Danmarks Tekniske Universitet

DTU

Analysis of Correlated Data
Assignment 4

Author

Marios-Dimitrios Lianos - s233558

November 29, 2023

Summary

At the outset of our statistical analysis, we investigated the influence of location and nitrogen levels, including their interactive effects on agricultural yield. Through a rigorous mixed-effects modeling approach, significant findings were uncovered that elucidate the dynamics of yield across different environmental and treatment conditions.
Our analysis confirmed that location is a significant factor in yield outcomes, with Knoxville exhibiting a notably higher yield than Jackson. The effect of nitrogen on yield was found to be substantial and positive, emphasizing the importance of this nutrient in crop production. Moreover, the relationship between nitrogen levels and yield was discovered to be non-linear, with a plateauing effect at higher nitrogen levels suggesting a point of diminishing returns. The interaction between location and the quadratic term of nitrogen also emerged as significant, indicating a complex interplay between these factors that affects yield differently depending on the location. This points to the potential need for location-specific fertilization strategies.

Contents

1 Introduction **1**

2 Exploratory Data Analysis **2**
2.1 Variables in the Dataset 2
2.2 Explorative Plots and Diagrams 3
2.3 Exploratory Model Analysis 5

3 Statistical Analysis **8**
3.1 Model Diagnostics 8
3.2 Results 15
3.3 Post-hoc Analysis 16

4 Conclusion **17**

5 Appendix A **18**

1 Introduction

Our study embarks on a comprehensive examination of the effects of various levels of nitrogen fertilizer on corn yield. This investigation was conducted across two distinct locations in Tennessee, encompassing a span of five years. The research involves the application of six different amounts of nitrogen fertilizer to a single variety of corn. The primary objective is to gather insightful data that could aid in optimizing the use of nitrogen fertilizer in corn cultivation.

2 Exploratory Data Analysis

2.1 Variables in the Dataset

- `loc` (Location): A categorical variable representing the two distinct locations in Tennessee Jackson and Knoxville. The inclusion of location allows us to examine how different geographical and environmental conditions influence corn yield.
- `year` (1962-1966): While primarily serving as a categorical variable, the year captures the temporal dimension of our study, spanning from 1962 to 1966. This aspect is crucial for understanding year-to-year variations and for making inferences that extend beyond the observed five-year period.
- `nitro` (Nitrogen Fertilizer Level): A numerical variable, nitro quantifies the amount of nitrogen fertilizer applied. Our analysis will not only consider the linear impact of nitrogen levels on yield but also explore higher-order terms (such as quadratic or cubic effects) to capture more complex relationships.
- `yield`: This is a continuous variable that represents the measured yield of corn.

It is crucial to highlight that the dataset appears to be balanced in terms of the combination of factors location (loc), year (year), and nitrogen level (nitro). Each unique combination of these three factors has an equal count, indicating that every level of nitrogen fertilizer is applied at each location for each year in the dataset.

- `Crossed Factors`: In the dataset, the factors are crossed, meaning that each level of one factor is combined with each level of the other factors.
- `Fixed Effects`: The Nitrogen Fertilizer Level (nitro), potentially including its higher-order terms, and the Location (loc) are the main fixed effects.
- `Random Effects`: The Year (year) is considered as a random effect in our model to account for variability over time. This approach helps capture potential influences of annual climatic and environmental differences on corn yield.

