Assignment 3

Cory J. Cascalheira

02/16/2021

# Assignment Instructions

1. Write up a response to the following question, using a linear regression: What influence, if any, does socioeconomic status (ESCS), gender (ST04Q01), and differences between students’ home language and that of the test (ST25Q01) have on students’ mathematics achievement (PV1MATH). Don’t forget to re-code gender and language variables. Also, after class on Monday, I realized that I accidentally coded gender wrong. I though male was 1 and female was 2, but that is backward: female =1 and male = 2, so we interpreted the regression coefficient backwards from what we should have.
   * 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?

# 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?

Wording for this answer is borrowed from Assignment 2, where relevant. Theoretically, we can assume that the scores on one test do not autocorrelate with the scores on a subsequent test (Rovai et al., 2014). 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). Thus, even though random selection was used, there is evidence of non-independent observations at the school and country levels, which may affect individual student data.

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

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

# Gender: 0 = female, 1 = male  
pisa$gender <- pisa$ST04Q01 - 1  
table(pisa$gender)

##   
## 0 1   
## 33714 31821

# Language: 0 = same, 1 = other  
pisa$language <- pisa$ST25Q01 - 1  
table(pisa$language)

##   
## 0 1   
## 57362 6100

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

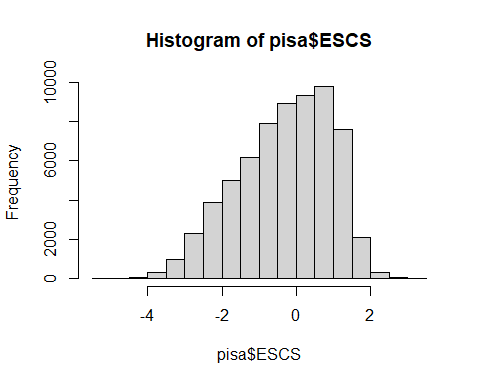
## lag Autocorrelation D-W Statistic p-value  
## 1 0.2684328 1.463128 0  
## Alternative hypothesis: rho != 0

Given that the D-W statistic is closer to 1 than 2 (i.e., 1.46) 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 and language are 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 and language.

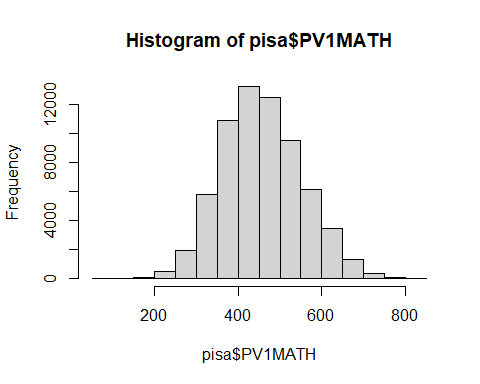
# ESCS - evidence of continuous level of measurement   
hist(pisa$ESCS)



length(unique(pisa$ESCS))

## [1] 697

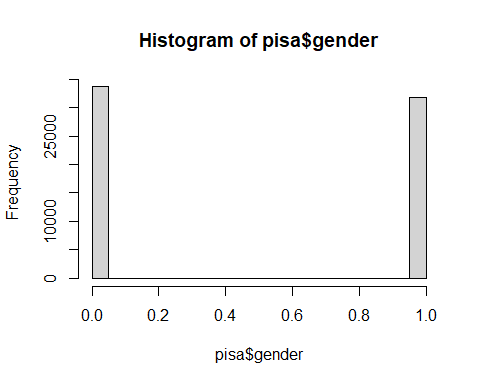
# PV1MATH - evidence of continuous level of measurement   
hist(pisa$PV1MATH)



length(unique(pisa$PV1MATH))

## [1] 5075

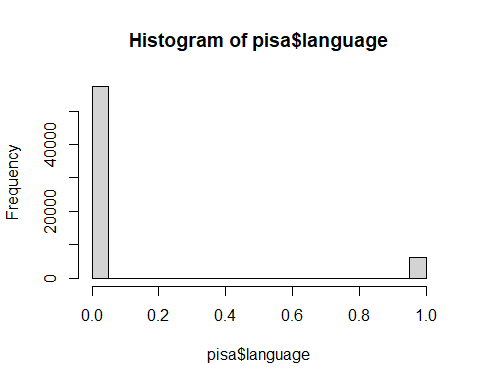
# Gender - evidence of discrete level of measurement   
hist(pisa$gender)



length(unique(pisa$gender))

