Stat 505 Assignment 2

Leslie 23 of 26 (-2 late) = 21

1. Fill in the blanks. You can use the verbatim environment

```
Source            df      SS       MS       F    p-value
---------------------------------------------------


Between groups   4    20.42      5.11    1.80    0.13

Within groups   216   613.14     2.84
---------------------------------------------------
Total           220   633.56
```

What generic null and alternative hypotheses are being tested? What do you conclude?

The null hypothesis is that the means between the four groups are the same, and the alterative hypothesis is that at least one of the means is different. There is not enough evidence to reject the null hypothesis, and we cannot conclude that any one of the means is different.

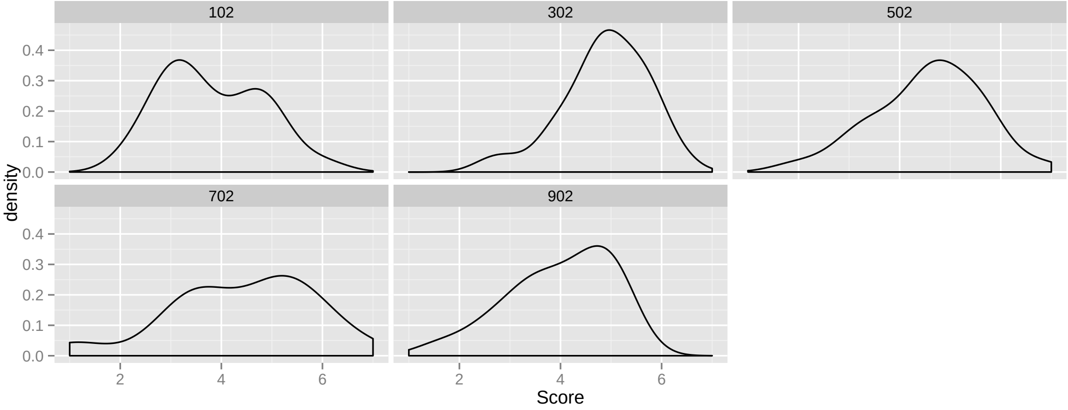
2. In *Journal of Computer-Mediated Communications*, 2008, a study was reported which assigned undergraduates to read a Facebook page and rate the attractiveness of the person. The pages were all identical expect for the "number of friends" reported which was either 102, 302, 502, 702, or 902. Read in these data and analyze to determine how people react to the reported number of friends.

```
friends <- read.table("~/Documents/Stat505/data files/friends.csv", h
```

(a) Show an appropriate plot of the data.

```
require(ggplot2)
qplot(x = Score, facets = ~Friends, data = friends, geom = "densit
```

This plot is great for comparing distributions, but not so good at comparing means. Boxplots or even dotplots would do that more clearly. (-1)



(b) Show the ANOVA table using Friends as a factor variable.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(Friends)	4	19.89	4.97	4.14	0.0034
Residuals	129	154.87	1.20		

(c) The R function `poly(x, degree)` builds an orthogonal polynomial. Using GroupNum as a continuous predictor, how high a polynomial degree can you use? How does the ANOVA table for this polynomial compare to the one in (2b)? Explain the connection. Would pvalues change if you used Friends instead of GroupNum?

You can use a polynomial of degree 2, 3, *or* 4 *because the degree must be less than the number of groups. The ANOVA table for this model is the same as the ANOVA table for the one in (2b). This makes sense, because you haven't added anything to the model. The only predictor in the model is the number of friends, which corresponds to groups* $1-5$. *Our first model treated the number of friends as a categorical variable, and this model treats group number as a continuous variable. Adding the squared, cubed, and quartic terms allows us to treat GroupNum(or number of friends) as a continuous variable, but it provides the same fit. We can see this because the ANOVA tables are exactly the same whether we treat friends as a categorical variable, or whether we use GroupNum as a continuous variable. Because GroupNum and friends contain the same information, we would get the same p-value in our ANOVA if we were to fit a linear model on poly(friends,4).*

perfectly correlated

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
poly(GroupNum, 4)	4	19.89	4.97	4.14	0.0034
Residuals	129	154.87	1.20		

We can see that the p-value does not change when Friends is used instead of GroupNum.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
poly(Friends, 4)	4	19.89	4.97	4.14	0.0034
Residuals	129	154.87	1.20		

