Assignment 2

Cory J. Cascalheira

02/03/2021

# Assignment Instructions

In a Word document, respond to the following and upload via Canvas:

1. Download the 2012 PISA multiple countries selected variables.sav file and open it in SPSS. Then, read through the 2012 PISA Math subset Codebook.docxPreview the document file to make sure you understand what all of the variables mean.
2. Write up a response to the following question, using a linear regression: What influence, if any, does socioeconomic status (ESCS) and gender (ST04Q01) have on students’ mathematics achievement (PV1MATH)?
   * Are assumptions of linear regression met?
     + Were data randomly collected?
     + Are individuals independent?
     + Are both the dependent variables (DV) and the independent variable (IV) continuous variables?
     + Are both DVs and the IV normally distributed?
     + Are residuals normally distributed?
     + Is there essentially no relationship between standardized residuals and the predicted variable?
   * Is the overall model statistically significant? (report ANOVA results, R, R2–we will talk more about this)
   * Are the ESCS and gender coefficients statistically significant? (report coefficient estimate and interpret it’s meaning, t statistic, p value, confidence interval)
     + If it is statistically significant, is it practically significant?
   * How would you answer the question?

# 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"))

## Assumption Checking

An evaluation of assumptions is helpful.

### Were data randomly collected?

Yes. [According to OECD](https://www.oecd.org/pisa/pisaproducts/SAMPLING-IN-PISA.pdf), the organization that administers the PISA, a stratified two-stage sample design is used and students have equal probability within schools to be part of the sample.

### Are individuals independent?

Theoretically, we can assume that the scores on one test do not autocorrelate with the scores on a subsequent test. For instance, assuming that students are not cheating, then Student A’s answers probably do not depend on Student B’s answers. At the country level, we would expect that the scores for each country do not automatically correlate with the scores of countries who took the test first.

Notably, however, we might expect that scores within the same school to be related because the students engaged in a similar curriculum taught by a set of teachers. Similarly, we might expect scores to be related within countries. For example, the Common Core standards for math education likely operate as an outside influence on student math achievement within the United States (i.e., all students received a similar education). Nonetheless, given the random selection of students, there is greater evidence of independent observations (Rovai et al., 2014).

Another method of verifying this assumption is the Durbin Watson test (Field et al., 2012).

First, we need to transform gender into a dummy variable. Next, we can define the model. Finally, we can execute the Durbin Watson test.

# Gender  
pisa$gender <- pisa$ST04Q01 - 1  
table(pisa$gender)

##   
## 0 1   
## 33714 31821

# Define the linear regression  
model <- lm(PV1MATH ~ ESCS + gender, data = pisa)  
  
# Check independence of observations  
durbinWatsonTest(model)

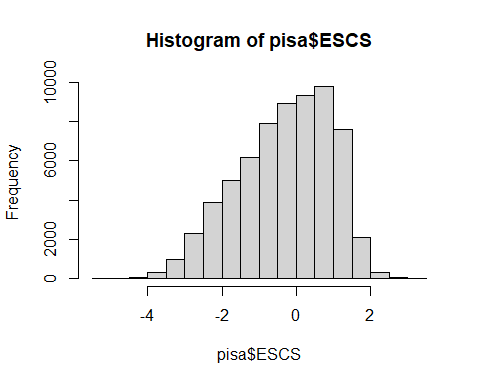
## lag Autocorrelation D-W Statistic p-value  
## 1 0.2730039 1.453991 0  
## Alternative hypothesis: rho != 0

Given that the D-W statistic is closer to 1 than 2 (i.e., 1.45) and the test was significant (Field et al., 2012), there is evidence of autocorrelation and, thus, the assumption of independent observations is not tenable.

### Are both the dependent variables (DV) and the independent variable (IV) continuous variables?

This assumption is not exactly tenable. Socioeconomic status and math achievement scores are continuous, but gender is dichotomous. However, multiple regression can still be executed with dummy coded dichotomous values (Rovai et al., 2014). Therefore, the variables are appropriate for this model after a transformation of gender.

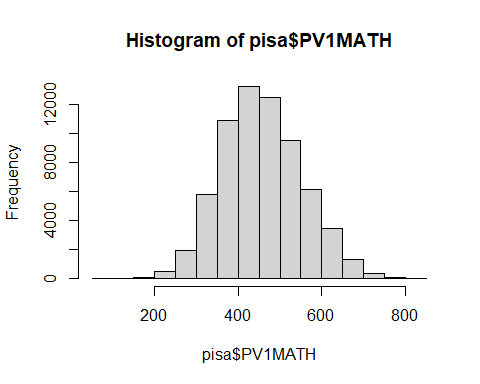
# ESCS evidence of continuous level of measurement   
hist(pisa$ESCS)



length(unique(pisa$ESCS))

