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".

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

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 the correlation coefficient of height and weight.

If we obtain a random sample of the data then each of the above parameters is subsituted 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:

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

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 seclectd 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

So, what does it mean to predict well?

Say I observe the predictors $X_1, \ldots, X_p$ and I want to predict Y.

Note: I will use X to denote the vector of all predictors.

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 all x the expected loss is minimized by the conditional expectation:

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

We usually call $f(x)$ the *regression function*.

Notice that if the linear regression model holds then

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

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 $Y_1, \ldots, Y_n$ 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.

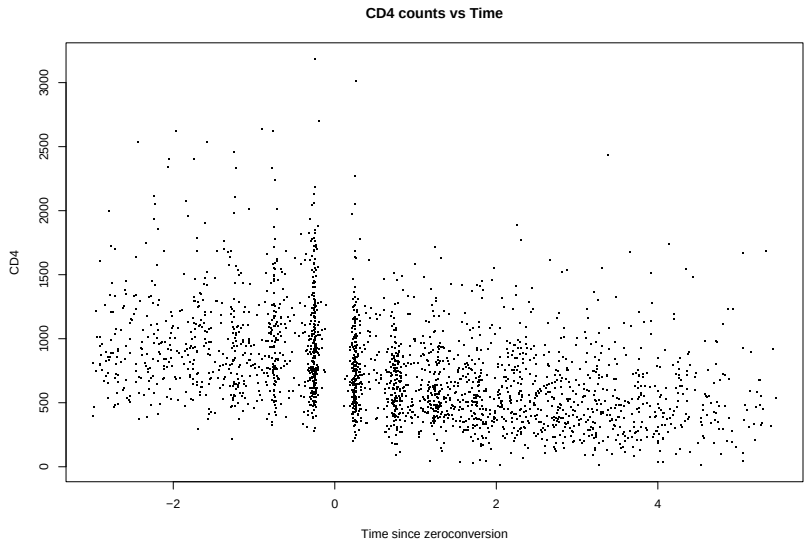
2.3 Scatterplot Smoothers

We will begin with a simple example were the regression function f will depend on a single predictor X but not necesarily linear.

Sometimes, the need to estimate f arises when investigators have to decide among various explanations for a physical phenomenon, and existing subject-knowledge or scientific theory says nothing about f. In this case we collect data to see what it says. Exhibiting some aspect of f may then imply the confirmation or revision of a given theory.

A scatter plot smoother is a tool for finding structure in a scatter plot: $(x_1, y_1), \ldots, (x_n, y_n)$

Figure 2.1: CD4 cell count since seroconversion for HIV infected men.



- Suppose that we consider $\mathbf{y} = (y_1, \ldots, y_n)'$ as the *response measurements* and $\mathbf{x} = (x_1, \ldots, x_n)'$ as the *design points*.
- We can think of $\mathbf{x}$ and $\mathbf{y}$ as outcomes of random variable X and Y. However, for scatter plot smoothers we don't really need stochastic assumptions,

it can be viewed as a descriptive tool.

- A scatter plot smoother can be defined as a function (remember the general definition of *function*) of $\mathbf{x}$ and $\mathbf{y}$ with domain at least containing the values in $\mathbf{x}$: $s = \mathbf{S}[\mathbf{y}|\mathbf{x}]$.
- There is usually a "recipe" that gives $s(x_0)$, which is the function $\mathbf{S}[\mathbf{y}|\mathbf{x}]$ evaluated at x_0, for all x_0. We will be calling x_0 the *target value* when we giving the recipe. Note: Some recipes don't give an $s(x_0)$ for all x_0, but only for the x's included in $\mathbf{x}$.

Note we will call the vector $\{s(x_1), \ldots, s(x_n)\}'$ as *the smooth* and gives us esimtates of f at $\mathbf{x}$.

Here is a simplistic example: For each $x \in \mathbf{x}$ define

$$s(x) = \text{ave}\{y_i; x_i = x\}.$$

What happens if the x_i are unique?

Since Y and X are, in general, non-categorical we don't expect to find many replicates at any given value of X. This means that we could end up with the data again, $s(x) = y$ for all x. Not very smooth!

Note: For convenience, through out this chapter, we assume that the data are sorted by X.

Many smoothers force $s(x)$ to be a smooth function of x. This is a fancy way of saying we think data points that are close (in x) should have roughly the same expectation.