2.2 Explorative Plots and Diagrams

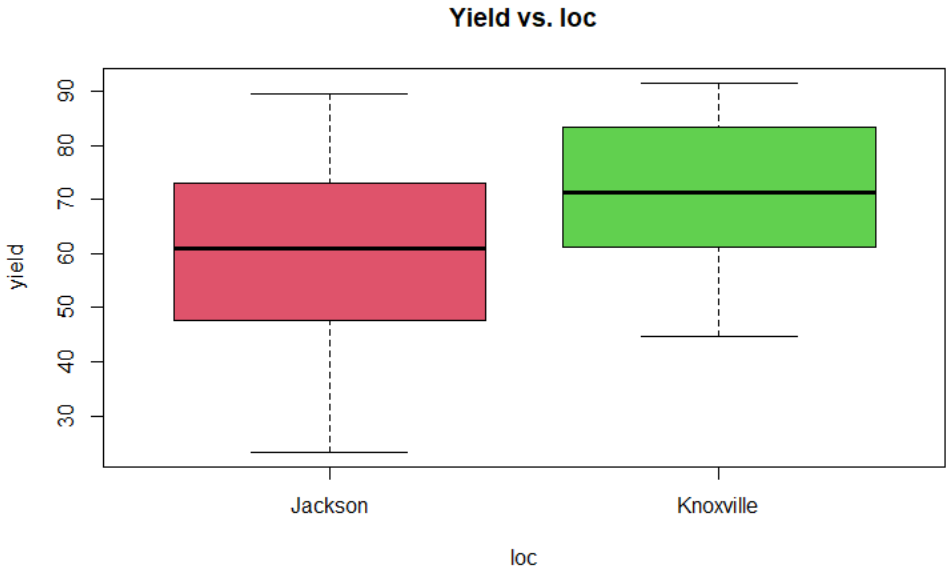

Figure 1: Boxplot Yield - Location

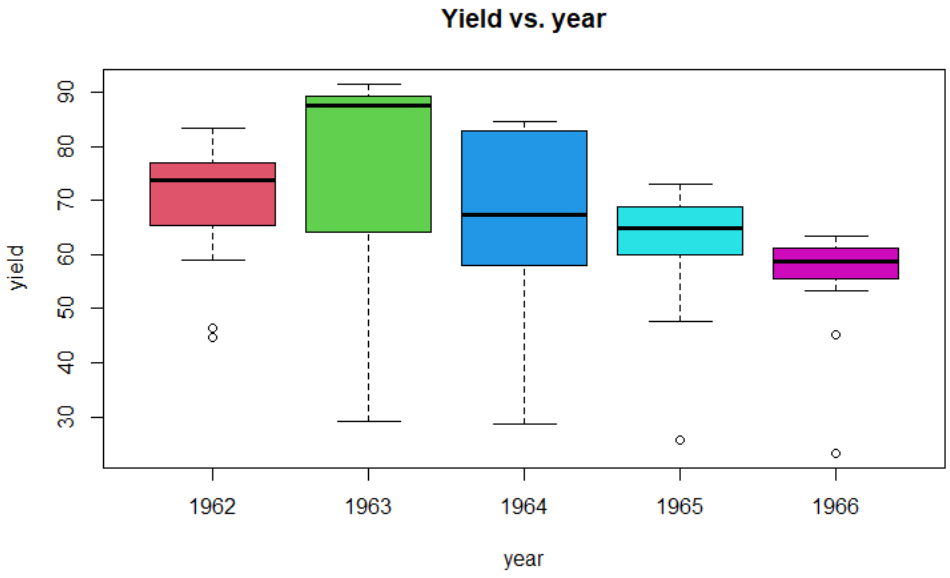

Figure 2: Boxplot Yield - Year

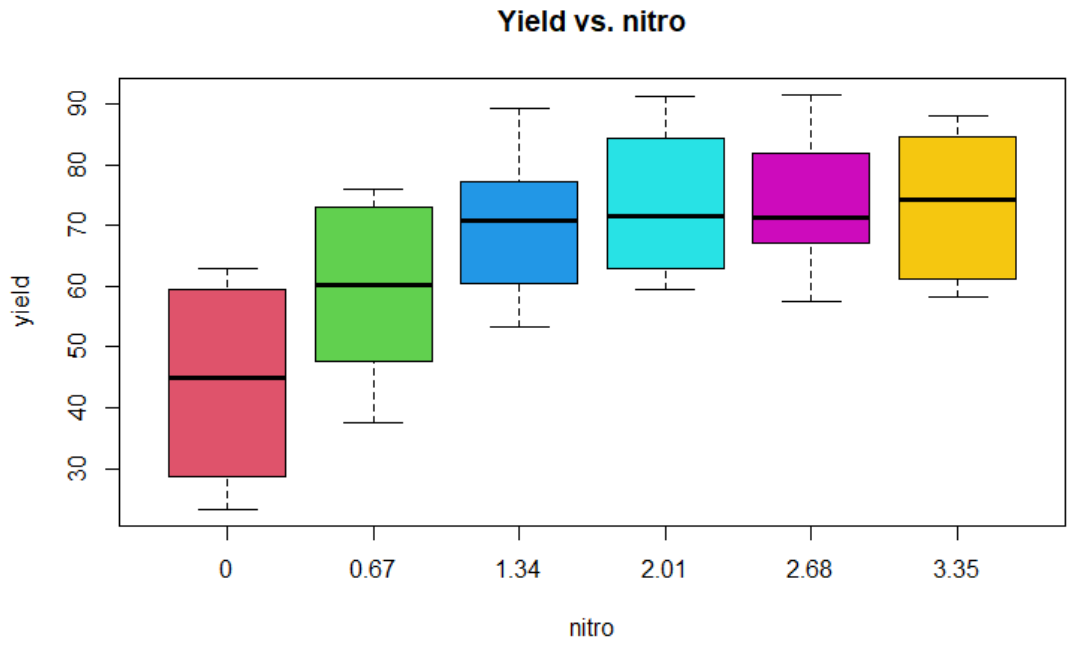

Figure 3: Boxplot Yield - Nitrogen

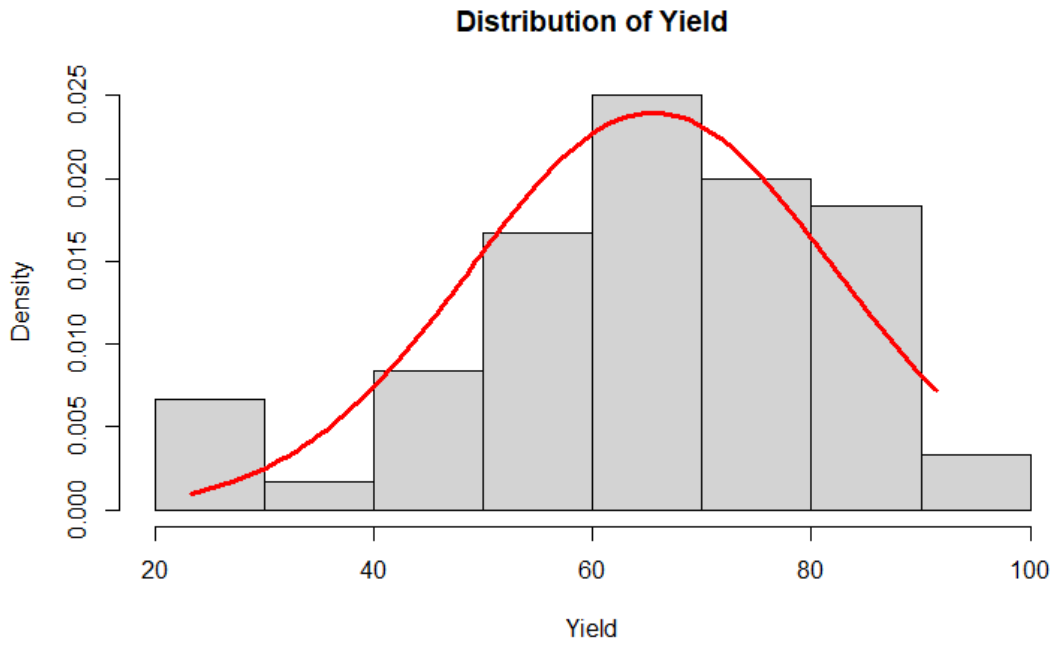

Figure 4: Histogram

2.3 Exploratory Model Analysis

Representation of the model:

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \delta_{ik} + (\alpha\beta)_{ij} + b_{k(i)} + c_{kl} + d_k + e_l + f_{k(j)} + g_{l(j)} + \epsilon_{ijkl} \quad (1)$$

Y_{ijkl}: is the observed yield,μ the overall mean yield. α_i is the fixed effect of the i-th level of loc, β_j is the fixed effect of the j-th level of nitro, $delta_{ik}$the quadratic effect of nitro at the i-th level of loc.$(\alpha\beta)_{ij}$ the interaction effect between loc and nitro. $b_{k(i)}$ is the random intercept and slope for nitro including its quadratic term for each combination of year:loc. c_{kl} is the random intercept for each combination of year:loc, d_k is the random intercept for each year:nitro combination. e_l is the random intercept for each year. $f_{k(j)}$ is the random slope for nitro for each year and $g_{l(j)}$ the random slope for nitro for each year:loc. ϵ_{ijkl} is the residual error.

The next procedure is a way to examine which terms can be dropped off from (1). When considering which terms to potentially exclude from a mixed-effects model, we typically start by examining the most complex terms and progressively move to simpler terms, ensuring at each step that the model still adequately fits the data.

First term that we will examine: $(1 + I(nitro^2) + nitro|year : loc)$

Model	logLik	AIC	BIC	Pr(>Chisq)
model2	-193.85	419.7	453.21	
model1	-182.95	409.9	455.97	0.001318 **
'log Lik.' 1.747123e-06				

Table 1: Anova for model2, model1

Then the model without: $(I(nitro^2) : loc)$

Model	logLik	AIC	BIC	Pr(>Chisq)
model2	-191.05	422.1	463.99	
model1	-182.95	409.90	455.97	0.0003031
'log Lik.' 2.679985e-06				

Table 2: Anova for model3, model1

From Tables: 1 and 2 we see that $(I(nitro^2) : loc)$ and $(1 + I(nitro^2)|year : loc)$ cannot be dropped.