## [1] 697

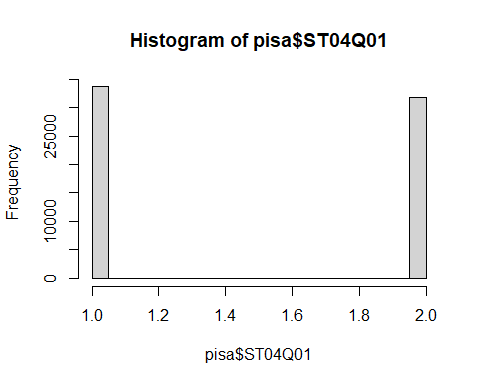
# PV1MATH evidence of continuous level of measurement   
hist(pisa$PV1MATH)



length(unique(pisa$PV1MATH))

## [1] 5075

# ST04Q01 evidence of discrete level of measurement   
hist(pisa$ST04Q01)



length(unique(pisa$ST04Q01))

## [1] 2

### Are both DVs and the IV normally distributed?

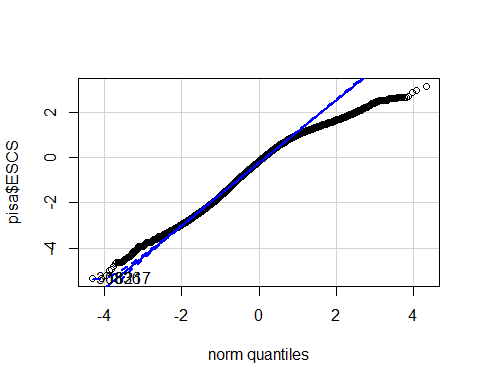
The variables are probably normal enough given the large sample size.

That is, the histograms above provide visual evidence for and against normality. Gender is a binomial distribution and is neither symmetrical nor bell-shaped; however, dichotomous predictors can be used in multiple regression when they are dummy coded to 0 and 1 (Rovai et al., 2014). Socioeconomic status appears negatively skewed, but is probably normal enough given that the sample size is large (*N* = 65,535). Math achievement scores appear symmetrical and bell-shaped.

### Are residuals normally distributed?

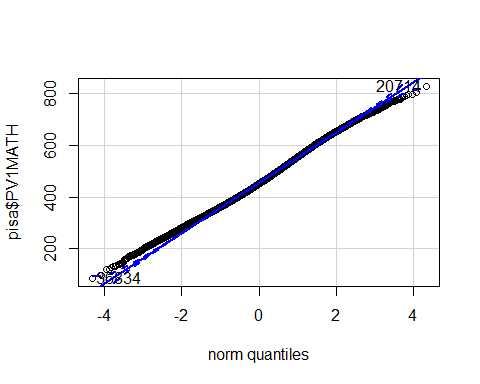
Based on a visual inspection of the Q-Q plots, the residuals of: \* ESCS are not normally distributed; \* PV1MATH are normally distributed; and \* ST04Q01 are not normally distributed.

# Q-Q plot of ESCS  
qqPlot(pisa$ESCS)



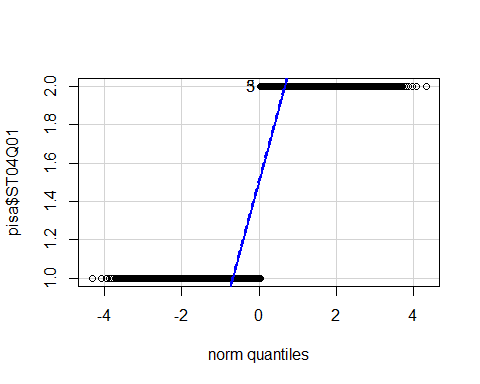
## [1] 30836 18217

# Q-Q plot of PV1MATH  
qqPlot(pisa$PV1MATH)



## [1] 20714 36834

# Q-Q plot of ST04Q01  
qqPlot(pisa$ST04Q01)



## [1] 3 5

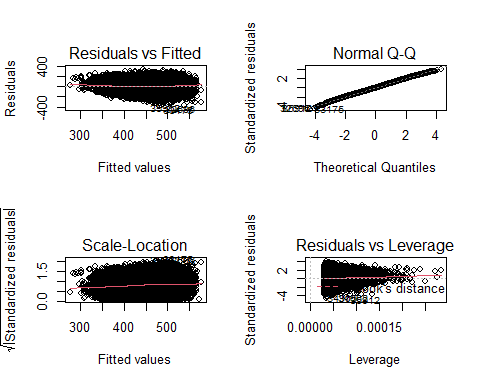
### Is there essentially no relationship between standardized residuals and the predicted variable?

