Chapter 2

Introduction

A common situation in applied sciences is that one has an independent variable or outcome Y and one or more dependent variable or covariates $X_1, \ldots, X_p$. One usually observes these variables for various "subjects".

Note: We use upper case to denote random variable. To denote actual numbers we use lower case. One way to think about it: Y has not happened yet, and when it does we see $Y = y$.

One may be interested in various things: What effects do the covariates have on the outcome? How well can we describe these effects? Can we predict the outcome using the covariates?, etc..

2.1 Linear Regression

Linear regression is the most common approach for describing the relation between predictors (or covariates) and outcome. Here we will see how regression relates to prediction.

Let's start with a simple example. Lets say we have a random sample of US males and we record their heights (X) and weights (Y).

Say we pick a random subject. How would you predict their weight?

What if I told you their height? Would your strategy for predicting change?

We can show mathematically that for a particular definition of "best", described below, the average is the best predictor of a value picked from that population. However, if we have information about a related variable then the conditional average is best.

One can think of the conditional average as the average weights for all men of a particular height.

In the case of weight and height, the data actually look bivariate normal (football shaped) and one can show that the best predictor (the conditional average) of weight given height is

$$\mathrm{E}[Y|X = x] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X) \tag{2.1}$$

with $\mu_X = \mathrm{E}[X]$ (average height), $\mu_Y = E[Y]$ (average weight), and where ρ is

Figure 2.1: Weight versus height.

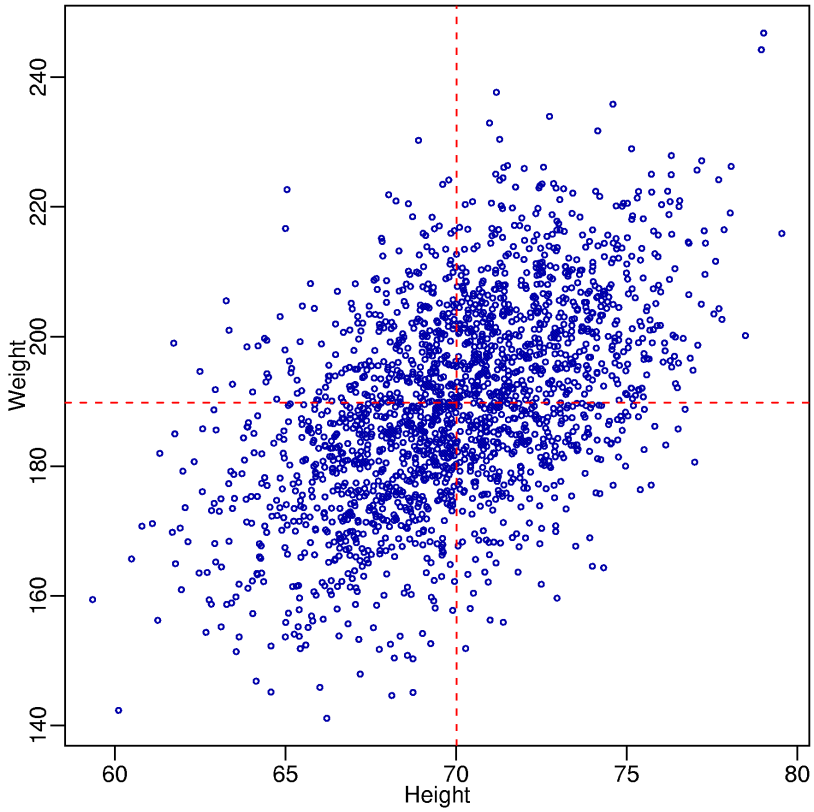
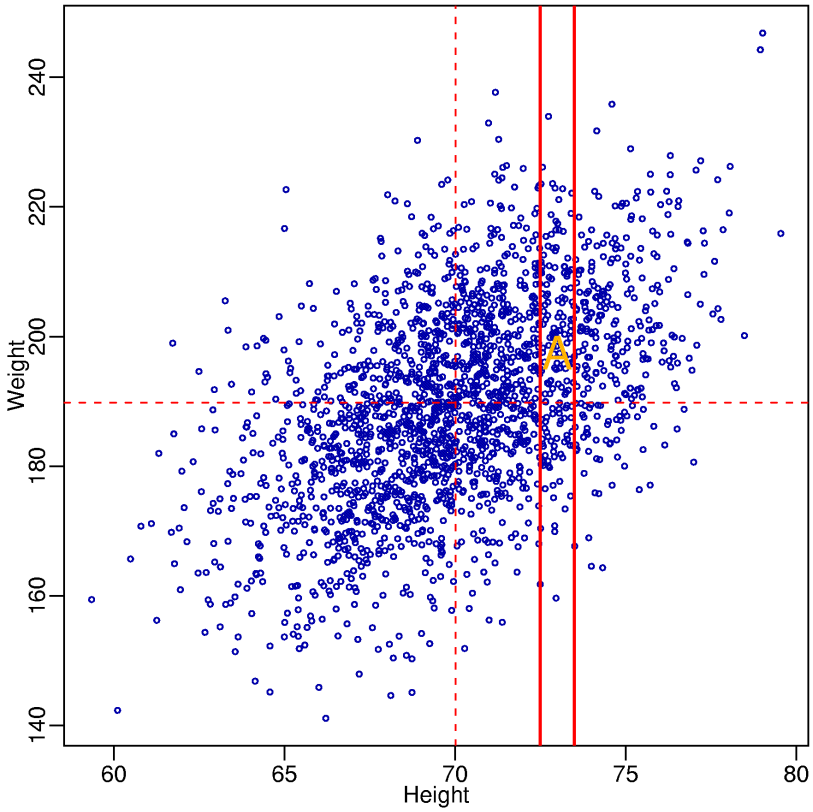


