Practicum #3

Due: Friday, December 3 @ 6pm

In these exercises, we will be using the iris data built into R. Begin by loading the data:

```
data(iris)
```

You can read about it using `?iris`

Problem 1:

Explore the iris data:

a. What variables are present? Are they continuous or categorical?
b. For the categorical variables, provide a table of how many samples there are in each category. Show the R code you used. (*hint: check out the `table()` function*)
c. Provide the range (minimum, maximum) of the continuous variables.
d. Another useful exploratory tool is the `pairs()` function, which plots the columns of a data frame against each other. Produce a scatter plot matrix (a pairs plot) of the continuous variables.

Problem 2:

Do iris flower sizes depend on species?

a. In class, we learned about the formula syntax in R, which uses the tilde character `~` to mean "as a function of." Many functions use this syntax, including the `boxplot()` function. For example, you can generate a boxplot for two variables with `boxplot(y ~ x)`. Using this syntax (or `ggplot2` if you prefer), Produce boxplots of sepal width broken down by iris species.
b. For each iris species, compute the mean sepal width (Show your work.)
c. For each iris species, compute the sample SD of the sepal width (Show your work.)
d. Using your answers above, fill out a complete ANOVA table for sepal width vs. species. (Show your work.)
e. Based on your table, what is the standard deviation of sepal width for *all* species? How does it compare to `sd(iris$Sepal.Width)`?
f. Based on your table, and using the `p*` family of functions (such as `pnorm` and `pt`), compute a p value for the ANOVA. Also provide, in words, an interpretation of the result.
g. In class, we learned about using `anova(lm(Y~X,data=myData))` to obtain an ANOVA analysis of `Y` versus the categorical variable `X` using the dataframe `myData`. Use this syntax to perform the ANOVA analysis you did by hand above. Do the results agree with yours?
h. Remember, the ANOVA is a single test which tells us if the three species's distribution of sepal widths came from the same overall distribution or from two or more distributions. However, if we reject the null hypothesis, it does not tell us which species differ. Use a TukeyHSD test to perform pairwise t-tests between species to see which, if any, comparisons are significantly different.
i. Compare a regular t-test for sepal width of versicolor vs. virginica. How do the p-values differ? Was this what you expected?

Problem 3

ANOVA assumes that the groups have equal variances. Let's check that assumption.

a. Produce boxplots of sepal **length** broken down by iris species. By eye, do you believe the variances are equal?
b. Although we learned about using the F statistic to compare MS(between) and MS(within), we can use an F test for a variety of things. For example, if we wanted to know if the variance of the setosa and virginica sepal lengths are equal, we can use an F test for equality of variance:

$$F = \frac{Var(Y)}{Var(X)}$$

- Perform an F test to compare two variances using the `var.test()` function. What is your interpretation of these results? (*hint: supply `var.test()` with the Sepal.Length for one species as `x` and the other species as `y`*)

c. If the equal variance or normality assumptions of ANOVA are violated, an alternative is to use the Kruskal-Wallis test, which is a generalization of the rank-sum test to > 2 groups. You can carry out a Kruskal-Wallis test in R using `kruskal.test()`. Read the help page and apply it here (*hint: you will not need `lm()`!*) – what do you conclude?
d. If (c) produced a significant result, do you think a TukeyHSD test would be applicable here to see which species are significantly different? Why or why not? If not, suggest an alternative. (You don't need to do the test)

Problem 4:

Next we'll consider the relationship between sepal length and sepal width

a. Refer to the scatterplot matrix you produced in problem 1(d). Do you expect sepal length and sepal width to be positively correlated, negatively correlated, or uncorrelated?
b. Compute the correlation between sepal length and sepal width. Are your expectations confirmed?
c. Now compute the correlation between sepal length and sepal width separately for each of the iris species. How do these compare to the overall correlation you observed in part (b)?
d. Using R's `lm()` function, fit a linear model that predicts sepal length as a function only of sepal width and assign it to `fit1`.
e. Repeat part (d), but now include an additional term to model the difference in average sepal length between species, assigning it to `fit2`.
f. As above, but now include the interaction between sepal width and species. Assign this to `fit3`.
g. Print out the summary of `fit3` and interpret the results. What are the results telling you?
h. Look at the summaries for `fit1`, `fit2`, and `fit3`. The "Multiple R-squared" value gives the square of the correlation between the observed dependent variable y_i (in this case, sepal length) and the estimated $\hat{y_i}$ from your model. We can use the squared correlation between them as a measure of how well our model fits. Which gives the best R^2?
i. Generally, adding more covariates will tend to produce a higher R^2. That doesn't necessarily mean it's a better model, however – you may be overfitting the data! A better way to select a model is to test if the variance explained by the larger model compensates for the degrees of freedom you introduce by adding covariates. Try `anova(fit1,fit2)`. Does `fit2` account for significantly more variance than `fit1`? What about `fit3`? Based on the ANOVA comparisons, which model would you choose?