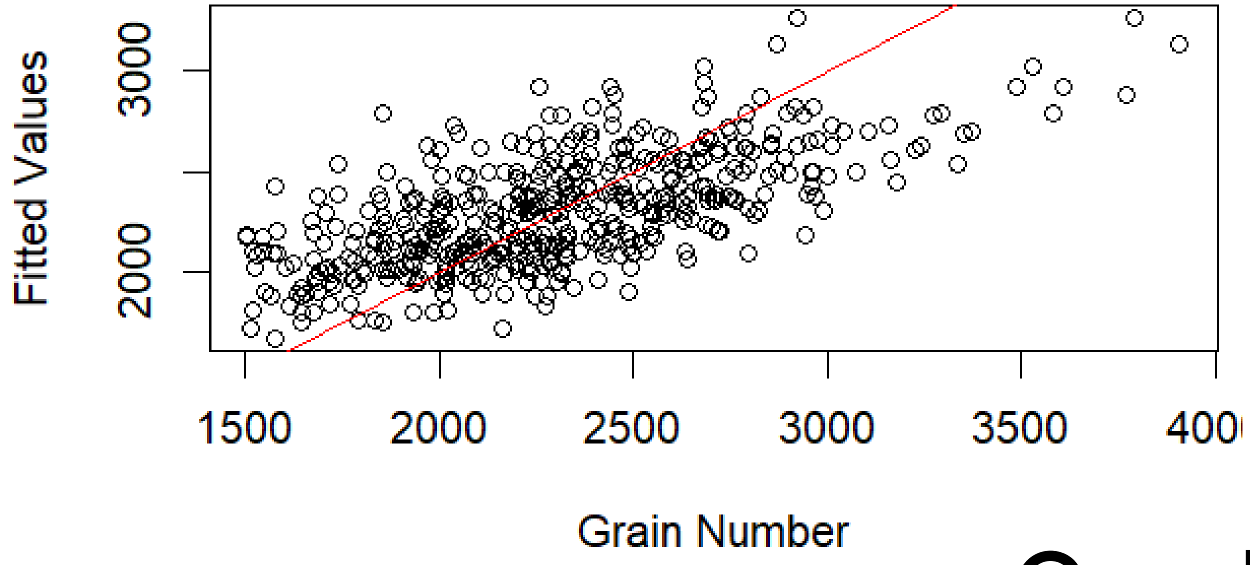
# HotWD_1 Graphical Comparison

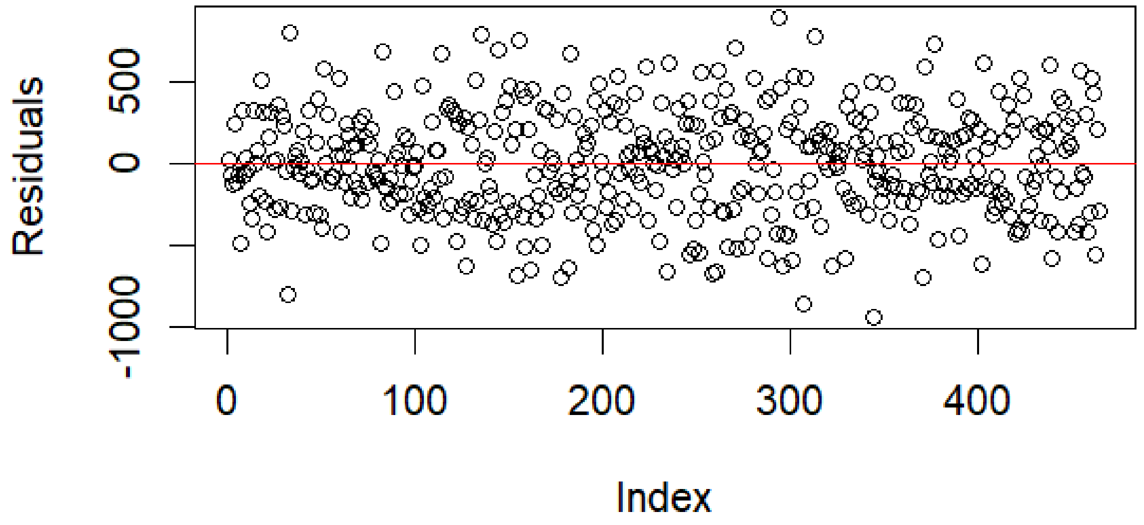
## Interpretation:

