Statistical Thinking for the 21st Century

Chapter 14

The General Linear Model

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.

In this chapter we will focus on a particular implementation of this approach, which is known as the *general linear model* (or GLM).

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 or predict** (usually referred to as Y)
- *independent variable*: This is a variable that we wish to use in order to **explain or predict** 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.

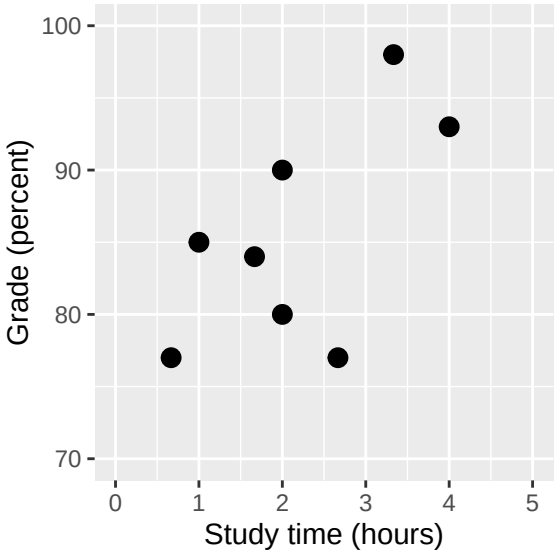


Figure 14.1: Relation between study time and grades

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.

As an example, let's generate some simulated data for the relationship between study time and exam grades (see Figure 14.1). 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 how to **decide** whether the correlation is statistically significant.

The correlation is quite high ($r = 0.632$), but it just barely reaches statistical significance because the sample size is so small.

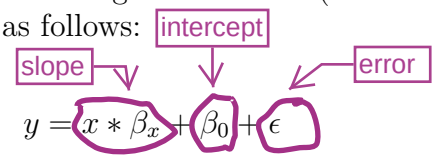
14.1 Linear regression

We can use the general linear model to describe the relation between two variables and to decide whether that relationship is statistically significant; in addition, the general linear model allows us to predict the value of the dependent variable given some new value(s) of the independent variable(s). Most importantly, the general linear model will allow us to build models that incorporate multiple independent variables, whereas correlation can only tell us about the relationship between two individual variables.

The specific version of the GLM that we use for this is referred to as as *linear regression*.

This is the slope; in regression, it's called the Beta weight because it tells us how much to weight our variable, *x*.

The simplest version of the linear regression model (with a single independent variable) can be expressed as follows:



The β_x value tells us how much we would expect y to change given a one-unit change in x. The intercept β_0 is an overall offset, which tells us what value we would expect y to have when $x = 0$; you may remember from our early modeling discussion that this is important to model the overall magnitude of the data, even if x never actually attains a value of zero. The error term ϵ refers to whatever is left over once the model has been fit. If we want to know how to predict y (which we call $\hat{y}$), then we can drop the error term:

$$\hat{y} = x * \beta_x + \beta_0$$

Note that this is simply the equation for a line, where β_x is the slope and β_0 is the intercept. Figure 14.2 shows an example of this model applied to the study time example.

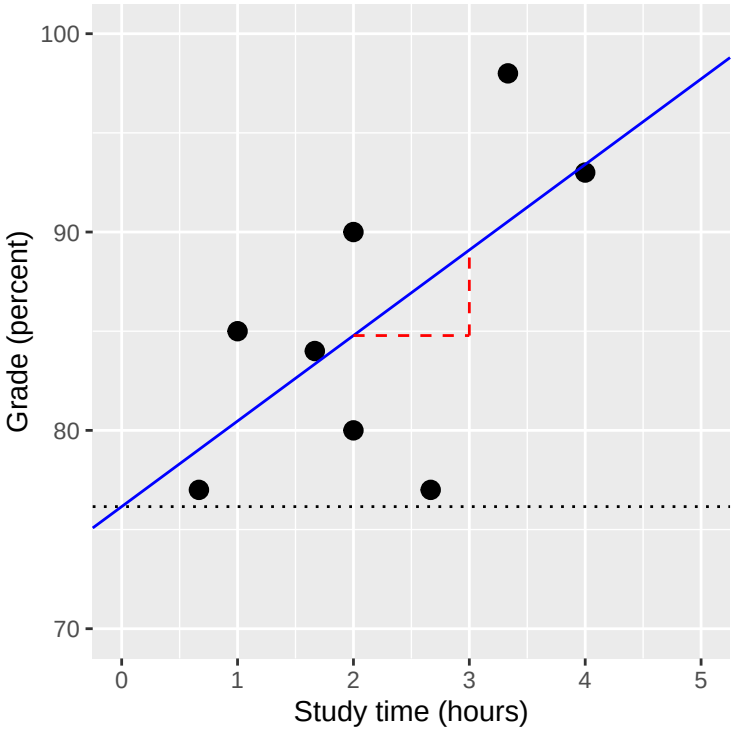


Figure 14.2: The linear regression solution for the study time data is shown in the solid line. The value of the intercept is equivalent to the predicted value of the y variable when the x variable is equal to zero; this is shown with a dotted line.

The value of Beta is equal to the slope of the line – that is, how much y changes for a unit change in x. This is shown schematically in the dashed lines, which show the degree of increase in grade for a single unit increase in study time.