Assignment 4.5

Cory J. Cascalheira

03/24/2021

**Important Note**: I retained information from the previous assignments given that we are building upon each step to create a more complex multiple regression model. Answers to the current assignment appear at the bottom of this document.

# Assignment Instructions

Question: Do males and females differ in their confidence in math and does this have different impacts on their math achievement?

For this time around, we are going to add the applied math variable to the set of predictors as well. The applied math variable might be referred to as “opportunity to learn” (OTL) because it asked students how frequently they experienced responding to applied math tasks in their classroom (1=frequently,2=sometimes,3=rarely,4=never). (we are adding this variable now, but will not pay attention to it much at the moment, but I want you to be familiar with it because it will appear in your midterm).

To understand if gender interacts with confidence, first create the interaction term.

IN R: This is a bit more simple. create each model separately and store them

model1<-lm(PV1MATH~ESCS+CONF+FEMALE+ENJOY+LANG+OTL, [put whatever your dataset is named here])

model2<-lm(PV1MATH~ESCS+CONF+FEMALE+ENJOY+LANG+OTL+FEMALE\*CONF, [put whatever your dataset is named here])

To compare their R2 values, look at the summary of each model separately. To look at whether model 2 is statistically significantly improved over model 1, run anova(model1,model2). To get a nice readout of the comparison of each model, load the sjPlot library, and run tab\_model(model1,model2). By the way, R outputs interaction variables by naming them FEMALE:CONF (the colon indicating that the two are multiplied together).

**Answer these questions**:

1. Does adding the interaction term (FEMALE\*CONFIDENCE) statistically significantly improve the model?
2. Does including the interaction term cause any multicollinearity? If so, what would your recommendation be to do to fix it?
3. Regardless of whether including the term matters, interpret the interaction term’s coefficient (if you have run things correctly, it should be about 6.2467). Make a table like we did in class to help you with this.

# Establish the Work Environment

Load all the dependencies and import the data.

# Dependencies  
library(car)  
library(psych)  
library(ppcor)  
library(tidyverse)  
library(rio)  
library(sjPlot)  
  
# Import  
pisa <- import("../data/2012 PISA multiple countries selected variables.sav") %>%  
 as\_tibble  
  
# Construct the APA theme for plots  
# Construct the APA theme  
apa\_theme <- theme\_bw() +  
 theme(panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank(),  
 panel.border = element\_blank(),  
 axis.line = element\_line(),  
 plot.title = element\_text(hjust = 0.5),  
 text = element\_text(size = 12, family = "sans"),  
 axis.text.x = element\_text(color = "black"),  
 axis.text.y = element\_text(color = "black"),  
 axis.title = element\_text(face = "bold"))

## Recode Data

First, I recoded the data to facilitate coding and interpretation.

# Rename pisa  
pisa\_1 <- pisa %>%  
 rename(country = CNT, language = ST25Q01, enjoy\_math = ST29Q04,  
 gender = ST04Q01, math\_career = ST48Q05, applied\_math = ST76Q01,  
 confidence = ST37Q05, math\_score = PV1MATH, ses = ESCS)  
  
# Recode values  
pisa\_2 <- pisa\_1 %>%  
 mutate(  
 # 0 = same; 1 = different  
 language = recode(language, `1` = 0, `2` = 1),  
 # 0 = male; 1 = female  
 gender = recode(gender, `1` = 1, `2` = 0)#,  
 # Had to remove these recoded variables because they affect the VIF  
 # Recode enjoy math so increasing numbers mean increased enjoyment  
 #enjoy\_math = recode(enjoy\_math, `1` = 4, `2` = 3, `3` = 2, `4` = 1),  
 # Recode confidence so increasing numbers mean increased confidence  
 #confidence = recode(confidence, `1` = 4, `2` = 3, `3` = 2, `4` = 1)  
 )  
pisa\_2

## # A tibble: 65,535 x 9  
## country math\_score ses language enjoy\_math gender math\_career applied\_math  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 CAN 492. 0.93 NA NA 1 NA 3  
## 2 CAN 394. -0.78 0 NA 1 NA 3  
## 3 CAN 390. -1.3 0 4 0 2 2  
## 4 CAN 504. 0.56 0 3 1 2 3  
## 5 CAN 466. -0.03 0 2 0 1 NA  
## 6 CAN 398. 0.74 0 4 1 2 2  
## 7 CAN 404. NA NA NA 1 NA NA  
## 8 CAN 406. -2.58 0 1 1 2 NA  
## 9 CAN 609. 0.88 0 1 0 1 NA  
## 10 CAN 452. 0.44 0 4 1 2 NA  
## # ... with 65,525 more rows, and 1 more variable: confidence <dbl>

## Assumption Checking

Other major assumptions (e.g., normality, linearity) were assumed met since we have worked with this dataset several times before.

## Model Comparison

First, we must construct the models to compare. The models must have the same number of participants.