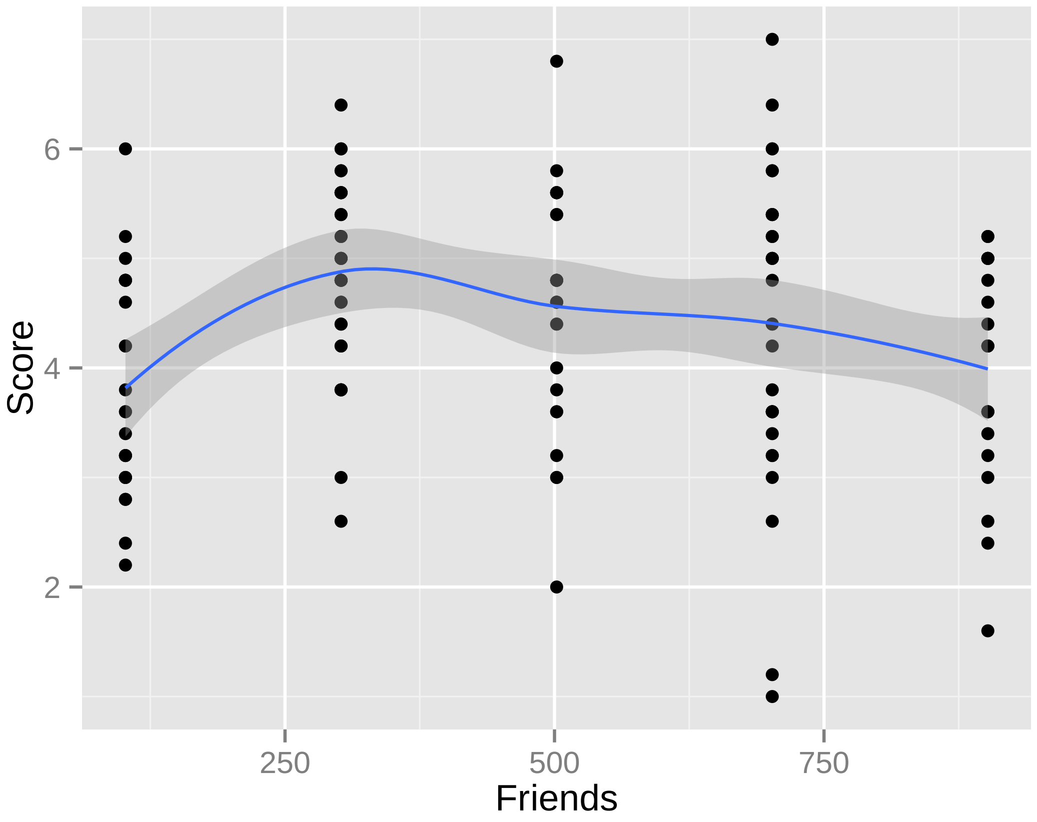
(d) The students used were a convenience sample of volunteers from a "large midwestern university". How does the sampling arrangement affect the scope of inference?

Because it was not a random sample, the scope of inference does not extend past those who participated in the study. The results cannot be extended to the entire student body at this university because we cannot be sure that those who volunteered were representative of all students at this "large midwestern university". good

(e) The students were randomly assigned to treatments. What difference does that make?

It doesn't change the scope of inference, but it does allow us to make causal inferences. Because students were randomly assigned to group one through five, we can say that the attractiveness scores are due to the experimental treatment (number of friends) and are not being dictated by the views of a particular group of people. ??

(f) Explain the relationship between number of friends and attractiveness. In particular, the researchers wondered if someone with more friends is always viewed as more attractive than someone with fewer friends. What do the data say about that?



This data suggests that those people at the extremes, with few friends or a lot of friends, have the lowest attractiveness scores, and those with an intermediate amount of friends have the highest scores. The group with 302 *friends has the highest mean attractiveness score of* 4.9*, and the group with* 502 *friends has the second highest mean attractiveness score of* 4.6*. The group with the fewest friends (*102*) has the lowest mean attractiveness score of* 3.8 *and the group with the most friends (*902*) has the second lowest mean attractiveness score of* 4.0*. The data does not suggest that people with more friends are viewed as more attractive.*

