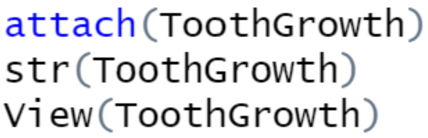
MTH 325, Lab #4, Spring 2023 Name ____

Instructions: Follow along with the tutorial portion of the lab. Replicate the code examples in R on your own, along with the demonstration. Then use those examples as a model to answer the questions/perform the tasks that follow. Copy and paste the results of your code to answer questions where directed. Submit your response file and the code used (both for the tutorial and part two). Your code file and your lab response file should each include your name inside. Be sure to follow the write-up directions in the Lab Directions file.

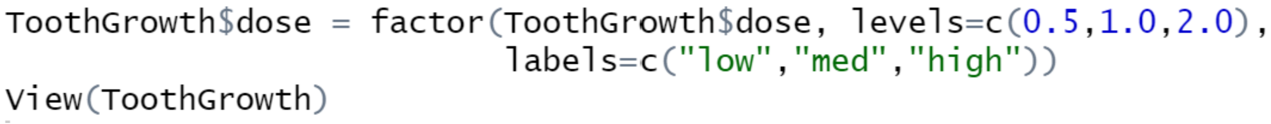
To start with, we will review ANOVA analysis in R, and we will compare the results to a linear regression model. To perform the analysis, we'll be using a dataset in R on tooth growth.



```
attach(ToothGrowth)
str(ToothGrowth)
View(ToothGrowth)
```

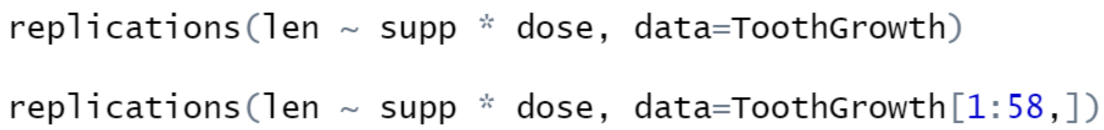
We begin by examining the data. The dataset has three variables. Len (length) is the variable we wish to predict. The two other variables are supp (supplement) with two levels VC (vitamin C) and OJ (orange juice), and dose that have three levels, 0.5, 1.0 and 2.0. To perform an ANOVA analysis, we'll need to convert the dose to a factor variable.

For the ANOVA analysis, we'll also relabel the levels of dosage to "low", "medium" and "high" labels.



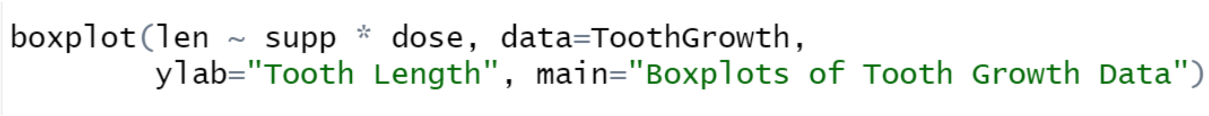
```
ToothGrowth$dose = factor(ToothGrowth$dose, levels=c(0.5,1.0,2.0),
                          labels=c("low","med","high"))
View(ToothGrowth)
```

We can check the number of replications for each variable and the interaction. What we want to ensure is that the observations are approximately equal to avoid the possibility of masking for events that have low numbers of representation in the data.



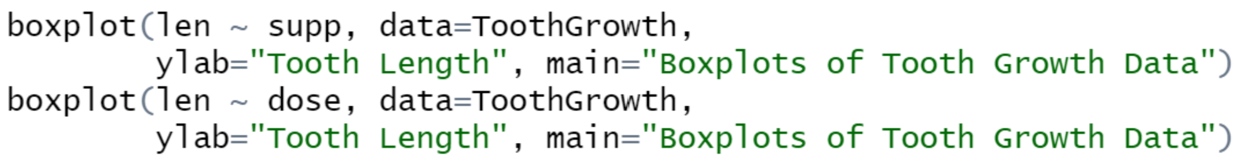
```
replications(len ~ supp * dose, data=ToothGrowth)

replications(len ~ supp * dose, data=ToothGrowth[1:58,])
```

Then we want to visualize the data to see if there is some potential for the variables to have an effect. We create a boxplot (this one uses base R). We can use this later to see if the model results agree with our initial assessment.



```
boxplot(len ~ supp * dose, data=ToothGrowth,
        ylab="Tooth Length", main="Boxplots of Tooth Growth Data")
```

