214C - Lab 1 - Multiple Regression Write-Up

Adam Garber

214C - Lab 1 - Multiple Regression Write-Up

# Introduction

Schools which include English Language Learners (ELLs; non-native English speakers) and schools comprised primarily of native English speakers may differ in academic performance based on a variety of socio-economic factors. This study utilizes a multiple regression model approach to investigate the effect of ELL status on test scores. Specifically, we investigate whether the relationship between 8th grade student test scores and school spending is moderated by ELL status after controlling for parent income. This study utilizes a public-use repository including school district-level educational data reported in 1998 from 220 school districts in Massachusetts. In this paper the term “school(s)” is used as shorthand to refer to “school-district(s)”. This paper includes the following sections: (1) review of descriptive statistics, (2) assessment of regression model assumptions, (3) description of the regression model specification, and (4) summary of the multiple regression results.

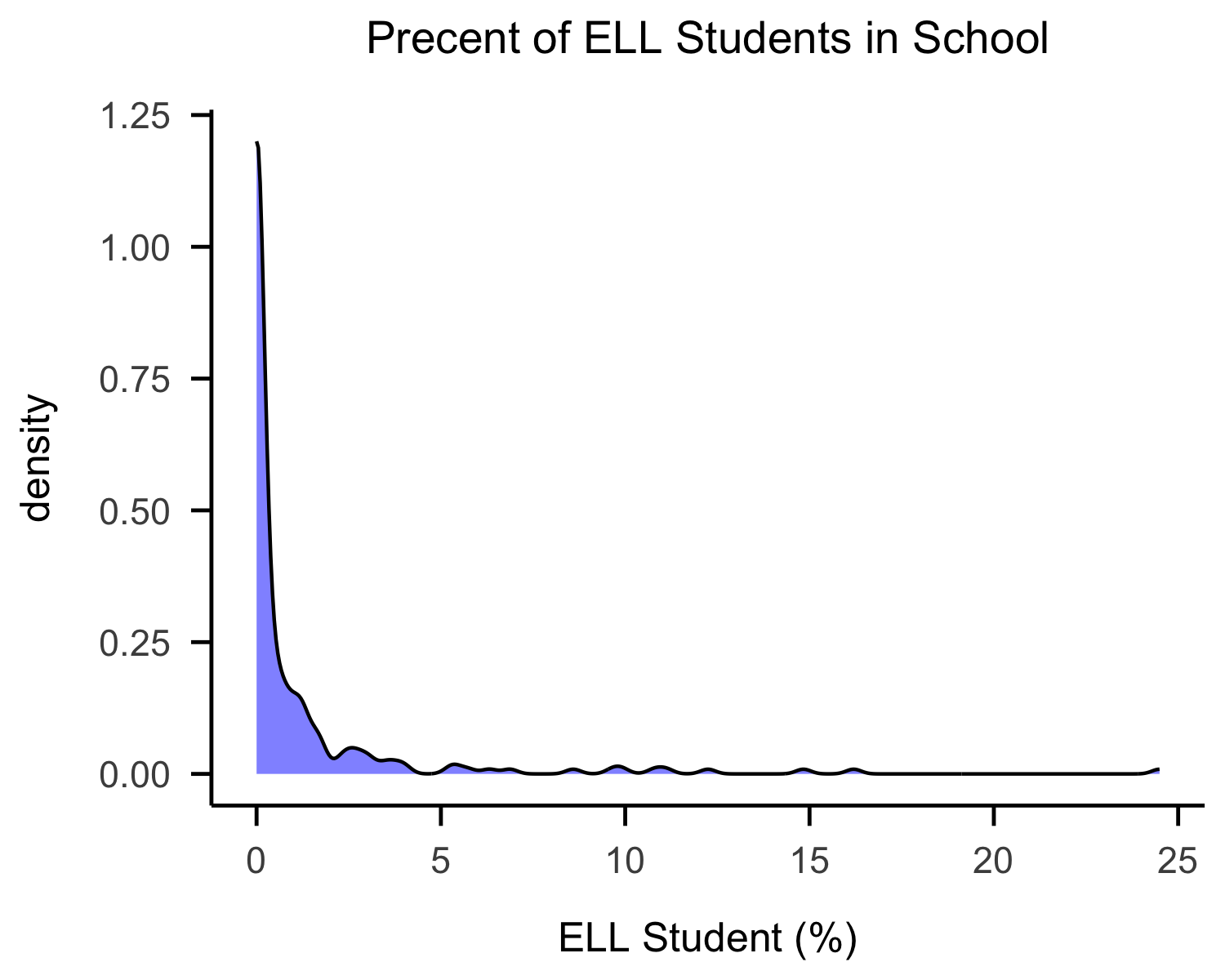
## Measures: Model Variables

Table 1 includes a description of each variable used in the multiple regression model. The variable ELL Status (english) reporting the percent of ELL students in a school was transformed into a dichotomous variable due to the heavily skewed distribution and clustering of the data at zero (i.e., floor effect; see Figure 1). In the transformed variable, schools which included at least one percent of ELL students were labeled “ELL” and schools composed of less than one percent ELL students were labeled as “NO ELL”.

Table 1:

*Description of variables included in the regression model*

| Variable Name | Description |
| --- | --- |
| Test Score | Average test scores for 8th grader students (1998) |
| School Spending | Total school spending per-pupil (dollars) |
| ELL Status | School includes ELL students (0=NO ELL, 1=ELL) |
| Parent Income | Average parent income (scale unavailable) |

 Figure 1. Distribution of percent of ELL students in school.

