STAT 222 Spring 2021 HW5

Matthew Zhao

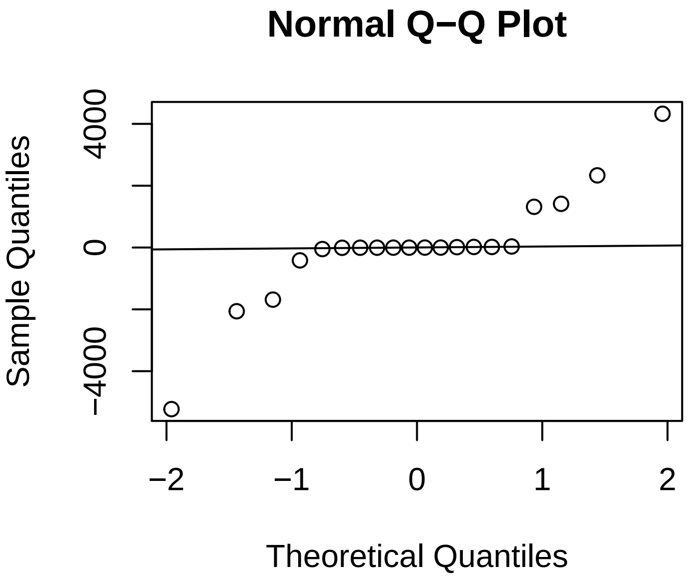
Question 1

```
insulate = read.table("http://www.stat.uchicago.edu/~yibi/s222/insulate.txt", h=T)
```

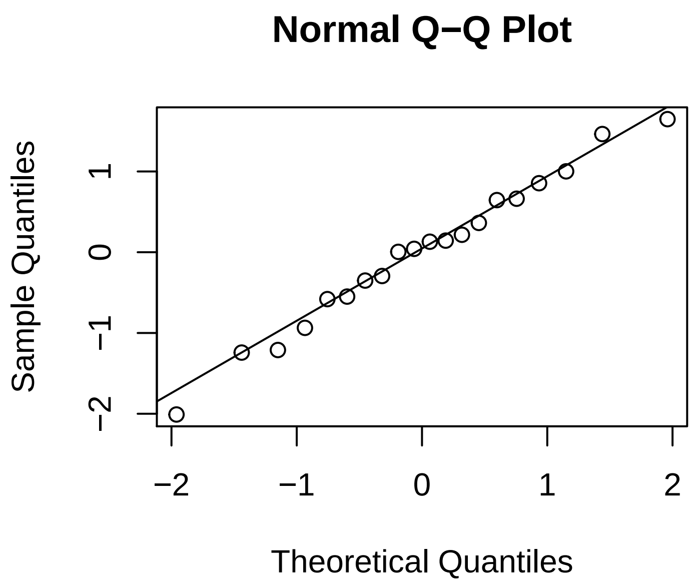
Q1a — 4 points

```
m1 = lm(failtime~as.factor(material),data=insulate)
m2 = lm(log(failtime) ~ as.factor(material), data=insulate)
```

```
qqnorm(m1$residuals)
qqline(m1$residuals)
```



```
qqnorm(m2$residuals)
qqline(m2$residuals)
```



These graphs suggest that our assumption that the data are not normally distributed is incorrect, since prior to transforming the response (failtime), the points did not fall on the line. After the log transformation of the response, the points fell on the line.

Q1b — 2 points

```
library(emmeans)
summary(emmeans(m1, "material", level=0.95))
##  material emmean   SE df lower.CL upper.CL
##         1  159.8 1021 15    -2016     2335
##         2    6.2 1021 15    -2169     2182
##         3 2941.7 1021 15      766     5117
##         4 5723.0 1021 15     3548     7898
##         5   10.8 1021 15    -2165     2186
##
## Confidence level used: 0.95
```

```
summary(emmeans(m2, "material", level=0.95))
##  material emmean    SE df lower.CL upper.CL
##         1   5.05 0.523 15    3.936     6.17
##         2   1.24 0.523 15    0.127     2.36
##         3   7.72 0.523 15    6.601     8.83
##         4   8.21 0.523 15    7.098     9.33
##         5   1.90 0.523 15    0.788     3.02
##
## Results are given on the log (not the response) scale.
## Confidence level used: 0.95
exp(summary(emmeans(m2, "material", level=0.95))[,5:6])
##     lower.CL  upper.CL
## 1   51.2128   476.846
## 2    1.1352    10.570
## 3  735.6844  6850.008
## 4 1210.0365 11266.734
## 5    2.1998    20.482
```

