EDUC 784: Regression

Peter Halpin

2022-03-02

# 1 About This Book

This “ebook” provides the course notes for EDUC 784. It is currently under development, so any feedback is appreciated (e.g., during class, via email, or the edit link in the header). This first chapter is just about how to use the book – the course content starts in Chapter 2.

## 1.1 Why this book?

There are a few goals of moving from “textbook + slides + exercises” to an ebook.

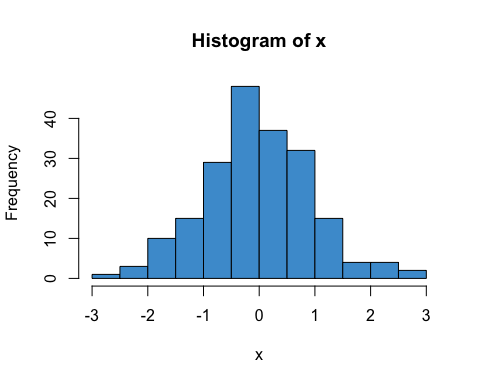
* To integrate course content (slides, readings, code, examples, and exercises) into one format, rather than having multiple files to sort through on Sakai.
* To address the perennial problem of choosing a textbook for this course – rather than having a required text, the goal is for this ebook to become the official course text. For supplementary texts, see the course syllabus.
* Most importantly, having a course text that is tightly aligned with the course content means that I can be more liberal in assigning readings as homework *before class*, so we can spend less time in lecture and more time discussing any questions you have about the readings, going through the examples in R together, and working on assignments.
* As a bonus, this book is another example of cool things you can do with R. It’s written in R (<https://bookdown.org>) – that is crazy, right??

## 1.2 How to use this book

The book combines lesson slides (Powerpoint / PDF) and R coding exercises (Rmarkdown / HTML) familiar from EDUC 710. You have already seen that the chapter sections of this book are quite short, closer to “slide sized” than “book-section sized.” This is so that they can double as course slides.

The main trick for incorporating R exercises is called “code folding.” An example of code folding is given on this page. Below, a histogram integrated into the text. By clicking on the button called “Show Code” on the top of the page, the R code that produced the histogram will also be visible. Notice that you may need to scroll horizontally to see all of the text in the code window. Also notice that when you hover your mouse over the code window, an icon appears in the top right corner – this lets you copy the block of code with one click.

# Here is some R code. You don't have to look at it when reading the book, but it is here when you need it  
x <- rnorm(200)  
hist(x, col = "#4B9CD3")



In summary, the basic workflow is as follows.

1. Before class, go through the assigned readings for conceptual understanding. You can skip all the code during your first reading.
2. We will go through the assigned readings again in class together, this time focusing on any questions you have and on doing R exercises.

Alright, let’s get to it!

# 2 Simple Regression

The focus of this course is linear regression with multiple predictors (AKA *multiple regression*), but we start by reviewing regression with one predictor (AKA *simple regression*).

## 2.1 An example from NELS

Figure 2.1 shows the relationship between Grade 8 Reading Achievement (percent correct on a reading test) and Socioeconomic Status (SES; a composite measure on a scale from 0-35). The data are a subsample of the 1988 National Educational Longitudinal Survey (NELS; see <https://nces.ed.gov/surveys/nels88/>).

# Load and attach the NELS88 data  
load("NELS.RData")  
attach(NELS)  
  
# Scatter plot  
plot(x = ses, y = achmat08, col = "#4B9CD3", ylab = "Reading Achievement (Grade 8)", xlab = "SES")  
  
# Run the regression model  
mod <- lm(achmat08 ~ ses)  
  
# Add the regression line to the plot  
abline(mod)

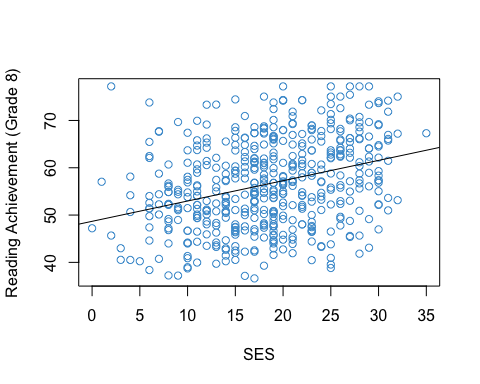


Figure 2.1: Reading Achievement and SES (NELS88).

The strength and direction of the linear relationship between the two variables is summarized by their correlation (specifically, the Pearson product moment correlation). In this sample, the correlation is

options(digits = 4)  
cor(achmat08, ses)

## [1] 0.3182

This is a moderate, positive correlation between Reading Achievement and SES. This correlation means that eighth graders from more well-off families (higher SES) also tended to do better in reading (higher Reading Achievement).

This relationship has been widely documented and discussed in education research (e.g., <https://www.apa.org/pi/ses/resources/publications/education>). **Please look over this web page and be prepared to share your thoughts about the relationship between SES and Achievement in class.**

## 2.2 The regression line

The line in the Figure 2.1 can be represented mathematically as

where

* denotes Reading Achievement
* denotes SES
* represents the values of on the line
* represents the regression intercept (the value of when )
* represents the regression slope (how much increases for each unit of increase in )

Note that represents the values of Reading Achievement in the data, whereas represents the values on the regression line. The difference is called a *residual*. The residuals for a subset of the data points in Figure 2.1 are shown in pink in Figure 2.2

# Get predicted values from regression model  
yhat <- mod$fitted.values  
  
# select a subset of the data  
set.seed(10)  
index <- sample.int(500, 30)  
  
# plot again  
plot(x = ses[index], y = achmat08[index], ylab = "Reading Achievement (Grade 8)", xlab = "SES")  
abline(mod)  
  
# Add pink lines  
segments(x0 = ses[index], y0 = yhat[index] , x1 = ses[index], y1 = achmat08[index], col = 6, lty = 3)  
  
# Overwrite dots to make it look at bit better  
points(x = ses[index], y = achmat08[index], col = "#4B9CD3", pch = 16)

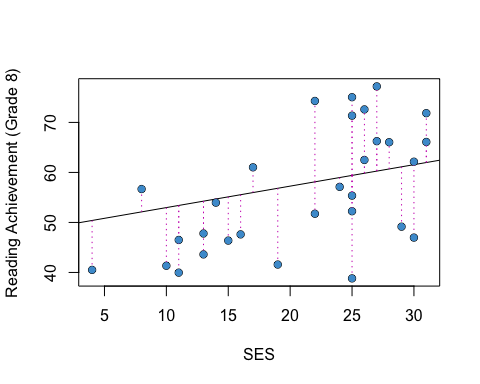


Figure 2.2: Residuals for a Subsample of the Example.

Notice that by definition. So, we can use either Equation (2.1) or Equation (2.2) to write out a regression model:

Both equations say the same thing, but Equation (2.2) lets us talk about the values of in the data, not just the predicted values.

## 2.3 OLS

Intuitively, one approach to “fitting a line to the data” is to select the parameters of the line (its slope and intercept) to minimize the residuals. In ordinary least squares (OLS) regression, we minimize a related quantity, the sum of squared residuals:

where indexes the respondents in the sample. OLS regression is very widely used and is the main focus of this course, although we will visit some other approaches (notably logistic regression) in the second half of the course.

Solving the minimization problem (i.e., doing the calculus) gives the following equations for the regression parameters

(If you aren’t familiar with the symbols in these equations, check out the review materials in Section 2.12 for a refresher.)

For the NELS example, the regression intercept and slope are, respectively:

coef(mod)

## (Intercept) ses   
## 48.6780 0.4293

**Please write down an interpretation of these numbers in terms of the line in Figure 2.1, and be prepared to share your answers in class!**

### 2.3.1 Correlation and regression

Note that if and are transformed to z-scores (i.e., to have mean of zero and variance of one), then

So, regression, correlation, and covariance are all very closely related when we consider only two variables at a time. This is why we didn’t make a big deal about simple regression in EDUC 710. But when we get to multiple regression (i.e., more than one variable), we will see that relationship between regression and correlation (and covariance) gets more complicated.

## 2.4 R-squared

In the previous section we saw that the predicted value of Educational Achievement increased by .43 units (about half a percentage point) for each unit of increase in SES. Another way to interpret this relationship is in terms of the proportion of variance in Reading Achievement that is associated with SES – i.e., to what extent are individual differences in Reading Achievement associated with, or explained by, individual differences in SES?

This question is represented graphically in Figure 2.3. The horizontal line denotes the mean of Reading Achievement. The difference between the indicated student’s Reading Achievement score and the mean can be divided into two parts.

* The black dashed line shows how much closer we get to the student’s score by considering instead of . This represents the extent to which this student’s Reading Achievement score is explained by the linear relationship with SES.
* The pink dashed line is the regression residual, which was introduced in Section 2.2. This is the variation in Reading Achievement that is “left over” after considering the linear relationship with SES.

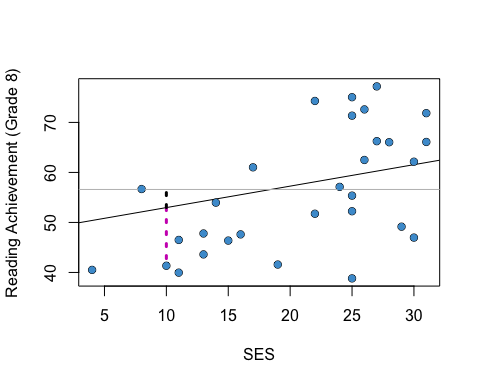


Figure 2.3: The Idea Behind R-squared.

The R-squared statistic summarizes the variation in Reading Achievement associated with SES (i.e., the black dashed line) relative to the total variation in Reading Achievement (i.e., black + pink) for all students in the sample. Aside from the regression parameters, R-squared is the most widely used statistic in regression analysis, so we will be seeing it a lot. Some authors call it the “coefficient of determination” instead of R-squared.

Using all of the cases from the example (Figure 2.1), the R-squared statistic is:

options(digits = 5)  
summary(mod)$r.squared

## [1] 0.10128

**Please write down an interpretation of this number and be prepared to share your answer in class!**

### 2.4.1 Derivation\*

To derive the R-squared statistic we work the numerator of the variance, which is called the total sum of squares.

It can be re-written using the predicted values :

The right hand side can be reduced to two other sums of squares using the rules of summation algebra (see the review in Section 2.12):

The first part is just from Section 2.3. The second part is called the regression sum of squares and denoted . Using this terminology we can re-write the above equation as

The R-squared statistic is

As discussed above, this is interpreted as the proportion of variance in that is explained by its linear relationship with .

## 2.5 The population model

In the NELS example, the population of interest is U.S. eighth graders in 1988. We want to be able to draw conclusions about that population based on the sample of eighth graders that participated in NELS. In order to do that, we make some statistical assumptions about the population, which are collectively referred to as the population model. We talk about how to check the plausibility of these assumptions in Chapter ??.

The regression population model has the following three assumptions, which are also depicted in the diagram below. Recall that the notation means that a variable has a normal distribution with mean and standard deviation .

1. Normality: The distribution of conditional on is normal for all values of .
2. Homoskedasticity: The conditional distributions have equal variances (also called homegeneity of variance).
3. Linearity: The means of the conditional distributions are a linear function of .

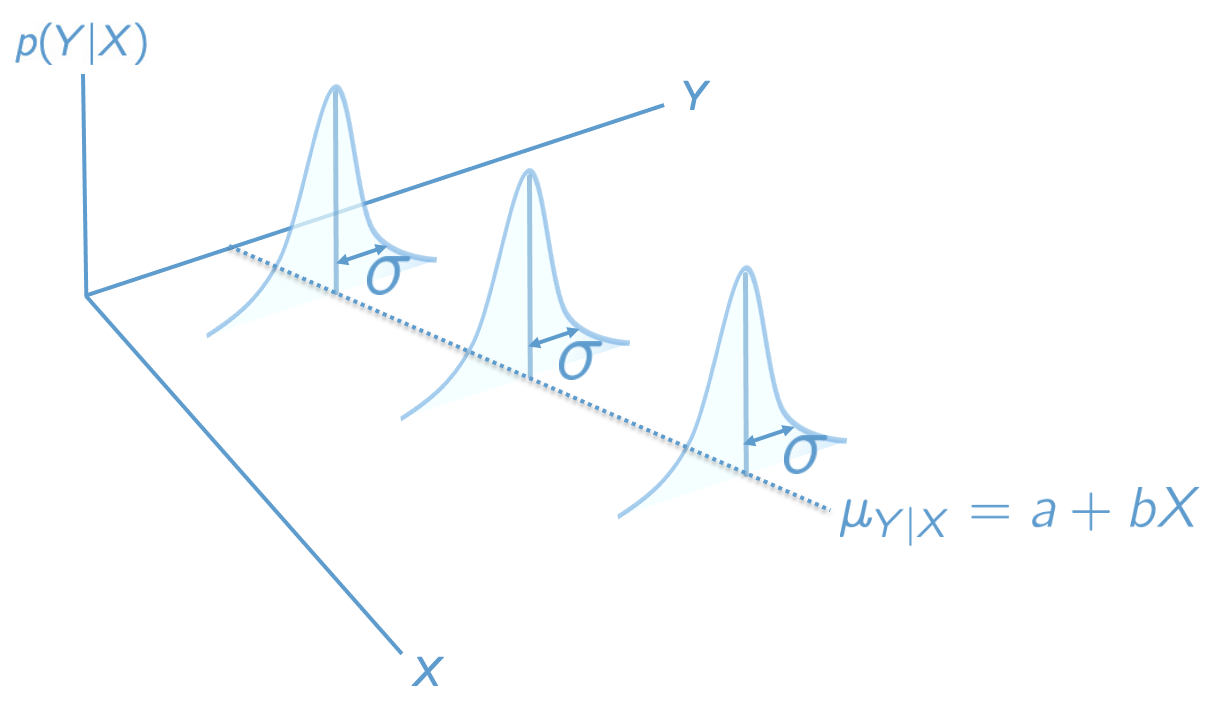


Figure 2.4: The Regression Population Model.

These three assumptions are summarized by writing

Sometimes it will be easier to state the assumptions in terms of the population residuals, , which subtract off the regression line: .

An additional assumption is usually made about the data in the sample – that they were obtained as a simple random sample from the population. We will see some ways of dealing with other types of samples later on the course (e.g., Chapter ??), but for now we can consider this a background assumption that applies to all of the procedures discussed in this course.

## 2.6 Clarifying notation

At this point we have used the mathematical symbols for regression (e.g., , ) in two different ways:

* In Section 2.2 they denoted sample statistics.
* In Section ?? they denoted population parameters.

The population versus sample notation for regression is a bit of a hot mess, but the following conventions are widely used.

| Concept | Sample statistic | Population parameter |
| --- | --- | --- |
| regression line |  |  |
| slope |  |  |
| intercept |  |  |
| residual |  |  |
| variance explained |  |  |

The “hats” always denote sample quantities, and the Greek letters (in this table) always denote population quantities, but there is some lack of consistency. For example, why not use instead of for the population slope? Well, is conventionally used to denote standardized regression coefficients in the *sample*, so its already taken (more on this in the Chapter 4 ).

One thing to note is that the hats are usually omitted from the statistics , , and if it is clear from context that we are talking about the sample rather than the population. This doesn’t apply to , because the hat is required to distinguish the predicted values from the data points.

Another thing to note is that while is often called the predicted value(s), is not usually referred to this way. It is called the conditional mean function or the conditional expectation function.

## 2.7 Inference for the slope

When the population model is true, is an unbiased estimate of . We also know the standard error of , which is equal to (cite:fox)

Using these two results, we can compute t-tests and confidence intervals for the regression slope in the usual way. These are summarized below. See the review in Section 2.12 for background information on bias, standard errors, t-tests, and confidence intervals.

### 2.7.1 t-tests

The null hypothesis can be tested against the alternative using the test statistic:

which has a t-distribution on degrees of freedom when the null hypothesis is true.

The test assumes that the population model is correct. The null hypothesis value of the parameter is usually chosen to be , in which case the test is interpreted in terms of the “statistical significance” of the regression slope.

### 2.7.2 Confidence intervals

For a given Type I Error rate, , the corresponding confidence interval is

where denotes the quantile of the -distribution with degrees of freedom.

For example, if is chosen to be , the corresponding confidence interval uses , the 2.5-th percentile of the t-distribution.

### 2.7.3 The NELS example

For the NELS example, the t-test of the regression slope is shown in the second row of the table below (we cover the rest of the output in the next few sections):

summary(mod)

##   
## Call:  
## lm(formula = achmat08 ~ ses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.600 -6.552 -0.148 6.023 27.663   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 48.6780 1.1282 43.15 < 2e-16 \*\*\*  
## ses 0.4293 0.0573 7.49 3.1e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.86 on 498 degrees of freedom  
## Multiple R-squared: 0.101, Adjusted R-squared: 0.0995   
## F-statistic: 56.1 on 1 and 498 DF, p-value: 3.13e-13

The corresponding confidence interval is

confint(mod)

## 2.5 % 97.5 %  
## (Intercept) 46.46146 50.89461  
## ses 0.31668 0.54184

**Please write down an interpretation of the t-test and confidence interval of the regression slope, and be prepared to share your answers in class!**

## 2.8 Inference for the intercept

The situation for the regression intercept is similar to that for the slope: the OLS estimate is unbiased and its standard error is (cite:fox)

The t-tests and confidence intervals are constructed in the way same as for the slope, with replacing in the notation of the previous slide. The t-distribution also has degrees of freedom for the intercept.

It is not usually the case that the regression intercept is of interest in simple regression. Recall that the intercept is the value of when . So, unless you have a hypothesis or research question about this particular value of (e.g., eighth graders with ), there isn’t a good rationale for testing the regression intercept.

When we get to multiple regression, we will see some examples of regression models where the intercept is meaningful, especially when we talk about categorical predictors in Chapter ?? and interactions in Chapter ??. But, for now, we can put it on the back burner.

**For sake of completeness, please take another look at the R output in the previous section and provide an interpretation of the t-test and confidence interval of the regression intercept, and be prepared to share your answers in class!**

## 2.9 Inference for R-squared

Inference for R-squared is quite a bit different than for the regression parameters. R-squared is a ratio of two sums of squares. We know from our study of ANOVA last semester that ratios of sums of squares are tested using an F-test, rather than a t-test. The F-test for (the population) R-squared is summarized below.

### 2.9.1 F-tests

The null hypothesis can be tested against the alternative using the test statistic:

which has a F-distribution on and degrees of freedom when the null is true. The test assumes that the population model is true. Confidence intervals for R-squared are generally not reported.

The R output from Section 2.7 is presented again below. **Please write down an interpretation of the F-test of R-squared and be prepared to share your answers in class!** Note that R uses the terminology “multiple R-squared” to refer to R-squared.

summary(mod)

##   
## Call:  
## lm(formula = achmat08 ~ ses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.600 -6.552 -0.148 6.023 27.663   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 48.6780 1.1282 43.15 < 2e-16 \*\*\*  
## ses 0.4293 0.0573 7.49 3.1e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.86 on 498 degrees of freedom  
## Multiple R-squared: 0.101, Adjusted R-squared: 0.0995   
## F-statistic: 56.1 on 1 and 498 DF, p-value: 3.13e-13

## 2.10 Power analysis

Statistical power is the probability of rejecting the null hypothesis, when it is indeed false. Rejecting the null hypothesis when it is false is sometimes called a “true positive,” meaning we have correctly inferred that a parameter of interest is not zero. Power analysis is useful for designing studies so that the statistical power / true positive rate is satisfactory. In practice, this comes down to having a large enough sample size.

Power analysis in regression is very similar to power analysis for the tests we studied last semester. There are four ingredients that go into a power analysis:

* The desired Type I Error rate, .
* The desired level of statistical power.
* The sample size, .
* The effect size, which for regression is Cohen’s f-squared statistic (AKA the signal to noise ratio):

In principal, we can plug-in values for any three of these ingredients and then solve for the fourth. But, as mentioned, power analysis is most useful when we solve for while planning a study. When solving for “prospectively,” the effect size should be based on reports of R-squared in past research. Power and are usually chosen to be .8 and .05, respectively.

When doing secondary data analysis (as in this class) there is not much point in solving for the sample size, since we already have the data. Instead, we can solve for the effect size. In the NELS example we have observations. The output below reports the smallest effect size we can detect with a power of .8 and . This is sometimes called the “minimum detectable effect size” (MDES). Note that and denote the degrees of freedom in the numerator and denominator of the F-test of R-squared, respectively.

library(pwr)  
pwr.f2.test(u = 1, v = 498, sig.level = .05, power = .8)

##   
## Multiple regression power calculation   
##   
## u = 1  
## v = 498  
## f2 = 0.015754  
## sig.level = 0.05  
## power = 0.8

**Please write down an interpretation of this power analysis, and be prepared to share your answers in class!**

## 2.11 Exercises

These exercises collect all of the R input used in this chapter in one step-by-step analysis, explain how the R input works, and provides some some additional exercises to work on. We will go through this material in class together.

### 2.11.1 The lm function

The functionlm, short for “linear model,” is used to estimate linear regressions using OLS. It also provides a lot of useful output.

The main argument that the user provides to the lm function is a formula. For the simple regression of Y on X, a formula has the syntax:

Y ~ X

Here Y denotes the outcome variable and X is the predictor variable. The tilde ~ just means “equals,” but the equals sign = is already used to assign values in R, so ~ is used in its place when writing a formula. We will see more complicated formulas as we go through the course. For more information on R’s formula syntax, see help(formula).

Let’s take a closer look using the following two variables from the NELS data, which is available on Sakai site for the course.

* achmat08: eighth grade math achievement (percent correct on a math test)
* ses: a composite measure of socio-economic status, on a scale from 0-35

# Load the data. Note that you can click on the .RData file and RStudio will load it  
# load("NELS.RData") #Un-comment this line to run  
  
# Attach the data: will dicuss this in class  
# attach(NELS) #Un-comment this line to run!  
  
# Scatter plot of math achievment against SES  
plot(x = ses, y = achmat08, col = "#4B9CD3")  
  
# Regress math achievement on SES; save output as "mod"  
mod <- lm(achmat08 ~ ses)  
  
# Print out the regression coefficients  
coef(mod)

## (Intercept) ses   
## 48.67803 0.42926

# add the regression line to the plot  
abline(mod)



Let’s do some quick calculations to check that the lm output corresponds the formulas for the slope and intercept in Section 2.3:

# Confirm that the slope from m is just the covariance divided by the variance of X  
cov\_xy <- cov(achmat08, ses)  
s\_x <- var(ses)  
b <- cov\_xy / s\_x  
b

## [1] 0.42926

# Confirm that the y-intercept is obtained from the two means and the slope  
xbar <- mean(ses)  
ybar <- mean(achmat08)  
  
a <- ybar - b \* xbar  
a

## [1] 48.678

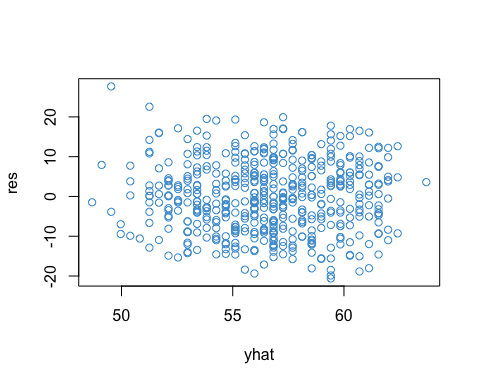
Let’s also check our interpretation of the parameters. If the answers to these questions are not clear, please make sure to ask in class!

* What is the predicted value of achmat08 when ses is equal to zero?
* How much do the predicted values of achmat08 increase for each unit of increase in ses?

### 2.11.2 Predicted values and residuals

The lm function also returns the residuals and the predicted values , which we can access using the $ operator. These are useful for various reasons, especially model diagnostics which we discuss later in the course. For now, lets take a look at the residual vs fitted plot.

yhat <- mod$fitted.values  
res <- mod$resid  
  
plot(yhat, res, col = "#4B9CD3")



cor(yhat, res)

## [1] -2.4274e-16

Note that the predicted values are uncorrelated with the residuals – this is always the case in OLS.

### 2.11.3 Variance explained

Above we found out that the regression coefficient was 0.4-ish. Another way to describe the relationships is by considering the amount of variation in that is associated with (or explained by) its relationship with . Recall that one way to do this is via the variance decomposition

from which we can compute the proportion of variation in Y that is associated with the regression model

Let’s compute “by hand” for our example.

# Compute the sums of squares  
ybar <- mean(achmat08)  
ss\_total <- sum((achmat08 - ybar)^2)  
ss\_reg <- sum((yhat - ybar)^2)  
ss\_res <- sum((achmat08 - yhat)^2)  
  
# Check that SS\_total = SS\_reg + SS\_res  
ss\_total

## [1] 43527

ss\_reg + ss\_res

## [1] 43527

# Compute R-squared  
ss\_reg/ss\_total

## [1] 0.10128

# Check that this is the same as the regression output:   
summary(mod)$r.squared

## [1] 0.10128

# Check that R-squared is really equal to the square of the PPMC  
cor(achmat08, ses)^2

## [1] 0.10128

### 2.11.4 Inference

At this point we can say that SES explained about 10% of the variation in eighth grade students’ math achievement, in our sample. However, we haven’t yet talked about statistical inference, or how we can make conclusions about a population based on a sample from that population.

Let’s use the summary function to test the coefficients in our model.

mod <- lm(achmat08 ~ ses)  
summary(mod)

##   
## Call:  
## lm(formula = achmat08 ~ ses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.600 -6.552 -0.148 6.023 27.663   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 48.6780 1.1282 43.15 < 2e-16 \*\*\*  
## ses 0.4293 0.0573 7.49 3.1e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.86 on 498 degrees of freedom  
## Multiple R-squared: 0.101, Adjusted R-squared: 0.0995   
## F-statistic: 56.1 on 1 and 498 DF, p-value: 3.13e-13

In the table, the t-test and p-values are for the null hypothesis that the corresponding coefficient is zero in the population. We can see that the intercept and slope are both significantly different from zero at the .05 level, however, the test of the slope is not very meaningful.

The text below the table summarizes the output for R-squared, including its F-test, it’s degrees of freedom, and the p-value. (We will talk about adjusted R-square in Chapter 4)

We can use the confint function to obtain confidence intervals for the regression coefficients. Use help to find out more about the confint function.

confint(mod)

## 2.5 % 97.5 %  
## (Intercept) 46.46146 50.89461  
## ses 0.31668 0.54184

Be sure to remember the correct interpretation of confidence intervals: *there is a 95% chance that the interval includes the true parameter value* (not: there is a 95% chance that the parameter falls in the interval). For example, there is a 95% chance that the interval [.31, .54] includes the true regression coefficient for SES.

### 2.11.5 Power analysis

Power analyses should ideally be done prospectively – i.e., before the data are collected. Since this class will work with secondary data analyses, most of our analyses will be retrospective. But don’t let this mislead you about the importance of statistical power – you should always do a power analysis before collecting data!!

To do a power analsyes in R, we can install and load the pwr package. If you haven’t installed an R package before it’s pretty straight forward – but just ask the instructor or a fellow student if you run into any issues.

# Install the package   
install.packages("pwr")

# Load the package by using the library command  
library("pwr")

# Use the help menu to see what the package does  
help("pwr-package")

To do a power analysis for linear regression, it is common to use Cohen’s as the effect size:

Recall that is the proportion of variance in explained by the model, and so is the proportion of variance not explained by the model. Thus, can be interpreted as a signal to noise ratio.

In addition to the effect size, we need to know the degrees of freedom for the F-test of R-square. The pwr functions use the following notation:

* u is the degrees of freedom in the numerator of an F-test.
* v is the degrees of freedom in the denominator of an F-test.

In simple regression, u = 1 and v = N - 2.

As an example of (prospective) power analysis, let’s find out many observations would be required to detect an effet size of R-square = .1, using and power = .8? To find the answer, enter the provided information into the pwr.f2.test function, and the function will solve for the “missing piece” – in this case .

# Use the provided values of R2, alpha, power (and u = 1) to solve for v = N - 2  
R2 <- .1  
f2 <- R2/(1-R2)  
pwr.f2.test(u = 1, f2 = f2, sig.level = .05, power = .8)

##   
## Multiple regression power calculation   
##   
## u = 1  
## v = 70.611  
## f2 = 0.11111  
## sig.level = 0.05  
## power = 0.8

In this example we find that . Since , so we know that a sample size of (rounded up to 73) is required to reject the null hypothesis that , when the true population value is , with a power of .8 and using a significance level of .05.

### 2.11.6 Additional exercises

If time permits, we will address these additional exercises in class. If you want more practice with R, please take a look at the review materials (on Sakai as of week 2; will be updated in the following section of the book as time permits).

These exercises replace achmat08 with

* achrdg08: eighth grade reading achievement (percent correct on a reading test)

Please answer the following questions using R.

* Plot achrdg08 against ses. Is there any evidence of nonlinearity in the relationship?
* What is the correlation between achrdg08 and ses? How does it compare to the correlation with Math and SES?
* How much variation in Reading is explained by SES? Is this more or less than for Math? Is the proportion of variance explained significant at the .05 level?
* How much do predicted Reading scores increase for a one unit of increase in SES? Is this a statistically significant at the .05 level?
* What are your overall conclusions about the relationship between Academic Achievement and SES in the NELS data?

If you want more practice with R please take a look at the Review materials (on Sakai as of week 2; will be updated in the following section of the book as time permits).

## 2.12 Review

### 2.12.1 Summation algebra

* Notation
* 3 Rules

### 2.12.2 Sample statistics

* Notation and formula’s
* R code

### 2.12.3 Bias and precision

* Sampling distributions
* Expected values and bias
* Standard errors and precision

### 2.12.4 t-tests

* Conceptual formula
* Null distribution
* Setting alpha
* Making a decision
* Interpretation
* R example

### 2.12.5 Confidence intervals

* Conceptual formula
* Interpretation
* Relation to hypothesis tests
* R example

### 2.12.6 F-tests

* Conceptual formula
* Interpretation
* R example

### 2.12.7 APA reporting

### 2.12.8 Working in R

* Hmmm

# 3 Interpretations of Regression

Before moving onto more complicated regression models, let’s consider why we might be interested in them first place. As discussed in the following sections, regression has three main uses:

* Prediction (focus on )
* Causation (focus on )
* Explanation (focus on )

By understanding these uses, you will have a better idea of how regression is applicable to your own research. Each of these interpretations also provides a different perspective on the importance of using multiple predictor variables, rather than only a single predictor.

## 3.1 Prediction

Prediction (etymology: “to make known beforehand”) means that we want to use to make a guess about . This use of regression makes the most sense when we know the value of before we know the value of .

When we are interested in using values of to make predictions about (yet unobserved) values of , we use as our guess. This is why is called the “predicted value” of .

When making predictions, we usually want some additional information about how good the predictions will be. In OLS regression, this information is provided by the prediction error variance (cite: Fox)

The prediction errors for the data in Figure 2.2 are represented in Figure 3.1 as a gray band around the regression line.

# Using a different plotting library that adds prediction error bands (need to double check computation)  
library(ggplot2)  
  
ggplot(NELS[index, ], aes(x = ses, y = achmat08)) +  
 geom\_point(color='#3B9CD3', size = 2) +  
 geom\_smooth(method = lm, color = "grey35") +  
 ylab("Reading Achievement (Grade 8)") +  
 xlab("SES") +  
 theme\_bw()

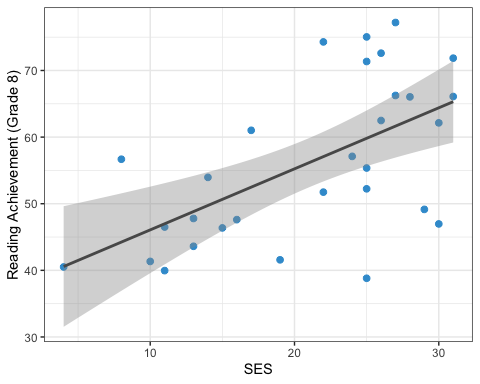


Figure 3.1: Prediction Error for Example Data.

Notice that the prediction error variance increases with – in other words, the larger the residuals (see Figure 2.2), the worse the prediction error. One way to reduce is to add more predictors into the model – i.e., multiple regression (elaborate).

### 3.1.1 More about prediction

Prediction was the original use of regression (see <https://en.wikipedia.org/wiki/Regression_toward_the_mean#History>). More recent methods developed in machine learning also focus mainly on prediction – although the methods used in machine learning are often more complicated than OLS regression, and the research context is usually quite different, the basic problem is the same. Machine learning has led to the use of out of sample predictions, rather than prediction error, as the main criterion for judging the quality of predictions made from a model. Machine learning has also introduced some new techniques for choosing which potential predictors to include in a model (i.e., “variable selection”). We will touch on these topics later in the course, although our main focus is OLS regression.

### 3.1.2 Regression toward the mean

Regression got its name from a statistical property of predicted scores called “regression toward the mean.” To explain this property, let’s assume and are z-scores (i.e., both variables have and ). Recall that this implies that and , so the regression equation reduces to

Since , the absolute value of the must be less than or equal to that of . And, since both variables have , this implies that is closer to the mean of than is to the mean of . This is what is meant by regression toward the mean.

## 3.2 Causation

A causal interpretation of regression means that that changing by one unit will change by units. Note that this is a statement about the population conditional mean function, not the sample predicted values.

This is a much stronger interpretation than prediction because it requires stronger assumptions. In particular, regression parameters can only be interpreted causally when all variables that are correlated with and are included as predictors in the model.

When a variable is left out, this is called *omitted variable bias*. This situation is nicely explained by Gelman and Hill (cite:Gelman), and a modified version of their discussion is provided below. This discussion is a bit technical, but the take-home messages are summarized in the following points.

* When a predictor variable that is correlated with and with is left out of a regression model, it is called an omitted variable.
* The problem is not just that we have an incomplete picture of how the omitted variable is related to . It is much more serious than this.
* Omitted variable bias means that the regression coefficients of the variables that *were not omitted* have the wrong value.
* The overall idea is basically the same as saying “correlation does not imply causation” or the notion of spurious correlations. It is also an example of what is called “endogeneity” in regression (etymology: originating from within).
* In order to mitigate omitted variable bias, we want to include all relevant predictors in our regression models – i.e., multiple regression

### 3.2.1 Omitted variable bias\*

We start by assuming a “true” regression model with two predictors. In the context of our example, this means that there is one other variable, in addition to SES, that is important for predicting Reading Achievement. Of course, there are many predictors of Reading Achievement (see Section ??), but we only need two to explain the problem of omitted variable bias.

Let’s write the “true” model as:

where is SES and is any other variable that is correlated with both and (e.g., number of books in the household).

Next, imagine that instead of using the model in (3.1), we analyze the data using the model with just SES. In our example, this would reflect a situation in which we don’t have data on the number of books in the house, so we have to make due with just SES, leading to the usual regression line (Section 2.2):

The basic problem of omitted variable bias is that – i.e., the regression parameter in the true model is not the same as the regression parameter in the model with only one predictor. This is perhaps surprising – leaving out the number of books in the household gives us the wrong regression parameter for SES!

To see why, start by writing as a function of .

where the regression parameters are written with Greek letters to distinguish them from the previous equations, and the residual is denoted instead of for the same reason.

Next we use Equation (3.3) to substitute for in Equation (3.1),

$$\begin{align}
Y & = a + b\_1 X\_1 + b\_2 X\_2 + \epsilon \\
Y & = a + b\_1 X\_1 + b\_2 (\alpha + \beta X\_1 + \nu) \\
Y & = \color{orange}{(a + \alpha)} + \color{green}{(b\_1 + b\_2\beta)} X\_1 + (\epsilon + \nu)
\qquad(3.4)
\end{align}$$

Notice that in the last line of Equation (3.4), is predicted using only , so it is equivalent to Equation (3.2). Based on this comparison, we can write

* $a^\* = \color{orange}{a + \alpha}$
* $b^\*\_1 = \color{green}{b\_1 + b\_2\beta}$

The equation for is what we are most interested in. It shows that the regression parameter in our one-parameter model, , is not equal to the “true” regression parameter using both predictors, .

This is what omitted variable means – leaving out in Equation (3.2) gives us the wrong regression parameter for . This is one of the main motivations for including more than one predictor variable in a regression model – i.e., to avoid omitted variable bias.

Notice that there two special situations in which omitted variable bias is not a problem:

* When the two predictors are not linearly related – i.e., .
* When the second predictor is not linearly related to – i.e., .

We will discuss the interpretation of these situations in class.