## Review of Descriptive Statistics

Table 2 presents the descriptive statistics for the variables included in the regression model. The outcome variable, 8th grade students’ average test scores (test\_scores) has a mean and a standard deviation of . The distribution of average test scores has minimal negative skew (as indicated by the similarity between mean and median values) and small negative kurtosis values. These shape statistics indicate that the outcome variable test scores is approximately normal. Test scores range from 641 to 747 points across schools in the sample. The predictor variable average school spending per student (spending) has a mean of and a standard deviation . The distribution of school spending has moderate positive skew, indicating that the right tail of the distribution is long or that the sample includes schools which reported high spending (relative to the mean). The kurtosis value is also positive and moderately high in value indicating that more schools are clustered at each tail of the distribution compared to a normal distribution. Average school spending per student (in dollars) ranged from 3465 to 9868 across schools in the sample. The predictor variable average parent income (income) has a mean of and a standard deviation of . The distribution of parent income is positively skewed and kurtotic with a similar distribution shape to the school spending variable. The highest and lowest average parent incomes reported were 9.7 and 46.85 respectively. Finally, for the distribution of schools with ELL students (english), 24% of schools reported ELL students and 76% of schools reported no ELL students (Table 3).

Table 2:

*Descriptive statistics for the Massachusetts school data*

|  | mean | sd | median | min | max | skew | kurtosis |
| --- | --- | --- | --- | --- | --- | --- | --- |
| test\_score | 698.41 | 21.05 | 698.00 | 641.00 | 747.00 | -0.20 | -0.12 |
| spending | 5,370.25 | 977.04 | 5,155.00 | 3,465.00 | 9,868.00 | 1.30 | 2.51 |
| income | 18.75 | 5.81 | 17.13 | 9.69 | 46.85 | 1.74 | 3.86 |

Table 3:

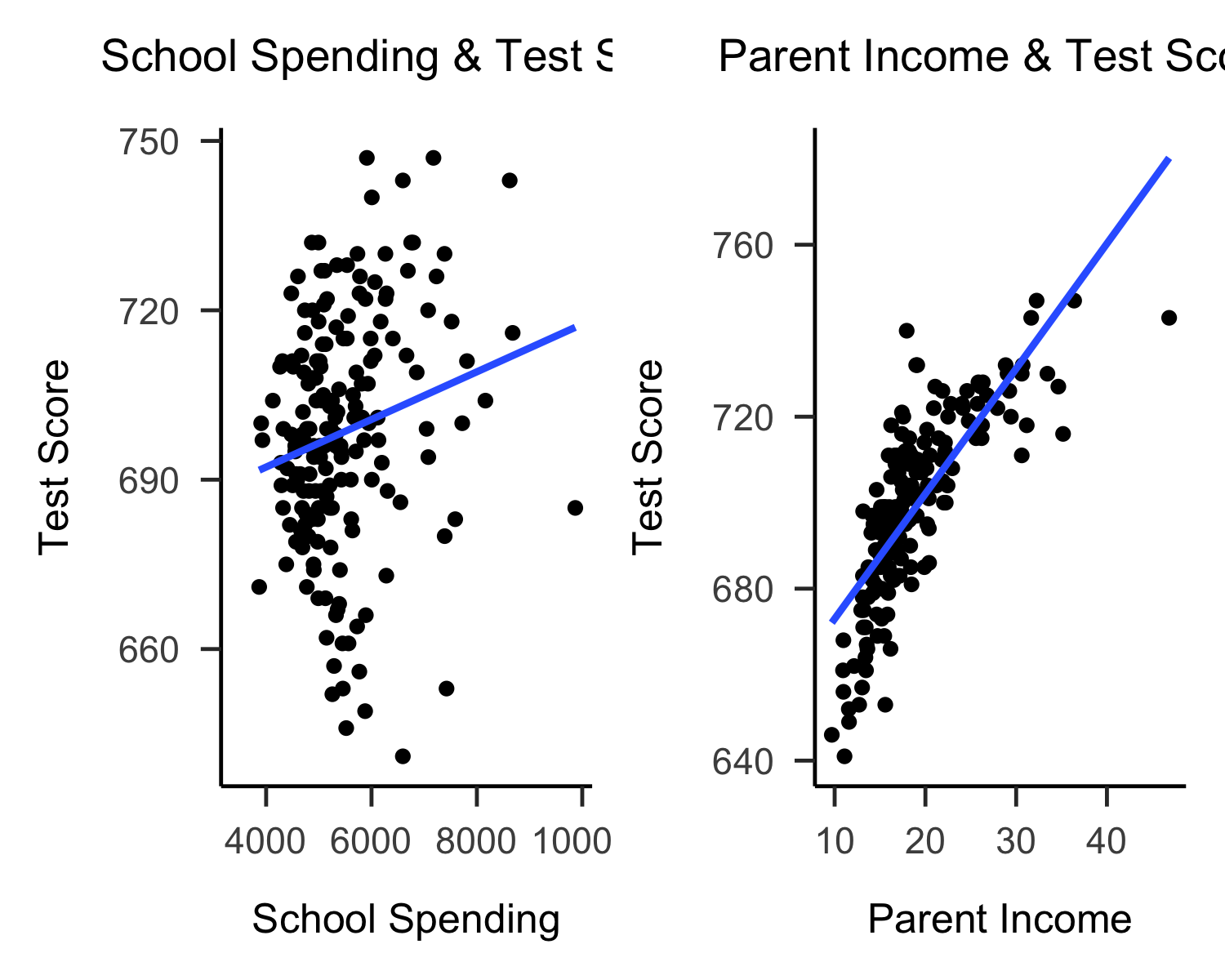
*Counts for schools with ELL students & NO ELL students*

| ELL Status | n | percent |
| --- | --- | --- |
| NO ELL | 167 | 0.76 |
| ELL | 53 | 0.24 |