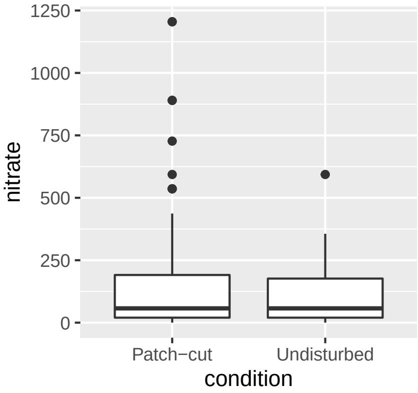
I believe that the second method is more reasonable because the m1 model violates our assumption of normal data and specifically has non-constant variance. As a result, the confidence intervals generated using this data would be incorrect since they were created assuming normal data. Since the data from the m2 model was normalized and then reverted, I would trust these confidence intervals more.

Question 2

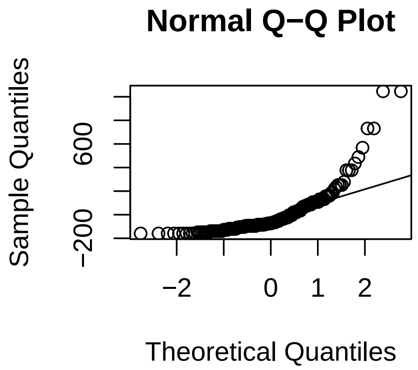
```
logging = read.table("http://www.stat.uchicago.edu/~yibi/s222/logging.txt", h=T)
```

Q2a — 2 points

```
library(ggplot2)
ggplot(logging, aes(x=condition, y=nitrate)) + geom_boxplot()
```



```
model = lm(nitrate~as.factor(condition),data=logging)
qqnorm(model$residuals)
qqline(model$residuals)
```

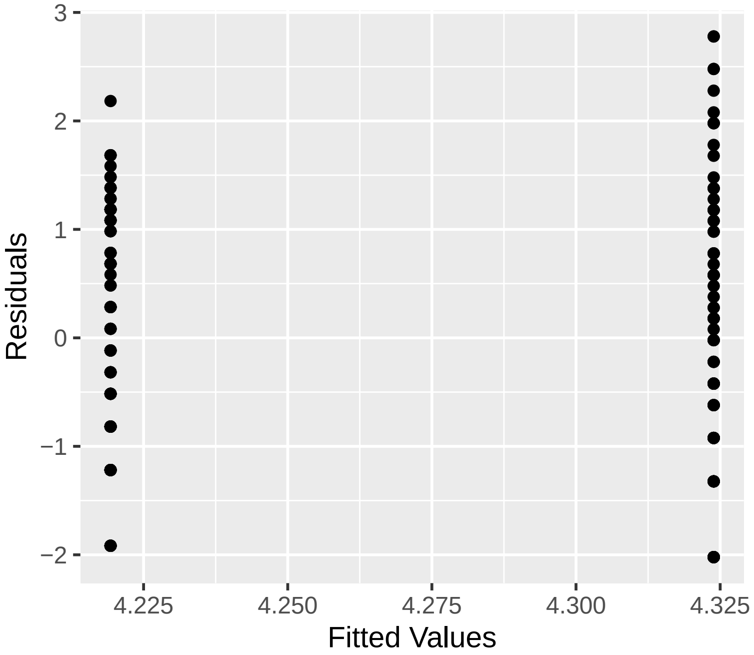


Right-skewed. This is confirmed by the boxplots which have large 75 percentiles far from the 50 percentiles.

