STAT 2132
Orly Olbum

Homework 7

1. Let $p_1, \dots, p_m$ be p-values for $H_{0,1}, \dots, H_{0,m}$
$p_j \sim U[0,1]$ for true $H_{0,j}$
$\pi_0 \in [0,1]$ is the fraction of m hypotheses that are true

ⓐ Show that Bonferroni procedure controls FWER regardless of dependence between $p_1, \dots, p_m$

For $\alpha \in (0,1)$,

$P(\exists j = 1, \dots, m$ such that $H_{0,j}$ is true and $p_j \le \alpha/m) = \pi_0 \alpha$

If we have m total null hypotheses, and m_0 true H_0's, then $\pi_0 = \frac{m_0}{m}$

FWER = prob of rejecting ≥ 1 true null hypothesis

$$= Pr\{\bigcup_{i=1}^{m_0} (p_i \le \frac{\alpha}{m})\}$$
$$\le \sum_{i=1}^{m_0} \{Pr(p_i \le \frac{\alpha}{m})\}$$
$$= m_0 \cdot \frac{\alpha}{m} = \pi_0 \alpha$$

Additionally, this control (Bonferroni) does not require p-values to be assumed independent | dependent for the equality to hold

ⓑ If $p_1, \dots, p_m$ are independent, show that Bonferroni procedure provides exact control of FWER for small α.
ie, for $m \ge 1$ and $\alpha \in (0,1)$,

$P(\exists j = 1, \dots, m$ such that $H_{0,j}$ is true and $p_j \le \frac{\alpha}{m}) = \pi_0 \alpha\{1 + o(1)\}$ as $\alpha \to 0$

As $\alpha \to 0$, $o(1) \to 0$ (like penalty for $\alpha \to 1$)

Say we are testing two hypotheses $\to m_0 = 2$, and

$$Pr\{\bigcup_{i=1}^{2} (p_i \le \frac{\alpha}{m})\} = m_0 \cdot \frac{\alpha}{m} - Pr\{(P_1 \le \frac{\alpha}{m}) \cap (P_2 \le \frac{\alpha}{m})\}$$

by properties of union | intersection

$$= 2 \cdot \frac{\alpha}{m} - Pr(P_1 \le \frac{\alpha}{m}) Pr(P_2 \le \frac{\alpha}{m} | P_1 \le \frac{\alpha}{m})$$
$$= 2 \cdot \frac{\alpha}{m} [1 - Pr(P_2 \le \frac{\alpha}{m} | P_1 \le \frac{\alpha}{m})]$$
$$= \pi_0 \alpha (1 - Pr(P_2 \le \frac{\alpha}{m} | p_1 \le \frac{\alpha}{m}))$$

→

1, continued

where the last term will approach 0 if the p-values are uncorrelated and 1 if they are correlated, which leads to the form $\pi_0 \alpha (1 - o(1))$.

ⓒ Based on the above results and the properties of conditional probability for independent events, (b) will not be necessarily true for dependend $p_1, \dots, p_m$. By inclusion-exclusion principal, we can expand beyond two tests.

2. $Y \in \mathbb{R}^n$ is a random vector with $E(Y) = 0$, $var(Y) = \Sigma(\theta)$ $\in \mathbb{R}^{n \times n}$ where $\theta \in \mathbb{R}^p$ is an unknown parameter. Assume $\Sigma(\theta)$ is continuously differentiable as a function of θ

→ there exist continuous matrix functions $M_j(\theta) \in \mathbb{R}^{n \times n}$ for $j = 1, \dots, p$ such that

$$\Sigma(\theta + \delta) - \Sigma(\theta) = \sum_{j=1}^{p} \delta_j M_j(\theta) + o(\|\delta\|_2)$$

→ Let $l(\theta) = -\frac{1}{2} \log[\det\{\Sigma(\theta)\}] - \frac{1}{2} Y^T \{\Sigma(\theta)\}^{-1} Y$ be log-likelihood for normal distribution.

→ No assumption of Y being normally distributed.