## [1] 2

# Language - evidence of discrete level of measurement   
hist(pisa$language)



length(unique(pisa$language))

## [1] 3

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

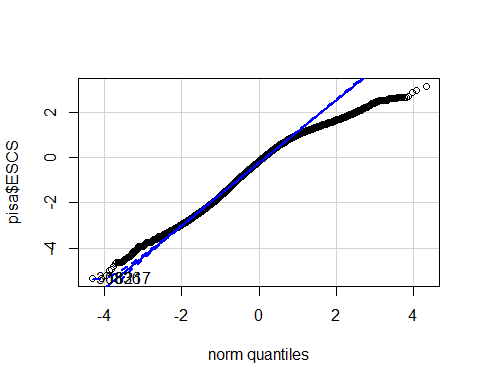
The variables are probably normal enough given the large sample size, even though normality is not strictly tenable.

That is, the histograms above provide visual evidence for and against normality. Gender and language are binomial distributions and are 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?

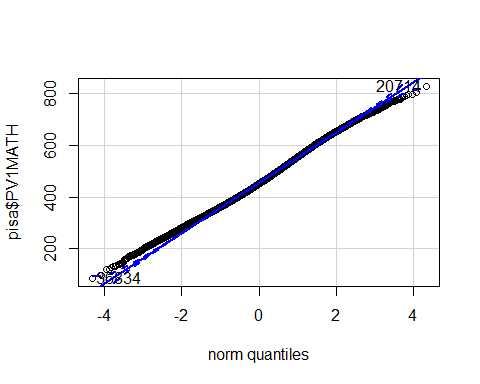
Based on a visual inspection of the Q-Q plots, the residuals of: \* ESCS are not normally distributed; \* PV1MATH are normally distributed; \* Gender are not normally distributed; and \* Language are not normally distributed.

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



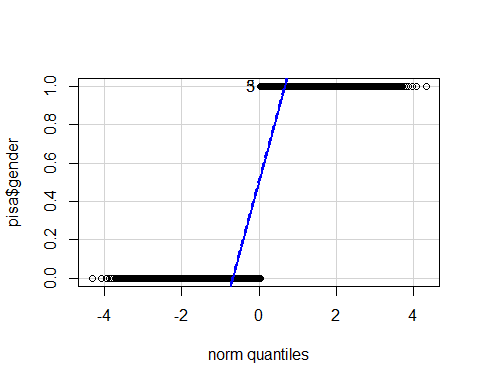
## [1] 30836 18217

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



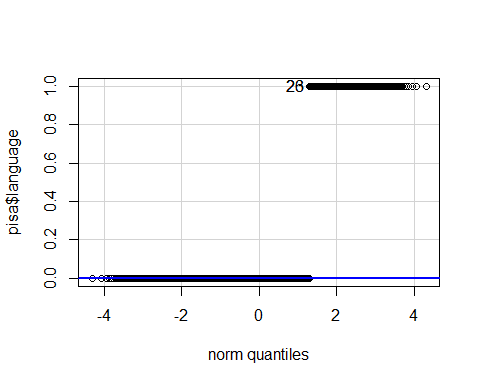
## [1] 20714 36834

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



## [1] 3 5

# Q-Q plot language  
qqPlot(pisa$language)



## [1] 23 26

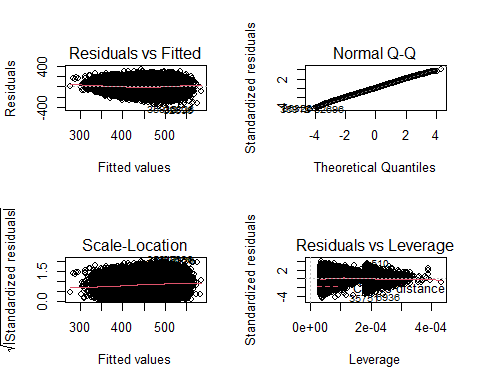
### Is there essentially no relationship between standardized residuals and the predicted variable?

