EDUC 784

Peter Halpin

2024-01-03

Table of contents

# Preface

These are the course notes for EDUC 784. Readings are assigned *before class.* Sections denoted with an asterisk (\*) are optional.

The notes contain **questions that are written in bold font**. The questions are also collected in a section called “Workbook” that appears towards the end of each chapter. During class time, we will discuss the Workbook questions, your answers, any additional question you have, etc. It is really important for you to do the readings, and write down your responses to the questions, before class. You won’t get much out of the lessons if you haven’t done this preparation.

Some chapters contain a section called “Exercises” that collects all of the R code from that chapter into a single overall workflow. **You don’t need to do the Exercises before class**, but you can if you want to. If a chapter doesn’t have an Exercises section, that means we will be working on an assignment together instead.

# 1. Review

This chapter is an exception to the overall format described in the Preface. It reviews some foundational material from EDUC 710 (Stat 1) that is useful for this course. There will be time to ask questions about the review material in the first class and second class, but there will not be time to review everything. So, if this review feels too short, you may also want to review your notes from EDUC 710.

Please read up to the Exercises in [Section 1.10](#sec-exercizes-1) before the first class. We will address any questions about this material in the first class and then begin the Exercises together.

## 1.1 Summation notation

Summation notation uses the symbol to stand-in for summation. For example, instead of writing

to represent the sum of the values of the variable in a sample of size , we can instead write:

The symbol means “add.” The symbol is called “Sigma” – it’s the capital Greek letter corresponding to the Latin letter “S”. The value is called the index, and is the starting value of the index and is the end value of the index. You can choose whatever start and end values you want to sum over. For example, if we just want to add the second and third values of , we write

When the start and end values are clear from context, we can use a shorthand notation that omits them. In the following, it is implicit that the sum is over all available values of (i.e., from to ):

## 1.2 Rules of summation

There are rules for manipulating summation notation that are useful for deriving results in statistics. These rules are things you learned about addition in grade school, but they are presented using summation notation. You don’t need to do mathematical proofs or derivations in this class, but you will occasionally see some derivations in these notes (mainly in the optional sections).

Here are the rules:

**Rule 1: Sum of a constant (multiplication)**. Summing the values of a constant is the same as multiplication. Specifically, if you add a constant to itself times, this just times the constant:

**Rule 2: Distributive property**. The sum of a variable times a constant is equal to the constant times the sum.

**Rule 3: Associative property**. It doesn’t matter what order we do addition in:

## 1.3 Sample statistics

Summation notation is useful for writing the formulas of statistics. The main statistics we use in the class are the mean, standard deviation, variance, covariance, and correlation. These are the building blocks for regression. Their symbols and formulas are presented below (using the shorthand summation notation). If you don’t remember their interpretation, you will need to go back to your Stat 1 notes.

* The mean
* The variance can be written as or sometimes using the symbol
* The standard deviation can be written or using the letter
* The covariance is a generalization of the variance to two variables, it describes how they co-vary:
* The correlation is the covariance divided by the product of the standard deviations of the variables. It takes on values between -1 and 1 and describes the strength and direction of the linear relationship between two variables.

For numerical examples see [Section 1.10.8](#sec-computing-stats-1).

## 1.4 Properties of sample statistics

The following are some useful properties of the sample statistics reviewed above. The properties tell us what happens to means, variances, covariances, and correlations when a variable is linearly transformed. We often linearly transform data (e.g., to compute percentages, proportions, z-scores, and in linear regression), so these rules turn out to be quite handy.

You can derive the properties using the rules of summation. For each property, the beginning of the derivation is shown. You should know the properties but completing the derivations is optional.

**Sum of deviations from the mean**. If we subtract the mean from each data point, we have what is called a deviation (or deviation score): . Deviation scores sum to zero: .

* Derivation

**Mean of a linear transformation**. If with known constants and , then

* Derivation:

**Variance of a linear transformation**. If with known constants and , then

* Derivation

**Mean and variance of a z-score**. The z-score (or standardized score) is defined as . Standardized scores are useful because they have and .

