**THE2**

**Assigned: 17 November 2017, 5:00pm**

**Due: 27 November 2017, 9:00am**

Take-home exam 2! You really made it. Remember, you are allowed and encouraged to refer to prior lab scripts, homework assignments, books/articles, and the internet. **You may *not* consult with other students.** Use the same general format as you do for homework assignments (see template document on the website). A few things to note:

* **Submit your files anonymously.** Instead of using the normal Lastname\_HWXX format, use the **last 4 digits of your student ID number** (e.g., “3180\_THE1.R”). *Include no identifying information in your files!*
* The only things that should appear in your Word doc are your write-ups and your figures (no reading questions this time because we’re so nice). Everything else can just be typed in your R script.
* Follow course norms for naming variables and models (or use names that will be crystal clear to us).
* You don’t need to include the question text next to the question number in your R script; the number is plenty for us.
* You can email us asking questions, but we won’t be as instructive as we would be on a homework assignment. This is an exam, after all.

Hopefully you’ll be giving thanks for what an enjoyable and interesting exam this is. Given the holiday, we *strongly recommend* that you do your best to complete the exam before break begins. A good approach might be aiming to finish by next Wednesday, then do a final check and submit on Sunday (this could allow you to catch any random errors). I think you will strongly overestimate your motivation to work on a take-home exam over a long weekend…

**Data Analysis**

**Study 1: Cats vs Dogs (42 points)**

A researcher is interested in settling an age-old debate: which animal is smarter, cats or dogs? She conducts a study to answer this question. She conceives of a novel task that can physically be solved by both animals. Furthermore, her paradigm allows her to create many different types of problems of varying difficulty for the animals to solve. She recruits cats and dogs from the Madison area to participate in the study. Because she is also curious which animal can learn better, she is including a condition variable: animals assigned to the Learning condition go through 10 trials of the task with feedback (i.e., they either receive a treat or not) before the testing phase, while those in the Control condition do not receive these feedback trials. The testing phase of the experiment involves 40 permutations of the task, and each pet is assigned a score out of 40 accordingly, based on the total number they are able to solve.

She has the following hypotheses:

1. Cats will perform better on the task than dogs overall.
2. This relationship will be qualified by an interaction, such that dogs will benefit more from the training than cats.

Codebook for Study 1

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Description** | **Values** |
| PetName | The name of the pet participating in the study. | Unique for each of X animals |
| Species | Whether the pet is a dog or a cat | 0 = dog 1 = cat |
| Condition | Whether the pet is in the learning or the control condition | 0 = control 1 = learning |
| Score | Score on the task | 0 – 40 |

1. Read in the data and perform some univariate and bivariate analyses. Write a couple sentences about these stats to give you a general sense of the data. [2]
2. Run a (single) command that shows the mean score broken down by species and condition. Do these means seem consistent with the hypotheses, at first glance? [1]
3. Run an additive model with the variables of interest. Briefly summarize the output of this model (still include appropriate statistics). [2]
4. Use this model to conduct a case analysis.
   1. Does it make sense look at leverage here? Why or why not? [1]
   2. Examine model outliers. Identify all points that are problematic here (list them by name, if any). For each you identify, comment briefly on why they are outliers, in terms of their values on the given variables. [2]
   3. Look at influence. What points have high influence on our parameter estimates? For each, identify whether they seem to be affecting a particular parameter. [2]
   4. Which pets should be removed, and why? For each, give some hypothetical reason why they might have obtained the score they did. [3]
   5. Store a copy of the complete data frame, then remove the individuals you’ve identified. From now on, use the data frame with the outliers removed. [1]
5. Process your variables as necessary, then run the focal interactive model to test the researcher’s hypotheses. [3]
   1. Do the data support hypothesis 1? Why or why not? Summarize briefly, including appropriate statistics. [3]
   2. Do the data support hypothesis 2? Create a quick and dirty plot to double-check before you write your answer. Then summarize briefly, including appropriate statistics. [4]
6. Test the effect of Species for pets in the Training condition. Summarize the result in real-world terms, including appropriate stats. [3]
7. Create a bar plot to display your results. Paste this graph into your Word doc and name it “Figure 1.” Brownie points for individuals who use interesting (i.e., not default) colors to display this information. [7]
8. Write a brief results section in your Word doc. No need to go into detail about outlier analyses (though you will want to mention them) or your test from Question 6, but do include some orienting information at the beginning and a one-sentence conclusion at the end (i.e., make sure it passes Mitch’s “makes sense to a random stranger” test). [8]