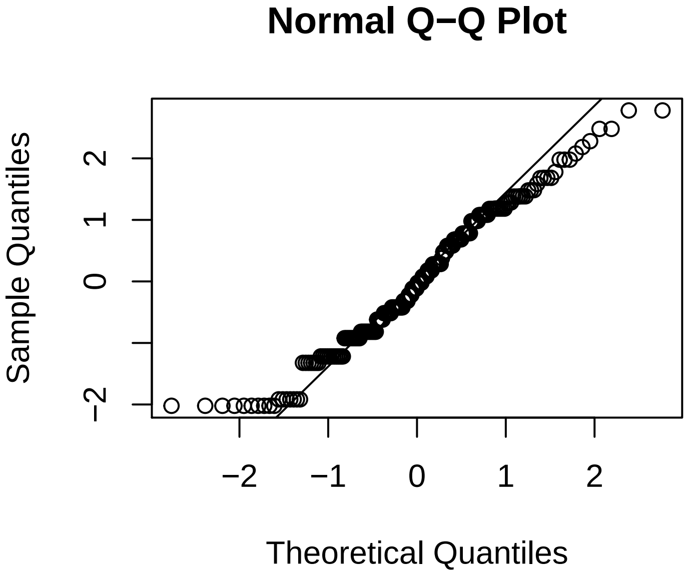
Q2b — 4 points

```
logging$lognitrate = log(10+logging$nitrate)
lmlog = lm(lognitrate ~ condition, data=logging)
```

```
ggplot(data = lmlog, aes(x = lmlog$fitted.values, y = lmlog$residuals)) +
  geom_point() + labs(x = "Fitted Values", y = "Residuals")
```



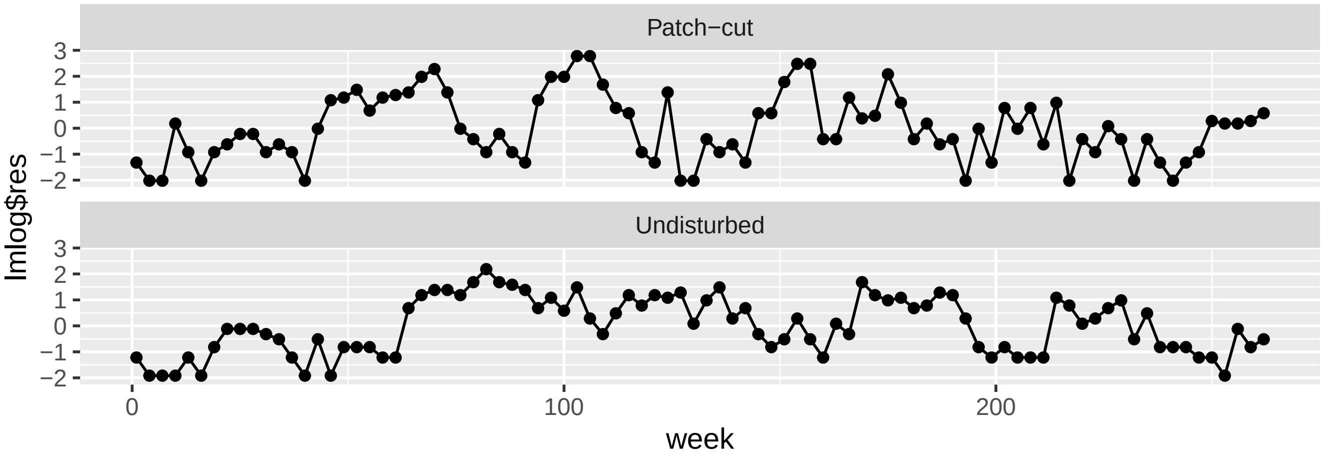
```
qqnorm(lmlog$residuals)
qqline(lmlog$residuals)
```



The plots tell us that both assumptions are false. The constant variability assumption is violated since the first group of fitted values has a lower variance than the second group of fitted values. The normality assumption is violated since the points do not fall on the line of the normal qq plot.

Q2c — 2 points

```
library(ggplot2)
ggplot(logging, aes(x=week, y=lmlog$res)) +
  geom_point() + geom_line() +
  facet_wrap(~condition, nrow=2)
```

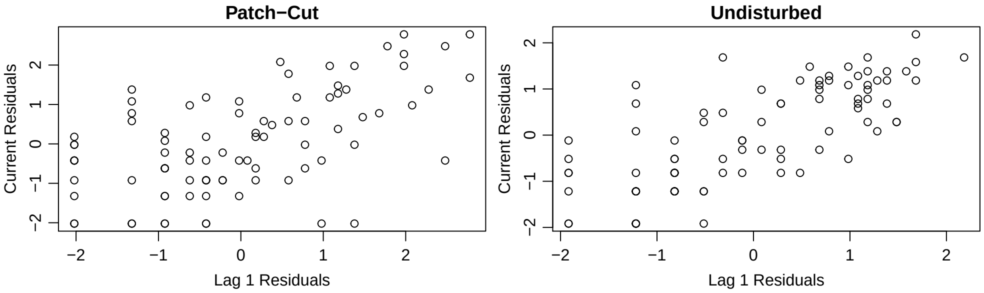


The graph for Undisturbed is fairly smooth, indicating positive autocorrelation and that there is serial dependence. The Patch-cut has smooth patches especially towards the beginning, indicating that there is also serial dependence.