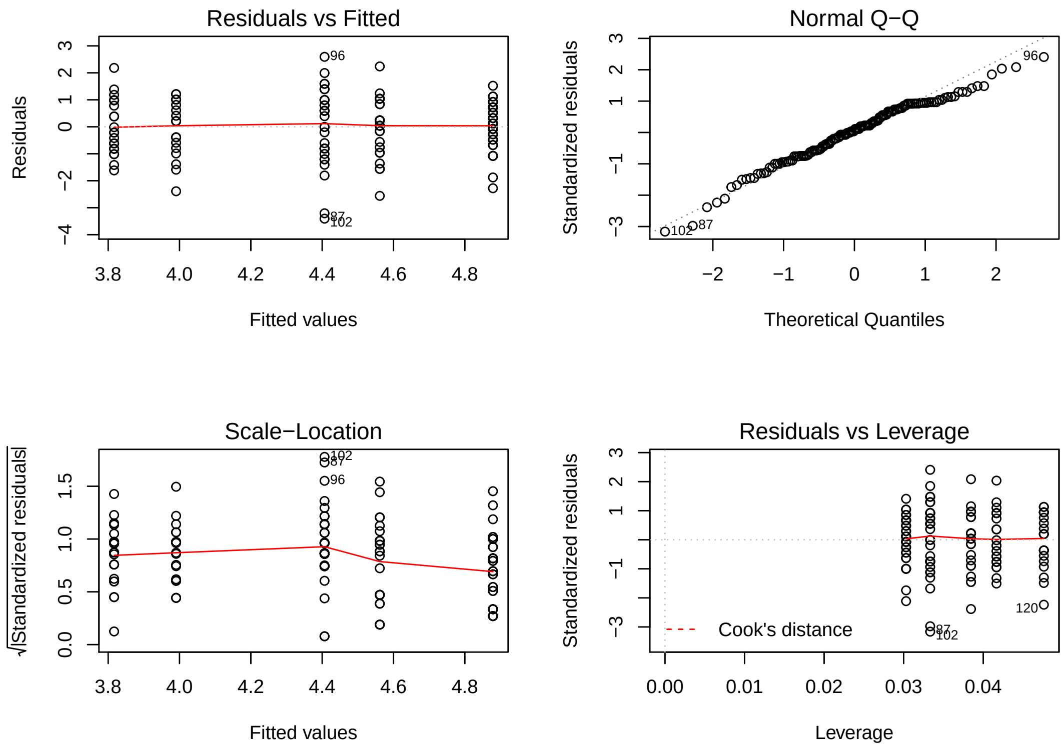
You can use the model to help here with interpretation

(g) Summarize your findings. Include a discussion of any problems with the "usual assumptions" and plots to back up your statements. (Do fix any problems.)

```
require(xtable)
xtable(summary(factor.fit))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.8167	0.2237	17.06	0.0000
factor(Friends)302	1.0621	0.2939	3.61	0.0004
factor(Friends)502	0.7449	0.3102	2.40	0.0177
factor(Friends)702	0.5900	0.3001	1.97	0.0514
factor(Friends)902	0.1738	0.3274	0.53	0.5964

```r
par(mfrow = c(2, 2))
plot(factor.fit)
```



According to the model where the numbers of friends is treated as a categorical variable, the treatment group that analyzed the person with 102 *friends gave the lowest mean attractiveness score of* 3.82*, and the* 302 *treatment had the highest mean attractiveness score of* $3.82 + 1.06 = 4.88$*. The* 402 *group had a mean score of* $3.82 + 0.74 = 4.56$*, the* 702 *group had a mean score of* $3.82 + 0.59 = 4.41$*, and the* 902 *group had the second lowest mean attractiveness score of* $3.82 + 0.17 = 3.99$*. This reinforces my previous summary that the people at the extremes (with the least and most number of friends) were viewed as the least attractive.* Means should have SE's attached.

The residual plots look good. We don't seem to have any problems with linearity. Normality and variance looks good. There are a few flagged points in the 702 group and one in the 902 group. Looking at these rows, these participants gave very low or high scores. I don't think there would be a problem with independence because each participant only rates one facebook person.

3. Researchers in Jordan are interested in plants useable for animal fodder which require little moisture. They tested four plant species in a greenhouse experiment varying the daily watering from 50 to 650 mm in 100 mm increments. Within eaach species, ater amounts were allocated at random. The response is dry biomass. Read in the data from here.