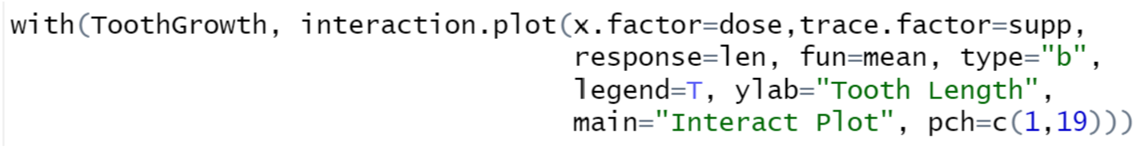
This plots the interaction groups. Plot each variable separately as well.



```
boxplot(len ~ supp, data=ToothGrowth,
        ylab="Tooth Length", main="Boxplots of Tooth Growth Data")
boxplot(len ~ dose, data=ToothGrowth,
        ylab="Tooth Length", main="Boxplots of Tooth Growth Data")
```

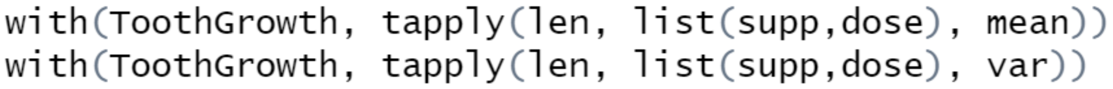
We want to use these graphs to assess whether the variances are equal within each variable. The spread of the boxplots should look similar within each graph.

We can make a plot of the interactions to see if there is an effect of the supplement type and the dosage on each other.



```
with(ToothGrowth, interaction.plot(x.factor=dose,trace.factor=supp,
                                   response=len, fun=mean, type="b",
                                   legend=T, ylab="Tooth Length",
                                   main="Interact Plot", pch=c(1,19)))
```

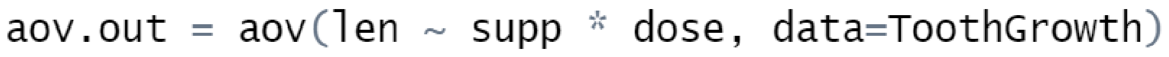
Before we perform the ANOVA, let’s look at a numerical summary of the data.



```
with(ToothGrowth, tapply(len, list(supp,dose), mean))
with(ToothGrowth, tapply(len, list(supp,dose), var))
```

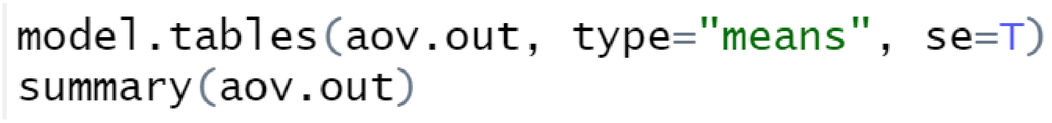
You can also examine other types of plots or numerical analyses such as normality plots (qqplots).

Finally, we are ready to look at the ANOVA analysis and the details of that model.



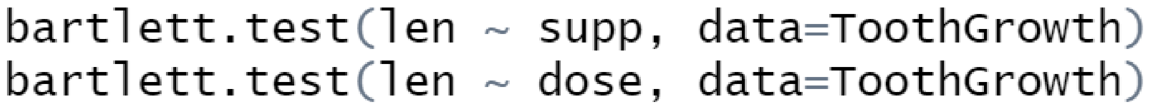
```
aov.out = aov(len ~ supp * dose, data=ToothGrowth)
```

We can look at the variable summaries (similar results to the output of the tapply functions above), and the model summary.



```
model.tables(aov.out, type="means", se=T)
summary(aov.out)
```

We can also conduct what is called a Bartlett test to determine if the variances are sufficiently equal to meet the model assumptions of our ANOVA test.



```
bartlett.test(len ~ supp, data=ToothGrowth)
bartlett.test(len ~ dose, data=ToothGrowth)
```

The null hypothesis for this test is that the variances are equal, so a high p-value is good for our model assumptions. A low p-value is bad.

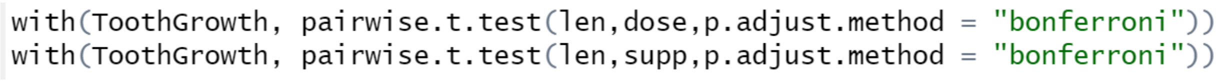
Get the model coefficients. Use this information to write an equation that models the response variable.

```
aov.out$coefficients
```

Perform your post hoc analysis such as residual plots, Tukey plots, etc.

```
TukeyHSD(aov.out)
plot(TukeyHSD(aov.out))
```

When conducting multiple hypothesis tests at the same time, some statistics are concerned about the possibility of the increasing risk of Type I errors just due to chance. There are various methods for accounting for this possibility using Bonferroni adjustments to the p-values.