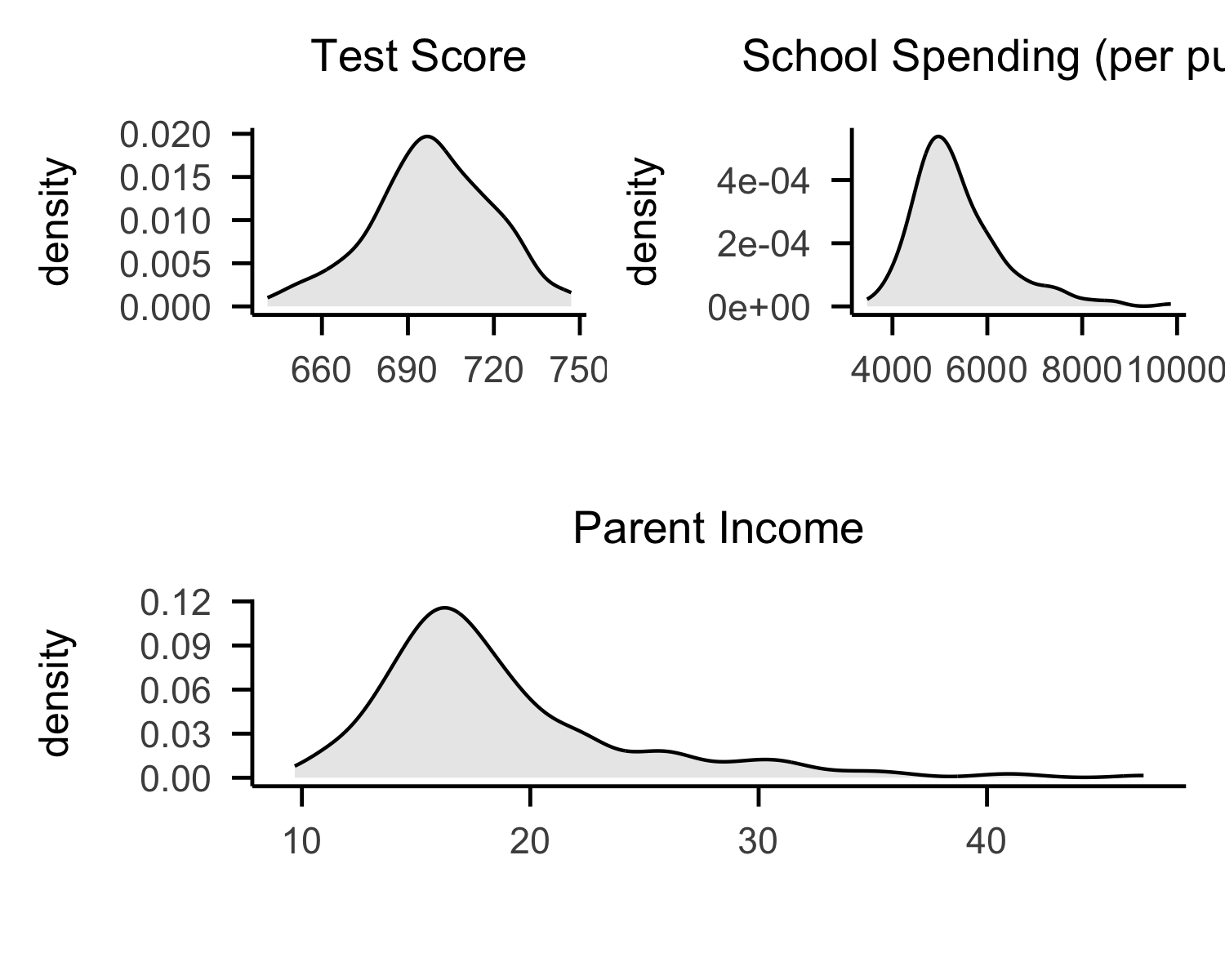
## Evaluation of Multiple Regression Assumptions

Prior to running a multiple linear regression, model assumptions were evaluated including, linearity, normality, multicollinearity, and homoscedasticity.