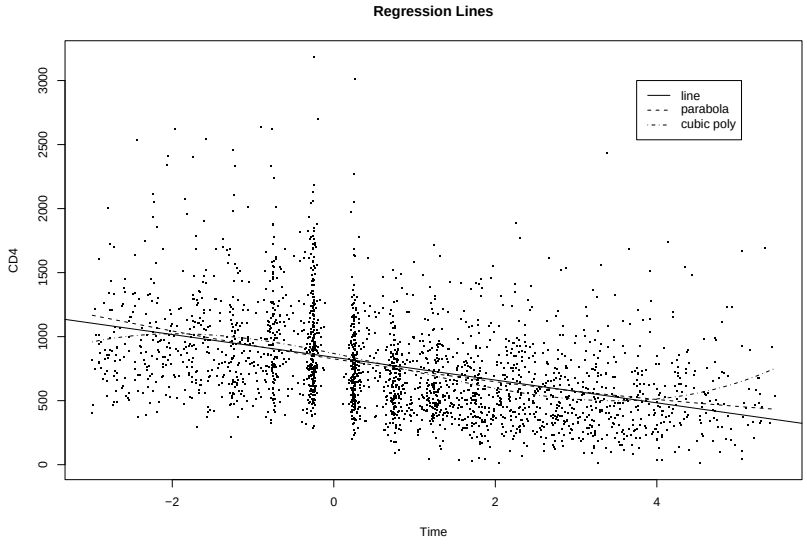
2.3.1 Parametric smoother

These are what you have seen already. We force a function defined by a few parameters on the data and use something like least squares to find the best estimates for the parameters.

For example, a regression line computed with least squares can be thought of as a smoother. In this case $S[\mathbf{y}|\mathbf{x}](x) = (1\, x)\,(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ with $\mathbf{X}$ a design matrix containing a column of 1's and $\mathbf{x}$.

The lack of flexibility of these types of smoother can make them provide misleading results.

Figure 2.2: CD4 cell count since seroconversion for HIV infected men.



2.3.2 Bin smoothers

A bin smoother, also known as a regressogram, mimics a categorical smoother by partitioning the predicted value into disjoint and exhaustive regions, then averaging the response in each region. Formally, we choose cut-points $c_0 < \ldots < c_K$ where $c_0 = -\infty$ and $c_K = \infty$, and define

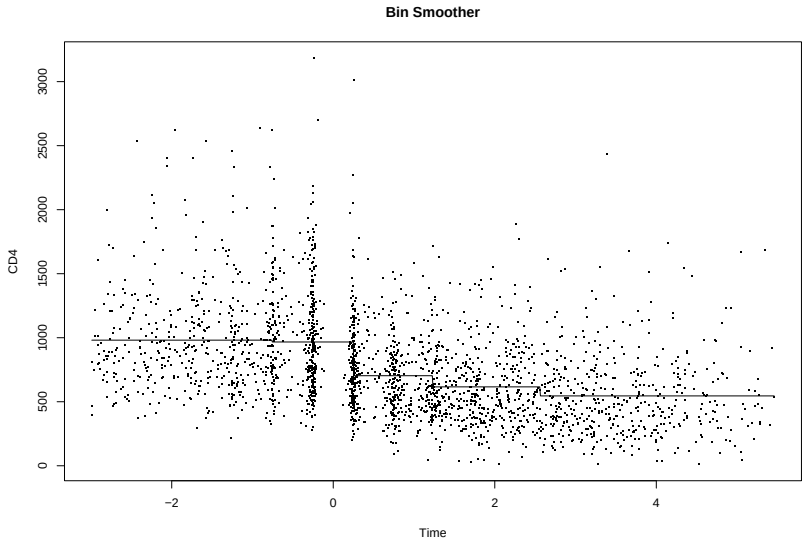
$$R_k = \{i; c_k \leq x_i < c_{k+1}\}; k = 0, \ldots, K$$

the indexes of the data points in each region. Then $S[\mathbf{y}|\mathbf{x}]$ is given by

$$s(x) = \text{ave}_{i \in R_k}\{y_i\} \text{ if } x \in R_k$$

Notice that the bin smoother will have discontinuities.

Figure 2.3: CD4 cell count since seroconversion for HIV infected men.



2.3.3 Running-mean/moving average

Since we have no replicates and we want to force $s(x)$ to be smooth we can use the motivation that under some stastical model, for any x_0 values of $f(x) = \mathrm{E}[Y|X = x]$ for x close to x_0 are similar.