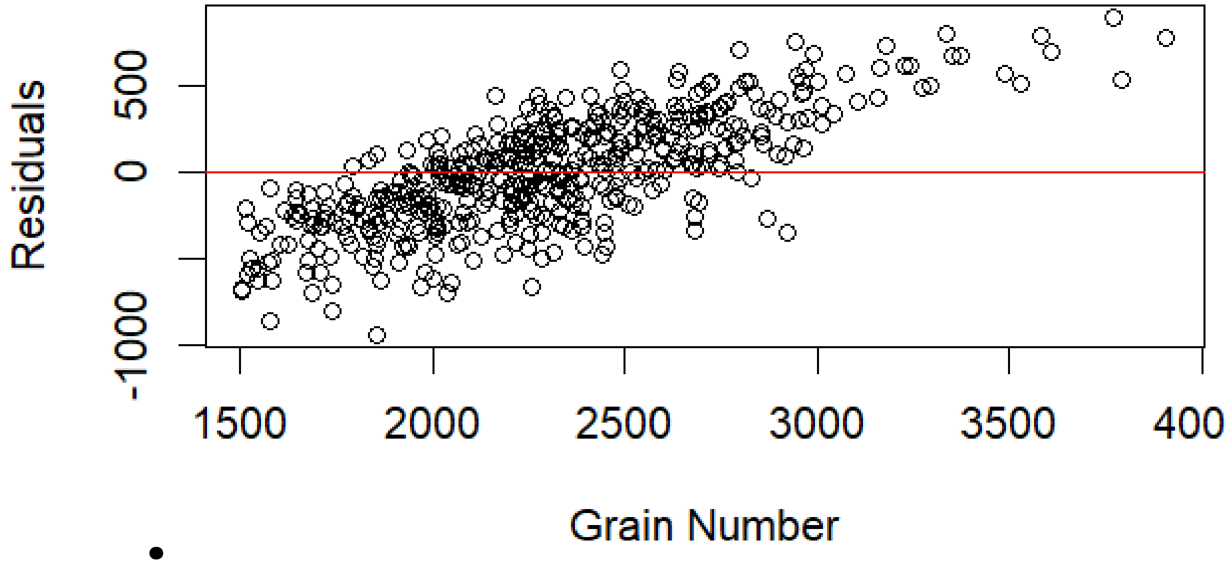
Linear Model mispecified for this dataset

Reducing variables increases a lot linearity in residuals

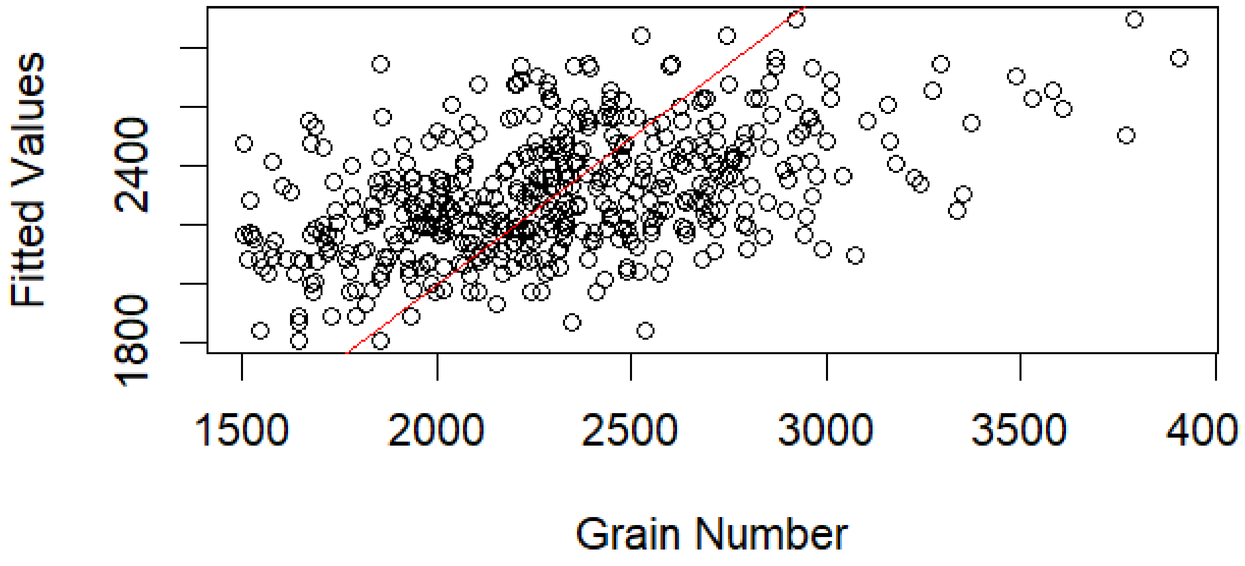
## Graphical Results: Standard Linear Model

