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 review up to the Exercises in Section **??** 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 Σ to stand-in for summation. For example, instead of writing

$$X_1 + X_2 + X_3 + ... + X_N$$

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

$$\sum_{i=1}^{N} X_i.$$

The symbol Σ means "add." The symbol is called "Sigma" – it's the capital Greek letter corresponding to the Latin letter "S". The value i is called the index, and 1 is the starting value of the index and N 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 X, we write

$$\sum_{i=2}^{3} X_i = X_2 + X_3.$$

When the start and end evalues are clear from context, we can use a shorthand notation that omits them. In the following, it is implicit that the sum is over the all available values of X (i.e., from 1 to N):

$$\sum_i X_i.$$

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 c is the same as multiplication. Specifically, if you add a constant c to itself N times, this just N times the constant:

$$\sum_{i=1}^{N} c = c + c + ... \tag{1.1}$$

$$= Nc \tag{1.2}$$

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

$$\sum_{i=1}^{N} cX_i = cX_1 + cX_2 + ... \tag{1.3}$$

$$= c(X_1 + X_2 + ...) \tag{1.4}$$

$$= c\sum_{i=1}^{N} X_i \tag{1.5}$$

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

$$\sum_{i=1}^{N}(X_i + Y_i) = (X_1 + Y_1) + (X_2 + Y_2) + ... \tag{1.6}$$

$$= (X_1 + X_2 + ...) + (Y_1 + Y_2 + ...) \tag{1.7}$$

$$= \sum_{i=1}^{N} X_i + \sum_{i=1}^{N} Y_i \tag{1.8}$$

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

$$\bar{X} = \frac{\sum_i X_i}{N}$$

- The variance can be written as $\text{var}(X)$ or sometimes using the symbol s^2

$$\text{var}(X) = \frac{\sum_i (X_i - \bar{X})^2}{N - 1}$$

- The standard deviation can be written $\text{SD}(X)$ or using the letter s

$$\text{SD}(X) = \sqrt{\text{var}(X)}$$

- The covariance is a generalization of the variance to two variables, it describes how they co-vary:

$$\text{cov}(X, Y) = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{N - 1}$$

- 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.

$$\text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}}$$

For numerical examples see Section **??**.

1.4 Some properties of sample statistics

The following are some useful properties of the sample statistics reviewed above. You can derive these 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): $d_i = X_i - \bar{X}$. It is always the case that $\sum_i d_i = 0$

- Derivation

$$\sum_i d_i = \sum_i (X_i - \bar{X}) = ...$$

Mean of a linear transformation. If $Y_i = A + BX_i$ with known constants A and B, then $\bar{Y} = A + B\bar{X}$

- Derivation:

$$\bar{Y} = \frac{\sum_i Y_i}{N} = \frac{\sum_i (A + BX_i)}{N} = ...$$

Variance of a linear transformation. If $Y_i = A + BX_i$ with known constants A and B, then $\text{var}(Y) = B^2\text{var}(X)$

- Derivation

$$\text{var}(Y) = \frac{\sum_i (Y_i - \bar{Y})^2}{N - 1} = \frac{\sum_i ((A + BX_i) - (A + B\bar{X}))^2}{N - 1} = ...$$

Mean and variance of a z-score. The z-score (or standardized score) is defined as $Z_i = (X_i - \bar{X})/\text{SD}(X)$. Standardized scores are useful because they have $\bar{Z} = 0$ and $\text{var}(Z) = 1$.

- Derivation: use the rules for linear transformation with $A = -\bar{X}/\text{SD}(X)$ and $B = 1/\text{SD}(X)$.