```
plantbio <- read.csv("~/Documents/Stat505/data files/plantBiomass.csv
```

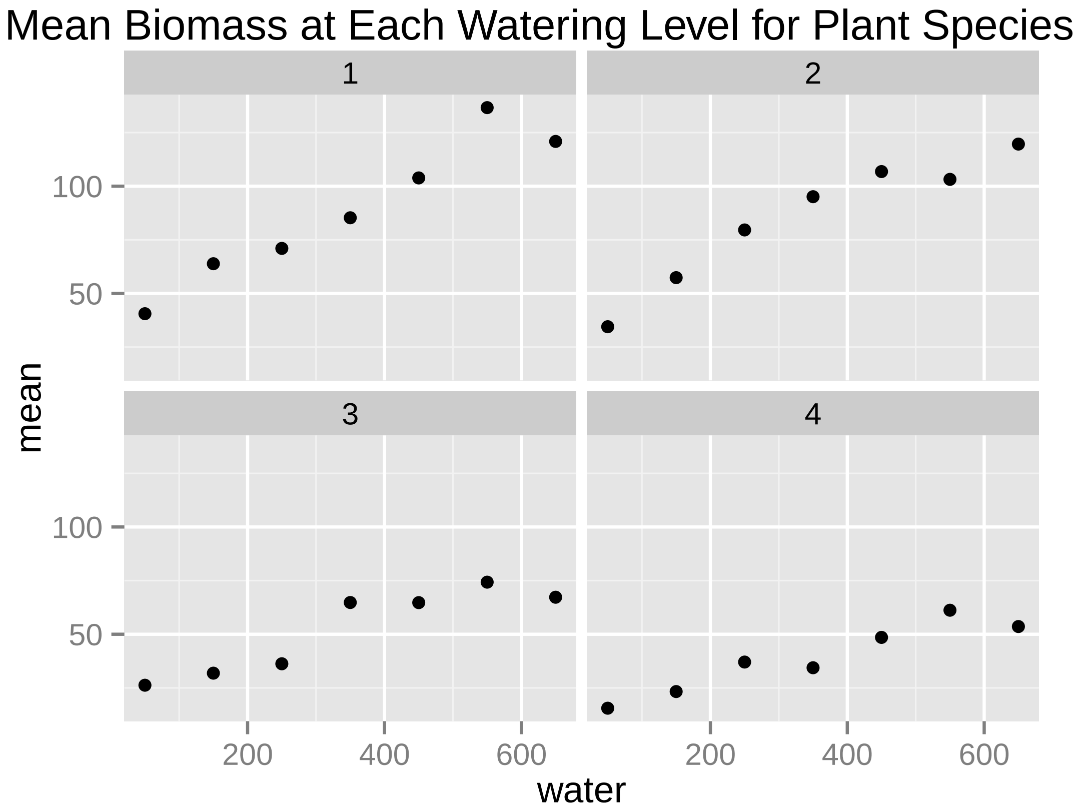
(a) Plot the data in a manner which allows us to easily compare mean biomass for each species as a function of water.

I'd like this in the appendix

```
x=0
for(i in seq(50,650,by=100)) {x[i] <- sapply(subset(plantbio,
 water==i & species == 1, select=biomass), mean)}
means1 <- x[seq(50,650,by=100)]

y=0
for(i in seq(50,650,by=100)) {y[i] <- sapply(subset(plantbio,
 water==i & species == 2, select=biomass), mean)}
means2 <- y[seq(50,650,by=100)]

v=0
for(i in seq(50,650,by=100)) {v[i] <- sapply(subset(plantbio,
 water==i & species == 3, select=biomass), mean)}
means3 <- v[seq(50,650,by=100)]

z=0
for(i in seq(50,650,by=100)) {z[i] <- sapply(subset(plantbio,
  water==i & species == 4, select=biomass), mean)}
means4 <- z[seq(50,650,by=100)]

waters <- seq(50,650,by=100)
plantmeans<- data.frame(cbind(species=
  c(rep(1,7),rep(2,7),rep(3,7),rep(4,7)),
  water=rep(waters,4), mean=c(means1,means2,means3,means4)))
```

ggplot actually has tools to extract means (or any statistic) for us.