Figure 2.2: Weight versus height. 73 inch men highlighted.



the correlation coefficient of height and weight.

If we obtain a random sample of the data then each of the above parameters is substituted by the sample estimates and we get a familiar expression:

$$\hat{Y}(x) = \bar{X} + r\frac{SD_Y}{SD_X}(x - \bar{X}).$$

Technical note: Because in practice it is useful to describe distributions of populations with continuous distribution we will start using the word *expectation* or the phrase *expected value* instead of average. We use the notation $\mathrm{E}[\cdot]$. If you think of integrals as sums then you can think of expectations as averages.

Notice that equation (2.1) can be written in this, more familiar, notation:

$$\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x$$

Because the conditional distribution of Y given X is normal we can write the even more familiar version:

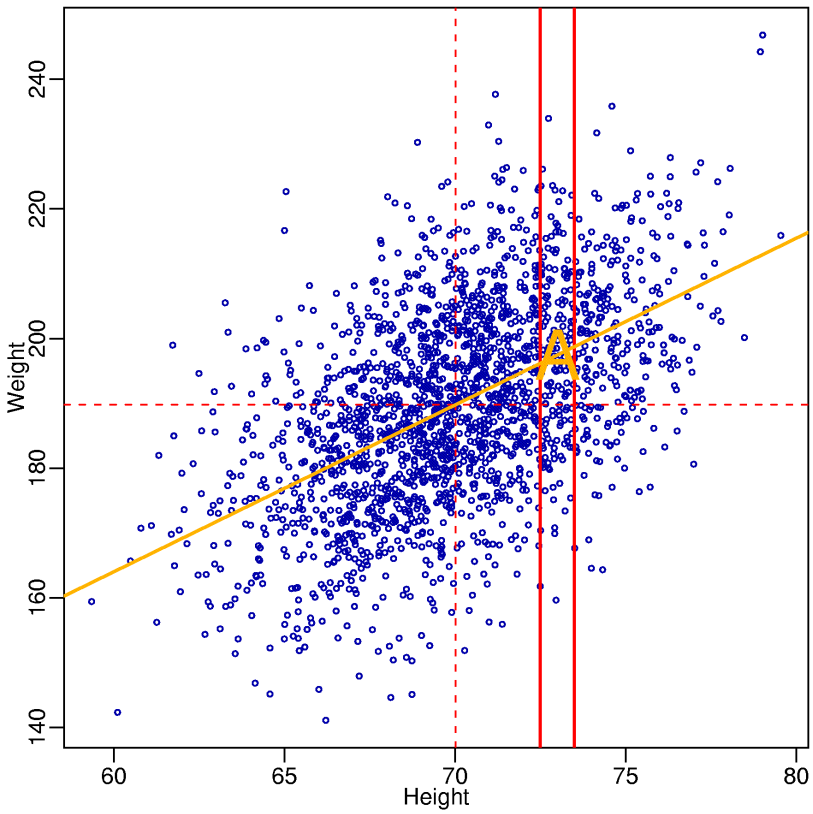
$$Y = \beta_0 + \beta_1 X + \epsilon$$

with ϵ a mean 0 normally distributed random variable that is independent of X. This notation is popular in many fields because β_1 has a nice interpretation and its typical (least squares) estimate has nice properties.

When more than one predictor exists it is quite common to extend this linear regression model to the multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon_i$$

Figure 2.3: Regression line.



with the ϵ_is unbiased (0 mean) errors independent of the X_j as before.

A drawback of these models is that they are quite restrictive. Linearity and additivity are two very strong assumptions. This may have practical consequences. For example, by assuming linearity one may never notice that a covariate has an effect that increases and then decreases. We will see various example of this in class.

Linear regression is popular mainly because of the interpret-ability of the parameters. However, the interpretation only makes sense if the model is an appropriate approximation of the natural data generating process. It is likely that the linear regression model from a randomly selected publication will do a terrible job at predicting results in data where the model was not trained on. Prediction is not really given much importance in many scientific fields, e.g. Economics, Epidemiology, and Social Sciences. In others fields, e.g. Surveillance and Finance, prediction is everything. Notice that in the fields where prediction is important regression is not as popular.