Covariance of linear transformation. If $Y_i = A + BX_i$ and $W_i = C + DU_i$ with known constants A, B, C, D, then $\text{cov}(Y, W) = BD\,\text{cov}(X, U)$

- Derivation

$$\text{cov}(Y, W) = \frac{\sum_i ((A + BX_i) - (A + B\bar{X}))((C + DU_i) - (C + D\bar{U}))}{N - 1} = ...$$

Correlation of linear transformation. If $Y_i = A + BX_i$ and $W_i = C + DU_i$ with known constants A, B, C, D, then $\text{cor}(Y, W) = \text{cor}(X, U)$ – 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 a statistic. These properties are defined in terms of the *sampling distribution* of a statistic. Recall that a sampling distribution arise from the following thought experiment: 1. Take a random sample of size N from a population of interest. 2. Compute a statistic using the sample data. It can be any statistic, but let's say the mean, $\bar{X}$, 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,

$$\bar{X}_1, \bar{X}_2, \bar{X}_3, \bar{X}_4, \bar{X}_5...$$

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 espeically interested in the following two statistics.

- **The expected value** of the mean, denoted $E(\bar{X})$, is the mean of the sampling distribution of the mean. That is a mouthful! That is why we say the expected value 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 $SE(\bar{X})$, 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 a statistic:

- **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: $E(\bar{X}) = \mu$), 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., $1/\text{SE}(\bar{X})^2$) 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.

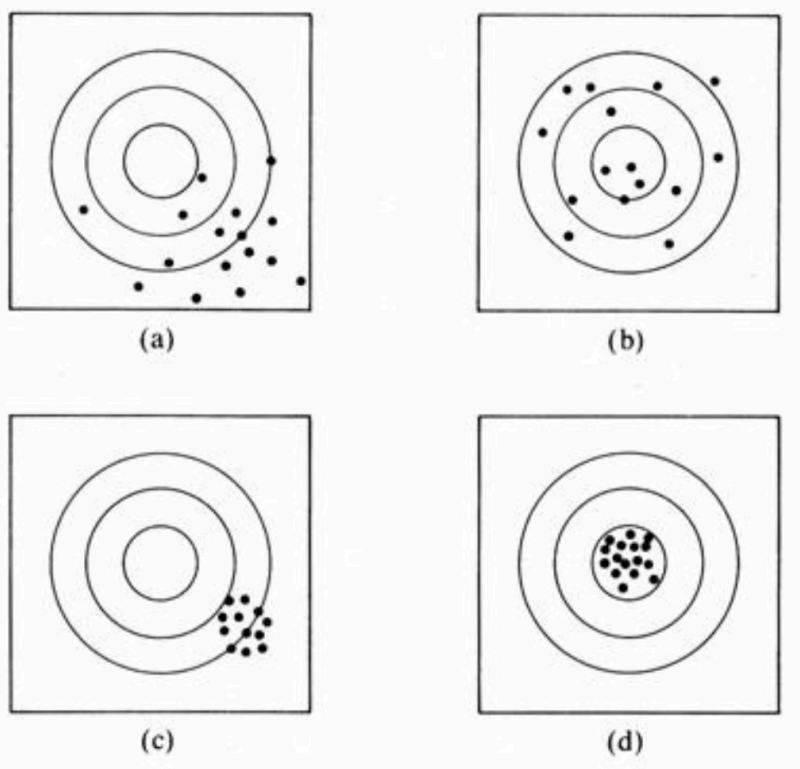

Figure 1.1: Bias and Precision

1.6 t-tests

The t-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 t-test is

$$t = \frac{\text{unbiased estimate} - \text{hypothesized value}}{\text{standard error}}$$

When we conduct a t-test, the basic rationale is as follows: "if the estimate statistic is close the hypothesized value of the parameter, then the numerator should be small relative to the standard error, and so t 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 t is close to zero, it means there was no effect.

In order to determine what values of t are "close to zero", we refer to its sampling distribution, which is called the t-distribution. The t-distribution tells what values of t 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 t has a "central" t-distribution.)