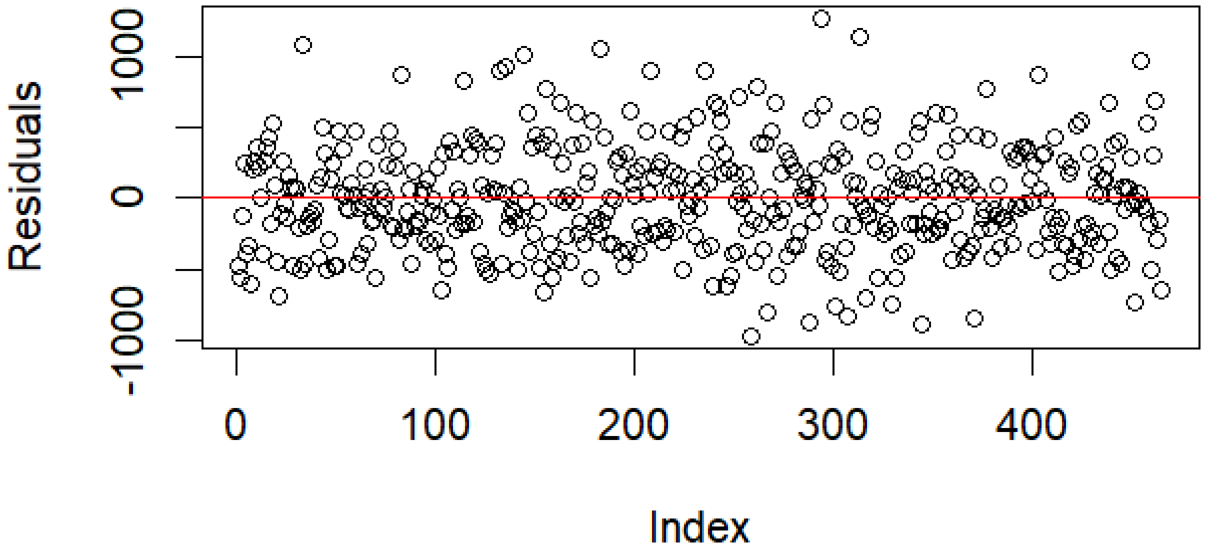


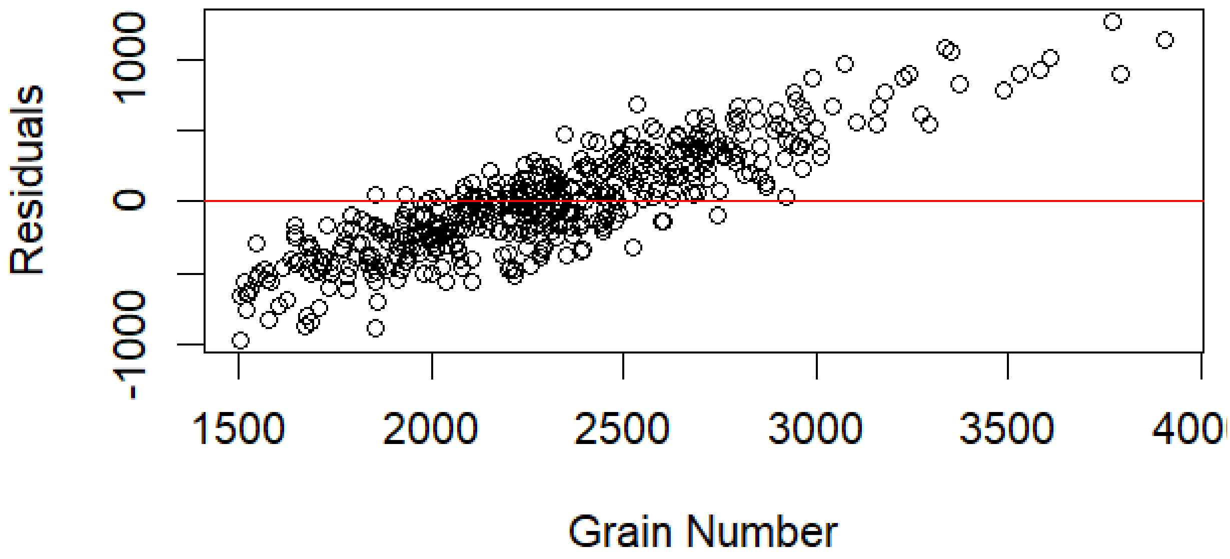




## Graphical Results: Backward Regression



# Linear Regression HotWD_2

### Standard Linear Model

Dataset: hot_WD_2

Adjusted R: 0.1792

### Lasso Penalized Model

Dataset: hot_WD_2

