EDUC 784: Regression

Peter Halpin

2022-01-26

# 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.

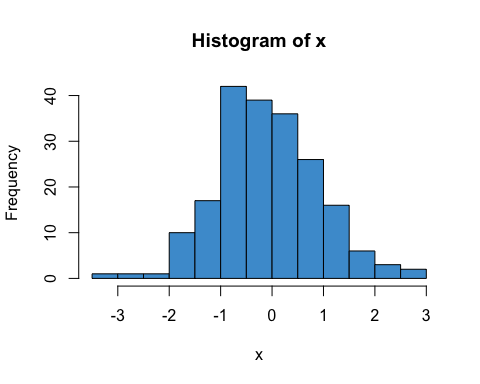
* To integrate course content (slides, readings, code, examples, and exercises) into one format, rather than having multiple files to sort through on Sakai.
* To address the perennial problem of choosing a textbook for this course – rather than having a required text, the goal is for this ebook to become 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)

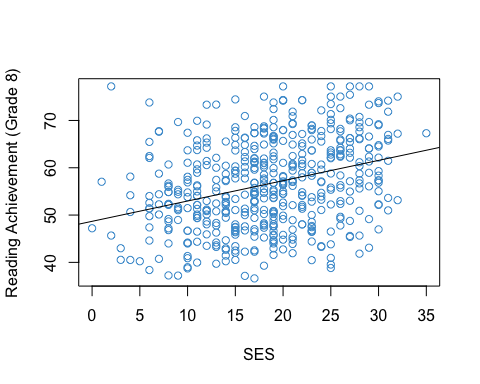


Figure 2.1: Reading Achievement and SES (NELS88).

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 = 2)  
cor(achmat08, ses)

## [1] 0.32

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)

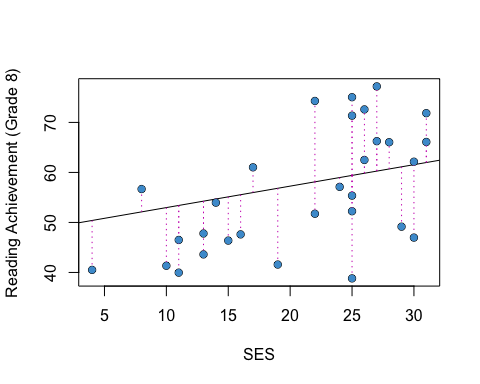


Figure 2.2: Residuals for a Subsample of the Example.

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   
## 48.68 0.43

**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.

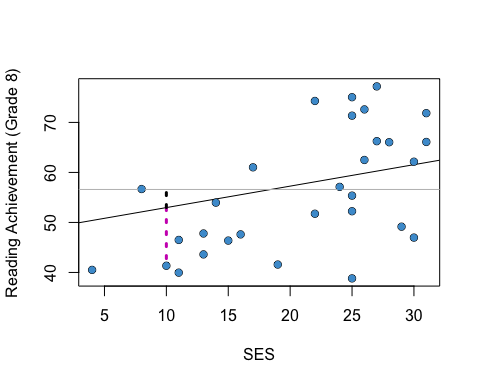


Figure 2.3: The Idea Behind R-squared.

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 .