ⓐ $V \in \mathbb{R}^{n \times n}$ is symmetric, PD matrix. For symmetric matrix $A \in \mathbb{R}^{n \times n}$, prove: (for $\epsilon > 0$, $u \in \mathbb{R}^n$)

$$\log\{\det(V + \epsilon A)\} - \log\{\det(V)\} = \epsilon \, Tr(AV^{-1}) + o(\epsilon)$$

$$u^T (V + \epsilon A)^{-1} u - u^T V^{-1} u = -\epsilon u^T V^{-1} A V^{-1} u + o(\epsilon)$$

$$\log\{\det(V + \epsilon A)\} = \log\{\prod_{j=1}^{p} (\nu_j + \epsilon a_j)\} - \log(\prod_{j=1}^{p} \nu_j)$$

where ν and a are eigenvalues of V & A, respectively

$$= \log\{\Sigma_j \nu_j + \epsilon \Sigma_j a_j\} - \log\{\Sigma_j \nu_j\}$$

$$= Tr(\log(V + \epsilon A)) - Tr(\log(V))$$

$$= \epsilon \, Tr(AV^{-1}) + o(\epsilon)$$ by log Taylor Expansion →

2, continued

$u^T(\underbrace{V+\epsilon A})^{-1}u - u^TV^{-1}u = -\epsilon u^TV^{-1}AV^{-1}u + o(\epsilon)$

$(V+\epsilon A)^{-1} = V^{-1} - (V+V\epsilon A^{-1}V)^{-1}$

$\to u^T[V^{-1} - (V+V\epsilon A^{-1}V)^{-1}]u - u^TV^{-1}u$

$= u^TV^{-1}u - u^T(V+V\epsilon A^{-1}V)^{-1}u^T - u^TV^{-1}u$

$= -\epsilon u^T(\frac{1}{\epsilon}V + VA^{-1}V)^{-1}u^T$

$= -\epsilon u^TV^{-1}AV^{-1}u^T + o(\epsilon)$

ⓑ Use part (a) to show that

$$[\nabla_\theta l(\theta)]_j = -\tfrac{1}{2}\operatorname{Tr}[M_j(\theta)\{\Sigma(\theta)\}^{-1}] + \tfrac{1}{2}Y^T\{\Sigma(\theta)\}^{-1}M_j(\theta)\{\Sigma(\theta)\}^{-1}Y$$

Conclude that a root of $\nabla_\theta l(\theta)$ is a suitable estimator for θ. Show that

$E\{\nabla_\theta l(\theta)\} = 0$, regardless of whether or not Y is normally distributed.

Second term: $u = Y$, $\Sigma(\theta) = V$, $M_j(\theta) = A$

$\Rightarrow Y^T(\Sigma(\theta) + \epsilon M_j(\theta))Y - Y^T\{\Sigma(\theta)\}^{-1}Y$

first term:

$\Rightarrow \log\{\det(\Sigma(\theta) + \epsilon M_j(\theta))\}$

$[\nabla_\theta l(\theta)]_j \Rightarrow$ take the gradient (derivative) at one index j

→

2, continued

If

$$\ell(\theta) = -\frac{1}{2}\log[\det\{\Sigma(\theta)\}] - \frac{1}{2}Y^T\{\Sigma(\theta)\}^{-1}Y$$

then our terms line up from what we proved in ⓐ

$$\nabla_\theta \ell(\theta) = \log\{\det(\Sigma(\theta) + \epsilon M_j(\theta))\}$$
$$+ Y^T(\Sigma(\theta) + \epsilon M_j(\theta))Y - Y^T\{\Sigma(\theta)\}^{-1}Y$$
$$= -\frac{1}{2}\mathrm{Tr}\{M_j(\theta)[\Sigma(\theta)]^{-1}\}$$
$$+ \frac{1}{2}Y^T\{\Sigma(\theta)\}^{-1}M_j(\theta)\{\Sigma(\theta)\}^{-1}Y$$

(from ⓐ, by Sherman-Morrison and properties of determinants and trace)