Removed variables: 177

Selected variables: 48

### Backward Regression

Dataset: Reduced hot_WD_2

Adjusted R: 0.3065

Significant variables: 13

AIC: 5876.3

**Common Significant Variables** (linear model vs backward on reduced set):

**No common significant variables**

Standard Linear Model had p-value of F-test very high

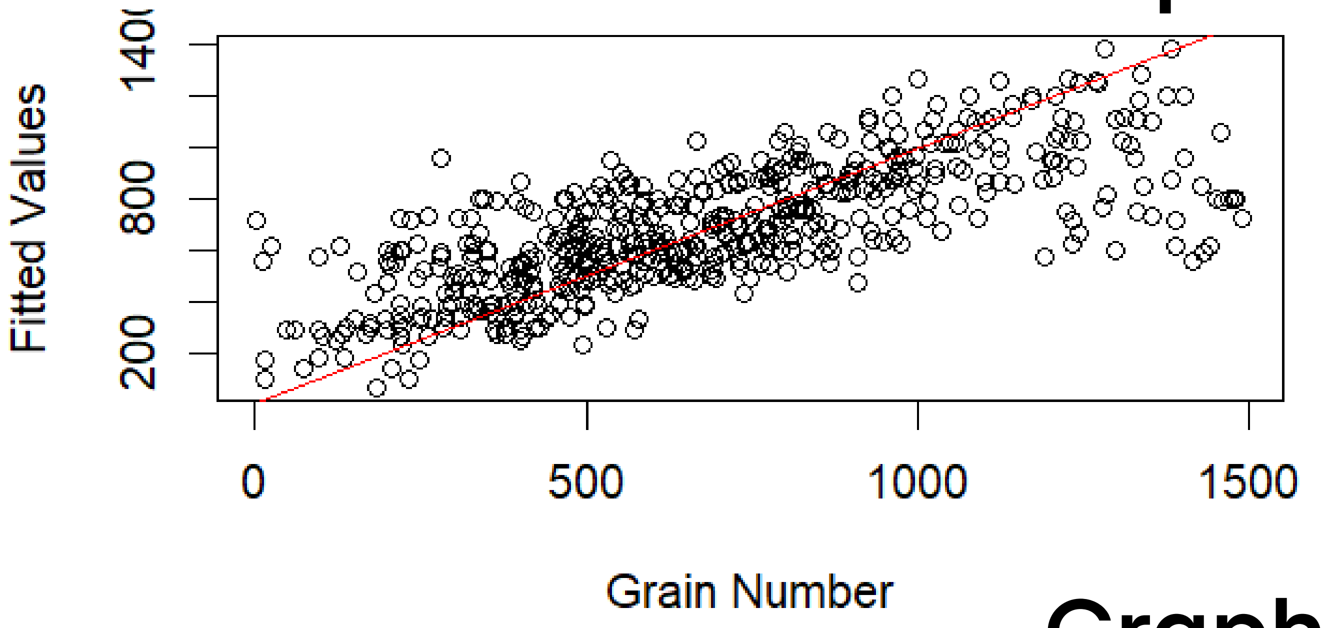
# HotWD_2 Graphical Comparison

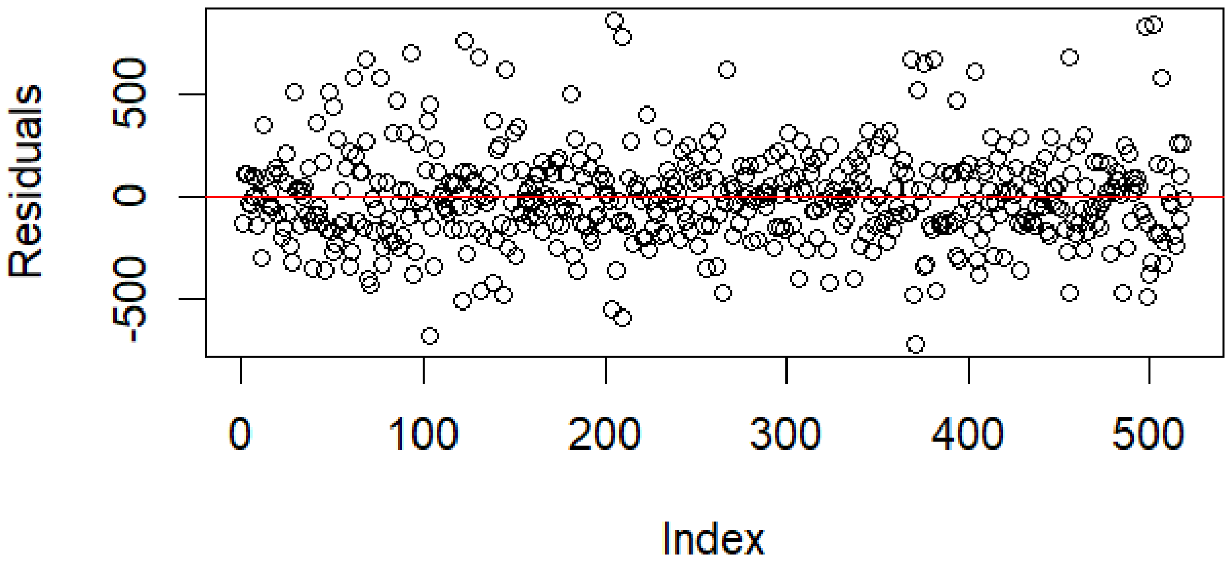
## Interpretation:

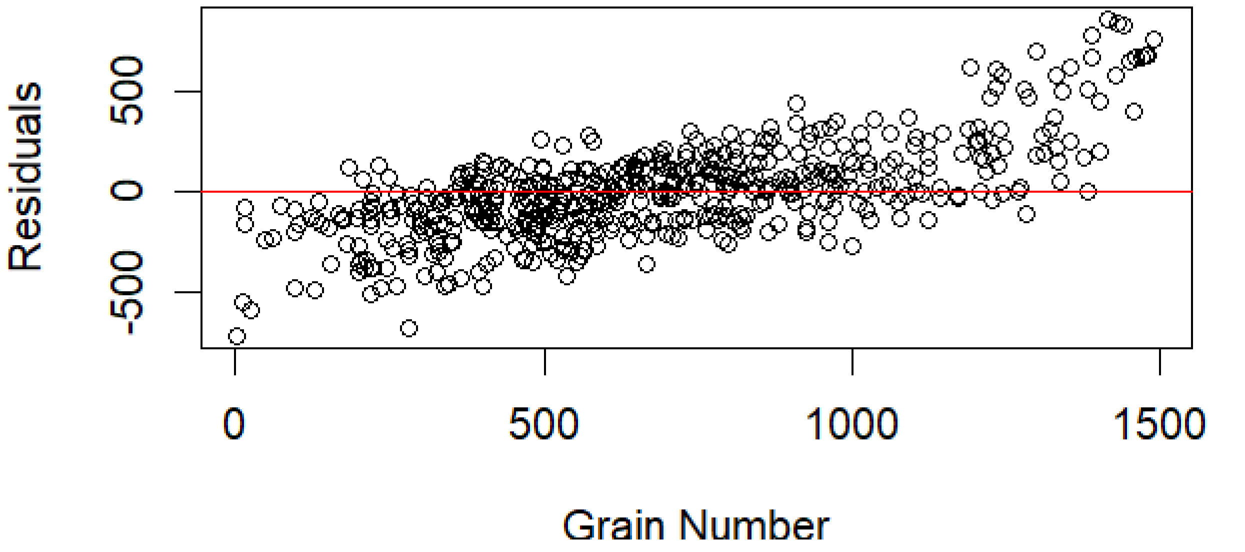
Linear Model fits better than fot hotWD_1