* Derivation: use the rules for linear transformation with and .

**Covariance of linear transformation**. If and with known constants , , , , then

* Derivation

**Correlation of linear transformation**. If and with known constants , , , , then – i.e., the correlation is not affected by linear transformations.

* Derivation: use the rules for variances and covariances of linear transformation and the formula for correlation.

## 1.5 Bias and precision

In this section we consider two more important properties of sample statistics. These properties are defined in terms of the *sampling distribution* of a statistic. Recall that a sampling distribution arises from the following thought experiment:

1. Take a random sample of size from a population of interest.
2. Compute a statistic using the sample data. It can be any statistic, but let’s say the mean, , for concreteness.
3. Write down the value of the mean, and then return the sample to the population.

After doing these 3 steps many times, you will have many values the sample mean,

The distribution of these sample means is called the sampling distribution (of the mean).

A sampling distribution is just like any other distribution – so it has its own mean, and its own variance, etc. These statistics, when computed for a sampling distribution, have special names. We are especially interested in the following two statistics.

* **The expected value** of the mean, denoted , is the mean of the sampling distribution of the mean. That is a mouthful! That is why we say the “expected value” or “expectation” of a statistic rather than the mean of a statistic. It’s called the expected value because it’s the average value over many samples.
* **The standard error** of the mean, denoted , is the standard deviation of the sampling distribution of the mean. It describes the sample-to-sample variation of the mean around its expected value.

Now for the two additional properties of sample statistics:

* **Bias**: If the expected value of a statistic is equal to a population parameter, we say that the statistic is an unbiased estimate of that parameter. For example, the expected value of the sample mean is equal to the population mean (in symbols: , so we say that the sample mean is an unbiased estimate of the population mean.
* **Precision**: The inverse of the squared standard error (i.e., ) is called the precision of a statistic. So, the less a statistic varies from sample to sample, the more precise it is. That should hopefully make intuitive sense. The main thing to know about precision is that it is usually increasing in the sample size – i.e., we get more precise estimates by using larger samples. Again, this should feel intuitive.

Below is a figure that is often used to illustrate the ideas of bias and precision. The middle of the concentric circles represent the target parameter (like a bull’s eye) and the dots represent the sampling distribution of a statistic. You should be able to describe each panel in terms of the bias and precision of the statistic.

|  |
| --- |
| Figure 1.1: Bias and Precision |

## 1.6 t-tests

The -test is used to make an inference about the value of an unknown population parameter. The test compares the value of an unbiased estimate of the parameter to a hypothesized value of the parameter. The conceptual formula for a -test is

When we conduct a -test, the basic rationale is as follows: “if the estimate is close the hypothesized value of the parameter, then the numerator should be small relative to the standard error, and so should be close to zero.”

Typically, the hypothesized value of the population parameter is equal to zero, in which case it is called a null hypothesis. The null hypothesis usually translates into a research hypothesis of “no effect” or “no relationship.” So, if is close to zero, it means there was no effect.

In order to determine what values of are “close to zero”, we refer to its sampling distribution, which is called the -distribution. The -distribution tells what values of are typical if the null hypothesis is true. (More specifically: if the sample statistic is normally distributed and its expected value equal to the hypothesized value, then has a “central” -distribution.)

Some examples of the -distribution are shown below. The x-axis denotes values of the statistic shown above, and is a parameter called the “degrees of freedom” (more on this below). You can see that the -distribution looks like a normal distribution centered a zero. So, when the null hypothesis is true, the expected value of is zero. Informally, we could say that values greater than are pretty unlikely, and values greater than are very unlikely. Keep in the mind that these are the values of we are expecting if the null hypothesis is true.

|  |
| --- |
| Figure 1.2: t-distribution (source: https://en.wikipedia.org/wiki/Student%27s\_t-distribution) |

More formally, we can compare the value of computed in a sample, denoted as , to a “critical value”, denoted as . The critical value is chosen so that the “significance level”, defined as , is sufficiently small.