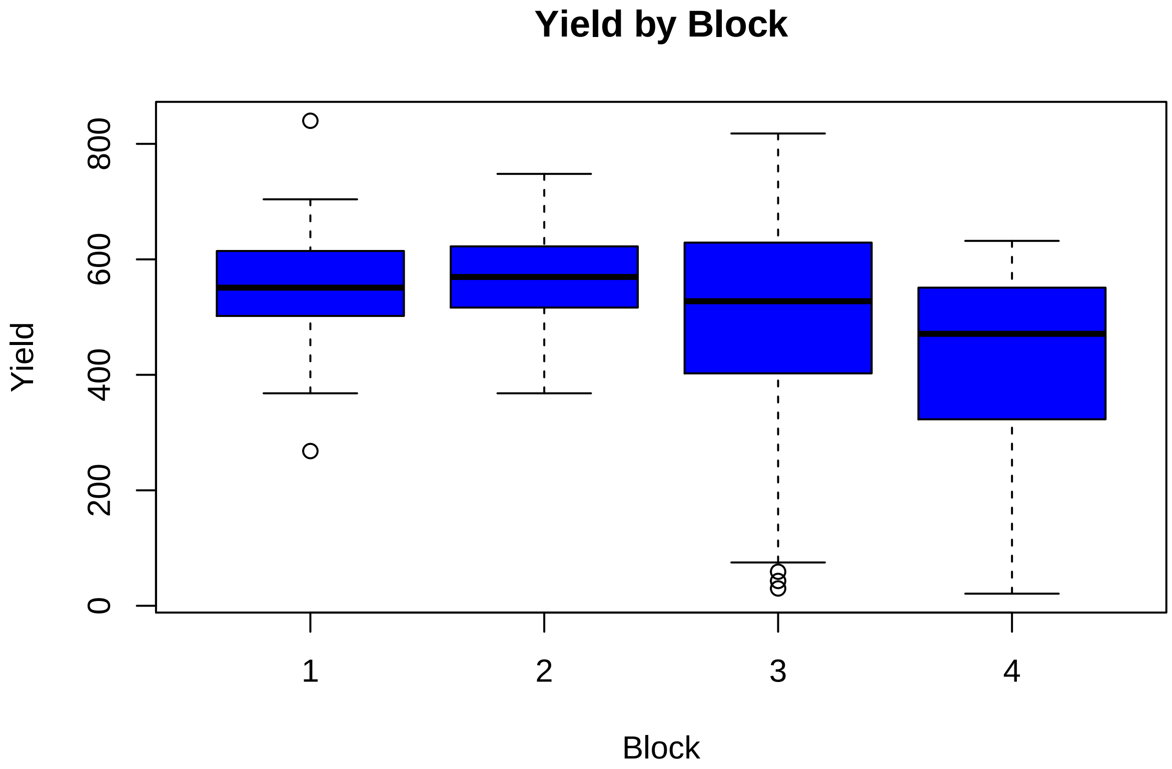
2132 HW7

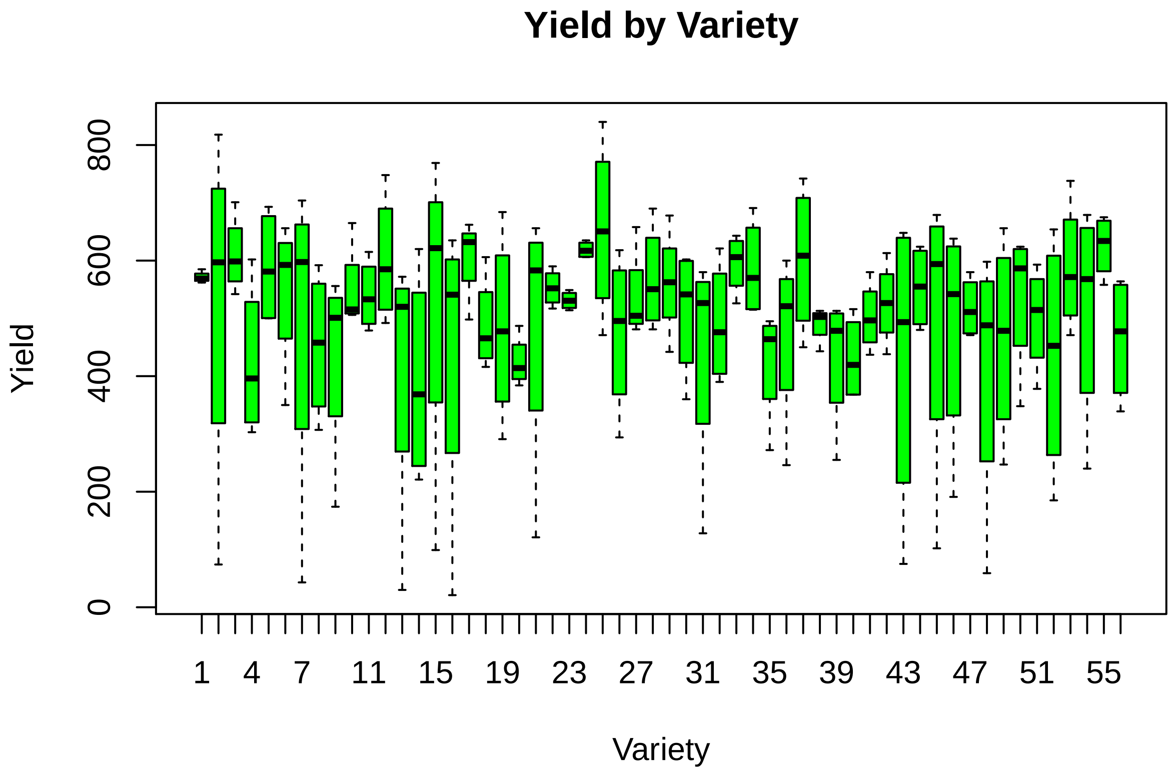
Problem 3

A study was done to compare the yields of 56 varieties of wheat in a randomized complete block design (RCBD) with four blocks of size 56. The data for the experiment are in the file "wheat56.txt". The four blocks are observations 1-56, 57-112, 113-168 and 169-224. The varieties, yields, latitudes and longitudes of each plot (latitudes and longitudes in unstated units) are given. Although the units are unstated, keep in mind that agricultural field trials like this are carried out at a single farm so that the weather is essentially the same at all plots. The labeling of the varieties as 1-56 "in order" in Block 1 is for convenience; you may assume that in fact the variety assignments were properly randomized in all four blocks.

(a)

Estimate the variety effects using the standard model for an RCBD treating blocks and varieties as fixed effects. Using appropriate tables and/or plots, summarize your findings about the differences between varieties. As part of your analysis, include an F-test for the hypothesis of no variety effects.





