Assignment 4

Cory J. Cascalheira

03/04/2021

# Assignment Instructions

Please complete the following:

1. Conduct a linear regression model for the following: PV1MATH = Intercept + ESCS\*X1 + Gender\*X2 + LANG\*X3 + ENJOYMATH\*X4
2. When you conduct the linear regression, under Options, check the box next to Collinearity Diagnostics
3. Include interpretation of the collinearity diagnostics (VIF and Tolerance) in your write up.

I will be posting a video soon about what these mean. They are also discussed in Field (2018).

# Establish the Work Environment

Load all the dependencies and import the data.

# Dependencies  
library(car)  
library(psych)  
library(tidyverse)  
library(rio)  
  
# 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 facilicate coding and interpretation.

# Rename pisa  
pisa\_1 <- pisa %>%  
 rename(country = CNT, language = ST25Q01, enjoy\_math = ST29Q04,  
 gender = ST04Q01, math\_career = ST48Q05, applied\_math = ST76Q01,  
 solve\_equation = ST37Q05, math\_score = PV1MATH, ses = ESCS)  
  
# Recode dichotomous 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),  
 # Recode enjoy math so increasing numbers mean increased enjoyment  
 enjoy\_math = recode(enjoy\_math, `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 1 0 2 2  
## 4 CAN 504. 0.56 0 2 1 2 3  
## 5 CAN 466. -0.03 0 3 0 1 NA  
## 6 CAN 398. 0.74 0 1 1 2 2  
## 7 CAN 404. NA NA NA 1 NA NA  
## 8 CAN 406. -2.58 0 4 1 2 NA  
## 9 CAN 609. 0.88 0 4 0 1 NA  
## 10 CAN 452. 0.44 0 1 1 2 NA  
## # ... with 65,525 more rows, and 1 more variable: solve\_equation <dbl>

## Assumption Checking

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

### 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.

First, we need to specify the model.

# Create the model  
model <- lm(math\_score ~ ses + gender + language + enjoy\_math, data = pisa\_2)

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.

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

## ses gender language enjoy\_math   
## 1.019888 1.005030 1.006297 1.018768

# Average VIF  
mean(vif(model))

## [1] 1.012496

# Tolerance  
1 / vif(model)

## ses gender language enjoy\_math   
## 0.9805001 0.9949950 0.9937428 0.9815776

None of the above values reached the thresholds suggested by Field et al. (2012), therefore multicollinearity is likely not a problem. The average VIF is barely greater than 1.

# Interpretation

# Summarize the model  
summary(model)

##   
## Call:  
## lm(formula = math\_score ~ ses + gender + language + enjoy\_math,   
## data = pisa\_2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -378.80 -55.64 -0.97 54.88 335.62   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 456.6267 1.2423 367.579 < 2e-16 \*\*\*  
## ses 36.4991 0.3229 113.047 < 2e-16 \*\*\*  
## gender -7.2765 0.8016 -9.077 < 2e-16 \*\*\*  
## language 10.0581 1.3610 7.390 1.49e-13 \*\*\*  
## enjoy\_math 7.1521 0.4429 16.148 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 81.56 on 41676 degrees of freedom  
## (23854 observations deleted due to missingness)  
## Multiple R-squared: 0.2407, Adjusted R-squared: 0.2407   
## F-statistic: 3304 on 4 and 41676 DF, p-value: < 2.2e-16

# Confidence intervals  
confint(model)

## 2.5 % 97.5 %  
## (Intercept) 454.191858 459.061551  
## ses 35.866242 37.131888  
## gender -8.847711 -5.705252  
## language 7.390562 12.725737  
## enjoy\_math 6.284013 8.020233

Results from the multiple regression were significant, *F*(4, 41,676) = 3,304, *p* < .001, and the model explained 24.07% of the variance in math achievement. All of the variables were significant predictors. Socioeconomic status (*t* = 113.05, *p* < .001) significantly contributed to the variance in math achievement, indicating that for every one unit increase in standardized socioeconomic status, a student’s score on the math achievement assessment would increase by 36.5 points (95% CI [35.87, 37.13]). Gender (*t* = 9.08, *p* < .001) and language (*t* = 10.06, *p* < .001) were also significant predictors, but the practical significance was limited: being female (versus male) was associated with a 7.28 point increase in math achievement (95% CI [5.71, 8.85]) and having a native language that was different from the test (versus the same) was associated with a 10.06 point increase in math achievement (95% CI [7.39, 12.73]). Finally, enjoyment of math exhibited a significant relationship with math achievement (*t* = 16.15, *p* < .001), indicating that participants tended to score 7.15 points higher (95% CI [6.28, 8.05]) for every one unit increase in math enjoyment. Again, math enjoyment is not considered practically significant given the scale of math achievement.

# References

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