Then we will examine: $(loc : nitro)$

Model	logLik	AIC	BIC	Pr(>Chisq)
model3	-184.01	410.03	454.01	
model1	-182.95	409.90	455.97	0.1445
'log Lik.' 0.002328155				

Table 3: Anova for model4, model1

$(1 + nitro|year : loc)$

Model	logLik	AIC	BIC	Pr(>Chisq)
model4	-184.01	404.03	441.73	
model3	-184.01	410.03	454.01	1
'log Lik.' 0.4469559				

Table 4: Anova for model5, model4

$(1|year : nitro)$

Model	logLik	AIC	BIC	Pr(>Chisq)
model5	-184.01	402.03	437.63	
model4	-184.01	404.03	441.73	1
'log Lik.' 0.5				

Table 5: Anova for model6, model5

$(1 + nitro|year)$

Model	logLik	AIC	BIC	Pr(>Chisq)
model6	-184.01	396.03	425.35	
model5	-184.01	402.03	437.63	1
'log Lik.' 0.4411685				

Table 6: Anova for model7, model6

From Tables: 3, 4, 5 and 6 we can exclude all of the above terms from our model.

Finally, the term: $(1|year : loc)$

Model	logLik	AIC	BIC	Pr(>Chisq)
model7	-184.01	394.03	421.25	
model6	-184.01	396.03	425.35	1
'log Lik.' 0.5				

Table 7: Anova for model8, model6

Finally, the term: $(1|year)$

Model	logLik	AIC	BIC	Pr(>Chisq)
model8	-184.01	392.03	417.16	
model7	-184.01	394.03	421.25	1
'log Lik.' 0.4998249				

Table 8: Anova for model8, model6

In Tables: 7 and 9 we see that we can drop both terms from the model.
As it occurs from the above examination, the final model in which we will conduct model diagnostics, to see if we need to further modify the model is:

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \delta_{ik} + (b_{0kl} + b_{1kl} \cdot \mathrm{I}(\mathrm{nitro}^2) + b_{2kl} \cdot \mathrm{nitro})_{ijkl} + \epsilon_{ijkl} \quad (2)$$

Y_{ijkl} is the observed yield, μ is the overall mean yield, α_i is the fixed effect of the i-th level of loc, and β_j is the fixed effect of the j-th level of nitro. The term δ_{ik} represents the interaction effect between the squared nitro and loc at the i-th level of loc. The random effects b_{0kl}, b_{1kl}, and b_{2kl} correspond to the random intercept, the random slope for the squared nitro, and the random slope for nitro within each combination of $year : loc$, respectively. Finally, ϵ_{ijkl} is the residual error.

For a better understanding of the final model, its representation in R is:

$model.lmer <- lmer(yield \ \ loc + nitro + I(nitro^2) : loc + (1 + I(nitro^2) + nitro|year : loc), data = assignment4)$

3 Statistical Analysis

3.1 Model Diagnostics

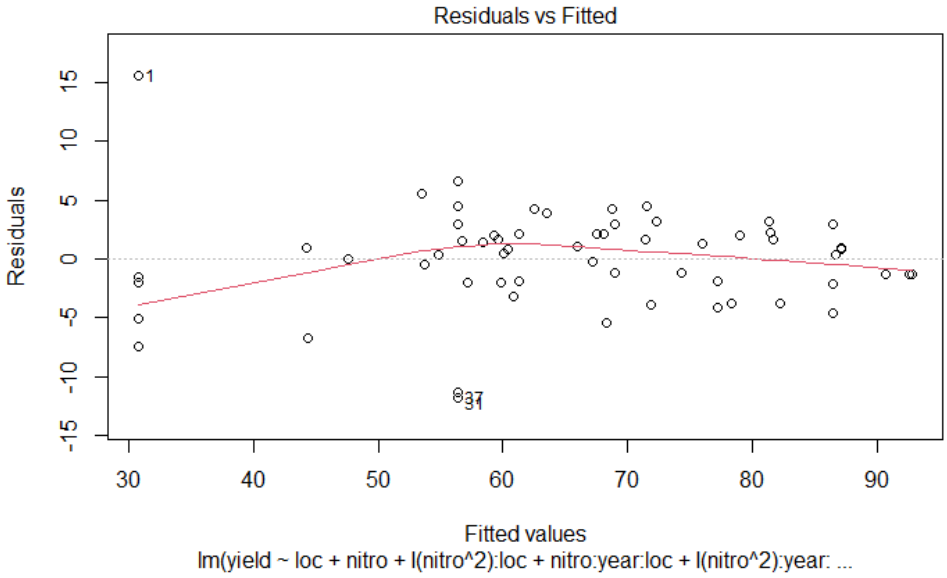

Figure 5: Residuals vs Fitted values

Fig: 5 and Fig: 6 display residuals that are largely symmetrical in their distribution, closely aligning with a normal distribution, except for a few notable outliers at the extremes. Fig: 9 and 10 show that locations and years do not have big differences in variability. In Fig: 11 we see that we have to make some adjustments in the coefficient of yield ($yield^{2.6}$).