```
require(ggplot2)
qplot(water, mean, facets = ~species, data = plantmeans, main = "M
```



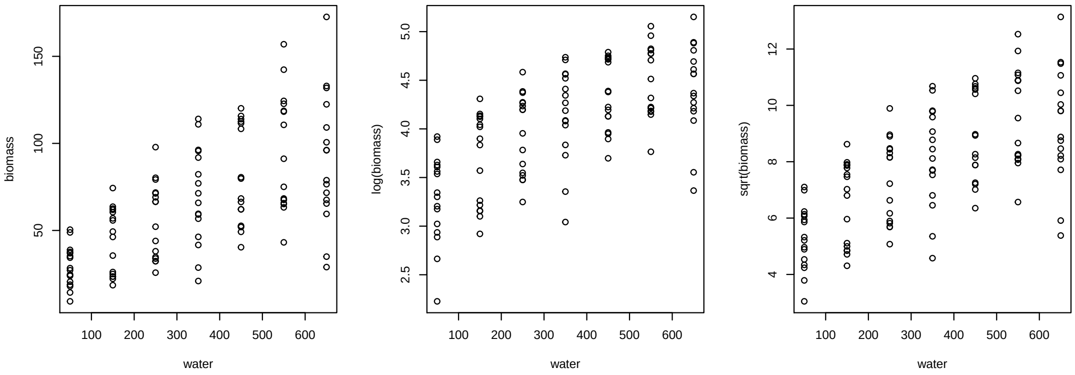
From these graphs, we can see that species one and two had more biomass than species three and four for all watering levels. The range in mean biomass for species one was $136.6 - 40.6 = 96$ *and the range in mean biomass for species three was* $74.3 - 26.2 = 48.1$. *We can see that the increase in biomass from additional watering was greater for species one than species three. A similar comparison* *could be made between species two and species four.*

(b) What questions would you ask about how the experiment was conducted which will help us determine if any assumptions were violated?

We need to check that our independence assumption is valid. We want to check that all other factors besides water and species are controlled. Here are some questions we should ask. Were other factors such as amount of sunlight, space, and amount of nutrients in the soil the same for each plant? What stage in the life cycle of the plants did they conduct this experiment? Were all the plants in the same stage of their lives? Were all the plants in their own pot? Did the plants come from different seed batches? Yes, that would set up correlations.

(c) Provide a full analysis of the effects of water on each plant species. Consider whether we should use water as a factor or as a continuous predictor. Examine the usual plots and fix any problems with the assumptions. Write a report to the researchers and include a discussion of the scope of inference for the results. Which plants appear to make the best use of limited amounts of water? How confident can we be that greenhouse results will carry over to the field?

```
par(mfrow = c(1, 3))
plot(biomass ~ water, data = plantbio)
plot(log(biomass) ~ water, data = plantbio)
plot(sqrt(biomass) ~ water, data = plantbio)
```



I am going to treat water as a factor because I don't see a linear relationship between water and biomass across all watering levels.

For species 1, 3, and 4, the mean biomass at the 650 watering level was lower than the mean biomass at the 550 watering level.

I explored a few different transformations. The log transformation seems to make the data less linear, and, looking at the figures above, the square root transformation didn't seem to change the linearity but did seem to make the variance more constant. I fit both models, one with sqrt(biomass) and the other without a transformation on biomass. As I predicted, the variance and normality of the residuals in the sqrt model looked better. That's the one I'll report. I started by including an interaction term, and then looked at the ANOVA table to see if it added anything to the model.

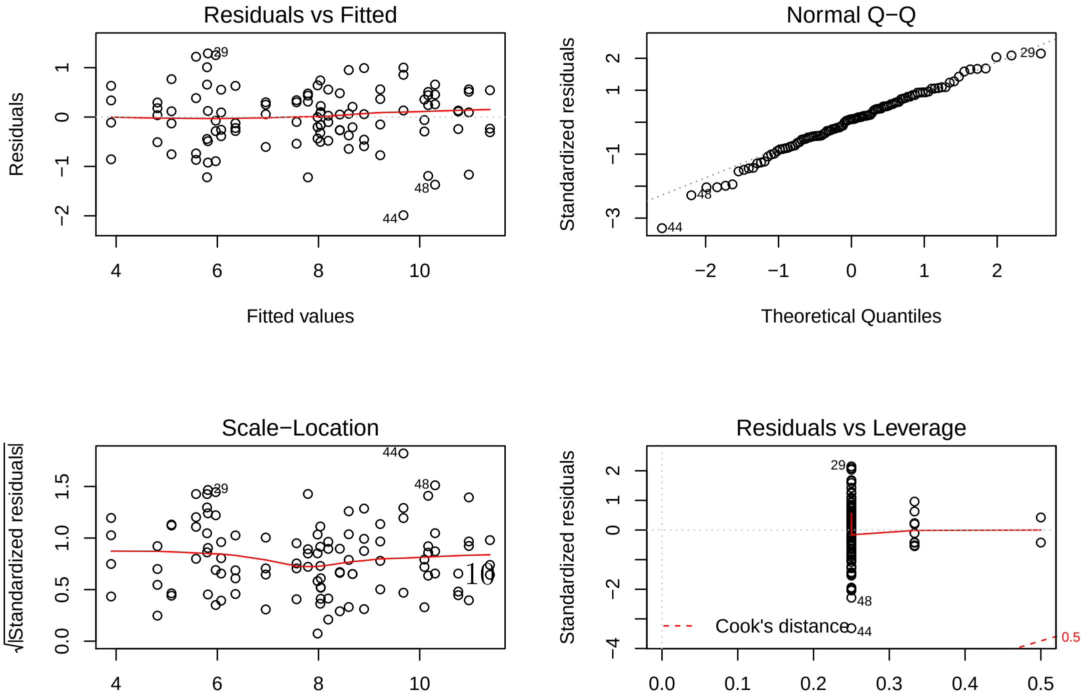
```
plant.ro <- plantbio[-c(23, 52, 53, 109, 110), ]
```