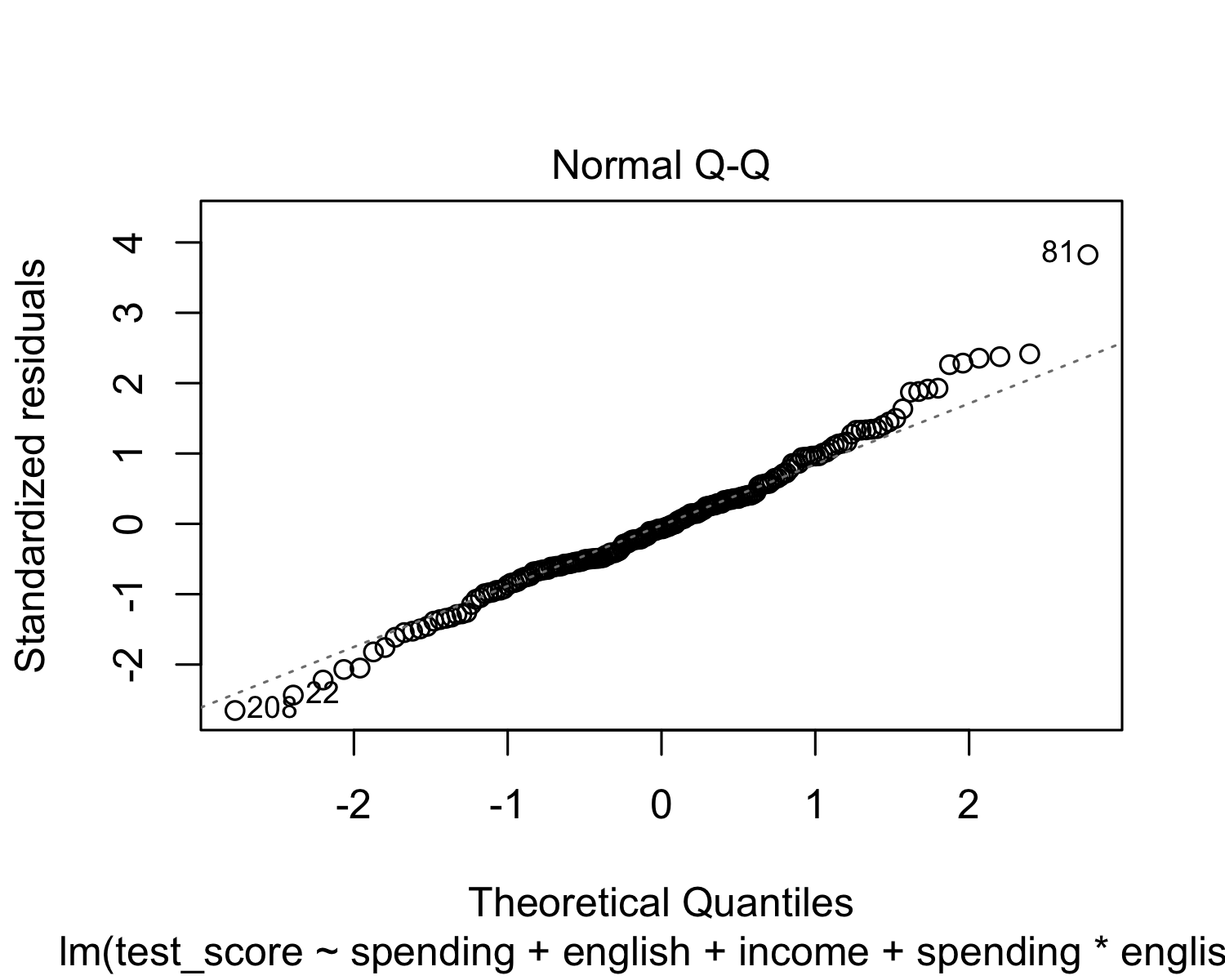
Linearity was assessed using bivariate plots to assess the relationship between the two continuous predictors (school\_spending and parent\_income) and the outcome variable (test\_scores). The left panel of Figure 2 displays the bivariate relationship between test scores and school spending with higher test scores being associated with higher school spending. This relationship is weak with high heterogeneity in test scores across the range of the school spending variable. The right panel of Figure 2, depicts a positive and moderately strong relationship between parent income and test scores. These plots provide sufficient evidence that the linearity assumption is satisfied for proceeding with the linear specification of these predictors.

 Figure 2: Bivariate scatterplots with fit line for continuous predictors by outcome variable.

Univariate normality was investigated using density plots to visualize the distribution of outcome and predictor variables (Figure 3). The outcome variable test scores on the left panel of Figure 1 can be seen to be approximately normal. School spending and income density plots reflect the positive skew and kurtosis values to be consistent with Table 2.

 Figure 3: Density plot of the outcome and predictor variables.

The multivariate normality assumption can be assessed using QQ-plots of the standardized residuals (Figure 4). The QQ-plot shows that the residuals are approximately normal with negligible violation to normality indicated by the points at either end of the plot slightly deviating form the line. Only a single residual appears to deviate significantly from normality. This provides evidence that the mulitvariate normality assumption is sufficiently satisfied to proceed with the regression analysis.

 Figure 4: QQ-plot of standardized residuals

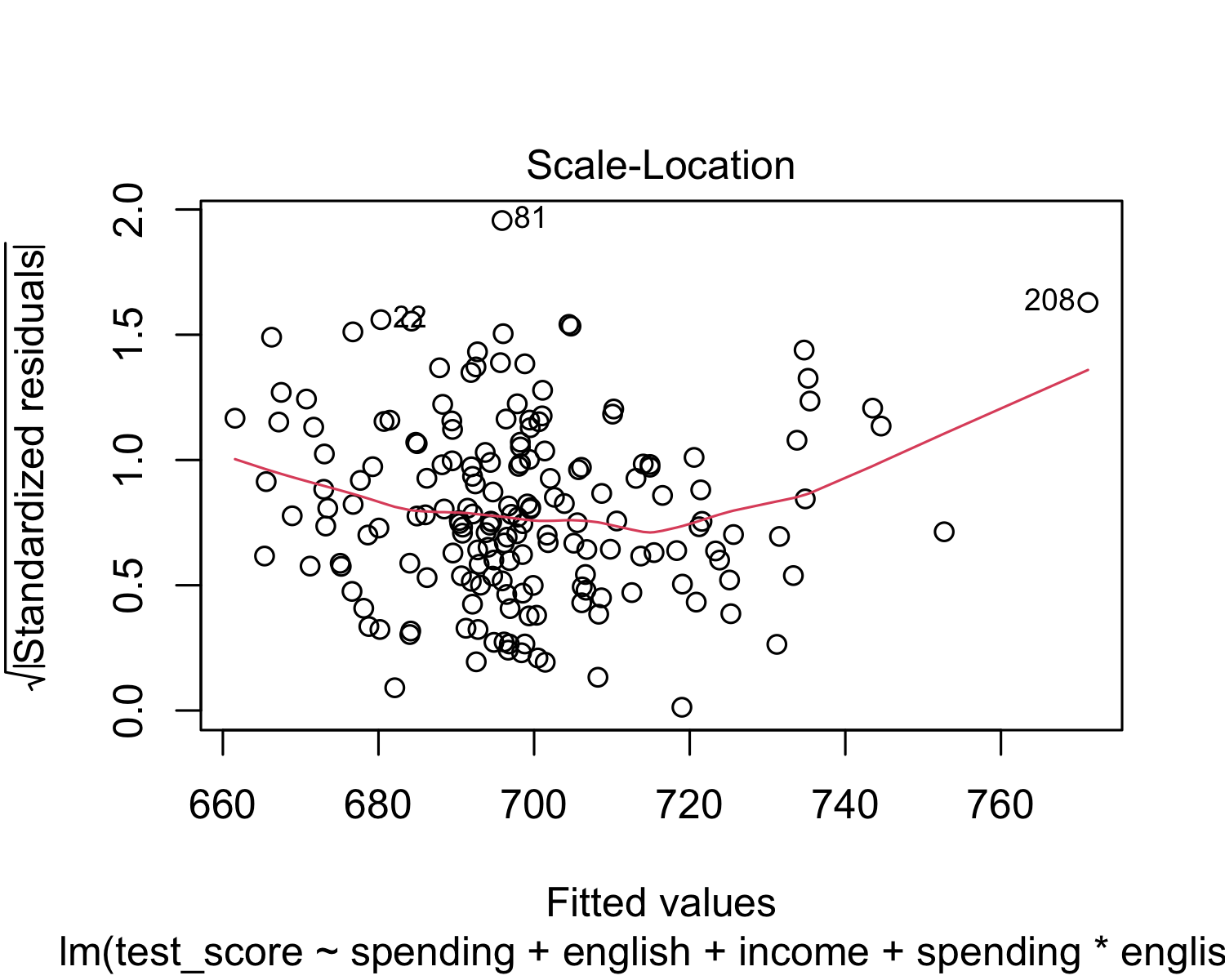
The multicollinearity assumption can be evaluated using a bivariate correlation table. The Pearson correlations shown in Table 4 are below the threshold indicating that no serious violations of multicollinearity are present. Furthermore, the variance inflation factor for the predictors are below the recommended threshold (VIF < 5) indicating that only moderate levels of multicollinearity exist for predictors in the model.