In the boxplot we can see that some varieties are clearly separated from the bulk of the group, and we may find a significant result when this model is run.

```
##              Df  Sum Sq Mean Sq F value   Pr(>F)    
## block         3  723630  241210  12.162 3.13e-07 ***
## Variety      55  954995   17364   0.875    0.712    
## Residuals   165 3272436   19833                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA output shows that block is a significant effect on Yield, and Variety is not, with F values of 12.16 and 0.88, respectively (with Variety treated as a factor). While the boxplot showed potentially some difference between Varieties, it was not enough to yield a significant result, and we conclude that there are no differenced between varieties as effects on yield. Could there be something else going on here?

(b)

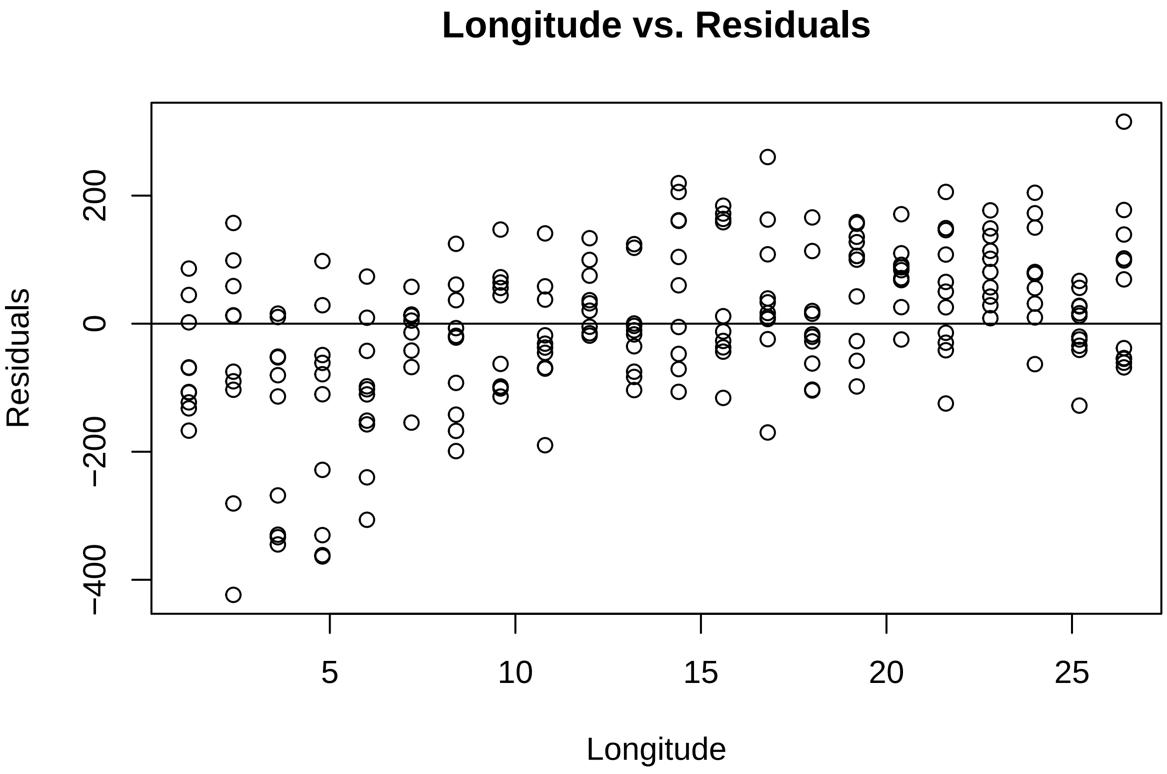
Find a 95% confidence interval for the mean yield of varieties 1-20 minus the mean yield of varieties 21-56.

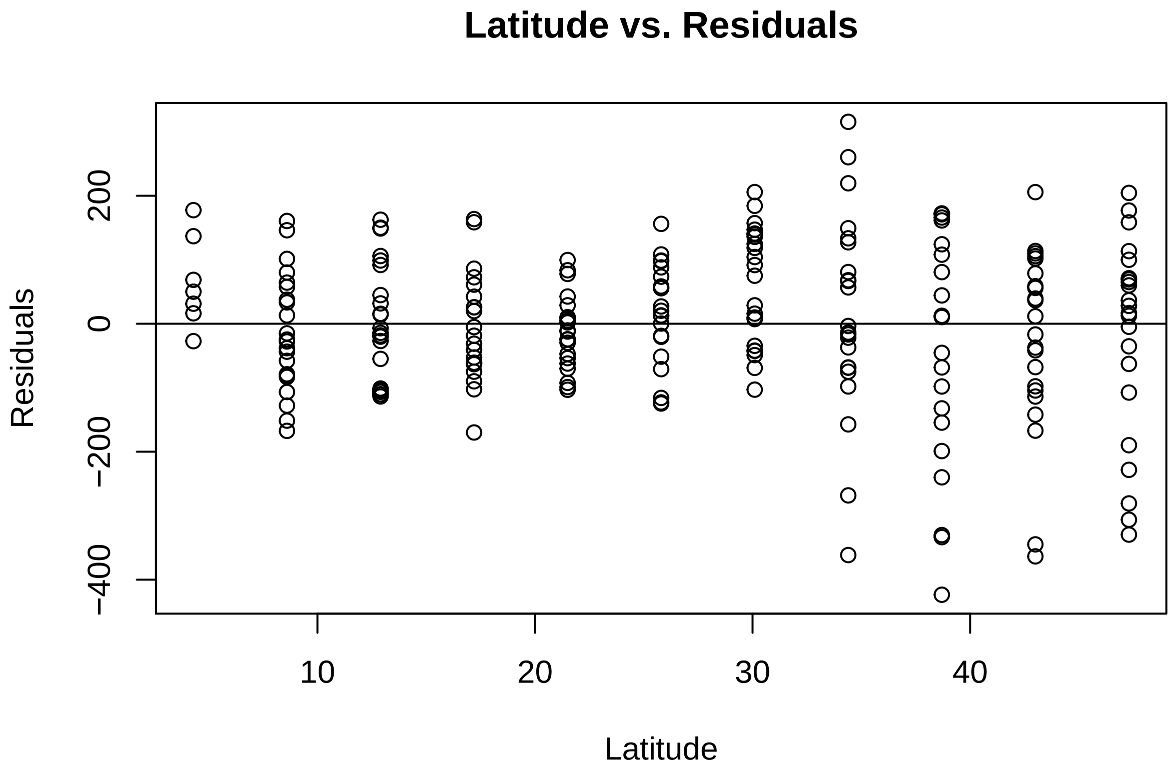
```
## [1] -87.55508  69.93008
```

The confidence interval is (-87.56, 69.93) which includes 0, which is what we would expect from the results above.

(c)

Plot the residuals as a function of the geographic coordinates of the plots. Discuss any patterns you see and comment on the reasonableness of the assumptions underlying the analyses in (a). Can you identify any varieties whose yields (relative to other varieties) might be over or underestimated because of the plots to which they were assigned? Comment.





When graphing the model residuals against longitude and latitude, we do not want to see any patterns. Latitude is somewhat scattered, but longitude does show a slight upward trend. Since they are not yet included in the model, a trend with the model residuals indicates that longitude may need to be included as a potential effect in the model. Lower longitudes seem to be underestimated while higher latitudes seem to be overestimated (i.e., the patterns).

A Tukey test of the model will show which varieties have which p-values. That is, which are contributing the results of the main F-test above.

(d)

Reanalyze the data including a linear function of the coordinates in your mean function. What effect does this change have on your inferences about variety effects? In particular, which estimated variety effects change the most from the analysis in (c)? Plot the residuals as a function of the geographic coordinates of the plots. To what extent are any problems you noted with the residual plot in (c) fixed?