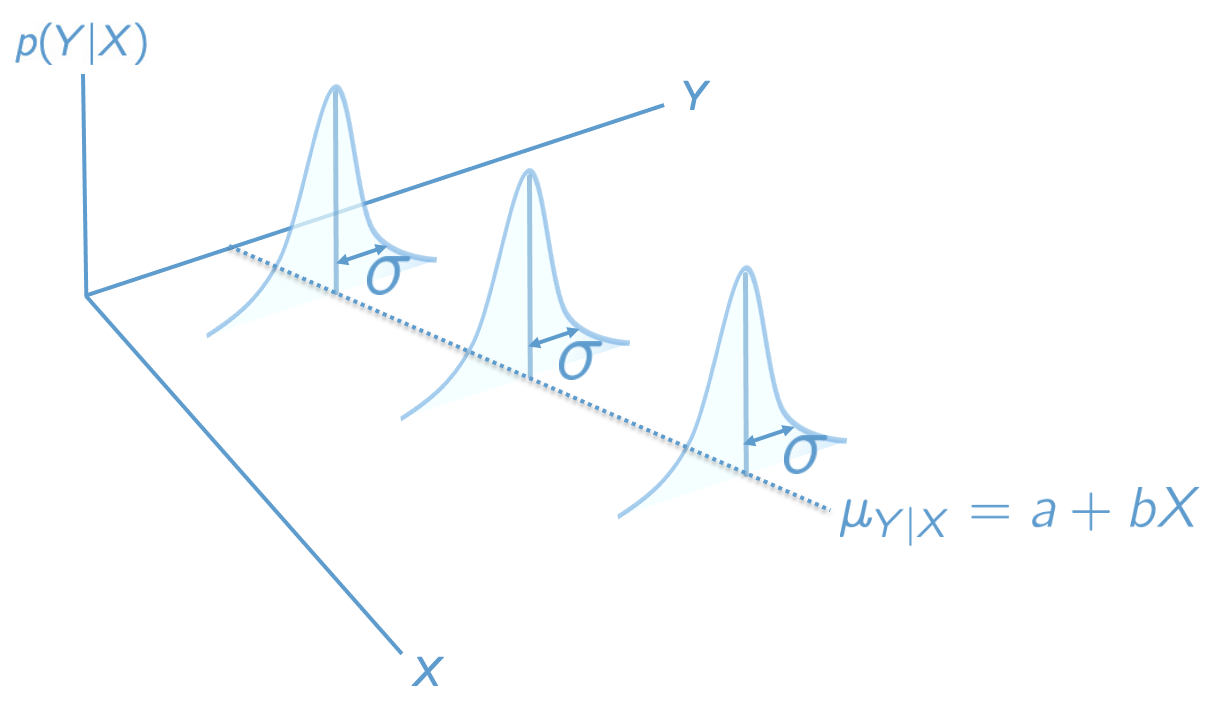


Figure 2.4: The Regression Population Model.

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 2.5 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).

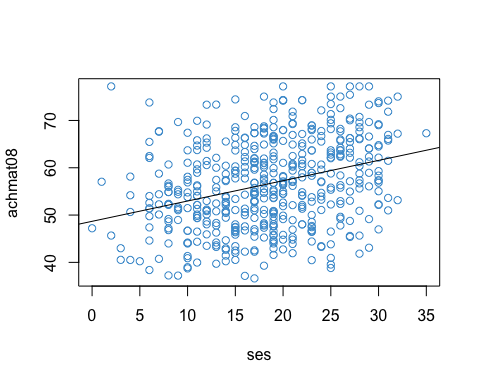
Let’s take a closer look using the following two variables from the NELS data, which is available on Sakai site for the course.

* 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

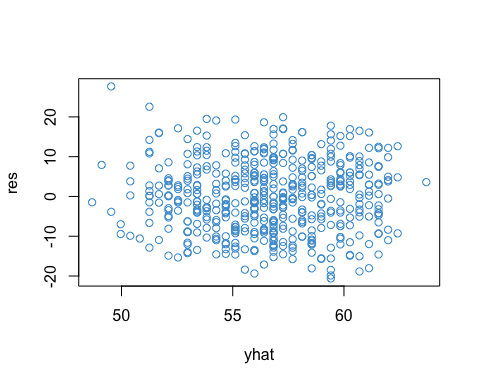
Let’s also check our interpretation of the parameters. If the answers to these questions are not clear, please make sure to ask in class!

* 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()

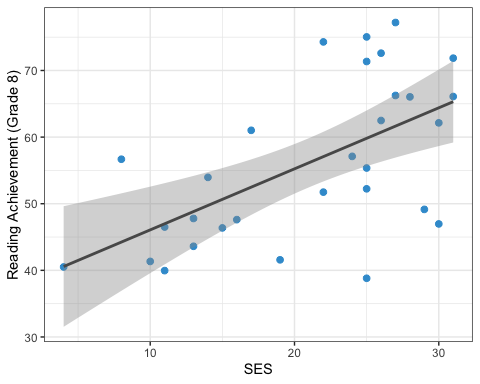


Figure 3.1: Prediction Error for Example Data.

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")

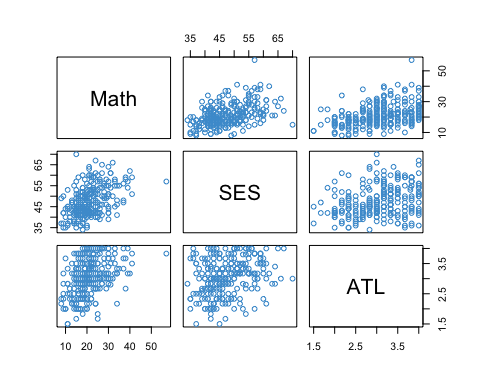


Figure 4.1: ECLS Example Data.