Table 4:

*Means, standard deviations, and correlations with confidence intervals*

|  | Variable | M | SD | 1 | 2 |
| --- | --- | --- | --- | --- | --- |
| X. | 1. test\_score | 698.41 | 21.05 |  |  |
| X..1 |  |  |  |  |  |
| X..2 | 2. spending | 5370.25 | 977.04 | .19\*\* |  |
| X..3 |  |  |  | [.05, .33] |  |
| X..4 |  |  |  |  |  |
| X..5 | 3. income | 18.75 | 5.81 | .78\*\* | .39\*\* |
| X..6 |  |  |  | [.71, .83] | [.27, .50] |
| X..7 |  |  |  |  |  |

*Note.* Note. M and SD are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). \* indicates p < .05. \*\* indicates p < .01.

The Homoscedasticity assumption is evaluated using a plot of the standardized residuals. Figure 5 shows horizontal but slightly curved fit line for the standardized residuals indicating that no serious violations of the assumption are present. The distribution of points in the plot are spread relatively evenly across points on the x-axis with reduced coverage at the higher end of the test score distribution.  Figure 5. Standardized residual plot

# Results

A multiple regression analysis was conducted to examine the relationship between test scores and school spending, parent income, and ELL status in a sample of schools. ELL status was hypothesized to moderate the relationship between average school spending and average test scores after controlling for parent income as shown in the linear regression equation below.

The multiple regression model with three predictors and interaction term explained 70% of the variance in average test scores, , , (see model equation).

Presented in Table 5, the regression model’s main effects and interaction term were all found to significantly predict test scores. The continuous predictors school spending and parent income were centered so that the intercept may be meaningfully interpreted. Therefore, the intercept estimate () is the average test score predicted for schools with school spending and income held constant at their respective sample means (; ) for “No ELL” schools. The main effects for average school spending (, , ) and ELL status (, , ) were both found to be significant after accounting for the interaction of school spending by ELL status and controlling for parent income. The school spending and parent income main effects, which are included in the interaction term, represent the condition when the value of the other term in the interaction is zero (i.e., when the interaction term has no effect). The main effect school spending indicates that for every 1-unit increase in school spending (from the sample mean), average test scores decrease by -.005 points for No ELL schools (ELL status = 0) with parent income held constant at the sample mean. The main effect for ELL status indicates that schools with ELL students have on average lower test scores by 12.34 points, when school spending is held constant at the sample mean (spending = 0) and parent income is held constant at the sample mean. The main effect for ELL status is a moderately sized negative effect (see standardized coefficients; Table 8) . In contrast, the main effect for school spending was relatively smaller in magnitude and negative. The control variable parent income, was found to have a positive relationship with test scores (, , ). This means that every 1-unit increase in test scores was associated with schools with 3.05 units higher average parent income.

Table 5:

*Regression model summary: Test score predicted by school spending, ELL, and parent income*

| Predictor |  | 95% CI |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Intercept | 701.520 | [699.482, 703.558] | 679.29 | 175 | < .001 |
| Spending cen | -0.005 | [-0.008, -0.002] | -3.47 | 175 | .001 |
| EnglishELL | -12.341 | [-16.764, -7.919] | -5.51 | 175 | < .001 |
| Income cen | 3.054 | [2.659, 3.450] | 15.23 | 175 | < .001 |
| Spending cen EnglishELL | 0.005 | [0.001, 0.009] | 2.64 | 175 | .009 |

Table 6 presents the standardized regression coefficients which can be used to compare magnitudes across the coefficients using a common scale. Here we can see, for example, that the effect size for average parent income () is nearly four times the magnitude of the main effect for school spending (). Similarly, the main effect for ELL status () is twice the magnitude of the main effect for school spending. The standardized regression coefficient for the interaction of school spending and ELL status () is considered a small sized effect in education research. However, in the current research context, this moderation effect may be considered meaningful due to its implication regarding test score achievement for historically marginalized populations.

Table 6.