Some examples of the t-distribution are shown below. The x-axis denotes values of the statistic t shown above, and ν is a parameter called the "degrees of freedom" (more on this below). You can see that the t-distribution looks like a normal distribution centered a zero. So, when the null hypothesis is true, the expected value of t is zero. Informally, we could say that values greater than ± 2 are pretty unlikely, and values greater than ± 4 are very unlikely. Keep in the mind that these are the values of t we are expecting if the null hypothesis is true.

More formally, we can compare the value of t computed in a sample, denoted as t_{obs}, to a "critical value", denoted as t^*. The critical value is chosen so that the "significance level", defined as $\alpha = \text{Prob}(|t| \geq t^*)$, is sufficiently small.

This 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 α 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 t would need to be very unlikely under the null hypothesis.

There are two equivalent ways of "formally" conducting a t-test.

1. Compare the observed value of t to the critical value. Specifically: if $t_{\text{obs}} > t^*$, reject the null hypothesis.

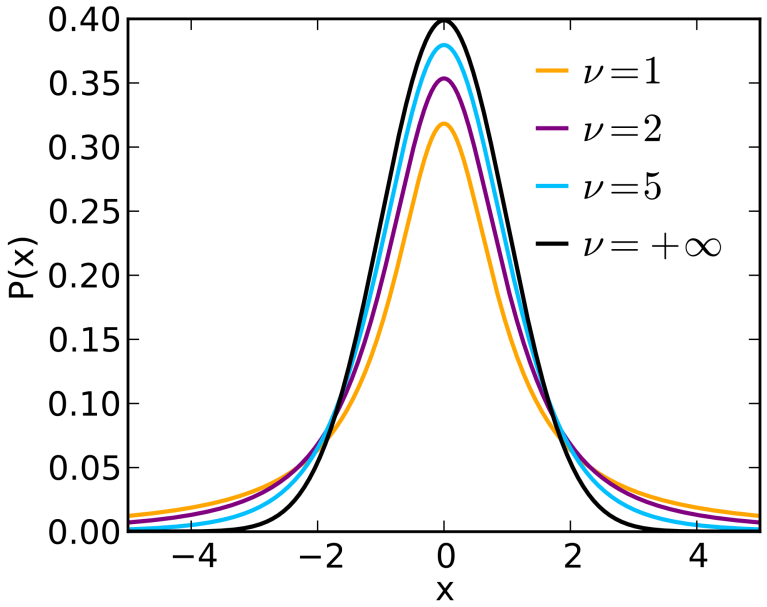

Figure 1.2: t-distribution (source: https://en.wikipedia.org/wiki/Student%27s_t-distribution)

2. Compare the significance level chosen by the research, α, to the "p-value" of the test, computed as $p = \text{Prob}(|t| \geq |t_{\text{obs}}|)$. Specifically: if $p < \alpha$, reject the null hypothesis.

Informally, both of these just mean that the absolute value of t should be pretty big (i.e., greater than t^*) 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 ν in Figure **??**. 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 t-distribution approaches a normal distribution. This means that that the differences between a t-test and a z-test is pretty minor in large samples (say $N \geq 100$).

However, when the degrees of freedom are small, the t-distribution has wider tails than the normal distribution. This is also shown in Figure **??**. The tails of the distribution are important when doing statistical tests, because we are interested knowing about large / unlikely values of t. So, in small samples (say $N < 100$), it is important to use the t-distribution.

1.7 Confidence intervals

A confidence interval uses the same equation as a t-test, except we solve for the population parameter rather than the value of t. Whereas a t-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:

$$\text{Interval} = \text{sample value} \pm t \times \text{standard error}.$$

We get the value of t from the t-distribution. In particular, if we want the interval to include the true population parameter $(1 - \alpha) \times 100\%$ of the time, then we choose t to be the $\alpha/2 \times 100$ percentile of the t-distribution. For example, if we set $\alpha = .05$, we will have a $(1 - \alpha) \times 100 = 95\%$ confidence interval by choosing t to be the $\alpha/2 \times 100 = 2.5$-th percentile of the t-distribution.