## 3.3 Explanation

In the social sciences, many uses of regression fall somewhere between prediction and causation. We want to do more than just predict outcomes of interest, but we often don’t have a basis for making strong assumptions required for causal interpretation of regression coefficients. This gray area between prediction and causation can be referred to as explanation.

In terms of our example, we might want to explain why eighth graders differ in there Reading Achievement in terms of a large number of potential predictors, such as

* Student factors
  + attendance
  + past academic performance in Reading
  + past academic performance in other subjects (Question: why include this? Hint: see previous section)
* School factors
  + their ELA teacher
  + the school they attend
  + their peers (e.g., the school’s catchment area)
* Home factors
  + SES
  + Number of books in the household
  + Maternal education

When the goal of an analysis is explanation, it usual to focus on the proportion of variation in the outcome variable that is explained by the predictors, i.e., R-squared. Later in the course we will see how we can systematically study the variance explained by individual predictors, or blocks of > 1 predictor (e.g., student factors, School factors), when we have many predictors / blocks in the model.

Note, that even a long list of predictors such as that above leaves out potential omitted variables. But, by including more than one predictor, we can get “closer” to a causal interpretation through a property of multiple regression called “statistical control.” Understanding what is meant by statistical control is the topic of the next chapter

# 4 Regression with Two Predictors

## 4.1 An example from ECLS

This section considers a subset of data from the 1998 Early Childhood Longitudinal Study (ECLS; <https://nces.ed.gov/ecls/>). We focus on the following three variables.

* Math Achievement in the first semester of Kindergarten. This variable can be interpreted as the number of questions out of 61 answered correctly on a math test. Don’t worry – the respondents in this study did not have to write a 61-question math test in the first semester of K! Students only answered a few of the questions and their scores were re-scaled to be out of the total of 61 questions afterwards.
* Socioecomonic Status (SES), which is a composite of household factors (e.g., parental education, household income) ranging from 30-72.
* Approaches to Learning (ATL), which is a teacher reported measure of behaviors that affect the ease with which children can benefit from the learning environment. It includes six items that rate the child’s attentiveness, task persistence, eagerness to learn, learning independence, flexibility, and organization. The items have 4 response categories (1-4), coded so that higher values represent more positive responses, and the scale is an unweighted average the six items.

More details about these variables are available in the ECLS user manual: <https://nces.ed.gov/ecls/data/ECLSK_K8_Manual_part1.pdf>.

In the scatter plots below, the panels are arranged in matrix format. The variable named on the diagonal appears on the vertical () axis in its row and the horizontal () axis in its column. For example, Math Achievement is on the vertical axis in the first row and the horizontal axis in the first column. Notice that plots below the diagonal are a mirror image of the plots above the diagonal.

load("ECLS250.RData")  
attach(ecls)

## The following object is masked from NELS:  
##   
## gender

example\_data <- data.frame(c1rmscal, wksesl, t1learn)  
names(example\_data) <- c("Math", "SES", "ATL")  
pairs(example\_data , col = "#4B9CD3")

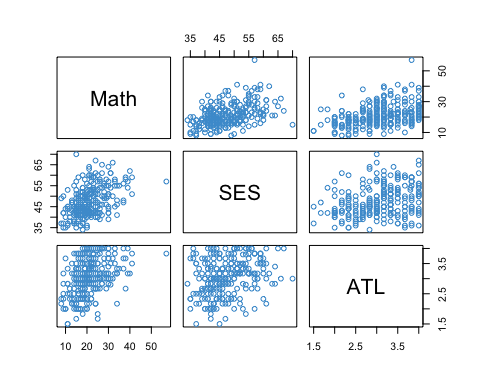


Figure 4.1: ECLS Example Data.

The format of Figure 4.1 is the same as that of the correlation matrix among the variables:

options(digits = 2)  
cor(example\_data)

## Math SES ATL  
## Math 1.00 0.44 0.40  
## SES 0.44 1.00 0.29  
## ATL 0.40 0.29 1.00

Again, notice that the entries below the diagonal are mirrored by the entries above the diagonal. We can see that all three variables are positively correlated. SES and ATL have similar correlations with Math Achievement (.44 and .40, respectively), and are also moderately correlated with each other (.29).

**If you have questions about how these plots and correlations are presented, please write them down now and share them class.**

## 4.2 The two-predictor model

In the ECLS example, we can think of Kindergarteners’ Math Achievement as the outcome variable, with SES and Approaches to Learning as potential predictors / explanatory variables. The multiple regression model for this example can be written as

where

* denotes the predicted Math Achievement scores.
* SES and ATL (it doesn’t matter which predictor we denote as or ).
* and are the regression slopes.
* The intercept is denoted by (rather than ).

Just like simple regression, the residual for Equation (4.1) is defined as and the model can be equivalently written as .

The correlations reported in Section 4.1 address how the three variables are (linearly) related in the ECLS data. Multiple regression lets us additionally address the following types of questions:

* Does ATL “add anything” to our understanding of Math Achievement, beyond SES alone?
* What is the relative importance of the two predictors?
* How much of the variance in Math Achievement do they explain?

As a first step towards answering these questions, the next section contrasts multiple regression with simple regression.

## 4.3 Multiple vs simple regression

At first glance, it might appear that simple regression and multiple regression are essentially the same thing. However, there is an important ingredient in multiple regression that is missing from simple regression. This section illustrates how the two approaches differ and asks you to think about what the missing ingredient might be.

Table 4.1 compares the output of three regression models using the ECLS example.

* “Multiple” is a two-predictor model that regresses Math Achievement on SES and ATL.
* “Simple (SES)” regresses Math Achievement on SES only.
* “Simple (ALT)” regresses Math Achievement on ALT only.

# Run models  
mod1 <- lm(Math ~ SES + ATL, data = example\_data)  
mod2a <- lm(Math ~ SES, data = example\_data)  
mod2b <- lm(Math ~ ATL, data = example\_data)  
  
# Collect output  
out1 <- c(coef(mod1), summary(mod1)$r.squared)  
out2a <- c(coef(mod2a), NA, summary(mod2a)$r.squared)  
out2b <- c(coef(mod2b), NA, summary(mod2b)$r.squared)[c(1,3,2,4)]  
out <- data.frame(rbind(out1, out2a, out2b))  
  
# Clean up names  
names(out) <- c(names(coef(mod1)), "R-squared")  
out$Model <- c("Multiple", "Simple (SES)", "Simple (ATL)")  
out <- out[c(5, 1:4)]  
row.names(out) <- NULL  
  
# Table   
options(knitr.kable.NA = '---')  
knitr::kable(out, caption = "Regression Coefficients and R-squared From the Three Models")

Table 4.1: Regression Coefficients and R-squared From the Three Models

| Model | (Intercept) | SES | ATL | R-squared |
| --- | --- | --- | --- | --- |
| Multiple | -6.05 | 0.35 | 3.5 | 0.27 |
| Simple (SES) | 0.62 | 0.44 | — | 0.19 |
| Simple (ATL) | 7.04 | — | 4.7 | 0.16 |

There are two main things to notice about the table:

* The regression coefficients in the multiple regression model do not equal the regression coefficients in the two simple regressions. This is an illustration of omitted variable bias, which we discussed in Section 3.2.
* The proportion of variance explained (R-squared) in the two-predictor model is less than the sum of the proportion of variance explained in the two simple models. Why is this weird? Well, we know that total variance of Math Achievement isn’t changing – i.e., is the same in all of the models. Therefore the R-squared values are all fractions with the same denominator, so they should be additive (e.g., ). But the values in the table don’t follow this pattern.

In summary, the regression coefficients and R-squared in the multiple regression model are different than what we would expect by extending simple regression in a naive way (i.e., by doing simple regression multiple times).

### 4.3.1 What is the missing ingredient?

Recall that in simple regression, the regression slope is just a repackaging of the correlation between the outcome and predictor (see Section 2.3). So,

* the “Simple (SES)” model considers the correlation between Math Achievement and SES, and
* the “Simple ATL” model considers the correlation between Math Achievement and ATL.

These two models leave out one of the correlations from Section 4.1 – which one? Bonus: Explain why this constitutes a case of omitted variable bias.

**Please write down your answers and be prepared to share them in class!**

## 4.4 OLS with two predictors

We can estimate the parameters of the two-predictor regression model in Equation (4.1) model using same approach as for simple regression, OLS. We do this by choosing the values of that minimize

Solving the mimmization problem leads to the following equations for the regression coefficients (the subscripts denote )

As promised, these equations are more complicated than for simple regression :) The next section addresses the interpretation of the regression coefficients.

## 4.5 Interpreting the coefficients

An important part of using multiple regression is getting the correct interpretation of the regression coefficients. The basic interpretation is that the slope parameter for SES represents how much predicted Math Achievement changes for a one unit increase of SES, *while holding ATL constant.* (The same interpretation holds when switching the predictors.) The important difference with simple regression is the “holding the other predictor constant” part, so let’s dig into it.

### 4.5.1 “Holding the other predictor constant”

We can start by revisitng the regression model in Equation (4.1):

If we increase SES () by one unit and hold ATL () constant, we get

The difference between and is how much the predicted value changes for a one unit increase in SES, while holding ATL constant:

So, the interpretation of the coefficients in multiple regression as “holding the other predictor(s) constant” is an immediate consequence of the model.

One draw back of this interpretation is that holding one predictor constant while changing another might not make sense in some applications. In fact, if the predictors are correlated, this means that changes in one predictor are associated with changes in the other. There following interpretation addresses this issue.

### 4.5.2 “Controlling for the other predictor”

The equations in for and in Section 4.4 admit another interpretation in terms of “controlling for the other predictor” (cite:Cohen). For example, the equation for is

$$

The correlation between the predictors is shown in red. Note that if the predictors are uncorrelated (i.e., $\color{red}{r^2\_{12}} = 0$) then

which is just the regression coefficient from simple regression (Section 2.3). In other words, the reason the formulas for the regression coefficients in the two-predictor model are more complicated than for simple regression is because they are “controlling for” or “accounting for” the relationship between the predictors. This argument is a bit of hand-wavy. It can be made more rigorous (cite:Pehazur), and we will discuss how in class if time permits.

### 4.5.3 The ECLS example

Below, the R output from the ECLS example is reported. **Please provide a written explanation of the regression coefficients for SES and ATL, using the interpretations above and / or any other interpretations you want to talk about. If you have questions about how to interpret the coefficients, also note them now. And, please be prepared to share your thoughts in class!**

summary(mod1)

##   
## Call:  
## lm(formula = Math ~ SES + ATL, data = example\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.10 -4.03 -1.07 3.29 29.54   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.0502 2.9003 -2.09 0.038 \*   
## SES 0.3512 0.0563 6.24 1.9e-09 \*\*\*  
## ATL 3.5212 0.6739 5.23 3.7e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.2 on 247 degrees of freedom  
## Multiple R-squared: 0.273, Adjusted R-squared: 0.267   
## F-statistic: 46.3 on 2 and 247 DF, p-value: <2e-16

## 4.6 Standardized coefficients

One question that arises in the interpretation of the example is the relative contribution of the two predictors to Kindergartener’s Math Achievement. In particular, the regression coefficient for ALT is 10 times larger than the regression coefficient for SES – does this mean that ALT is 10 times more important than SES?

The short answer is, “no.” ALT is on a scale of 1-4 whereas SES ranges over a much larger set of values. In order to make the regression coefficients more comparable, we can standardize the variables so that they have the same variance. Many researchers go a step further and standardize all of the variables to be z-scores with M = 0 and SD = 1. The resulting regression coefficients are often called -coefficients or -weights.

Comparison with Equations (4.2) shows that and

For the ECLS example, the -weights are reported below. Notice that, while the regression coefficients (and their standard errors) have changed compared to the unstandardized output reported in Section 4.5, much of the output is the same (t-tests, p-values, R-squared, its F-test).

# Unlike other software, R doesn't have a convenience functions for beta coefficients.   
z\_example\_data <- as.data.frame(scale(example\_data))  
z\_mod <- lm(Math ~ SES + ATL, data = z\_example\_data)  
summary(z\_mod)

##   
## Call:  
## lm(formula = Math ~ SES + ATL, data = z\_example\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.959 -0.560 -0.149 0.457 4.104   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.22e-16 5.42e-02 0.00 1   
## SES 3.53e-01 5.67e-02 6.24 1.9e-09 \*\*\*  
## ATL 2.96e-01 5.67e-02 5.23 3.7e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.86 on 247 degrees of freedom  
## Multiple R-squared: 0.273, Adjusted R-squared: 0.267   
## F-statistic: 46.3 on 2 and 247 DF, p-value: <2e-16

**Please write down an interpretation of the of beta coefficients in the above output. Your interpretation should include reference to the fact that the variables have been standardized. Please be prepared to share your interpretation / questions in class!**

There are a number of potential pitfalls of using Beta coefficients to “ease” the comparison of regression coefficients. In the context of our example, we might wonder whether the overall cost of raising a child’s Approaches to Learning by 1 SD is comparable to the overall cost of raising their family’s SES by 1 SD. In general, putting variables on the same scale is only a superficial way of making comparisons among their regression coefficients.

## 4.7 (Multiple) R-squared

R-squared in multiple regression has the same general formula and interpretation as in simple regression:

As shown in the previous sections, the R-squared for the ECLS example is equal to .273. **Please write down your interpretation of this value and be prepared to share your answer in class.**

As discussed below, we can also say a bit more about R-squared in multiple regression.

### 4.7.1 The multiple correlation

, the square-root of , is called the *multiple correlation* because

It is the correlation between the observed values and the predicted values. In simple regression, the multiple correlation is just the same the regular correlation coefficient . But in multiple regression, it is the correlation between the observed values and a linear combination of the values (i.e., ), so it gets a special name.

### 4.7.2 Relation with simple regression

Like the regression coefficients in Equation (4.2), the equation for R-squared can also be written in terms of the correlations among the three variables:

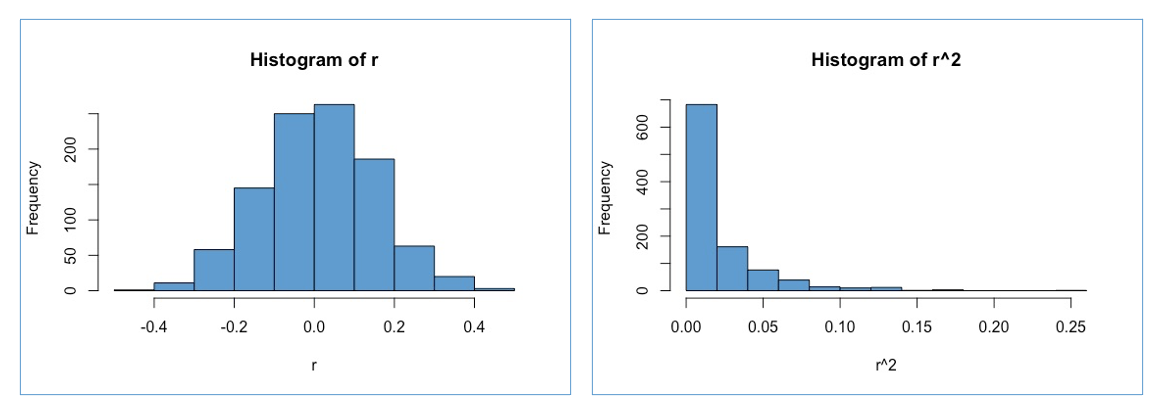
If the correlation between the predictors is zero, then we have the simplified formula

When the predictors are correlated, either positively or negatively, it can be show that

This explains the relationship among the R-squared values in Table 4.1. The sum of the R-squared values in the simple models is only equal to the R-squared value in the two-predictor when the predictors are not correlated. Otherwise, it the sum is larger than the multiple R-squareds.

### 4.7.3 Adjusted R-squared

The sample R-squared is an upwardly biased estimate of the population R-squared. The cause of the bias for the case of simple regression when is illustrated in the figure below ( is the population correlation).



(#fig:adjusted)Sampling Distribution of and when $ ho = 0$.

For the “un-squared” correlation, , the sample distribution is centered at the true value , so it is an unbiased estimate of .

But for the squared correlation, , the mean of the sampling distribution is slightly above zero because all of the random deviations from the population value are in the same direction (because they have been squared). So it is an upwardly biased (i.e., too large) estimate of .

The adjusted R-squared corrects this bias. The formula is:

where is the number of predictors in the model. It can be seen that the adjustment is larger when

* The number of predictors is large relative to the sample size .
* R-squared is closer to zero.

So, roughly speaking, the adjustment will be more severe when there are a lot of predictors in the model, but the predictors don’t explain a lot of variation in the outcome. In general, adjusted R-squared should be reported whenever it would lead to different substantive conclusions than the un-adjusted value.

## 4.8 Inference for the slopes

There isn’t really any thing new that about inference with multiple regression, except the formula for the standard errors:

In this formula, denotes the number of predictors and is the R-squared that results from regressing predictor on the other predictors (without the variable).

Notice that the first part of the standard error (before the “”) is the same as simple regression (see Section 2.7). The last part, which includes , is unique to multiple regression.

The standard errors can be used to construct t-tests and confidence intervals using the same approach as simple regression (see Section 2.7). The degrees of freedom for the t-distribution are

The R output for the ECLS example is presented (again) below. **Please write down your conclusions about the statistical significance of the predictors and be prepared to share your answer in class.**

summary(mod1)

##   
## Call:  
## lm(formula = Math ~ SES + ATL, data = example\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.10 -4.03 -1.07 3.29 29.54   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.0502 2.9003 -2.09 0.038 \*   
## SES 0.3512 0.0563 6.24 1.9e-09 \*\*\*  
## ATL 3.5212 0.6739 5.23 3.7e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.2 on 247 degrees of freedom  
## Multiple R-squared: 0.273, Adjusted R-squared: 0.267   
## F-statistic: 46.3 on 2 and 247 DF, p-value: <2e-16

### 4.8.1 Comments on precision

We can use Equation (4.3) to understand the factors that influence the size of the standard errors. Recall that standard errors describe the sample-to-sample variability of a statistic. If there is a lot sample-to-sample variability, the statistic is said to be imprecise.

* The standard errors *decrease* with
  + The sample size,
  + The proportion of variance in the outcome explained by the predictors,
* The standard errors *increase* with
  + The number of predictors,
  + The proportion of variance in the predictor that is explained by the other predictors,

So, large sample sizes and small residual variance () lead to high precision in multiple regression. On the other hand, including many highly correlated predictors in the model leads to less precision. In particular, the situation where approaches the value of is called *multicollinearity* (or just *collinearity* with 2 predictors). We will talk about multicollinearity in more detail in Chapter 6.

## 4.9 Inference for R-squared

The R-squared statistic in multiple regression tells us how much variation in the outcome is explained by all of the predictors together. If the predictors do not explain any variation, then the population R-squared is equal to zero.

Notice that implies (in the population). So, testing the significance of R-squared is equivalent to testing whether any of the regression parameters are non-zero. When we addressed ANOVA last semester, we called this the omnibus hypothesis. But in regression analysis, it is usually just referred to as a test of R-squared.

The null hypothesis can be tested using the statistic

which has an F-distribution on and degrees of freedom when the null hypothesis is true.

**Using the R output reported in the previous section, please write down your conclusion about the statistical significance of the R-squared statistic in the ECLS example.**

## 4.10 Workbook

This section collects the questions asked in this chapter. We will discuss these questions in class. If you haven’t written down / thought about the answers to these questions before class, the lesson will not be very useful for you! So, please attempt to engage with each question by writing down one or more answers, asking clarifying questions, posing follow up questions, etc.

Section 4.1

* If you have questions about the interpretation of a correlation matrix (below) or pairwise plots (see Section 4.1), please write them down now and share them class.

cor(example\_data)

## Math SES ATL  
## Math 1.00 0.44 0.40  
## SES 0.44 1.00 0.29  
## ATL 0.40 0.29 1.00

Section 4.3

* The two simple regression models (Math ~ SES and Math ~ ATL) leave out one of the correlations from Section 4.1 – which one? Bonus: Explain why this constitutes a case of omitted variable bias.

Section 4.5

* Below, the R output from the ECLS example is reported. Please provide a written explanation of the regression coefficients for SES and ATL, using the interpretations from Section?? and / or any other interpretations you want to talk about. If you have questions about how to interpret the coefficients, also note them now. And, please be prepared to share your thoughts in class!

summary(mod1)

##   
## Call:  
## lm(formula = Math ~ SES + ATL, data = example\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.10 -4.03 -1.07 3.29 29.54   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.0502 2.9003 -2.09 0.038 \*   
## SES 0.3512 0.0563 6.24 1.9e-09 \*\*\*  
## ATL 3.5212 0.6739 5.23 3.7e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.2 on 247 degrees of freedom  
## Multiple R-squared: 0.273, Adjusted R-squared: 0.267   
## F-statistic: 46.3 on 2 and 247 DF, p-value: <2e-16

Section 4.6

* Please write down an interpretation of the of beta coefficients in the above below. Your interpretation should include reference to the fact that the variables have been standardized. Please be prepared to share your interpretation / questions in class!

summary(z\_mod)

##   
## Call:  
## lm(formula = Math ~ SES + ATL, data = z\_example\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.959 -0.560 -0.149 0.457 4.104   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.22e-16 5.42e-02 0.00 1   
## SES 3.53e-01 5.67e-02 6.24 1.9e-09 \*\*\*  
## ATL 2.96e-01 5.67e-02 5.23 3.7e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.86 on 247 degrees of freedom  
## Multiple R-squared: 0.273, Adjusted R-squared: 0.267   
## F-statistic: 46.3 on 2 and 247 DF, p-value: <2e-16

Section 4.7

* The R-squared for the ECLS example is equal to .273. Please write down your interpretation of this value and be prepared to share your answer in class.

Section 4.8

* Look at the R output for the ECLS example above again (either one). Please write down your conclusions about the statistical significance of the predictors and be prepared to share your answer in class.

Section 4.9

* Look at the R output for the ECLS example above again (either one). Please write down your conclusion about the statistical significance of the R-squared statistic in the ECLS example.

## 4.11 Exercises

These notes provide an overview of regression with two variables in R.

### 4.11.1 The ECLS250 data

Let’s start by getting our example data loaded into R. We will be using a subset of cases from the Early Childhood Longitudinal Survey 1998-1998 (ECLS-K). Here is a description of the data from the official NCES codebook (page 1-1 of <https://nces.ed.gov/ecls/data/ECLSK_K8_Manual_part1.pdf>):

*The ECLS-K focuses on children’s early school experiences beginning with kindergarten and ending with eighth grade. It is a multisource, multimethod study that includes interviews with parents, the collection of data from principals and teachers, and student records abstracts, as well as direct child assessments. In the eighth-grade data collection, a student paper-and-pencil questionnaire was added. The ECLS-K was developed under the sponsorship of the U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics (NCES). Westat conducted this study with assistance provided by Educational Testing Service (ETS) in Princeton, New Jersey.*

*The ECLS-K followed a nationally representative cohort of children from kindergarten into middle school. The base-year data were collected in the fall and spring of the 1998–99 school year when the sampled children were in kindergarten. A total of 21,260 kindergartners throughout the nation participated*.

The subset of the ECLS-K data used in this class was obtained from the link below. The codebook for these data is available in our Resources folder. Note that we will be using only a small subset of the full ECLS2577 data for this example

<http://routledgetextbooks.com/textbooks/_author/ware-9780415996006/data.php>

Let’s load ECLS-K data into R. Make sure to download the file ECLS250.RData from this week’s resources folder and save the file in your working directory – check out the R exercises from our first lesson for a refresher of how to do this.

detach(ecls); detach(NELS) # remove previously attached data  
load("ECLS250.RData") # load new example  
attach(ecls) # attach   
  
# knitr and kable are just used to print nicely -- you can just use head(ecls[, 1:5])   
knitr::kable(head(ecls[, 1:5]))

| caseid | gender | race | c1rrscal | c1rrttsco |
| --- | --- | --- | --- | --- |
| 960 | 2 | 1 | 28 | 58 |
| 113 | 1 | 8 | 14 | 39 |
| 1828 | 1 | 1 | 22 | 50 |
| 1693 | 1 | 1 | 21 | 50 |
| 643 | 2 | 1 | 14 | 39 |
| 772 | 1 | 1 | 21 | 49 |

The naming conventions for these data are bit challenging.

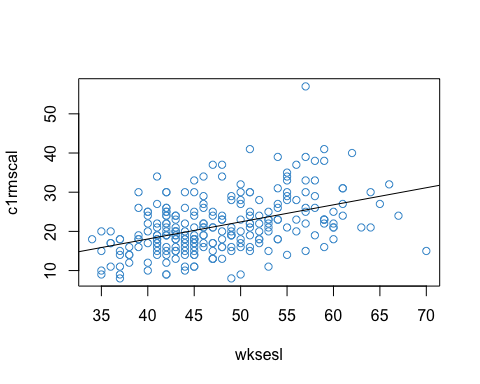
* Variable names begin with c, p, or t depending on whether the respondent was the child, parent, or teacher. Variables that start with wk were created by the ECLS using other data sources available in during the kindergarten year of the study.
* The time points (1-4 denoting fall and spring of K and Gr 1) appear as the second character.
* The rest of the name describes the variable.

The variables we will use for this illustration are:

* c1rmscal: Child’s score on a math assessment, in first semester of Kindergarten . The scores can be interpreted as number of correct responses out of a total of approximately 80 math exam questions.
* wksesl: An SES composite of household factors (e.g., parental education, household income) ranging from 30-72.
* t1learn: Approaches to Learning Scale (ATLS), teacher reported in first semester of kindergarten. This scale measures behaviors that affect the ease with which children can benefit from the learning environment. It includes six items that rate the child’s attentiveness, task persistence, eagerness to learn, learning independence, flexibility, and organization. The items have 4 response categories (1-4), coded so that higher values represent more positive responses, and the scale is an unweighted average the six items.

To get started lets produce the simple regression of Math with SES. This is another look at the relationship between Academic Achievement and SES that we discussed in Chapter 2. If you do not feel comfortable running this analysis or interpreting the output, take another look at Section 2.11.

plot(x = wksesl, y = c1rmscal, col = "#4B9CD3")  
mod <- lm(c1rmscal ~ wksesl)  
abline(mod)



summary(mod)

##   
## Call:  
## lm(formula = c1rmscal ~ wksesl)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.131 -4.355 -0.849 3.678 31.536   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.6159 2.7393 0.22 0.82   
## wksesl 0.4359 0.0567 7.68 3.6e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.5 on 248 degrees of freedom  
## Multiple R-squared: 0.192, Adjusted R-squared: 0.189   
## F-statistic: 59 on 1 and 248 DF, p-value: 3.61e-13

cor(wksesl, c1rmscal)

## [1] 0.44

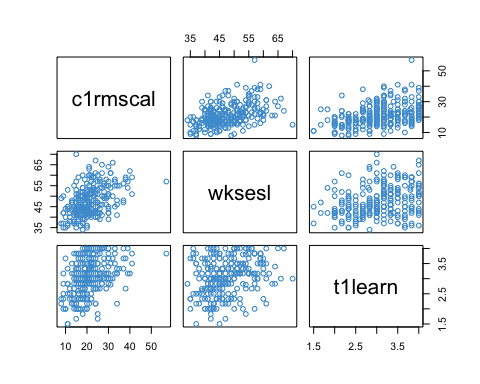
### 4.11.2 Multiple regression with lm

Let’s tale a look at “zero-order” relationship among the three variables. This type of descriptive, two-way analysis is a good way to get familiar with your data before getting into multiple regression analysis. We can see that the variables are all moderately correlated and their relationships appear reasonably linear.

# Use cbind to create a data.frame with just the 3 variables we want to examine  
data <- cbind(c1rmscal, wksesl, t1learn)  
  
# Correlations  
cor(data)

## c1rmscal wksesl t1learn  
## c1rmscal 1.00 0.44 0.40  
## wksesl 0.44 1.00 0.29  
## t1learn 0.40 0.29 1.00

# Scatterplots  
pairs(data, col = "#4B9CD3")



In terms of input, multiple regression with lm is just as simple as for a single predictor. The only difference is the model formula. To include more predictors in a formula, just include them on the right hand side, separated by at + sign.

* e.g, Y ~ Χ1 + Χ2

For our example, let’s consider the regression of math achievement on SES and Approaches to Learning. We’ll save our result as mod1 which is short for “model one.”

mod1 <- lm(c1rmscal ~ wksesl + t1learn)  
summary(mod1)

##   
## Call:  
## lm(formula = c1rmscal ~ wksesl + t1learn)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.10 -4.03 -1.07 3.29 29.54   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.0502 2.9003 -2.09 0.038 \*   
## wksesl 0.3512 0.0563 6.24 1.9e-09 \*\*\*  
## t1learn 3.5212 0.6739 5.23 3.7e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.2 on 247 degrees of freedom  
## Multiple R-squared: 0.273, Adjusted R-squared: 0.267   
## F-statistic: 46.3 on 2 and 247 DF, p-value: <2e-16

We can see from the output that regression coefficient for t1learn is about 3.5. This means that, as the predictor increases by a single unit, children’s predicted math scores increase by 3.5 points (out of 80), after controlling for the SES. You should be able to provide a similar interpretation of the regression coefficient for wksesl. Together, both predictors accounted for about 27% of the variation in students’ math scores. In education, this would be considered a pretty good coefficient of determination.

We will talk about the statistical tests later on. For now let’s consider the relationship with simple regression.

### 4.11.3 Relations between simple and multiple regression

First let’s consider how the two simple regression compare to the multiple regression with two variables. Here is the relevant output:

# Compare the multiple regression output to the simple regressions  
mod2a <- lm(c1rmscal ~ wksesl)  
summary(mod2a)

##   
## Call:  
## lm(formula = c1rmscal ~ wksesl)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.131 -4.355 -0.849 3.678 31.536   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.6159 2.7393 0.22 0.82   
## wksesl 0.4359 0.0567 7.68 3.6e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.5 on 248 degrees of freedom  
## Multiple R-squared: 0.192, Adjusted R-squared: 0.189   
## F-statistic: 59 on 1 and 248 DF, p-value: 3.61e-13

mod2b <- lm(c1rmscal ~ t1learn)  
summary(mod2b)

##   
## Call:  
## lm(formula = c1rmscal ~ t1learn)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.40 -4.21 -1.00 3.77 31.84   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.039 2.148 3.28 0.0012 \*\*   
## t1learn 4.730 0.693 6.83 6.7e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.6 on 248 degrees of freedom  
## Multiple R-squared: 0.158, Adjusted R-squared: 0.155   
## F-statistic: 46.6 on 1 and 248 DF, p-value: 6.66e-11

The important things to note here are

* The regression coefficients from the simple models ( and ) are larger than the regression coefficients from the two-predictor model. Can you explain why? (Hint: see Section 4.5.)
* The R-squared terms in the two simple models (.194 + .158 = .352) add up to more than the R-squared in the two-predictor model (.274). Again, take a moment to think about why before reading on. (Hint: see Section 4.7.)

### 4.11.4 Inference with 2 predictors

Let’s move on now to consider the statistical tests and confidence intervals provided with the lm summary output.

For regression with more than one predictor, both the t-tests and F-tests have a very similar construction and interpretation as with simple regression. The main differences are (see Sections 4.8 and 4.9):

* The degrees of freedom for both tests now involve , the number of predictors.
* The standard error of the b-weight is more complicated, because it involves the inter-correlation among the predictors.

We can see that for mod1 that both b-weights are significant at the .05 level, and so is the R-square. As mentioned previously, it is not usual to interpret or report results on the regression intercept unless you have a special reason to do so (e.g., see Chapter 5).

# Revisting the output of mod1  
summary(mod1)

##   
## Call:  
## lm(formula = c1rmscal ~ wksesl + t1learn)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.10 -4.03 -1.07 3.29 29.54   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.0502 2.9003 -2.09 0.038 \*   
## wksesl 0.3512 0.0563 6.24 1.9e-09 \*\*\*  
## t1learn 3.5212 0.6739 5.23 3.7e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.2 on 247 degrees of freedom  
## Multiple R-squared: 0.273, Adjusted R-squared: 0.267   
## F-statistic: 46.3 on 2 and 247 DF, p-value: <2e-16

### 4.11.5 APA reporting of results

In terms of writing out the results, there are many formatting styles used in social sciences. As one example, the convention for APA style is to write the coefficient, followed by the test statistic (with its degrees of freedom) and then the p-value. It is also conventional to use 2 decimal places, unless more decimal places are needed to address rounding error.

Here is how we might write out the results of our regression using APA format:

* The regression of Math Achievement on SES was positive and statistically significant at the .05 level ().
* The regression of Math Achievement on Approaches to Learning was also positive and statistically significant at the .05 level ().
* Together both predictors accounted for about 27% of the variation in Math Achievement ( = .274, adjusted = .268), which was also statistically significant at the .05 level ().

Instead of, or in addition to, the statistical tests, we could include the confidence intervals for the regression coefficients. It is not usual to report confidence intervals on R-squared.

confint(mod1)

## 2.5 % 97.5 %  
## (Intercept) -11.76 -0.34  
## wksesl 0.24 0.46  
## t1learn 2.19 4.85

* The 95% confidence interval on the regression coefficient of Math achievement on SES was . For Approaches to Learning, the 95% confidence interval was .

When we have a regression model with many predictors, or are comparing among different models, it is more usual to put all the relevant statistics in a table rather than writing them out one by one. We will see how to do that later on in the course.

For more info on APA format, see the APA publications manual (<https://www.apastyle.org/manual>).

# 5 Categorical Predictors

So far we have considered examples in which we regress a continuous outcome variable (e.g., Math Achievement) on one or more continuous predictors (SES, Approaches to Learning). In this chapter we consider how regression can be used with categorical predictors.

In an experimental context, the canonical example of a categorical predictor is treatment status (e.g., 1 = treatment group, 0 = control group). Examples of other categorical predictors commonly used in education research include:

* Geographical region / school district (Orange, CH-C, Wake, …)
* Type of school (public, private, charter, religious)
* Which classroom, teacher, or school a student was assigned to
* Gender (if recorded as categorical)
* Race / ethnicity (if recorded as categorical)
* Free / reduced price lunch status
* English language learner status
* Individualized learning plan status
* …

This chapter will focus on the topic of “contrast coding” (also called “effect coding” or “dummy coding”). In particular, we will

* Address the special case of a single binary predictor
* Show some ways that this approach generalizes to categorical predictors with more than 2 categories, specifically
  + reference-group coding (called treatment contrasts in R)
  + deviation coding (called sum contrasts in R).

The main thing to know about the different approaches to contrast coding is that they each lead to a different interpretation of the coefficients in the regression model. In this chapter we will use a two-step procedure to work out how to interpret the regression coefficients. We will apply this two-step approach to the types of contrast coding listed above, and you can use the same approach to work out the interpretation of other contrasts that you may encounter in your research (there are many different types of contrasts out there!).

Along the way we will see that regression includes as special cases the independent samples t-tests of means and one-way ANOVA procedure we discussed last semester. In the next chapter we will address how to combine categorical and continuous predictors in the same model.

## 5.1 Focus on interpretation

Categorical predictors are challenging to understand, because, depending on the contrast coding used, the model results can appear quite different.

For example, the two models below uses the same data and the same variables (Math Achievement regressed on Urbanicity), but their regression coefficients have different values – Why? Because Urbanicity used different contrast coding.

The lm output doesn’t tell us what kind of coding was used for our categorical variables – we need to know what is going on “under the hood” so that we can interpret the output correctly.

Notice that nothing has changed with respect to the computation of standard errors, R-squared, t-tests, F-tests, or p-values – all of this is the same as the previous chapters. The difference between these two models is just how the categorical predictor is interpreted.

detach(ecls)  
load("NELS.RData")  
attach(NELS)  
  
# run model with default contrast (treatment / dummy coding)  
egA <- lm(achrdg08 ~ urban)  
  
# change to sum / deviation contrasts and run again  
contrasts(urban) <- contr.sum(n = 3)  
colnames(contrasts(urban)) <- c("Rural", "Suburban")  
egB <- lm(achrdg08 ~ urban)  
  
# print  
summary(egA)

##   
## Call:  
## lm(formula = achrdg08 ~ urban)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.30 -6.16 0.21 6.79 15.56   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 54.993 0.688 79.88 <2e-16 \*\*\*  
## urbanSuburban 0.668 0.912 0.73 0.464   
## urbanUrban 3.128 1.048 2.98 0.003 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.76 on 497 degrees of freedom  
## Multiple R-squared: 0.019, Adjusted R-squared: 0.0151   
## F-statistic: 4.82 on 2 and 497 DF, p-value: 0.00841

summary(egB)

##   
## Call:  
## lm(formula = achrdg08 ~ urban)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.30 -6.16 0.21 6.79 15.56   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 56.258 0.402 139.90 <2e-16 \*\*\*  
## urbanRural -1.265 0.565 -2.24 0.026 \*   
## urbanSuburban -0.597 0.530 -1.13 0.260   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.76 on 497 degrees of freedom  
## Multiple R-squared: 0.019, Adjusted R-squared: 0.0151   
## F-statistic: 4.82 on 2 and 497 DF, p-value: 0.00841

detach(NELS)

## 5.2 Data and social constructs

Before getting into the math, let’s consider some conceptual points.