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

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). What do you think about these relationships in the context of the current debate about funding universal pre-K in the United States?

## 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:

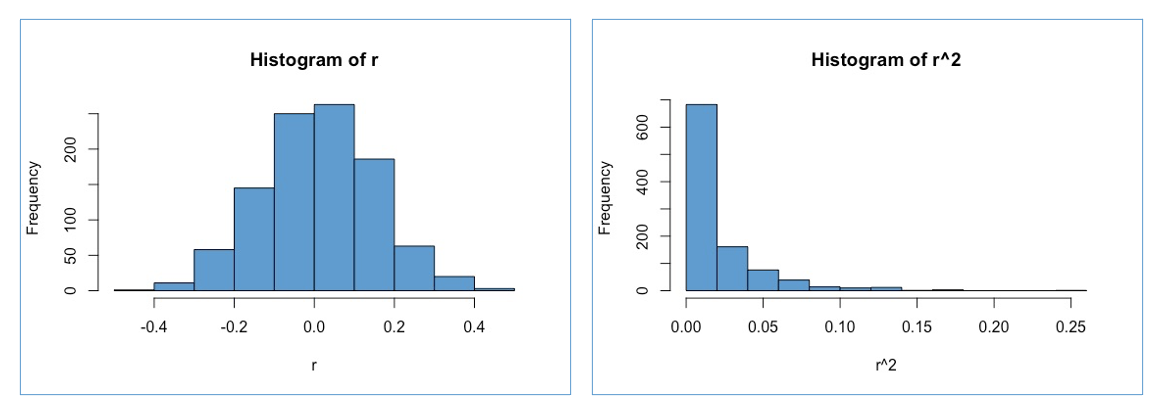
If the correlation between the predictors is zero, then we have the simplified formula

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 ??.

## 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.

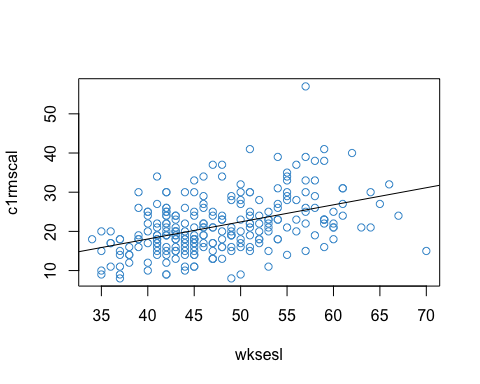
* Variable names begin with c, p, or t depending on whether the respondent was the child, parent, or teacher. Variables that start with wk were created by the ECLS using other data sources available in during the kindergarten year of the study.
* The time points (1-4 denoting fall and spring of K and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

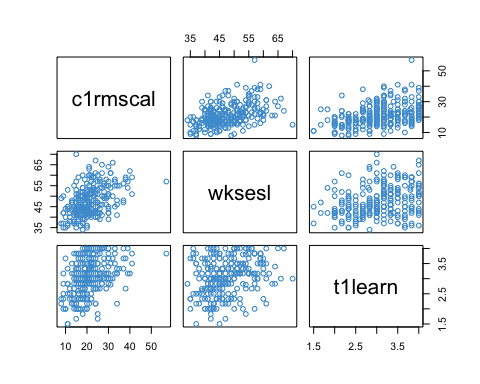
### 4.11.2 Multiple regression with lm

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
* ELL status
* ILP 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 coefficients in models with categorical predictors. we will apply 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 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)

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

# 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: 8.8 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: 8.8 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

## 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.

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 we are 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")