The significance level is chosen by the researcher. It represents our tolerance for false positives or Type I Errors – i.e., incorrectly rejecting the null hypothesis, or incorrectly concluding there is an effect when there isn’t one. When we set to a small number, we are saying that we want the probability of a false positive to be small. This means we are going to need strong evidence before we reject the null hypothesis – i.e., the value of would need to be very unlikely under the null hypothesis.

There are two equivalent ways of “formally” conducting a -test.

1. Compare the observed value of to the critical value. Specifically: if , reject the null hypothesis.
2. Compare the significance level chosen by the research, , to the “p-value” of the test, computed as . Specifically: if , reject the null hypothesis.

Informally, both of these just mean that the absolute value of should be pretty big (i.e., greater than ) before we reject the null hypothesis.

One last thing before moving on: the t-distribution has a single parameter called its “degrees of freedom”, which is denoted as in [Figure 1.2](#fig-tdist). The degrees of freedom are always an increasing function of the sample size, with larger samples leading to more degrees of freedom. When the degrees of freedom approach , the -distribution approaches a normal distribution. This means that the difference between a -test and a -test is pretty minor in large samples (say ).

However, when the degrees of freedom are small, the -distribution has wider tails than the normal distribution. This is also shown in [Figure 1.2](#fig-tdist). The tails of the distribution are important when doing statistical tests, because we are interested knowing about large / unlikely values of . So, in small samples (say ), it is important to use the -distribution.

## 1.7 Confidence intervals

A confidence interval uses the same equation as a -test, except we solve for the population parameter rather than the value of . Whereas a -test lets us make a guess about specific value of the parameter of interest (i.e., the null-hypothesized value), a confidence interval gives us a range of values that include the parameter of interest, with some degree of “confidence.”

Confidence intervals have the general formula:

We get the value of from the -distribution. In particular, if we want the interval to include the true population parameter of the time, then we choose to be the percentile of the -distribution. For example, if we set , we will have a confidence interval by choosing to be the -th percentile of the -distribution.

As mentioned, -tests and confidence intervals are closely related. In particular, if the confidence interval includes the value , this is the same as retaining the null hypothesis that the parameter is equal to . This should make intuitive sense. If the confidence interval includes , we are saying that it is a reasonable value of the population parameter, so we should not reject that value. This equivalence between tests and confidence intervals assumes we use the same level of for both of them.

In summary, if the confidence interval includes zero, we retain the null hypothesis at the stated level of . If the confidence interval does not include zero, we reject the null hypothesis at the stated level of .

## 1.8 F-tests

The -test is used to infer if two independent variances have the same expected value. This turns out to be useful when we analyze the variance of a variable into different sources (i.e., Analysis of Variance or ANOVA).

A variance can be defined as a sum-of-squares divided by its degrees of freedom. For example, the sample variance is just a sum-of-squared deviations from the sample mean (a sum of squares) divided by (its degrees of freedom).

The generic formula for an F-test is the ratio of two variances:

where denotes sums-of-squares and denotes degrees of freedom.

Just the like -test, the -test is called by the letter “F” because it has an -distribution when the null hypothesis is true (i.e., when the variances have the same expected value). The plot below shows some examples of - distributions. These distributions tell us the values of that are likely, if the null hypothesis is true.

|  |
| --- |
| Figure 1.3: F-distribution (source: https://en.wikipedia.org/wiki/F-distribution) |

The F distribution has two parameters, which are referred to as the “degrees of freedom in the numerator” and the “degrees of freedom in the denominator” (in the figure, d1 and d2, respectively). We always write the numerator first and then the denominator . So, the green line in the figure is “an -distribution on 10 and 1 degrees of freedom”, which means the in the numerator is 10 and the in the denominator is 1.

We use an -test the same way we use a -test – we set a significance level and use this level to determine how large the value of needs to be for us to reject the null hypothesis. The main difference is that is non-negative, because it is the ratio of squared numbers. We don’t usually compute confidence intervals for statistics with an -distribution.

