Lab 7: Quant by Cat Interaction

SDS358: Applied Regression Analysis

Michael J. Mahometa, Ph.D.

.

> "There are more things in heaven and earth, Horatio, than are dreamt of in your [statistical model]."
>
> *Shakespeare – Hamlet*

Introduction

The basic idea of Lab is as follows: Answer a research question with the provided dataset. Each week, that research question (and data) will change depending on the topic we've covered the prior class days. Once we're done with Lab, you'll have a Lab Assignment, that will look a lot like the Lab: a research question you'll need to answer given some data. In Lab, you'll learn the procedure for answering the research question. For the Lab Assignment, you'll do that procedure for a grade (independently).

To help answer the research question, we'll follow some basic steps that we'll repeat throughout the semester:

- Reflect on the Question: Figure out the variables of interest, and the technique that's required.
- Analyze the Data: Perform the steps required for the technique.
- Draw Conclusions: Use the information that you got from the prior step to answer the research question in a concise, logical manner.

Let's get started:

Primary Research Question:

Among single adult learners participating online courses at a major university, when controlling for Age and Gender, does Job Type significantly moderate the effect of Social Support on overall Happiness?

Step1: Reflect on the Question:

Download the syntax and data files from Canvas.

Let's load in our SDSRegressionR package so that we can use some of it's functions later:

```
#Load our class package
library(SDSRegressionR)
```

Next, we'll load in the data. Be sure to use the basic file structure we talked about the first Lab: Put your syntax in a folder specific to this Lab. Then, make a "data" folder in that same place - use lowercase. If you do that, then all of this syntax will work like a charm.

For this lab, we'll need to subset:

```
work <- read_csv("data/workers.csv")
sing <- work %>%
  filter(Marital.status == "Single")
```

Check the Data:

To make sure that we're working with the right data, and that we're all looking at it the same way, we'll answer some basic questions about the data before moving on:

1. How many observations are in the dataset for the model ("**sing**")?

2. What was the Happiness score for the first participant 30 or older?

3. Of the first 10 participants, how many had a Religiosity score under 50?

These questions can be answered simply by looking at the dataset once it's loaded in:

```
View(sing)
```

Check the Variables of Interest

Let's find the variables that we need to answer the primary research question:

1. Which variable tells us the Happiness of a participant?
   - What type of variable is this?
   - What scale is this variable on?
2. What is the "variable of interest" for this model?
   - What is the moderator for this mode?
3. List *all* the variables that will go into this model

Again, these can be answered by looking at the dataframe, and with the help of the *names()* function. Also, the codebook for the data frame is our friend. You can open this in R or Excel. Remember, R is case-sensitive.

For **categorical data** we *really* need to look at the codebook. Some variables may be coded as 0/1, and we won't really see that by looking at the data alone. We can also run `table()` and `levels()` to help with this also.

```
names(sing)
```

```
##  [1] "SubID"             "Age"               "Female"
##  [4] "College.grad"      "Have.child"        "Marital.status"
##  [7] "Job"               "Religiosity"       "Social.support"
## [10] "Life.satisfaction" "Happiness"         "Stressors"
## [13] "Depression"
```

```
#Gender
table(sing$Female)
```

```
##
##  0  1
## 12 79
```

```
#Job Type
table(sing$Job)
```

```
##
##        Academic    Professional SupportServices
##              45              21              25
```

Reflect on the Method

The last part of Reflect on the Question asks about the method or technique we'll use.

1. We will use Multiple Linear Regression with an interaction to answer this Lab question. Why?
2. We'll need to dummy code a categorical variable. Why?
3. We'll use Anova (type III) impact of the overall interaction. Why?
4. We will need to run _____ regression models to fully explain the simple slopes of the interaction. Why?

Step2: Analyze the Data

In this step, we'll run the provided syntax and answer some questions about the output to help us prepare for the final step.