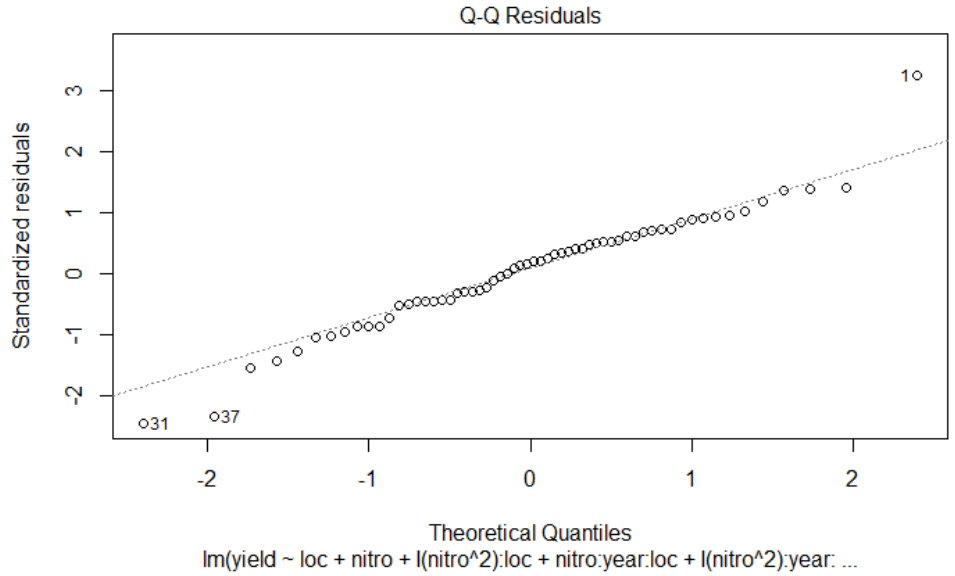

Figure 6: Q-Q Residuals

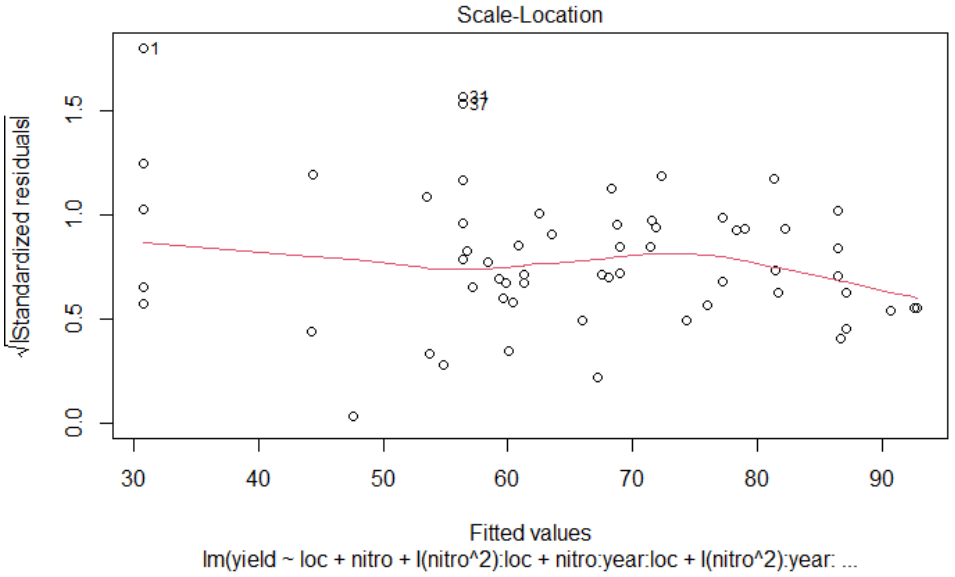

Figure 7: Scale-Location

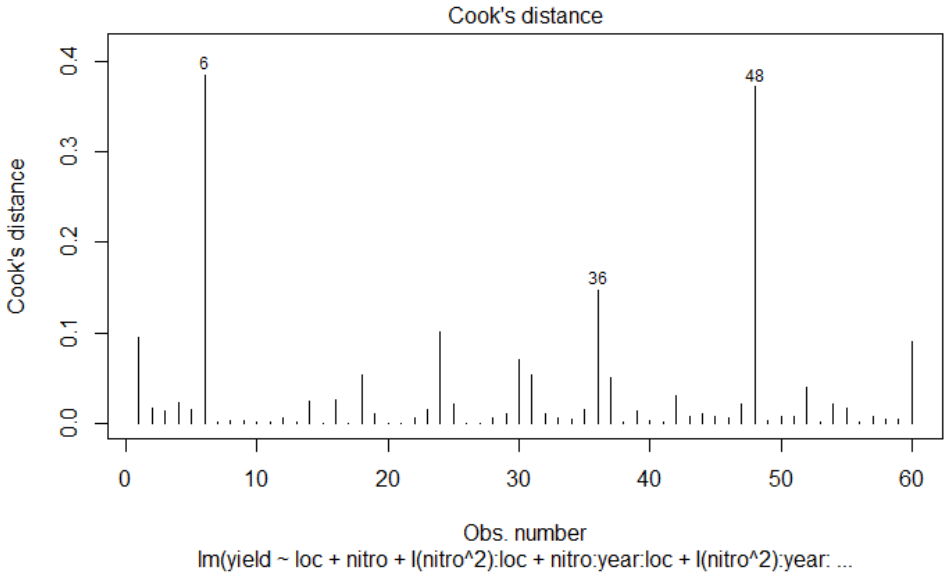

Figure 8: Cooks distance

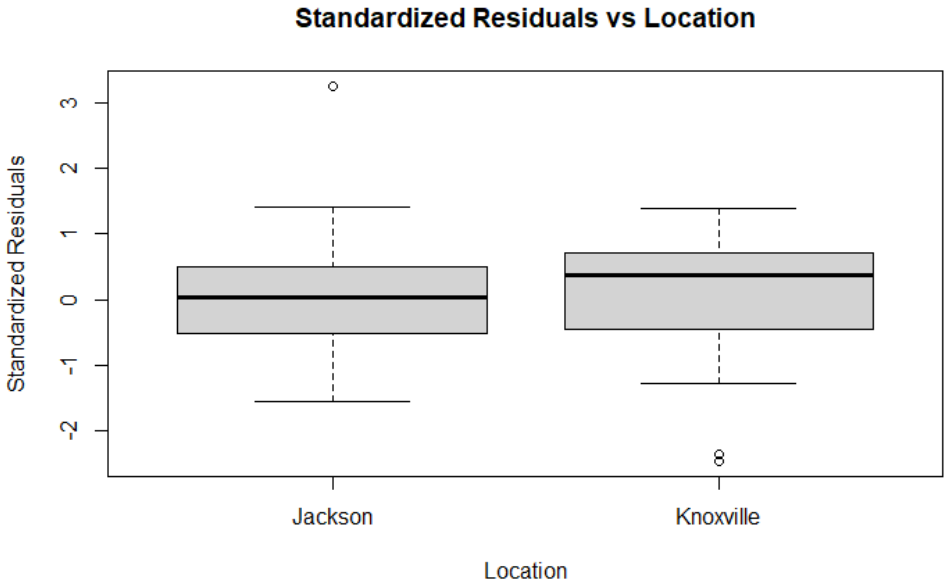

Figure 9: Studentized Residuals vs Location

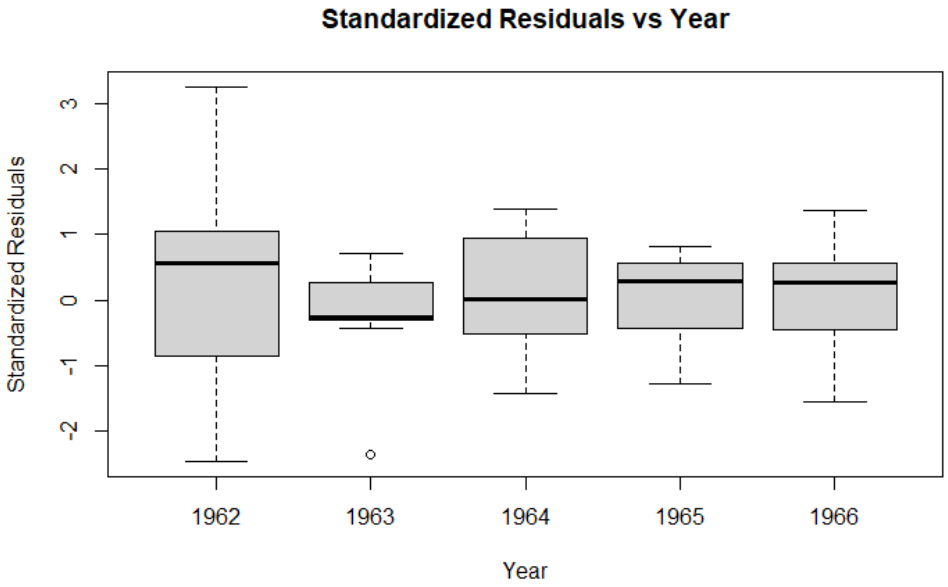

Figure 10: Studentized Residuals vs Year

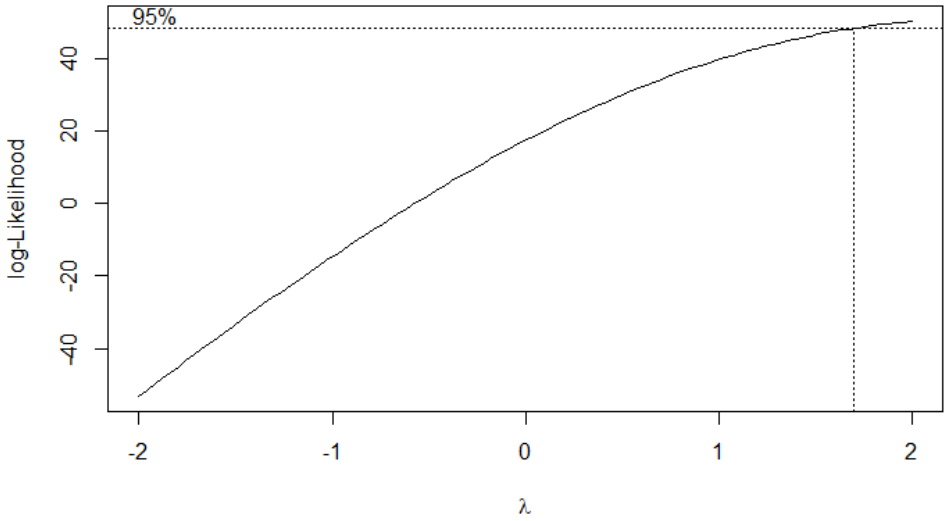

Figure 11: Box Cox

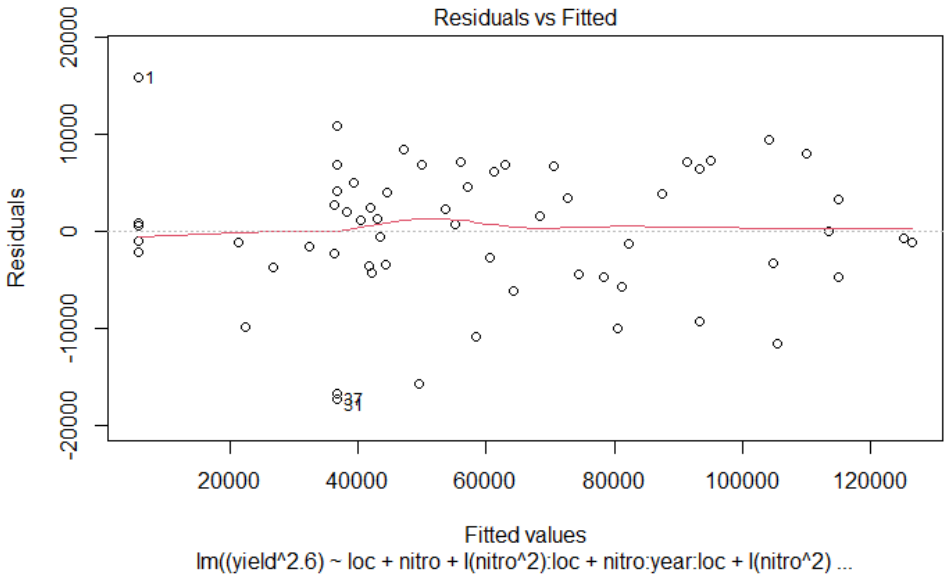

Figure 12: Residuals vs Fitted values

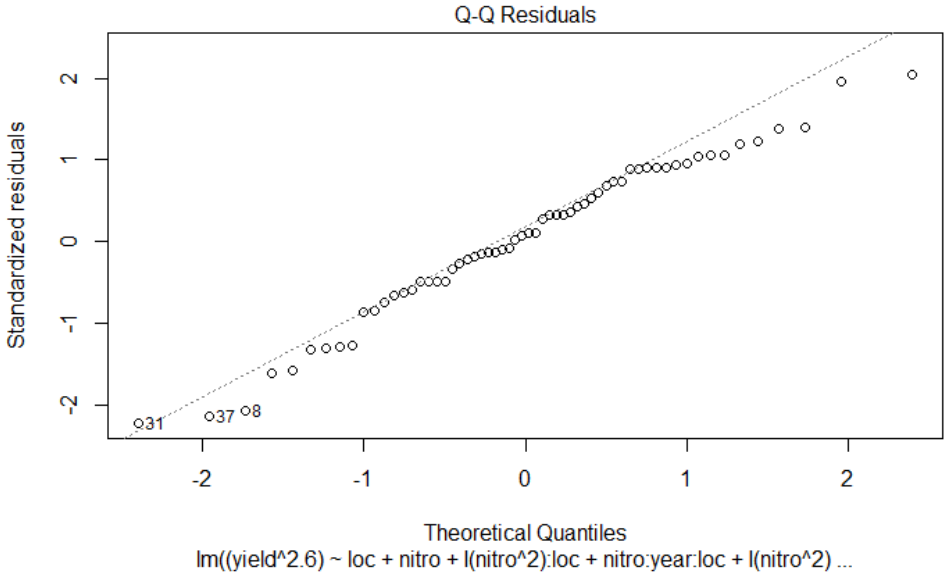

Figure 13: Q-Q Residuals

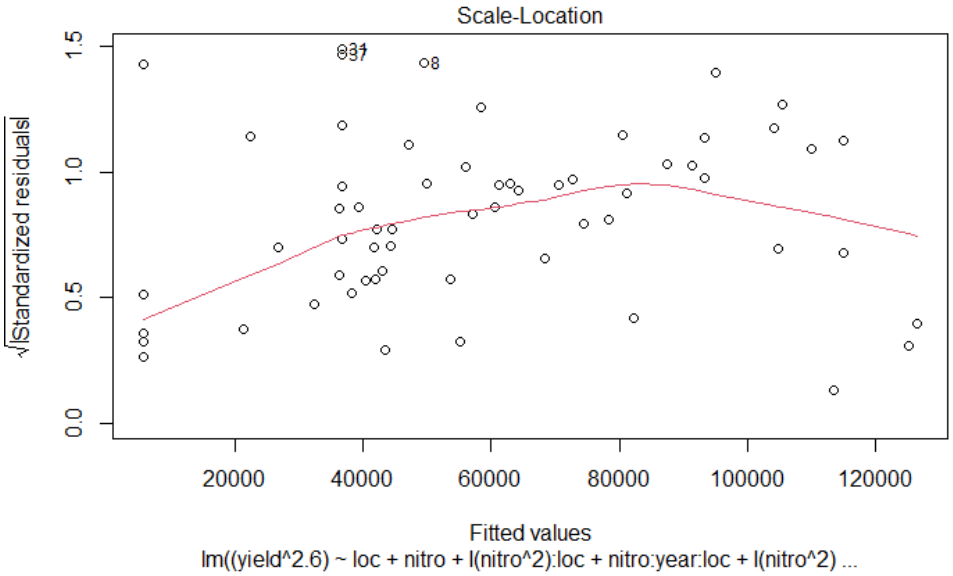

Figure 14: Scale-Location

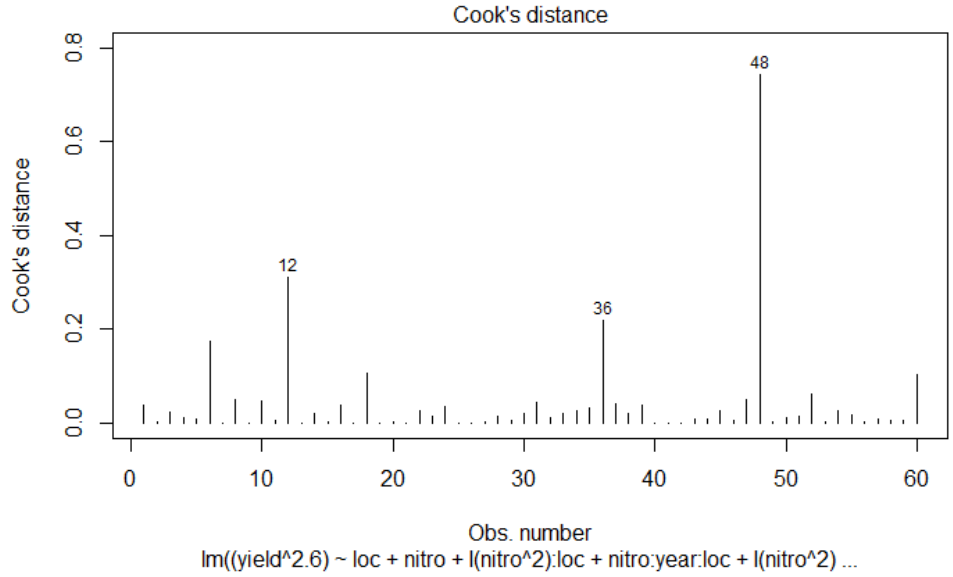

Figure 15: Cooks ddistance