```
bmassFit <- lm(sqrt(biomass)~factor(water)*factor(species), data=p

require(xtable)
xtable(summary(bmassFit))
```

```
xtable(anova(bmassFit))
```

```
par(mfrow=c(2,2))
plot(bmassFit)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.3584	0.3463	18.36	0.0000
factor(water)150	1.6230	0.4897	3.31	0.0014
factor(water)250	2.0623	0.4897	4.21	0.0001
factor(water)350	2.8619	0.4897	5.84	0.0000
factor(water)450	3.8086	0.4897	7.78	0.0000
factor(water)550	5.0297	0.5290	9.51	0.0000
factor(water)650	4.6131	0.4897	9.42	0.0000
factor(species)2	-0.5443	0.4897	-1.11	0.2697
factor(species)3	-1.2645	0.4897	-2.58	0.0117
factor(species)4	-2.4577	0.4897	-5.02	0.0000
factor(water)150:factor(species)2	0.1285	0.6926	0.19	0.8532
factor(water)250:factor(species)2	1.0239	0.6926	1.48	0.1433
factor(water)350:factor(species)2	1.0025	0.6926	1.45	0.1517
factor(water)450:factor(species)2	0.6811	0.6926	0.98	0.3284
factor(water)550:factor(species)2	-0.0818	0.7481	-0.11	0.9132
factor(water)650:factor(species)2	-0.3337	0.7208	-0.46	0.6447
factor(water)150:factor(species)3	-1.1373	0.6926	-1.64	0.1045
factor(water)250:factor(species)3	-1.1877	0.6926	-1.71	0.0903
factor(water)350:factor(species)3	0.0799	0.6926	0.12	0.9084
factor(water)450:factor(species)3	-0.8581	0.6926	-1.24	0.2190
factor(water)550:factor(species)3	-1.5260	0.7208	-2.12	0.0374
factor(water)650:factor(species)3	-1.5144	0.6926	-2.19	0.0317
factor(water)150:factor(species)4	-0.7075	0.6926	-1.02	0.3101
factor(water)250:factor(species)4	0.1129	0.6926	0.16	0.8709
factor(water)350:factor(species)4	-0.9654	0.6926	-1.39	0.1672
factor(water)450:factor(species)4	-0.7517	0.6926	-1.09	0.2810
factor(water)550:factor(species)4	-1.1398	0.7208	-1.58	0.1178
factor(water)650:factor(species)4	0.1584	0.7743	0.20	0.8385



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(water)	6	229.68	38.28	79.81	0.0000
factor(species)	3	170.57	56.86	118.54	0.0000
factor(water):factor(species)	18	15.97	0.89	1.85	0.0331
Residuals	79	37.89	0.48		

Looking at the anova table, we can see that the interaction term has an F statistic of 1.85 *and a p-value of* 0.0331*. At a significance level of* 0.05*, we will keep the interaction term because it indicates that the interaction term does add to the model.*

I removed five outliers from the data, one from species 1*, two from species* 2 *and two from species* 4*. The pair of outliers from species* 4 *had very low biomass in the* 650 *watering group. I'm guessing that either these two plants were recording errors or they responded drastically to overwatering. I also removed two outliers from species* 2*, one that had low biomass in the* 550 *watering group, and another that had high biomass in the* 650 *watering group. All of the outliers I removed were from the two highest watering levels, showing that we see more variation in plant growth for higher watering levels.*

If I did the removal (which I'm not incliined to do), I'd report 2 analyses, with

After applying the transformation and removing the outliers, the residual plots look good.

and without the outliers

Here is my report to the researchers. Species 1 *and* 2 *had more biomass than species* 3 *and* 4 *at all watering levels. At a significance level of 0.05, we saw that species* 3 *had a lower mean biomass than species* 1 *at the first watering level (*$t = -2.58, p = 0.012$*). Species* 4 *also had a lower mean biomass than species* 1 *at the first watering level (*$t = -5.02, p < 0.00001$*). We cannot say that species* 2 *had a lower mean biomass than species* 1 *at the first watering level at a significance level of* 0.05 *(*$t = -1.11, p = 0.27$*). We can say that species* 1 *and* 2 *made better use of limited amounts of water than did species* 3 *and* 4*.*

We do see an increase in mean biomass for all species when the water level increases from 50 *to* 150*. There is strong evidence that* *the 550 water level results in more growth in all species than the 50 water level ($t = 9.51, p < 0.00001$). Species 1, 3, and 4 all had less biomass in the 650 watering level than they did in the 550 watering level, indicating that these species are susceptible to overwatering.*

The scope of inference of the experiment does not extend beyond the population from which you took the sample. Thus, the results cannot be generalized to plants outside of a greenhouse. Outside plants could be subject to other factors (such as predation) that were not taken into account in this experiment.