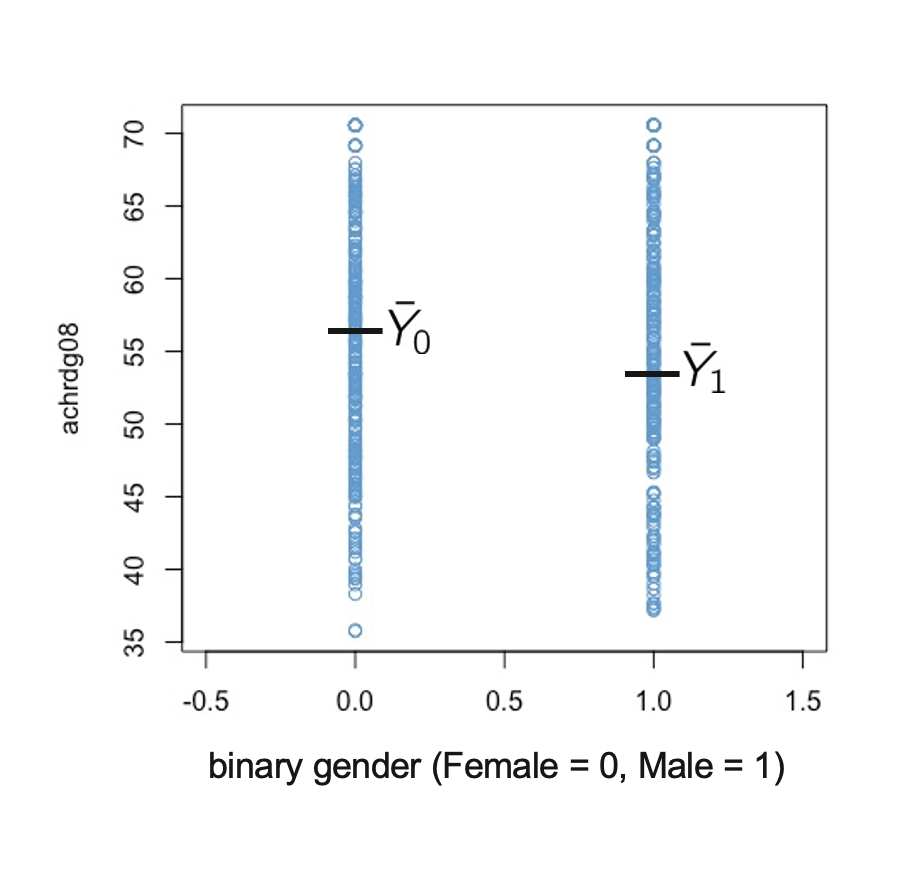


Figure 5.1: Reading Achievement on Binary Gender.

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:

# 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   
## 56.47 -0.92

**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: 8.8 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):

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

##   
## Two Sample t-test  
##   
## data: achrdg08 by binary\_gender  
## t = 1, df = 498, p-value = 0.2  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.64 2.48  
## sample estimates:  
## mean in group 0 mean in group 1   
## 56 56

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")

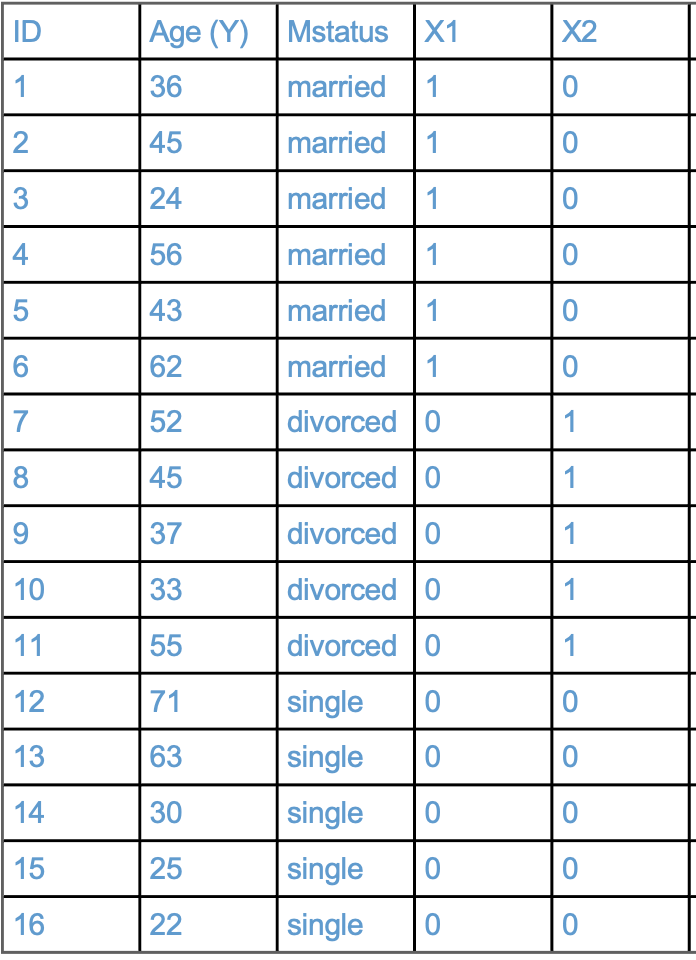


Figure 5.2: Toy Martital Status Example.

