Lab 15 - Multivariate Regression & Interpretation

Snow Christensen

November 30, 2017

Complete the following exercises below and include all code used to find the answers. Knit together the PDF document and commit both the Lab 15 RMD file and the PDF document to Git. Push the changes to GitHub so both documents are visible in your public GitHub repository.

1. Select a second explanatory variable from your dataset that you think has implications for the theoretical association of your focal relationship.

a. Describe the theoretical reasoning for selecting this variable.

I want to look at the effect of race on the association between NOASSERT and AUTHSI because race affects how people are socialized and also how they are percieved and/or attacked by others.

b. What type of relationship do you think this variable has with your focal variables? Given that, what do you expect to happen to your focal relationship when it is added to the model?

I've looked into this relationship a little before but I would be interested to see more clearly what this relationship is with the addition of race as a category for analysis. Based on my past analyses I do not think race will have a strong effect but I want to know for sure.

c. Is it a continuous or categorical variable? What implications does this have for a multivariate regression equation?

It is a categorical variable so I will need to make sure it is coded as a factor so it will be used correctly in the regression.

d. Conduct a multivariate linear regression with this additional explanatory variable and save the model object. Print out the full results by calling `summary()` on your model object.

```
library(haven)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#load full dataset
data <- read_por("H:/Projects/sexual_violence_college-master/ICPSR_03212/DS0001/03212-0001-Data.por")

#data subset
data_ID <- data %>%
  dplyr::select(CODENUM, RACE, NOASSERT, PLEASE, PRESSSI, AUTHSI)


#convert variables to numerics
data_ID$RACE <- as.numeric(data_ID$RACE)
data_ID$NOASSERT <- as.numeric(data_ID$NOASSERT)
```

```r
data_ID$PRESSSI <- as.numeric(data_ID$PRESSSI)
data_ID$PLEASE <- as.numeric(data_ID$PLEASE)
data_ID$AUTHSI <- as.numeric(data_ID$AUTHSI)

#create multivariate regression with NOASSERT and AUTHSI controlling for RACE
MR.lm <- lm( NOASSERT ~ AUTHSI + RACE ,
data = data_ID )

#print summary of multivariate regression
summary(MR.lm)
```

```
##
## Call:
## lm(formula = NOASSERT ~ AUTHSI + RACE, data = data_ID)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1595 -0.7204 -0.5216  0.2796  3.4784
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.48005    0.19341   7.652 3.47e-14 ***
## AUTHSI       0.43912    0.18201   2.413    0.016 *
## RACE        -0.19878    0.04388  -4.530 6.35e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9023 on 1531 degrees of freedom
##   (46 observations deleted due to missingness)
## Multiple R-squared:  0.01656,    Adjusted R-squared:  0.01528
## F-statistic: 12.89 on 2 and 1531 DF,  p-value: 2.803e-06
```

e. Describe the results of the multivariate analysis, highlighting:

- the apparent association between the control variable and the focal response variable

The association between the control variable and the focal response variable is negative and not very strong because the coefficient is only -0.19878.

- how the focal association changed when you incorporatedthe control variable The focal association increased by about .02, but other than that it is the same.
- the implications of these results for your focal association The implications do not change much in terms of my analysis of the focal association, except for the fact that I can now rule out race as a large contributor to the association.

f. How well does this model fit the data? Is it an improvement over the bivariate model? Why or why not?

The data fits about the same as it did in the bivariate model. It is better generally in the sense that we are accounting for race so we can be sure of the effects of race on this association. However, at the same time it is not that useful because the association is pretty much the same as it was in the bivariate model but now there are more variables in this regression.

2. Select any additional variables you want to incorporate into your final model. For each additional variable added to the model answer the following questions:

a. Describe the theoretical reasoning for selecting this variable.

I want to look at the effect of race on the association between PLEASE and PRESSSI because race affects how people are socialized and also how they are percieved and/or attacked by others.

b. What type of relationship do you think this variable has with your focal variables? Given that, what do you expect to happen to your focal relationship when it is added to the model?