First, some terminology. *Binary* means the a variable can take on only two values: 1 and 0. If a variable takes on two values but these are represented with other numbers (e.g., 1 and 2) or with non-numeric values (“male,” “female”), it is called *dichotomous* rather than binary. Otherwise stated, a binary variable is a dichotomous variable whose values are 1 and 0. The term *categorical* is used to describe variables with two or more categories.

Note that encoding a variable as dichotomous does not imply that the underlying social construct is dichotomous. For example, we can encode educational attainment as a dichotomous variable indicating whether a person has graduated high school or not. This does not imply that educational attainment has only two values in real life, or even that educational attainment is best conceptualized in terms of years of formal education. Nonetheless, for many outcomes of interest it can be meaningful to consider whether individuals have completed high school (e.g., <https://www.ssa.gov/policy/docs/research-summaries/education-earnings.html>)

In general, the way that a variable is encoded in a dataset is not a statement about reality – it reflects a choice made by researchers about how to represent reality. In particular, we are often faced with less-than-ideal encodings of so-called demographic variables in quantitative data. For example, both NELS and ECLS conceptualize gender as dichotomous and use a limited set of categories for race. These representations are not well aligned with current literature on gender and racial identity.

Nonetheless, I would argue that these categorical variables have utility, *especially* in the study of social inequality. Here is an example of why I think gender qua “male/female” is a flawed but important consideration in global education: <https://www.unicef.org/education/girls-education>.

**Please take a moment to write down your thoughts on the tensions that arise when conceptualizing social constructs such as gender or race as categorical, and I will invite you to share you thoughts in class.**

## 5.3 Binary predictors

Let’s start our interpretation of categorical predictors with the simplest case: a single binary predictor.

Figure 5.1 illustrates the regression of Reading Achievement in Grade 8 (achrdg08) on a binary encoding of Gender (female = 0, male = 1) using the NELS data. There isn’t a lot going on the plot! However, we can see the conditional distributions of Reading Achievement for each value of Gender, and the means of the two groups are indicated.

knitr::include\_graphics("images/reading\_on\_gender.png")

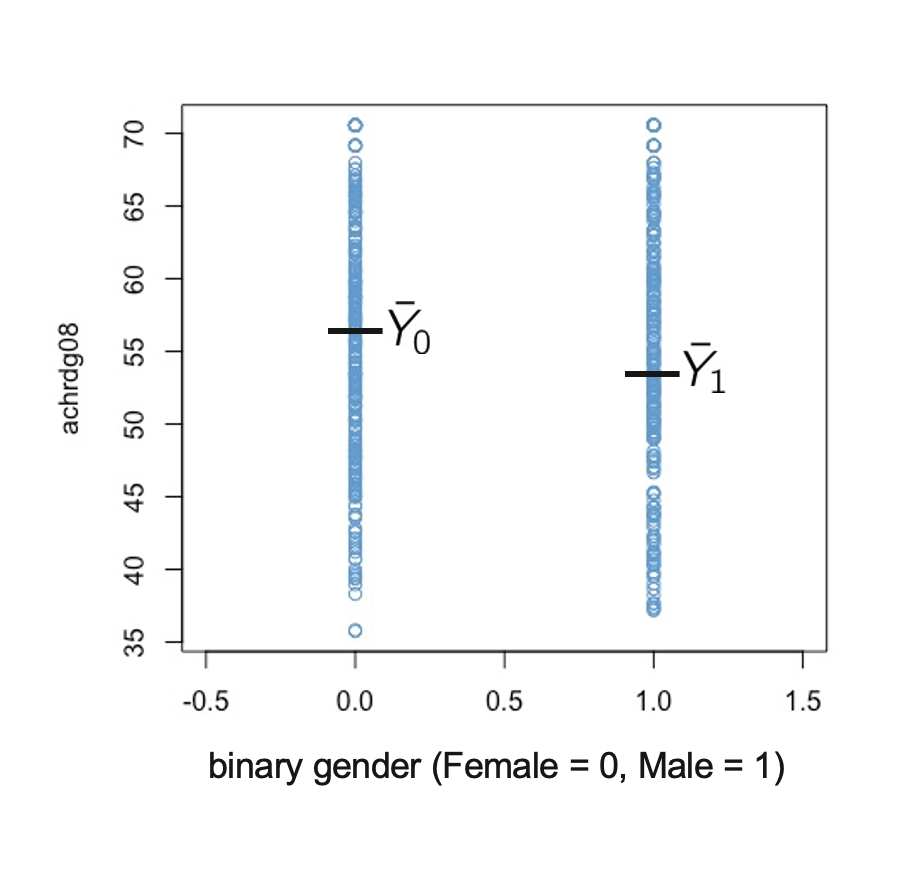


Figure 5.1: Reading Achievement on Binary Gender.

In this situation, the simple regression equation from Section 2.2 still holds

but can only take on one of two values: 0 or 1. The question we want to answer is how to interpret the regression coefficients in this context. The general strategy for approaching this kind of problem has two steps:

* *Step 1.* Plug the values for into the regression equation.
* *Step 2.* Solve for the model parameters in terms the predicted values.

Looking at Equation (5.1) we can conclude that intercept () is equal to the predicted value of Reading Achievement for Females, and Equation (5.2) shows that the regression slope () is equal to the difference between predicted Reading Achievement for Males and Females.

For a single categorical predictor, the predicted values for each category are just the group means on the variable. So, using the notation of Figure 5.1 we can re-write Equations (5.1) and (5.2) as

Note that we will use the equivalence between the predicted values for each category and the mean of the corresponding group throughout this chapter. This equivalence holds when there is only one categorical predictor in the model, and no other predictors. Additional predictors are discussed in the next chapter.

For the example data, regression coefficients are:

attach(NELS)

## The following object is masked \_by\_ .GlobalEnv:  
##   
## urban

# convert "Female / Male" coding to binary  
gender <- NELS$gender  
binary\_gender <- (gender == "Male")\*1  
mod\_binary <- lm(achrdg08 ~ binary\_gender)  
coef(mod\_binary)

## (Intercept) binary\_gender   
## 56.4678 -0.9223

detach(NELS)

**Please take a moment and write down how these two numbers are related to Figure 5.1. In particular, what is equal to, what is equal to, and what is their difference equal to?**

### 5.3.1 Relation with t-tests

Simple regression with a binary predictor is equivalent to conducting an independent samples t-test in which the variable (Gender) is the grouping variable and the variable (Reading Achievement) is the outcome. The following output illustrates this.

For the regression model:

summary(mod\_binary)

##   
## Call:  
## lm(formula = achrdg08 ~ binary\_gender)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.728 -6.147 0.378 6.976 15.005   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 56.468 0.534 105.70 <2e-16 \*\*\*  
## binary\_gender -0.922 0.793 -1.16 0.25   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.83 on 498 degrees of freedom  
## Multiple R-squared: 0.00271, Adjusted R-squared: 0.000708   
## F-statistic: 1.35 on 1 and 498 DF, p-value: 0.245

For the independent samples t-test (with homogeneity of variance assumed):

attach(NELS)

## The following objects are masked \_by\_ .GlobalEnv:  
##   
## gender, urban

t.test(achrdg08 ~ binary\_gender, var.equal = T)

##   
## Two Sample t-test  
##   
## data: achrdg08 by binary\_gender  
## t = 1.2, df = 498, p-value = 0.2  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.6354 2.4801  
## sample estimates:  
## mean in group 0 mean in group 1   
## 56.47 55.55

detach(NELS )

Although the two functions produce different output, we can see that the pattern of values in the two sets of output corresponds to Equations (5.3) and (5.4). In particular, the t-value and p-value for binary\_gender in the regression output is equivalent to the t-value and p-value in the two sample t-test (other than the sign of the t-value).

**If you have any questions about the relation between these two sets of output, please note them now and be prepared ask them in class.**

### 5.3.2 Summary

When doing regression with a binary predictor:

* The intercept is equal to the mean of the group coded “0”
* The regression coefficient is equal to the mean difference between the groups
* Testing is equivalent to testing the mean difference
  + i.e., regression with a binary variable is the same as a t-test of means for independent groups

## 5.4 Reference-group coding

Now that we know how regression with a binary predictor works, let’s consider how to extend this approach to categorical predictors with categories. There are many ways to do this, and the general topic is variously called “contrast coding,” “effect coding,” or “dummy coding.”

The basic idea is to represent the categories of a predictor in terms of dummy variables. Binary coding of a dichotomous predictor is one example of this: We represented a categorical variable with categories using binary predictor.

The most common approach to contrast coding is called *reference-group* coding. In R, the approach is called *treatment contrasts* and is the default coding for categorical predictors.

It is called reference-group coding because:

* The researcher chooses a reference-group.
* The intercept is interpreted as the mean of the reference-group.
* The regression coefficients are interpreted as the mean differences between the other groups and the reference-group.

Note that reference-group coding is a generalization of binary coding. In the example from Section 5.3:

* Females were the reference-group.
* The intercept was equal to the mean Reading Achievement for females.
* The regression coefficient was equal to the mean difference between males and females.

The rest of this section considers how to generalize this approach to greater than 2 groups.

### 5.4.1 A hypothetical example

Figure 5.2 presents a toy data example. The data show the Age and marital status (Mstatus) of 16 hypothetical individuals. Marital status is encoded as

* Single (never married)
* Married
* Divorced

knitr::include\_graphics("images/marital\_status1.png")

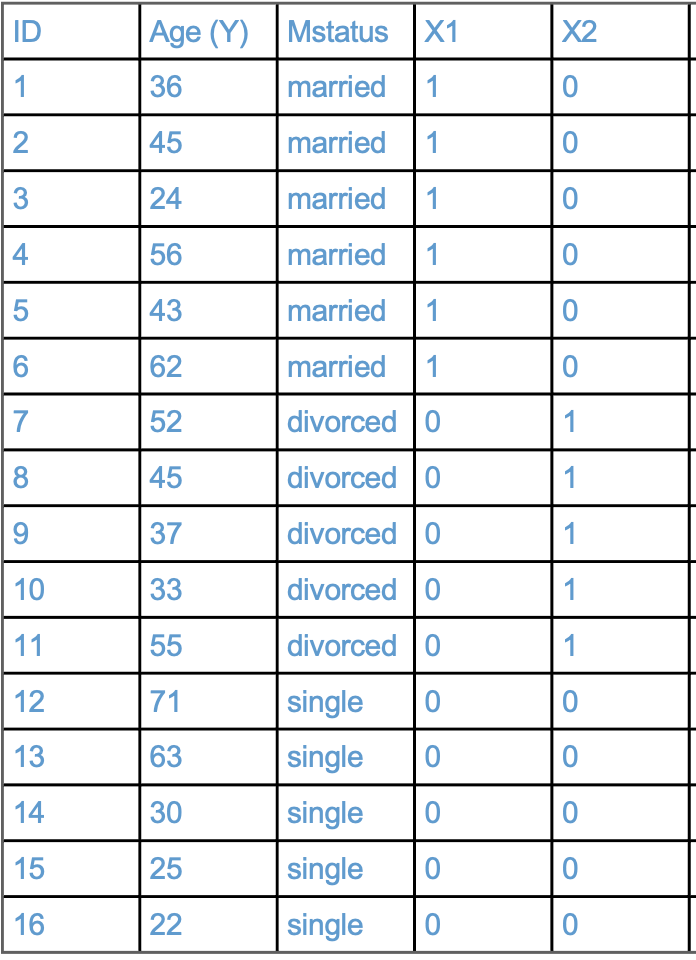


Figure 5.2: Toy Martital Status Example.

These 3 categories are represented by two binary variables, denoted and .

* is a binary variable that is equal to 1 when Mstatus is “married,” and equal to 0 otherwise.
* is a binary variable that is equal to 1 when Mstatus is “divorced,” and equal to 0 otherwise.

The binary variables are often called *dummies* or *indicators*. For example, is a dummy or indicator for married respondents.

In reference-group coding, the group that does not have a dummy variable is the reference group. It is also the group that is coded zero on all of the included dummies.

**What is the is reference-group for this example? Please write down your answer and be prepared to share it in class.**

### 5.4.2 Interpreting the regression parameters

Regressing Age on the dummies we have:

In order to interpret the regression coefficients we can use the same two steps as in Section 5.3

* *Step 1.* Plug the values for the -variables into the regression equation.
* *Step 2.* Solve for the model parameters in terms the predicted values.

**Using the above equations, please write down an interpretation of the regression parameters for the hypothetical example. (Note: this question is not asking for a numerical answer, it is just asking you to put the above equations into words.)**

### 5.4.3 categories

Figure 5.2 extends the toy data example by adding another category for Mstatus (“widowed”).

knitr::include\_graphics("images/marital\_status2.png")

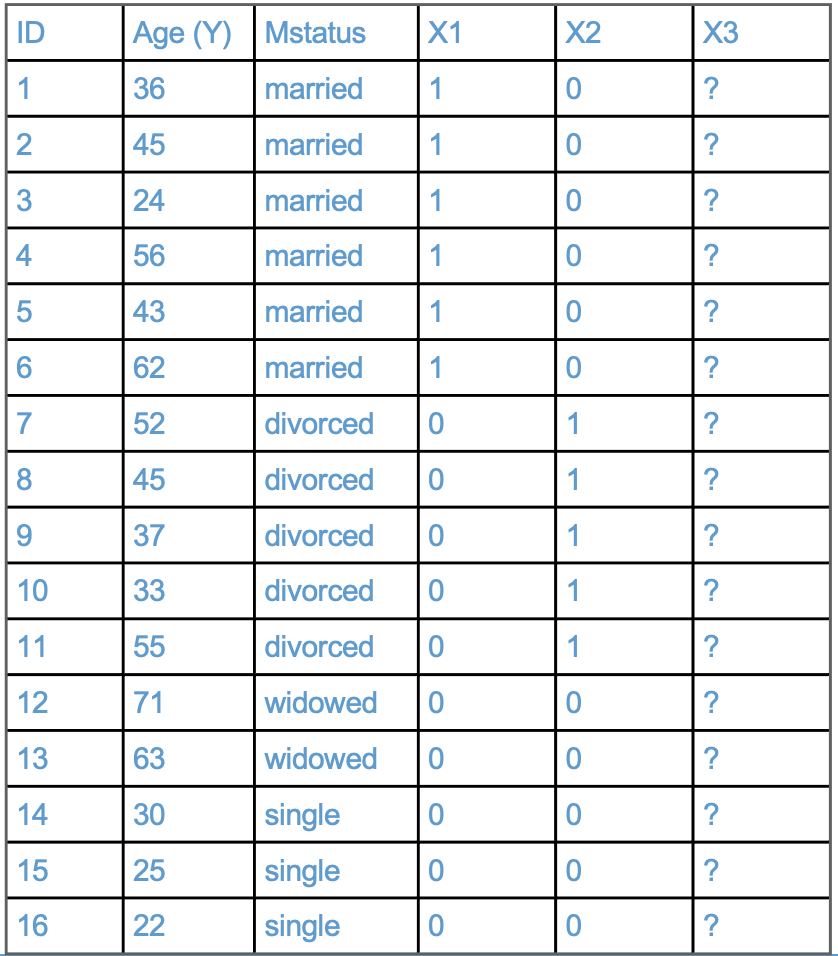


Figure 5.3: Toy Martital Status Example, Part 2.

**Please work through the following questions and be prepared to share your answers in class**

* **How should be coded so that “single” is the reference-group?**
* **Using the two-step approach illustrated above, write out the interpretation of the model parameters in the following regression equation:**
* **Bonus: What would happen if we included dummies for all 4 categories of Mstatus in the model?**

### 5.4.4 Summary

In reference-group coding with a single categorical variable:

* The reference is group is chosen by the analyst – it is the group that is coded zero on all dummies, or the one that has its dummy left out of the dummies used in the model.
* The intercept is interpreted as the mean of the reference group.
* The regression coefficients of the dummy variables are interpreted as the difference between the mean of the indicated group and the mean of the reference group.

## 5.5 Deviation coding

In some cases there is a clear reference group (e.g., in experimental conditions, comparisons are made to the control group). But in other cases, it is not so clear what the reference group should be. In both of the examples we have considered, the choice of reference group was arbitrary.

In such cases it can be preferable to use different types of contrast coding that do not require a reference group. One approach to getting rid of the reference-group is called *deviation coding*. In R this is called *sum-to-zero constrasts*, or *sum* contrasts for short. )

In deviation coding:

* The intercept is equal to the mean of the predicted values for each category i.e,

$$

* The regression coefficients compare each group to the intercept.

The main difference compared to reference-group coding is the interpretation of the intercept – it is no longer an arbitrarily chosen reference-group, but instead represents the mean of the predicted values. For a single predictor, this is equal to overall mean on , when the groups have equal sample size ():

This equation says that, when the groups have equal sample sizes, the deviation-coded intercept is equal to the overall mean . Consequently, the regression coefficients are interpreted as the deviation of each group mean from the overall mean. This is why it is called deviation coding.

When the groups have unequal sample size, the situation is a bit more complicated. In particular, we have to weight the predicted values in Equation (5.5) by the group sample sizes. This is addressed in 5.5.4 (optional). To clarify that intercept in deviation coding is not always equal to , we refer to it as an “unweighted mean” of group means / predicted scores.

Note that there are still only regression coefficients. So one group gets left out of the analysis, and the researcher has to chose which one. This is a shortcoming of deviation coding, which is addressed in the Section 5.5.5 (optional)

### 5.5.1 A hypothetical example

The International Development and Early Learning Assessment (IDELA) is an assessment designed to measure young children’s development in literacy, numeracy, social-emotional, and motor domains, in international settings. Figure 5.4 shows the countries in which the IDELA had been used as of 2017 (<https://www.savethechildren.net/sites/default/files/libraries/GS_0.pdf>.)

knitr::include\_graphics("images/idela\_map.png")



Figure 5.4: IDELA Worldwide Usage, 2017.

If our goal was to compare countries’ IDELA scores, it would be difficult to agree on which country should serve as the reference group to which the others are compared. Therefore, it would be preferable to avoid the problem of choosing a reference group altogether. In particular, deviation coding let’s us compare each country’s mean IDELA score to the (unweighted) mean over all of the countries.

Figure 5.5 presents a toy data example. The data show the IDELA scores and Country for 16 hypothetical individuals. The countries considered in this example are

* Ethiopia
* Vietnam
* Boliva

These 3 countries are represented by two binary variables, denoted and .

* is a dummy for Ethiopia
* is a dummy for Vietnam

knitr::include\_graphics("images/idela1.png")

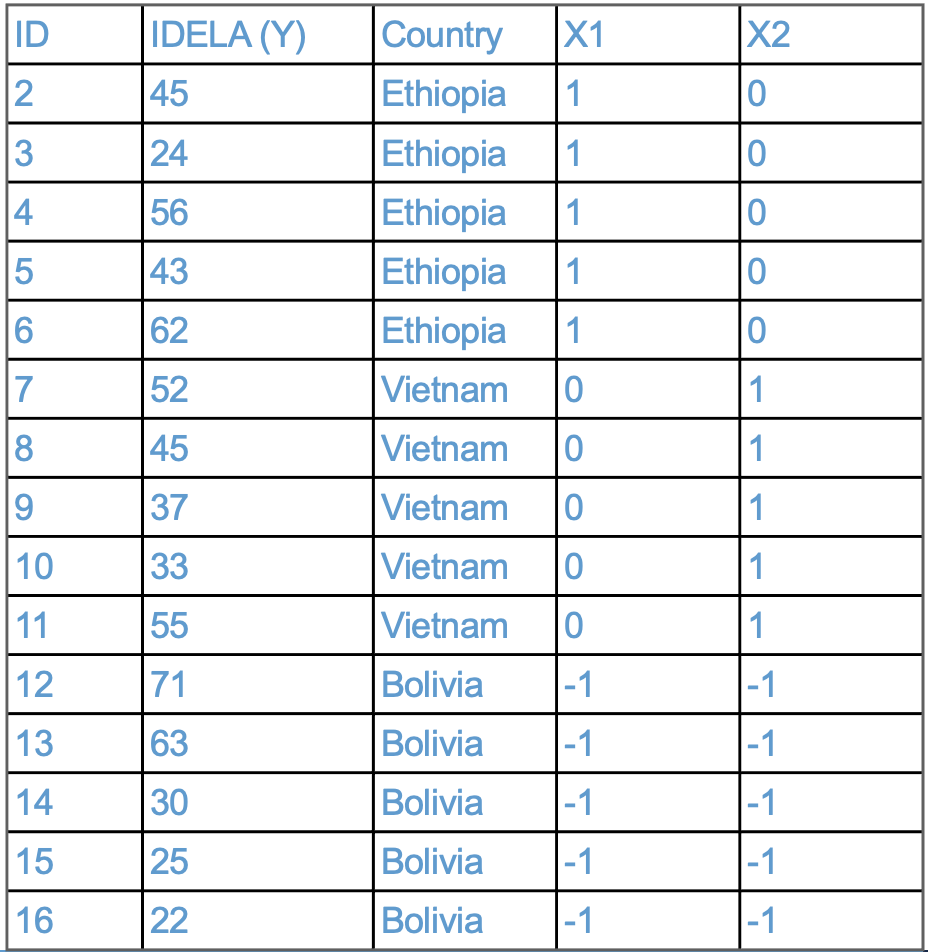


Figure 5.5: Toy IDELA Example.

Note that the dummy variables are different than for the case of reference-group coding discussed Section 5.4. In deviation coding, the dummies always take on values . The same group must receive the code for all dummies. The group with the value is the group that gets left out of the analysis.

### 5.5.2 Interpreting the regression parameters

Regressing IDELA on the dummies we have:

In order to interpret the regression coefficients we proceed using the same two steps as in Section 5.3 and Section 5.4

* *Step 1.* Plug the values for the -variables into the regression equation.
* *Step 2.* Solve for the model parameters in terms the predicted values.

$$$$

**Using the above equations, please write down an interpretation of the regression parameters for the hypothetical example. In particular, what do you think about using the unweighted mean of countries’ predicted IDELA scores as the comparison point? Is this meaningful? Would another approach be better?**

### 5.5.3 Summary

In deviation coding with a single categorical variable:

* The intercept is interpreted as the unweighted mean of the groups’ means, which is equal to the overall mean on the variable when the groups have equal sample sizes.
* The regression coefficients of the dummy variables are interpreted as the difference between the mean of the indicated group and the unweighted mean of the groups.
* There are still only regression coefficients, so one group gets left out (see extra material for how to get around this).

### 5.5.4 Extra: Deviation coding with unequal sample sizes\*

When groups have unequal sample size, the unweighted mean of the group means is not the overall mean of the Y variable. This is not always a problem. For example, in the international comparisons example, it is reasonable (i.e., democratic) that each country should receive equal weight, even if the size of their populations differ.

However, if you want to compare each groups’ mean to the overall mean on , deviation coding can be adjusted by replacing the value with the ratio of indicated group’s sample size to the omitted group’s sample size. An example for 3 groups is shown below.

You can use the 2-step procedure to show that this coding, called *weighted deviation coding*, results in

Replacing with you can also show that , using the rules of summation algebra. Unlike the case for the unweighted mean, this relationship holds regardless of the sample sizes of the groups.

### 5.5.5 Extra: Deviation coding all groups included\*

Another issue with deviation coding is that it requires leaving one group out of the model. This is a shortcoming of the approach. As a work around, one can instead use the following approach. Note that this approach will affect the value, statistical significance, and interpretation of R-squared, so you should only use it if you aren’t interested in reporting R-squared.

* Step A: Standardize the variable to have M = 0 and SD = 1.
* Step B: Compute binary dummy variables (reference-group coding) for all groups,
* Step C: Regress on the dummy variables, without the intercept in the model

It is easy to show that the regression coefficients are just the means of the indicated group. Since the overall mean of is zero (see Step A), the group means can be interpreted as deviations from the overall mean on .

Note that to omit the intercept in R, you can use the formula syntax

Y ~ -1 + X1 + ...

in the lm function. The -1 removes the intercept from the model. Again, keep in mind that this will make the R-squared uninterpretable.

## 5.6 Workbook

This section collects the questions asked in this chapter. We will discuss these questions in class. If you haven’t written down / thought about the answers to these questions before class, the lesson will not be very useful for you! So, please engage with each question by writing down one or more answers, asking clarifying questions, posing follow up questions, etc.

**Section 5.2**

* Please take a moment to write down your thoughts on the tensions that arise when conceptualizing social constructs such as gender or race as categorical, and I will invite you to share you thoughts in class.

**Section 5.3**

* Please take a moment and write down how these two numbers in the R output are related to the Figure (reproduced below). In particular, what is equal to, what is equal to, and what is their difference equal to?

# regression coefficients from Reading Achievement on Binary Gender  
coef(mod\_binary)

## (Intercept) binary\_gender   
## 56.4678 -0.9223

knitr::include\_graphics("images/reading\_on\_gender.png")

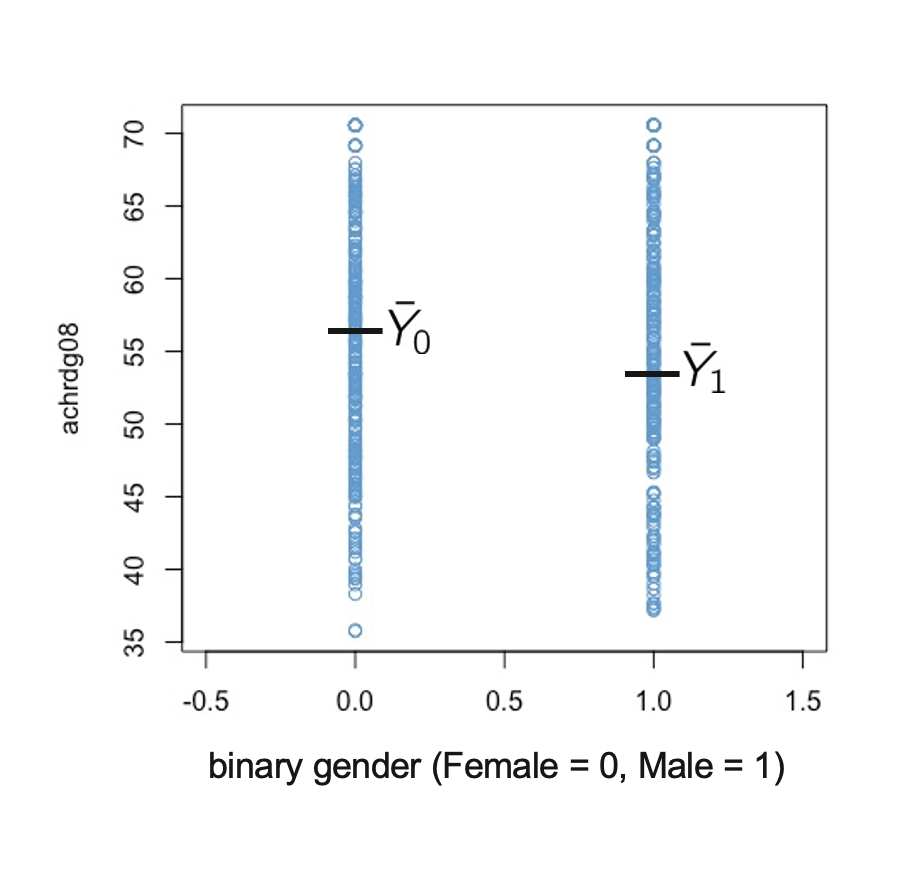


Figure 5.6: Reading Achievement on Binary Gender.

* If you have any questions about the relation between regression with a binary predictor and an independent samples t-test (output below), please note them now and be prepared ask them in class.

summary(mod\_binary)

##   
## Call:  
## lm(formula = achrdg08 ~ binary\_gender)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.728 -6.147 0.378 6.976 15.005   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 56.468 0.534 105.70 <2e-16 \*\*\*  
## binary\_gender -0.922 0.793 -1.16 0.25   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.83 on 498 degrees of freedom  
## Multiple R-squared: 0.00271, Adjusted R-squared: 0.000708   
## F-statistic: 1.35 on 1 and 498 DF, p-value: 0.245

t.test(achrdg08~binary\_gender, var.equal = T, data = NELS)

##   
## Two Sample t-test  
##   
## data: achrdg08 by binary\_gender  
## t = 1.2, df = 498, p-value = 0.2  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.6354 2.4801  
## sample estimates:  
## mean in group 0 mean in group 1   
## 56.47 55.55

**Section 5.4**

* What is the is reference-group in the example below? Please write down your answer and be prepared to share it in class.

knitr::include\_graphics("images/marital\_status1.png")

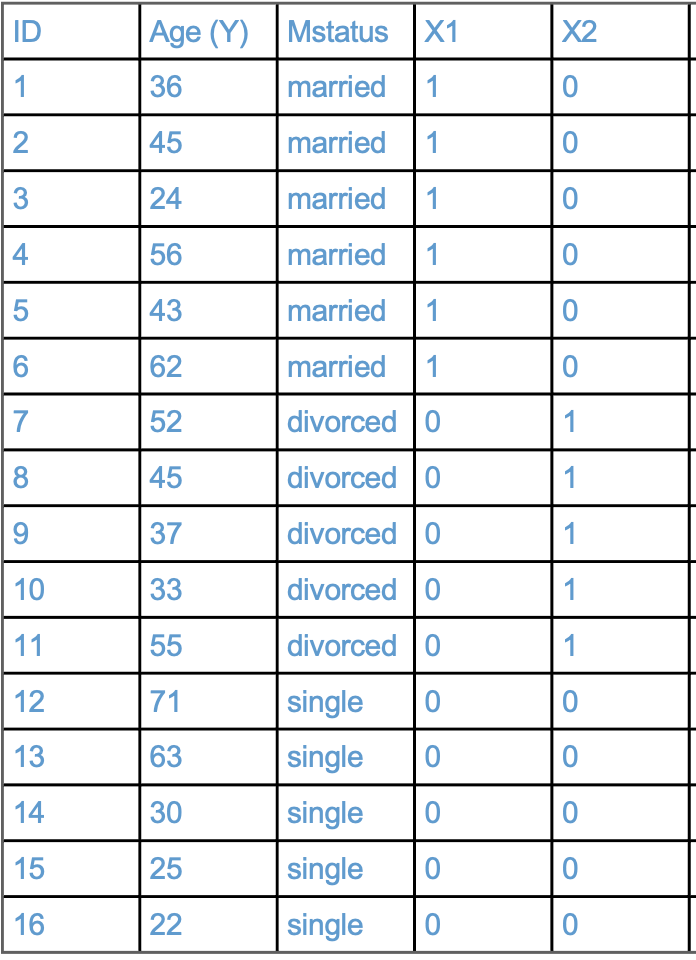


Figure 5.7: Toy Martital Status Example.

* Using the equations below, please write down an interpretation of the regression parameters for the hypothetical example. (Note: this question is not asking for a numerical answer, it is just asking you to put the above equations into words.)
* Using the example below:
  + How should be coded so that “single” is the reference-group?
  + Using the two-step approach illustrated above, write out the interpretation of the model parameters in the following regression equation:
  + Bonus: What would happen if we included dummies for all 4 categories of Mstatus in the model?

knitr::include\_graphics("images/marital\_status2.png")

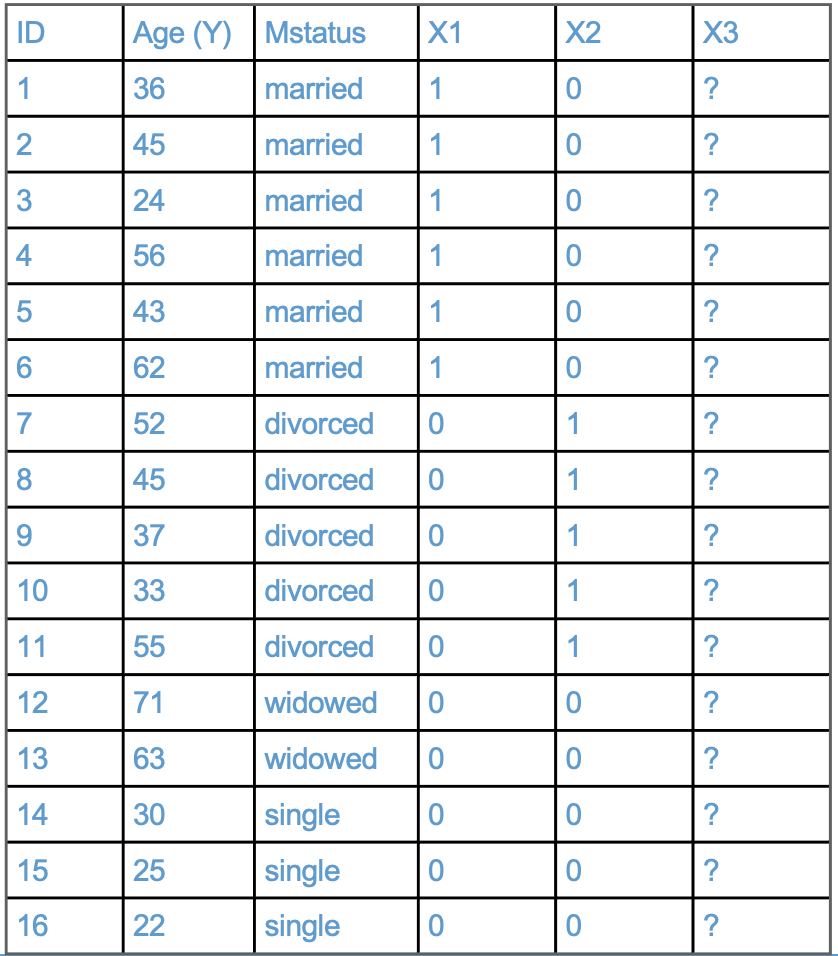


Figure 5.8: Toy Martital Status Example, Part 2.

**Section 5.5**

* Using the equations (below), please write down an interpretation of the regression parameters for the hypothetical example. In particular, what do you think about using the unweighted mean of countries’ predicted IDELA scores as the comparison point? Is this meaningful? Would another approach be better?

$$$$

## 5.7 Exercises

These exercises provide an overview of contrast coding with categorical predictors in R. Two preliminary topics are also discussed: linear regression with a binary predictor, and the factor data class in R.

### 5.7.1 A single binary predictor

Regression with a single binary predictor is equivalent to an independent groups t-test of a difference between means. Let’s illustrate this using reading achievement in grade 8 (achrdg08) and gender in the NELS dataset. Note that, although the construct gender need not be conceptualized as dichotomous or even categorical, the variable gender reported in NELS data is dichotomous, with values “Female” and “Male.”

R treats variables like gender as “factors.” Factors have special properties that we are going to work with later on, but for now let’s recode gender to binary\_gender by setting Female = 0 and Male = 1

### 5.7.2 Recoding a factor to numeric

load("NELS.RData")  
attach(NELS)  
  
# Note that gender is non-numeric -- it is a factor with levels "Female" and "Male"  
head(gender)

## [1] Male Female Male Female Male Female  
## Levels: Female Male

class(gender)

## [1] "factor"

# A trick to create a binary indicator for males   
binary\_gender <- (gender == "Male") \* 1   
  
# Check that the two variables are telling us the same thing  
table(gender, binary\_gender)

## binary\_gender  
## gender 0 1  
## Female 273 0  
## Male 0 227

It is often hassle to get from factor to numeric or vice versa. Above we used the trick of first creating from the factor a logical vector (gender == "Male") and then coercing the logical vector to binary by multiplying it by 1. Other strategies can be used. See help(factor) for more information.

### 5.7.3 Binary predictors and independent samples t-tests

Now let’s get back to regressing reading achievement (achrdg08) on binary\_gender, and comparing this to an independent groups t-test using the same two variables.