*Standardized coefficients for multiple regression model.*

## # Standardization method: refit  
##   
## Parameter | Std. Coef. | 95% CI  
## -------------------------------------------------  
## (Intercept) | 0.12 | [ 0.02, 0.22]  
## spending | -0.23 | [-0.36, -0.10]  
## englishELL | -0.56 | [-0.77, -0.35]  
## income | 0.82 | [ 0.71, 0.92]  
## spending:englishELL | 0.25 | [ 0.06, 0.43]

The relationship between school spending and test scores was found to be moderated by ELL status after controlling for parent income. This is indicated by the significant interaction estimated in the model for school spending by ELL status (, , ). This means that for schools with ELL students the slope of school spending is .005 units higher relative to the slope for schools with no ELL students (. This moderation effect is depicted below using a simple slopes plot (Figure 6). The variable average parent income has been centered so that the plot presents the moderation with parent income held constant at the sample average. As seen in the plot (blue line), test scores decrease with increased school spending for schools with No ELL status after controlling for parent income and this slope is found to be statistically different from zero (i.e., negative relation; see Table 7). In contrast, the simple slope for schools with ELL students (orange line) is slightly positive but this slope estimate was found not to be significantly different from zero (i.e., no relation; see Table 7).

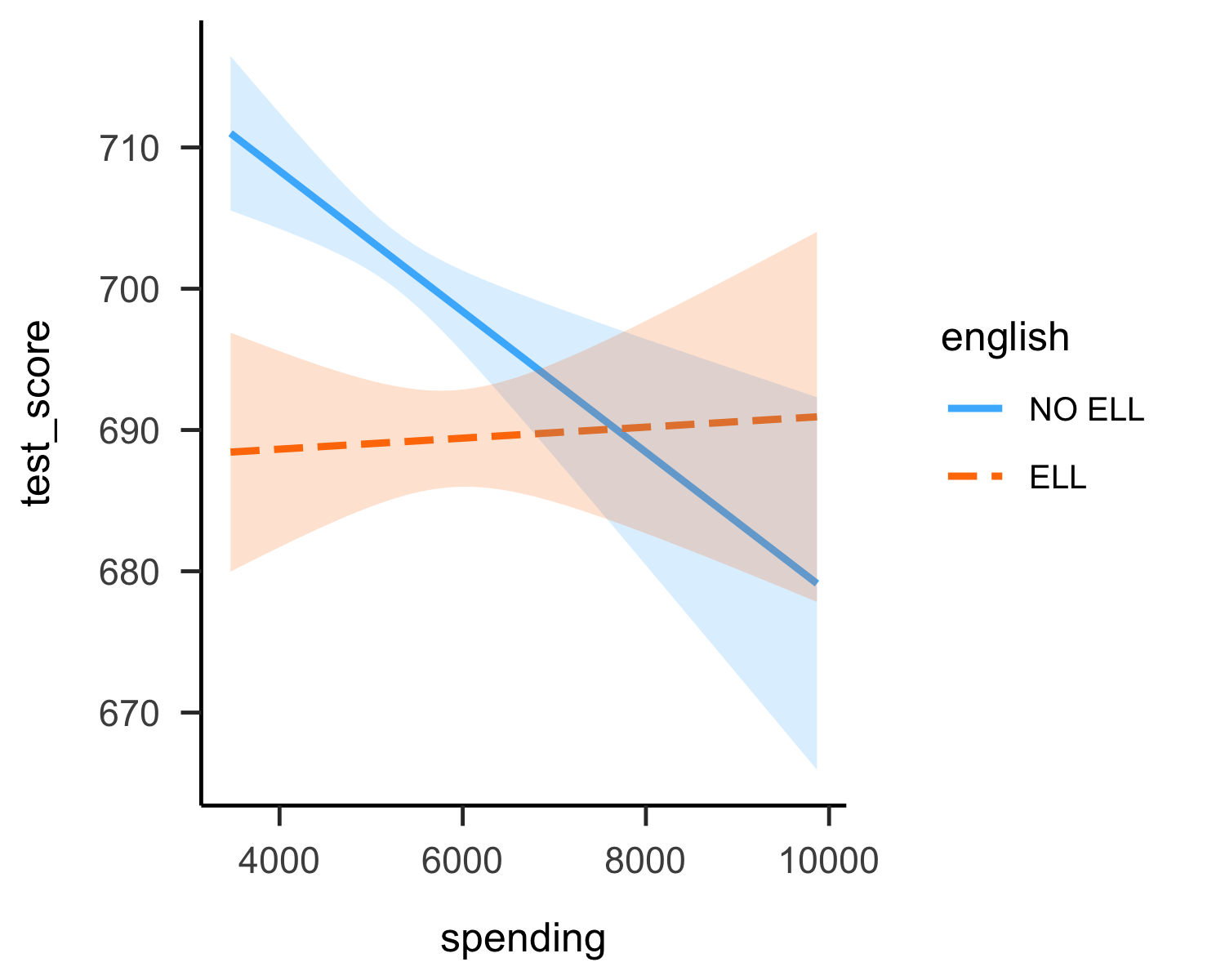
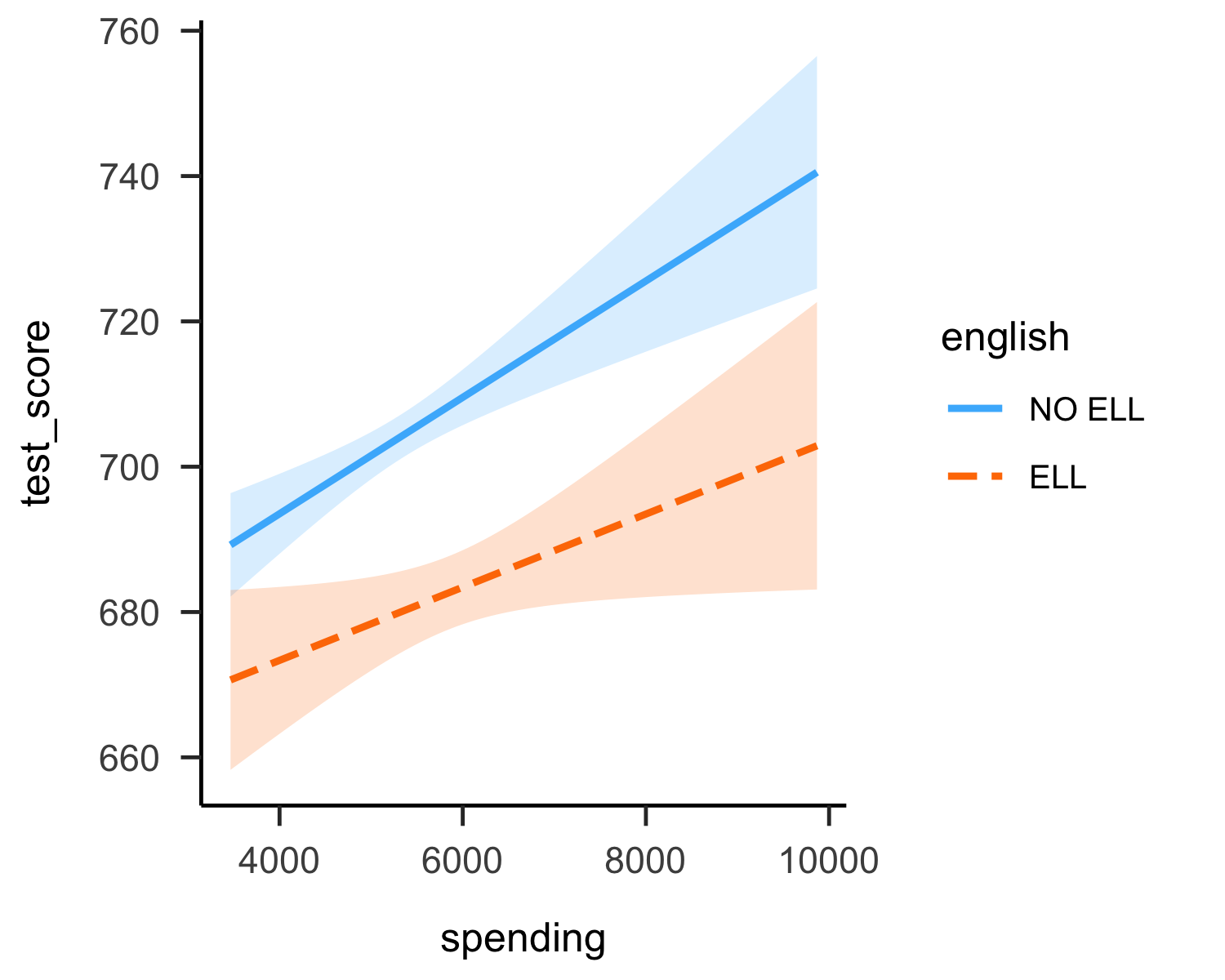
 Figure 6. Simple slope plot of ELL status moderating the relationship between school spending and test scores presented at the sample average of parent income