2.2 Prediction

Methods for prediction can be divided into two general groups: continuous outcome and discrete outcomes.

When the data is discrete we will refer to it as *classification*. Other terms are *discriminant analysis*, *machine learning*, *supervised learning*.

When the data is continuous we will refer to it as *regression*. Other terms are *smoothing* and *curve estimation*.

These seem very different but they have some in common. In this class we will talk about the commonalities but in general we will discuss these two cases separately.

The main common characteristic in both cases we observe predictors $X_1, \ldots, X_p$ and we want to predict the outcome Y.

Note: I will use X to denote the vector of all predictors. So X_i are the predictors for the i-th subject and can include age, gender, ethnicity, etc...

Note: Given a prediction method we will use $f(x)$ to denote the prediction we would get if the predictors are $X = x$.

Q: What are examples of prediction problems?

So, what does it mean to predict well? Let's look at the continuous data case first.

If I have a prediction $f(X)$ based on the predictors X how do I define a "good prediction" mathematically. A common way of defining closeness is with Euclidean distance:

$$L\{Y, f(X)\} = \{Y - f(X)\}^2. \tag{2.2}$$

We sometime call this the *loss function*.

Notice that because both Y and $f(X)$ are random variables so is (2.2). Minimizing a random variable is meaningless because its not a number. A common thing to

do is minimize the average loss or the **expected prediction error**:

$$\mathrm{E}_X \mathrm{E}_{Y|X}[\{Y - f(X)\}^2|X]$$

For a given x the expected loss is minimized by the conditional expectation:

$$f(x) = \mathrm{E}[Y|X = x]$$

so it all comes down to getting a good estimate of $\mathrm{E}[Y|X = x]$. We usually call $f(x)$ the *regression function*.

Note: For discrete problems we usually want a plausible prediction. Note $f(x)$ is typically a continous number and not a class. We can take an extra step and define a prediction rule. For example for binary outcome we can say: if $f(x) > 0.5$ I predict a 1, otherwise predict 0. However, it is useful to change the loss function. More on this later.

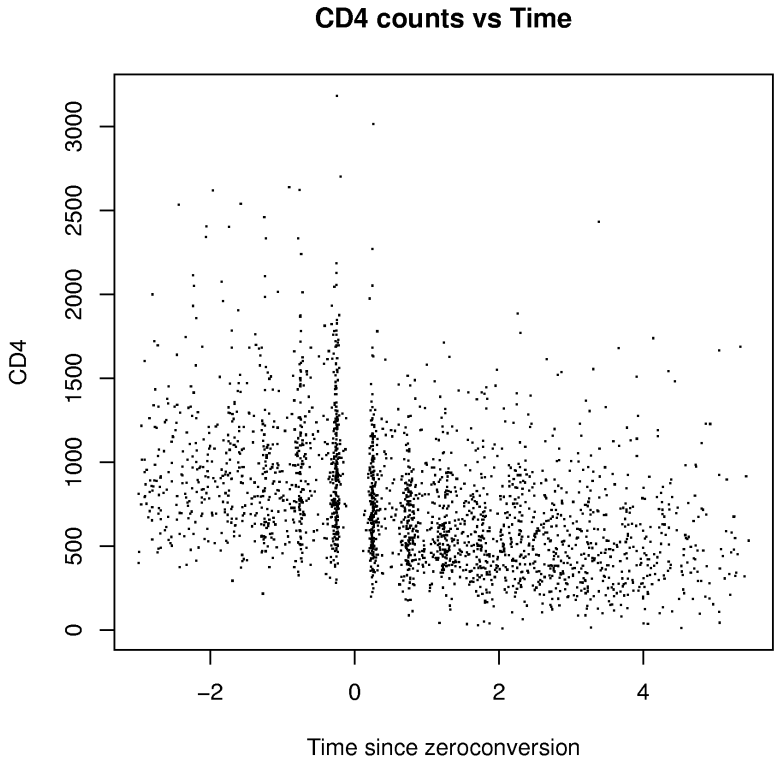
Notice that if the linear regression model holds then

$$f(X) = \mathrm{E}[Y|X_1 = x_1, \ldots, X_p = x_p] = \beta_0 + \sum_{j=1}^{p} x_j \beta_j.$$

For Gaussian models the solution is the same as for least squares and MLE. However, many times it is hard to believe that the linear regression model holds. A simple example comes from AIDS research. See Figure 2.4.

Technical note: It should be noted that for some designed experiments it does not make sense to assume the X are random variables. In this case we usually assume we have "design points" $x_{1i}, \ldots, x_{pi}, i = 1, \ldots, n$ and non-IID observations

Figure 2.4: CD4 Data



$Y_1, \ldots, Y_n$ for each design point. In most cases, the theory for both these cases is very similar if not the same. These are called the *random design model* and *fixed design model* respectively.