Q2d — 2 points

```
res.patch = subset(lmlog$res, logging$condition == "Patch-cut")
res.undisturb = subset(lmlog$res, logging$condition == "Undisturbed")
par(mai=c(.6,.6,.3,.01),mgp=c(2,.7,0))
plot(res.patch[1:87], res.patch[2:88], ylab="Current Residuals",
      xlab="Lag 1 Residuals", main="Patch-Cut")
cor(res.patch[1:87], res.patch[2:88])
## [1] 0.57549

plot(res.undisturb[1:87], res.undisturb[2:88], ylab="Current Residuals",
      xlab="Lag 1 Residuals", main="Undisturbed")
cor(res.undisturb[1:87], res.undisturb[2:88])
## [1] 0.75092
```



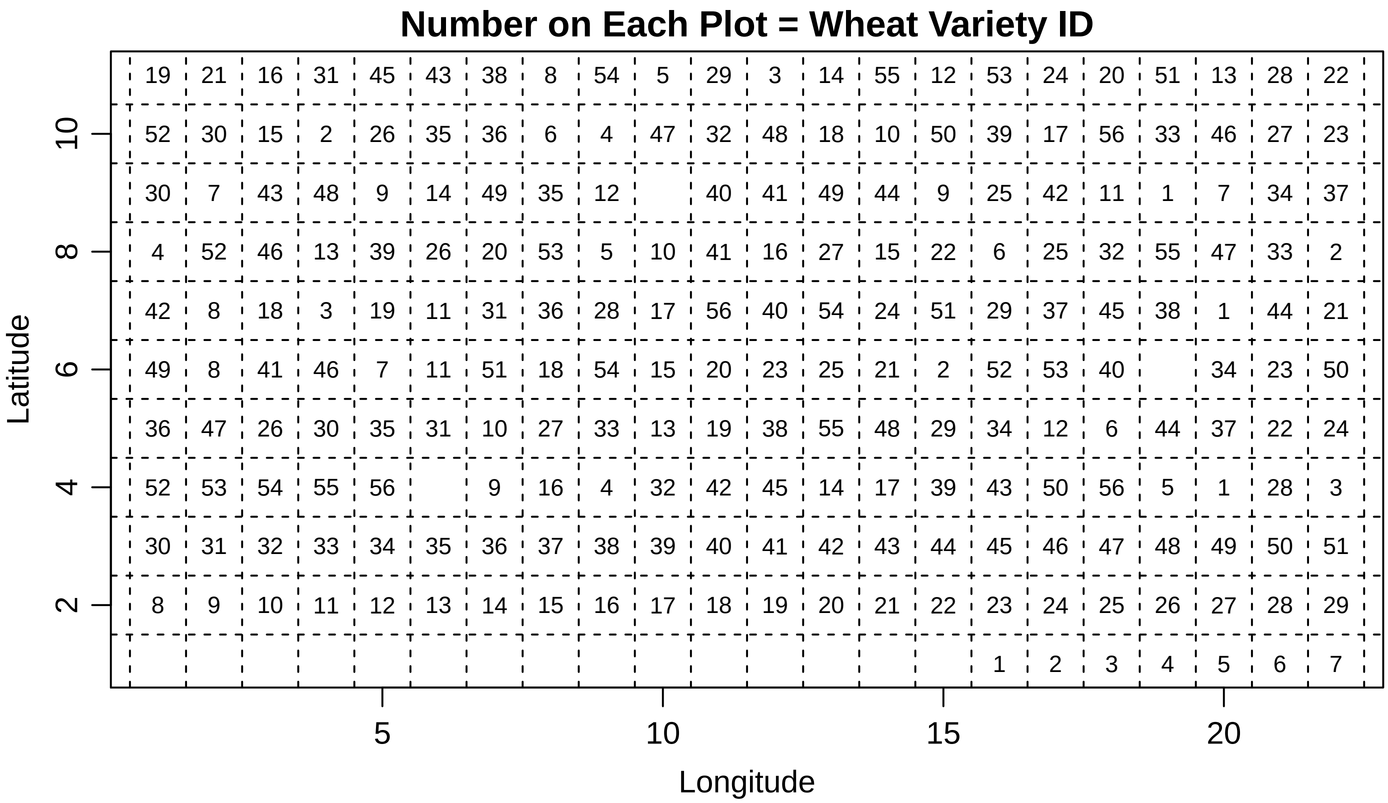
Both plots exhibit a weak positive correlation, indicating serial dependence.