I've looked into this association a little before but I would be interested to see more clearly what this relationship is with the addition of race as a category for analysis. Based on my past analyses I do not think race will have a strong effect but I want to know for sure.

c. Is it a continuous or categorical variable? What implications does this have for a multivariate regression equation?

It is a categorical variable so I will need to make sure it is coded as a factor so it will be used correctly in the regression.

d. Conduct a multivariate linear regression by adding one explanatory variable at a time and save the model objects. Print out the full results by calling `summary()` on each model object.

```
#create multivariate regression with NOASSERT and AUTHSI controlling for RACE
MR.lm2 <- lm(PLEASE ~ PRESSSI + RACE ,
data = data_ID )

#print summary of multivariate regression
summary(MR.lm2)
```

```
##
## Call:
## lm(formula = PLEASE ~ PRESSSI + RACE, data = data_ID)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1387 -0.9386 -0.1387  0.3712  3.6810
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.04825    0.10089  20.303  < 2e-16 ***
## PRESSSI      0.20010    0.06195   3.230  0.00126 **
## RACE        -0.30979    0.04918  -6.299 3.91e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.007 on 1530 degrees of freedom
##   (47 observations deleted due to missingness)
## Multiple R-squared:  0.03111,    Adjusted R-squared:  0.02984
## F-statistic: 24.56 on 2 and 1530 DF,  p-value: 3.159e-11
```

e. Describe the results of the multivariate analysis, highlighting:

- the apparent association between each additional control variable and the focal response variable

The association between the control variable and the focal response variable is negative with a slope of -0.30979 so it is not very strong.

- how the focal association changed when you incorporated each control variable

The focal association is stronger now because the coefficient increased by about .1 however it is still not very strong.

- the implications of these results for your focal association The results of this are important because they address the potential impact of race on the association, however, they do not change the relationship between focal explanatory and response variables that much.

f. How well does the full (all explanatory variables included) model fit? Are any of the other models you ran a better fit? Explain how you came to the conclusion you did.

The data fits about the same as it did in the bivariate model. It is better generally in the sense that we are accounting for race so we can be sure of the effects of race on this association. However, at the same time it is not that useful because the association is pretty much the same as it was in the bivariate model but now there are more variables in this regression.

g. Select the model that you think best fits the data. Provide a brief synopsis of the analysis of your data using this model and describe the implications for the theoretical arguments you set out to test.

```
#create multivariate regression with NOASSERT and AUTHSI controlling for RACE
MR.lm2 <- lm(PLEASE ~ PRESSSI + RACE ,
data = data_ID )

#print summary of multivariate regression
summary(MR.lm2)
```

```
##
## Call:
## lm(formula = PLEASE ~ PRESSSI + RACE, data = data_ID)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1387 -0.9386 -0.1387  0.3712  3.6810
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.04825    0.10089  20.303  < 2e-16 ***
## PRESSSI      0.20010    0.06195   3.230  0.00126 **
## RACE        -0.30979    0.04918  -6.299 3.91e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.007 on 1530 degrees of freedom
##   (47 observations deleted due to missingness)
## Multiple R-squared:  0.03111,    Adjusted R-squared:  0.02984
## F-statistic: 24.56 on 2 and 1530 DF,  p-value: 3.159e-11
```

I think the multivariate regressions best fit the data because they provide the same information as the bivariate regression except I am also accounting for race, which is always an important factor. In this model we can see that there is a positive association between PLEASE and PRESSSI, although surprisingly weak, and that there is actually a slightly negative association between RACE and PRESSSI and this association does not affect the association between PLEASE and PRESSSI. This is important for attempting to find a pattern for who is being targetted by sexual predators. It helped disprove my original hypotheses that 1) characteristic traits such as PLEASE and NOASSERT would play a large role in determining who was victimized, and 2) RACE of the victim would play a large role in who was being targetted.These findings also help disprove some of the stereotypes surrounding women and sexual assault, especially when discussing the concept of victim blaming.