The first problem was inspired by Mooney et al. (2010) who studied evolutionary trade-offs in milkweed species. The second problem was inspired by research done in Andy Suarez’s lab. Recent work to understand caste determination in ants was published by Smith et al. (2008).

Note that extensive R code is provided here, so you can replicate graphs and analyses. But you actually did not need to show all of your code to get full credit. ANOVA tables are good to show. But otherwise, I did not expect to see much else. It was more important to interpret the results in your text.

1. **Summary:** An experiment was conducted to test plant defense theory, according to which plants would have to choose between investing their resources for fast growth, or for resistance. Multiple linear regression was utilized to analyze these data. The raw biomass measurements were log-transformed, so as to compare rates of growths across the different species. The analysis showed that biomass was affected by both the fertilizer treatment and the aphid predator treatments, and that these effects were species-specific (i.e. there were significant interactive effects species:predator and species:fertilizer). These species-specific responses to predator treatment and to fertilizer treatment were later analyzed with linear regression. As predicted by theory, the species’ response to fertilizer was positively correlated to the species’ benefit from aphid predators (p=.007). However, the species log-biomass did not correlate to the species’ benefit from aphid predators (p=.52).

(a) Plots of log transformed biomass for all species are given below. The first plot shows how biomass differs greatly across species. The effects of the predator treatment and of the fertilizer treatment seem to be pretty small when data from all species are combined together (not shown), but these effects are clearly consistent across species (second plot): reducing aphid predator appears to cause a increase in biomass, and fertilizer treatment also seems to increase biomass, which all make sense. Furthermore, there seems to be no interactive effect between the fertilizer and the predator treatments, within any of the 14 species (parallel lines in the third plot).

---

1Evolutionary trade-offs in plants mediate the strength of trophic cascades. Science 327:1642-1644.
library(lattice)
library(MASS)
plant=read.table('plantgrowth.txt', header=T)
plant$species=factor(plant$species)
plant$fertilizer=factor(plant$fertilizer)
plant$plant=factor(plant$plant)
plant$species = reorder(plant$species, plant$biomass)
plot(log(biomass)~species, plant)
xyplot(log(biomass)~fertilizer|species, groups=predator, plant, type=c('a'), lty=1:2, auto.key=list(columns=2))
with(plant, interaction.plot(species, predator, log(biomass), yaxt="n", ylab="mean biomass (g)"))
bb = c(100, 200, 400, 800, 1000)
axis(2, at=log(bb), labels=bb, las=1)
with(plant, interaction.plot(species, fertilizer, log(biomass), yaxt="n", ylab="mean biomass (g)"))
axis(2, at=log(bb), labels=bb, las=1)

(b) Biomass was log-transformed for analysis. The residual and diagnostic plots look much better with this transformation, showing a homogeneous variance and normal distribution for residuals (diagnostic plots not shown). Furthermore, an additive increase in log-biomass corresponds to a multiplicative increase, which seem to be more comparable across species of very different size. For instance, a 200% increase can be interpreted as an increase of comparable size, across plants of various sizes from grasses to shrubs.

Based on plant defense theory, different types of the plants may react to environmental conditions quite differently. These different reactions essentially correspond to interactive terms between species and predator, and between species and fertilizer. So I definitely want to include species:predator and species:fertilizer in the full model. I will also test the presence of the other two- or three-way interactions. Note that I will not consider including the replication number plant into the model, since this number is quite meaningless: there is nothing in common for plants labeled with same number. This plant number was just a number from 1 to 5 sequentially assigned to units from each treatment and species, and used for randomization in the field.

Several selection methods were used. Search strategies included forward and backward selection. Model selection used the AIC and F-tests. All selection methods lead to the same final linear model, where log(biomass) is explained by species (14 categories), fertilizer treatment, predator treatment and the 2-way interactions between species:fertilizer (F-test, $p < .0001$) and species:predator (F-test, $p < .00001$). The 2-way interaction between fertilizer and predator treatments was not significant (F-test, $p = 0.9977$). The main effects of species, predator and fertilizer treatments were all statistically significant (F-tests, $p < 2.2 \times 10^{-16}$ for each).
Stat 572  
Second Midterm Exam  
April 15, 2010

```r
fit=lm(log(biomass)~ species + predator + fertilizer + species:fertilizer+species:predator, data=plant)
```