```
##               Df  Sum Sq Mean Sq F value   Pr(>F)    
## block          3  723630  241210  19.285 9.44e-11 ***
## Variety       55  954995   17364   1.388   0.0595 .  
## Longitude      1 1205282 1205282  96.365  < 2e-16 ***
## Latitude       1   28438   28438   2.274   0.1335    
## Residuals    163 2038716   12507                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] 156
```

The p_vals table shows the p-values associated with the Tukey test for the first model (just block and Variety) and the second model (additionally longitude and latitude). 1's in the "diff" column indicate the same p-value, and 0's indicate a different p-value, which means a different test result from Tukey in the second model. There are 156 Variety/block combinations that are different in the second model after including longitude and latitude.

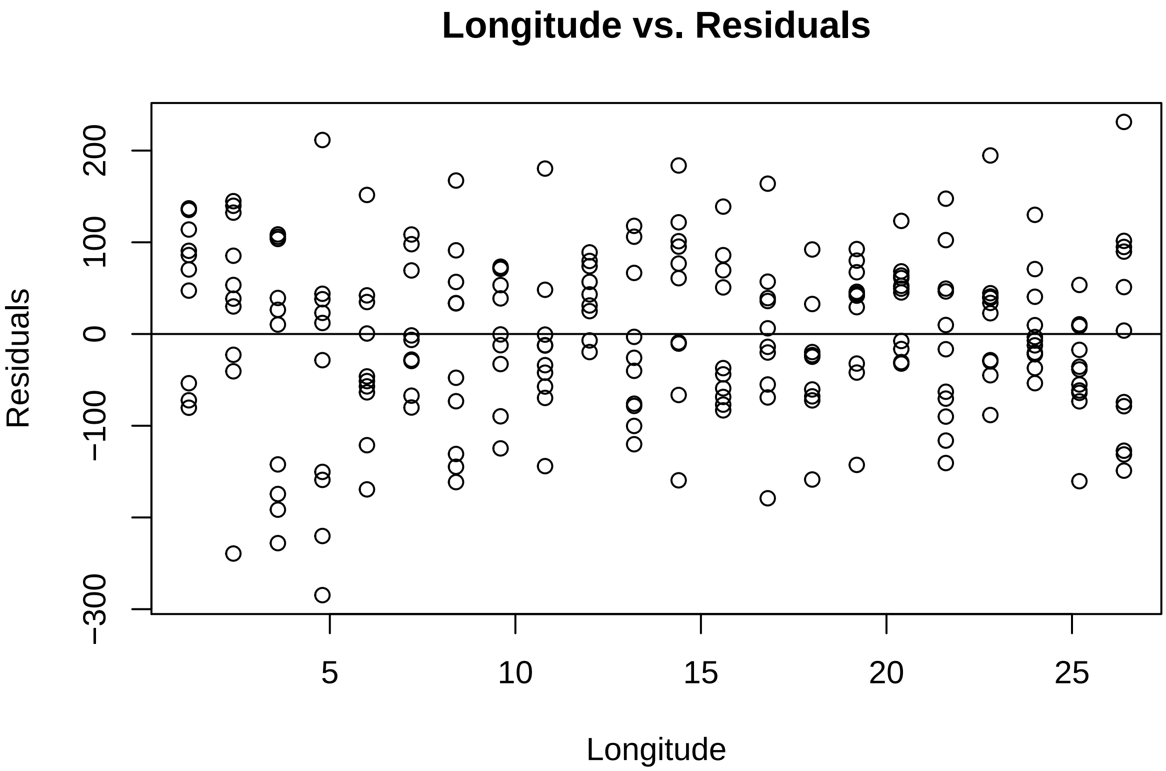
(e)

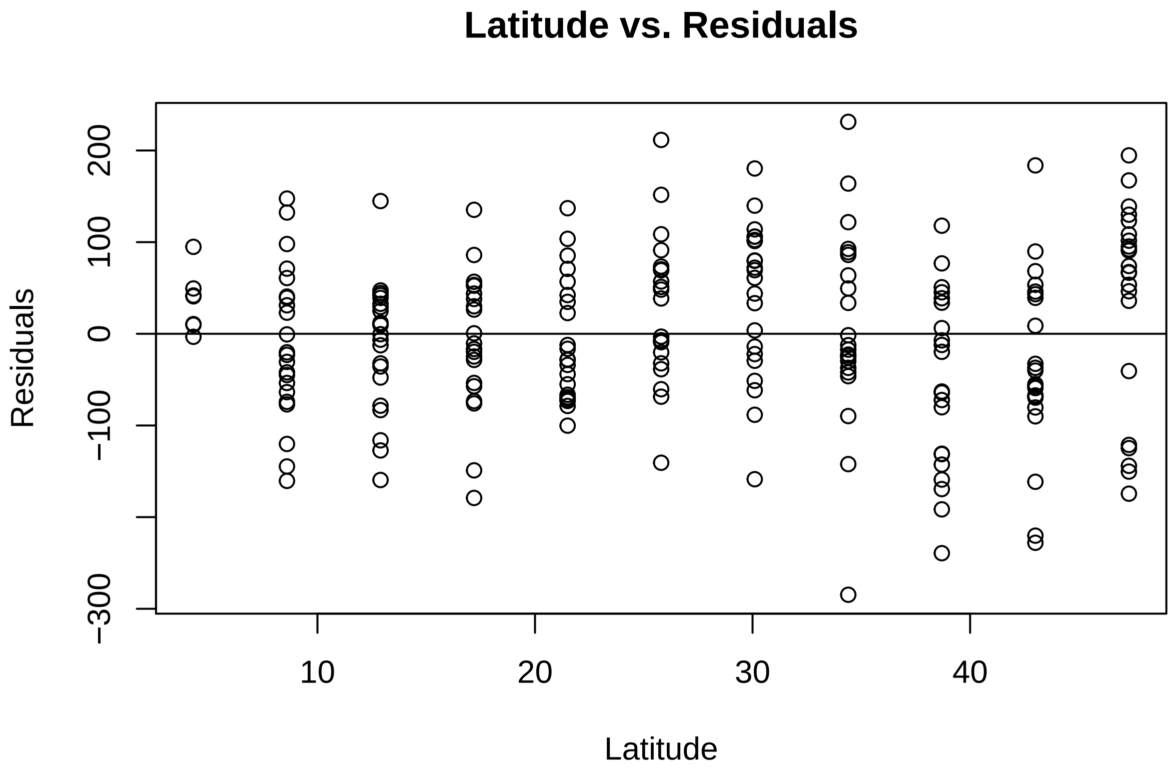
Answer the same questions as in (d), but this time including a quadratic function of the coordinates (i.e., a second order polynomial in latitude and longitude) in your mean function.

