Lab 6: Drawing conclusions in linear regression models

Create a short reproducible report answering the questions below, The report should be less than 3 pages, including all figures, and should be submitted as both PDF and Rmd formats. You do not need to show your code in the PDF report.

This lab is due at 5pm on Thursday, November 17th. You should submit your assignment, in the form of both a knitted RMarkdown PDF as well as the .Rmd file that created the PDF, by uploading them to your personal Google Drive folder that is shared with the TA and the instructor. While you may collaborate with other students on this assignment, you must write up your own code and answers to the questions. Absolutely no cutting and pasting of any portion of the answers. This assignment, like the others, will be worth 50 points.

Introduction to Hypothesis Testing

A central question in regression analysis is "How can we formally evaluate statistical evidence about associations between the observed variables, while controlling for the presence of other factors?"

One important method to evaluating evidence is using confidence intervals. This approach allows us to give an interval of possible values (often centered on our "best guess" of a parameter) that we think is likely to cover the true value. However, another common approach used to assess models is to use hypothesis testing. Hypothesis testing is described in nice intuitive detail in Kaplan, Chapters 13-15, and this lab will highlight some of the key points covered in those chapters.

Hypothesis testing relies on a certain logic that enables us to weigh evidence present in our dataset. In particular, we end up calculating p-values, which represent the probability of observing the data that we did observe, given that a particular hypothesis (our null hypothesis, or H_0) is true. Oftentimes, statistical theory can gives us tidy results that help us determine what the world would look like if H_0 were true. Othertimes, we might need to rely on our intuition and computation to show us what the H_0 version of the world would look like.

Getting started

Let's use the FEV data again and focus on the relationship betwen the fev variable with the age, height, and smoke covariates. To start, let's load the data into our current R session:

```
library(Hmisc)
library(mosaic)
library(ggplot2)
getHdata(FEV)
```

Exercise 1 Let's say that we are interested in exploring the model that uses age, smoke, height and sex to predict the outcome. Fit this model. Interpret the coefficients. For now, don't give a confidence interval or interpret our uncertainty.

```
fm <- lm(fev ~ age + smoke + height + sex, data=FEV)
```

Exercise 2 We are interested in testing a hypothesis about whether smoking is associated with fev. One way to address this question is to set up a hypothetical world where we know that smoking does not have an association with fev time. One way to achieve this is to randomly shuffle or permute the values of the covariate. By randomly shuffling these values, we are breaking any association that might exist. Test out this shuffling method and see for yourself what the data looks like when we

assign a random value of smoking from our dataset to each net time. First off, what is the relationship between FEV and smoking status in the real data? Explain the trend that you see in the data. Second, describe the changes you see when you shuffle smoking status.

```
qplot(smoke, fev, data=FEV) + geom_boxplot()
qplot(shuffle(smoke), fev, data=FEV) + geom_boxplot()
```

Exercise 3 We know that by "shuffling" smoking status, we can create single datasets that represent a reality where smoking is not related to any other of the variables, including fev. But this is a noisy process (remember, *data collection is where the randomness happens*), and each dataset will look different. Use the `shuffle()` function and the `do()` syntax from Lecture 9 to run a simulation in which you fit a set of regression models from Exercise 1 where smoke has been shuffled. Try do-ing this for different numbers of times. Pick a number of repetitions (N) that is big enough so that when you repeat the experiment the results don't appreciably change, but that runs pretty quickly overall. Hint: by "results don't change" you could think about plotting the estimates of one of the coefficients.

```
s <- do(100) * lm(fev ~ age + shuffle(smoke) + height + sex, data=FEV)
ggplot(s) + geom_density(aes(smokecurrent.smoker))
sum(abs(s$smokecurrent.smoker)>=abs(coef(fm)["smokecurrent smoker"]))
```

Exercise 4 In the last exercise you created what we could call a "null distribution" of coefficient estimates. This means that you created a distribution of what the coefficients might have looked like if the null hypothesis, i.e. that $\beta_{smoke} = 0$ or that smoking is unrelated to fev, were true. Plot this null distribution and compare it to the value of the coefficient from when you fit the model to the original data. Write code that calculates the percentage of the null distribution that is further from zero (in absolute value) than the fitted coefficient from the original dataset. That number has a name! It's a p-value from what is called a permutation test. Congratulations! You've just performed a hypothesis test. Interpret the results of your test and make a statement about the evidence for or against smoking having an effect on FEV after controlling for the other variables in the model.

Exercise 5 There is another way to perform the same hypothesis test, weighing evidence for or against the impact of smoking on FEV. This uses the R output from the original fitted model and relies on the statistical approximation to the sampling distribution of β_{smoke}. Essentially, we can assume that the sampling distribution for β_{smoke} under the null hypothesis is a Student's t distribution, with mean zero and standard deviation equal to the standard error for the coefficient reported by R (see Kaplan Ch 15.5). So when you type `summary(fm)` and you look at the coefficient table, you will see a p-value for each coefficient separately. Compare the p-value in this table to the p-value you obtained in Exercise 4. How similar/different are they?

Exercise 6 State in 1-2 sentences your conclusions from the 'canned' analysis performed in R using the `lm()` function. Is there evidence that smoking impacts FEV, once adjusting for other varibles? Justify your conclusions using a confidence interval (hint: you can use the `confint()` function) and a p-value. Are these conclusions similar to your conclusions from your permutation-based analysis of this data from Exercise 3 and 4? If not, discuss how your permutation-based model was different and why that might have contributed to different results.

Extra Credit

Read Kaplan sections 15.2 and 15.3 and consider the following output from these two analysis of variance tables in R:

```
anova(fm)
fm1 <- lm(fev ~ age + height + sex + smoke, data=FEV)
anova(fm1)
```

The mathematical equations for these two models are equivalent, so why does smoke have a different p-value in these two models? Which do you think better represents the true association of smoking with FEV?