Q2e — 2 points

I do not. This is because the data violates a number of our assumptions, meaning a standard ANOVA F-test is not suitable. Specifically, because the data show serial dependence (positive autocorrelation), the MSE will underestimate the actual variance of the noise. This will cause the F statistic to be artificially large and make it more likely that we will have a lower p value and reject the null i.e. the likelihood of Type I Error is higher.

Question 3

```r
wheat56 = read.table("http://www.stat.uchicago.edu/~yibi/s222/wheat56.txt", h=T)
par(mai=c(.6,.6,.3,.1),mgp=c(2,.7,0))     # reducing the margin of plot
plot(wheat56$Longitude, wheat56$Latitude, type="n",
     xlab="Longitude", ylab="Latitude", main="Number on Each Plot = Wheat Variety ID")
for(i in 0:22){abline(v=i+0.5, lty=2)}    # adding the vertical grid lines
for(j in 0:11){abline(h=j+0.5, lty=2)}    # adding the horizontal grid lines
text(wheat56$Longitude, wheat56$Latitude, labels = wheat56$Variety, cex=0.75)
```

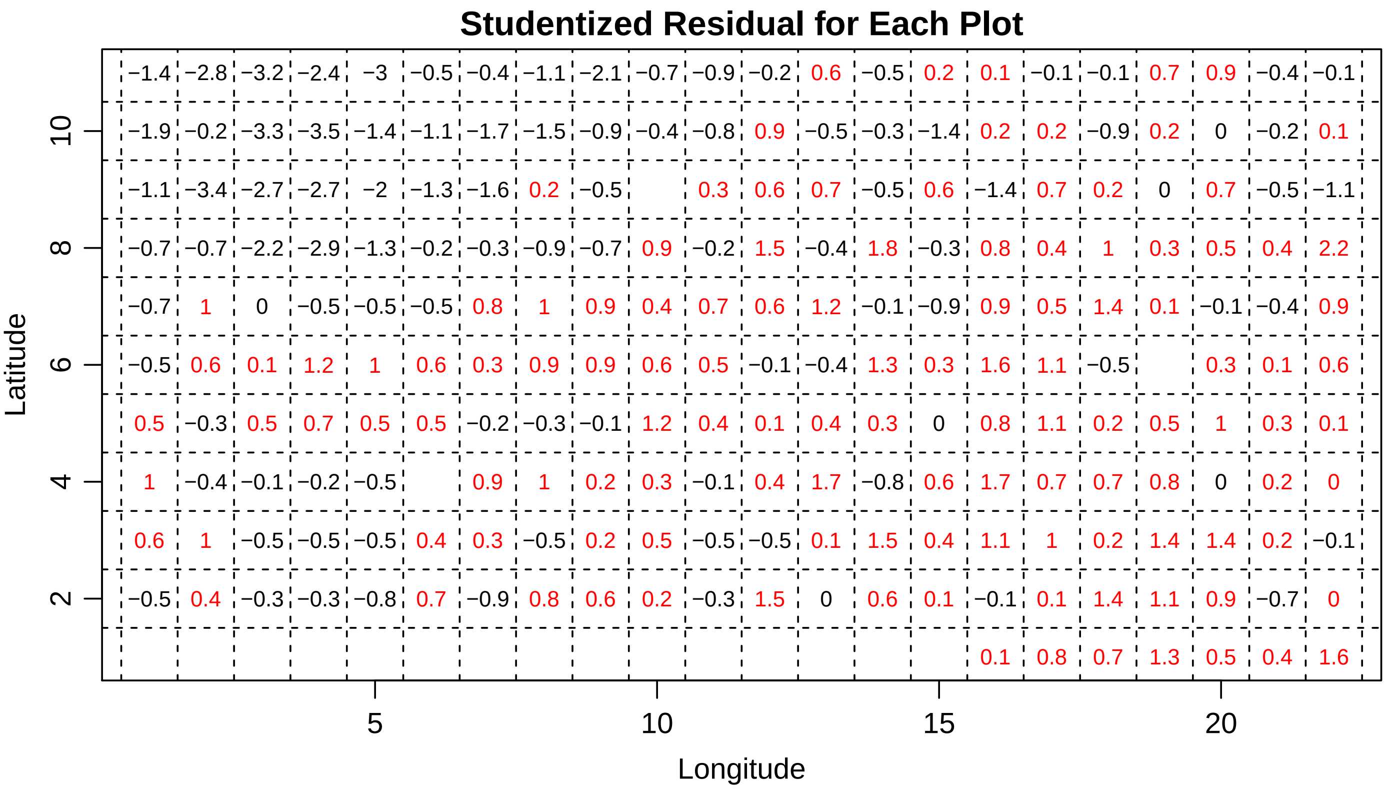


19	21	16	31	45	43	38	8	54	5	29	3	14	55	12	53	24	20	51	13	28	22
52	30	15	2	26	35	36	6	4	47	32	48	18	10	50	39	17	56	33	46	27	23
30	7	43	48	9	14	49	35	12		40	41	49	44	9	25	42	11	1	7	34	37
4	52	46	13	39	26	20	53	5	10	41	16	27	15	22	6	25	32	55	47	33	2
42	8	18	3	19	11	31	36	28	17	56	40	54	24	51	29	37	45	38	1	44	21
49	8	41	46	7	11	51	18	54	15	20	23	25	21	2	52	53	40		34	23	50
36	47	26	30	35	31	10	27	33	13	19	38	55	48	29	34	12	6	44	37	22	24
52	53	54	55	56		9	16	4	32	42	45	14	17	39	43	50	56	5	1	28	3
30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51
8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