Table 7.

*Simple slope estimates with significance tests*

## SIMPLE SLOPES ANALYSIS   
##   
## Slope of spending when english = ELL:   
##   
## Est. S.E. t val. p  
## -------- -------- -------- --------  
## 0.0004 0.0016 0.2418 0.8093  
##   
## Slope of spending when english = NO ELL:   
##   
## Est. S.E. t val. p  
## --------- -------- --------- --------  
## -0.0050 0.0014 -3.4721 0.0007

The negative simple slope for schools with no ELL students is surprising because school spending would be expected to increase test scores regardless of population and it is possible that this effect (slope) is being suppressed or reduced by parent income. This theory can be tested by removing the variable parent income from the model and re-plotting the moderation effect. The simple slopes plot for school spending by ELL status without controlling for parent income is shown in Figure 7. Here we can see that school spending now has a positive relationship with test scores for both schools with and without ELL students and the interaction term is no longer significant (, , ).

 Figure 7: Simple slope plot presenting moderation result after removing the control variable parent income

## Discussion

The current study provides a preliminary exploration of how ELL status relates to school spending and 8th grade test scores. The finding that school spending is negatively related to test scores for schools without ELL students after controlling for parent income is interesting. This seems to suggest that school spending may not be benefiting schools which include ELL students in Massachusetts school districts. Further research is warranted to understand the reasons for this unexpected result. A theory which may explain this result, is that schools with higher spending (i.e., more affluent districts) may have school climates that have adverse effects for students who are non-native speakers (i.e., schools where discrimination is prevalent). To explore such a theory further research is warranted that includes measures of school climate as a moderating and/or mediating variable in the model. Limitations of this analysis include small sample size (*N*=220), unavailability of exogenous variables potentially salient to the research question (e.g, school climate), and lack of experimental control. A contribution of this study is that it suggests directions for future research such as the investigation of how school climate relates to test scores and school spending for historically underprivileged populations.

## References

Data Source: <https://search.r-project.org/CRAN/refmans/AER/html/MASchools.html>

# R Code: All Analyses, Plots, and Tables

# Installing tinytex & papaja packages  
  
#install.packages("tinytex")  
#tinytex::install\_tinytex()  
#install.packages("devtools")  
#devtools::install\_github("crsh/papaja")  
  
#Load libraries   
library("papaja")  
library(knitr)  
library(janitor)  
library(ggpubr)  
library(psych)  
library(tidyverse)  
library(apaTables)  
library(patchwork)  
  
# Read in data  
ma\_schools <- read\_csv("https://raw.githubusercontent.com/ejvanholm/DataProjects/master/MASchools.csv")  
  
model\_data <- ma\_schools %>%   
 select(test\_score = score8, spending = exptot, income, english) %>%   
 mutate\_at(vars("english"),   
 ~ case\_when(  
 english < 1 ~ "NO\_ELL",  
 english >= 1 ~ "ELL" )) %>%   
 mutate(english = factor(english,  
 labels = c("NO ELL", "ELL"),  
 levels = c("NO\_ELL", "ELL")))  
  
# Table 1  
  
variables <- tribble(  
 ~"Variable Name", ~"Description",   
#------------------|---------------------------------------------|,  
 "Test Score", "Average test scores for 8th grader students (1998)",   
 "School Spending", "Total school spending per-pupil (dollars)",   
 "ELL Status", "School includes ELL students (0=NO ELL, 1=ELL)",   
 "Parent Income", "Average parent income (scale unavailable)")   
  
apa\_table(  
 variables,  
 caption = "Description of Variables Included in the Regression Model.")  
  