We can examine the relationship between residuals and the predicted values graphically. Since a visual inspection of the fitted values vs. residuals plot is randomly distributed around the zero line and does not appear cone-like (Field et al., 2012), there appears to be no evidence of heteroscedasticity (i.e., variance of the residuals is not the same for all levels of the predictor; Rovai et al., 2014). Therefore, the data appear to meet the assumption of linearity.

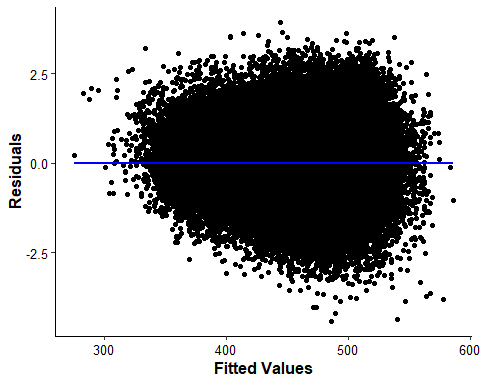
The residuals themselves seem normally distributed as evidence by the histogram and Q-Q plot.

In summary, there appears to be no relationship between the the standardized residuals and the predicted variable, math achievement.

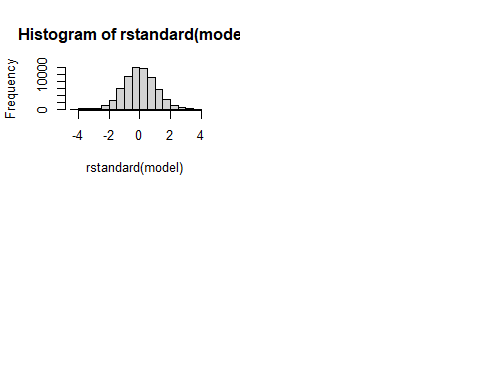
# Create plots  
par(mfrow = c(2, 2))  
plot(model)



# Remove missing values  
pisa %>%  
 filter(!is.na(PV1MATH), !is.na(gender), !is.na(ESCS), !is.na(language)) %>%  
 # 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 + language, data = pisa)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -366.17 -56.47 -1.07 55.65 325.85   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 465.5733 0.4789 972.117 < 2e-16 \*\*\*  
## ESCS 35.6948 0.2638 135.334 < 2e-16 \*\*\*  
## gender 7.3260 0.6586 11.123 < 2e-16 \*\*\*  
## language 8.9117 1.1210 7.949 1.9e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 82.71 on 63209 degrees of freedom  
## (2322 observations deleted due to missingness)  
## Multiple R-squared: 0.2295, Adjusted R-squared: 0.2295   
## F-statistic: 6276 on 3 and 63209 DF, p-value: < 2.2e-16

# Confidence intervals  
confint(model)

## 2.5 % 97.5 %  
## (Intercept) 464.634605 466.51200  
## ESCS 35.177849 36.21176  
## gender 6.035134 8.61690  
## language 6.714423 11.10893

# Descriptive statistics  
pisa %>%  
 select(PV1MATH, ESCS, gender, language) %>%  
 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  
## language 4 63462 0.10 0.29 0.00 0.00 0.00 0.00 1.00 1.00 2.74  
## kurtosis se  
## PV1MATH -0.20 0.37  
## ESCS -0.50 0.00  
## gender -2.00 0.00  
## language 5.51 0.00

## Answer the Question

Results from the multiple regression were significant, *F*(3, 63,209) = 6,276, *p* < .001, and the model explained 22.95% of the variance in math achievement. All of the variables were significant predictors. Socioeconomic status (*t* = 135.33, *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 35.69 points (95% CI [35.18, 36.21]). Gender (*t* = 11.12, *p* < .001) and language (*t* = 7.95, *p* < .001) were also significant predictors, but the practical significance was limited: being male (versus female) was associated with a 7.32 point increase in math achievement (95% CI [6.03, 8.62]) and having a native language that was different from the test (versus the same) was associated with a 8.91 point increase in math achievement (95% CI [6.71, 11.11]).

# 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.