As mentioned, t-tests and confidence intervals are closely related. In particular, if the confidence interval includes the value 0, this is the same as retaining the null hypothesis that the parameter is equal to 0. This should make intuitive sense. If the confidence interval includes 0, we are saying that it is a reasonable value of the population parameter, so we should not reject that value. This relationship assumes we use the same level of α for both the test and confidence interval.

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

1.8 F-tests

The F-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 $N - 1$ (its degrees of freedom).

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

$$F = \frac{SS_A/df_A}{SS_B/df_B},$$

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

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

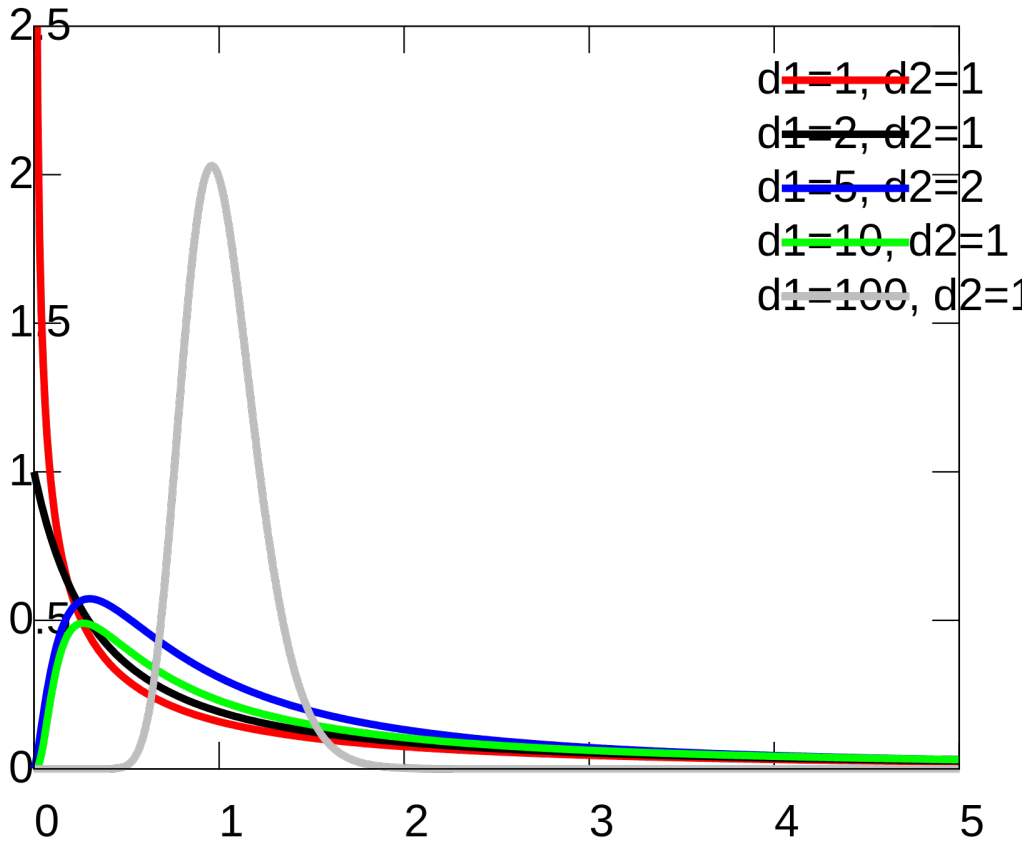

Figure 1.3: F-distribution (source: https://en.wikipedia.org/wiki/F-distribution)

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 df first and then the denominator df. So, the green line in the figure is "an F-distribution on 10 and 1 degrees of freedom", which means the df in the numerator is 10 and the df in the denominator is 1.

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

usually compute confidence intervals for statistics with an F distribution.

1.9 APA reporting

It is important to be able 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 Table 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).