# Regression with a binary variable  
mod1 <- lm(achrdg08 ~ binary\_gender)  
summary(mod1)

##   
## Call:  
## lm(formula = achrdg08 ~ binary\_gender)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.728 -6.147 0.378 6.976 15.005   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 56.468 0.534 105.70 <2e-16 \*\*\*  
## binary\_gender -0.922 0.793 -1.16 0.25   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.83 on 498 degrees of freedom  
## Multiple R-squared: 0.00271, Adjusted R-squared: 0.000708   
## F-statistic: 1.35 on 1 and 498 DF, p-value: 0.245

# Compare to the output of t-test  
t.test(achrdg08 ~ binary\_gender, var.equal = T)

##   
## Two Sample t-test  
##   
## data: achrdg08 by binary\_gender  
## t = 1.2, df = 498, p-value = 0.2  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.6354 2.4801  
## sample estimates:  
## mean in group 0 mean in group 1   
## 56.47 55.55

Note that:

* The intercept in mod1 is equal the mean of the group coded 0 (females) in the t-test:
* The b-weight in mod1 is equal to difference between the means:
* The t-test of the b-weight, and its p-value, are equivalent to the t-test mean difference (except for the sign):

In summary, a t-test of a b-weight of a binary predictor is equal equivalent to a t-test of the mean difference.

### 5.7.4 Reference-group coding

As discussed in Section 5.4, the basic idea of contrast coding is to replace a categorical variable with categories with “dummy” variables. In reference-group coding, the dummy variables are binary, and the resulting interpretation is:

* The intercept is interpreted as the mean of the reference group. The reference is group is chosen by the analyst – it is the group that is coded zero on all dummies, or the group whose dummy variable is “left out” of the dummy variables.
* The regression coefficients of the dummy variables are interpreted as the difference between the mean of the indicated group and the mean of the reference group

Reference-group coding is the default contrast coding in R. However, in order for contrast coding to be implemented, our categorical predictor needs to be represented in R as a factor.

### 5.7.5 More about factors

Factors are the way that R deals with categorical data. If you want to know if your variable is a factor or not, you can use the functions class or is.factor. Let’s illustrate this with the urban variable from NELS.

# Two ways of checking what type a variable is   
class(urban)

## [1] "factor"

is.factor(urban)

## [1] TRUE

# Find out the levels of a factor using "levels"  
levels(urban)

## [1] "Rural" "Suburban" "Urban"

# Find out what contrast coding R is using for a factor "constrasts"  
contrasts(urban)

## Suburban Urban  
## Rural 0 0  
## Suburban 1 0  
## Urban 0 1

In the above code, we see that urban is a factor with 3 levels and the default reference-group contrasts are set up so that Rural is the reference group. Below we will show how to change the contrasts for a variable.

If we are working with a variable that is not a factor, but we want R to treat it as a factor, we can use the factor command. Let’s illustrate this by turning binary\_gender back into a factor.

# Change a numeric variable into a factor  
class(binary\_gender)

## [1] "numeric"

factor\_gender <- factor(binary\_gender)  
class(factor\_gender)

## [1] "factor"

levels(factor\_gender)

## [1] "0" "1"

We can also use the levels function to tell R what labels we want it to use for our factor. levels should be assigned a text vector with length equal to the number of levels of the variable. The entries of the assigned vector will be the new level names of the factor.

# Change the levels of factor\_gender to "F" and "M"  
levels(factor\_gender) <- c("Female", "Male")  
levels(factor\_gender)

## [1] "Female" "Male"

### 5.7.6 Back to reference-group coding

OK, back to reference coding. Let’s see what lm does when we regress achrdg08 on urban.

mod2 <- lm(achrdg08 ~ urban)  
summary(mod2)

##   
## Call:  
## lm(formula = achrdg08 ~ urban)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.30 -6.16 0.21 6.79 15.56   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 54.993 0.688 79.88 <2e-16 \*\*\*  
## urbanSuburban 0.668 0.912 0.73 0.464   
## urbanUrban 3.128 1.048 2.98 0.003 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.76 on 497 degrees of freedom  
## Multiple R-squared: 0.019, Adjusted R-squared: 0.0151   
## F-statistic: 4.82 on 2 and 497 DF, p-value: 0.00841

In the output, we see that two regression coefficients are reported, one for Suburban and one for Urban. As discussed in Section 5.4, these coefficients are the mean difference between the indicated group and the reference group (Rural).

We can see that Urban students scores significantly higher than Rural students (3.125 percentage points), but there was no significant difference between Rural and Suburban students.

The intercept is the mean of the reference group (rural) – about 55% on the reading test.

Note the R-squared – Urbanicity accounts for about 2% of the variation reading achievement. As usual, the F-test of R-squared has degrees of freedom and , but now (the number of predictors) is equal to – the number of categories minus one.

### 5.7.7 Changing the reference group

What if we wanted to use a group other than Rural as the reference group? We can chose a different reference group using the cont.treatment function. This function takes two arguments

* n tells R how many levels there
* base tells R which level should be the reference group

# The current reference group is Rural  
contrasts(urban)

## Suburban Urban  
## Rural 0 0  
## Suburban 1 0  
## Urban 0 1

# Chance the reference group to the Urban (i.e., the last level)  
contrasts(urban) <- contr.treatment(n = 3, base = 3)  
contrasts(urban)

## 1 2  
## Rural 1 0  
## Suburban 0 1  
## Urban 0 0

Note that when we first ran contrasts(urban), the column names were names of the levels. But after changing the reference group, the column names are just the numbers 1 and 2. To help interpret the lm output, it is helpful to name the contrast levels appropriately

# Naming our new contrasts  
colnames(contrasts(urban)) <- c("Rural", "Suburban")  
contrasts(urban)

## Rural Suburban  
## Rural 1 0  
## Suburban 0 1  
## Urban 0 0

Now we are ready to run our regression again, this time using a different reference group.

mod3 <- lm(achrdg08 ~ urban)  
summary(mod3)

##   
## Call:  
## lm(formula = achrdg08 ~ urban)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.30 -6.16 0.21 6.79 15.56   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 58.120 0.790 73.56 <2e-16 \*\*\*  
## urbanRural -3.128 1.048 -2.98 0.003 \*\*   
## urbanSuburban -2.460 0.991 -2.48 0.013 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.76 on 497 degrees of freedom  
## Multiple R-squared: 0.019, Adjusted R-squared: 0.0151   
## F-statistic: 4.82 on 2 and 497 DF, p-value: 0.00841

Compared to the output from mod2, note that

* The intercept now represents the mean reading scores of the Urban group, because this is the new reference group.
* The regression coefficients now represent the mean difference between the indicated group with the new reference group.
* The R-square and F stay the same – in other words, the total amount of variation explained by the variable urban does not change, just because we changed the reference group.

### 5.7.8 Deviation coding

It is possible to change R’s default contrast coding to one of the other built-in contrasts (see help(contrasts) for more information on the built in contrasts).

For instance, to change to deviation coding, we use R’s contr.sum function and tell it how many levels there are for the factor (n). In deviation coding, the intercept is equal to the unweighted mean of the predicted values, and the regression coefficients are difference between the indicated group and the unweighted mean.

contrasts(urban) <- contr.sum(n = 3)  
contrasts(urban)

## [,1] [,2]  
## Rural 1 0  
## Suburban 0 1  
## Urban -1 -1

# As above, it is helpful to name the contrasts using "colnames"  
colnames(contrasts(urban)) <- c("Rural", "Suburban")

Now we are all set to use deviation coding with lm.

mod4 <- lm(achrdg08 ~ urban)  
summary(mod4)

##   
## Call:  
## lm(formula = achrdg08 ~ urban)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.30 -6.16 0.21 6.79 15.56   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 56.258 0.402 139.90 <2e-16 \*\*\*  
## urbanRural -1.265 0.565 -2.24 0.026 \*   
## urbanSuburban -0.597 0.530 -1.13 0.260   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.76 on 497 degrees of freedom  
## Multiple R-squared: 0.019, Adjusted R-squared: 0.0151   
## F-statistic: 4.82 on 2 and 497 DF, p-value: 0.00841

Note the following things about the output:

* The regression coefficients compare each group’s mean to the unweighted mean of the groups. Rural and Suburban students are below the unweighted mean, but the difference is only significant for Rural.
* Although the output looks similar to that of mod3, the coefficients are all different and they all have different interpretations. The lm output doesn’t tell us this, we have to know what is going on under the hood.
* The R-square does not change from mod2 – again, the type of contrast coding used doesn’t affect how much variation is explained by the predictor.

### 5.7.9 Extra: Relation to ANOVA

As our next exercise, let’s compare the output of lm and the output of aov – R’s module for Analysis of Variance. If regression and ANOVA are really doing the same thing, we should be able to illustrate it with these two modules. Note that if you check out help(aov), it explicitly states that aov uses the lm function, so, finding that the two approaches give similar output shouldn’t be a big surprise!

# Run our model as an ANOVA  
aov1 <- aov(achrdg08 ~ urban)  
  
# Compare the output with lm  
summary(aov1)

## Df Sum Sq Mean Sq F value Pr(>F)   
## urban 2 741 370 4.82 0.0084 \*\*  
## Residuals 497 38163 77   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(mod2)

##   
## Call:  
## lm(formula = achrdg08 ~ urban)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.30 -6.16 0.21 6.79 15.56   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 54.993 0.688 79.88 <2e-16 \*\*\*  
## urbanSuburban 0.668 0.912 0.73 0.464   
## urbanUrban 3.128 1.048 2.98 0.003 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.76 on 497 degrees of freedom  
## Multiple R-squared: 0.019, Adjusted R-squared: 0.0151   
## F-statistic: 4.82 on 2 and 497 DF, p-value: 0.00841

If we compare the output from aov to the F-test reported by lm we see that the F-stat, degrees of freedom, and p-value are all identical. If we compute eta-squared from aov, we also find that it is equal to the R-squared value from lm.

# Compute eta-squared from aov output  
741 / (741 + 38163)

## [1] 0.01905

In short, ANOVA and regression are doing the same thing: R-squared is the same as omega-squared and the ANOVA omnibus F-test is the same as the F-test of R-squared. The main difference is that lm focuses on contrasts for analyzing and interpreting the group differences, whereas ANOVA focuses on the F-test of the omnibus hypothesis and the procedures for analyzing group difference are conducted as a follow-up step.

### 5.7.10 Extra: Group-mean coding

As a step towards addresses the issues with deviation coding, let’s consider another coding procedure. If we omit the intercept term, all of the reference-group coded dummies can be included and the regression coefficients now correspond to means of each group. We omit the intercept using -1 in the model formula.

# Omit the intercept using -1  
mod5 <- lm(achrdg08 ~ -1 + urban)  
summary(mod5)

##   
## Call:  
## lm(formula = achrdg08 ~ -1 + urban)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.30 -6.16 0.21 6.79 15.56   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## urbanRural 54.993 0.688 79.9 <2e-16 \*\*\*  
## urbanSuburban 55.660 0.598 93.1 <2e-16 \*\*\*  
## urbanUrban 58.120 0.790 73.6 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.76 on 497 degrees of freedom  
## Multiple R-squared: 0.976, Adjusted R-squared: 0.976   
## F-statistic: 6.82e+03 on 3 and 497 DF, p-value: <2e-16

Note the following things about the output:

* The coefficients are no **longer interpreted mean differences**, just the raw means.
* The R-square and F-test **do** change from mod2. When the intercept is omitted, R-squared can no longer interpretated as a proportion of variance unless the variable variable is centered. When the intercept is omitted, it also changes the degrees of freedom in the F-test, because there are now 3 predictors instead of 2. In general, when the intercept is omitted, R-square and its F-test do not have the usual interpretation. When reporting R-squared and its F-test, we should use a model with the intercept.

By itself, group mean coding is not very interesting uninteresting – we don’t usually want to test whether the group means are different from zero. However, we will see in the next section that it can be used to provide an alternative to deviation coding.

### 5.7.11 Extra: Weighted versus unweighted deviation coding

This section presents a “hack” for addressing the two main issues with deviation coding noted in Section 5.5. This is a hack in the sense that we are working around R’s usual procedures rather than replacing them with a completely new procedure. The result of this hack is to provide deviation coding in which comparisons are made to grand mean on the outcome for all categories, not just categories. As a side effect, the F-test for the R-squared statistic is no longer correct, so you should not use this approach to report R-squared.

First, note that the group sample sizes are not equal for ubran

table(urban)

## urban  
## Rural Suburban Urban   
## 162 215 123

Because of this, the unweighted mean of the group means (i.e., the intercept in mod4 above) is not equal to the overall mean on achrdg08 – we can see that the overall mean and the unweighted group means are slightly different for these data:

group\_means <- tapply(achrdg08, INDEX = urban, FUN = mean)  
group\_means

## Rural Suburban Urban   
## 54.99 55.66 58.12

# Compare the overall mean and the unweighted mean of the group means  
# Center the outcome variable  
ybar <- mean(achrdg08)  
ybar

## [1] 56.05

mean(group\_means)

## [1] 56.26

Note that the intercept in mod4 is not equal to the overall mean, ybar, but is instead equal to the unweighted average of the group means, mean(group\_means). the difference isnt very big in this example, but it can be quite drastic with highly unequal sample sizes.

If you would like to compare the groups to the overall mean when the sample sizes are unequal, the simplest way to do this in R is by first centering the Y variable and then using group mean coding.

After centering, the grand mean of is zero, and so the group means represent deviations from the grand mean (i.e., deviations from zero) and the tests of the regression coefficients are tests of whether the group means are different from the grand mean.

# Change the contrasts back to reference group  
contrasts(urban) <- contr.treatment(n = 3, base = 1)  
colnames(contrasts(urban)) <- c("Suburban", "Urban")  
  
# Center the outcome variable  
dev\_achrdg08 <- achrdg08 - ybar  
  
# Run the regression with group mean coding  
mod6 <- lm(dev\_achrdg08 ~ -1 + urban)  
summary(mod6)

##   
## Call:  
## lm(formula = dev\_achrdg08 ~ -1 + urban)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.30 -6.16 0.21 6.79 15.56   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## urbanRural -1.056 0.688 -1.53 0.126   
## urbanSuburban -0.389 0.598 -0.65 0.516   
## urbanUrban 2.071 0.790 2.62 0.009 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.76 on 497 degrees of freedom  
## Multiple R-squared: 0.019, Adjusted R-squared: 0.0131   
## F-statistic: 3.22 on 3 and 497 DF, p-value: 0.0226

Note the following things about the output:

* The regression coefficients compare each group’s mean to the overall mean on the outcome. Urban students are significantly above average, Rural and Suburban students are below average but the difference is not significant.
* The R-square is the same as mod2 because the Y variable is centered, but the F test **does** change from mod2, because there are now 3 variables included in the model. In general, when the intercept is omitted, R-square and its F-test do not have the usual interpretation. So, when reporting R-squared and F-test, we should use a model with the intercept, rather than the approach outlined here.

# 6 Interactions

In statistics, the term *interaction* means that the relationship between two variables depends on a third variable. In the context of regression, we are usually interested in the situation where the relationship between the outcome and a predictor depends on the value of another predictor . This situation is also referred to as *moderation* or sometimes as *effect heterogeneity*.

Interactions are a “big picture” idea with a lot conceptual power, especially when describing topics related to social inequality or “gaps.” Some examples of interactions are:

* The relationship between wages and years of education depends on gender. (<https://en.wikipedia.org/wiki/Gender_pay_gap>)
* The relationship between reading achievement and age depends on race (<https://cepa.stanford.edu/educational-opportunity-monitoring-project/achievement-gaps/race/>)
* The effect of COVID-19 school shutdowns on academic achievement depends on SES. (<https://www.mckinsey.com/industries/education/our-insights/covid-19-and-student-learning-in-the-united-states-the-hurt-could-last-a-lifetime>)

This chapter starts by considering what happens when both categorical and continuous predictors are used in a model, and uses this combination of predictors as a way of digging into the math behind interactions. Later sections will consider what happens when we have interactions between two continuous predictors, or two categorical predictors.

## 6.1 An example from NELS

There is well-documented gender gap in STEM achievement by the end of high school. The t-test reported below illustrates the gap in Math Achievement using the NELS data.

load("NELS.RData")  
t.test(achmat12 ~ gender, var.equal = T, data = NELS)

##   
## Two Sample t-test  
##   
## data: achmat12 by gender  
## t = -4.6, df = 498, p-value = 7e-06  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -4.527 -1.798  
## sample estimates:  
## mean in group Female mean in group Male   
## 55.47 58.63

In this chapter, our first goal is to use linear regression to better understand this gender gap in Math Achievement. To do this, we consider a third variable, Reading Achievement. The plot below shows the relationship between Math Achievement and Reading Achievement estimated just for males (Blue), and the same relationship estimated just for females (Black).

attach(NELS)  
females <- gender == "Female"  
males <- gender == "Male"  
  
mod1 <- lm(achmat12[females] ~ achrdg12[females])  
mod2 <- lm(achmat12[males] ~ achrdg12[males])  
  
# Plot reading and math for females  
plot(achrdg12[females], achmat12[females], xlab = "Reading", ylab = "Math")  
abline(mod1, lwd = 2)  
  
# Add points and line for males  
points(achrdg12[males], achmat12[males], col = "#4B9CD3", pch = 2)  
abline(mod2, col = "#4B9CD3", lwd = 2)  
  
# Add a legend  
legend(x = "topleft", legend = levels(gender), pch = c(1, 2), col = c(1, "#4B9CD3"))

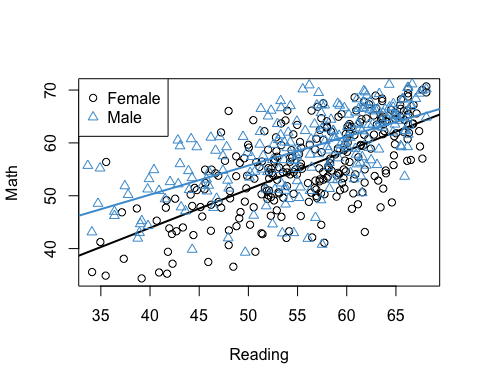


Figure 6.1: Math Achievement, Reading Achievement, and Gender.

detach(NELS)

**Take a minute to think about what this plot is telling us about the relationships among Math Achievement, Reading Achievement, and Gender.**

* **Is the gender gap in math constant?**
* **Is the relationship between math and reading the same for males and females?**
* **Can you provide a summary of the plot in terms of what it says about the gender gap in Math Achievement?**

**Please write down your answers to these questions and be prepared to share them in class.**

Note that in the figure above, we estimated two separate simple regression models, one just for males and one just for females. In the next few sections, we will work our way towards a single multiple regression model that can be used to represent the relationships among these three variables.

## 6.2 Binary + continuous

As a first step, let’s consider what happens when we include both Gender and Reading Achievement as predictors of Math Achievement in our usual multiple regression equation:

where

* is Math Achievement in grade 12
* is Reading Achievement in grade 12
* is Gender (binary, with Female = 0 and Male = 1)

Note that this model does **not** include an interaction between the two predictors – we are first going to consider what is “missing” from the usual regression model, and then use this to motivate inclusion of another predictor that represents the interaction. To get an initial sense of what is missing, the model in Equation (6.1) is plotted Figure 6.2 – can you spot the difference with Figure 6.1?

attach(NELS)  
mod3 <- lm(achmat12 ~ achrdg12 + gender, data = NELS)  
a\_females <- coef(mod3)[1]  
b\_females <- coef(mod3)[2]  
  
a\_males <- a\_females + coef(mod3)[3]  
b\_males <- b\_females   
  
# Plot reading and math for females  
plot(achrdg12[females], achmat12[females], xlab = "reading", ylab = "math")  
abline(a\_females, b\_females, lwd = 2)  
  
# Add points and line for males  
points(achrdg12[males], achmat12[males], col = "#4B9CD3", pch = 2)  
abline(a\_males, b\_males, col = "#4B9CD3", lwd = 2)  
  
# Add a legend  
legend(x = "topleft", legend = levels(gender), pch = c(1, 2), col = c(1, "#4B9CD3"))

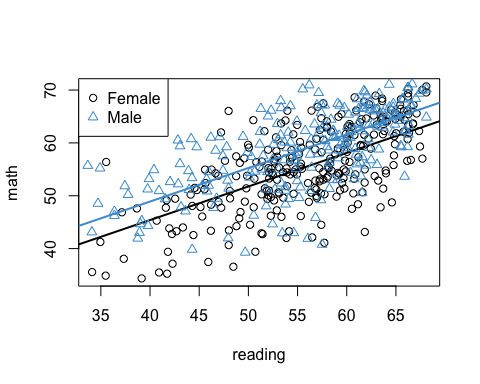


Figure 6.2: Math Achievement, Reading Achievement, and Gender (No Interaction).

detach(NELS)

In order to interpret our multiple regression model, we can take the same two-step approach we used to interpret categorical predictors in Chapter 5. First, we plug in values for the categorical predictor, then we use the resulting equations solve for quantities of interest. In particular,

The equations for and are referred to *simple trends* or *simple slopes*. These describe the regression of Math on Reading, simply for Males, or simply for Females. This usage of the term “simple” is related to the simple regression model with only one predictor, but it is also used more widely. The difference between the two simple regression equations is, in this context, the predicted gender gap in Math Achievement.

From these equations we can see that:

* The simple trends for Females and Males have different intercepts, because the regression coefficient for Gender () is added to the intercept for Males.
* The simple trends for Females and Males have the same slope, meaning that the regression lines are parallel (see Figure 6.1).
* The difference between the predicted values (i.e., the predicted gender gap in Math Achievement) is a constant, and is equal to the regression coefficient for Gender ()

The regression coefficients for the example data are shown below. **Please use these numbers to provide an interpretation of the simple trends and the gender gap in Math Achievement for the NELS example** (Don’t worry about statistical significance, just focus on the meaning of the coefficients.)

coef(mod3)

## (Intercept) achrdg12 genderMale   
## 19.9812 0.6355 3.5017

### 6.2.1 Marginal means

Before moving on to discuss what is missing from the regression model in Equation (6.1), it is important to note that the interpretation of the regression coefficient on gender has changed from what we discussed in Chapter 5. In Chapter 5, we noted that the regression coefficient on a dummy variables can be interpreted as the mean of the group indicated by the dummy (e.g. the mean of Math Achievement for Males, or for Females). However, when additional predictors are included in the regression model, this relationship no longer holds in general.

To see this we can compare the output of the t-test in Section 6.1 with the regression output shown above. In the t-test, the group means for Math Achievement were 55.47 for Females and 58.63 for males, so the mean difference was

However, the regression coefficient on gender in the multiple regression model above is equal to . Why the difference?

We actually saw the answer to this question in Chapter 4 when we discussed the interpretation of regression coefficients in multiple regression. In the multiple regression model, the regression coefficient on Gender controls for (“partials out”) the relationship between Reading Achievement and Gender. So, the multiple regression coefficient represents the relationship between Gender and Math, after controlling for Reading, whereas the group mean difference does not control for Reading. These two values will be the same only if the two predictors are uncorrelated.

In order to emphasize the distinction between “raw” group means computed from the data and the predicted group means obtained from a multiple regression model, the latter are referred to as *marginal means*, or sometimes as *adjusted means* or *least squares means*.

### 6.2.2 Summary

In a regression model with one continuous and one binary predictor (and no interaction):

* The model results in two regression lines, one for each value of the binary predictor. These are called the simple trends.
* In a standard multiple regression model, the simple trends are parallel but can have a different intercept; the difference in the intercepts is equal to regression coefficient of the binary variable.
* The difference between the simple trends is often called a “gap,” and the gap is also equal to the regression coefficient of the binary variable.
* It is important to note that the predicted group means for the binary variable are no longer equal to the “raw” group means computed directly from the data, because the predicted group means now control for the intercorrelation among the predictors. The predicted group means are called marginal means to emphasize this distinction.

## 6.3 Binary + continuous + interaction

In the previous section, the multiple regression model led us to conclude that the gender gap in Math Achievement is constant and equal to 3.50 (i.e., the regression coefficient on Gender). As we saw in Figure 6.2, this implies that the simple trends are parallel. However, these conclusions do not agree with what we saw in Section 6.1 when we fitted two simple regression model to the NELS data. As we discuss in this section, what is missing from the multiple regression model in Equation (6.1) is the interaction between Gender and Reading.

Mathematically, an interaction is just the product between two variables. Equation (6.2) shows how to include this product in our multiple regression model – we just take the product of the two predictors and add it into the model as a third predictor:

For the NELS example, this regression model is depicted in Figure 6.3. Note the the simple trends are no longer parallel and the regression lines agree exactly with what we had in Section 6.1. So, as promised, we have now arrived at a single multiple regression model that captures the relationships among Math Achievement, Reading Achievement, and Gender.

attach(NELS)  
# Interaction via hard coding  
genderXachrdg12 <- (as.numeric(gender) - 1) \* achrdg12  
mod4 <- lm(achmat12 ~ achrdg12 + gender + genderXachrdg12)  
  
a\_females <- coef(mod4)[1]  
b\_females <- coef(mod4)[2]  
  
a\_males <- a\_females + coef(mod4)[3]  
b\_males <- b\_females + coef(mod4)[4]  
  
# Plot reading and math for females  
plot(achrdg12[females], achmat12[females], xlab = "reading", ylab = "math")  
abline(a\_females, b\_females, lwd = 2)  
  
# Add points and line for males  
points(achrdg12[males], achmat12[males], col = "#4B9CD3", pch = 2)  
abline(a\_males, b\_males, col = "#4B9CD3", lwd = 2)  
  
# Add a legend  
legend(x = "topleft", legend = levels(gender), pch = c(1, 2), col = c(1, "#4B9CD3"))

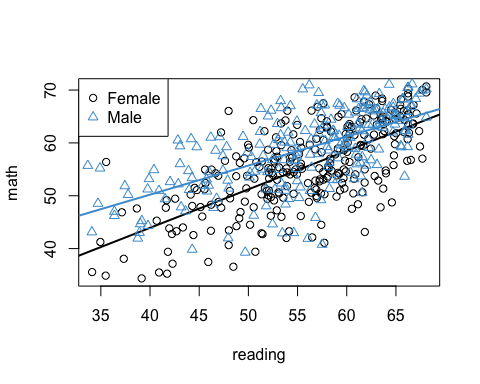


Figure 6.3: Math Achievement, Reading Achievement, and Gender (No Interaction).

detach(NELS)

To see what the model says about the data, let’s work through the model equations using our two-step procedure. As usual, we first plug in values for the categorical predictor, then we use the resulting equations solve for quantities of interest (the simple trends and the gender gap).

Similar to the results in Section 6.2, we can that see that

* The simple trends for Females and Males have different intercepts, because the regression coefficient for Gender () is added to the intercept for Males.

However, in contrast to Section 6.2,

* The simple trends for Females and Males no longer have same slope, because the regression coefficient for the interaction () is added to the slope for Males.
* The difference between the predicted values (i.e., the predicted gender gap in Math Achievement) is no longer constant, but is instead a function of . In particular, the gender gap in Math changes by units for each unit of change in Reading.

This last point is especially important in the context of our example. The gender gap in Math Achievement is a function of Reading Achievement. This is the mathematical meaning behind the concept of an interaction – the relationship between two variables (Math and Gender) is changing as a function of a third variable (Reading). This is what is means when we say, e.g., that the relationship between Gender and Math depends on Reading.

### 6.3.1 Choosing the moderator

The concept of an interaction is “symmetrical” in the sense that we can chose which variable goes in the “depends on” clause. For example, it is equally valid to say

* the relationship between Math and Gender depends on Reading, or
* the relationship between Math and Reading depends on Gender.

Whichever variable appears in the “depends on” clause is called the *moderator*, and the other two variables are called the focal variables. The researcher chooses which variable to treat as the moderator when interpreting an interaction. The overall idea here is to “break down” the interaction in the way that is most compatible with the research question(s).

For example, our research question was about the gender gap in STEM Achievement. So, our focal relationship is between Math and Gender, and Reading would be our moderator. Taking this approach, it would make sense to focuses our interpretation on difference – i.e., the predicted gender gap. So, we could say that the relationship between Math and Gender depends on Reading, or, more specifically, that the predicted gender gap in Math changes by units for each unit of increase in Reading.

By contrast, if we were more interested in the relationship between Math and Reading, then we could treat Gender as the moderator. Here it would make sense to focus our interpretation on the simple trends, and in particular on the slope parameters in these equations. For example, we might say:

* For females, predicted Math Achievement changed by units for each unit of increase in Reading,
* whereas for males, the predicted change was () units for each unit of increase in Reading.

This might feel less intuitive than talking about the gender gap, but the two interpretations are mathematicaly equivalent. Its just a matter of whether you want to interpret the regression coefficient with reference to the gender gap, or with reference to the simple trends.

### 6.3.2 Back to the example

The regression coefficients for the example data are shown below. **Please use these numbers to provide an interpretation of the interaction between Gender and Reading.** (Don’t worry about statistical significance, just focus on the meaning of the coefficients.)

coef(mod4)

## (Intercept) achrdg12 genderMale genderXachrdg12   
## 14.8031 0.7282 13.3933 -0.1779

Some potential answers are hidden below, but don’t peak until you have tried it for yourself!

# Gender gap  
# General: The gender gap in Math is smaller for students who are also strong in Reading  
# Specific: The gender gap in Math Achievement decreases by .18 percentage points for each percentage point increase Reading Achievement  
  
# Simple slopes  
# General: The relationship between Math and Reading is stronger (i.e. has a larger slope) for Females than for Males  
# Specific:  
# For Females, Math scores are predicted to increase by .72 percentage points for each percentage point increase in Reading Achievement  
# For Males, Math scores are predicted to increase by .55 percentage points for each percentage point increase in Reading Achievement

### 6.3.3 Centering the continuous predictor

You may have noticed that coefficient on gender was wildly different in regression models with and without the interaction. In the model without the interaction (Section 6.2) the coefficient on Gender was 3.50, and in the model with the interaction (above), it was 13.40. So in one model, the “effect” of being Male was a 3.5 percentage points gain on a Math exam, but in the other model, it was a 13.40 percentage point gain. Why this huge difference in the effect of Gender?

The answer can be seen in the equation for the gender gap. In the model **without the interaction**, the gender gap was constant and equal to the regression coefficient on Gender (denoted as in the model):

But in the regression model **with the interaction**, the gender gap was a linear function of Reading and the regression coefficient on Gender is actually the intercept for the linear relationship.

So, in the model with the interaction, is the gender gap for students who score zero on Reading Achievement. Since the lowest score on Reading Achievement was around 35, the intercept in this equation (i.e., the regression coefficient on gender, ) is not very meaningful.

One way to address this situation is to center the Reading variable so that it has a mean of zero. To do this, let

denote the deviation scores for (i.e., the mean-centered version of ). Then we just regress Math Achievement on rather than . Working through the equations shows that the gender gap is now

Since when , the regression coefficient on Gender () is now interpretable as the gender gap in Math Achievement, for students with average Reading Achievement. This is a much more interpretable number than the coefficient in the original interaction model!

In the example, this approach yields the following model parameters:

attach(NELS)  
# compute the deviation scores for reading  
reading\_dev <- achrdg12 - mean(achrdg12, na.rm = T)   
  
# Run the interaction model as above  
genderXreading\_dev <- (as.numeric(gender) - 1) \* reading\_dev  
  
mod5 <- lm(achmat12 ~ reading\_dev + gender + genderXreading\_dev)  
coef(mod5)

## (Intercept) reading\_dev genderMale genderXreading\_dev   
## 55.2944 0.7282 3.4993 -0.1779

detach(NELS)

The regression coefficient for Gender is now pretty close to what it was in the original model without the interaction, but the interpretation is different. Notice that the intercept in the multiple regression model with Reading centered has changed compared to the previous model in which Reading was not centered. However, the centering did not affect the regression coefficient for Reading or the interaction.

**Please write down your interpretation of the intercept and the regression coefficient for Gender in the above regression output, and be prepared to share your answer in class**.

### 6.3.4 Summary

The interaction between two variables is just their product. When this product is added as a predictor in a multiple regression model with one continuous and one binary predictor:

* The model again results in two regression lines, one for each value of the binary predictor, but we get regression lines with different intercepts *and different slopes*.
* The difference in slopes is equal to the regression coefficient on the interaction term.
* The difference (“gap”) between the regression lines changes as a linear function of the continuous predictor, and this change is again equal to the regression coefficient on the interaction term.
* These last two points are equivalent ways of stating the central idea behind a (two-way) interaction: the relationship between two variables changes as a function of a third variable.
* When interpreting an interaction, the research chooses which pair of variables will be the “focal relationship” and which variable will be the moderator.
* Centering the continuous predictor can be helpful for ensuring that the regression coefficient on the binary variable remains interpretable in the presense of an interaction.

## 6.4 Inference for interactions

Compared to the two simple regression models discussed in Section 6.1, the real power of the multiple regression model is that it facilitates statistical inference about the simple trends and “gaps.”

For example, in the standard summary output for the regression model, we now have a test of whether the interaction terms is statistically significant. You should feel comfortable interpreting the regression coefficients in this output based on the work we have done so far in this chapter, but now is a good time double check. The interpretation of the standard errors, tests of statistical significance, R-squared, etc, are all the same as in Chapter 4.

summary(mod5)

##   
## Call:  
## lm(formula = achmat12 ~ reading\_dev + gender + genderXreading\_dev)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.058 -3.786 0.501 4.077 16.289   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 55.2944 0.3513 157.41 < 2e-16 \*\*\*  
## reading\_dev 0.7282 0.0470 15.49 < 2e-16 \*\*\*  
## genderMale 3.4993 0.5214 6.71 5.3e-11 \*\*\*  
## genderXreading\_dev -0.1779 0.0651 -2.73 0.0065 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.8 on 496 degrees of freedom  
## Multiple R-squared: 0.462, Adjusted R-squared: 0.459   
## F-statistic: 142 on 3 and 496 DF, p-value: <2e-16

Note that the output above doesn’t answer all of the questions we might have about a interaction. For example, since we are interested in the gender gap in Math Achievement, we might want to know for which students the gap is statistically significant. In particular, does the gap disappear for students with higher levels of Reading Achievement? We can answer this type of question by testing marginal effects, which are a generalization of the marginal means discussed in section 6.2. In general, marginal effects are useful when the goal is to “follow up” a significant interaction in which the focal predictor is categorical, as opposed to continuous.

Note that marginal effects are only of interest when the interaction is significant – if the interaction is not significant, then we know that the relationship between the focal variables doesn’t depend on the moderator, so there is nothing left to say about that dependence.

### 6.4.1 Marginal effects

There are three main types of marginal effects. To explain these three approaches, first let’s write the gender gap in Math Achievement using a slightly more compact notation.

Then we having the following marginal effects that can be computed:

* Marginal effects at the mean (MEM): Report the gap at the mean value of
* Average marginal effect (AVE): Report the average of the marginal effect:
* Marginal effects at represenative values (MERV): Report the marginal effect for a range of “interesting” values

MEM and AVE are equivalent for linear regression models (but we will visit the distinction again when we get to logistic regression). We already have used this approach when we centered the continuous predictor in order to interpret the regression coefficient on Gender – this coefficient was the MEM / AVE.

In this section we focus on MERV, which is the more widely used approach. One usual choice for the “interesting values” is the quartiles of , which is reported below. Another popular choice is the mean of plus or minus 1 SD.

# Install the package if you haven't already done so  
# install.packages("emmeans")  
  
# Load the package into memory  
library(emmeans)  
  
# Fit the model using R's formula syntax for interaction '\*'  
mod6 <- lm(achmat12 ~ gender\*achrdg12, data = NELS)  
  
# Use the emmeans function to get the gender means on math, broken down by reading  
gap <- emmeans(mod6, specs = "gender", by = "achrdg12", cov.reduce = quantile)  
  
# Test whether the differences are significant  
contrast(gap, method = "pairwise")

## achrdg12 = 31.8:  
## contrast estimate SE df t.ratio p.value  
## Female - Male -7.74 1.637 496 -4.728 <.0001   
##   
## achrdg12 = 51.2:  
## contrast estimate SE df t.ratio p.value  
## Female - Male -4.27 0.593 496 -7.207 <.0001   
##   
## achrdg12 = 57.0:  
## contrast estimate SE df t.ratio p.value  
## Female - Male -3.25 0.529 496 -6.138 <.0001   
##   
## achrdg12 = 61.7:  
## contrast estimate SE df t.ratio p.value  
## Female - Male -2.41 0.658 496 -3.659 0.0003   
##   
## achrdg12 = 68.1:  
## contrast estimate SE df t.ratio p.value  
## Female - Male -1.28 0.967 496 -1.321 0.1872

The output shows the gender gap in Math Achievement for 5 values of Reading Achievement. The values of Reading Achievement are its 5 quartiles. **Please examine the statistical significance of the gender gap in Math Achievement at the 5 quartiles of Reading Achievement and make a conclusion whether the gap “dissapeared” for students with higher levels of Reading Achievement. Please be prepared to share your answer in class!**

Note that the computations going on “under the hood” for testing marginal means are pretty complicated. You can read more details here: <https://cran.r-project.org/web/packages/emmeans/vignettes/basics.html>

### 6.4.2 Simple trends

Another way to follow-up a significant interaction is to examine the statistical significance of the simple trends. As discussed in Section 6.3, the simple trends aren’t very meaningful in the context of our example, so this section is just about illustrating the technique, not about adding anything to our discussion of the gender gap in STEM. In general, testing simple trends can be useful when following-up a significant interaction in which the focal predictor is continuous, rather than categorical.