# Select variables of interest  
pisa\_compare <- pisa\_2 %>%  
 select(math\_score, ses, gender, language, enjoy\_math, confidence, applied\_math)  
  
# Remove missing values  
pisa\_compare\_1 <- na.omit(pisa\_compare)  
  
# Model 1  
model\_1 <- lm(math\_score ~ ses + gender + language + enjoy\_math + confidence + applied\_math,   
 data = pisa\_compare\_1)  
  
# Summarize Model 1  
summary(model\_1)

##   
## Call:  
## lm(formula = math\_score ~ ses + gender + language + enjoy\_math +   
## confidence + applied\_math, data = pisa\_compare\_1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -336.6 -52.2 0.9 52.0 301.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 522.8504 2.2424 233.169 < 2e-16 \*\*\*  
## ses 31.6648 0.4577 69.178 < 2e-16 \*\*\*  
## gender -8.0879 1.0907 -7.415 1.26e-13 \*\*\*  
## language 6.6796 1.8646 3.582 0.000341 \*\*\*  
## enjoy\_math -0.4734 0.6478 -0.731 0.464876   
## confidence -32.4355 0.7125 -45.525 < 2e-16 \*\*\*  
## applied\_math 3.9555 0.7124 5.552 2.85e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 76.98 on 20050 degrees of freedom  
## Multiple R-squared: 0.3213, Adjusted R-squared: 0.3211   
## F-statistic: 1582 on 6 and 20050 DF, p-value: < 2.2e-16

# Model 2  
model\_2 <- lm(math\_score ~ ses + gender + language + enjoy\_math + confidence + applied\_math + confidence\*gender,   
 data = pisa\_compare\_1)  
  
# Summarize Model 2  
summary(model\_2)

##   
## Call:  
## lm(formula = math\_score ~ ses + gender + language + enjoy\_math +   
## confidence + applied\_math + confidence \* gender, data = pisa\_compare\_1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -335.74 -52.10 0.87 51.85 307.07   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 528.2939 2.5155 210.014 < 2e-16 \*\*\*  
## ses 31.7191 0.4576 69.313 < 2e-16 \*\*\*  
## gender -18.6301 2.4664 -7.553 4.42e-14 \*\*\*  
## language 6.6808 1.8636 3.585 0.000338 \*\*\*  
## enjoy\_math -0.5676 0.6478 -0.876 0.380893   
## confidence -35.5315 0.9640 -36.859 < 2e-16 \*\*\*  
## applied\_math 4.0070 0.7121 5.627 1.86e-08 \*\*\*  
## gender:confidence 6.2467 1.3110 4.765 1.90e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 76.94 on 20049 degrees of freedom  
## Multiple R-squared: 0.3221, Adjusted R-squared: 0.3219   
## F-statistic: 1361 on 7 and 20049 DF, p-value: < 2.2e-16

Adding the interaction term increased the multiple from 32.13% (model 1) to 32.21% (model 2). In other words, the model with the interaction term explained an additional .08% of the variance in math scores.

Now we can compare the models.

# Are the models significantly different from one another?  
anova(model\_1, model\_2)

## Analysis of Variance Table  
##   
## Model 1: math\_score ~ ses + gender + language + enjoy\_math + confidence +   
## applied\_math  
## Model 2: math\_score ~ ses + gender + language + enjoy\_math + confidence +   
## applied\_math + confidence \* gender  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 20050 118821312   
## 2 20049 118686902 1 134410 22.705 1.902e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Yes, the ANOVA test revealed that the models are significantly different from one another, *F*(1, 20,049) = 22.705, *p* < .001.

Find the adjusted multiple R-squared, but do not print the results due to a formatting issue.

# Show table in R studio but not in the Word document  
tab\_model(model\_1, model\_2)

### Collinearity

Collinearity is a subset of multicollinearity applied to two predictors (Field et al., 2012). Multicollinearity occurs when two or more predictors are highly correlated within a regression model. As values of collinearity increase, problems are introduced into the model, such as:

* unstable beta weights, which might fit the sample but not represent the population;
* less contribution to multiple correlation (i.e., *R*) and variance explained because collinear variables share too much common variance; and
* difficulty assessing which predictors are most important.

The variance inflation factor (VIF) is a measure of “whether a predictor has a strong linear relationship with other predictor(s)” (Field et al., 2012, p. 276). Values of 10 are problematic. If the average VIF is greater than 1, then multicollinearity is likely a problem.

Tolerance, according to Field and colleagues, is the reciprocal of VIF. Problems occurs when tolerance is less than 0.1 and multicollinearity is more likely a problem when tolerance is less than 0.2.

Here on the collinearity statistics for model 1.