## 1.9 APA reporting

It is important to write up the results of statistical analyses in a way that other people will understand. For this reason, there are conventions about how to report statistical results. In this class, we will mainly use tables and figures (formatted in R) rather than inline text. But sometimes reporting statistics using inline text unavoidable, in which case this course will use APA formatting. You don’t need to use APA in this class, but you should be familiar with some kind of conventions for reporting statistical results in your academic writing.

The examples below illustrate APA conventions. We haven’t covered the examples, they are just illustrative of the formatting (spacing, italics, number of decimal places, whether or not to use a leading zero before a decimal, etc). More details are available online (for example, [here](https://psych.uw.edu/storage/writing_center/stats.pdf)).

* Jointly, the two predictors explained about 22% of the variation in Academic Achievement, which was statistically significant at the .05 level ().
* After controlling for SES, a one unit of increase in Maternal Education was associated with units of increase in Academic Achievement ().
* After controlling for Maternal Education, a one unit of increase in SES was associated with units of increase in Academic Achievement. This was a statistically significant relationship ().

## 1.10 Exercises

This section will walk through some basics of programming with R. We will get started with this part of the review in the first class. You don’t need to do it before class.

If you are already familiar with R, please skim through the content and work on getting the NELS data loaded. If you are not familiar with R, or would like to brush up your R skills, you should work through this section.

### 1.10.1 General info about R

Some things to know about R before getting started:

* R is case sensitive. It matters if you useCAPS or lowercase in your code.
* Each new R command should begin on its own line.
* Unlike many other programming languages, R commands do **not** need to end with punctuation (e.g., ; or .).
* R uses the hash tag symbol (#) for comments. Comments are ignored by R but can be helpful for yourself and others to understand what your code does. An example is below.

# This is a comment. R doesn't read it.  
# Below is a code snippet. R will read it and return the result.   
2 + 2

[1] 4

* R’s working memory is cumulative. This means that you have to run code in order, one line after the next. It also means that any code you run is still hanging around in R’s memory until you clear it away using rm or the brush icon in R Studio - make sure to ask about how to do this in class if you aren’t sure.

### 1.10.2 The basics

As we have just seen, R can do basic math like a calculator. Some more examples are presented in the code snippets below. R’s main math symbols are

* + addition
* - subtraction or negative numbers
* \* multiplication
* / division (don’t use \)
* ^ or \*\* exponentiation

2 \* 2

[1] 4

# Remember pedmas? Make sure to use parentheses "()",   
# not brackets "[]" or braces "{}"  
(2 - 3) \* 4 / 5

[1] -0.8

# Exponentiation can be done two ways  
2^3

[1] 8

2\*\*3

[1] 8

# Square roots are "squirt". Again, make sure to use "()",   
# not brackets "[]" or braces "{}"  
sqrt(25)

[1] 5

# Logs and exponents, base e (2.718282....) by default   
log(100)

[1] 4.60517

exp(1)

[1] 2.718282

# We can override the default log by using the "base" option  
log(100, base = 2)

[1] 6.643856

# Special numbers...  
pi

[1] 3.141593

### 1.10.3 The help function

The help function is your best friend when using R. If we want more info on how to use an R function (like log), type:

help(log)

If you don’t exactly remember the name of the function, using ??log will open a more complete menu of options.

### 1.10.4 Logicals and strings

R can also work with logical symbols that evaluate to TRUE or FALSE. R’s main logical symbols are

* == is equal to
* != is not equal to
* > greater than
* < less than
* >= greater than or equal to
* <= less than or equal to

Here are some examples:

2 + 2 == 4

[1] TRUE

2 + 2 == 5

[1] FALSE

2 + 3 > 5

[1] FALSE

2 + 3 >= 5

[1] TRUE

The main thing to note is that the logical operators return TRUE or FALSE as their output, not numbers. There are also symbols for logical conjunction (&) and disjunction (|), but we won’t get to those until later.