We can understand simple trends in reference to what they add to the standard summary output for the lm function (reported at the beginning of Section 6.4):

* The regression coefficient on the continuous predictors tells us about simple trend for the group designated as zero on the binary predictor (e.g, simply for females).
* The regression coefficient on the interaction term tells whether the simple trends differ for the two groups (e.g., whether the simple trend for males differs from the simple trend for females).

Note that what is missing, or implicit, in this output is a test of the simple trend for the group designated as one on the binary predictor. Although we can compute the simple trend for this group from the output, the output does not provide a statistical test of whether this trend is different from zero. So, in our example, the summary output doesn’t tell us whether the simple trend for males is different from zero.

The test of the simple trends for the example are reported below. Again, these aren’t super interesting in the context of our example, but you should **check your understanding of simple trends by writing down an interpretation of the output below**.

# Use the emtrends function to get the regression coefficients on reading, broken down by gender  
simple\_slopes <- emtrends(mod6, var = "achrdg12", specs = "gender")  
test(simple\_slopes)

## gender achrdg12.trend SE df t.ratio p.value  
## Female 0.728 0.0470 496 15.487 <.0001   
## Male 0.550 0.0451 496 12.208 <.0001

### 6.4.3 A note on plotting

Another nice advantage of having everything in one model is that we can level-up our plotting. Check out this plot from the visreg package (and its only line of code!). You should be able to draw a similar conclusion from the plot as you did from looking at the MERVs.

# Install the package if you haven't already done so  
# install.packages("visreg")  
# Load the package into memory  
library(visreg)  
  
mod6 <- lm(achmat12 ~ gender\*achrdg12, data= NELS)  
visreg(mod6, xvar = "achrdg12", by = "gender", overlay = TRUE)

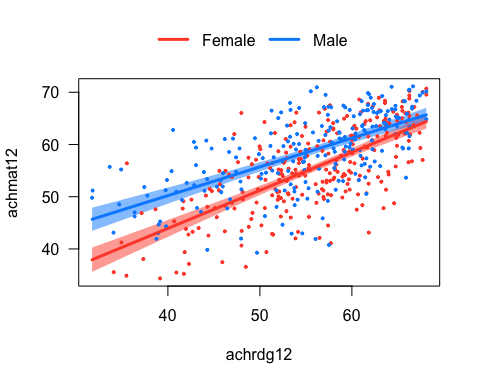


Figure 6.4: Example of a plot using the visreg package.

### 6.4.4 Summary

When making inferences about an interaction:

* Often we can get all of the information we need from the regression coefficient on the interaction term, and its associated test of significance. If the interaction isn’t significant, we stop there. But if the interaction is significant, we may want to report more information about how the focal relationship depends on the moderator.
* When the focal predictor is categorical it can be interesting to “follow up” a significant interaction by taking a closer look at the statistical significance of the marginal means (e.g, how the gender gap in Math changes as a function of Reading)
* When the focal predictor is continuous, it can be interesting to “follow up” a significant interaction by taking a closer look at the statistical significance of the simple trends / simple slopes.

## 6.5 Two continuous predictors

In this section we address the situation in there is an interaction between two continuous predictors. The regression equation and overall interpretation is the same as the previous sections – e.g., the relationship between Y and X1 changes as a function of X2.

However, there are also some special details that crop up when considering an interaction between two continuous predictors. In this section we will address:

* The importance of centering the two predictors. When there are two continuous predictors, centering helps interpret the coefficients on the predictors (just like in Section 6.3), and can additionally be helpful for reducing the correlation between the predictors and the interaction.
* How to follow-up a significant interaction using simple trends. Because the predictors are continuous, the focus is on simple trends rather than marginal means.

First, we introduce an new example.

### 6.5.1 Another NELS example

To illustrate an interaction between two continuous predictors, let’s replace Gender with SES in our previous analysis. Apologies that this new example is mainly for convenience and doesn’t represent a great research question about, e.g., why the relationships between Math and Reading might change as a function of SES.

The example data and overall approach with SES as the moderator are illustrated below (note that the values 9, 19, and 28 are 10th, 50th, and 90th percentiles SES, respectively).

#Interaction without centering   
mod7 <- lm(achmat12 ~ achrdg12\*ses, data = NELS)  
  
# Note that band = F removes the confidence intervals  
visreg(mod7, xvar = "achrdg12", by = "ses", overlay = TRUE, band = F)

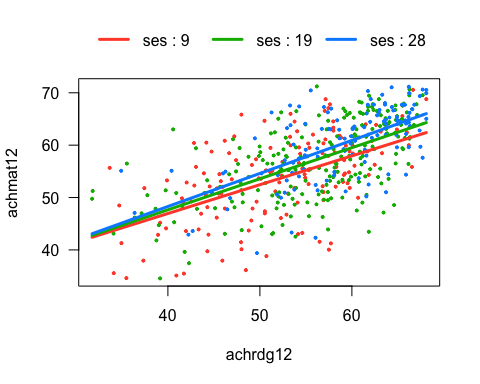


Figure 6.5: Math (achmat), Reading (achrdg), and SES

### 6.5.2 Centering the predictors

Centering the predictors facilitates the interpretation of their regression coefficients in the presences of an interaction, just as it did Section 6.3. In particular, the coefficients and in the regression model

can be interpreted in terms of the following simple trends:

$$

For example, is the relationship between and , when is equal to zero.

In general, setting a variable to the value of zero may not be meaningful. But, when setting a *centered* variable (i.e., a deviation score) to zero, this is equivalent to setting the original variable to its mean. So, if the variables in Equation (6.3) were centered, we could say that is the relationship between and , when is equal to its mean. Again, this is just the same trick as Section 6.3, but this time both predictors are continuous and both are centered.

There is another reason for centering continuous predictors when there is an interaction in the model, and this has to do with reducing the correlation among the predictors and their interaction. In general, the interaction term will be positively correlated with both predictors if (a) the predictors themselves are positively correlated and (b) the predictors take on strictly positive (or strictly negative) values. Highly correlated predictors lead to redundant information the model, so we generally want to avoid this situation (this is technically called *multicollinearity* and we discuss it in more detail in a later Chapter). Centering can “break” the correlation between the preditors and the interaction, thereby making the predictors less redundant with their interaction.

To see how this works, let’s take a look at Figure 6.6. The left hand panel shows that SES and its interaction with reading are highly correlated. This is because (a) SES and Reading are themselves positively correlated, and (b) both SES and Reading take on strictly positive values. As mentioned, the interaction term will be positively correlated with both predictors whenever these two conditions hold. (The figure just shows the correlation with SES, but the same situation holds for Reading.)

attach(NELS)  
# Correlation without centering  
r <- cor(ses, achrdg12\*ses)  
  
# Plot  
par(mfrow = c(1, 2))  
title <- paste0("correlation = ", round(r, 3))  
plot(ses, achrdg12\*ses, col = "#4B9CD3", main = title, xlab = "SES", ylab = "SES X Reading")  
  
achrdg12\_dev <- achrdg12 - mean(achrdg12)  
ses\_dev <- ses - mean(ses)  
r <- cor(ses\_dev, achrdg12\_dev\*ses\_dev)  
  
# Plot  
title <- paste0("correlation = ", round(r, 3))  
plot(ses\_dev, achrdg12\_dev\*ses\_dev, col = "#4B9CD3", main = title, xlab = "SES Centered", ylab = "SES Centered X Reading Centered")

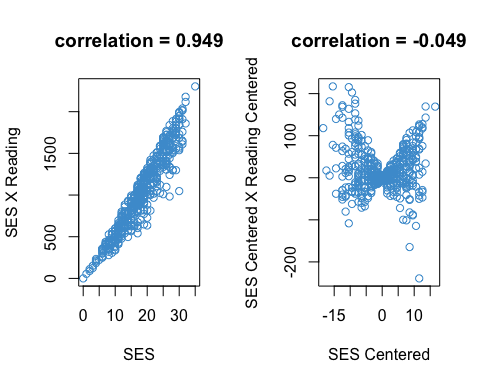


Figure 6.6: Correlation Between SES and SES X Reading, With and Without Centering

detach(NELS)

We can see in the right hand panel of Figure 6.6 how centering the two predictors “breaks” the linear relationship between SES and its interaction with Reading. After centering, the relationship between the SES and its interaction is now highly non-linear, and the correlation is approximately zero. Again, the same is true for the relationship between Reading and the interaction, but the figure only shows the situation for SES. The upshot of all this is that the predictors will be less redundant with the interaction term – i.e., centering reduces multicollinearity between the “main effects” of the predictors and their interaction.

The two sets of output below show the regression of Math on Reading, SES, and their interaction. The first ouptut does not center the predictors, but the second output does (the \_dev notation denotes the centered predictors).

We can see that SES is a significant predictor in the centered model but not in the “un-centered” model. This has to do with changing the interpretation of the coefficient (it now represents the relationship between Math and SES for students with average Reading), and also reflects the fact that SES is no longer so highly correlated with the interaction term after centering.

attach(NELS)  
# Without centering  
mod7 <- lm(achmat12 ~ achrdg12\*ses)  
summary(mod7)

##   
## Call:  
## lm(formula = achmat12 ~ achrdg12 \* ses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.134 -3.894 0.728 4.130 15.015   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 25.84920 5.38980 4.80 2.1e-06 \*\*\*  
## achrdg12 0.51160 0.09943 5.15 3.9e-07 \*\*\*  
## ses -0.10011 0.29121 -0.34 0.73   
## achrdg12:ses 0.00427 0.00520 0.82 0.41   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.03 on 496 degrees of freedom  
## Multiple R-squared: 0.418, Adjusted R-squared: 0.415   
## F-statistic: 119 on 3 and 496 DF, p-value: <2e-16

# With centering  
mod8 <- lm(achmat12 ~ achrdg12\_dev\*ses\_dev)  
summary(mod8)

##   
## Call:  
## lm(formula = achmat12 ~ achrdg12\_dev \* ses\_dev)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.134 -3.894 0.728 4.130 15.015   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 56.82622 0.28691 198.07 <2e-16 \*\*\*  
## achrdg12\_dev 0.59031 0.03610 16.35 <2e-16 \*\*\*  
## ses\_dev 0.13730 0.04149 3.31 0.001 \*\*   
## achrdg12\_dev:ses\_dev 0.00427 0.00520 0.82 0.412   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.03 on 496 degrees of freedom  
## Multiple R-squared: 0.418, Adjusted R-squared: 0.415   
## F-statistic: 119 on 3 and 496 DF, p-value: <2e-16

detach(NELS)

**To check your understanding of the output above, please provide an interpretation of all four regression coefficients in the centered model. Your interpretations should make reference to the situation where one or both predictors are equal to zero (see Equation (6.3) above) and should also mentioned the interpretation of the value of zero for the centered variables**

Note that coefficient on interaction, the R-squared, and their associated tests do not change value. This is discussed in more detail in the extra material at the end of this section, but it is sufficient to note that centering only affects the interpretation of the main effects (and the intercept, of course).

### 6.5.3 Simple trends

Centering helps us interpret the “main effects” of the individual predictors, but we haven’t yet discussed how to interpret the interaction term when both predictors are continuous. In the case with a binary predictor, we had two different regression lines. So what do we get when both predictors are continuous?

The usual way to answer this question is an extension of the MERV approach to marginal effects, discussed in Section 6.4. The basic idea is to present the relationship between the two focal variables, for a selection of values of the moderator. The is shown in Figure ?? above. As with MERV, the choice of values of the moderator is up to the researcher, but usual choices are

* The quartiles of the moderator (i.e., the five number summary) or subset thereof
* M 1 SD of the moderator
* A selection of percentiles (visreg uses the 10th, 50th, and 90th)

These are all doing very similar things, so the choosing among them isn’t very important.

Although the interaction between Reading and SES was not significant in our example model, let’s break down the interaction with SES as the moderator just to see how this approach works. The first part output below shows the simple slopes for the three values of SES shown in Figure 6.5 (i.e., the 10th, 50th, and 90th deciles). We can see that the simple slopes are all different from zero (and the non-significant interaction tells us that they are not different from one-another).

# Break down interaction with SES as moderator  
simple\_slopes <-emtrends(mod7, var = "achrdg12", specs = "ses", at = list(ses = c(9, 19, 28)))  
test(simple\_slopes)

## ses achrdg12.trend SE df t.ratio p.value  
## 9 0.550 0.0583 496 9.429 <.0001   
## 19 0.593 0.0365 496 16.251 <.0001   
## 28 0.631 0.0639 496 9.878 <.0001

### 6.5.4 Summary

When regressing an outcome on two continuous predictors and their interaction, the overall interpretation of the model is same as discussed in Section 6.3, but:

* It is useful to center both predictors, both to facilitate the interpretation of the regression coefficients on the predictors (the “main effects”), and to reduce the correlation between the predictors and their interaction (i.e., reduce multicollinearity).
* When following up a significant interaction, the usual approach is to report the simple trends between the focal variables for a selection of values of the moderator (e.g., a selection of percentiles). The example illustrated how to do this even though the interaction was not significant, but you shouldn’t follow up a non-significant interaction.

### 6.5.5 Extra: How centering works\*

It might seem that centering both predictors to improve the chances of getting significant main effects is a dubious practice. However, using the centered or the un-centered variables doesn’t really make a difference in term of what predictors are in the model are. The follow algebra show why, using for the centered variables:

$$

The second line of the equation shows that we are not changing what we regress on – i.e., the predictors are still and . We are re-packaging the intercept and main effects, which is exactly the purpose of this approach. But, centering does not change the regression coefficient for the interaction – it is still interpreted with respect to the un-centered variables.

## 6.6 Two categorical predictors

This section addresses interactions between two categorical predictors. Up until now, we have looked at interactions only for categorical predictors that are dichotomous. In this section we address an example in which one of the categorical predictors has more than two levels. This requires combining what we learned about contrast coding (Chapter 5) with what we have learned about interactions. One nice aspect of interactions among categorical predictors is that we usually don’t need to use procedures like marginal effects to follow up significant interactions, so long as we make good use of contrast coding.

In experimental (as opposed to observational) settings, interactions among categorical predictors fall under the much larger topic of ANOVA and experimental design. The analysis we look at in this section is a two-way between-subjects ANOVA, meaning that there are two categorical predictors considered, as well as their interaction, and both predictors are cross-sectional. ANOVA is a big topic and is not the focus of this course. However, we will discuss how to summarize the results of our analysis in an ANOVA table, and consider how this differs from the standard regression approach.

### 6.6.1 An example from ECLS

For this topic we will switch over to the ECLS data and examine how SES and Pre-K attendance interact to predict Math Achievement at the beginning of Kindergarten. The variables we will examine are

* Math Achievement at the beginning of K (c1rmscal). This is the number of correct questions on a test with approximately 70 items.
* Whether the child attended Pre-K (p1center). This is a binary variable that indicates pre-K attendance.
* SES, coded as quintiles (wksesq5). We will denote this variable as SES, but keep in mind it is quintiles in this example (e.g., SES = 1 are the respondents with SES between the minimum and the first quintile).

Coding SES as quintiles allows us to consider it as a categorical predictor with 5 levels. This is convenient for our illustration of interactions between categorical predictors. It is also a widely-used practice in policy research, because SES often has non-linear relationships with outcome variables of interest, and these relationships can be more easily captured by treating SES as a categorical variable.

In this analysis, our focus will be whether the “effect” of Pre-K on Math Achievement depends on (i.e., is moderated by) the child’s SES. The research question is relevant to the discussion we previously had concerning the paper shared by Andrea (<http://dx.doi.org/10.1037/dev0001301>). Also note that I will use the term “effect” in this section to simplify language, but we know that Pre-K attendance was not randomly assigned in ECLS, so please keep in mind that this terminology is not strictly correct.

The relationship among the three variables is summarized in the visreg plot below. We can see that the effect of Pre-K on Math Achievement appears to differ as a function of SES – i.e., there is an interaction between Pre-K and SES. Our goal in this section is to produce an analysis corresponding to the Figure. As in Section 6.2, we will start with a model that does not include the interaction as a first step. **Before moving on, please take a moment to write down how Figure 6.7 would be different if there was no interaction between Pre-K and SES**.

load("ECLS2577.Rdata")  
ecls$prek <- factor(2 - ecls$p1center)  
ecls$wksesq5 <- factor(ecls$wksesq5)  
mod <- lm(c1rmscal ~ prek\*wksesq5, data = ecls)  
visreg::visreg(mod, xvar = "wksesq5", by = "prek",   
 partial = F, rug = F, overlay = T,   
 strip.names = T, xlab = "SES",   
 ylab = "Math Achievement in K")

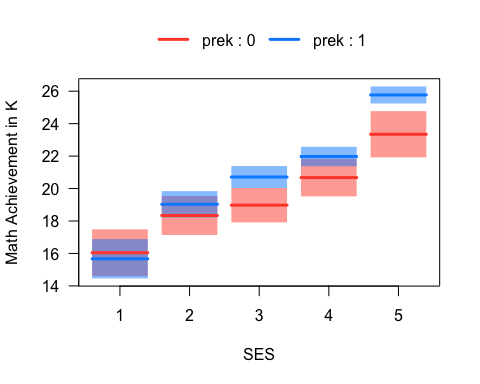


Figure 6.7: Math Achievement, Pre-K Attendence, and SES

### 6.6.2 The “no-interaction” model

In order to represent a model with multiple categorical predictors, it is helpful to change our notation from the usual and to something like the following:

In in this notation, the predictor variables are indicators (binary dummies) that use the variable names rather than , , etc. The variable is just the indicator for Pre-K attendance, as defined above. The variable is an indicator for the j-th quintile of SES. Note that both variables use reference-group coding, as discussed in Chapter 5.

We can interpret the coefficients in this model in using the same two-step procedure described in Chapter 5. Since there are many terms in the model, things are going to start getting messy quickly, so brace yourself for some long equations (but simple math!).

The main points about the interpretation of this model are as follows.

* The intercept is the predicted value of Math Achievement for students in the first SES quintile who did not attend Pre-K. This corresponds to the blue line in the first column of Figure 6.7.
* The effect of Pre-K attendance for students in the first SES quintile is equal to . This corresponds to the difference between the red and blue lines in the first column of Figure 6.7.
* Because the model in Equation (6.4) does not include an interaction, we know that it implies the effect of pre-K is constant over levels of SES. Below we consider SES = 2, but the same approach works for the other levels of SES.

This equation says that the difference between the red and blue lines in the second column of Figure 6.7 is the same as the difference in the first column – i.e., they both equal . Similar calculations show that Equation (6.4) implies that the effect of Pre-K is constant over all levels of SES. Hence, we need to change the model (i.e., add an interaction)

Before moving on, note that similar calculations can be used to show that (a) the “effect” of being in any SES quintile, as compared to the first quintile, is just the regression coefficient on the indicator variable for that quintile, and (b) this effect is constant over levels of Pre-K attendance.

### 6.6.3 Adding the interaction(s)

We have just seen that Equation (6.4) implies that the effect of Pre-K is constant over levels SES, and vise versa. In order to address our research question about whether the relationship between Pre-K attendance and Math Achievement depends on children’s SES, we will need to add something to the model – an interaction (surprise!).

We know that interactions are just products (multiplication) of predictor variables. But, since SES is represented as 4 dummies, this means we need 4 products in order to represent the interaction of Pre-K with SES. The resulting model can be written:

As you can see, our model is getting pretty full! Even though we are only considering two distinct “conceptual” variables, we have 9 coefficients in our regression model.

Again, there are a few main things to notice:

* The interpretation of the intercept has not changed. It still corresponds to the blue line in the first column of Figure 6.7.
* The regression coefficient on is still the “effect” of Pre-K for students in the first SES quintile. It corresponds to the difference betwee the red and blue line in the first column of Figure 6.7. This is because all the variables are equal to zero for students in the first SES quintile, and so all of the interaction terms in Equation (6.5) are equal to zero for this case.
* The effect of Pre-K is no longer constant over levels of SES. Again we will focus on SES = 2, but the same approach works for the other levels of SES.

The last line shows that the “effect” of Pre-K for students in the second SES quintile is . This is not the same as for the effect for students in the first quintile, which was just . In other words, the difference between the red and blue lines in the first column of Figure 6.7 (i.e., ) is not equal to the difference in the second column (i.e., ). Consequently, our new model with the interaction better reflects the example data.

The same approach shows that the effect of Pre-K at each level of SES results in a similar equation:

This pattern makes it clear that, to isolate effect of each interaction (i.e., through ), we need to subtract off – i.e., we need to subtract off the effect of Pre-K for students in the first SES quintile. In this sense, the interpretation of is quite similar the interpretation of the intercept in regular reference-group coding (see Section 5.7.4). It is the “reference effect” or baseline to which the interaction terms are compared.

For example

* The interaction between Pre-K and the second SES quintile is the additional effect pre-K has on Math Achievement for students in the second SES quintile, *as compared to the effect in the first SES quintile.*
* The interaction between Pre-K and the third SES quintile is the additional effect pre-K has on Math Achievement for students in the 3rd SES quintile, *as compared to the effect in the first SES quintile.*
* etc etc.

Mathematically, the interaction terms are represented as “differences-in-differences.” For example,

$$

This looks quite complicated but it is just an extension of reference-group coding. This equation is saying that the “reference effect” or “baseline” for interpreting the interaction () is the effect of Pre-K in the first SES quintile (i.e., ). As noted above, all of the interaction terms have the same reference effect.

### 6.6.4 Back to the example

That last section was a lot to take in, so let’s put some numbers on the page to check our understanding. The output below shows the summary for a model that regresses Math Achievement on Pre-K, SES, and their interaction. **Please write down an interpretation of magnitude, direction, and statistical significance of each regression coefficient in this output (including the intercept), and be prepared to share your answers in class.** Remember that wksesq5 is the variable code for the SES quintiles – the digit that follows the variable code indicates the level of variable. It may be helpful to refer to Figure 6.7 in your interpretations.

mod <- lm(c1rmscal ~ prek\*wksesq5, data = ecls)  
summary(mod)

##   
## Call:  
## lm(formula = c1rmscal ~ prek \* wksesq5, data = ecls)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.768 -4.768 -0.975 3.955 31.232   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.045 0.735 21.82 < 2e-16 \*\*\*  
## prek1 -0.373 0.960 -0.39 0.6973   
## wksesq52 2.293 0.957 2.40 0.0166 \*   
## wksesq53 2.930 0.913 3.21 0.0013 \*\*   
## wksesq54 4.631 0.944 4.91 9.9e-07 \*\*\*  
## wksesq55 7.299 1.034 7.06 2.2e-12 \*\*\*  
## prek1:wksesq52 1.064 1.212 0.88 0.3801   
## prek1:wksesq53 2.109 1.154 1.83 0.0679 .   
## prek1:wksesq54 1.671 1.168 1.43 0.1527   
## prek1:wksesq55 2.797 1.234 2.27 0.0235 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.9 on 2567 degrees of freedom  
## Multiple R-squared: 0.162, Adjusted R-squared: 0.159   
## F-statistic: 55.2 on 9 and 2567 DF, p-value: <2e-16

### 6.6.5 The ANOVA approach

The output in the previous section is detailed enough that it is not usually required to follow-up a significant interactions among categorical predictors using marginal effects. However, the summary output omits some information we might be interested in. For example, what is the overall effect of Pre-K (i.e., is there a significant difference in Math Achievement for students who attended Pre-K or not, regardless of their level of SES)? Similarly, what is the overall effect of SES, regardless of Pre-K attendance. We can answer these questions by asking how much variance was explained by each predictor, rather than only examining the effect of the individual dummy variable. This is the ANOVA approach we discussed last semester, but now applied to two categorical predictors. `

The ANOVA table for our example is below. Note that it leads to slightly different conclusions than the regression output above. We will discuss the discrepancies between the ANOVA and regression output in class, so please **write down your interpretation of the ANOVA table below and be prepared to share you thoughts in class.**

anova(mod)

## Analysis of Variance Table  
##   
## Response: c1rmscal  
## Df Sum Sq Mean Sq F value Pr(>F)   
## prek 1 3434 3434 72.14 <2e-16 \*\*\*  
## wksesq5 4 19914 4978 104.58 <2e-16 \*\*\*  
## prek:wksesq5 4 299 75 1.57 0.18   
## Residuals 2567 122198 48   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## 6.7 Workbook

This section collects the questions asked in this chapter. We will discuss these questions in class. If you haven’t written down / thought about the answers to these questions before class, the lesson will not be very useful for you! So, please engage with each question by writing down one or more answers, asking clarifying questions, posing follow up questions, etc.

**Section 6.1**

attach(NELS)  
mod1 <- lm(achmat12[females] ~ achrdg12[females])  
mod2 <- lm(achmat12[males] ~ achrdg12[males])  
  
# Plot reading and math for females  
plot(achrdg12[females], achmat12[females], xlab = "Reading", ylab = "Math")  
abline(mod1, lwd = 2)  
  
# Add points and line for males  
points(achrdg12[males], achmat12[males], col = "#4B9CD3", pch = 2)  
abline(mod2, col = "#4B9CD3", lwd = 2)  
  
# Add a legend  
legend(x = "topleft", legend = levels(gender), pch = c(1, 2), col = c(1, "#4B9CD3"))

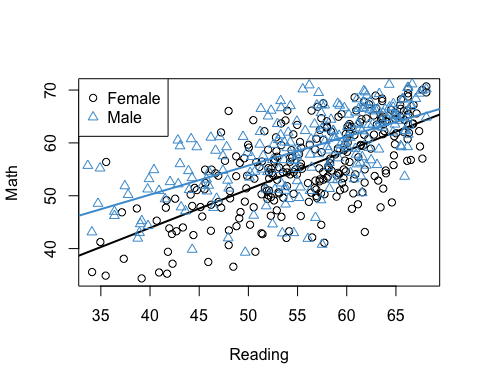


Figure 6.8: Math Achievement, Reading Achievement, and Gender.

detach(NELS)

Take a minute to think about what this plot is telling us about the relationships among Math Achievement, Reading Achievement, and Gender.

* Is the gender gap in math constant?
* Is the relationship between math and reading the same for males and females?
* Can you provide a summary of the plot in terms of what it says about the gender gap in Math Achievement?

**Section 6.2**

* No interaction model: The regression coefficients for the example data are shown below. Please use these numbers to provide an interpretation of the simple trends and the gender gap in Math Achievement for the NELS example. (Don’t worry about statistical significance, just focus on the meaning of the coefficients.)

mod3 <- lm(achmat12 ~ achrdg12 + gender, data = NELS)  
coef(mod3)

## (Intercept) achrdg12 genderMale   
## 19.9812 0.6355 3.5017

**Section 6.3**

* Interaction model: The regression coefficients for the example data are shown below. Please use these numbers to provide an interpretation of the interaction between Gender and Reading. (Don’t worry about statistical significance, just focus on the meaning of the coefficients.)

attach(NELS)  
genderXachrdg12 <- (as.numeric(gender) - 1) \* achrdg12  
mod4 <- lm(achmat12 ~ achrdg12 + gender + genderXachrdg12)  
coef(mod4)

## (Intercept) achrdg12 genderMale genderXachrdg12   
## 14.8031 0.7282 13.3933 -0.1779

detach(NELS)

* Interaction model with centered continuous predictor: Please write down your interpretation of the intercept and the regression coefficient for Gender in the regression output below.

attach(NELS)  
# compute the deviation scores for reading  
reading\_dev <- achrdg12 - mean(achrdg12, na.rm = T)   
  
# Run the interaction model as above  
genderXreading\_dev <- (as.numeric(gender) - 1) \* reading\_dev  
  
mod5 <- lm(achmat12 ~ reading\_dev + gender + genderXreading\_dev)  
coef(mod5)

## (Intercept) reading\_dev genderMale genderXreading\_dev   
## 55.2944 0.7282 3.4993 -0.1779

detach(NELS)

**Section 6.4**

* The output shows the gender gap in Math Achievement for 5 values of Reading Achievement. The values of Reading Achievement are the 5 quartiles. Please examine the statistical significance of the gender gap in Math Achievement at the 5 quartiles of Reading Achievement and make a conclusion whether the gap “dissapeared” for students with higher levels of Reading Achievement.

# Install the package if you haven't already done so  
# install.packages("emmeans")  
  
# Load the package into memory  
library(emmeans)  
  
# Fit the model using R's formula syntax for interaction '\*'  
mod6 <- lm(achmat12 ~ gender\*achrdg12, data= NELS)  
  
# Use the emmeans function to get the gender means on math, broken down by reading  
gap <- emmeans(mod6, specs = "gender", by = "achrdg12", cov.reduce = quantile)  
  
# Test whether the differences are significant  
contrast(gap, method = "pairwise")

## achrdg12 = 31.8:  
## contrast estimate SE df t.ratio p.value  
## Female - Male -7.74 1.637 496 -4.728 <.0001   
##   
## achrdg12 = 51.2:  
## contrast estimate SE df t.ratio p.value  
## Female - Male -4.27 0.593 496 -7.207 <.0001   
##   
## achrdg12 = 57.0:  
## contrast estimate SE df t.ratio p.value  
## Female - Male -3.25 0.529 496 -6.138 <.0001   
##   
## achrdg12 = 61.7:  
## contrast estimate SE df t.ratio p.value  
## Female - Male -2.41 0.658 496 -3.659 0.0003   
##   
## achrdg12 = 68.1:  
## contrast estimate SE df t.ratio p.value  
## Female - Male -1.28 0.967 496 -1.321 0.1872

**Section 6.5**

* To check your understanding of centering with two continuous predictors, please provide an interpretation of all four regression coefficients in the centered model (below). Your interpretations should make reference to the situation where one or both predictors are equal to zero (see Equation (6.3)) and should also mentioned the interpretation of the value of zero for the centered variables.

attach(NELS)  
mod8 <- lm(achmat12 ~ achrdg12\_dev\*ses\_dev)  
summary(mod8)

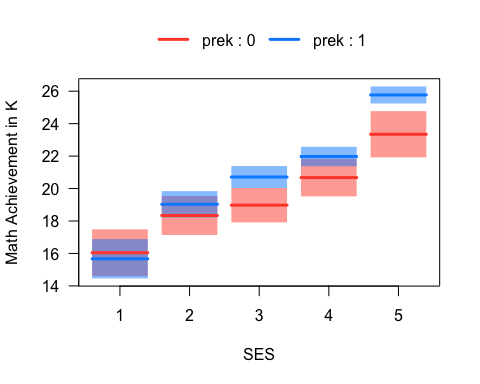
##   
## Call:  
## lm(formula = achmat12 ~ achrdg12\_dev \* ses\_dev)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.134 -3.894 0.728 4.130 15.015   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 56.82622 0.28691 198.07 <2e-16 \*\*\*  
## achrdg12\_dev 0.59031 0.03610 16.35 <2e-16 \*\*\*  
## ses\_dev 0.13730 0.04149 3.31 0.001 \*\*   
## achrdg12\_dev:ses\_dev 0.00427 0.00520 0.82 0.412   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.03 on 496 degrees of freedom  
## Multiple R-squared: 0.418, Adjusted R-squared: 0.415   
## F-statistic: 119 on 3 and 496 DF, p-value: <2e-16

detach(NELS)

**Section 6.6**

* Please take a moment to write down how the Figure below would be different if there was no interaction between Pre-K and SES.

load("ECLS2577.Rdata")  
ecls$prek <- factor(2 - ecls$p1center)  
ecls$wksesq5 <- factor(ecls$wksesq5)  
mod <- lm(c1rmscal ~ prek\*wksesq5, data = ecls)  
visreg::visreg(mod, xvar = "wksesq5", by = "prek",   
 partial = F, rug = F, overlay = T,   
 strip.names = T, xlab = "SES",   
 ylab = "Math Achievement in K")



* Please write down an interpretation of magnitude, direction, and statistical significance of each regression coefficient in this output (including the intercept), and be prepared to share your answers in class. Remember that wksesq5 is the variable code for the SES quintiles – the digit that follows the variable code indicates the level of variable.

summary(mod)

##   
## Call:  
## lm(formula = c1rmscal ~ prek \* wksesq5, data = ecls)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.768 -4.768 -0.975 3.955 31.232   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.045 0.735 21.82 < 2e-16 \*\*\*  
## prek1 -0.373 0.960 -0.39 0.6973   
## wksesq52 2.293 0.957 2.40 0.0166 \*   
## wksesq53 2.930 0.913 3.21 0.0013 \*\*   
## wksesq54 4.631 0.944 4.91 9.9e-07 \*\*\*  
## wksesq55 7.299 1.034 7.06 2.2e-12 \*\*\*  
## prek1:wksesq52 1.064 1.212 0.88 0.3801   
## prek1:wksesq53 2.109 1.154 1.83 0.0679 .   
## prek1:wksesq54 1.671 1.168 1.43 0.1527   
## prek1:wksesq55 2.797 1.234 2.27 0.0235 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.9 on 2567 degrees of freedom  
## Multiple R-squared: 0.162, Adjusted R-squared: 0.159   
## F-statistic: 55.2 on 9 and 2567 DF, p-value: <2e-16

* Write down your interpretation of the ANOVA table below and be prepared to share you thoughts in class.

anova(mod)

## Analysis of Variance Table  
##   
## Response: c1rmscal  
## Df Sum Sq Mean Sq F value Pr(>F)   
## prek 1 3434 3434 72.14 <2e-16 \*\*\*  
## wksesq5 4 19914 4978 104.58 <2e-16 \*\*\*  
## prek:wksesq5 4 299 75 1.57 0.18   
## Residuals 2567 122198 48   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## 6.8 Exercises

These exercises provide an overview of how to add interactions using the lm function, how to center continuous predictors, and how to follow-up significant interactions with the emmeans package.

### 6.8.1 Binary + continuous + interaction

There are multiple ways of implementing interactions in R.

* We can “hard code” new variables into our data (e.g., the product of a binary gender variable and reading)
* We can use R’s formula notation for single term interactions (:)
* We can use R’s formula notation for factorial interactions (\*)

The following three models illustrate how to use these three approaches, and show that they all producing the same output. In general, the \* syntax is the easiest to use, so we will stick with that one going forward. The variables used in the example are from the NELS data:

* achmat12 is Mat Achievement (percent correct on a mat test) in grade 12.
* achrdg12 is Reading Achievement (percent correct on a reading test) in grade 12.
* gender is dichotomous encoding of gender with values Male and Female (it is not a binary variable, but a factor, as discussed in Section ??).

load("NELS.RData")  
attach(NELS)  
# Interaction via hard coding  
genderXreading <- (as.numeric(gender) - 1) \* achrdg12  
mod1 <- lm(achmat12 ~ achrdg12 + gender + genderXreading)  
summary(mod1)

##   
## Call:  
## lm(formula = achmat12 ~ achrdg12 + gender + genderXreading)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.058 -3.786 0.501 4.077 16.289   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 14.8031 2.6492 5.59 3.8e-08 \*\*\*  
## achrdg12 0.7282 0.0470 15.49 < 2e-16 \*\*\*  
## genderMale 13.3933 3.6583 3.66 0.00028 \*\*\*  
## genderXreading -0.1779 0.0651 -2.73 0.00652 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.8 on 496 degrees of freedom  
## Multiple R-squared: 0.462, Adjusted R-squared: 0.459   
## F-statistic: 142 on 3 and 496 DF, p-value: <2e-16

# Interaction via `:` operator  
mod2 <- lm(achmat12 ~ achrdg12 + gender + achrdg12:gender)  
summary(mod2)

##   
## Call:  
## lm(formula = achmat12 ~ achrdg12 + gender + achrdg12:gender)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.058 -3.786 0.501 4.077 16.289   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 14.8031 2.6492 5.59 3.8e-08 \*\*\*  
## achrdg12 0.7282 0.0470 15.49 < 2e-16 \*\*\*  
## genderMale 13.3933 3.6583 3.66 0.00028 \*\*\*  
## achrdg12:genderMale -0.1779 0.0651 -2.73 0.00652 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.8 on 496 degrees of freedom  
## Multiple R-squared: 0.462, Adjusted R-squared: 0.459   
## F-statistic: 142 on 3 and 496 DF, p-value: <2e-16

# Interaction via `\*` operator  
mod3 <- lm(achmat12 ~ achrdg12\*gender)  
summary(mod3)

##   
## Call:  
## lm(formula = achmat12 ~ achrdg12 \* gender)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.058 -3.786 0.501 4.077 16.289   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 14.8031 2.6492 5.59 3.8e-08 \*\*\*  
## achrdg12 0.7282 0.0470 15.49 < 2e-16 \*\*\*  
## genderMale 13.3933 3.6583 3.66 0.00028 \*\*\*  
## achrdg12:genderMale -0.1779 0.0651 -2.73 0.00652 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.8 on 496 degrees of freedom  
## Multiple R-squared: 0.462, Adjusted R-squared: 0.459   
## F-statistic: 142 on 3 and 496 DF, p-value: <2e-16

Before moving on, check your interpretation of the coefficients in the models. In particular, what does the regression coefficient on the interaction term mean?