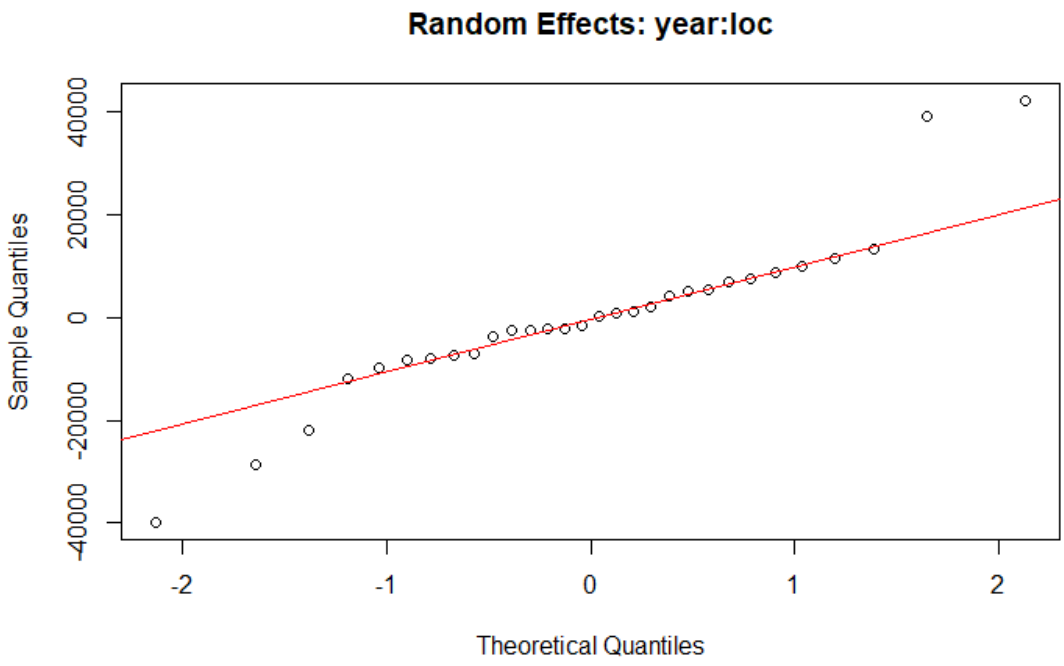

Figure 16: QQ Plot for Random Effects

The Q-Q plot suggest that the assumption of normality for the random effects in the mixed-effects model is reasonable. While there are minor deviations, they do not appear to be significant enough to suggest major violations of the normality assumption.

3.2 Results

Term	Estimate	Std. Error	Pr(>|t|)
(Intercept)	5773.673	4301.616	0.207690
locKnoxville	30880.395	5120.605	0.000313 ***
nitro	43055.575	9145.092	0.001097 **
$locJackson : I(nitro^2)$	-7030.070	1852.625	0.003663 **
$locKnoxville : I(nitro^2)$	-9406.315	1852.625	0.000515 ***

Table 9: Linear Model Analysis Results

Term	NumDF	DenDF	F value	Pr(>F)
loc	1	7.9997	36.368	0.0003126 ***
nitro	1	9.0322	22.166	0.0010968 **
loc:I(nitro^2)	2	13.3223	17.428	0.0001911 ***

Table 10: ANOVA Summary Table

From Table: 9 we see that *locKnoxville* has a significant positive effect on *yield* compared to the baseline location. The *nitrogen* level has a significant positive effect on yield ($p < 0.01$). Higher levels of nitrogen are associated with an increase in yield. $locJackson : I(nitro^2)$ and $locKnoxville : I(nitro^2)$ both interactions of the quadratic term of nitro with locations are significant ($p < 0.01$). This suggests that the effect of increasing nitro levels on yield varies by location and is nonlinear.

In Table: 10 Both *loc* and $loc : I(nitro^2)$ are highly significant predictors of *yield*.