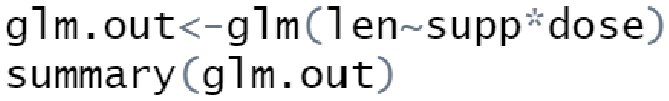
```
with(ToothGrowth, pairwise.t.test(len,dose,p.adjust.method = "bonferroni"))
with(ToothGrowth, pairwise.t.test(len,supp,p.adjust.method = "bonferroni"))
```

Does this analysis raise any potential concerns? If so, how significant are they?

To obtain the analysis plots, you can plot the model. Note: you'll need to press enter several times for all the plots to appear.

```
plot(aov.out)
```

Compare your ANOVA analysis to the general linear model of the same data.

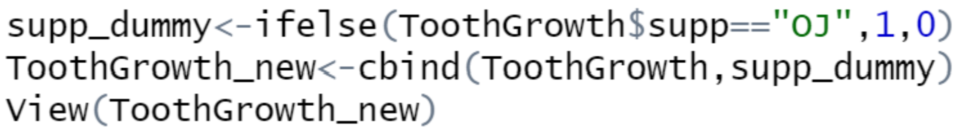


```
glm.out<-glm(len~supp*dose)
summary(glm.out)
```

We want to compare our model to a regression model. So, we will reload the data.

```
data("ToothGrowth")
View(ToothGrowth)
```

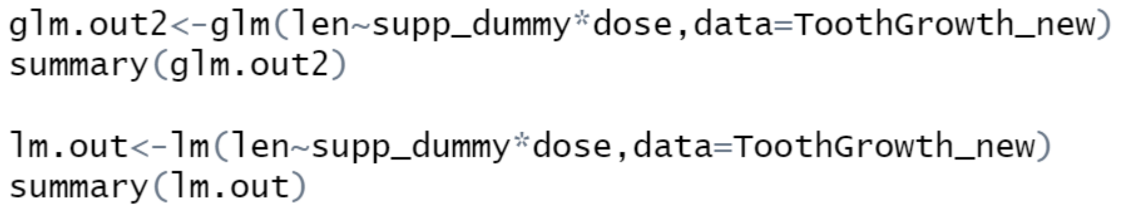
We will say more about using dummy variables in a future lecture, but for now, since supp (supplement) has only two values, we are going to replace those levels with 0 for VC and 1 for OJ. Since these will now be numbers, along with our original dosage levels, we'll be able to perform regression on the data. We do the replacement with the following commands.



```
supp_dummy<-ifelse(ToothGrowth$supp=="OJ",1,0)
ToothGrowth_new<-cbind(ToothGrowth,supp_dummy)
View(ToothGrowth_new)
```

This code creates a new dataframe which we'll use from now on. And we'll use supp_dummy in our analysis instead of the original supp variable.

Compare the results of the analysis using glm and lm. Is there a difference?



```
glm.out2<-glm(len~supp_dummy*dose,data=ToothGrowth_new)
summary(glm.out2)

lm.out<-lm(len~supp_dummy*dose,data=ToothGrowth_new)
summary(lm.out)
```

Tasks:

Continue your analysis of the ToothGrowth data. Create your model plots (residual plots, etc.) to test the assumptions of your regression model, following the examples in this lab and in previous regression labs. Describe your model equations for the ANOVA model and the linear regression model. How do they compare with each other, particularly with respect to interpretations and p-values. Write up your comparison of the two approaches to this data. Include graphs as appropriate to illustrate your process and analysis. Which model do you think works the best for this data? Why?

Resources:

1. https://homepages.inf.ed.ac.uk/bwebb/statistics/Factorial_ANOVA_in_R.pdf
2. Discovering Statistics Using R. Andy Field, Jeremy Miles, Zoe Field. (2012)
3. https://book.stat420.org/applied_statistics.pdf
4. https://scholarworks.montana.edu/xmlui/handle/1/2999
5. https://www.rstudio.com/resources/cheatsheets/
6. https://data-flair.training/blogs/hypothesis-testing-in-r/
7. https://www.scribbr.com/statistics/anova-in-r/
8. https://r-graph-gallery.com/84-tukey-test.html
9. https://www.real-statistics.com/one-way-analysis-of-variance-anova/basic-concepts-anova/
10. https://statisticsglobe.com/combine-two-vectors-into-data-frame-in-r
11. https://bookdown.org/steve_midway/DAR/understanding-anova-in-r.html