  
# Figure 1  
  
ggdensity(ma\_schools$english, fill = "blue",title = "Precent of ELL Students in School",  
 ggtheme = theme\_apa(), xlab="ELL Student (%)")  
  
# Table 2: Descriptives  
df <- model\_data %>% select(-english)  
  
summary <- round(describe(df),2) %>%   
 select(mean, sd, median, min, max, skew, kurtosis)   
  
apa\_table(  
 summary,  
 caption = "Descriptive Statistics for the Massachusetts School Data",  
 align = c("l", "r", "r", "r", "r", "r", "r", "r")  
)  
  
# Table 3  
counts <- tabyl(model\_data$english) %>%   
 rename("ELL Status" = "model\_data$english")  
  
apa\_table(  
 counts,  
 caption = "Counts for schools with ELL students & NO ELL students"  
)  
  
# Figure 2: Bivariate plots  
  
b1 <- model\_data %>%   
 ggplot(aes(spending, test\_score)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se =F) +  
 labs(x = "School Spending", y = "Test Score") +   
 labs (title = "School Spending & Test Score") +  
 theme\_apa()  
  
  
b2 <- model\_data %>%   
 ggplot(aes(income, test\_score)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se =F) +  
 labs(x = "Parent Income", y = "Test Score") +   
 labs (title = "Parent Income & Test Score") +  
 theme\_apa()  
  
(b1+b2)   
  
# Figure 3: Density plots  
  
g1 <- ggdensity(model\_data$test\_score, fill = "lightgray",title = "Test Score", xlab = "",  
 ggtheme = theme\_apa())  
  
g2 <- ggdensity(model\_data$spending, fill = "lightgray",title = "School Spending (per pupil)",  
 xlab = "", ggtheme = theme\_apa())  
  
g3 <- ggdensity(model\_data$income, fill = "lightgray",title = "Parent Income",  
 xlab = "",ggtheme = theme\_apa())  
  
(g1+g2)/(g3)  
  
# Regression model  
reg\_model <- lm(test\_score ~ spending + english + income + spending\*english,  
 data=model\_data)  
  
# Figure 4: QQ-plot  
plot(reg\_model, 2)  
  
# Table 4: Correlation table   
  
cortable <- apa.cor.table(model\_data[,1:3], filename = NA)  
  
apa\_table(cortable$table.body,  
 caption = cortable$table.title,  
 note = cortable$table.note,  
 font\_size = "footnotesize",  
 row.names = T)  
  
# Variance inflation factor (VIF)  
library(car)  
no\_mod\_model <- lm(test\_score ~ spending + income + english, data=model\_data)  
  
vif(no\_mod\_model)  
  
# Figure 5: Standardized residuals  
plot(reg\_model, 3)  
  
  
# To present regression coefficient results we will center predictors `spending` & `income`  
# Predictors are centered so that the intercept is meaningful to interpret   
# Without centering, the intercept represents schools with (`spending=$0` & `income=$0`).   
# Schools with \*zero\* average spending/income do not exist in the sample (values are not plausible).   
# With centering intercept is for schools with spending=5370.3 & income=18.8 (mean values) & ELL=0 (No ELL)  
centered\_data <- model\_data %>%   
 mutate(spending\_cen = scale(spending, scale = FALSE), # center spending variable  
 income\_cen = scale(income, scale = FALSE)) # center income variable  
  
cen\_model <- lm(test\_score ~ spending\_cen + english + income\_cen + spending\_cen\*english,  
 data=centered\_data)  
  
  
# Regression equation (no coefficients)  
library(equatiomatic)  
  
extract\_eq(cen\_model)  
  
# Regression equation (with coefficients)  
extract\_eq(cen\_model, use\_coefs = TRUE, digits=3)  
  
#Table 5: Regression Output Summary  
apa\_lm <- apa\_print(cen\_model, digits = 3)  
  
apa\_table(  
 apa\_lm$table,  
 caption = "Regression Model Summary. Test Score predicted by School Spending, ELL (%), Parent Income",  
 align = c("l", rep("r",5))  
)  
  
# Table 6: Standardized regression coefficients  
library(effectsize)  
  
parameters::standardize\_parameters(reg\_model)   
  
# Regression model with ONLY the control variable `income` centered for interaction plot (`income` is fixed at the sample mean)  
# `spending` is NOT centered so that the x-axis of the interaction plot depicts the range of school spending on the variables original scale (`spending` range; 3465-9868)  
  
centered\_data <- model\_data %>%   
 mutate(income\_cen = scale(income, scale = FALSE))  
  
cen\_model <- lm(test\_score ~ spending + english + income\_cen + spending\*english,  
 data=centered\_data)  
   
# Create moderation plot (simple slopes)  
library(interactions)  
  
interact\_plot(cen\_model, pred = spending, modx = english, interval = TRUE, int.width = 0.9, data = centered\_data) + theme\_apa()  
  
# Table 7: Simple slopes (significance tests)  
library(sandwich)  
  
sim\_slopes(cen\_model, pred = spending, modx = english, johnson\_neyman = FALSE,  
 digits = 4)  
  
# Moderation plot: After removing `income`  
  
no\_income\_model <- lm(test\_score ~ school\_spending\*english, data=model\_data)  
  
interact\_plot(no\_income\_model, pred = school\_spending, modx = english, interval = TRUE, int.width = 0.9, data = model\_data) + theme\_apa()