```
##                  Df  Sum Sq Mean Sq F value   Pr(>F)    
## block             3  723630  241210  21.446 9.72e-12 ***
## Variety          55  954995   17364   1.544  0.01958 *  
## Longitude         1 1205282 1205282 107.163  < 2e-16 ***
## I(Longitude^2)    1  118603  118603  10.545  0.00142 ** 
## Latitude          1   25471   25471   2.265  0.13431    
## I(Latitude^2)     1  112289  112289   9.984  0.00189 ** 
## Residuals       161 1810792   11247                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] 0
```

The new model fit with the second order terms for long/lat have led to a significant variety effect. Now, the residual plots show no apparent pattern and we can rest a bit easier about the model fit.


Longitude vs. Residuals
Residuals
200
100
0
−100
−300
5
10
15
20
25
Longitude



None of the p-values from the Tukey test in the new model (quadratic longitude and latitude terms) are different than in the second model. Additionally, squaring these covariates had no effect on the model output either.

#(f) *Do you think the design used for this study was well-chosen? Discuss any problems you see and describe how the study might have been designed differently to avoid or reduce these problems.*

If there is a way to assign one value to location as opposed to longitude and latitude, that may help account for this variable as an effect on yield. For example, setting up plots and treating it as a factor for the model. If we look at longitude and latitude separately, we may be able to pinpoint a row or column that acts differently but not a specific location. However, in the later model with the quadratic terms we did end up seeing the significant result for Variety that we expected earlier, so maybe we are okay with this model (i.e., well-chosen).

