Lab 6: Categorical Predictors

SDS358: Applied Regression Analysis

Michael J. Mahometa, Ph.D.

.

> ”A little [statistics] is a dangerous thing. Drink deep, or taste not the Pierian spring.”
>
> *Alexander Pope*

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, Gender, Child status, Job Type, and Social Support, what predictors significantly impact the outcome of 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:

```r
#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 mean Happiness score for the first participant 30 or older?

3. Of the first 10 participants, how many had a Social Support score under 20?

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 are the independent variables for the research question?
3. Which of the control variables for the research question are categorical?

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() to help with this as well.

```
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"
```

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 to answer this Lab question. Why?

2. We'll need to dummy code a categorical variable. Why?

3. We'll use Sequential Regression to evaluate the impact of the categorical variable. Why?
4. We will also need to apply a correction to the post-hoc evaluations for the categorical variable. 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) ####
#Lab6: Categorical Variables

install.packages("devtools") #if needed...
devtools::install_github("MichaelJMahometa/SDSRegressionR", force=TRUE)

library(SDSRegressionR)

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

#Examine the categorical variable:
table(sing$Job, useNA = "always")

#Recode into dummy variables:
#Job Type
sing <- sing %>%
  mutate(Academic = case_when(Job == "Academic" ~ 1,
                              Job != "Academic" ~ 0),
         Professional = case_when(Job == "Professional" ~ 1,
                                  Job != "Professional" ~ 0),
         SupportServices = case_when(Job == "SupportServices" ~ 1,
                                     Job != "SupportServices" ~ 0))

#Run the model with DUMMIES (SupportServices as reference)
hap <- lm(Happiness ~ Age + Female + Have.child + Academic + Professional +
             Social.support, data=sing)
summary(hap)

#Run the model with FACTOR VARIABLE (SupportServices as reference)
sing <- sing %>%
  mutate(Job_f = factor(Job, levels=c("SupportServices", "Academic", "Professional")))

hap_f <- lm(Happiness ~ Age + Female + Have.child + Job_f + Social.support, data=sing)
summary(hap_f)

#Check the model...
library(car)
vif(hap_f)
residFitted(hap_f)
cooksPlot(hap_f, key.variable = "SubID", print.obs=TRUE, sort.obs = TRUE)
threeOuts(hap, key.variable = "SubID")
```

```r
#Drop the outliers
g_sing <- sing %>%
  filter(SubID %not in% c(...)) #Let's discuss

#Rerun
hap2_f <- lm(Happiness ~ Age + Female + Have.child + Job_f + Social.support, data=g_sing)
summary(hap2_f)

lmBeta(hap2_f)
pCorr(hap2_f)

#Overall ANOVA
library(car)
Anova(hap2_f, type="III")

#Job R^2
lmSingleR2(hap2_f, "Job_f")

#Post-hoc exploration
library(emmeans)
hap2_f_mn <- emmeans(hap2_f, "Job_f")
hap2_f_mn
pairs(hap2_f_mn, adjust="none") #Too little
pairs(hap2_f_mn, adjust="bonferroni") #Too much
pairs(hap2_f_mn, adjust="holm") #Just right
```

Question 1

The table() function was used on a categorical predictor. How many levels of this variable are there?________ This will require the use of ________ dummy coded variables to be created.

Question 2

What was the purpose of creating three dummy variables, when we only need ________? (Fill in the blank and answer resulting question.)

Question 3

We removed ________ outliers, because of ________.

Question 4

The categorical variable of "Job" showed a ________ impact on the prediction of Happiness: F(________, ________) = ________, p ________ 0.05.

Quesiton 5

This categorical variable required ________ post-hoc comparisons.

Question 6

Bonferroni-Holm adjusted post-hoc comparisons showed a _______ difference between Academic and Professional outcomes (t(_______) = _______), a _______ difference between Academic and SupportServices outcomes (t(_______) = _______), and a _______ difference between Professional and SupportServices outcomes (t(_______) = _______).

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).

Among single adult learners participating online courses at a major university, when controlling for Age, Gender, Child status, Job Type, and Social Support, what predictors significantly impact the outcome of overall Happiness?

> Our primary research question investigated the predictive impact of several items on the outcome of Happiness among single adult learners. These underlying items included age, gender, child status, job type, and social support. The overall model was significant, F(________, ________) = ________, $p < 0.05$. Of the predictor variables in the model, job type, age, and social support were ________ predictors of Happiness. Social support showed a significant positive impact on quality of life (t(________) = ________, $p < 0.05$), and could account for ________ % of unique variance in the model. Job type also showed a significant impact (F(________, ________) = ________, $p < 0.05$), and accounted for ________ % of the variance in Happiness. Age, although significant and positive, only accounted for ________ % of the variance in happiness.
>
> Bonferroni-Holm adjusted post-hoc investigations of Job type showed that there was a significant difference in Happiness between Academic and Support Service participants (t(________) = ________, $p < 0.05$) as well as between Academic and Professional participants (t(________) = ________, $p < 0.05$). However, there was no significant difference in Happiness between Professional and Support Service participants (t(________) = ________, $p < 0.05$).

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!