EDUC 784: Week 2 Readings

Peter Halpin

2022-01-14

# 2 Simple Regression

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

## 2.1 An example from NELS

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

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

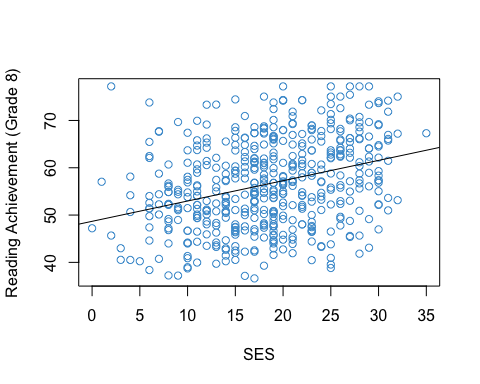


Figure 2.1: Reading Achievement and SES (NELS88).

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

options(digits = 2)  
cor(achmat08, ses)

## [1] 0.32

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

This relationship has been widely documented and discussed in education research (e.g., <https://www.apa.org/pi/ses/resources/publications/education>). Please share your thoughts on this relationship in class.

## 2.2 The regression line

The line in the Figure 2.1 can be represented mathematically as

where

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

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

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

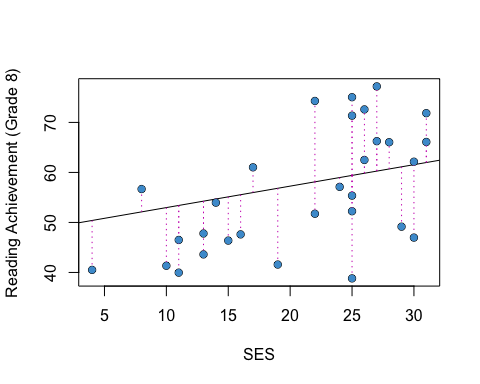


Figure 2.2: Residuals for a Subsample of the Example.

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

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

## 2.3 OLS

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

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

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

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

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

coef(mod)

## (Intercept) ses   
## 48.68 0.43

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

### 2.3.1 Correlation and regression

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

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

## 2.4 R-squared

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

This question is represented graphically in Figure 2.3. The horizontal line denotes the mean of Reading Achievement. The difference between the indicated student’s Reading Achievement score and the mean can be divided into two parts. The black dashed line shows how much closer we get to the student’s score by considering instead of . This represents the extent to which this student’s Reading Achievement score is explained by the linear relationship with SES. The pink part is the regression residual, which was introduced in Section 2.2. This is the variation in Reading Achievement that is “left over” after considering the linear relationship with SES.

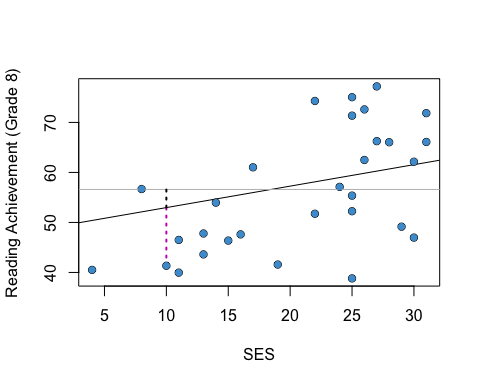


Figure 2.3: The Idea Behind R-squared.

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

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

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

## [1] 0.10128

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

### 2.4.1 Derivation\*

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

It can be re-written using the predicted values :