Analysis of Variance Table

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>RSS</th>
<th>AIC</th>
<th>F Value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>8.4476</td>
<td>-896.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ predator:fertilizer</td>
<td>1</td>
<td>0.0000</td>
<td>8.4476</td>
<td>-894.25</td>
<td>0.0000</td>
</tr>
<tr>
<td>- species:predator</td>
<td>13</td>
<td>1.6106</td>
<td>10.0582</td>
<td>873.39</td>
<td>3.4905</td>
</tr>
<tr>
<td>- species:fertilizer</td>
<td>13</td>
<td>1.9723</td>
<td>10.4199</td>
<td>863.50</td>
<td>4.2743</td>
</tr>
</tbody>
</table>

This ANOVA table is sequential, but since this is a completely balanced factorial design, the order in the sequence does not matter. This analysis confirms that plants have different responses to the environmental changes.

Note that averaging the biomass of all 5 plants in each treatment group was a great loss of information. It resulted in no df left to test the 3-way interaction, and a loss of power to test 2-way interactions. This was detrimental to answer (d), since 2-way interactions captured how differently species responded to the fertilizer and predator treatments.

(c) To answer this question, we can either use the model above, or use the data from species 3 only.

```r
> summary(fit)
Coefficients:  
Estimate Std. Error t value  Pr(>|t|)
(Intercept) 4.862990  0.072967 66.647 < 2e-16 ***
species3 1.220733  0.103190 11.830 < 2e-16 ***
predatorreduced -0.218865  0.084254 -2.598  0.009971 **
fertilizer5 0.045375  0.084254  0.539  0.590702
species3:fertilizer5 0.334077  0.119154  2.804  0.005468 **
species3:predatorreduced -0.091445  0.119154 -0.767  0.443573
```

# or, alternatively, re-fit a model on species 3 only:

```r
> plant.3=subset(plant,species=='3')  
> fit.3=lm(log(biomass)~predator+fertilizer, plant.3)  
> coef(summary(fit.3))
```

# yet another method: use the full model with species 3 as the reference species (not shown here)

From my model above, which assumes no interaction between predator and fertilizer effects, I predict for species 3 that the effect of reducing aphid predators on log(biomass) is $-0.2188 - 0.0914 = -0.31$. This means that reducing aphid predators caused a reduction of plants biomass by about 31% (more exactly, a reduction by $(1 - \exp(-0.31)) \times 100\% = 27\%$) on average.

Similarly, based on my model, the effect of adding fertilizer on log(biomass) is estimated at $0.0453 + 0.3340 = 0.38$ for species 3, which means that 5g of fertilizer caused an increase of biomass by about 38% (or more exactly, by $(\exp(0.38) - 1) \times 100\% = 46\%$) on average.

(d) The three measures for each species are given in the table below along with the scatter plots between growth measures and resistance measure. These measures can be obtained from the average log-biomass in each condition, from the data set itself, or can be obtained from the model obtained in (b). It would not be appropriate to obtain these measures from model with no interactions, because otherwise the same measures would be obtained for all species, and the interesting information would be lost. The measure of resistance to aphids is calculated as how much the plant’s log
biomass decreased on average due to the reduction of aphid predators. Note that in the table below, we find the same coefficients for species 3 as those calculated in (c).

<table>
<thead>
<tr>
<th>species</th>
<th>mean log biomass</th>
<th>response to fertilizer</th>
<th>resistance to aphids</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>4.78</td>
<td>0.05</td>
<td>-0.22</td>
</tr>
<tr>
<td>1</td>
<td>5.11</td>
<td>0.06</td>
<td>-0.20</td>
</tr>
<tr>
<td>13</td>
<td>5.39</td>
<td>0.08</td>
<td>-0.23</td>
</tr>
<tr>
<td>10</td>
<td>5.39</td>
<td>0.34</td>
<td>-0.39</td>
</tr>
<tr>
<td>6</td>
<td>5.43</td>
<td>-0.11</td>
<td>0.04</td>
</tr>
<tr>
<td>9</td>
<td>5.44</td>
<td>0.03</td>
<td>-0.21</td>
</tr>
<tr>
<td>4</td>
<td>5.64</td>
<td>0.32</td>
<td>-0.63</td>
</tr>
<tr>
<td>8</td>
<td>5.71</td>
<td>0.49</td>
<td>-0.23</td>
</tr>
<tr>
<td>11</td>
<td>5.74</td>
<td>0.22</td>
<td>-0.28</td>
</tr>
<tr>
<td>12</td>
<td>5.94</td>
<td>0.32</td>
<td>-0.45</td>
</tr>
<tr>
<td>5</td>
<td>6.02</td>
<td>0.26</td>
<td>-0.40</td>
</tr>
<tr>
<td>3</td>
<td>6.12</td>
<td>0.38</td>
<td>-0.31</td>
</tr>
<tr>
<td>14</td>
<td>6.53</td>
<td>0.37</td>
<td>-0.37</td>
</tr>
<tr>
<td>2</td>
<td>6.86</td>
<td>0.07</td>
<td>-0.18</td>
</tr>
</tbody>
</table>