??

assuming this is a random sample?

(d) Suppose that we fit a model with an intercept and a slope for each plant.

$$y_i = \beta_0 + \beta_1 x_i + \alpha_{0j[i]} + \alpha_{1j[i]} x_i + \epsilon_i$$

In this notation, i is the row number, $j[i]$ is the plant species of the plant in row $i (j = 1, \ldots, 4)$, β's are overall effects, and α's are adjustments for each plant species. Fit this model in R and show the estimated coefficients. What combination of Greek letters is estimated by each coefficient shown? Which coefficients does R effectively set to zero? Plot the four lines. Here's one way to do multiple lines.

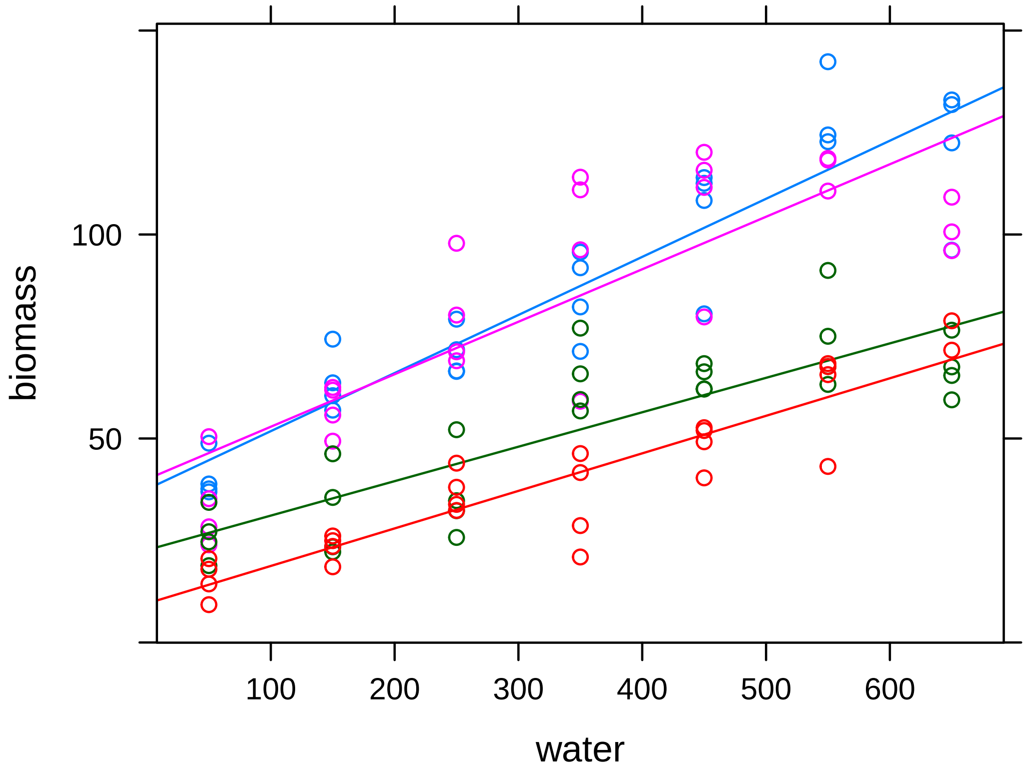
β_0 is the intercept for the first plant species. β_1 is the slope for the first plant species. $\alpha_{0j[i]}$ is the adjustment to the intercept for species j in row i, ($j = 1, 2, 3, 4$). $\alpha_{1j[i]}$ is the adjustment to the slope for species j in row i. R effectively sets $\alpha_{0,1[i]}$ and $\alpha_{1,1[i]}$ equal to zero. Let's look at each value in the linear model summary table.

```r
require(lattice)
xyplot(biomass ~ water, data = plant.ro,
       group = species, type=c("p","r"))
```