**Study 2 (58 points)**

A graduate student (cough cough) has started baking as a form of stress relief. As his skills have improved, he has begun to wonder about the science behind the perfect pie, and decides to upend his research program and completely refocus to the qualities of a good pastry treat. With grant money from NIH (curiously), he hires renown cookery writer Mary Berry (of *Great British Bake Off* fame) to be the judge of the baked goods. For his initial investigation, the grad student decides to focus on seasonally-appropriate pumpkin pie. He varies the amount of fat (i.e., butter, lard, or margarine) used in the pastry shell, the amount of pumpkin pie spice added to their filling mixture, and the balance of sweetened condensed milk and canned pumpkin (see codebook). He only uses amounts he can find in existing recipes, so no pies are particularly extreme on any dimension. The grad student also notes whether the pie has a “soggy bottom” (i.e., the crust underneath the filling is moist and undercooked rather than thin and crispy). Due to the large number of variables, he bakes X pies in order to have good statistical power. Mary Berry is paid to taste each of these pies, and rates them on three variables: how “scrummy” they are, how “beautiful” they are, and how “dodgy” they are (reverse-coded). The scale is anchored at 1 and 7, and precise to the tens place between these (e.g., a score of 2.3 is possible)

The grad student has the following hypotheses:

1. More fat in the crust will have a positive effect on Mary Berry’s rating, but this will only be the case if there is no soggy bottom.
2. More spice in the filling will have a positive effect on Mary Berry’s rating, and this effect will be stronger for pies with less canned pumpkin/more sweetened condensed milk.

Codebook for Study 2

|  |  |  |
| --- | --- | --- |
| **Variable name** | **Description** | **Range** |
| PieID | ID of a given pie | 1 – 128 |
| Fat | Amount of fat used in the crust, in oz | 2 – 12 |
| Spice | Amount of pumpkin pie spice used, in tsp | 0 – 4 |
| Pumpkin | Amount of canned pumpkin used, in oz | 12 – 20 |
| SCM | Amount of sweetened condensed milk used, in oz | 12 – 20 |
| SgyBtm | Whether the pie has a “soggy bottom” | 0 = does not 1 = does |
| Scrummy | How “scrummy” Mary considers the pie to be | 1 – 7 |
| Beauty | How “beautiful” Mary considers the pie to be | 1 – 7 |
| Dodgy | How “dodgy” Mary considers the pie to be | 1 – 7 |

1. Read in the data and perform some univariate and bivariate analyses. Write a couple sentences about these stats to give you a general sense of the data. [2]
2. The graduate student cares more about the balance of pumpkin to sweetened condensed milk (SCM) than he does the raw amount of each.
   1. This will be very clear to you when you do two things: compute the correlation between them and look at the values you’d obtain if you added them together for each pie. How would you describe how these variables are related? [2]
   2. Make a new variable called “Filling” by subtracting the amount of SCM from the amount of pumpkin. Report the minimum and maximum values of this new variable and what these values mean *in plain English*. [2]
   3. To avoid later confusion, delete variables “Pumpkin” and “SCM.” [1]
3. That’s not the only data processing to be done. We also need to know whether Mary’s rating scale has good reliability.
   1. Check the alpha value, bearing in mind that “dodgy” is reversed scored. Comment briefly on your result. [2]
   2. Combine all three variables into one composite score called “Rating” and comment on the univariate stats for this measure. [2]
4. Run an *additive* model predicting Rating from Filling and Spice. Use this model to complete model assumptions. [2]
   1. Check the three model assumptions we can examine using the modelAssumptions command and briefly comment on each. [5]
   2. Based on these results, would you suggest a transformation? If you were to transform, which variable would make the most sense, based on its univariate distribution and its definition? [2]
   3. What Box-Cox value does R recommend for transforming this variable? In plain English, how would this transformation affect the scores? [1]  
        
      *Continue* ***without*** *doing a transformation of this variable.*
5. Process your variables as necessary, then run the proper interactive model to test hypothesis 1. [3]
   1. Are the data consistent with hypothesis 1? Why or why not? Make a quick and dirty plot to reinforce your conclusion. Summarize briefly, including statistics. [4]
   2. Is there a positive effect of fat for pies with a soggy bottom? [3]
6. Process your variables as necessary, then run the proper interactive model to test hypothesis 2. [3]
   1. Are the data consistent with hypothesis 2? Why or why not? Make a quick and dirty plot to reinforce your conclusion. Summarize briefly, including statistics. [4]
   2. Interpret *b1, b2,* and *b3* in plain English, just one sentence for each parameter estimate. [3]
7. Choose one of these hypotheses and create a publication-quality graph. Make an informed choice about which variable to put on the x axis and which to use to define the lines. Plot 2-3 lines at meaningful or useful values of the variable used to define the lines. Paste this graph in your Word doc and label it “Figure 2.” Brownie points for individuals who use interesting (i.e., not default) colors to display this information (or other interesting and helpful formatting choices). [7]
8. Write a brief results section. Set up the study, briefly comment on your model assumptions process (just 1-2 sentences), report results from your tests of the researchers’ hypotheses, and write a simple conclusion at the end. Like before, make sure the write-up passes Mitch’s “makes sense to a random stranger” test. [10]