How do we define close? A formal definition is the *symmetric nearest neighborhood*

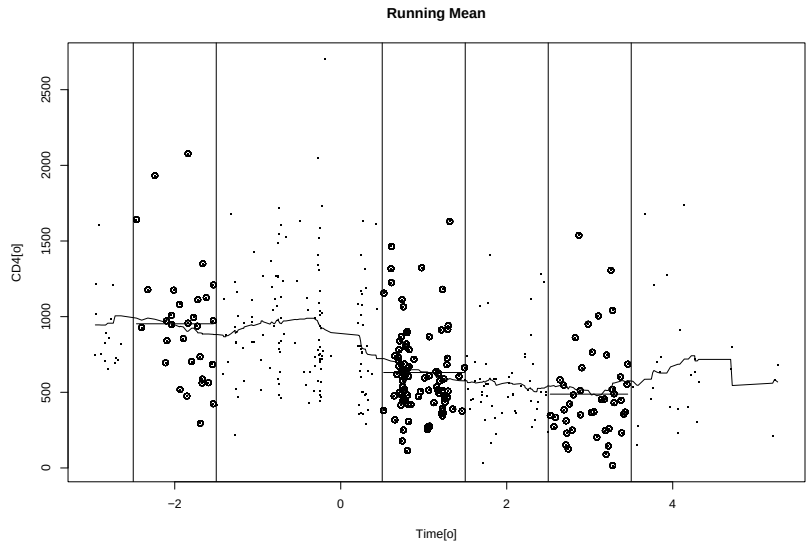
$$N^S(x_i) = \{\max(i-k,1), \ldots, i-1, i, i+1, \min(i+k,n)\}$$

We may now define running mean as:

$$s(x_i) = \text{ave}_{j \in N^S(x_i)}\{y_j\}$$

We can also forget about the symmetric part and simply define the nearest k neighbors.

Figure 2.4: CD4 cell count since seroconversion for HIV infected men.



This usually too wiggly to be considered useful. Why do you think?

Notice we can also fit a line instead of a constant. This procedure is called running-line.

Can you write out the recipe for $s(x_i)$ for the running-line smoother?

2.3.4 Kernel smoothers

One of the reasons why the previous smoothers is wiggly is because when we move from x_i to x_{i+1} two points are usually changed in the group we average. If the new two points are very different then $s(x_i)$ and $s(x_{i+1})$ may be quite different. One way to try and fix this is by making the transition smoother. That's the idea behind kernel smoothers.

Generally speaking a kernel smoother defines a set of weights $\{W_i(x)\}_{i=1}^n$ for each x and defines

$$s(x) = \sum_{i=1}^{n} W_i(x) y_i.$$

We will see that most scatter plot smoothers can be considered to be kernel smoothers in this very general definition.

What is called a kernel smoother in practice has a simple approach to represent the weight sequence $\{W_i(x)\}_{i=1}^n$ by describing the shape of the weight function $W_i(x)$ by a density function with a scale parameter that adjusts the size and the form of the weights near x. It is common to refer to this shape function as a *kernel* K. The kernel is a continuous, bounded, and symmetric real function K which integrates to one,

$$\int K(u)\, du = 1.$$

For a given scale parameter h, the weight sequence is then defined by

$$W_{hi}(x) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}$$

Notice: $\sum_{i=1}^{n} W_{hi}(x_i) = 1$

The kernel smoother is then defined for any x as before by

$$s(x) = \sum_{i=1}^{n} W_{hi}(x) Y_i.$$

Notice: if we consider x and y to be observations of random variables X and Y then one can get an intuition for why this would work because

$$E[Y|X] = \int y f_{X,Y}(x,y)\, dy / f_X(x),$$

with $f_X(x)$ the marginal distribution of X and $f_{X,Y}(x,y)$ the joint distribution of (X,Y), and

$$s(x) = \frac{n^{-1}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) y_i}{n^{-1}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}$$

Because we think points that are close together are similar, a kernel smoother usually defines weights that decrease in a smooth fashion as one moves away from the target point.

Running mean smoothers are kernel smoothers that use a "box" kernel. A natural candidate for K is the standard Gaussian density. (This is very inconvenient computationally because its never 0). This smooth is shown in Figure 2.5 for $h = 1$ year.

In Figure 2.6 we can see the weight sequence for the box and Gaussian kernels for three values of x.