- Jointly, the two predictors explained about 22% of the variation in Academic Achievement, which was statistically significant at the .05 level ($R^2 = .22, F(2, 247) = 29.63, p < .001$).
- After controlling for SES, a one unit of increase in Maternal Education was associated with $b = 1.33$ units of increase in Academic Achievement ($t(247) = 5.26, p < .001$).
- After controlling for Maternal Education, a one unit of increase in SES was associated with $b = 0.32$ units of increase in Academic Achievement. This was a statistically significant relationship ($t(247) = 2.91, p < .01$).

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 use`CAPS` 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 hashtag 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 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

Often we want to work with multiple numbers or other objects at once. 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 **??** to Section **??**

```
# 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
```

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

```
[1] 24
```

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

```
[1] TRUE
```

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

```
[1] TRUE
```

```r
# Compute the mean
mean(x)
```

```
[1] 14.33333
```

```r
# Compute the variance and sd
var(x)
```

```
[1] 17.46667
```

```r
sd(x)
```

```
[1] 4.179314
```

```r
# Compute the covariance and correlation
cov(x, y)
```

```
[1] 10.66667
```

```
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, lets 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 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 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 datasetby their names directly

```r
# Attach the data set
attach(NELS)

# Print the first 10 values of the 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:

```r
detach(NELS)
```

1.10.10 Preview of next week

Figure **??** 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.

```r
# 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)
```

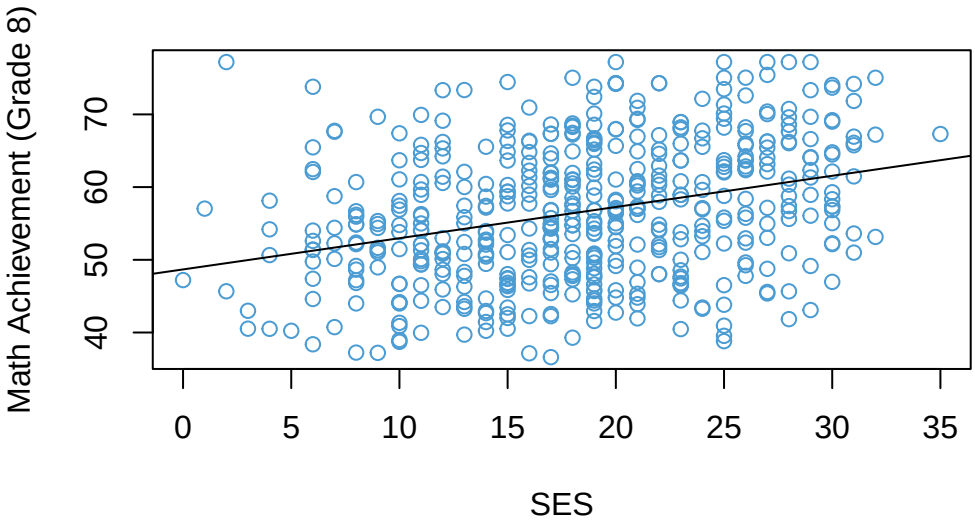

Figure 1.4: Math Achievement and SES (NELS88).

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 = "