### 6.8.2 Centering continuous predictors

As noted in Section 6.3, the regression coefficient on Gender is not very interpretable when the interaction in the model, because it now corresponds to the gender gap when achrdg12 = 0. We can fix this issue by re-scaling achrdg12 so that zero has a meaningful value. One easy and widely used approach is to center achrdg12 at its mean, or, stated otherwise, to work with the deviation scores achrdg12 instead of the “raw” score. Let’s see what happens

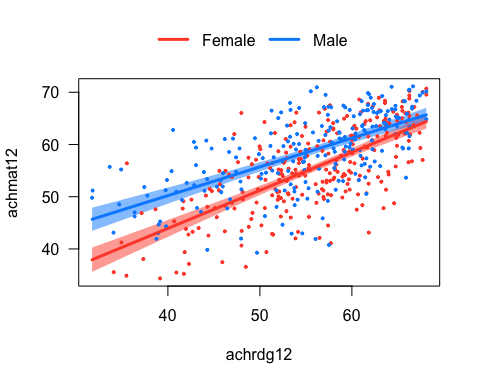
# Re-run the model with reading centered at its mean  
achrdg12\_dev <- achrdg12 - mean(achrdg12)  
mod4 <- lm(achmat12 ~ achrdg12\_dev\*gender)  
summary(mod4)

##   
## Call:  
## lm(formula = achmat12 ~ achrdg12\_dev \* gender)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.058 -3.786 0.501 4.077 16.289   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 55.2944 0.3513 157.41 < 2e-16 \*\*\*  
## achrdg12\_dev 0.7282 0.0470 15.49 < 2e-16 \*\*\*  
## genderMale 3.4993 0.5214 6.71 5.3e-11 \*\*\*  
## achrdg12\_dev:genderMale -0.1779 0.0651 -2.73 0.0065 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.8 on 496 degrees of freedom  
## Multiple R-squared: 0.462, Adjusted R-squared: 0.459   
## F-statistic: 142 on 3 and 496 DF, p-value: <2e-16

Note that the intercept and the regression coefficient on gender have changed values compared to mod3. What is the interpretation of these coefficients in the new model?

Next, let’s plot our model with the interaction term. One advantage of having everything in a single model is that we can level-up our plotting! The following code uses the visreg package. Note that the error bands in the plot are produced using the standard errors from emmeans, which is discussed in the following section. We

# Install the package if you haven't already done so  
# install.packages("visreg")  
  
# Load the package into memory  
library(visreg)  
  
visreg(mod3, xvar = "achrdg12", by = "gender", overlay = TRUE)



If you want to know more about how visreg works, type help(visreg).

### 6.8.3 Breaking down a significant interaction

If an interaction is significant, then we usually want to report a bit more information about how the focal relationship changes as a function of the moderators. There are two main ways to do this:

* Marginal effects (aka marginal means, least squares means, adjusted means): This approach is used when the focal predictor is categorical and we want to compare means across the categories, as a function of the moderator.
* Simple trends (aka simple slopes): This approach is used when the focal predictor is continuous and we want to examine the slopes of the simple trends as a function of the moderator.

Usually, the researcher will chose one or the other approach, whichever is best suited to address the research questions of interest. Our example was motivated by consideration of the gender gap in STEM (i.e., the relationship between a STEM and a categorical predictor), so the marginal effects approach is better suited. We will also illustrate simple trends, just to show how that approach works.

### 6.8.4 Marginal effects

Let’s break the interaction by asking how the relationship between Math and Gender (i.e., the gender achievement gap in Math) changes as a function of Reading. This can be done using emmeans package, and the main function in that pacakge is also called emmeans.

The three main arguments for the emmeans function:

* object – the output of lm. This is the first argument
* specs – which factors(s) in the model we want the means of
* by – which other predictors break the means down by

We can use emmeans to compute the marginal effect at the mean (MEM) as follows:

# Install the package if you haven't already done so  
 # install.packages("emmeans")  
  
# Load the package into memory  
library(emmeans)  
  
# Use the emmeans function to get the gender means on math, broken down by reading  
gap <- emmeans(mod3, specs = "gender", by = "achrdg12")  
summary(gap)

## achrdg12 = 55.6:  
## gender emmean SE df lower.CL upper.CL  
## Female 55.3 0.351 496 54.6 56.0  
## Male 58.8 0.385 496 58.0 59.5  
##   
## Confidence level used: 0.95

# Test whether the difference is significant  
contrast(gap, method = "pairwise")

## achrdg12 = 55.6:  
## contrast estimate SE df t.ratio p.value  
## Female - Male -3.5 0.521 496 -6.712 <.0001

In the above output, we only get one Gender difference in Math, and that is computed for the value of achrdg12 = 55.60188, which is the mean values of Reading. As noted, this is called the marginal effect at the mean (MEM).

It is often more helpful to report Gender difference for multiple different values of achrdg12, which is called MERV (marginal effects at representative values). While there are many ways to do chose the values, one convenient approach approach is to use the quartiles of achrdg12. This is accomplished using the cov.reduce argument of emmeans as follows.

# Use the the covarate reduce option of emmeans with the quantile function  
gap\_quartiles <- emmeans(mod3, specs = "gender", by = "achrdg12", cov.reduce = quantile)  
summary(gap\_quartiles)

## achrdg12 = 31.8:  
## gender emmean SE df lower.CL upper.CL  
## Female 37.9 1.186 496 35.6 40.3  
## Male 45.7 1.129 496 43.5 47.9  
##   
## achrdg12 = 51.2:  
## gender emmean SE df lower.CL upper.CL  
## Female 52.1 0.412 496 51.3 52.9  
## Male 56.4 0.426 496 55.6 57.2  
##   
## achrdg12 = 57.0:  
## gender emmean SE df lower.CL upper.CL  
## Female 56.3 0.355 496 55.6 57.0  
## Male 59.6 0.393 496 58.8 60.3  
##   
## achrdg12 = 61.7:  
## gender emmean SE df lower.CL upper.CL  
## Female 59.8 0.447 496 58.9 60.6  
## Male 62.2 0.482 496 61.2 63.1  
##   
## achrdg12 = 68.1:  
## gender emmean SE df lower.CL upper.CL  
## Female 64.4 0.674 496 63.1 65.7  
## Male 65.7 0.693 496 64.3 67.0  
##   
## Confidence level used: 0.95

# Test whether the gender difference in math achievement is significant at each quartile of reading achievement  
contrast(gap\_quartiles, method = "pairwise")

## achrdg12 = 31.8:  
## contrast estimate SE df t.ratio p.value  
## Female - Male -7.74 1.637 496 -4.728 <.0001   
##   
## achrdg12 = 51.2:  
## contrast estimate SE df t.ratio p.value  
## Female - Male -4.27 0.593 496 -7.207 <.0001   
##   
## achrdg12 = 57.0:  
## contrast estimate SE df t.ratio p.value  
## Female - Male -3.25 0.529 496 -6.138 <.0001   
##   
## achrdg12 = 61.7:  
## contrast estimate SE df t.ratio p.value  
## Female - Male -2.41 0.658 496 -3.659 0.0003   
##   
## achrdg12 = 68.1:  
## contrast estimate SE df t.ratio p.value  
## Female - Male -1.28 0.967 496 -1.321 0.1872

At this point, you should be able to summarize your conclusions about the gender gap in Math and how it depends on Reading.

### 6.8.5 Simple trends

Next we will show how to use emtrends to test the conditional or “simple” slopes of Math on Reading, given Gender. As mentioned, this approach is not very well suited to the example, but we are going through it here just to illustrate how to do this type of analysis.

The three main arguments for emtrends are

* object – the output of lm. This is the first argument
* var – which continuous predictor in the model we want the slopes of
* specs – which factor predictor(s) in the model to break the trend down by

Let’s see how it works.

# Use the emtrends function to get the regression coefficients on reading, broken down by gender  
simple\_slopes <- emtrends(mod3, var = "achrdg12", specs = "gender")  
summary(simple\_slopes)

## gender achrdg12.trend SE df lower.CL upper.CL  
## Female 0.728 0.0470 496 0.636 0.821  
## Male 0.550 0.0451 496 0.462 0.639  
##   
## Confidence level used: 0.95

test(simple\_slopes)

## gender achrdg12.trend SE df t.ratio p.value  
## Female 0.728 0.0470 496 15.487 <.0001   
## Male 0.550 0.0451 496 12.208 <.0001

The foregoing analysis tells us how the relationship between reading and math changes as a function of gender, and, in particular, whether the simple slopes are significant for males and females.

### 6.8.6 Two continuous predictors

Interactions with continuous predictors are basically the same as for continuous and categorical. One main issue is that we should always center the predictors, not only to facilitate interpretation of the regression coefficients, but also to reduce the correlation between the main effects and the interaction.

For an example, let’s replace gender with SES from our previous analysis. Apologies that this new example is mainly for convenience and doesn’t represent a great research question about, e.g., about why the relationships between math and reading might change as a function of SES!

Here we will focus on how centering affects the results of a regression with interactions among continuous predictors.

# Without centering  
mod5 <- lm(achmat12 ~ achrdg12\*ses)  
summary(mod1)

##   
## Call:  
## lm(formula = achmat12 ~ achrdg12 + gender + genderXreading)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.058 -3.786 0.501 4.077 16.289   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 14.8031 2.6492 5.59 3.8e-08 \*\*\*  
## achrdg12 0.7282 0.0470 15.49 < 2e-16 \*\*\*  
## genderMale 13.3933 3.6583 3.66 0.00028 \*\*\*  
## genderXreading -0.1779 0.0651 -2.73 0.00652 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.8 on 496 degrees of freedom  
## Multiple R-squared: 0.462, Adjusted R-squared: 0.459   
## F-statistic: 142 on 3 and 496 DF, p-value: <2e-16

# With centering  
achrdg12\_dev <- achrdg12 - mean(achrdg12)  
ses\_dev <- ses - mean(ses)  
mod6 <- lm(achmat12 ~ achrdg12\_dev\*ses\_dev)  
summary(mod2)

##   
## Call:  
## lm(formula = achmat12 ~ achrdg12 + gender + achrdg12:gender)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.058 -3.786 0.501 4.077 16.289   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 14.8031 2.6492 5.59 3.8e-08 \*\*\*  
## achrdg12 0.7282 0.0470 15.49 < 2e-16 \*\*\*  
## genderMale 13.3933 3.6583 3.66 0.00028 \*\*\*  
## achrdg12:genderMale -0.1779 0.0651 -2.73 0.00652 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.8 on 496 degrees of freedom  
## Multiple R-squared: 0.462, Adjusted R-squared: 0.459   
## F-statistic: 142 on 3 and 496 DF, p-value: <2e-16

We can see that, while both models account for the same overall variation in math, SES is significant in the centered model. This has to do both with changing the interpretation of the coefficient (it now represents the relationship between math and reading for students with average reading) and because it is no longer so highly redundant with the interaction term.

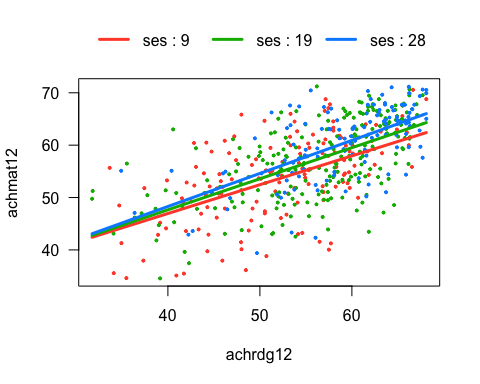
Although the interaction with SES was not significant in either model, let’s break down the interaction with emtrends just to see how it works. This time we will use the at option rather than the ’cov.reduce` option to break down the interaction.

# Break down interaction with SES as moderator  
simple\_slopes <-emtrends(mod5, var = "achrdg12", specs = "ses", at = list(ses = c(9, 19, 28)))  
summary(simple\_slopes)

## ses achrdg12.trend SE df lower.CL upper.CL  
## 9 0.550 0.0583 496 0.435 0.665  
## 19 0.593 0.0365 496 0.521 0.664  
## 28 0.631 0.0639 496 0.506 0.757  
##   
## Confidence level used: 0.95

Finally let’s summarize our (non significant) interaction with a nice plot. Note that visreg breaks the interaction down at the 10th, 50th, and 90th percentile of the by variable. You can overwrite the defaults using the breaks argument (see help(visreg)).

# Note that band = F removes the confidence intervals  
visreg(mod5, xvar = "achrdg12", by = "ses", overlay = TRUE, band = F)



### 6.8.7 Two categorical predictors

For this topic we will switch over to the ECLS data and examine how SES and Pre-K attendance interact to predict Math Achievement at the beginning of Kindergarten. The variables we will examine are

* Math Achievement at the beginning of K (c1rmscal). This is the number of correct questions on a test with approximately 70 items.
* Whether the child attended Pre-K (p1center). This is a binary variable that indicates pre-K attendance.
* SES, coded as quintiles (wksesq5). We will denote this variable as SES, but keep in mind it is quintiles in this example (e.g., SES = 1 are the respondents with SES between the minimum and the first quintile).

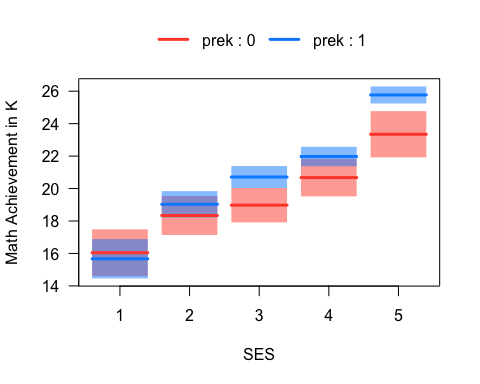
The regression model is as follows. Note that both variables need to be converted to factors in R, so that R will treat them as categorical variables. Also recall that in R the default contrast coding for categorical predictors is reference-group coding.

load("ECLS2577.Rdata")  
ecls$prek <- factor(2 - ecls$p1center)  
ecls$wksesq5 <- factor(ecls$wksesq5)  
mod <- lm(c1rmscal ~ prek\*wksesq5, data = ecls)  
summary(mod)

##   
## Call:  
## lm(formula = c1rmscal ~ prek \* wksesq5, data = ecls)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.768 -4.768 -0.975 3.955 31.232   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.045 0.735 21.82 < 2e-16 \*\*\*  
## prek1 -0.373 0.960 -0.39 0.6973   
## wksesq52 2.293 0.957 2.40 0.0166 \*   
## wksesq53 2.930 0.913 3.21 0.0013 \*\*   
## wksesq54 4.631 0.944 4.91 9.9e-07 \*\*\*  
## wksesq55 7.299 1.034 7.06 2.2e-12 \*\*\*  
## prek1:wksesq52 1.064 1.212 0.88 0.3801   
## prek1:wksesq53 2.109 1.154 1.83 0.0679 .   
## prek1:wksesq54 1.671 1.168 1.43 0.1527   
## prek1:wksesq55 2.797 1.234 2.27 0.0235 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.9 on 2567 degrees of freedom  
## Multiple R-squared: 0.162, Adjusted R-squared: 0.159   
## F-statistic: 55.2 on 9 and 2567 DF, p-value: <2e-16

To facilitate interpretation of the ouput, you can refer to the plot below. Each regression coefficient in the output corresponds to a feature of this plot.

visreg::visreg(mod, xvar = "wksesq5", by = "prek",   
 partial = F, rug = F, overlay = T,   
 strip.names = T, xlab = "SES",   
 ylab = "Math Achievement in K")



In order to summarize the model as an ANOVA table, we can use the following code. Note that the ANOVA output tests the variance explained (i.e., R-squared) of the original variables, and does not include dummy variables.

anova(mod)

## Analysis of Variance Table  
##   
## Response: c1rmscal  
## Df Sum Sq Mean Sq F value Pr(>F)   
## prek 1 3434 3434 72.14 <2e-16 \*\*\*  
## wksesq5 4 19914 4978 104.58 <2e-16 \*\*\*  
## prek:wksesq5 4 299 75 1.57 0.18   
## Residuals 2567 122198 48   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# 7 Model Building

Up until this point we have considered models with only a handful of predictors. This chapter addresses how to scale-up to ‘real life’ applications in which we may have many predictors.

There are two basic approaches to model building with many predictors. A *simulataneous* model is a single regression model that includes all predictors of interest. When there are many predictors, this is sometimes referred to the “kitchen sink” approach – just put everything in the model (including the kitchen sink). Conceptually, this approach has the advantage that it mitigates omitted variable bias and therefore helps justify the causal interpretation of regression coefficients. However, in practice it can lead to computational problems, especially when highly correlated predictors are included in the model. We discuss this situation in Section 7.4

*Hierarchical* model building is an alternative approach in which predictors are added into a model sequentially. A hierarchical approach is most appropriate when the predictors can be partitioned into conceptual “blocks,” and the blocks are added into the model one at a time. A block is just a conceptual grouping of one or more predictors, and we will some examples as well as in Sections 7.1 and 7.3) The rationale for adding the blocks of predictors sequentially is to isolate the proportion of variance in the outcome (i.e., R-squared) attributable to each block.

Let’s illustrate the difference between these two approaches with a hypothetical example. In the example, we seek to explain Reading Achievement using predictors that can be conceptually grouped into student and curriculum factors.

* Student factors
  + age
  + previous achievement
  + parental education
  + SES
* Curriculum factors
  + in-classroom reading (hrs)
  + peer discussion (hrs)
  + reading materials (level)
  + homework (hrs)

In a simultaneous model, all predictors would be entered into a regression model at the same time and we would interpret the regression coefficients and R-squared from that one model.

In a hierarchical model, we could instead proceed as follows.

* First we run a regression model with just the student factors. These variables are known to predict Reading Achievement, but are not malleable from the perspective of educational policy. That is, we know they are important for predicting the outcome, but they aren’t the types of factors that can be intervened upon via schooling. These types of predictors are often called “control variables.” What counts as a control variable is a largely non-statistical consideration that depends on the research context.
  + After we run the regression model with this first block of predictors, we record the R-squared statistic. We don’t care about the value of the regression coefficients at this point, since we know we have left out the curriculum factors.
* Next we run a regression model with student factors *and* the curriculum factors, so that our model now includes all of the predictors. By comparing the the R-squared of the model in this second step (curriculum and student factors) to the R-squared of the model in the first step (only student factors), we can determine how much additional variation in Reading Achievement was explained by the curriculum factors. If there is a substantial amount additional of variation explained by the curriculum factors, we can reason that intervening on the curriculum may be a good way to improve Reading Achievement.

It is important to note the following points about this example:

* The second step in the hierarchical model is the same the simultaneous model – they both include all of the predictors. This is true in general. The final step of a hierarchical model is always the same as the corresponding simultaneous model.
* So, what does the hierarchical model give us that the simultaneous model doesn’t? The change in R-squared attributable to each additional block of predictors, controlling for the previous block(s) of predictors. This “change-in-R-squared” statistic is the main focus of this chapter.
* We usually interpret the regression coefficients in the simultaneous model only (i.e., in the last step of the hierarchical model), because any model that doesn’t include all of the relevant predictors is subject to omitted variable bias.
* It is important to distinguish model building from variable selection. Model building is about research-question-driven strategies for including a fixed set of predictors in a regression model. Variable selection is about “weeding out” predictors that are not useful. So, in model building, the final model should always include all of the potential predictors, but in variable selection, the final model may include only a subset of the potential predictors.
* Variable selection is an important topic that is the focus of a lot current research in machine learning (e.g, LASSO). There was also a large number of “step-wise” regression procedures proposed for variable selection in last century, but they didn’t work (i.e., different procedures resulted in different models and there were no good criteria by which to choose amongst them). If you have a research scenario that requires variable selection, this problem is best dealt with through machine learning. If you are interested, check out this resource: <https://glmnet.stanford.edu> and feel free to ask questions in class.
* A note on terminology: In some research areas, “hierarchical” is used to refer to multilevel models (e.g., “hierarchical linear models” in sociology). That is not the usage here. It would make sense to refer to the types of models we are calling “hierarchical” as “step-wise,” but, as mentioned, the term “step-wise” already refers to a large number flawed techniques for variable selection. So… here we are.

One last thing – in general, you should build models that answer your research questions. This chapter illustrates the main statistical techniques in terms of hierarchical regression, but you can use those techniques in other ways as well.

## 7.1 Hierarhical models

The key ideas of hierarchical modeling are:

1. Partition the predictors into conceptual “blocks.” Each block contains at least one predictor, but they usually contain more than one (e.g., “control variables”). No predictor can be included in more than one block.
2. Build a series of regression models in which the blocks are entered into the model sequentially. For this to be *hierarchical* model building it is required that each new model in the series includes the blocks from all previous models. So, with three blocks we would have:
3. Compare the R-squared’s of each subsequent model to determine whether the new block “adds anything” to the model. Denoting the R-squared values of the three models as , the focal quantities are:
   * Compute to evaluate the contribution of Block 1
   * Compute to evaluate contribution of Block 2, controlling for Block 1
   * Compute to evaluate contribution of Block 3, controlling for Block 1 & 2.

The symbol is read “delta R-squared” and the subscripts denote which two models are being compared.

The main advantage of hierarchical modeling is that we partition the variance explained by all of the variables together ( in this example) into parts contributed uniquely by each block:

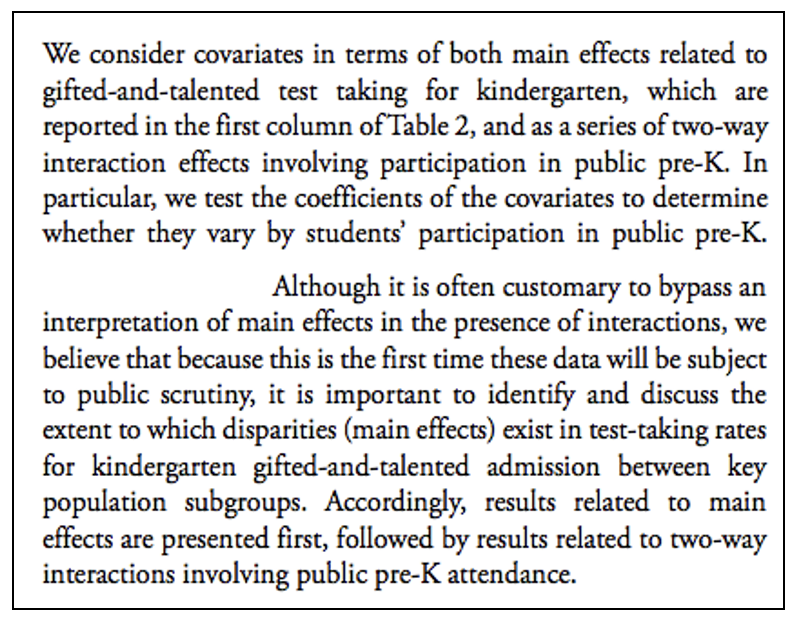
Sometimes we write as , since is the contribution of the first block of predictors after controlling for a model with no predictors (“Model 0”). This makes the notation a bit more consistent:

There are many other approaches to model building, but it is the above equation that uniquely defines *hierarhical* modeling building. In short, hierarchical modeling is about partitioning the total variance explained by a set of predictors (here denoted ) in to the variance uniquely attributable to different blocks of predictors (the s).

### 7.1.1 Example of blocks

We have already seen some examples of blocks of predictors.

* Categorical predictors with categories are represented as blocks of dummy variables. The dummies are conceptually equivalent to a single predictor.
* Interactions are blocks of predictors. Often interactions are entered into a regression model on separate step, after the “main effects” have been entered. Below is an excerpt that explains this rational (cite: Lu & Wienberg (2016). Public Pre-K and Test Taking for the NYC Gifted-and-Talented Programs: Forging a Path to Equity. Educational Researcher, 45, 36-47):



More generally, blocks are any subset of predictors about which you want to ask a research question. It can be just one variable, or it can be a collection of conceptually related variables.

**Please take a moment to write down one or more example of a block of variables from your area of research, and I will invite you to share your examples in class.**

### 7.1.2 Nested models

Whatever your blocks are, it is important that they are used to make a sequence of *nested models*. The idea behind nesting is depicted in Figure 7.1.

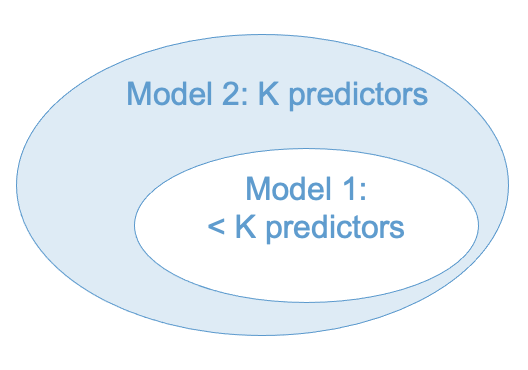


Figure 7.1: Nested Models

In words: Model 1 is nested within Model 2 if Model 1 can be obtained from Model 2 by removing some predictors from Model 2. We can think of this like Matryoshka nesting dolls – Model 1 is contained within Model 2 in the sense that all of the predictors in Model 1 are also in Model 2. In the Figure, this is indicated by Model 1 being contained within Model 2.

Not all models that can be made from a set of predictors are nested. The difference between nested and non-nested models is illustrated in Figure 7.2. (The partially overlapping ovals are intended to indicate that the models share some predictors, but each model also has some unique predictors.)

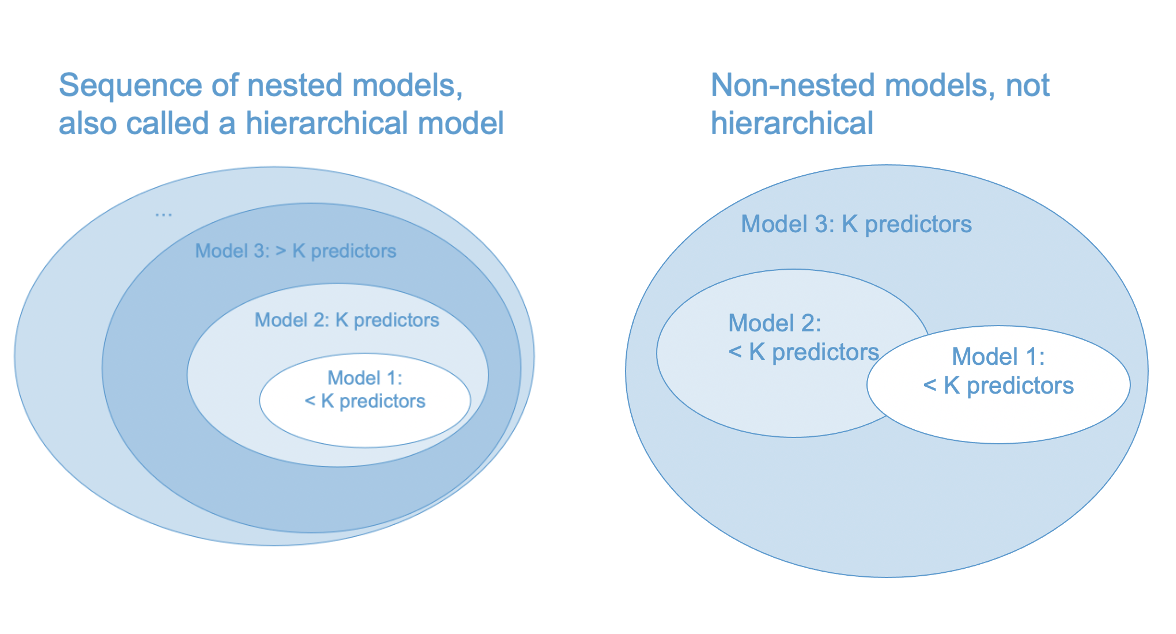


Figure 7.2: Nested vs Non-Nested Models

A hierarchical model is a series of nested regression models combined with a specific sequence of statistics. It is important that the models are nested, otherwise we can’t compute statistical tests for the statistics (more on this in Section 7.2). In short, we can compare Model 1 and Model 2 in the left hand panel of Figure 7.2, but not in the right hand panel.

(Side note: In the right hand panel, We could compare Model 1 vs Model 3 and Model 2 vs Model 3, but this would not be a hierarchical model because the s wouldn’t add up to the of Model 3.)

### 7.1.3 Some practical advice

Before moving on to the math, it might be helpful to consider some practical advice about hierarchical modeling. The advice offered by the late Jacob Cohen is summarized below.

* The hierarchical approach provides an effective strategy of inference if variables are entered into the model according to their relevance to the study.
* The variables that are most central to the study should be entered first, those that are secondarily of interest should be entered second, and those that fall into the category, “I wonder if” or “just in case” should be entered last.
* This principle may be succinctly stated as “least is last.” That is, “when research factors can be ordered as to their centrality, those of least relevance are appraised last in the hierarchy, and their results taken as indicative rather than conclusive” (Cite:CCWA 5.7.3).

Figure 7.3 maps Cohen’s strategy onto our diagram for nested models. One addendum to this approach is that “control variables” are usually entered first. These are variables that have been established as important in past research, but are not of central interest to the present study. The rationale for entering them first is that new studies should add something beyond what has been established by past research.

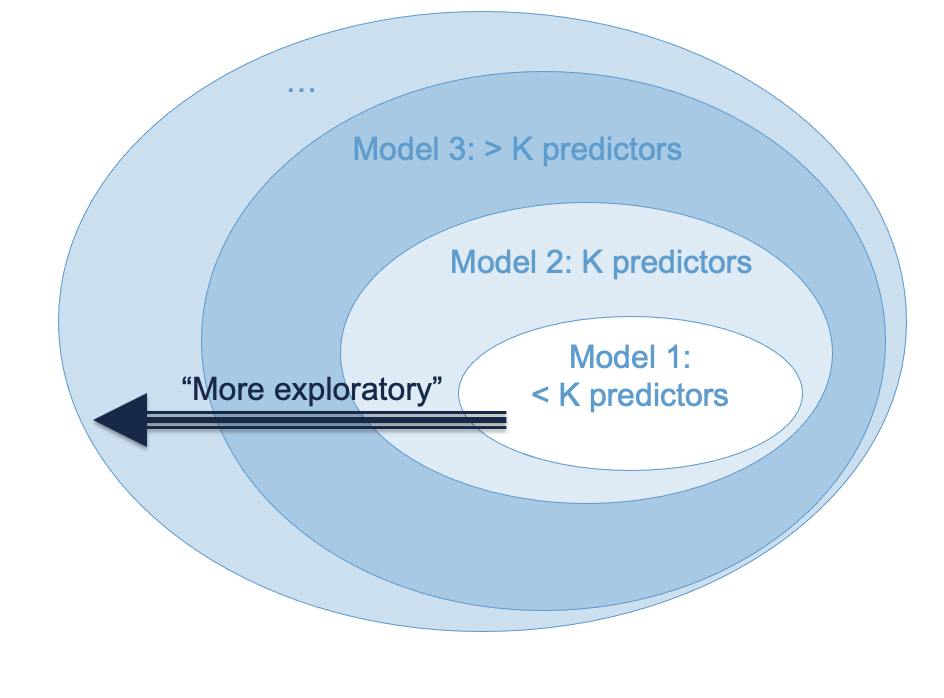


Figure 7.3: Cohen’s ‘Least is Last’ Principle

## 7.2

We have already seen how to compute – by subtracting the R-squared value of a nested model from that of a nesting model. This is a very widely used way of comparing nested regression models.

When adding predictors into a model one at a time (i.e., blocks with only a single predictor), testing is equivalent to testing the b-weight of the added predictor. But when adding multiple predictors into a model, provides important information that is not captured by any of the statistics we have considered so far.

### 7.2.1 Inference for

Let Model 1 be a regression model with predictors that is nested within a second model, Model 2, with predictors. Define and . To test the hypothesis versus we can use the test statistic

which has an F-distribution on and degrees of freedom, when the null hypothesis true. This test assumes that Model 2 satisfies the linear regression population model (see Section 2.5).

## 7.3 A worked example

To illustrate hierarchical modeling, let’s use the ECLS data to address whether reading achievement at the beginning of Kindergarten (c1rrscal) is related to SES (wksesl), parental (mother’s and father’s) education (wkmomed and wkdaded, respectively), and attendance in center-based care before K (p1center; pre-K status, for short).

This example is intended to be a realistic illustration of how to use the techniques we have covered up to this point, and also shows how to report results in a regression table. But there are some caveats that should be mentioned before starting:

* We used some tricks from later chapters to deal with non-linearity, and
* keep in the mind the subset of ECLS data we are using is not a representative sample.

### 7.3.1 Statement of research questions

We will address the following research questions.

* RQ1: Does SES predict Reading Achievement upon entry to K?
* RQ2: Does Parental Education (mother’s and father’s) predict Reading Achievement, after controlling for SES?
* RQ3: After controlling for both SES and Parental Education, is Pre-K Status associated with better Reading Achievement in K?
* RQ4: Does the relationship of SES or Parental Education to Reading Achievement change as function of Pre-K Status – in more causal language: does Pre-K participation reduce pre-existing disparities in Reading Achievement upon entry to K?

### 7.3.2 Modeling building strategy

The research questions were addressed via hierarchical regression modeling with the following blocks of predictors.

* Block 1: wksesl
* Block 2: wkmomed and wkdaded
* Block 3: plcetner
* Block 4: Interactions between p1center and the other predictors

To aid the interpretation of regression coefficients in the presence of interactions, all continuous variables (predictors and outcome) were transformed to z-scores. plcenter was coded as binary indicator for pre-K status.

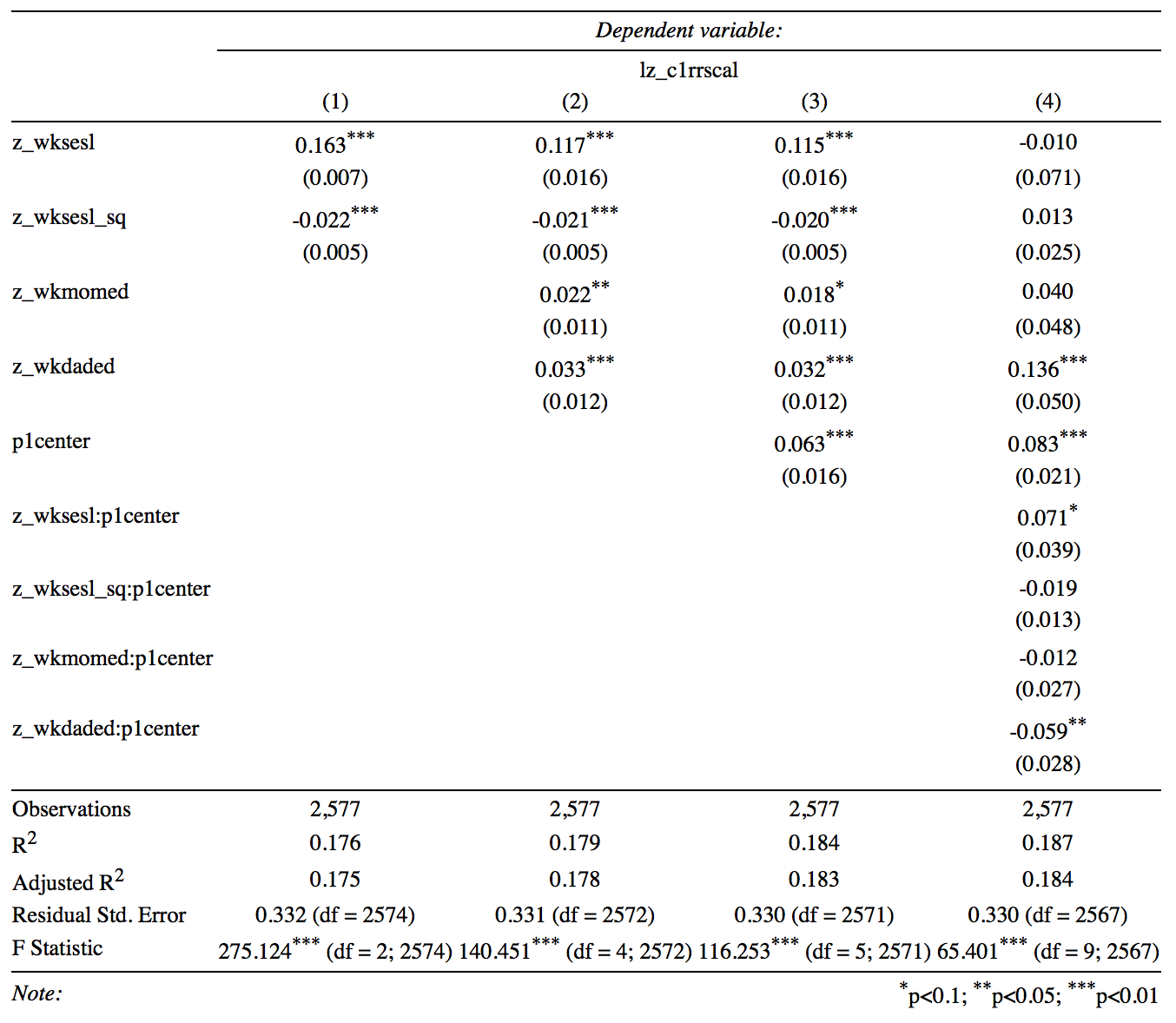
The blocks were entered into the model in the indicated order. The RQs were addressed by examining the R-square change () for each corresponding block. (In practice, we would also interpret the sign and direction of the significant regression coefficients in the final model, but we will leave that as an exercise.) A significance level of .05 was used for all tests reported in this study.