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

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

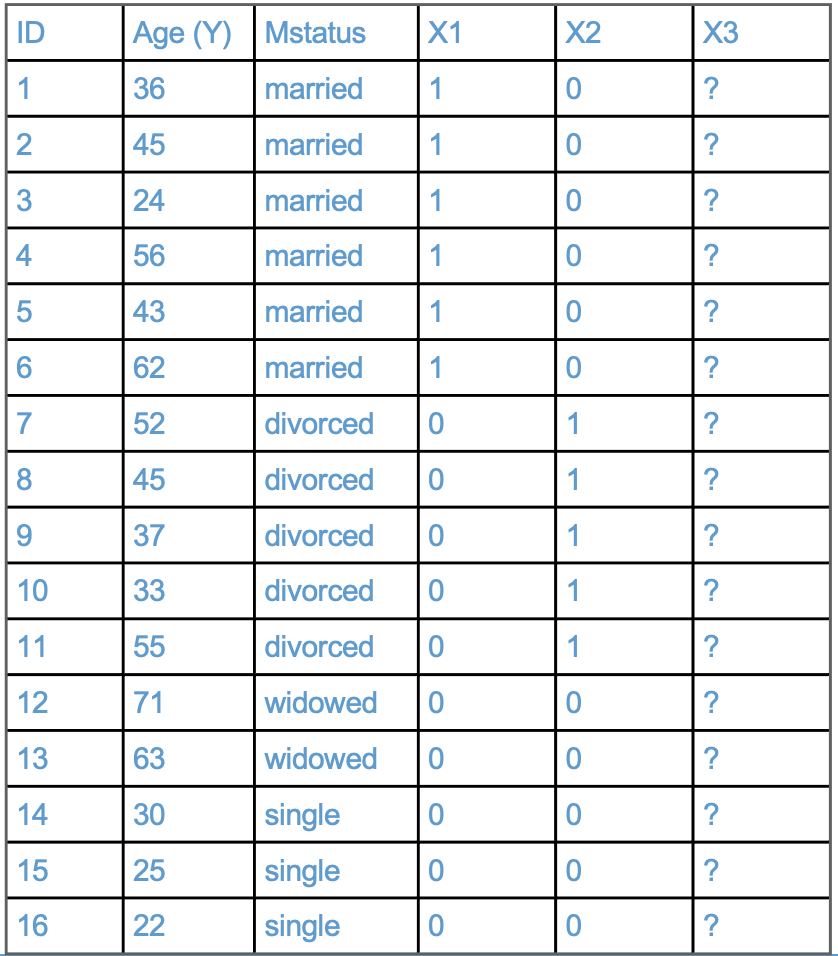


Figure 5.3: Toy Martital Status Example, Part 2.

**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, :

So, when the groups have equal sample sizes, the intercept is equal to the overall mean , and the regression coefficients are 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")