															1	2	3	4	5	6	7

```r
lmwheat = lm(Yield ~ as.factor(Variety), data = wheat56)
```

Q3a — 2 points

```r
par(mai=c(.6,.6,.3,.1),mgp=c(2,.7,0)) # reducing the margin of plot
plot(wheat56$Longitude, wheat56$Latitude, type="n",
     xlab="Longitude", ylab="Latitude", main="Studentized Residual for Each Plot")
for(i in 0:22){abline(v=i+0.5, lty=2)} # adding the vertical grid lines
for(j in 0:11){abline(h=j+0.5, lty=2)} # adding the horizontal grid lines
text(wheat56$Longitude, wheat56$Latitude,
     labels = round(rstudent(lmwheat),1), cex=0.75,
     col = 1+(rstudent(lmwheat)>0))
```



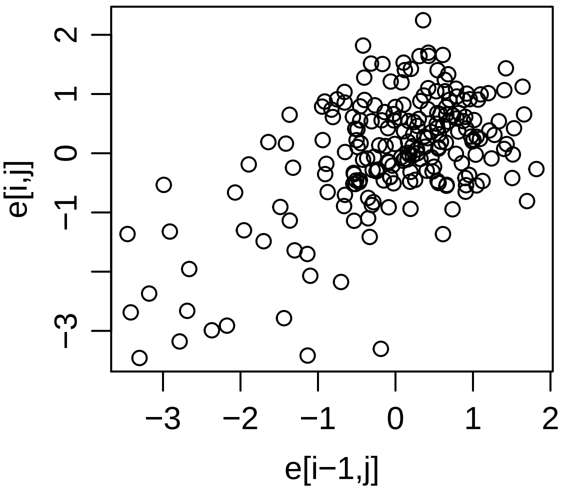
−1.4	−2.8	−3.2	−2.4	−3	−0.5	−0.4	−1.1	−2.1	−0.7	−0.9	−0.2	0.6	−0.5	0.2	0.1	−0.1	−0.1	0.7	0.9	−0.4	−0.1
−1.9	−0.2	−3.3	−3.5	−1.4	−1.1	−1.7	−1.5	−0.9	−0.4	−0.8	0.9	−0.5	−0.3	−1.4	0.2	0.2	−0.9	0.2	0	−0.2	0.1
−1.1	−3.4	−2.7	−2.7	−2	−1.3	−1.6	0.2	−0.5		0.3	0.6	0.7	−0.5	0.6	−1.4	0.7	0.2	0	0.7	−0.5	−1.1
−0.7	−0.7	−2.2	−2.9	−1.3	−0.2	−0.3	−0.9	−0.7	0.9	−0.2	1.5	−0.4	1.8	−0.3	0.8	0.4	1	0.3	0.5	0.4	2.2
−0.7	1	0	−0.5	−0.5	−0.5	0.8	1	0.9	0.4	0.7	0.6	1.2	−0.1	−0.9	0.9	0.5	1.4	0.1	−0.1	−0.4	0.9
−0.5	0.6	0.1	1.2	1	0.6	0.3	0.9	0.9	0.6	0.5	−0.1	−0.4	1.3	0.3	1.6	1.1	−0.5		0.3	0.1	0.6
0.5	−0.3	0.5	0.7	0.5	0.5	−0.2	−0.3	−0.1	1.2	0.4	0.1	0.4	0.3	0	0.8	1.1	0.2	0.5	1	0.3	0.1
1	−0.4	−0.1	−0.2	−0.5		0.9	1	0.2	0.3	−0.1	0.4	1.7	−0.8	0.6	1.7	0.7	0.7	0.8	0	0.2	0
0.6	1	−0.5	−0.5	−0.5	0.4	0.3	−0.5	0.2	0.5	−0.5	−0.5	0.1	1.5	0.4	1.1	1	0.2	1.4	1.4	0.2	−0.1
−0.5	0.4	−0.3	−0.3	−0.8	0.7	−0.9	0.8	0.6	0.2	−0.3	1.5	0	0.6	0.1	−0.1	0.1	1.4	1.1	0.9	−0.7	0
															0.1	0.8	0.7	1.3	0.5	0.4	1.6

