**Psychology 610/710**

**HW 2, Week 2**

Save your answers to the data analysis in your R script, adding comments where appropriate. Save your answers to the reading questions in a Microsoft Word document (or similar format). Email to the TAs by 5:00pm Wednesday. Use naming format “Lastname\_HW#”

**Data analysis**

For this homework, we will use the “stray\_animals.csv” file.

This dataset contains (fictitious) statistics about the number of stray animals (in thousands) in each of the 50 US states. The dataset contains information about the number of stray dogs, cats, and other animals measured in year 1, as well as information about the number of stray cats and dogs measured in year 5, after a nation-wide TNR (trap-neuter-return) program was launched in an attempt to reduce the number of stray cats. Finally, the dataset also includes information about the number of animal shelters per 10,000 people in each state.

1. Read in the data and inspect them (summaries, structure, descriptive stats).
2. Create an appropriate plot that shows whether the number of stray dogs in year one varies as a function of the number of animal shelters in the state. Copy and paste this graph into your word document.
	1. Add a best fitting line.
	2. Does the relationship appear positive or negative? Strong or weak?
3. Go through the following steps to create a composite score that reflects the overall numbers of stray animals, referred to as the “stray index” in these questions:
	1. Standardize the measures for stray dogs in year 1, stray cats in year 1, and other stray animals in year 1, individually.
	2. Check to make sure you standardized these variables correctly.
	3. Before making a composite score, check the reliability. We need to know whether it makes sense to combine these three variables and use them as a single measure. Comment on the reliability. Should we consider dropping any of these groups from the stray index? Which one(s)? For now, don’t drop any.
	4. Create a new variable (Strays1) representing the average of these standardized scores.
	5. Comment on the researchers’ decision to standardize then average these scores. Does their decision make sense or not? How does this measure differ (in practical terms) from one in which the researchers average the scores without standardizing them? Type your answer to these questions in your R-script (as comments).

Now the researchers want to investigate if the number of stray cats decreased from year 1 to year 5.

1. Create a histogram of the number of stray cats in year 1. Label your histogram properly. Do the same thing for the number of stray cats in year 5. Copy and paste the graphs into your word document.
2. Create a variable that represents the change in number of stray cats from year 1 to year 5. Name this variable “CatsChange.”
3. What is the formula for each of the two models you should compare if you want to know whether the change in the number of stray cats is different from zero? Type your answers to this question in your R-script (as comments).
4. Compute the SSE for each model using a step-by-step brute force method, like we did in lab. Remember to label clearly (SSE0, SSE1). Make sure your script gives the output of the SSE so I can see it.
5. Compare the two models by calculating their respective *F* statistics and *p* values (also the slow version like we did in lab). Which model is better? Why? Interpret the result in plain English (i.e., what does the result say regarding the effectiveness of the program?).
6. Even though there was no targeted program to change the number of stray dogs between year 1 and year 5, the researchers are curious to see if the number of stray dogs changed during that period. Create a variable to represent this change in the number of stray dogs.
7. Using the lm command, estimate the simple model (mean-only) for the change in the number of stray dogs.
	1. What does the b0 value represent?
	2. What does the significance test of the b0 mean?
	3. This command tells R to do a model comparison. What are the two models being compared?
	4. What can we conclude based on the results of our test?

**Reading Questions**

Answer these questions in your Word document.

1. You are using a database of pet food consumption to determine how much food cats eat per day, on average. You notice that there are a number of households containing multiple cats. What core model assumption is being violated here? How could we fix this problem?
2. Define Type I and Type II error. For each type, give an example of how a researcher could reduce the risk of committing the error.
3. In practical terms, what does the numerator of the *F* statistic equation represent? The denominator?

In the word document, please tell us how long it took you to complete this assignment.