Not excessive linearity in residuals

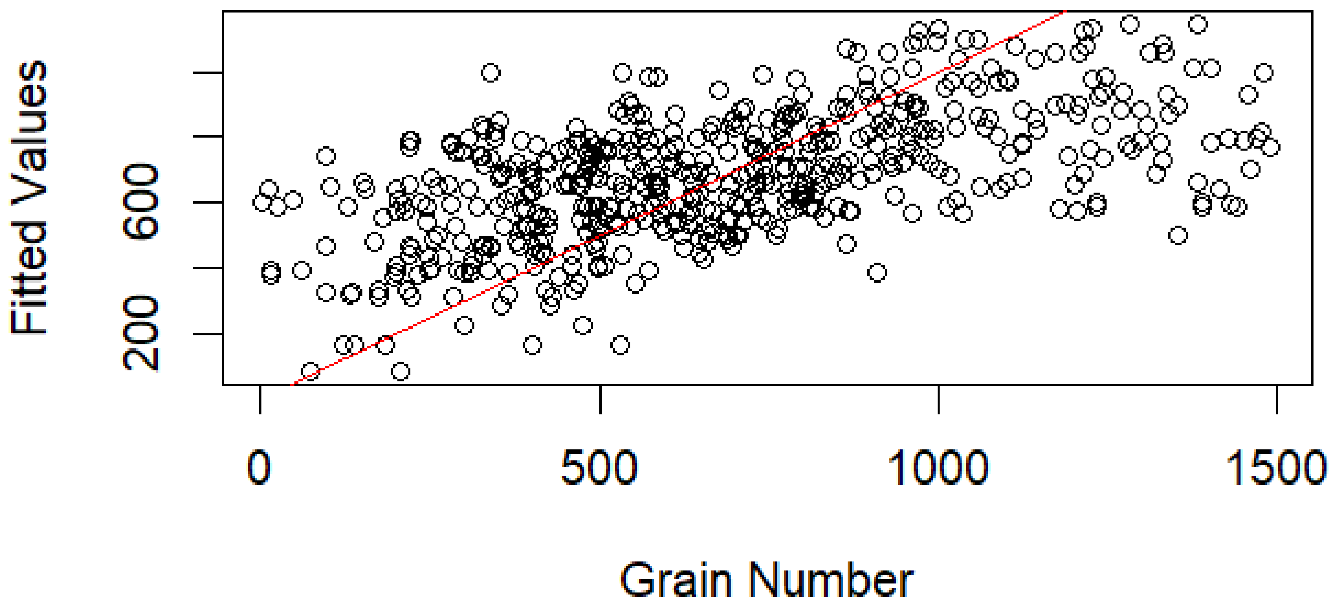
## Graphical Results: Standard Linear Model