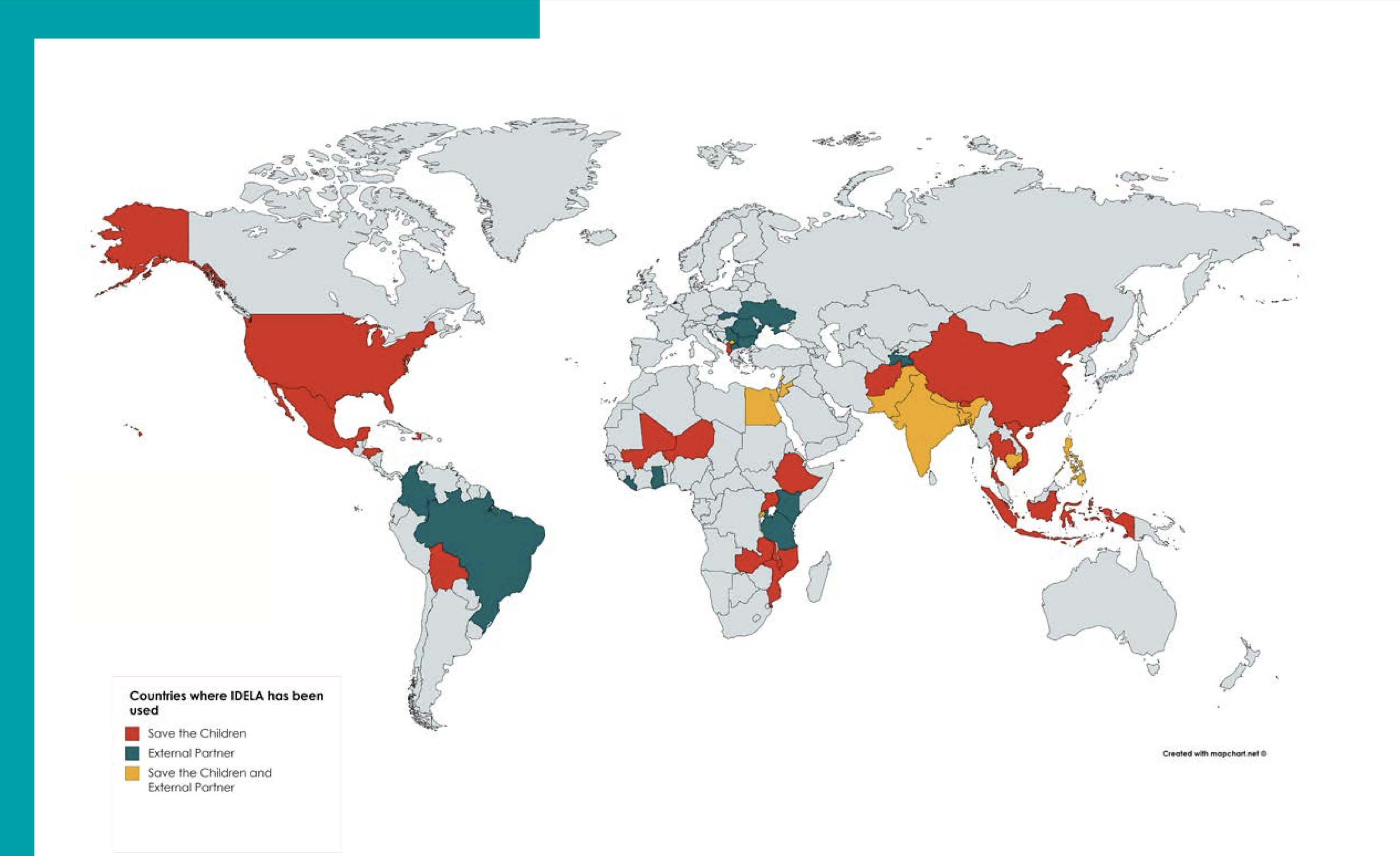


Figure 5.4: IDELA Worldwide Usage, 2017.

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 along the lines of Section ??. 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")

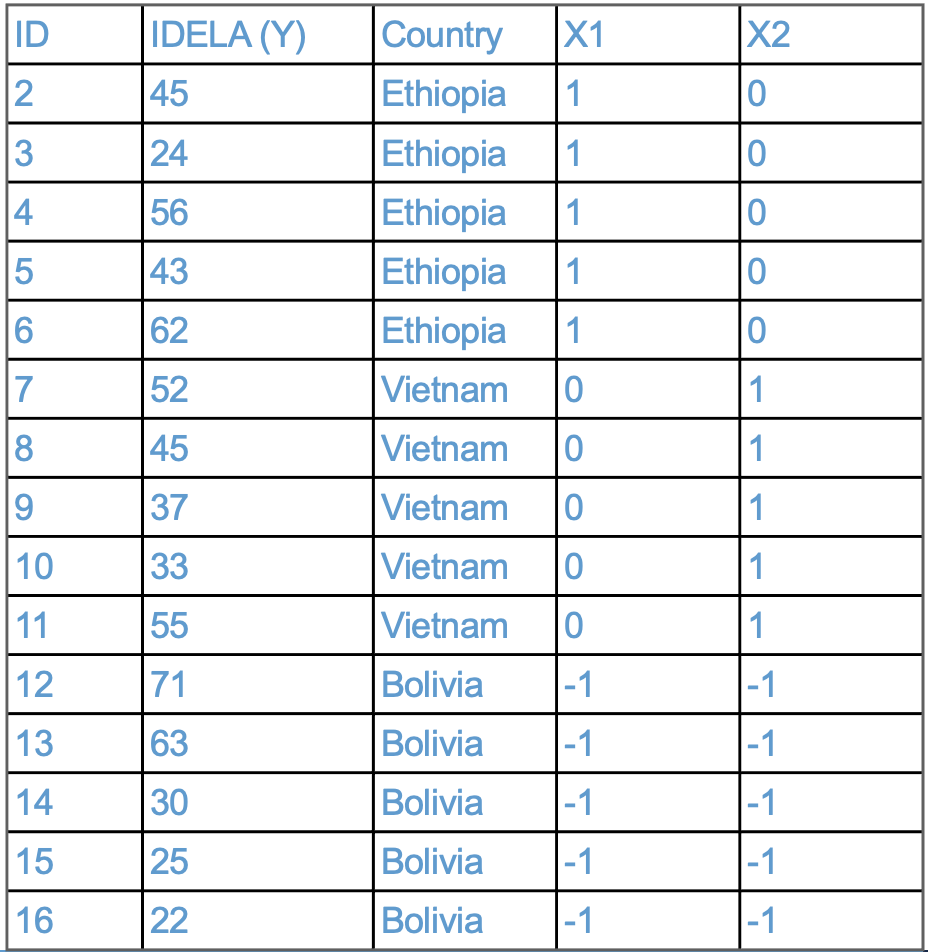


Figure 5.5: Toy IDELA Example.