There is a spatial pattern of the residuals. Specifically, we can see a clustering of positive (black) residuals in the upper left and bottom left with the negative (red) residuals primarily in the middle and bottom right. Additionally, the residuals are highly correlated with their neighboring plots, with black plots and red plots near others of the same color.

Q3b — 1 point

```
res.array = array(NA, dim = c(22,11))
for(i in 1:22){
  for(j in 1:11){
    if(sum(wheat56$Longitude == i & wheat56$Latitude == j)>0){
      res.array[i,j] = rstudent(lmwheat)[wheat56$Longitude == i & wheat56$Latitude == j]
    }
  }
}
round(res.array, digits=1)
##        [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
##   [1,]   NA -0.5  0.6  1.0  0.5 -0.5 -0.7 -0.7 -1.1  -1.9  -1.4
##   [2,]   NA  0.4  1.0 -0.4 -0.3  0.6  1.0 -0.7 -3.4  -0.2  -2.8
##   [3,]   NA -0.3 -0.5 -0.1  0.5  0.1  0.0 -2.2 -2.7  -3.3  -3.2
##   [4,]   NA -0.3 -0.5 -0.2  0.7  1.2 -0.5 -2.9 -2.7  -3.5  -2.4
##   [5,]   NA -0.8 -0.5 -0.5  0.5  1.0 -0.5 -1.3 -2.0  -1.4  -3.0
##   [6,]   NA  0.7  0.4   NA  0.5  0.6 -0.5 -0.2 -1.3  -1.1  -0.5
##   [7,]   NA -0.9  0.3  0.9 -0.2  0.3  0.8 -0.3 -1.6  -1.7  -0.4
##   [8,]   NA  0.8 -0.5  1.0 -0.3  0.9  1.0 -0.9  0.2  -1.5  -1.1
##   [9,]   NA  0.6  0.2  0.2 -0.1  0.9  0.9 -0.7 -0.5  -0.9  -2.1
##  [10,]   NA  0.2  0.5  0.3  1.2  0.6  0.4  0.9   NA  -0.4  -0.7
## [11,]   NA -0.3 -0.5 -0.1  0.4  0.5  0.7 -0.2  0.3  -0.8  -0.9
## [12,]   NA  1.5 -0.5  0.4  0.1 -0.1  0.6  1.5  0.6   0.9  -0.2
## [13,]   NA  0.0  0.1  1.7  0.4 -0.4  1.2 -0.4  0.7  -0.5   0.6
## [14,]   NA  0.6  1.5 -0.8  0.3  1.3 -0.1  1.8 -0.5  -0.3  -0.5
## [15,]   NA  0.1  0.4  0.6  0.0  0.3 -0.9 -0.3  0.6  -1.4   0.2
## [16,]  0.1 -0.1  1.1  1.7  0.8  1.6  0.9  0.8 -1.4   0.2   0.1
## [17,]  0.8  0.1  1.0  0.7  1.1  1.1  0.5  0.4  0.7   0.2  -0.1
## [18,]  0.7  1.4  0.2  0.7  0.2 -0.5  1.4  1.0  0.2  -0.9  -0.1
## [19,]  1.3  1.1  1.4  0.8  0.5   NA  0.1  0.3  0.0   0.2   0.7
## [20,]  0.5  0.9  1.4  0.0  1.0  0.3 -0.1  0.5  0.7   0.0   0.9
## [21,]  0.4 -0.7  0.2  0.2  0.3  0.1 -0.4  0.4 -0.5  -0.2  -0.4
## [22,]  1.6  0.0 -0.1  0.0  0.1  0.6  0.9  2.2 -1.1   0.1  -0.1
```

```
par(mai=c(.6,.6,.1,.1),mgp=c(2,.7,0))    # for reducing the margin of plots
plot(as.vector(res.array[1:21,1:11]),as.vector(res.array[2:22,1:11]),
     xlab="e[i-1,j]", ylab="e[i,j]")
```