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

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

The R-squared statistic is

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

## 2.5 The population model

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

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

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

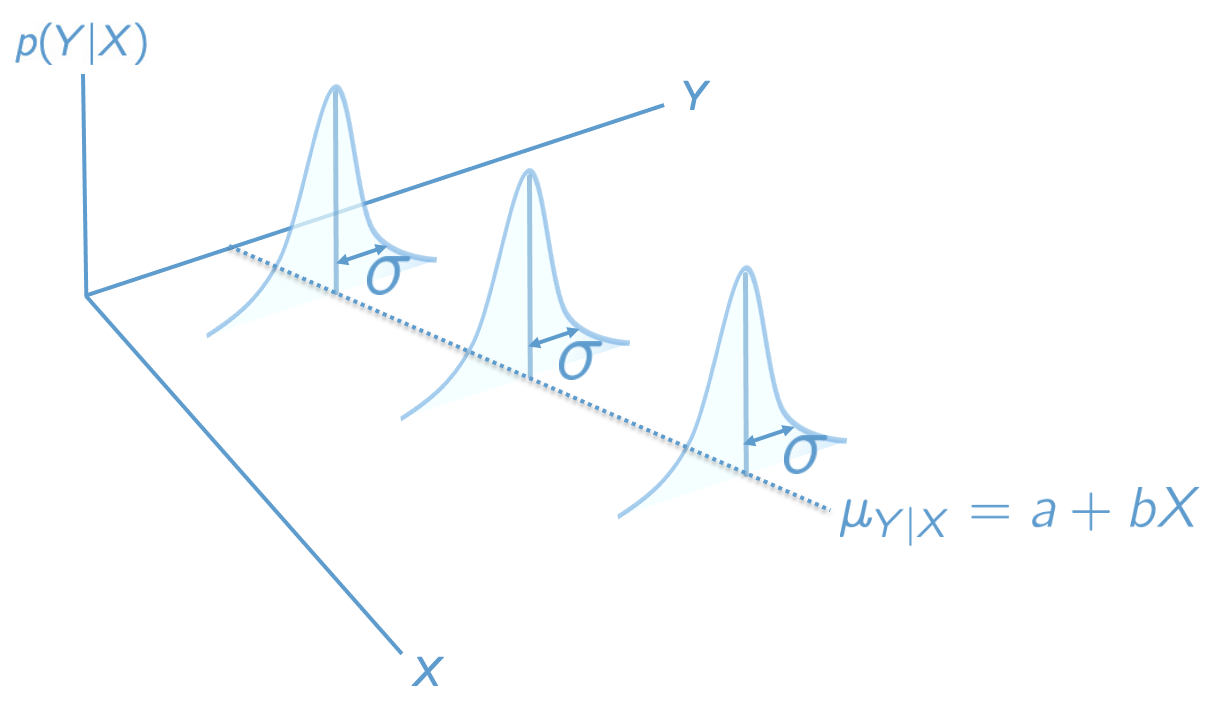


Figure 2.4: The Regression Population Model.

These three assumptions are summarized by writing

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

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

## 2.6 Clarifying notation

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

* In Section 2.2 they denoted sample statistics.
* In Section 2.5 they denoted population parameters.

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

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

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

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

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

## 2.7 Inference for the slope

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

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 ?? for background information on bias, precision, t-tests, and confidence intervals.

### 2.7.1 t-tests

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

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

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

### 2.7.2 Confidence intervals

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

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

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

### 2.7.3 The NELS example

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

summary(mod)

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

The corresponding confidence interval is

confint(mod)

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

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

## 2.8 Inference for the intercept

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

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

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

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

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

## 2.9 Inference for R-squared

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

### 2.9.1 F-tests

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

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

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

summary(mod)

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

## 2.10 Power analysis

Statistical power is the probability of rejecting the null hypothesis, when it is indeed false. Rejecting the null hypothesis when it is false is sometimes called a “true positive”, meaning we have correctly inferred that a parameter of interested 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 since, since we already have the data. Instead, we can solve for the effect size. In the NELS example we have observations. The output below reports the smallest effect size we can detect with a power of .8 and . This is sometimes called the “minimum detectable effect size” (MDES). Note that and denote the degrees of freedom in the numerator and denominator of the F-test of R-squared, respectively.

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

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

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

## 3. Interpretations of Regression

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

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

## 3.1 Regression and prediction

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

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

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

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

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

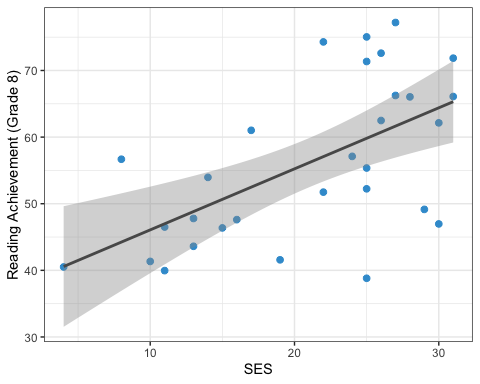


Figure 3.1: Prediction Error for Example Data.

Notice that the prediction error variance increases with – in other words, the larger the residuals (Figure 2.2), the worse the prediction error. As we will see in this Section (??), one way to reduce is to add more predictors into the model – i.e., multiple regression.

### 3.1.1 More about Prediction

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

## 3.2 Regression and causation

A causal interpretation of regression means that that changing by one unit will change by units. Note that this is an assumption about the population model, specifically the conditional expectation function.

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

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

* When a predictor variable that is correlated with and is left out of a regression model, this is called omitted variable bias.
* The overall idea is basically the same as saying “correlation does not imply causation” or the notion of spurious correlations.
* This is also an example of what is called “endogeneity” in regression (etymology: originating from within).
* In order to avoid omitted variable bias, we want to include more than one predictor in our regression models – i.e., multiple regression.

### 3.2.1 Omitted variable bias\*

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

Let’s write the “true” model as:

$$

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

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

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

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

(Side note: by adding the residual into the model for , we ensure the equality holds for not just – more on this later.)

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

$$

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

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

This last equation for is what we are interested in. It shows that the regression parameter in our data analytic model using just SES, , is not equal to the “true” regression parameter using both predictors, .

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

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

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

We will discuss the interpretation of these situations in class.

## 3.3 Regression and explanation

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

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

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

Even this long list leaves out potential omitted variables. But, by including more than one predictor, we can get “closer” to a causal interpretation through “statistical controls” (See Chapter 3).

When the goal of an analysis is explanation, it usual to focus on the proportion of variation in the outcome variable that is explained by the predictor(s), i.e., R-squared.