In addition to numbers and logicals, R can work with text (also called “strings”). We wont use strings a lot but they are worth knowing about.

"This is a string in R. The quotation marks tell R the input is text."

[1] "This is a string in R. The quotation marks tell R the input is text."

### 1.10.5 Assignment (naming)

Often we want to save the result of a calculation so that we can use it later on. In R, this means we need to assign the result a name. Once we assign the result a name, we can use that name to refer to the result, without having to re-do the calculation that produced the result. For example:

x <- 2 + 2

Now we have given the result of 2 + 2 the name “x” using the assignment operator, <-.

Note that R no longer prints the result of the calculation to the console. If we want to see the result, we can type x

# To see the result a name refers to, just type the name  
x

[1] 4

We can also do assignment with the = operator.

y = 2 + 2  
y

[1] 4

It’s important to note that the = operator also gets used in other ways (e.g., to override default values in functions like log). Also, the math interpretation of “=” doesn’t really capture what is happening with assignment in computer code. In the above code, we are not saying that “2 + 2 equals y.” Instead, we are saying, “2 + 2 equals 4 and I want to refer to 4 later with the name ‘y’.”

Almost anything in R can be given a name and thereby saved in memory for later use. Assignment will become a lot more important when we name things like datasets, so that we can use the data for other things later on.

A few other side notes:

* Names cannot include spaces or start with numbers. If you want separate words in a name, consider using a period ., an underscore \_, or CamelCaseNotation.
* You can’t use the same name twice. If you use a name, and then later on re-assign that same name to a different result, the name will now only represent the new result. The old result will no longer have a name, it will be lost in the computer’s memory and will be cleaned up by R’s garbage collector. Because R’s memory is cumulative, it’s important to keep track of names to make sure you know what’s what.
* R has some names that are reserved for built-in stuff, like log and exp and pi. You can override those names, but R will give a warning. If you override the name, this means you can’t use the built-in functions until you delete that name (e.g., rm(x)).

### 1.10.6 Pop-quiz

1. In words, describe what the following R commands do.
   * x <- 7
   * x = 7
   * x == 7
   * 7 -> x
   * 7 > x

* Answers: Check the commands in R.

### 1.10.7 Vectors

In research settings, we often want to work with multiple numbers at once (surprise!). R has many data types or “objects” for doing this, for example, vectors, matrices, arrays, data.frames, and lists. We will start by looking at vectors.

Here is an example vector, containing the sequence of integers from 15 to 25.

# A vector containing the sequence of integers from 15 to 25  
y <- 15:25  
y

[1] 15 16 17 18 19 20 21 22 23 24 25

When we work with a vector of numbers, sometimes we only want to access a subset of them. To access elements of a vector we use the square bracket notation []. Here are some examples of how to index a vector with R:

# Print the first element of the vector y  
# Note: use brackets "[]" not parens"()"  
y[1]

[1] 15

# The first 3 elements  
y[1:3]

[1] 15 16 17

# The last 5  
y[6:11]

[1] 20 21 22 23 24 25

We can also access elements of a vector that satisfy a given logical condition.

# Print the elements of the vector y that are greater than the value 22  
y[y > 22]

[1] 23 24 25

This trick often comes in handy so its worth understanding how it works. First let’s look again at what y is, and what the logical statement y > 22 evaluates to:

# This is the vector y  
y

[1] 15 16 17 18 19 20 21 22 23 24 25

# This is the logical expression y > 22  
y > 22

[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE TRUE TRUE

We can see that y > 22 evaluates to TRUE or FALSE depending on whether the correspond number in the vector y is greater than 22. When we use the logical vector as an index – R will then return all the values for which y > 22 is TRUE.

In general, we can index a vector y with any logical vector of the same length as y. The result will return only the values for which the logical vector is TRUE.