Preliminary examination of the model residuals indicated violation of the assumption of linearity. To address this, the outcome variable was log-transformed and a quadratic term was added for wksesl (wksesl\_sq). These procedures for checking and addressing non-linearity are discussed in the coming chapters.

### 7.3.3 Results

The following table reports the results of the hierarchical model. The table was produced using the stargazer package in R, and is representative of the usual format for reporting regression output from hierarchical regression models. The rows are the variables and the columns are the series of nested models. The values reported for each variable are the regression coefficients, with the standard errors in parenthesis. **Please take a moment to look over the table and write down any questions you have about its interpretation**

load("ECLS2577.RData")  
attach(ecls)  
  
# Recode variables  
p1center <- 2 - p1center  
z\_c1rrscal <- scale(c1rrscal)  
z\_wksesl <- scale(wksesl)  
z\_wksesl\_sq <- z\_wksesl^2  
z\_wkmomed <- scale(wkmomed)  
z\_wkdaded <- scale(wkdaded)  
log\_c1rrscal <- log(z\_c1rrscal - min(z\_c1rrscal) + 1)  
  
# Run models  
# models  
mod1 <- lm(log\_c1rrscal ~ z\_wksesl + z\_wksesl\_sq)  
mod2 <- lm(log\_c1rrscal ~ z\_wksesl + z\_wksesl\_sq + z\_wkmomed + z\_wkdaded)  
mod3 <- lm(log\_c1rrscal ~ z\_wksesl + z\_wksesl\_sq + z\_wkmomed + z\_wkdaded + p1center)  
mod4 <- lm(log\_c1rrscal ~ (z\_wksesl + z\_wksesl\_sq + z\_wkmomed + z\_wkdaded)\*p1center)  
  
# Regression Table  
#stargazer::stargazer(mod1, mod2, mod3, mod4, type = 'html', omit = "Constant")  
detach(ecls)



The above table reports the for each model but does not provide the statistics or the corresponding F-tests. The statistics were

R2\_1 <- summary(mod1)$r.squared  
R2\_2 <- summary(mod2)$r.squared  
R2\_3 <- summary(mod3)$r.squared  
R2\_4 <- summary(mod4)$r.squared  
  
delta\_rsquared <- c(R2\_1, R2\_2 - R2\_1, R2\_3 - R2\_2, R2\_4 - R2\_3)  
Models <- paste0("Model ", 1:4)  
names(delta\_rsquared) <- Models  
delta\_rsquared

## Model 1 Model 2 Model 3 Model 4   
## 0.176122 0.003150 0.005124 0.002133

The F-tests of the s for each model are give are reported below. The rows of the table below the report the F-test of for each model. The test for Model 1 is omitted.

# Tests of R-square change  
knitr::kable(cbind(Models, anova(mod1, mod2, mod3, mod4)))

| Models | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
| --- | --- | --- | --- | --- | --- | --- |
| Model 1 | 2574 | 283.1 | — | — | — | — |
| Model 2 | 2572 | 282.0 | 2 | 1.0825 | 4.971 | 0.0070 |
| Model 3 | 2571 | 280.3 | 1 | 1.7606 | 16.168 | 0.0001 |
| Model 4 | 2567 | 279.5 | 4 | 0.7329 | 1.683 | 0.1512 |

**Before moving on, please write down your a conclusions about which blocks had statistically significant values of .**

### 7.3.4 Interpretation

The focus of this section is on the interpretation of the statistics and their significance. In practice, we would also interpret the sign and direction of the significant regression coefficients in the final model. The outcome variable was log transformed but we ignore that detail here.

* RQ1: SES had significant linear and quadratic relationships with Reading Achievement. Together, both predictors explained about 17.6% of the variance in log Reading (.
* RQ2: After controlling for SES, Parental Education explained a small but statistically significant proportion of the variance in Reading Achievement (),
* RQ3: After controlling for SES and Parental Education, Pre-K status explained an additional .5% of the variation in reading achievement ().
* RQ4: There was mixed evidence about the interactions between Pre-K Status and the other predictors. The F-test of R-squared change indicated that the interaction terms did not explain a significant amount of variation in reading (). However, the interaction with paternal education was significant at the .05 level. As shown below, re-running Model 4 with only the paternal education interaction, it was found that this effect was not significant at the .05 level (). It may be concluded that there were no significant interactions with Pre-K status.

mod4A <- lm(log\_c1rrscal ~ z\_wksesl + z\_wksesl\_sq + z\_wkmomed + z\_wkdaded\*p1center)  
summary(mod4A)

##   
## Call:  
## lm(formula = log\_c1rrscal ~ z\_wksesl + z\_wksesl\_sq + z\_wkmomed +   
## z\_wkdaded \* p1center)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.8875 -0.2315 -0.0003 0.2086 1.3488   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.79583 0.01516 52.49 < 2e-16 \*\*\*  
## z\_wksesl 0.11516 0.01632 7.06 2.2e-12 \*\*\*  
## z\_wksesl\_sq -0.01964 0.00531 -3.70 0.00022 \*\*\*  
## z\_wkmomed 0.01842 0.01117 1.65 0.09933 .   
## z\_wkdaded 0.04277 0.01702 2.51 0.01202 \*   
## p1center 0.06027 0.01616 3.73 0.00020 \*\*\*  
## z\_wkdaded:p1center -0.01402 0.01637 -0.86 0.39183   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.33 on 2570 degrees of freedom  
## Multiple R-squared: 0.185, Adjusted R-squared: 0.183   
## F-statistic: 97 on 6 and 2570 DF, p-value: <2e-16

In summary, it was found that SES and Parental Education were significant predictors of children’s Reading Achievement upon entry to Kindergarten, indicating that pre-existing disparities in reading due to economic factors are apparent at even this very early age. Attendance in Pre-K explained an additional .5% of variance in Reading Achievement and was associated with increased Reading Achievement; however, there was no evidence that participation in Pre-K reduced pre-existing disparities in Reading Achievement associated with the economic factors (i.e., there were no interactions between Pre-K participation and the economic factors).

## 7.4 Too many predictors?

What is better: more predictors or fewer? While there is no definitive answer to this question, there are two points that we should keep in mind when adding variables into a model:

* The ratio of sample size to number of predictors: .
* The correlation among the predictors (AKA multicollinearity).

In order to understand the effect of these factors on multiple regression we need the formula for the standard error of a regression coefficient, which is presented below:

Also recall that the standard error is the inverse of the precision, and that is the R-squared from the regression of the -th predictor on the remaining predictors.

### 7.4.1

We can see from the equation for that increasing the number of predictors leads the term to get smaller. Consequently, the ratio

will get larger (and hence the standard error will get larger) if the increase in the number of predictors is not compensated for by an increase in In other words, predictors that do not add explanatory power to the model (i.e., do not increase R-squared) will lead to reduced precision for all of the regression coefficients. A similar story applies to the adjusted R-squared statistic discussed in Section 4.7.

So long as N >> K, the effect of the number of predictors is negligible. But when the number of predictors gets close to the sample size, there is a problem. Many people seek “rules of thumb” for the ratio of observations to predictors. The only hard rule is , but if -ish you should start worrying about the number of predictors in your model. If you really want to know how large your sample size needs to be, do a power analysis (Section 2.10).

### 7.4.2 Multicollinearity

Regardless of sample size, when a predictor is highly correlated with the other predictors, its precision is negatively affected via the term . When the situation is called multicollinearity. Note that multicollinearity can occur even if none of the pairwise correlations among the predictors is large – it has to do with the relation between predictor and *all* of the other predictors, which is why it is called *multi*collinearity.

The term is called the *tolerance* and its reciprocal is called the variance inflation factor (VIF).

Note that the is in the standard error of the regression coefficient.

There is no exact cut off for , but think of it as telling you how many times the variance of a regression coefficient (i.e., its squared standard error) increases due to the predictor’s linear association with the other predictors.

* If VIF = 5, the variance is 5 times larger than it would be without multicollinearity.
* If VIF = 10, the variance is 10 times larger than it would be without multicollinearity.
* If VIF > 10, it is a good idea to reconsider whether the predictor needs to be in the model.

In general, it is a good practice to check VIF as part of model diagnostics, which is the topic of the next chapter. For our example, the VIF statistics for Model 3 and Model 4 are reported below.

Model 3:

car::vif(mod3)

## z\_wksesl z\_wksesl\_sq z\_wkmomed z\_wkdaded p1center   
## 6.290 1.206 2.950 3.377 1.054

Model 4

car::vif(mod4)

## z\_wksesl z\_wksesl\_sq z\_wkmomed   
## 28.158 6.447 12.800   
## z\_wkdaded p1center z\_wksesl:p1center   
## 13.816 1.812 28.320   
## z\_wksesl\_sq:p1center z\_wkmomed:p1center z\_wkdaded:p1center   
## 7.562 12.517 14.012

It is apparent that including the interaction terms led to a high degree of multicollinearity for some regression coefficients in the model. Based on this information, as well as the non-singificant for the interaction block, it would be advisable to focus our interpretation on Model 3 rather than Model 4.

## 7.5 Workbook

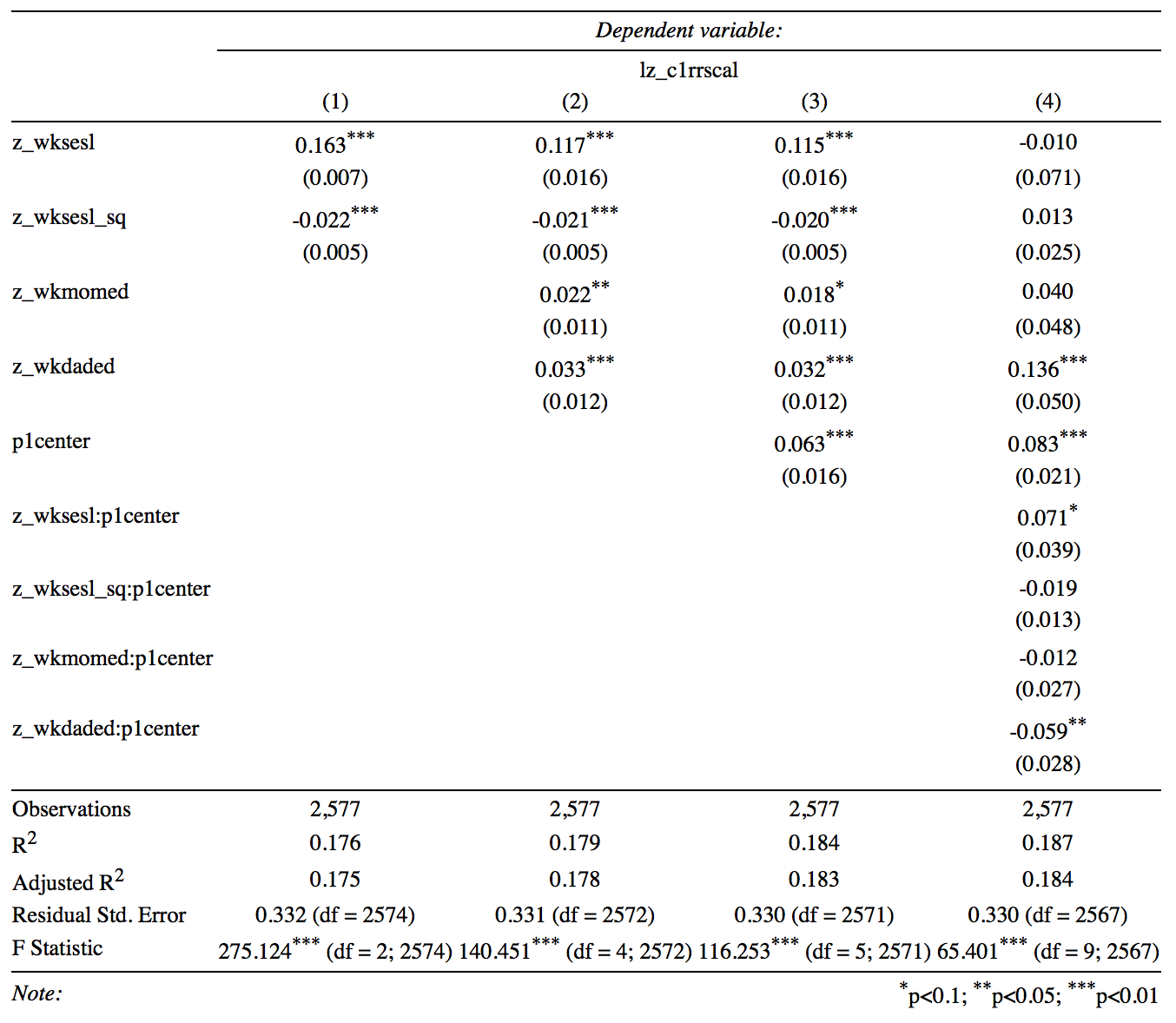
This section collects the questions asked in this chapter. We will discuss these questions in class. If you haven’t written down / thought about the answers to these questions before class, the lesson will not be very useful for you! So, please engage with each question by writing down one or more answers, asking clarifying questions, posing follow up questions, etc.

**Section 7.1**

Please take a moment to write down one or more example of a block of variables from your area of research, and I will invite you to share your examples in class.

**Section 7.3**

* Look over the table below and write down any questions you have about its interpretation.



* Please write down your a conclusions about which blocks had statistically significant values of .

# Tests of R-square change  
knitr::kable(anova(mod1, mod2, mod3, mod4))

| Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
| --- | --- | --- | --- | --- | --- |
| 2574 | 283.1 | — | — | — | — |
| 2572 | 282.0 | 2 | 1.0825 | 4.971 | 0.0070 |
| 2571 | 280.3 | 1 | 1.7606 | 16.168 | 0.0001 |
| 2567 | 279.5 | 4 | 0.7329 | 1.683 | 0.1512 |

## 7.6 Exercises

The exercises reproduce the worked example in Section 7.3. For additional details on the interpretation of the output, please see that section.

First we set up the variables

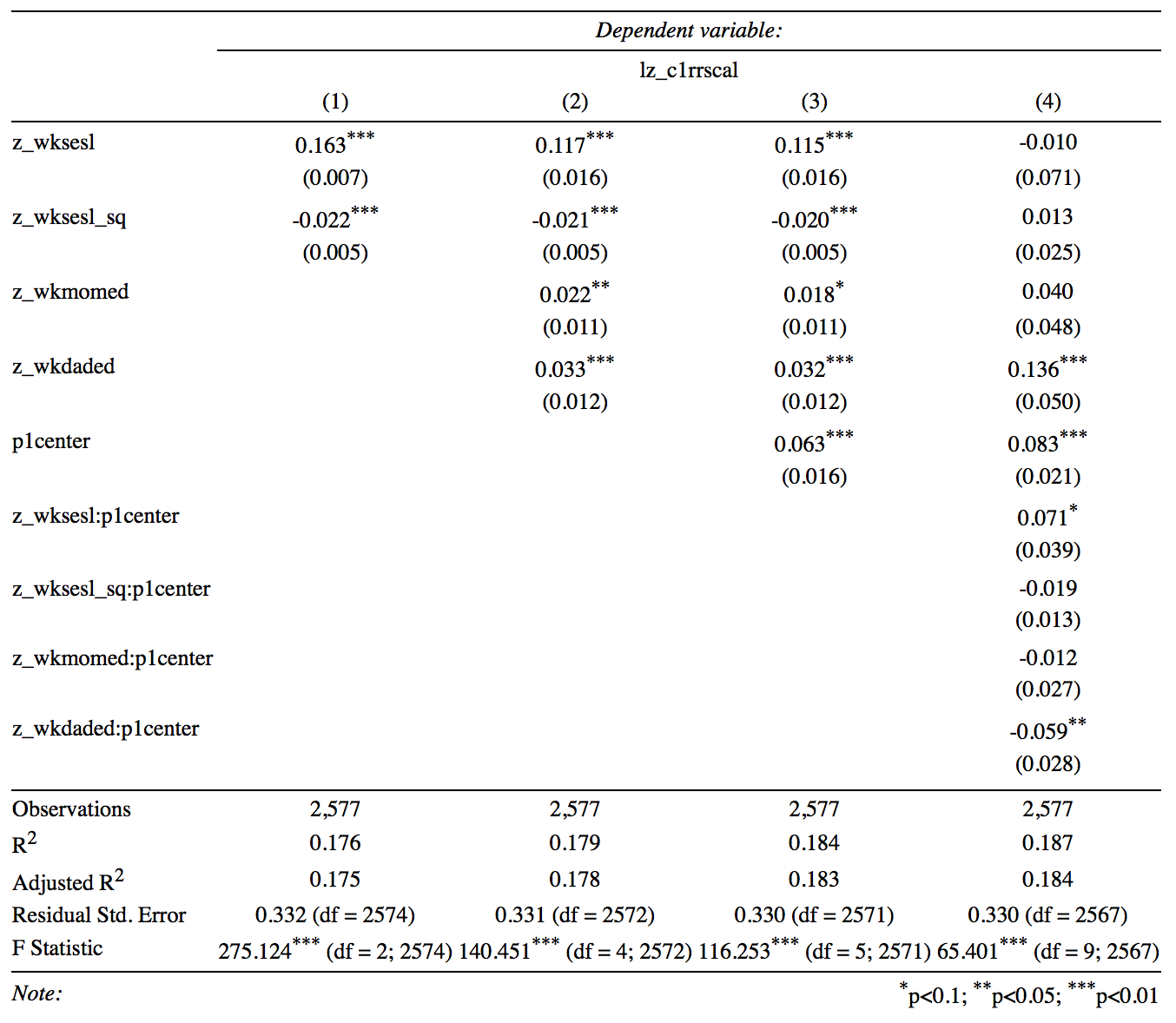
rm(list = ls()) # clean up environment  
load("ECLS2577.RData")  
attach(ecls)  
  
# Recode variables  
p1center <- 2 - p1center  
z\_c1rrscal <- scale(c1rrscal)  
z\_wksesl <- scale(wksesl)  
z\_wksesl\_sq <- z\_wksesl^2  
z\_wkmomed <- scale(wkmomed)  
z\_wkdaded <- scale(wkdaded)  
log\_c1rrscal <- log(z\_c1rrscal - min(z\_c1rrscal) + 1)

Next, run the models.

# Run models  
mod1 <- lm(log\_c1rrscal ~ z\_wksesl + z\_wksesl\_sq)  
mod2 <- lm(log\_c1rrscal ~ z\_wksesl + z\_wksesl\_sq + z\_wkmomed + z\_wkdaded)  
mod3 <- lm(log\_c1rrscal ~ z\_wksesl + z\_wksesl\_sq + z\_wkmomed + z\_wkdaded + p1center)  
mod4 <- lm(log\_c1rrscal ~ (z\_wksesl + z\_wksesl\_sq + z\_wkmomed + z\_wkdaded)\*p1center)  
  
# Regression Table  
#stargazer::stargazer(mod1, mod2, mod3, mod4, type = 'html', omit = "Constant")  
detach(ecls)

The stargazer table doesn’t format nicely in markdown, so I attach a screen shot instead. We will go through some different types of output with stargazer in class You must install the stargazer package for the function to work.

knitr::include\_graphics("images/stargazer.png")



The above table reports the for each model but does not provide the statistics or the corresponding F-tests. The statistics were

R2\_1 <- summary(mod1)$r.squared  
R2\_2 <- summary(mod2)$r.squared  
R2\_3 <- summary(mod3)$r.squared  
R2\_4 <- summary(mod4)$r.squared  
  
delta\_rsquared <- c(R2\_1, R2\_2 - R2\_1, R2\_3 - R2\_2, R2\_4 - R2\_3)  
Models <- paste0("Model ", 1:4)  
names(delta\_rsquared) <- Models  
delta\_rsquared

## Model 1 Model 2 Model 3 Model 4   
## 0.176122 0.003150 0.005124 0.002133

The F-tests of the s for each model are give are reported below. The rows of the table below the report the F-test of for each model. The test for Model 1 is omitted.

# Tests of R-square change  
knitr::kable(cbind(Models, anova(mod1, mod2, mod3, mod4)))

| Models | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
| --- | --- | --- | --- | --- | --- | --- |
| Model 1 | 2574 | 283.1 | — | — | — | — |
| Model 2 | 2572 | 282.0 | 2 | 1.0825 | 4.971 | 0.0070 |
| Model 3 | 2571 | 280.3 | 1 | 1.7606 | 16.168 | 0.0001 |
| Model 4 | 2567 | 279.5 | 4 | 0.7329 | 1.683 | 0.1512 |

The variance inflation factors for Model 3 and Model 4 are obtained use the vif function of the car package (you must install the package for this code to work)

car::vif(mod3)

## z\_wksesl z\_wksesl\_sq z\_wkmomed z\_wkdaded p1center   
## 6.290 1.206 2.950 3.377 1.054

car::vif(mod4)

## z\_wksesl z\_wksesl\_sq z\_wkmomed   
## 28.158 6.447 12.800   
## z\_wkdaded p1center z\_wksesl:p1center   
## 13.816 1.812 28.320   
## z\_wksesl\_sq:p1center z\_wkmomed:p1center z\_wkdaded:p1center   
## 7.562 12.517 14.012

# 8 Assumption checking

In Section 2.5 we introduced the population model for linear regression. But how do we know whether this model applies to our data? That is the question we address in this chapter. In particular, we discuss data analyses that can be used to better understand whether the population assumptions of linear regression are consistent with our data. This is referred to *assumption checking*.

Why do we care about assumption checking? Well, if the population assumptions of linear regression are not met, any of the following can happen:

* Regression coefficients and R-squared could be biased
* Standard errors could be too small (or too large)
* t- and F-tests could be too small (or too large)
* p-values could be too small (or too large)
* Confidence intervals could be too small (or too large)

Basically, all of the numbers we get in the regression output from R could be wrong.

The philosophy of assumption checking is not to verify the assumptions or prove beyond doubt that they are true – i.e., we are not *testing* assumptions, we are just checking them. Checking means we want to know if there are any problems with a regression model that might affect our conclusions (e.g., which predictors are significant). So, for example, we don’t really care if the residuals are normally distributed – we just care if there are any violations of this assumption that might be affecting the results our analysis. If there is no evidence that an assumption is problematic, we proceed with the main analysis as intended. We address ways to deal with “violations” of model assumptions in Chapter 9.

Assumption checking generally involves plotting the residuals from a regression model and trying to infer what the plots tell us about the population model. The focus of this chapter is to introduce you to these plots and how to interpret them. Sometimes assumption checking can feel a bit like reading tea leaves, because interpreting plots can be pretty subjective. Honing your interpretation of will happen gradually with experience.

*Regression diagnostics* are a related topic that we also introduce this chapter. These are procedures for detecting outliers. Unlike assumption checking, which focuses on the model *per se*, diagnostics focus on whether individual data points are having an undue influence on the model (e.g., the predicted values or the parameter estimates). We only cover the basics behind diagnostics in this chapter and the material is optional (i.e., it will not be assessed in this course, but you may find it useful for your research).

Outlier detection can be useful for identifying influential data points, but it is almost never the the case that data points should be omitted because they are outliers. Unless you can find something specifically wrong with a data point (e.g., a data entry error) you should not omit data. A better way to deal with outliers is by using statistical procedures that are specifically designed to deal with them, called *robust statistics*. Robust regression is an advanced topic that we won’t get to in this course, but check out this resource if you are interested and feel free to ask questions in class: <https://cran.r-project.org/web/views/Robust.html>.

## 8.1 Recap of population model

Before moving, let’s recap the population model for linear regression. This was introduced for simple linear regression in Section 2.5. To restate the assumptions for multiple linear regression we will use the vector notation

to represent the predictor variables. For our purposes, the vector is just a list of the predictors in a model.

1. Normality: The distribution of conditional on is normal for all values of .
2. Homoskedasticity: The conditional distributions have equal variances (also called homegeneity of variance).
3. Linearity: The means of the conditional distributions are a linear function of .

These three assumptions can be summarized in term of the model residuals , which subtract off the regression line and therefore have mean of zero, but otherwise have the same distribution as :

These assumptions about the residuals are presented graphically in Figure 8.1. It is instructive to compare this with Figure 2.4. In particular, note that the “X” axis is replaced by , which is an approach for plotting residuals used later in this chapter.

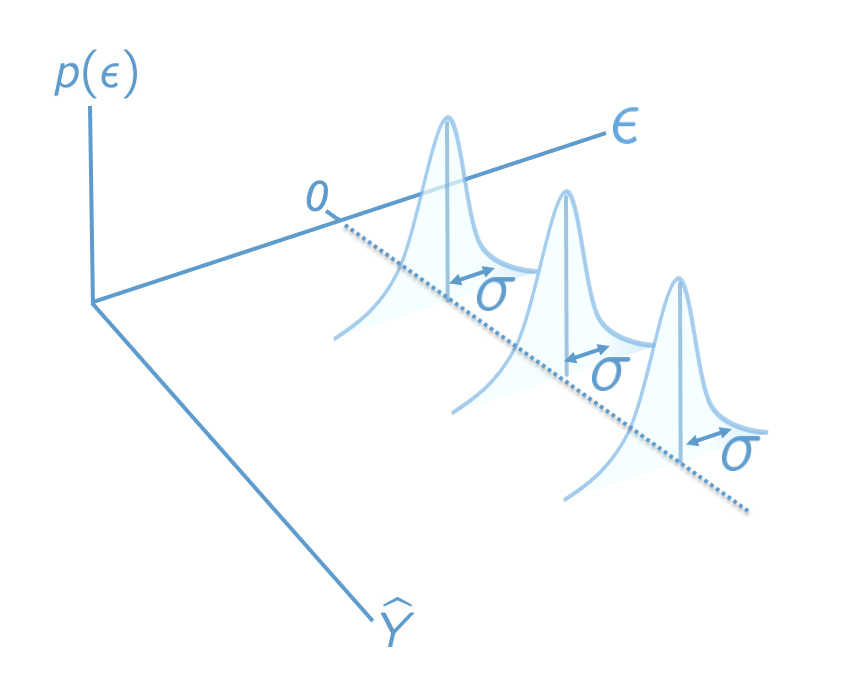


Figure 8.1: The Multiple Regression Population Model.

## 8.2 Linearity

This assumption is about whether the regression function is “really” a line or if it could be better represented as some other relationship. A classic example is shown below. The example is from Anscombe’s quartet, which we address in more detail in Section 8.6

# non linearity in Anscombe's second example  
attach(anscombe)  
mod <- lm(y2 ~ x2)  
  
# Take a look at the raw data  
par(mfrow = c(1, 2))  
plot(x2, y2, col = "#4B9CD3", xlab = "X", ylab = "Y")  
abline(mod)  
  
# Compare to the residual vs fitted plot  
plot(mod, which = 1)

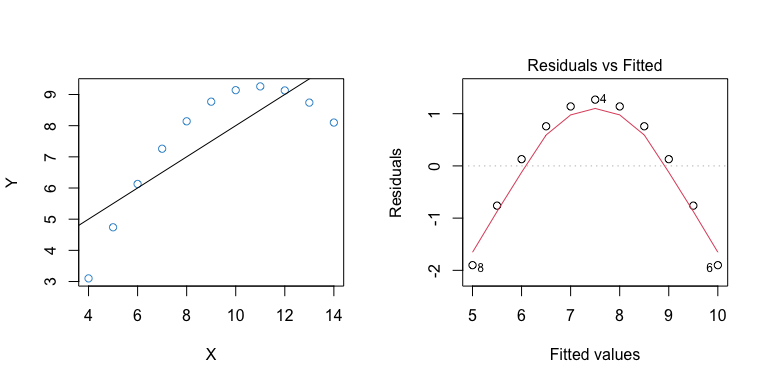


Figure 8.2: Anscombe’s second dataset

detach(anscombe)

The left hand panel of Figure 8.2 shows the scatter plot of the example data. It is pretty obvious that the relationship between and is not linear.

The right hand panel shows the residuals versus the predicted (“fitted”) values from the regression of on . This is the sample analogue of the population model in Figure 8.1. It is important to note the following about this plot:

* The key idea is that deviations from the regression line in the left hand panel correspond to deviations from the horizontal line (i.e., Residuals = 0) in the right hand panel.
* The non-linear trend is still apparent in the right hand panel, but now the nonlinearity is with reference to Residuals = 0. Recall that the residuals should all be centered around this horizontal line if the population model is true (see Figure 8.1).
* The red line in the right hand panel is a locally weighted smoothed (“lowess”) regression line – it follows whatever trend is in the residuals without assuming the trend is linear.
* The overall interpretation of the residual vs fitted plot is as follows:
  + if the red line follows the horizontal line (Residuals = 0), we conclude that the the assumption of linearity is not problematic for the data.
  + If the red line deviates systematically from the horizontal line, this is evidence that the assumption of lineary is not met.

In this Figure 8.2, the assumption is clearly not met. In fact, this is so obvious that it doesn’t require looking at the residual versus fitted plot at all. So, why do we use the residual versus plot? The answer is, when we have predictors, the “raw” data are not so easy to interpret. In general, the assumption is much more easily checked using the residual versus fitted plot, even if the patterns are a bit harder to interpret.

This situation is illustrated with reference to Figure 8.3. Note that in this example the regression model has 2 predictors. We could produce a 3-D plot in which is represented by a plane, but the versus approach will not extend beyond 2 predictors. On the other hand, the residual versus fitted plot can be used with any number of predictors.

# Example: regression c1rmscal on ses\_orig and t1learn  
load("ECLS250.RData")  
mod2 <- lm(c1rmscal ~ ses\_orig + t1learn, data = ecls)  
plot(mod2, which = 1)

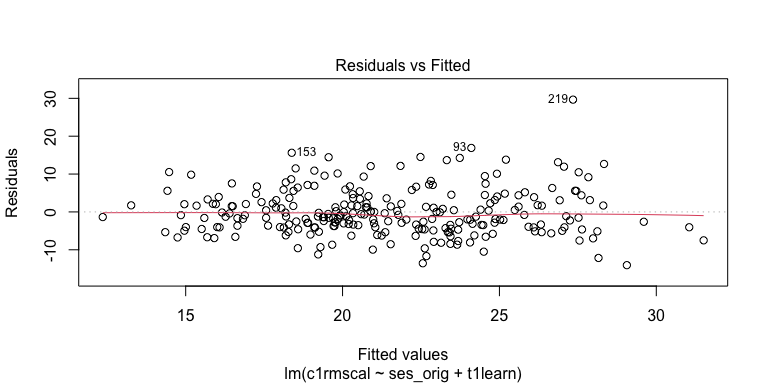


Figure 8.3: An example from ECLS.

**Please write down whether you think the linearity assumption is problematic for this second example, and be sure to explain why with reference to Figure 8.3.**

Before moving on, let’s take a look at a few more examples. In each of these examples, **please write down whether you think the linearity assumption is problematic and explain why with reference to the plots.**. Hint: be careful not to over-interpret the lowess line in the tails of the plots, where only a few data points can have a big impact on the local trend. Focus your interpretation on the bulk of the data, and whether it shows a systemic trend away from the horizontal line at 0.

# Example: regression c1rmscal on ses\_orig and t1learn  
set.seed(101)  
par(mfrow = c(2, 2))  
for(i in 1:4) {  
 x <- rnorm(200)  
 e <- rnorm(200)  
 y <- x + e   
 plot(lm(y ~ x), which = 1)  
}

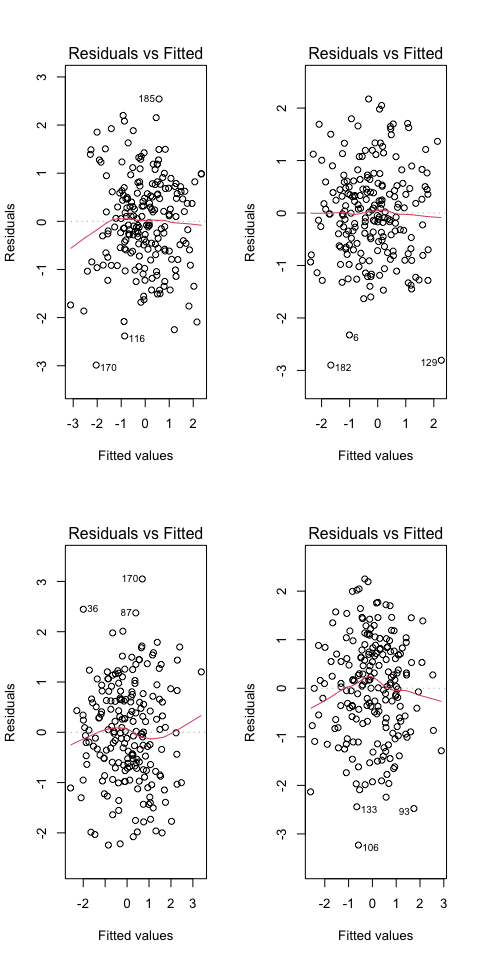


Figure 8.4: More examples.

### 8.2.1 Summary

To check the assumption of linearity, we can use a residual vs predicted (fitted) plot.

* If the plot does not show a systematic trend other than a horizontal line at Residuals = 0, then there is no evidence against the assumption.
* If the residuals do show a trend away from 0, then we should worry about the assumption.
* Don’t over interpret the tails of the lowess (red) lines in the R plots.
* If the assumption is violated: consider adding quadratic or other non-linear terms to the model (Chapter 9), or non-linear transformations of the variable (Chapter 10).

## 8.3 Homoskedasticity

This assumption means that the variance of the residuals should not change as a function of the predicted values. Because we are again concerned with residuals and predicted values, we can re-use the same plot we used to check linearity. However, we are no longer interested in whether the lowess trend (red line) systematically deviates from zero – we are interested in whether the range of the residuals (on the vertical axis) changes over the predicted values (on the horizontal axis).

Figure 8.5 illustrates two data sets in which the assumption of linearity is met, but the right hand panel shows evidence of heteroskedasticity. This is apparent by observing the range of the residuals over values of .

# homoskedastic example  
set.seed(1)  
x <- sort(rnorm(250))  
e <- rnorm(250)  
y <- x + e  
mod3 <- lm(y~x)  
par(mfrow = c(1, 2))  
plot(mod3, which = 1)  
  
# Heteroskedastic example  
y2 <- y  
y2[x > 0] <- x[x > 0] + 3\* e[x > 0]  
y2[x < -1] <- x[x < -1] + .3\* e[x < -1]  
mod4 <- lm(y2~x)  
plot(mod4, which = 1)

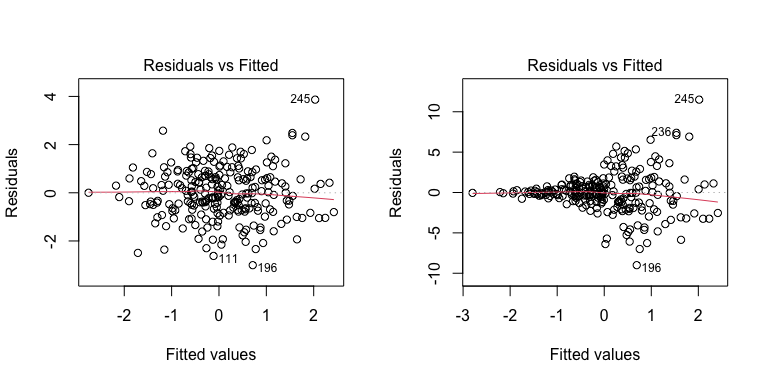


Figure 8.5: Illustration of homo- and heteroskedasicity.