3.3 Post-hoc Analysis

Contrast	Estimate	SE	df	t.ratio	p.value
Jackson - Knoxville	-24214	5937	8	-4.078	0.0035

Table 11: Contrast Analysis Results

After we back-transformed EMM_s

Location	EMM	SE	df	Lower CL	Upper CL
Jackson	68.00970	33.02035	2.452388	57.99423	76.09372
Knoxville	77.75062	33.02035	2.452388	69.98148	84.44259

Table 12: Estimated Marginal Means for Location

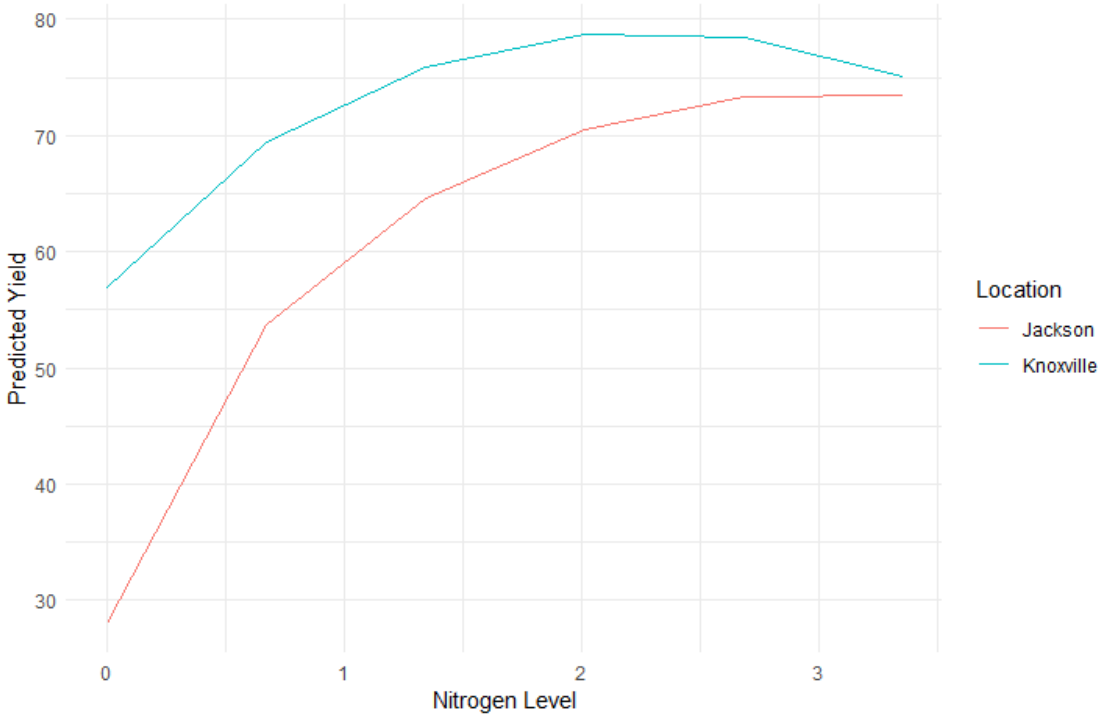

Figure 17

The post-hoc analysis indicates that *location* is a significant factor in predicting yield. Knoxville has a higher estimated yield compared to Jackson. The difference in yield between the two locations is statistically significant, as evidenced by the non-overlapping confidence intervals. The plot further supports this finding, showing that at any given level of nitrogen, the predicted yield is higher in Knoxville than in Jackson. Lastly, with higher amounts applied of nitrogen fertilizer, the yield increases.

4 Conclusion

Firstly, the estimated marginal means (EMM_s) from the post-hoc analysis indicate a statistically significant difference in yield between Jackson and Knoxville, with Knoxville having a higher yield than Jackson. This is supported by the contrast estimate, which shows a significant negative value when comparing Jackson to Knoxville, with a low p-value of 0.0035, suggesting strong evidence against the null hypothesis of no difference between the two locations.
The fixed effects within the model reveal that both $locKnoxville$ and $nitro$ are significant predictors of yield, with p-values well below the 0.01 threshold, indicating a strong association with the yield. Additionally, the interaction terms $locJackson : I(nitro^2)$ and $locKnoxville : I(nitro^2)$ are both significant, which implies that the effect of the quadratic term of nitrogen on yield varies by location, and this relationship is not merely additive but more complex.
The ANOVA summary provides further insights into the importance of these terms. The loc term has an F-value of 36.368 with a highly significant p-value, confirming that location is a major factor influencing yield. The $nitro$ term and the interaction $loc : I(nitro^2)$ also have significant F-values, underscoring the importance of nitrogen level and its non-linear relationship with yield across different locations.
The Q-Q plots for random effects suggest that the random effects included in the model for $year : loc$ combinations fit the normality assumption reasonably well. This is indicated by the points closely following the reference line in the middle range, though with some deviation at the tails, which is typical in many datasets and not necessarily a concern unless extreme.
The predictive plot showing the relationship between yield and nitrogen levels for Jackson and Knoxville depicts a non-linear response to nitrogen application, with yield increasing up to a certain point before plateauing or slightly decreasing, suggesting diminishing returns at higher nitrogen levels.