### 1.10.8 Computing sample stats

The following are examples of statistical operations you can do with vectors of numbers. These examples follow closely to [Section 1.1](#sec-summation-1) to [Section 1.4](#sec-properties-1)

# Making a vector with the "c" command (combine)   
x <- c(10, 9, 15, 15, 20, 17)  
  
# Find out how long a vector is (i.e., the sample size)  
length(x)

[1] 6

# Add up the elements of a vector  
sum(x)

[1] 86

# Add up the elements of a subset of a vector  
sum(x[2:3])

[1] 24

# Check the distributive rule  
sum(x\*2) == sum(x) \* 2

[1] TRUE

# Check the associative rule  
y <- c(5, 11, 11, 19, 13, 15)  
sum(x) + sum(y) == sum(x + y)

[1] TRUE

# Compute the mean  
mean(x)

[1] 14.33333

# Compute the variance  
var(x)

[1] 17.46667

# Compute the standard deviation  
sd(x)

[1] 4.179314

# Compute the covariance  
cov(x, y)

[1] 10.66667

# Compute the correlation  
cor(x, y)

[1] 0.5457986

### 1.10.9 Working with datasets

Most of the time, we will be reading-in data from an external source. The easiest way to do this is if the data is in the .RData file format. Then we can just double-click the .Rdata file and Rstudio will open the file, or we can use the load command in the console – both do the same thing.

To get started, let’s load the NELS data. The data are a subsample of the 1988 National Educational Longitudinal Survey (NELS; see <https://nces.ed.gov/surveys/nels88/>).

This data and codebook are available on the Canvas site of the course under “Files/Data/NELS” and are linked in the “Module” for Week 1. You need to download the data onto your machine and then open the data file (e.g., by clicking it, or double-clicking, or whatever you do to open files on your computer). That will do the same thing as the following line of code

#This is what happens when you double-click NELS.RData  
load("NELS.RData")

The function dim reports the number of rows (500 persons) and columns (48 variables) for the data set.

dim(NELS)

[1] 500 48

If you want to look at the data in a spreadsheet, use the following command. It won’t render anything in this book, but you can see what it does in R. (You may need to install XQuartz from https://www.xquartz.org if you are using a Mac.)

View(NELS)

If you want to edit the data set using the spreadsheet, use edit(NELS). However, R’s spreadsheet editor is pretty wimpy, so if you want to edit data in spreadsheet format, use Excel or something.

Working with data is often made easier by “attaching”” the dataset. When a dataset it attached, this means that we can refer to the columns of the dataset by their names.

# Attach the data set  
attach(NELS)  
  
# Print the first 10 values of the NELS gender variable  
gender[1:10]

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

**Warning about attaching datasets.** Once you attach a dataset, all of the column names in that dataset enter R’s working memory. If the column names in your dataset were already used, the old names are overwritten. If you attach the same dataset more than once in the same session, R will print a warning telling you that the previously named objects have been “masked” – this won’t affect your analyses, but it can be irritating.

The basic point: we should only attach each dataset once per R session. Once you are done using a data set it is good practice to detach it:

detach(NELS)

### 1.10.10 Preview of next week

[Figure 1.4](#fig-nels-1) shows the relationship between Grade 8 Math Achievement (percent correct on a math test) and Socioeconomic Status (SES; a composite measure on a scale from 0-35). Once you have reproduced this figure, you are ready to start the next chapter.

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

|  |
| --- |
| Figure 1.4: Math Achievement and SES (NELS88). |

# 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 @ref(fig:fig1) shows the relationship between Grade 8 Math Achievement (percent correct on a math 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 = "Math Achievement (Grade 8)", xlab = "SES")  
  
# Run the regression model  
mod <- lm(achmat08 ~ ses)  
  
# Add the regression line to the plot  
abline(mod)

|  |
| --- |
| Math Achievement and SES (NELS88). |

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

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

[1] 0.3182

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

This relationship between SES and academic achievement 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 this relationship.**

## 2.2 The regression line

The line in the Figure @ref(fig:fig1) can be represented mathematically as

where

* denotes Math Achievement
* denotes SES
* 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 Math Achievement in the data, whereas represents the values computed from the regression equation (i.e., the values on the regression line). The difference is called a *residual*. The residuals for a subset of the data points in Figure @ref(fig:fig1) are shown in pink in Figure @ref(fig:fig2)

# 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 = "Math 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)

|  |
| --- |
| Residuals for a Subsample of the Example. |

Notice that by definition. So, we can use either Equation @ref(eq:yhat) or the equation below to write out a regression model:

Both equations say the same thing, but Equation @ref(eq:y) lets us talk about the values of in the data, not just the predicted values.