Note that the dummy variables are different than for the case of reference coding discussed Section ??. 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 analogous to the reference group – but rather than being the reference group, it 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 ??

* *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 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 of the group means, this relationship holds regardless of the sample sizes in 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 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 attempt to 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 related to the Figure. 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   
## 56.47 -0.92

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

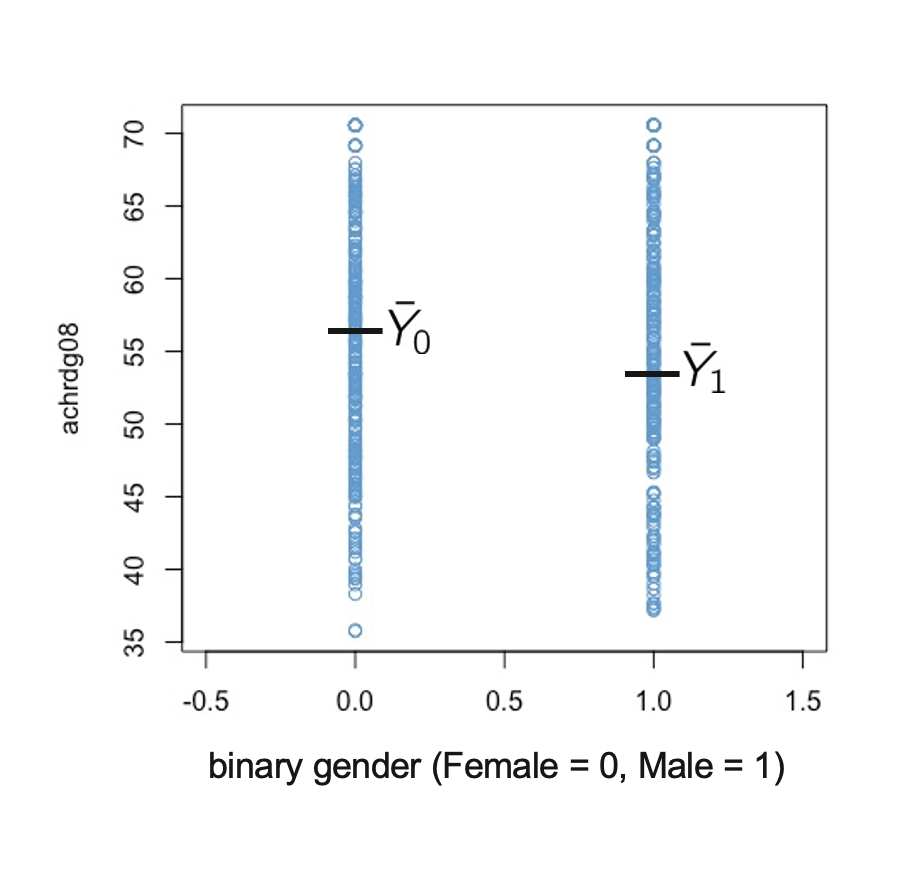


Figure 5.6: Reading Achievement on Binary Gender.

* If you have any questions about the relation between these two sets of 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: 8.8 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)

##   
## Two Sample t-test  
##   
## data: achrdg08 by binary\_gender  
## t = 1, df = 498, p-value = 0.2  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.64 2.48  
## sample estimates:  
## mean in group 0 mean in group 1   
## 56 56

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")