In these plots, each species is represented by its number. The first plot shows that there is no clear relationship between the species’ resistance to aphids and how large the plant is in term of average log biomass. The second plot shows the response to fertilizer against the resistance to aphids.
log biomass ($r^2 = 0.03, p = 0.52$). It could be that plant biomass does not provide an representative measure of the plant’s growth strategy (below-ground biomass was ignored for instance). The second plot shows a clear negative relationship between how much growth was gained from fertilizer and the measure of resistance to aphids ($r^2 = 0.46, p = 0.007$). It is in accord with the theory that predicts those fast-growing plants (high scores of growth measure) would have less resistance to herbivores and would benefit more from the presence of aphid predators.

```r
> fit.d1=lm(growth1~resistance, measure)
> summary(fit.d1)
Multiple R-squared: 0.03475, Adjusted R-squared: -0.04569
F-statistic: 0.432 on 1 and 12 DF, p-value: 0.5234
> fit.d2=lm(growth2~resistance, measure)
> summary(fit.d2)
Multiple R-squared: 0.4618, Adjusted R-squared: 0.4169
F-statistic: 10.3 on 1 and 12 DF, p-value: 0.00751
layout(matrix(1:2,1,2))
par(mar=c(3,3,2,.5), mgp=c(1.3,.3,0), tck=-.01)
plot(growth1~resistance, cex.lab=1.2, type="n", data=measure, ylab="mean log biomass", xlab="resistance to aphids")
text(y=growth1, x=resistance, labels=measure$species)
abline(lm(growth1~resistance, measure))
plot(growth2~resistance, cex.lab=1.2, type="n", data=measure, ylab="mean response to fertilizer", xlab="resistance to aphids")
text(y=growth2, x=resistance, labels=measure$species)
abline(lm(growth2~resistance, measure))
```

If the raw biomass differences had been used instead of the log-transformed biomass, a stronger negative relationship would have been found between growth and resistance, and a significant relationship would have been found between biomass and predator benefit. But this last relationship could be just an artifact that a 20% increase (say) due to predator presence represents a larger increase in terms of raw biomass for a large plant (like a shrub) than for a small plant (like a grass). To compare species on a more equal footing, using the log-biomass seems more appropriate here. The residual plots of the 2 simple linear regressions above looked okay.

It was adequate to look at the effects on the raw biomass and then to express these fertilizer/predator effects in terms of % decrease or % increase. This method resulted in almost equivalent graphs to those above, up to a change in scale.

Another adequate method was to divide up the set of 14 species into a set of “slow” growers (7 species) and a set of “fast” growers (7 species) based on their response to fertilizer, for instance. Then, a 2-sample t-test could be used to see if the mean resistance to aphids is the same for slow and fast growers. The same qualitative conclusion would have been reached. Other students have used yet other adequate methods, based on dividing up the set of species into ‘speed’ categories.
2. **Summary:** An experiment was conducted to understand which of diet and/or genetic background influence caste determination in a particular colony of a particular species of ants. Logistic regression showed that both diet and genetic factors are linked to caste determination (Likelihood ratio test, \( p = 0.0002 \) for diet and \( p = 0.028 \) for father-line). Since this was essentially an observational study, many diet measurements are highly correlated. Therefore, it is not clear which diet variable is most influential and the analysis cannot state any causal relationships.

(a) Percent Nitrogen, C:N ratio and \( \delta^{15}N \) are highly correlated (see figure below) and all three predictors seem to be related to the proportion of workers (WF) and of reproductive females (RF). \( \delta^{13}C \) does not seem to be related to the proportion of RF (plot not shown). The proportion of WF and RF seemed to vary between fathers, but samples sizes are relatively low from each father (from 14 to 19). Father 2 and 5 sired a higher proportion of RF than the other fathers.

(b) Since 3 of the diet measurements were highly correlated, and since the sample size is not super large, I will attempt to include only one variable at a time among percentN, CtoN and \( \delta^{15}N \). Logistic
regression will be used to explain caste based on the following predictors: father (5 categories), $\delta^{13}C$, and one of the other 3 diet variables. Model selection will start with a full model that has all 2-way and possibly the 3-way interaction. I will use AIC and likelihood ratio tests.

