DISCUSSION 3

1 Problem

Read the dataset from

1. Examine the mass for each colony graphically. Summarize your observations.

2. Find the mean and standard deviation of the Headwidth and the mass for each colony. Recall that the function `tapply(var, factor, fun)` will apply the function `fun` to a variable `var`, separately for each level of the factor `factor`.

3. Plot Headwidth versus log of mass using a different color or plotting symbol for each colony. Also (either in separate plots or new plots), use a different color or plotting symbol for each genus. Add a legend to your plot(s), either with the function `legend` in R or manually.

4. Based on the visual inspection of these plots, which model best predicts Headwidth?

5. Fit following models to predict Headwidth from the log of mass and these additional explanatory variables:
   - none
   - colony
   - colony and an interaction colony-log of mass
   - genus

   For each model, write the equation that predicts the Headwidth for colony 5 in the style above. Find a numerical prediction for the Headwidth of an ant in colony 4 with mass value 40 for each model.

6. Use the last model above to predict the Headwidth of an ant from each colony where the mass is the average for that colony.

7. Consider model 2 and model 4. Which is nested in the other one? To determine how many ants were sampled in each colony/genus level, the `table` function might be useful. `table(pred1, pred2)` will provide a table with a row for each level of `pred1`, a column for each level of `pred2`, and will fill in the table with the number of units sampled at each level combination.

8. Use F-test to compare model 1, model 2 and model 3. Use F-test to compare model model 1, model 2 and model 4.

9. Based on the tests in question 8, select a preferred model. Does this choice match your answer to question 4? How many parameters are needed to describe this model? What percentage of the variability in Headwidth is explained by this model?

10. Compare the $R^2$ values of all models above. Does your selected model have the largest $R^2$ value? Why or why not?

11. Make a residual plot for the model you selected, and list any assumption that might not be met for the regression analysis. If you think all assumptions are met, then briefly explain why you think so.
12. Using the model selected above, re-analyze the data using the non-transformed mass values instead of using the log of mass. Which analysis shows the most linear relationship? Does any of these two analyses suffer from non-homogeneous variance? Is any of these two analyses influenced by a few highly influential points? Overall, was it a good choice to use the square root mass rather than the original mass values to predict Headwidth?

- Outliers and Influential points
  1. Regression outliers: unusual Y values
  2. High leverage points: unusual X values
  3. Influential points: points that significantly effect the fit (check Cook’s distance)

2 Solutions

```r
ants=read.table("ants.txt",header=T)
str(ants)
#transform those categorical variables into factors.
ants$Genus=factor(ants$Genus)
ants$Colony=factor(ants$Colony)

#1
plot(Mass~Colony,ants)

#2
with(ants,tapply(Headwidth,Colony,mean))
with(ants,tapply(Headwidth,Colony,sd))
with(ants,tapply(Mass,Colony,mean))
with(ants,tapply(Mass,Colony,sd))

#3
par(mfrow=c(1,1))
with(ants,plot(Headwidth~log(Mass),pch=as.numeric(Colony)))
legend("topleft",pch=1:10,legend=levels(ants$Colony))
with(ants,plot(Headwidth~log(Mass),pch=as.numeric(Genus)))
legend("topleft",pch=1:10,legend=levels(ants$Genus))

#5
fit1=lm(Headwidth~log(Mass),data=ants)
fit2=lm(Headwidth~log(Mass)+Colony,data=ants)
fit3=lm(Headwidth~log(Mass)+Colony,data=ants)
fit4=lm(Headwidth~log(Mass)+Genus,data=ants)
new1=data.frame(Mass=40)
new2=data.frame(Mass=40, Colony="4")
new3=data.frame(Mass=40, Genus="1")
predict(fit1,new1)
predict(fit2,new2)
predict(fit3,new2)
predict(fit4,new3)
```
table(ants$Colony)
table(ants$Genus)
new=data.frame(  
  Mass=rep(with(ants,tapply(Mass,Colony,mean)),2),  
  Genus=rep(c("1","2"),each=10))  
predict(fit4, new)

#7  
table(ants$Genus,ants$Colony)
summary(fit1))

#8  
aova(fit1,fit2,fit3)  
aova(fit1,fit4,fit2)

#9  
fit0=lm(Headwidth~1, data=ants)  
aova(fit0,fit4)

#10  
summary(fit1)  
summary(fit2)  
summary(fit3)  
summary(fit4)

#11  
par(mfrow=c(2,2))  
plot(fit4)

#12  
fit5=lm(Headwidth~Mass+Genus, data=ants)  
plot(fit5)