CHAPTER OVERVIEW

26: The General Linear Model

Learning Objectives

- Describe the concept of linear regression and apply it to a bivariate dataset
- Describe the concept of the general linear model and provide examples of its application
- Describe how cross-validation an allow us to estimate the predictive performance of a model on new data

Remember that early in the book we described the basic model of statistics:

$$outcome = model + error$$

where our general goal is to find the model that minimizes the error, subject to some other constraints (such as keeping the model relatively simple so that we can generalize beyond our specific dataset). In this chapter we will focus on a particular implementation of this approach, which is known as the *general linear model* (or GLM). You have already seen the general linear model in the earlier chapter on Fitting Models to Data, where we modeled height in the NHANES dataset as a function of age; here we will provide a more general introduction to the concept of the GLM and its many uses.

Before we discuss the general linear model, let's first define two terms that will be important for our discussion:

- *dependent variable*: This is the outcome variable that our model aims to explain (usually referred to as *Y*)
- *independent variable*: This is a variable that we wish to use in order to explain the dependent variable (usually referred to as *X*).

There may be multiple independent variables, but for this course we will focus primarily on situations where there is only one dependent variable in our analysis.

A general linear model is one in which the model for the dependent variable is composed of a *linear combination* of independent variables that are each multiplied by a weight (which is often referred to as the Greek letter beta - β), which determines the relative contribution of that independent variable to the model prediction.

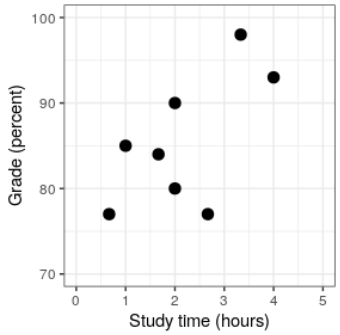

Figure 26.1: Relation between study time and grades

As an example, let's generate some simulated data for the relationship between study time and exam grades (see Figure 26.1). Given these data, we might want to engage in each of the three fundamental activities of statistics:

- *Describe*: How strong is the relationship between grade and study time?
- *Decide*: Is there a statistically significant relationship between grade and study time?
- *Predict*: Given a particular amount of study time, what grade do we expect?

In the last chapter we learned how to describe the relationship between two variables using the correlation coefficient, so we can use that to describe the relationship here, and to test whether the correlation is statistically significant using the `cor.test()` function in R:

```
##
##  Pearson's product-moment correlation
##
## data:  df$grade and df$studyTime
## t = 2, df = 6, p-value = 0.05
## alternative hypothesis: true correlation is greater than 0
## 95 percent confidence interval:
##  0.014 1.000
## sample estimates:
##  cor
## 0.63
```

The correlation is quite high, but just barely reaches statistical significance because the sample size is so small.

26.1: Linear Regression
26.2: Fitting More Complex Models
26.3: Interactions Between Variables
26.4: Beyond Linear Predictors and Outcomes
26.5: Criticizing Our Model and Checking Assumptions
26.6: What Does “Predict” Really Mean?
26.7: Suggested Readings
26.8: Appendix