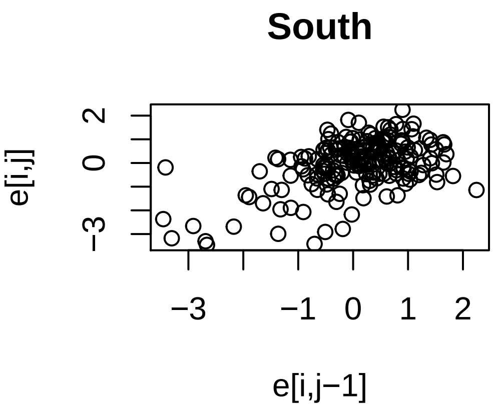


```
cor(as.vector(res.array[1:21,1:11]),as.vector(res.array[2:22,1:11]),
    use="complete.obs")
## [1] 0.55837
```

There is strong evidence of spatial dependence given the positive trend and clustering in the top right. Additionally, the correlation value itself is fairly larger than zero at 0.55, indicating spatial dependence.

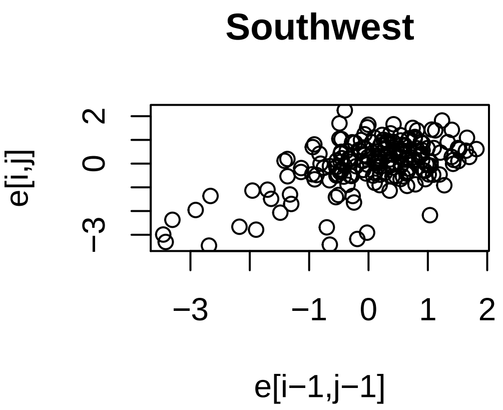
Q3c — 3 points

```
# South
plot(as.vector(res.array[1:22,1:10]),as.vector(res.array[1:22,2:11]),
     xlab="e[i,j-1]", ylab="e[i,j]", main = "South")
```



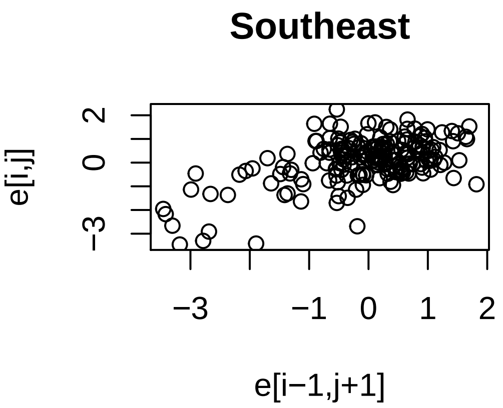
```
cor(as.vector(res.array[1:22,1:10]),as.vector(res.array[1:22,2:11]),
    use="complete.obs")
## [1] 0.51096
```

```
# Southwest
plot(as.vector(res.array[1:21,1:10]),as.vector(res.array[2:22,2:11]),
     xlab="e[i-1,j-1]", ylab="e[i,j]", main = "Southwest")
```



```
cor(as.vector(res.array[1:21,1:10]),as.vector(res.array[2:22,2:11]),
    use="complete.obs")
## [1] 0.55051
```

```
plot(as.vector(res.array[1:21,2:11]),as.vector(res.array[2:22,1:10]),
     xlab="e[i-1,j+1]", ylab="e[i,j]", main = "Southeast")
```



```
cor(as.vector(res.array[1:21,2:11]),as.vector(res.array[2:22,1:10]),
    use="complete.obs")
```

```
## [1] 0.54855
```

Yes. All of the correlations are relatively positive (>0.5), indicating that there is a positive relationship between the residual of a plot and its neighbors.