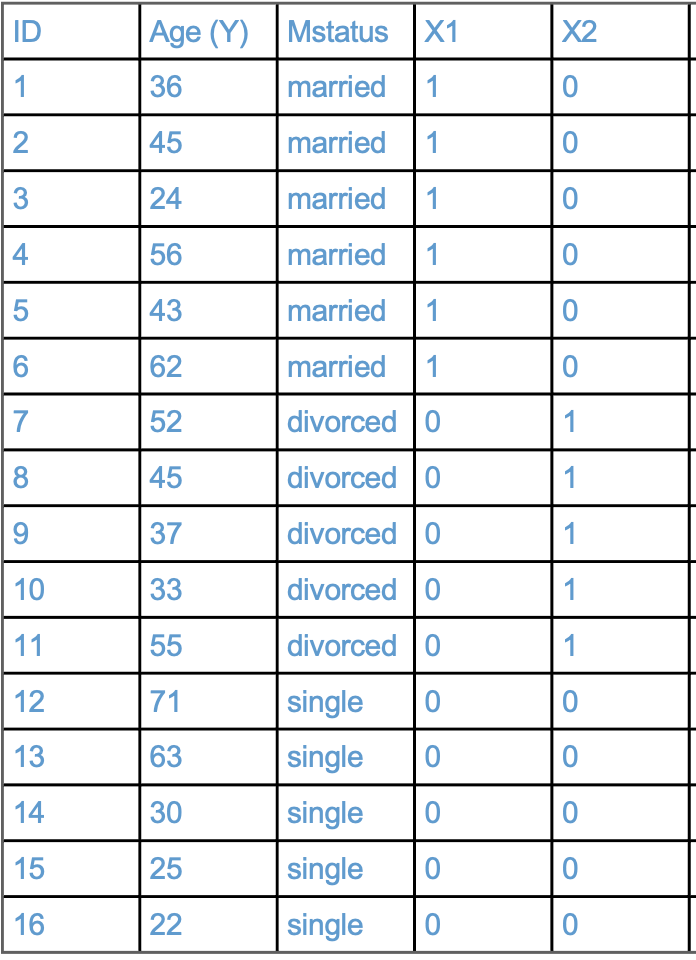


Figure 5.7: Toy Martital Status Example.

* 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")

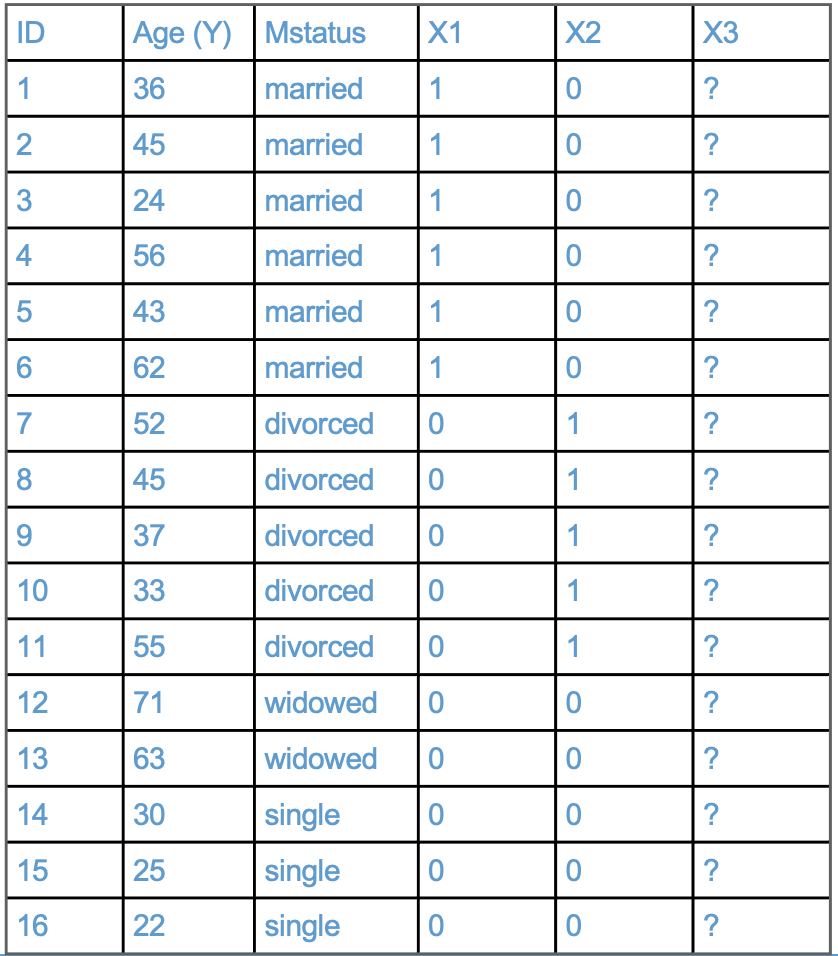


Figure 5.8: Toy Martital Status Example, Part 2.

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