**Note:** even though many tests are performed with the function `anova()` in R, not all these tests are ANOVA tests. Here, the proper name was a ‘likelihood ratio’ test. The proper name of the test was not a chi-square test either. The LR test is performed by comparing the likelihood ratio statistic to a chi-square distribution, but the actual name of the method is still a ‘likelihood ratio’ test.

My best model was found to include father (LRT, $X^2 = 10.9$, df=4, $p = 0.028$) and percentN (LRT, $X^2 = 13.9$, df=1, $p = 0.0002$). Their interaction was not significant (LRT, $p = 0.67$). Other good models included the C:N ratio (LRT, $X^2 = 12.8$, df=1, $p = 0.0003$) or $\delta^{15}N$ (LRT, $X^2 = 13.1$, df=1, $p = 0.0003$) as the diet predictor instead of %N. Note that the model with an interactive effect between father-line and $\delta^{15}N$ lead to a perfect separation of worker females and reproductive females from father 1 (WF if $\delta^{15}N < 4.3$ and RF if $\delta^{15}N > 4.4$), which may suggest that $\delta^{15}N$ is a great determinant of caste. However, it was not the case in any other father-line, therefore I chose not to include the interaction between father-line and $\delta^{15}N$ when these 2 predictors were combined.

```
> full=glm(caste.isqueen~father*percentN*d13C,family=binomial,data=ant)
> library(MASS)
> stepAIC(full,scope=~father*percentN+CtoN*d15N*d13C,test="Chisq")
caste.isqueen ~ father + percentN
Df Deviance AIC LRT Pr(Chi)
<none> 96.079 108.08
+ d15N 1 95.606 109.61 0.4736 0.4913498
+ d13C 1 95.862 109.86 0.2168 0.6414646
+ CtoN 1 96.048 110.05 0.0309 0.8605228
- father 4 106.977 110.98 10.8983 0.0277305 *
+ father:percentN 4 93.749 113.75 2.3304 0.6752447
- percentN 1 109.966 119.97 13.8865 0.0001942 ***
```

```
> fit = glm(caste.isqueen~father+percentN,family=binomial,data=ant)
> summary(fit)
> sum(residuals(fit, type="pearson")^2)/79
[1] 1.077988
```

Overall, there is strong evidence that diet and father-line are linked to caste determination. There was no lack of fit in my final logistic model, as evidenced by a dispersion very close to 1 (1.08) and by a lack-of-fit test (residual deviance = 96.08, $p = 0.25$ by parametric bootstrapping).

i. Due to high correlations between pairs of diet variables, each variable was significant when included alone, but became non-significant when another diet variable was included. Therefore, we cannot tell which of %N, C:N ratio of $\delta^{15}N$ is most influential on caste determination. When both $\delta^{13}C$ and %N were included in the model (along with father-line), $\delta^{13}C$ was not significant but %N was still significant. This suggests that $\delta^{13}C$ is not linked to caste determination, other than just from its correlation with the other diet variables. Furthermore, the effect of $\delta^{13}C$ was then estimated to be negative, which goes against expectations (fortunately this negative coefficient was not-significantly different from zero).

ii. From the estimated coefficients associated with fathers, father-line 2 appears to be most biased toward producing reproductive females and father-line 1 appears to be most biased toward producing worker females, given equal nutrition. The odds for a RF in father-line 2 are $\exp(2.1437) = 8.53$ times as large as the odds in father-line 1, given equal %N. For instance, at the average 12.8%N, the probability of a RF is 21.8% from father 1 and 70.4% from father 2. The difference is 48.6 percentage points, which is quite large!
```r
> fit = glm(caste.isqueen~father+percentN, family=binomial, data=ant)
> summary(fit)
Coefficients:

             Estimate  Std. Error z value Pr(>|z|)  
(Intercept)  -7.1041      2.0347  -3.492 0.00048 ***  
father2      2.1437       0.8996   2.383 0.01717 *   
father3      0.4519       0.8724   0.518 0.60444          
father4      0.7204       0.8910   0.809 0.41878          
father5      1.8873       0.8799   2.145 0.03196 *   
percentN     0.4555       0.1420   3.208 0.00134 **            

> newdat = data.frame(father=factor(1:2), percentN=rep(12.8,2))
> predict(fit, newdat, type="response")
   1     2
0.2186 0.7047
> 70.4 - 21.8
[1] 48.6
```