Table 1: Linear Model Summary Interpretation

	Greek Letter
Intercept for Species 1	$E[y_{1[i]}] = \beta_0 + \alpha_{0,1[i]} = \beta_0$
Slope for Species 1	$E[y_{1[i]} \mid x_i = a+1] - E[y_{1[i]} \mid x_i = a]$ $= \beta_0 + \alpha_{0,1[i]} + \beta_1(a+1) + \alpha_{1,1[i]}(a+1)$ $-(\beta_0 + \alpha_{0,1[i]} + \beta_1 a + \alpha_{1,1[i]} a) = \beta_1 + \alpha_{1,1[i]} = \beta_1$
Adjustment to the Intercept for Species j	$E[y_{j[i]}] - E[y_{1[i]}] = \beta_0 + \alpha_{0,j[i]} - \beta_0 = \alpha_{0,j[i]}$ -alpha_{01}
Adjustment to the Slope for Species j	$E[Y_{j[i]} \mid x = a+1] - E[y_{j[i]} \mid x = a]$ $-(E[y_{1[i]} \mid x = a+1] - E[y_{1[i] \mid x=a}])$ $= \beta_0 + \beta_1(a+1) + \alpha_{0,j[i]} + \alpha_{1,j[i]}(a+1)$ $-(\beta_0 + \beta_1 a + \alpha_{0,j[i]} + \alpha_{1,j[i]} a)$ $-(\beta_0 + \beta_1(a+1) - \beta_0 - \beta_1 a) = \alpha_{1,j[i]}$ -alpha_{11}

(-2)



```
## group splits the data, type='r' draws regression line, 'p'
```

R Code

```
# this is equivalent to \SweaveOpts{...}
opts_chunk$set(fig.width=5, fig.height=4, out.width='\\linewidth', dev='p
options(replace.assign=TRUE,width=112, digits = 3, max.print="72",
        show.signif.stars = FALSE)
```

```
friends <- read.table("~/Documents/Stat505/data files/friends.csv",head=T
```

```
require(ggplot2)
qplot(x = Score, facets=~Friends, data = friends,
      geom = 'density')
```

```
factor.fit <- lm(Score~factor(Friends), data=friends)
require(xtable)
print(xtable(anova(factor.fit)), table.placement =
 getOption("xtable.table.placement", "H"))
```

```
group.fit <- lm(Score~(poly(GroupNum, 4)), data=friends)
require(xtable)
print(xtable(anova(group.fit)), table.placement =
  getOption("xtable.table.placement", "H"))
```

```
check.fit <- lm(Score~(poly(Friends,4)),data=friends)
require(xtable)
print(xtable(anova(check.fit)), table.placement =
  getOption("xtable.table.placement", "H"))
```

```
qplot(Friends, Score, data=friends,
  geom=c("point","smooth"))
```

```r
require(xtable)
xtable(summary(factor.fit))
par(mfrow=c(2,2))
plot(factor.fit)
```

```r
plantbio <- read.csv("~/Documents/Stat505/data files/plantBiomass.csv",
                    head=T)
```

```r
x=0
for(i in seq(50,650,by=100)) {x[i] <- sapply(subset(plantbio,
 water==i & species == 1, select=biomass), mean)}
means1 <- x[seq(50,650,by=100)]

y=0
for(i in seq(50,650,by=100)) {y[i] <- sapply(subset(plantbio,
 water==i & species == 2, select=biomass), mean)}
means2 <- y[seq(50,650,by=100)]

v=0
for(i in seq(50,650,by=100)) {v[i] <- sapply(subset(plantbio,
 water==i & species == 3, select=biomass), mean)}
means3 <- v[seq(50,650,by=100)]

z=0
for(i in seq(50,650,by=100)) {z[i] <- sapply(subset(plantbio,
  water==i & species == 4, select=biomass), mean)}
means4 <- z[seq(50,650,by=100)]

waters <- seq(50,650,by=100)
plantmeans<- data.frame(cbind(species=
  c(rep(1,7),rep(2,7),rep(3,7),rep(4,7)),
  water=rep(waters,4), mean=c(means1,means2,means3,means4)))
```

```
require(ggplot2)
qplot(water, mean, facets=~species, data = plantmeans,
main  = "Mean Biomass at Each Watering Level for Plant Species 1-4")
```

```
par(mfrow=c(1,3))
plot(biomass~water, data=plantbio)
plot(log(biomass)~water, data=plantbio)
plot(sqrt(biomass)~water, data=plantbio)
```

```
plant.ro <- plantbio[-c(23,52,53,109,110),]
```

```
bmassFit <- lm(sqrt(biomass)~factor(water)*factor(species), data=plant.ro

require(xtable)
xtable(summary(bmassFit))
xtable(anova(bmassFit))
par(mfrow=c(2,2))
plot(bmassFit)
```