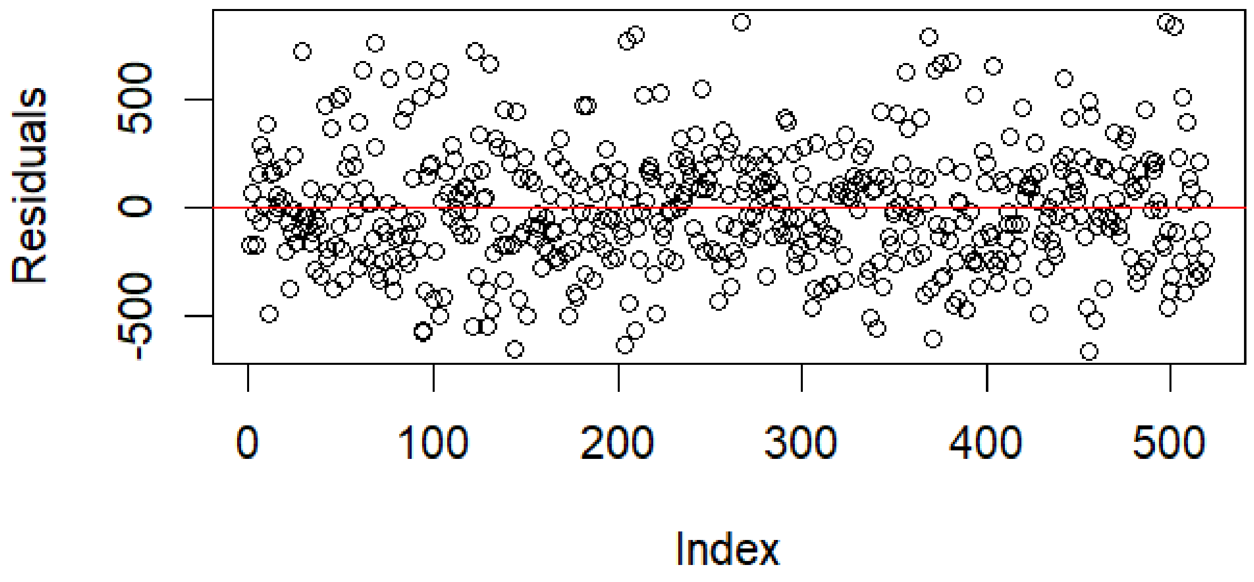


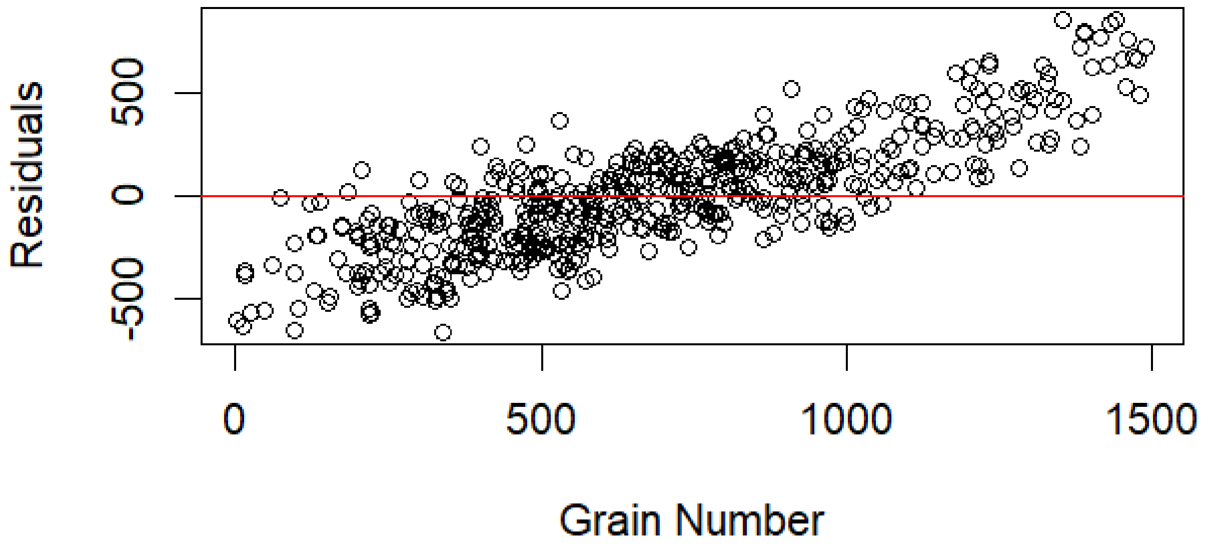




## Graphical Results: Backward Regression



# General Results

No Common Variables between hotWD_1 and hotWD_2

Linearity in Residuals increases as reducing number of variables.
Linearity in Residuals probably due to omitted variables.

No Common Variables between coolWW and HotWW

# ElasticNet

Penalization method that **combines Lasso and Ridge regression methods**

- **Handles correlations better than Lasso alone**
- **Performs variable selection**
- **Reduces overfitting**

## Results

- There aren’t significant differences in selecting the restricted dataset applying lasso or elastic net with parameter alpha=0.5.
- Some differences emerge in datasets Hot_WD_1 and Hot_WD_2, but they don’t resolve or worsen previous results

# Conclusions

## We can summarize our results in some points:

- positive relation between water availability and grain.number
- linear behaviour is more evident in better environmental conditions
- variability of results increases as temperature increases and water availability decreases
- For high temperatures and low water, linear model fits better for less productive plants: in extreme conditions SNPs can play a fudamental role in defining plant productivity, but the impact of environment is stronger

# Thank you