5 Appendix A

```
# Libraries
library(ggplot2)
library(lmerTest)
library(emmeans)
library(dplyr)
library(MASS)
library(car)
library(multcomp)

# Reading in data:
assignment4 <- read.table("assignment4.txt", header = TRUE, sep = "\t")
str(assignment4)
head(assignment4)

# Converting variables to factors
assignment4$loc <- as.factor(assignment4$loc)
assignment4$year <- as.factor(assignment4$year)
assignment4$nitro <- as.factor(assignment4$nitro)
assignment4$nitro <- as.numeric(as.character(assignment4$nitro))

# Check if Factors are Balanced
assignment4 %>%
  group_by(loc, year, nitro) %>%
  summarise(count = n()) %>%
  ungroup() %>%
  arrange(count)


# ---------------------------------------------------------------------------


#---------Diagrams and Plots--------#

par(mfrow=c(1,1))

# Boxplots for each categorical variable
for (var_name in c('loc', 'year', 'nitro')) {
  k <- length(levels(as.factor(assignment4[,var_name]))) + 1
  boxplot(as.formula(paste("yield ~", var_name)),
          col = 2:k, main = paste("Yield vs.", var_name),
          data = assignment4)
```

```
}

# Histogram for Yield
f <- function(x) {
  dnorm(x, mean = mean(assignment4$yield), sd = sd(assignment4$yield)
     )
}
hist(assignment4$yield, xlab='Yield', probability=T, main = "
   Distribution of Yield")
curve(f, from = min(assignment4$yield), to = max(assignment4$yield),
   lwd=3, col="red", add=T)

par(mfrow=c(1,1))

#
   ----------------------------------------------------------------------------



# Fixed effects and interactions
model1 <- lmer(yield ~ loc + nitro + loc:nitro + I(nitro^2):loc
# Random effects
+ (1 + I(nitro^2) + nitro | year:loc)
+ (1 + nitro | year:loc)
+ (1 | year:loc)
+ (1 | year:nitro)
+ (1 + nitro | year)
+ (1 | year),
  data = assignment4)

print(summary(model1), corr = FALSE)
ranova(model1)
anova(model1)


model2 <- update(model1, ~. -(1+I(nitro^2)+nitro|year:loc))
ranova(model2)
anova(model1,model2)
0.5*(1-pchisq(2*(logLik(model1)-logLik(model2)),1))
#p = 0.001318 ** we won't drop (1+I(nitro^2)+nitro|year:loc), 'log
   Lik.' 1.747123e-06

model2 <- update(model1, ~. -(I(nitro^2):loc))
ranova(model2)
anova(model1,model2)
0.5*(1-pchisq(2*(logLik(model1)-logLik(model2)),1))
```

```
#p = 0.0003031 *** we won't drop (I(nitro^2):loc), 'log Lik.' 2.67998
   5e-06

model3 <- update(model1, ~. -(loc:nitro))
ranova(model3)
anova(model1,model3)
0.5*(1-pchisq(2*(logLik(model1)-logLik(model3)),1))
#p = 0.1445 we drop (loc:nitro), 'log Lik.' 0.002328155

model4 <- update(model3, ~. - (1 + nitro | year:loc))
ranova(model4)
anova(model3, model4)
0.5*(1-pchisq(2*(logLik(model3)-logLik(model4)),1))
#p = 1 so we drop (1 + nitro | year:loc), 'log Lik.' 0.4469559

model5 <- update(model4, ~. -(1 | year:nitro))
ranova(model5)
anova(model4,model5)
0.5*(1-pchisq(2*(logLik(model4)-logLik(model5)),1))
#p = 1 so we drop (1 | year:nitro), 'log Lik.' 0.5

model6 <- update(model5, ~. -(1 + nitro | year))
ranova(model6)
anova(model5,model6)
0.5*(1-pchisq(2*(logLik(model5)-logLik(model6)),1))
#p = 1 so we drop (1 + nitro | year), 'log Lik.' 0.4411685

model7 <- update(model6, ~. -(1 | year:loc))
ranova(model7)
anova(model6,model7)
0.5*(1-pchisq(2*(logLik(model6)-logLik(model7)),1))
#p = 1 so we drop (1 | year:loc), 'log Lik.' 0.5

model8 <- update(model7, ~. -(1 | year))
ranova(model8)
anova(model7,model8)
0.5*(1-pchisq(2*(logLik(model7)-logLik(model8)),1))
#p = 1 so we drop (1 | year), 'log Lik.' 0.4998249


#
   ------------------------------------------------------------------------

#Model Diagnostics #1

```

```
model1 <- lmer(yield ~ loc + nitro + I(nitro^2):loc + (1 + I(nitro^2)
    + nitro | year:loc), data = assignment4)

model.lm <- lm(yield ~ loc + nitro + I(nitro^2):loc + nitro:year:loc
   + I(nitro^2):year:loc, data = assignment4)
par(mfrow=c(1,1))
plot(model.lm, which=1:4)
par(mfrow=c(1,1))
par(mfrow=c(1,1))
stdresid = rstandard(model.lm)
with(assignment4, plot(stdresid ~ nitro,
                       xlab = "Nitrogen Level",
                       ylab = "Standardized Residuals",
                       main = "Standardized Residuals vs Nitrogen
                          Levels"))
with(assignment4, plot(stdresid ~ loc,
                       xlab = "Location",
                       ylab = "Standardized Residuals",
                       main = "Standardized Residuals vs Location"))
with(assignment4, plot(stdresid ~ year,
                       xlab = "Year",
                       ylab = "Standardized Residuals",
                       main = "Standardized Residuals vs Year"))
par(mfrow=c(1,1))

boxcox(model.lm)
par(mfrow = c(1, 1))

#Model Diagnostics #2

model2.lm <- lm((yield^2.6) ~ loc + nitro + I(nitro^2):loc + nitro:
   year:loc + I(nitro^2):year:loc, data = assignment4)
par(mfrow=c(1,1))
plot(model2.lm,which=1:4)
par(mfrow=c(1,1))
boxcox(model2.lm)
par(mfrow = c(1, 1))

model.lmer <- lmer((yield^2.6) ~ loc + nitro + I(nitro^2):loc + (1 +
   I(nitro^2) + nitro | year:loc), data = assignment4)
par(mfrow=c(1,1))
# QQ plot for the random intercepts and slopes for nitro at each year
   :loc combination
qqnorm(unlist(ranef(model.lmer)$`year:loc`), main="Random Effects:
   year:loc")
qqline(unlist(ranef(model.lmer)$`year:loc`), col="red")
```

```
#
    ------------------------------------------------------------------------

#Results

model <- lmer((yield^2.6) ~ loc + nitro + I(nitro^2):loc + (1 + I(
    nitro^2) + nitro | year:loc), data = assignment4)
summary(model)
ranova(model)
anova(model)
#
    ------------------------------------------------------------------------



#Post-hoc

# Conduct post-hoc analysis for the fixed effect 'loc'
emm_loc <- emmeans(model, specs = pairwise ~ loc)
# Summary of estimated marginal means for 'loc'
print(summary(emm_loc))
# Pairwise comparisons of estimated marginal means for 'loc' with
    Tukey adjustment
pairwise_loc <- pairs(emm_loc, adjust = "tukey")
print(pairwise_loc)

back_transformed_emm_loc <- summary(emm_loc)$emmeans^(1/2.6)
# Print the back-transformed EMMs
print(back_transformed_emm_loc)
# Plot the back-transformed EMMs
plot(emm_loc, xlab = 'Location', ylab = 'Back-transformed Estimated
    Marginal Means')

# Here, we use the entire range of 'nitro' values present in the
    original dataset
nitro_vals <- unique(assignment4$nitro)
locs <- levels(assignment4$loc)
years <- levels(assignment4$year)
new_data <- expand.grid(nitro = nitro_vals, loc = locs, year = years)

# Predict the yield on the transformed scale
new_data$Yield_pred_transformed <- predict(model, newdata = new_data,
     re.form = NA)

# Back-transform the predictions to the original yield scale
new_data$Yield_pred <- new_data$Yield_pred_transformed^(1/2.6)

```

```
199 # Plot the predictions
200 ggplot(new_data, aes(x = nitro, y = Yield_pred, color = loc)) +
201   geom_line() +
202   labs(x = "Nitrogen Level", y = "Predicted Yield", color = "Location
        ") +
203   theme_minimal()
```