# Check the variance inflation factor (VIF)  
vif(model\_1)

## ses gender language enjoy\_math confidence applied\_math   
## 1.090256 1.004645 1.011493 1.133422 1.182644 1.010739

# Average VIF  
mean(vif(model\_1))

## [1] 1.0722

# Tolerance  
1 / vif(model\_1)

## ses gender language enjoy\_math confidence applied\_math   
## 0.9172157 0.9953767 0.9886372 0.8822836 0.8455628 0.9893749

None of the above values reached the thresholds suggested by Field et al. (2012), therefore multicollinearity is likely not a problem.

Here are the collinearity statistics for model 2.

# Check the variance inflation factor (VIF)  
vif(model\_2)

## ses gender language enjoy\_math   
## 1.090933 5.142571 1.011493 1.134478   
## confidence applied\_math gender:confidence   
## 2.167259 1.010973 6.072169

# Average VIF  
mean(vif(model\_2))

## [1] 2.518554

# Tolerance  
1 / vif(model\_2)

## ses gender language enjoy\_math   
## 0.9166468 0.1944553 0.9886372 0.8814625   
## confidence applied\_math gender:confidence   
## 0.4614122 0.9891464 0.1646858

Collinearity is a problem in model 2 because the VIF = 16.95 for gender and VIF = 17.74 for the interaction term; both are greater than the recommended threshold of 10 (Field et al., 2012). The tolerance statistic yields further evidence of collinearity. The tolerance for gender (.059) and the interaction term (.056) are less than 0.1, indicating collinearity.

# Questions for Assignment 6

## Does adding the interaction term (FEMALE\*CONFIDENCE) statistically significantly improve the model?

The ANOVA test revealed that the models are significantly different from one another, *F*(1, 20,049) = 22.705, *p* < .001. Adding the interaction term increased the multiple from 32.13% (model 1) to 32.21% (model 2). In other words, the model with the interaction term explained an additional .08% of the variance in math scores. The interaction term improves the model statistically, but its practical significance is not tenable. Indeed, the statistical significance is likely attributable to the large sample size (i.e., the analysis is overpowered).

## Does including the interaction term cause any multicollinearity? If so, what would your recommendation be to do to fix it?

Without recoded variables of confidence and applied math, the answer is no! I need to discuss this problem with Chris. I think the issue is that, by recoding, I was basically treating the levels of confidence and applied math as reversed scored items, which does not make sense and changes the meaning of the variables. I was trying to fix the directionality so that higher scores indicated more confidence but, instead, created an issue.

Here is my previous answer for sake of completeness:

Yes. Collinearity is a problem in model 2 because the VIF = 16.95 for gender while the VIF = 17.74 for the interaction term; both are greater than the recommended threshold of 10 (Field et al., 2012). The tolerance statistic yields further evidence of collinearity. The tolerance for gender (.059) and the interaction term (.056) are less than 0.1, indicating collinearity.

Given that the model with the interaction term evinces no practically significant improvement, I would eliminate the interaction term and retain model 1 for the sake parsimony. Another option would be the removal of the gender variable. If I were interested in the interaction of gender and confidence for theoretical reasons, then there may be an argument for this second recommendation.

## Regardless of whether including the term matters, interpret the interaction term’s coefficient (if you have run things correctly, it should be about 6.2467). Make a table like we did in class to help you with this.

Model 2 is given by the following equation:

MathScore = 360.40 + (31.72 \* SES) + (-12.60 \* Gender) + (6.68 \* Language) + (0.57 \* EnjoyMath) + (29.28 \* Confidence) + (4.01 \* AppliedMath) + (6.25 \* Gender \* Confidence)

Note that I recoded confidence such that 1 = low confidence and 4 = high confidence. Note that the table below include gender even though gender was not a significant predictor in the model (which I would remove prior to interpreting the interaction coefficient for publication).

# Pull the essential variables  
variables <- model\_2$coefficients[c(3,6,8)]  
variables

## gender confidence gender:confidence   
## -18.630123 -35.531502 6.246745

data.frame(  
 Confidence = c("Not at all", "Not very", "Confident", "Very confident"),  
 Male = c(variables[2] \* 1, variables[2] \* 2, variables[2] \* 3, variables[2] \* 4),  
 Female = c(variables[1] + (variables[2] \* 1) + (variables[3] \* 1 \* 1),   
 variables[1] + (variables[2] \* 2) + (variables[3] \* 1 \* 2),   
 variables[1] + (variables[2] \* 3) + (variables[3] \* 1 \* 3),   
 variables[1] + (variables[2] \* 4) + (variables[3] \* 1 \* 4))  
)

## Confidence Male Female  
## 1 Not at all -35.5315 -47.91488  
## 2 Not very -71.0630 -77.19964  
## 3 Confident -106.5945 -106.48440  
## 4 Very confident -142.1260 -135.76915

In summary, confidence matters more for male students than for female students because greater confidence increases male math scores more than it increases female math scores.

# References

Field, A., J. Miles, & Z. Field. (2012). *Discovering statistics using R*. SAGE Publications Ltd.