~~~
```{r setup, include = FALSE}
knitr::opts_chunk$set(echo = FALSE)
setwd("C:/Users/orlyo/OneDrive/Desktop/Grad School/Spring 2021/2. STAT 2132 - Applied Stat. Methods 2/Homeworks/HW7")

library(dvmisc)

wheat = read.csv("wheat.csv")
wheat$Variety = as.factor(wheat$Variety)
wheat$obs = seq.int(nrow(wheat))
wheat$block[wheat$obs <= 56] = 1
wheat$block[wheat$obs > 56 & wheat$obs <= 112] = 2
wheat$block[wheat$obs > 112 & wheat$obs <= 168] = 3
wheat$block[wheat$obs > 168 & wheat$obs <= 224] = 4
wheat$block = as.factor(wheat$block)
```

## Problem 3

# (a)
```{r}
boxplot(Yield ~ block, data = wheat, main = "Yield by Block", xlab = "Block", ylab = "Yield", col = "blue", border = "black")
boxplot(Yield ~ Variety, data = wheat, main = "Yield by Variety", xlab = "Variety", ylab = "Yield", col = "green", border = "black")
```

```{r, warning = FALSE}
fit = aov(Yield ~ block + Variety, data = wheat)
summary(fit)
```

# (b)
```{r, warning = FALSE}
mod = lm(Yield ~ block + Variety, data = wheat)
mu1 = mean(fit$fitted.values[as.numeric(wheat$Variety) <= 20])
mu2 = mean(fit$fitted.values[as.numeric(wheat$Variety) > 20])
t = qtukey(p = 0.95, nmeans = 2, df = 54)/sqrt(2)
n1 = 20
n2 = 36
sig.sq = sqrt(get_mse(mod)*((1/n1) + (1/n2)))
lo = (mu1 - mu2) - t*sig.sq
hi = (mu1 - mu2) + t*sig.sq
ci = c(lo, hi)
ci
```
~~~

~~~
# (c)
```{r, warning = FALSE}
plot(wheat$Longitude, fit$residuals, main = "Longitude vs. Residuals", xlab = "Longitude", ylab = "Residuals")
abline(h = 0)
plot(wheat$Latitude, fit$residuals, main = "Latitude vs. Residuals", xlab = "Latitude", ylab = "Residuals")
abline(h = 0)

tukey1 = TukeyHSD(fit)
variety1 = data.frame(tukey1$Variety[,4])
# View(variety1)
```

# (d)
```{r, warning = FALSE}
fit2 = aov(Yield ~ block + Variety + Longitude + Latitude, data = wheat)
summary(fit2)

tukey2 = TukeyHSD(fit2)
variety2 = data.frame(tukey2$Variety[,4])
# View(variety2)

p_vals = cbind(round(variety1, 3), round(variety2, 3))
colnames(p_vals) = c("p1", "p2")

p_vals$diff = with(p_vals, ifelse(p1 == p2, 1, 0))
# View(p_vals)

n = nrow(p_vals)
n - sum(p_vals$diff)
```

# (e)
```{r, warning = FALSE}
fit3 = aov(Yield ~ block + Variety + Longitude + I(Longitude^2) + Latitude + I(Latitude^2), data = wheat)
summary(fit3)

tukey3 = TukeyHSD(fit3)
variety3 = data.frame(tukey3$Variety[,4])
# View(variety3)

p_vals_update = cbind(p_vals, round(variety3, 3))
colnames(p_vals_update) = c("p1", "p2", "p3")

p_vals_update$diff = with(p_vals_update, ifelse(p3 == p3, 1, 0))
# View(p_vals_update)

m = nrow(p_vals_update)
~~~

~~~
m - sum(p_vals_update$diff)
```

```{r}
plot(wheat$Longitude, fit3$residuals, main = "Longitude vs. Residuals", xlab = "Longitude", ylab = "Residuals")
abline(h = 0)
plot(wheat$Latitude, fit3$residuals, main = "Latitude vs. Residuals", xlab = "Latitude", ylab = "Residuals")
abline(h = 0)
```
~~~