# Run the regression model
mod <- lm(achmat08 ~ ses)

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

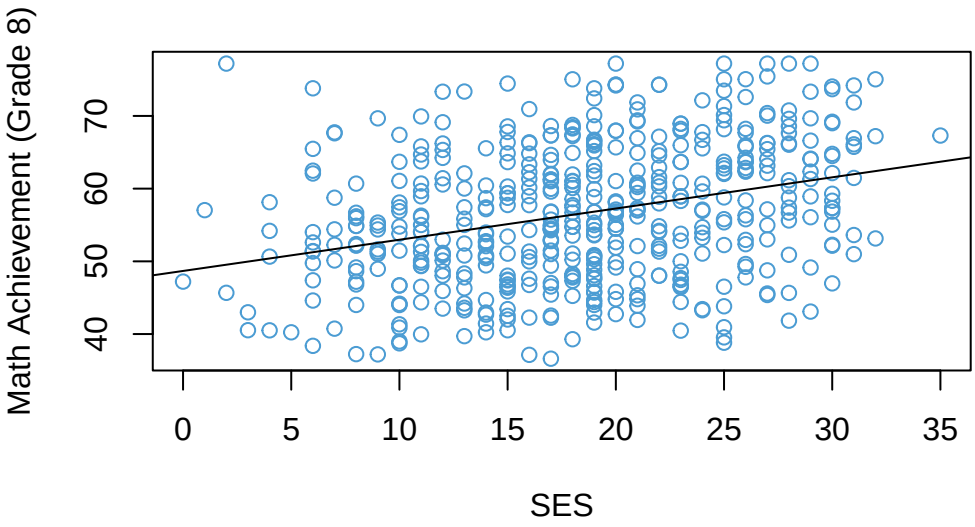

Figure 2.1: Math Achievement and SES (NELS88).

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

$$\widehat{Y} = a + bX(\#eq : yhat)$$

where

- Y denotes Math Achievement
- X denotes SES
- a represents the regression intercept (the value of $\widehat{Y}$ when $X = 0$)
- b represents the regression slope (how much $\widehat{Y}$ increases for each unit of increase in X)

Note that Y represents the values of Math Achievement in the data, whereas $\widehat{Y}$ represents the values computed from the regression equation (i.e., the values on the regression line). The difference $e = Y - \widehat{Y}$ 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 =

# Overwrite dots to make it look at bit better
points(x = ses[index], y = achmat08[index], col = "#4B9CD3", pch = 16)
```

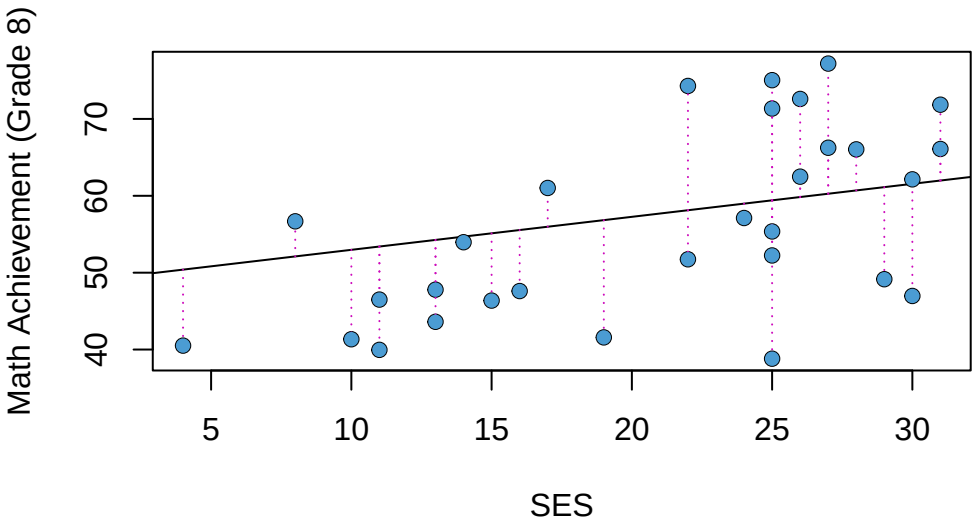

Figure 2.2: Residuals for a Subsample of the Example.

Notice that $Y = \widehat{Y} + e$ by definition. So, we can use either Equation @ref(eq:yhat) or the equation below to write out a regression model:

$$Y = a + bX + e. (\#eq : y) \tag{2.1}$$

Both equations say the same thing, but Equation @ref(eq:y) lets us talk about the values of Y 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,

$$\widehat{MATH} = a + b(SES). (\#eq : read) \tag{2.2}$$

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, r_{XY} is much clearer than $r_{SES,MATH}$ – 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.