Another way to write out the model is using the variable names (or abbreviations) in place of the more generic “X, Y” notation. For example,

This notation is more informative about the specific variables in the example. But it is also more clunky and doesn’t lend itself to other mathematical expressions. For example, is much clearer than – in general, we want most of the text on the “baseline”, not the subscripts or superscripts.

You should be familiar with all 3 ways of presenting regression equations and you are free to use whatever approach you like best in your own writing.

## 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 in the second half of the course.

Solving the minimization problem (setting derivatives to zero) 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 @ref(review-2) for a refresher.)

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

coef(mod)

(Intercept) ses   
 48.6780 0.4293

**Please write down an interpretation of these numbers in terms of the line in Figure** @ref(fig:fig1)**, 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

(You can check these results by plugging the value 0 for the means and the value 1 for the variance in the equations above.)

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 – its basically just the same thing as correlation. 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 Math 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 Math Achievement that is associated with SES – i.e., to what extent are individual differences in Math Achievement associated with, or explained by, individual differences in SES?

This question is represented graphically in Figure @ref(fig:rsquared). The horizontal line denotes the mean of Math Achievement. The difference between the indicated student’s Math 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 Math Achievement score () by considering the predicted values () instead of the overall mean (). This represents the extent to which this student’s Math Achievement is explained by the linear relationship with SES.
* The pink dashed line is the regression residual, which was introduced in Section @ref(regression-line-2). This is the variation in Math Achievement that is “left over” after considering the linear relationship with SES.

|  |
| --- |
| The Idea Behind R-squared. |

The R-squared statistic averages the variation in Math Achievement associated with SES (i.e., the black dashed line) relative to the total variation in Math Achievement (i.e., black + pink) for all students in the sample. R-squared is a 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 @ref(fig:fig1)), 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 @ref(review-2)):

The first part is just from Section @ref(ols-2). 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 @ref(chapter-8).

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 values of conditional on , denoted , are normally distributed (the figure shows these distributions for three 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 .

|  |
| --- |
| 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 subtracts 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, 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 @ref(regression-line-2) they denoted sample statistics.
* In Section @ref(population-model-2) 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 @ref(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 are often called the predicted values, is not usually referred to this way. It is called the conditional mean function or the conditional expectation function. We will introduce some other notations for and later in the course.

**Please be prepared for a pop quiz on notation during class!**

## 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 @ref(review-2) 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 , or 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, just by 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 ), you won’t be interested in this test (even though R always provides it).

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 @ref(chapter-5) and interactions in Chapter @ref(chapter-6). But, for now, we can put it on the back burner.

## 2.9 Inference for R-squared

Inference for R-squared is quite a bit different than for the regression parameters. As we saw in section @ref(rsquared-2), 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 @ref(inference-for-slope-2) 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 the output 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 the output 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

**What is the MDES for the NELS example? Please be prepared to share your answer in class.**

## 2.11 Workbook

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

**Section** @ref(example-2)

# Scatter plot  
plot(x = ses, y = achmat08, col = "#4B9CD3", ylab = "Math Achievement (Grade 8)", xlab = "SES")  
  
# Run the regression model  
mod <- lm(achmat08 ~ ses)  
  
# Add the regression line to the plot  
abline(mod)

|  |
| --- |
| Math 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 the plot above, the correlation is

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

[1] 0.3182

This correlation means that eighth graders from more well-off families (higher SES) also tended to do better in Math (higher Math Achievement).