We can examine the relationship between residuals and the predicted values. Since a visual inspection of the fitted values vs. residuals plot is not random distributed around the zero line and appears cone-like (Field et al., 2012), there appears to be evidence of heteroscedasticity (i.e., variance of the residuals is not the same for all levels of the predictor; Rovai et al., 2014) but the data seem to meet the assumption of linearity.

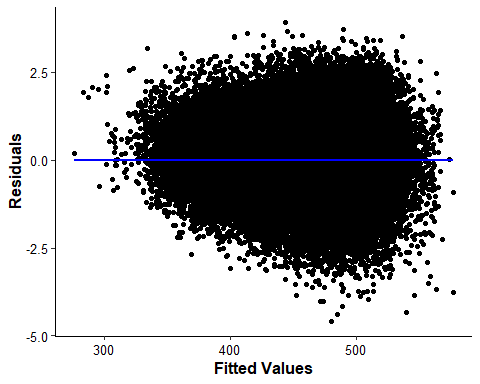
The residuals themselves, however, seem normally distributed as evidence by the histogram and Q-Q plot.

In summary, there appears to be a relationship between the the standardized residuals and the predicted variable, math achievement.

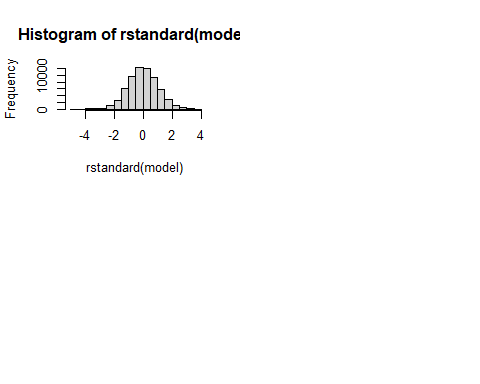
# Create plots  
par(mfrow = c(2, 2))  
plot(model)



# Remove missing values  
pisa %>%  
 filter(!is.na(PV1MATH), !is.na(gender), !is.na(ESCS)) %>%  
 # Create a plot of standarized residuals vs. fitted values only  
 ggplot(aes(model$fitted.values, rstandard(model))) +  
 geom\_point() +  
 geom\_smooth(method = "lm", colour = "Blue") +   
 labs(x = "Fitted Values", y = "Residuals") +  
 apa\_theme



# Histogram of standardized residuals  
hist(rstandard(model))



## Significance

Summarize the model to check for significance.

# Summary of the model  
summary(model)

##   
## Call:  
## lm(formula = PV1MATH ~ ESCS + gender, data = pisa)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -381.65 -56.70 -1.05 55.95 326.22   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 465.7358 0.4640 1003.63 <2e-16 \*\*\*  
## ESCS 35.5138 0.2619 135.62 <2e-16 \*\*\*  
## gender 6.6423 0.6549 10.14 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 83.16 on 64622 degrees of freedom  
## (910 observations deleted due to missingness)  
## Multiple R-squared: 0.2234, Adjusted R-squared: 0.2234   
## F-statistic: 9294 on 2 and 64622 DF, p-value: < 2.2e-16

# Descriptive statistics  
pisa %>%  
 select(PV1MATH, ESCS, gender) %>%  
 describe()

## vars n mean sd median trimmed mad min max range skew  
## PV1MATH 1 65535 456.71 94.84 451.64 454.50 97.01 85.85 828.41 742.56 0.20  
## ESCS 2 64625 -0.31 1.25 -0.17 -0.24 1.36 -5.33 3.13 8.46 -0.43  
## gender 3 65535 0.49 0.50 0.00 0.48 0.00 0.00 1.00 1.00 0.06  
## kurtosis se  
## PV1MATH -0.2 0.37  
## ESCS -0.5 0.00  
## gender -2.0 0.00

## Answer the Question

The overall multiple regression model was significant, *F*(2, 64,622) = 9,294, *p* < .001. The linear combination of socioeconomic status and gender are reasonable predictors of student math achievement. For every one unit increase in standardized SES, a student’s math achievement is predicted to rise by 35.51 points. Student identification as male (i.e., gender = 1) is associated with a 6.64 point increase in match achievement. The model explained 22.34% of the variance in math achievement.

However, caution is necessary when interpreting these results. Although many assumptions were met, there is some evidence of autocorrelation, negative skew in SES scores and residuals, and heteroscedasticity.

**Questions for class:**

* Is there a rule of thumb for autocorrelation?
* Make sure the large sample size makes the distribution of SES normal enough.
* How much of a problem is heteroscedasticity?
  + How would we report it?

# References

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

Rovai, A. P., J. D. Baker, & M. K. Ponton. (2014). *Social science research design and statistics: A practitioner’s guide to research methods and IBM SPSS analysis* (2nd ed). Watertree Press.