Here's the syntax you'll need (from the .R syntax file):

```
#### Here is the R script you will use:  (remember that # indicates a comment) ####
#Lab7: Categorical Interaction

library(SDSRegressionR)

#Import data...
work <- read_csv("data/workers.csv")
sing <- work %>%
  filter(Marital.status == "Single")

#Examine the categorical variable(s):
#Remember, it's good to do this for ALL categorical variables
table(sing$Female)
table(sing$Job)

#Run the Factoring for ALL categorical variables:
# (Male as Reference)
# (SupportServices as reference)
sing <- sing %>%
  mutate(Gender = factor(Female, levels=c(0,1), labels=c("Male", "Female")),
         JobType = factor(Job, levels=c("SupportServices", "Academic", "Professional")))

#Initial Model:
#Run the model (SupportServices as reference)
hap <- lm(Happiness ~ Age + Gender + JobType + Social.support +
            JobType*Social.support, data=sing)
summary(hap)

#Check the diagnostics/outliers
residFitted(hap)
library(car)
vif(hap)
c <- cooksPlot(hap, key.variable="SubID", print.obs=TRUE, sort.obs=TRUE, save.cutoff=TRUE)
c
threeOuts(hap, key.variable="SubID")

#Get good data
g_sing <- sing %>%
  filter(SubID %not in% c(...))

#Re-run the model:
hap2 <- lm(Happiness ~ Age + Gender + JobType + Social.support +
            JobType*Social.support, data=g_sing)
summary(hap2)

#Check the overall interaction significance
library(car)
```

```r
Anova(hap2, type="III")

#Simple Slopes
library(emmeans)
ref_grid(hap2)
emmeans(hap2, "Social.support", at=list(Social.support = c(0,1)), by="JobType")
job_means <- emmeans(hap2, "Social.support", at=list(Social.support = c(0,1)), by="JobType")
job_means #just easier to read

#Test of Simple Slopes
pairs(job_means, reverse=TRUE)
job_slopes <- pairs(job_means, reverse=TRUE)
job_slopes

# (difference of differences -- Interaction terms)
pairs(update(job_slopes, by=NULL), reverse=TRUE, adjust="none")

#CI Plot (for fun)
library(emmeans)
mns <- summary(emmeans(hap2, "Social.support",
                       at=list(Social.support = seq(0,60,1)), by="JobType"))
simpleScatter(g_sing, x=Social.support, y=Happiness, ptalpha = 0,
              title="Social Support and Happiness",
              subtitle = "by Employment Group") +
  geom_line(data=mns, aes(x=Social.support, y=emmean, color=JobType)) +
  geom_ribbon(data=mns, aes(y=emmean, ymin=lower.CL, ymax=upper.CL, group=JobType),
              alpha=0.3) +
  #Change to your group names and number of groups
  scale_colour_manual(name = "Groups",
                      values =c("blue", "red", "green"),
                      #IMPORTANT: Same order below as the factor()
                      labels = c("SupportServices", "Academic", "Professional"))

#Or a straight ggplot...but this falls a little short...
ggplot(g_sing,aes(x=Social.support,y=Happiness,color=JobType)) +
  stat_smooth(method=lm, se = FALSE, fullrange=TRUE) +
  geom_point() +
  labs(title="Job Type Interaction", x="Social.support", y="Happiness") +
  theme_bw()
```

Question 1

How many dummy variables were needed for the categorical variable? How many interaction terms were needed in the model?

Question 2

Because we had no observations that were outliers on *all three* diagnostic tools, we took all observations that had a Cook's Distance greater than $(4/n - k - 1)$; number of observations: _____. These outliers represented _____% of the original data.

Question 3

Job Type _____ moderated the relationship between Social Support and Happiness: F(_____, _____) = _____, p _____ 0.05.

Quesiton 4

The simple slope of Social.support for the Support group was _____ , t(_____) = _____, p _____ 0.05.

Quesiton 5

The simple slope of Social.support for the Academic group was _____ , t(_____) = _____, p _____ 0.05.

Quesiton 6

The simple slope of Social.support for the Professional group was _____ , t(_____) = _____, p _____ 0.05.

Question 7

The impact of Social.support on Happiness between the Support and the Professional group was _____ different (t(_____) = _____, p _____ 0.05).

Question 8

The impact of Social.support on Happiness between the Support and the Academic group was _____ different (t(_____) = _____, p _____ 0.05).

Question 9

The impact of Social.support on Happiness between the Professional and the Academic group was _____ different (t(_____) = _____, p _____ 0.05).

Step3: Draw Conclusions

The final step is for us to Draw Conclusions. We'll take the syntax we've been given from Analyze the Question, run it, then examine the output. The questions from the prior step help set us up for the Draw Conclusions part.

We'll "fill in the blanks" in a canned paragraph for the Lab. For the Lab Assignment, you'll need to come up with a similar paragraph all on your own (please don't steal mine).

Primary Research Question
Among single adult learners participating online courses at a major university, when controlling for Age and Gender, does Job Type significantly moderate the effect of Social Support on overall Happiness?

> Our primary research question investigated the possible moderation of Job Type on the relationship between Social support and Happiness, after controlling for age and gender. The overall model was significant, F(_____, _____) = _____, $p < 0.05$, R^2 = _____. There was a significant interaction between Job Type and Social Support, F(_____, _____) = _____, $p < 0.05$, indicating that Job Type moderated the relationship between Social Support and Happiness. Evaluation of simple slopes showed a significant impact of Social Support on Happiness for _____ (b = _____, t(_____) = _____, $p < 0.05$) and _____ (b = _____, t(_____) = _____, $p < 0.05$), but not _____ (b = _____, t(_____) = _____, p = _____).
>
> The impact of Social support on Happiness between the Academic and Support group was _____ different (t(_____) = _____, p _____ 0.05), as well as between the Academic and Professional group (t(_____) = _____, p _____ 0.05). However, there was no difference in impact between the Support and Professional groups (t(_____) = _____, p = _____).

Lab Assignment

Now, with the tools at your disposal (the R syntax from Lab, and the logic of proceeding through the three steps of answering the research question), you'll have a Lab Assignment to complete (independently). For now, the Lab Assignment is to be completed in Canvas. It will follow the basic structure, and lead to the same place - answering the research question with a concise paragraph as in Draw Conclusions.

Good Luck!