This relationship between SES and academic achievement 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 this relationship.

**Section** @ref(ols-2)

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

coef(mod)

(Intercept) ses   
 48.6780 0.4293

Please write down an interpretation of these numbers in terms of the line in Figure @ref(fig:fig1), and be prepared to share your answers in class.

**Section** @ref(rsquared-2)

Using all of the cases from the example (Figure @ref(fig:fig1)), 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.

**Section** @ref(notation-2)

Please be prepared for a pop quiz on notation during class!

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

**Section** @ref(inference-for-slope-2)

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!

**Section** @ref(inference-for-rsquared-2)

Using the same output as above, please write down an interpretation of the F-test of R-squared and be prepared to share your answers in class. Note that the output uses the terminology “multiple R-squared” to refer to R-squared.

**Section** @ref(power-2)

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 the output $u $ 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

What is the MDES for the NELS example? Please be prepared to share your answer in class.

## 2.12 Exercises

These exercises collect all of the R input used in this chapter into a single step-by-step analysis. It explains how the R input works, and provides some additional exercises. We will go through this material in class together, so you don’t need to work on it before class (but you can if you want.)

### 2.12.1 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:

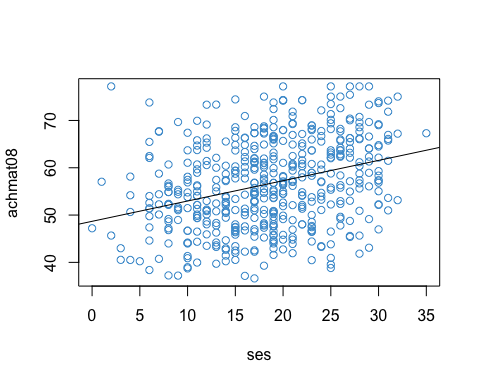
Y ~ X

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.

* 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)  
  
# Add the regression line to the plot  
abline(mod)



# Print out the regression coefficients  
coef(mod)

(Intercept) ses   
 48.67803 0.42926

Let’s do some quick calculations to check that the lm output corresponds the formulas for the slope and intercept in Section @ref(ols-2):

We won’t usually do these kind of “manual” calculations, but it is a good way consolidate knowledge presented in the readings with the output presented by R. It is also useful to refresh our memory about some useful R functions and how the R language works.

# Confirm that the slope from lm is equal to 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 does the predicted value of achmat08 increase for each unit of increase in ses?

### 2.12.2 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

The R-squared for the example is presented in the output below. You should be able to provide an interpretation of this number, so if it’s not clear make sure to ask in class!

# R-squared from the example  
summary(mod)$r.squared

[1] 0.10128

As above, 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

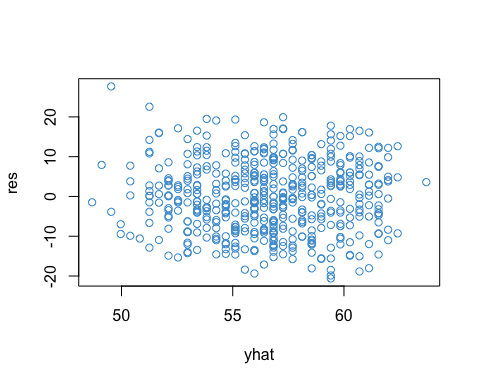
# Check that R-squared is really equal to the square of the PPMC  
cor(achmat08, ses)^2

[1] 0.10128

### 2.12.3 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 just take a look at the residual vs fitted plot to illustrate the code.

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



cor(yhat, res)

[1] -2.9515e-16

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

### 2.12.4 Inference

Next let’s talk about statistical inference, or how we can make conclusions about a population based on a sample from that population.

We can use the summary function to test the coefficients in our model.

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 intercept is not very meaningful (why?).

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 also 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.

### 2.12.5 Power analysis

Power analyses should ideally be done 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 analsyis 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.

### 2.12.6 Additional exercises

If time permits, we will address these additional exercises in class.

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?