To make it clearer what aspect of these plots is relevant for evaluating the assumption of homoskedasticity, the same figures are replicated below, but this time with blue lines represented my own “eye-balling” of the range of the residuals. In the left plot, the two lines are parallel, meaning the range is constant. In the right plot, the two lines form a cone, meaning the the range of the residuals increases for larger values of .

Note that I didn’t draw any lines in the tail ends of the plots – this is because there are fewer observations in the tails, so it is harder to make a judgment about the range of values. To avoid “reading the tea leaves” I focus on the range of values of for which there are sufficient observations to judge the range of the residuals.

To repeat, *the blue lines are just there for your reference*, to highlight the relevant information in the plot. You wouldn’t generally include these lines in the plot.

# homoskedastic example with ref lines  
par(mfrow = c(1, 2))  
plot(mod3, which = 1)  
segments(x0 = -1.5, y0 = 2, x1 = 1.5, y1 = 2, col = "#4B9CD3", lty = 2, lwd = 3)  
segments(x0 = -1.5, y0 = -2, x1 = 1.5, y1 = -2, col = "#4B9CD3", lty = 2, lwd = 3)  
  
# Heteroskedastic example  
plot(mod4, which = 1)  
segments(x0 = -1.5, y0 = 1, x1 = 1.5, y1 = 8, col = "#4B9CD3", lty = 2, lwd = 3)  
segments(x0 = -1.5, y0 = -1, x1 = 1.5, y1 = -8, col = "#4B9CD3", lty = 2, lwd = 3)

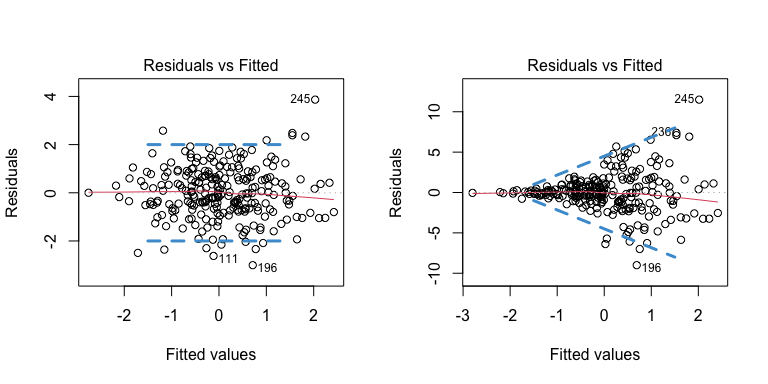


Figure 8.6: Illustration of Homo- and Heteroskedasicity, with Reference Lines

# remove y2 from memory to avoid naming conflicts later on  
rm(y2)

Let’s take another look at the plots in Figure 8.4. **Please write down whether you think the homoskedasticity assumption is problematic and explain why with reference to the plots**.

### 8.3.1 Summary

To check the assumption of homoskedasticity, we can use a residual versus fitted plot (again).

* If the range of values on the vertical axis appears approximately constant over the horizontal axis, then there is no evidence against the assumption.
* If the range does show a pattern (e.g., cone, bow-tie, ellipse), then we should worry about this assumption.
* If the assumption is violated, it will affect the standard errors of the regression coefficients as well as the resulting t-tests and confidence intervals. There are heteroskedasticity robust standard errors that can address this issue (Chapter 9).

## 8.4 Normality

There are many ways to check this normality of a set of data, but a widely applicable technique is a qq plot (short for quantile-quantile plot). A qq plot compares the quantiles (e.g., percentiles) of two different distributions.

For our assumption, we want to compare the quantiles of our standardized residuals to the quantiles of a standard normal distribution. Standardizing means the residuals should have variance equal to one, and, combined with the other population assumptions of linear regression, this implies that the residuals should have a standard normal distribution (see Section 8.1).

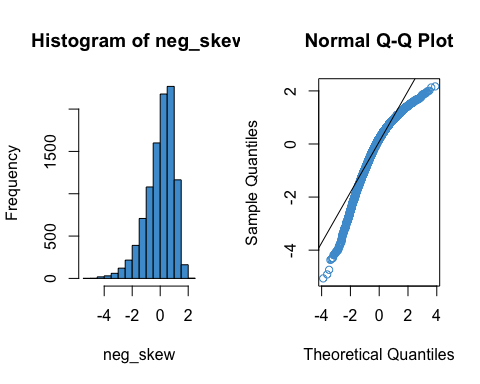
Since qq plots might not be something you have seen before, we’ll take a look at a few examples. Each figure below pairs a histogram and qq plot. In the qq plot, data points should fall on the diagonal line if the data are normally distributed. Focus on how the pattern in the histogram shows up as in deviations from the diagonal line the qq plot. We will discuss the interpretation of these patterns together in class. \*Please write down any questions you have about the interpretation of the qq plots and we can address them together in class.\*\*

Also, it should be emphasized that the line in the normal qqplot is *not a regression line!* It is just the diagonal line , and the bulk of the data points should fall on that line if the data were drawn from a normal distribution.

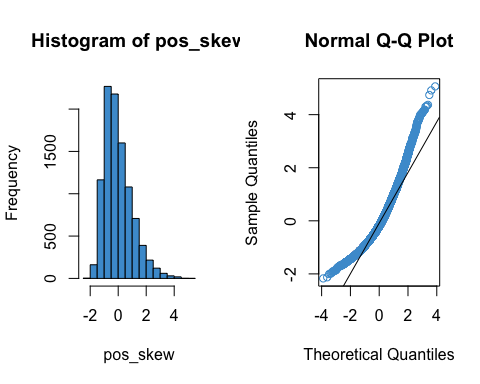
# Comparing histograms and q-q plots  
distributions <- read.csv("distributions.csv")  
names(distributions)[2] <- "normal"  
attach(distributions)  
dist\_names <- names(distributions)  
  
# Normal   
par(mfrow = c(1, 2))  
hist(normal, col = "#4B9CD3")  
qqnorm(normal, col = "#4B9CD3")  
qqline(normal)



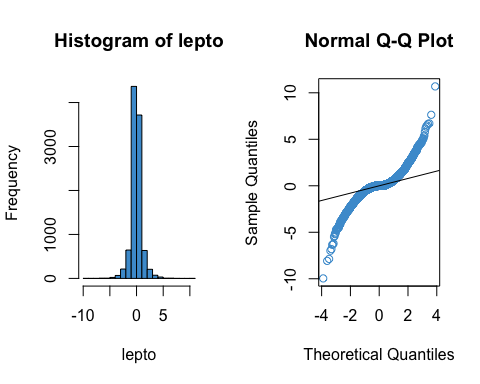
# Negative skew  
par(mfrow = c(1, 2))  
hist(neg\_skew, col = "#4B9CD3")  
qqnorm(neg\_skew, col = "#4B9CD3")  
qqline(neg\_skew)



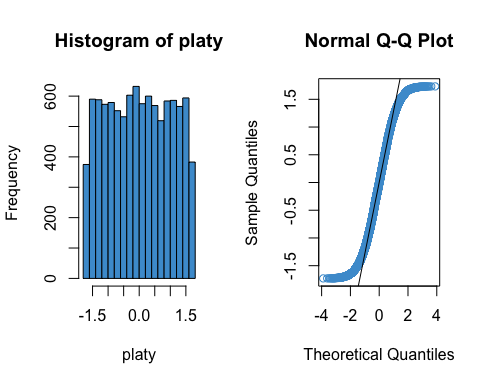
# Positive skew  
par(mfrow = c(1, 2))  
hist(pos\_skew, col = "#4B9CD3")  
qqnorm(pos\_skew, col = "#4B9CD3")  
qqline(pos\_skew)



# Leptokurtic  
par(mfrow = c(1, 2))  
hist(lepto, col = "#4B9CD3")  
qqnorm(lepto, col = "#4B9CD3")  
qqline(lepto)



# Platykurtic  
par(mfrow = c(1, 2))  
hist(platy, col = "#4B9CD3")  
qqnorm(platy, col = "#4B9CD3")  
qqline(platy)



Next, let’s consider a more realistic example using the default plotting from the lm function. **Please write down whether you think the normality assumption is problematic for the data in Figure @reff(fig:normality1), and be sure to explain why with reference to the plot.** Hint: if you think the data are non-normal, you should be able to interpret the pattern of deviations with reference to the examples given above (e.g. skew, kurtosis).

# Example: regression c1rmscal on ses\_orig and t1learn  
plot(mod2, which = 2)

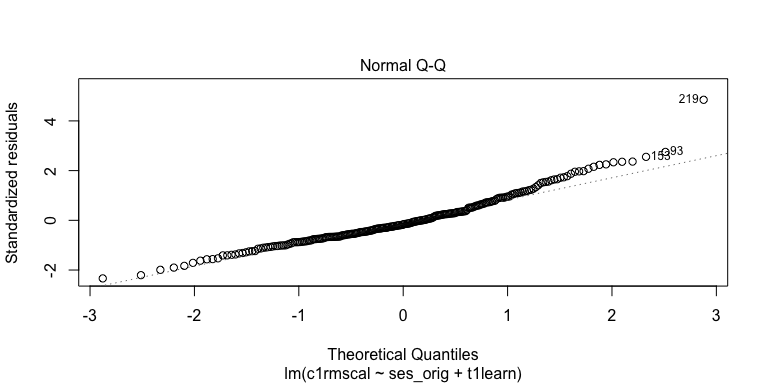


Figure 8.7: An example from ECLS.

Practice makes perfect, so lets work through a few more examples. In each of the examples in Figure 8.8, **please write down whether you think the normality assumption is problematic and explain why with reference to the plots.**

# Example: regression c1rmscal on ses\_orig and t1learn  
set.seed(101)  
par(mfrow = c(2, 2))  
for(i in 1:4) {  
 x <- rnorm(200)  
 e <- rnorm(200)  
 y <- x + e   
 plot(lm(y ~ x), which = 2)  
}

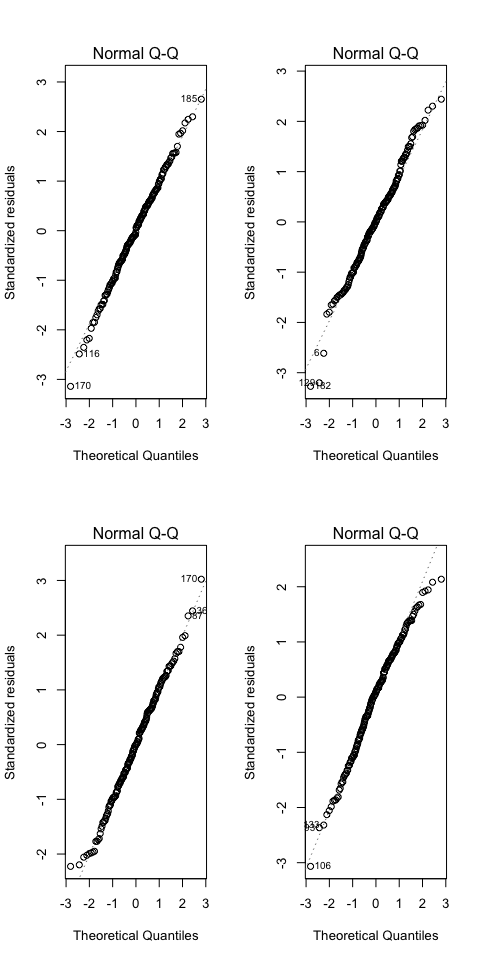


Figure 8.8: More examples.

### 8.4.1 Summary

To check the assumption of normality, we can use a qq plot of the standardized residuals against the standard normal distribution.

* If the points from the qq plot follows the line Y = X, then there is no evidence against the assumption
* If the residuals do show a trend off of the diagonal line, then we should worry about the assumption
* If the assumption is violated, the central limit theorem implies that significance tests used in OLS regression will be robust to violations of normality when sample sizes are large.
  + Practically this means N / K > 30 (but other guidelines are used too)
  + For some specific types of violations, notably positive skew, it is also common to transform the Y variable (Chapter 10)
* (The line in a normal qqplot is *not a regression line!*)

## 8.5 A worked example

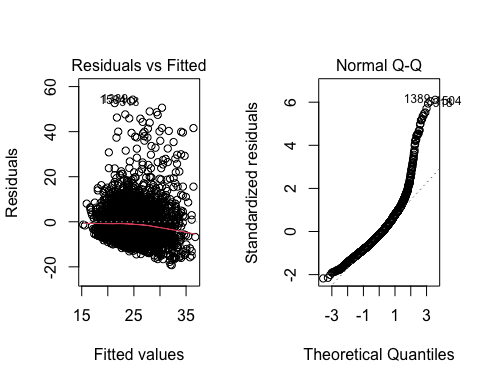
To illustrate these assumption checking procedures, let’s revisit the example from Section 7.3. In that example, it was noted that there was evidence that one (or more) of the assumptions were problematic. Let’s take a look at why this was the case.

For this example, we will use the ECLS data to regress reading achievement at the beginning of Kindergarten (c1rrscal) on SES (wksesl), parental (mother’s and father’s) education (wkmomed and wkdaded, respectively), and attendance in center-based care before K (p1center). This is model 3 from Section 7.3 (we don’t consider the model with the interactions).

The workflow for assumption checking requires first running the model and then producing the residual vs fitted plot and a qqplot of the residuals. If the plots look OK, we go ahead and interpret the model results. If the plots don’t look OK, we give up and wonder why we ever bothered with regression in the first place. Just kidding :) – the next two chapters of these notes address how to deal with violations of the assumptions.

After running the model, we obtain the following two plots from the regression output. **For each of the three population assumptions of linear regression, please write down whether you think the assumption is problematic and explain why with reference to the plots.**

# Run model  
load("ECLS2577.RData")  
mod5 <- lm(c1rrscal ~ factor(p1center) + wksesl + wkmomed + wkdaded, data = ecls)  
  
# Check assumptions  
par(mfrow = c(1,2))  
plot(mod5, 1)  
plot(mod5, 2)



To foreshadow the next couple of chapters, here is what the plots looked like after dealing with the problematic assumptions. The data still exhibit heteroskedasticity, but we can address that using heteroskedasticity consistent standard errors (Chapter 9).

# Run model  
attach(ecls)  
log\_c1rrscal <- log(c1rrscal - min(c1rrscal) + 1)  
wksesl\_sq <- wksesl^2  
mod6 <- lm(log\_c1rrscal ~ factor(p1center) + wksesl + wksesl\_sq + wkmomed + wkdaded)  
  
# Check assumptions  
par(mfrow = c(1,2))  
plot(mod6, 1)  
plot(mod6, 2)

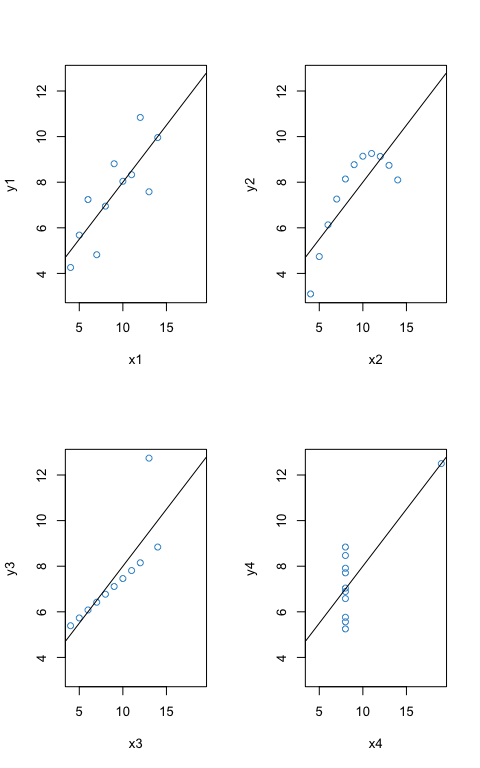


## 8.6 Diagnostics\*

This section is optional and is currently under construction. It focuses on walking through the code so the “hide code” button is currently omitted.

Like assumption checking, regression diagnostics also make extensive use of regression residuals, but this time the objective is to identify individual data points that are “outliers” with respect to the model. To work through the basic concepts of regression diagnostics, let’s again use the Anscombe’s quartet.

# Plotting Anscombe's quartet  
attach(anscombe)  
par(mfrow = c(2,2))  
ymax <- max(anscombe[,5:8])  
ymin <- min(anscombe[,5:8])  
xmax <- max(anscombe[,1:4])  
xmin <- min(anscombe[,1:4])  
  
  
plot(x1, y1, col = "#4B9CD3", xlim = c(xmin, xmax), ylim = c(ymin, ymax))  
abline(lm(y1 ~ x1))  
  
plot(x2, y2, col = "#4B9CD3", xlim = c(xmin, xmax), ylim = c(ymin, ymax))  
abline(lm(y2 ~ x2))  
  
plot(x3, y3, col = "#4B9CD3", xlim = c(xmin, xmax), ylim = c(ymin, ymax))  
abline(lm(y3 ~ x3))  
  
plot(x4, y4, col = "#4B9CD3", xlim = c(xmin, xmax), ylim = c(ymin, ymax))  
abline(lm(y4 ~ x4))



The interesting thing about these four examples is that they all have the same univariate and bivariate summary statistics – e.g., the same mean, variance, covariance, correlation, and regression coefficients. But first example is the only one that would be suitable for analysis by using these statistics. The second example shows a non-linear relationship, which we addressed in Section 8.2. In this section we will focus on the last two examples, since they have clear outliers.

### 8.6.1 Leverage

Leverage describes how “unusual” a data point is on the X variable(s) – i.e., how far from the mean it is on each predictor. The function hatvalues computes the leverage for each data point. Let’s check out the leverage for the 3rd and 4th examples from Anscombe’s quartet.

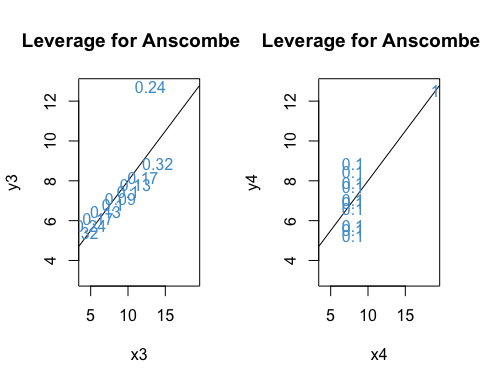
# leverage for Anscombe 3  
anscombe3 <- lm(y3 ~ x3)  
leverage3 <- hatvalues(anscombe3)  
  
# Take a look at the leverage values for each data point  
leverage3

## 1 2 3 4 5 6 7 8 9 10   
## 0.10000 0.10000 0.23636 0.09091 0.12727 0.31818 0.17273 0.31818 0.17273 0.12727   
## 11   
## 0.23636

# Show the leverage values in the scatter plot using the function "text"  
par(mfrow = c(1,2))  
  
plot(x3, y3, col = "white", xlim = c(xmin, xmax), ylim = c(ymin, ymax), main = "Leverage for Anscombe 3")  
abline(lm(y3 ~ x3))  
  
text(y3 ~ x3, labels = round(leverage3, 2), col = "#4B9CD3")  
  
# leverage for Anscombe 4  
anscombe4 <- lm(y4 ~ x4)  
leverage4 <- hatvalues(anscombe4)  
  
# Take a look at the leverage values for each data point  
leverage4

## 1 2 3 4 5 6 7 8 9 10 11   
## 0.1 0.1 0.1 0.1 0.1 0.1 0.1 1.0 0.1 0.1 0.1

# Show the leverage values in the scatter plot using the function "text"  
plot(x4, y4, col = "white", xlim = c(xmin, xmax), ylim = c(ymin, ymax), main = "Leverage for Anscombe 4")  
abline(lm(y4 ~ x4))  
  
text(y4 ~ x4, labels = round(leverage4, 2), col = "#4B9CD3")



Recall from the lesson that

* h should be smaller (closer to 0) for values closer to the mean of X
* The maximum value of h is 1

Based on the plots, we can see that the largest leverage is for the outlier in Anscombe 4.

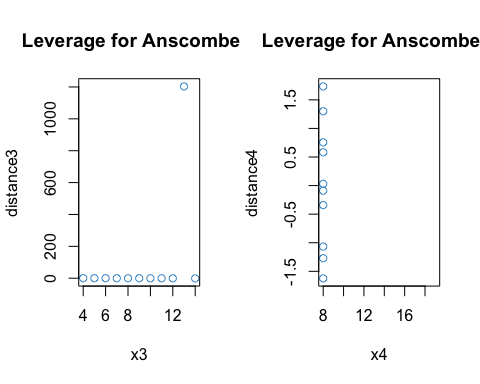
### 8.6.2 Distance (residuals)

Distance is about the size of the residuals. In order to judge the size of a residual, it helps to use the (externally) studentized residuals rather than the “raw” residuals. Because the studentized residual have a t-distribution on degrees of freedom, a rough ballpark for interpreting studentized residuals is that

* Values around +/- 2 are considered large.
* Values beyond +/- 3 are considered very large.

Let’s see what we have for our examples:

# Distance for Anscombe 3  
distance3 <- rstudent(anscombe3)  
  
par(mfrow = c(1,2))  
plot(x3, distance3, main = "Leverage for Anscombe 3", col = "#4B9CD3")  
  
# Distance for Anscombe 4  
distance4 <- rstudent(anscombe4)  
plot(x4, distance4, main = "Leverage for Anscombe 4", col = "#4B9CD3")



Clearly, the notion of distance is useful for describing what the problem is with Anscombe’s 3rd example. For the 4th example, the outlying data point is omitted because it has leverage of exactly 1, which means that the studentized residuals are undefined (divide by zero; R notes this in the console).

### 8.6.3 Influence

Influence describes how much the model results would change if a data point were omitted. Roughly, the conceptual relationships among influence, distance, and leverage are given by the following equation:

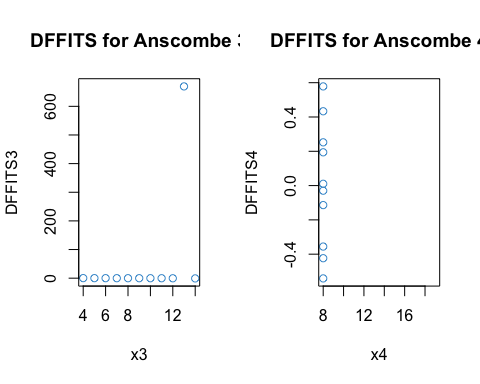
This equation tells us that, for a data point to have high influence, it must be a large distance from the regression line (have a large residual) *and* have high leverage (be far away from the mean on ).

There are a number of ways of computing influence. Like externally studentized residuals, they are all deletion statistics, or statistics computed using a “leave-one-out” approach.

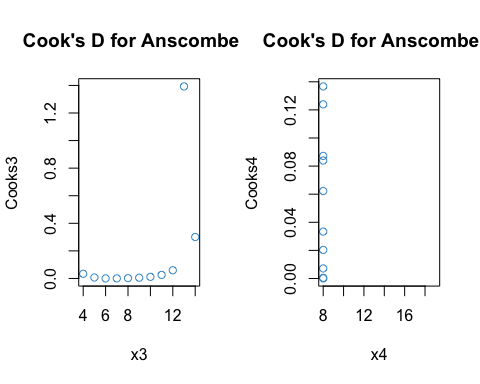
Influence statistics can also be classified into global versus local. Global approaches consider how a data point affects the predicted values. Local approaches consider how a data point affects the value of a specific regression coefficient.

Let’s start with DFFITS and Cook’s distance, two measures of global influence.

# DFFITS for Anscombe 3  
DFFITS3 <- dffits(anscombe3)  
  
par(mfrow = c(1,2))  
plot(x3, DFFITS3, main = "DFFITS for Anscombe 3", col = "#4B9CD3")  
  
# Distance for Anscombe 4  
DFFITS4 <- dffits(anscombe4)  
plot(x4, DFFITS4, main = "DFFITS for Anscombe 4", col = "#4B9CD3")



# Cook's distance for Anscombe 3  
Cooks3 <- cooks.distance(anscombe3)  
  
par(mfrow = c(1,2))  
plot(x3, Cooks3, main = "Cook's D for Anscombe 3", col = "#4B9CD3")  
  
# Distance for Anscombe 4  
Cooks4 <- cooks.distance(anscombe4)  
plot(x4, Cooks4, main = "Cook's D for Anscombe 4", col = "#4B9CD3")



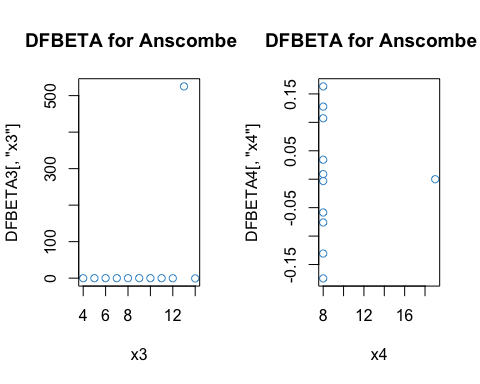
These statistics are similar but Cook’s distance is more interpetable. Values greater than 1 are considered indicative of high influence. We can see that the outlier in the 3rd example is highly influence. The outlier in the 4th example is “NA” because h = 1 hence there is a divide by zero problem.

For local measures of influence, the interpretation is roughly the same as global measures with simple regression (i.e., a single predictor). For multiple regression, local measures can provided additional insight to consider which regression coefficients are most influenced by an outlier.

# DFBETAs distance for Anscombe 3  
DFBETA3 <- dfbetas(anscombe3)  
  
#Take a look at the output: We get values for each coefficient, including the intercept  
DFBETA3

## (Intercept) x3  
## 1 -4.626e-03 -4.413e-02  
## 2 -3.713e-02 1.864e-02  
## 3 -3.579e+02 5.253e+02  
## 4 -3.290e-02 1.142e-18  
## 5 4.916e-02 -1.172e-01  
## 6 4.897e-01 -6.674e-01  
## 7 2.700e-02 -2.088e-02  
## 8 2.409e-01 -2.089e-01  
## 9 1.374e-01 -2.314e-01  
## 10 -1.970e-02 1.342e-02  
## 11 1.054e-01 -8.740e-02

# Plots for the regression coefficients  
par(mfrow = c(1,2))  
plot(x3, DFBETA3[,"x3"], main = "DFBETA for Anscombe 3", col = "#4B9CD3")  
  
# Distance for Anscombe 4  
DFBETA4 <- dfbetas(anscombe4)  
plot(x4, DFBETA4[,"x4"], main = "DFBETA for Anscombe 4", col = "#4B9CD3")



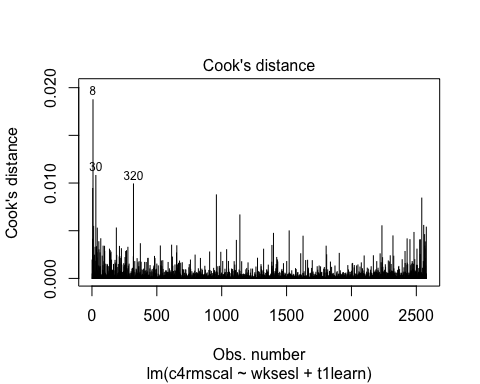
The output for DFBETA3 looks a lot like DFFITS3 (because we only have one predictor).

For DFBETA4, it is strange that our unusual data point (x4 = 18) was not identified as problematic – keep in mind that when this data point is removed, the variance of X is zero and so the regression coefficient for the leave-one-out model is not defined. Still, it is not clear why R reports the value as zero, rather than omitted.

### 8.6.4 A more realistic example

As a more realistic example, let’s consider Question 3 from Assignment 1 using the graphical output from lm. Note that the graphical output uses the internally studentized residuals, and refers to these as “standardized residuals.”

model <- lm(c4rmscal ~ wksesl + t1learn, data = ecls)  
  
# Influence via Cook's distance  
plot(model, which = 4)



We can see that, although R automatically labels the 3 data points with the highest values of Cook’s D, none of the data points are actually close to the cut off value of 1. In other words, none of the data points in this example have an undue influence on the results of the regression model.

## 8.7 Workbook

This section collects the questions asked in this chapter. We will discuss these questions in class. If you haven’t written down / thought about the answers to these questions before class, the lesson will not be very useful for you! So, please engage with each question by writing down one or more answers, asking clarifying questions, posing follow up questions, etc.

**Section 8.2**

* Please write down whether you think the linearity assumption is problematic for the example below, and be sure to explain why with reference to the figure.

plot(mod2, which = 1)

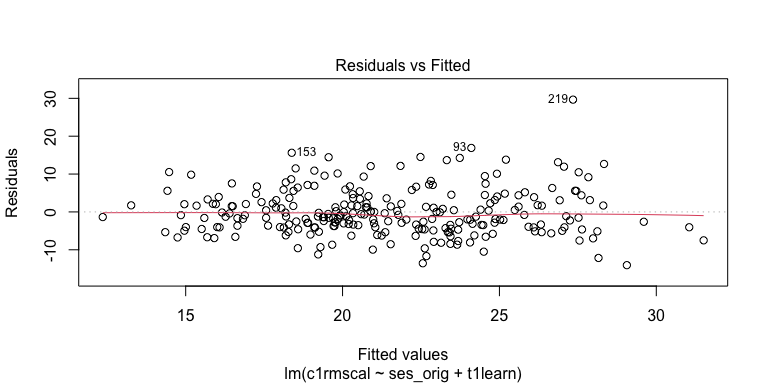


Figure 8.9: An example from ECLS.

* For each of the four examples below, please write down whether you think the linearity assumption is problematic and explain why with reference to the plots. Hint: be careful not to over-interpret the lowess line in the tails of the plots, where only a few data points can have a big impact on the local trend. Focus your interpretation on the bulk of the data, and whether it shows a systemic trend away from the horizontal line at 0.

# Example: regression c1rmscal on ses\_orig and t1learn  
set.seed(101)  
par(mfrow = c(2, 2))  
for(i in 1:4) {  
 x <- rnorm(200)  
 e <- rnorm(200)  
 y <- x + e   
 plot(lm(y ~ x), which = 1)  
}

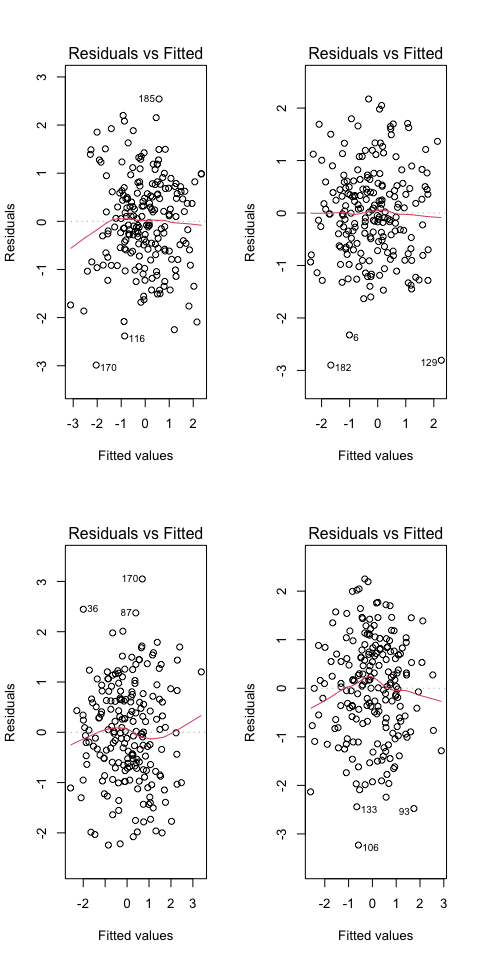


Figure 8.10: More examples.

**Section 8.3**

* Let’s take another look at the four plots in the above figure. For each plot, please write down whether you think the homoskedasticity assumption is problematic and explain why with reference to the plot.

**Section 8.4**

* Please write down whether you think the normality assumption is problematic for the data in the figure below, and be sure to explain why with reference to the plot. Hint: if you think the data are non-normal, you should be able to interpret the pattern of deviations (e.g. skew, kurtosis).

# Example: regression c1rmscal on ses\_orig and t1learn  
plot(mod2, which = 2)

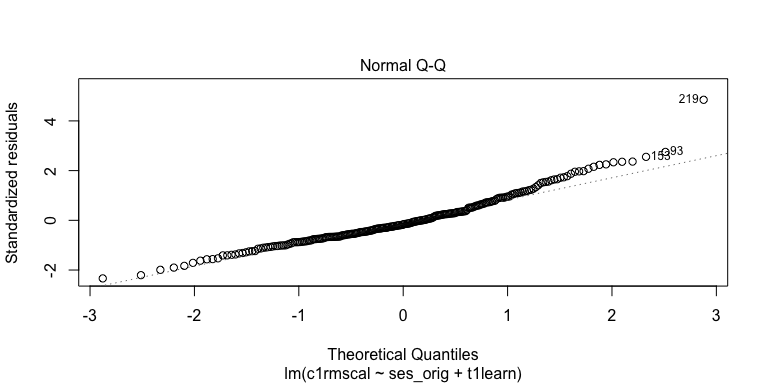


Figure 8.11: An example from ECLS.

* In each of the examples below, please write down whether you think the normality assumption is problematic and explain why with reference to the plots.

# Example: regression c1rmscal on ses\_orig and t1learn  
set.seed(101)  
par(mfrow = c(2, 2))  
for(i in 1:4) {  
 x <- rnorm(200)  
 e <- rnorm(200)  
 y <- x + e   
 plot(lm(y ~ x), which = 2)  
}

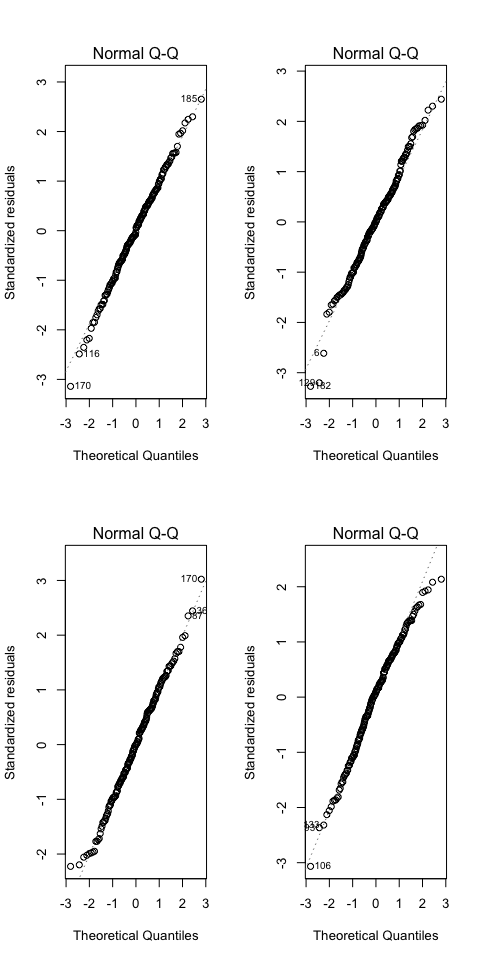
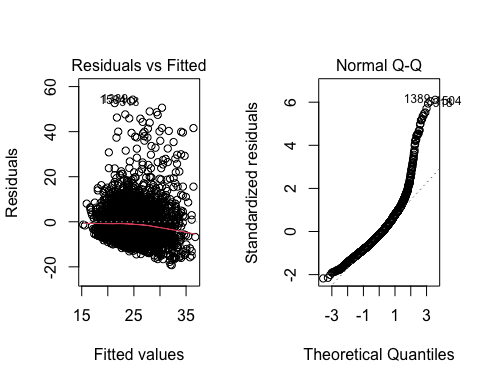


Figure 8.12: More examples.

**Section 8.5**

* For each of the three population assumptions of linear regression, please write down whether you think the assumption is problematic for the example shown below, and explain why with reference to the plots.

par(mfrow = c(1,2))  
plot(mod5, 1)  
plot(mod5, 2)



## 8.8 Exercises

There isn’t much new in terms of R code in this chapter. Once we run a model with lm, we just call the plot function on the lm output to produce the graphics requires for assumption checking. This section shows these steps for the worked example in Section 8.5. Please refer to that section for interpretation of the output.

# Run model  
load("ECLS2577.RData")  
mod5 <- lm(c1rrscal ~ factor(p1center) + wksesl + wkmomed + wkdaded, data = ecls)  
  
# Puts both plots in one figure  
par(mfrow = c(1,2))   
  
# Check assumptions  
plot(mod5, 1)  
plot(mod5, 2)

