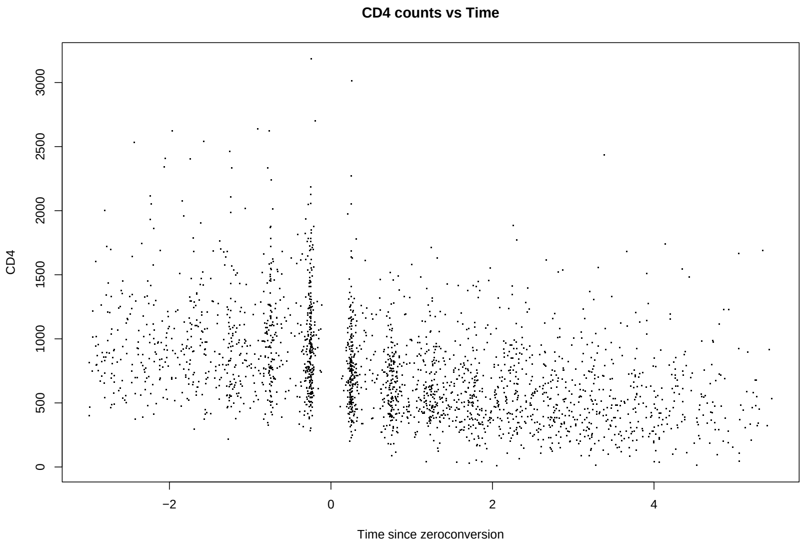
Chapter 2

Overview of various smoothers

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 $\mathbf{x} = (x_1, \dots, x_n)'$ 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 considered 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 $\{s(x_1), \dots, s(x_n)\}'$ as *the smooth*.

Here is a stupid example: If we assume a random desing model and take expectations over the empirical distribution $\hat{F}$, defined by the observations, we have for any $x_0 \in \{x_1, \dots, x_n\}$,

$$E_{\hat{F}}[Y|X = x_0] = \text{ave}\{y_i; x_i = x_0\}.$$

Define $s(x_0) = \mathrm{E}_{\hat{F}}[Y|X = x_0]$. 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, $s(x_0) = y_0$ for all x_0. 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.

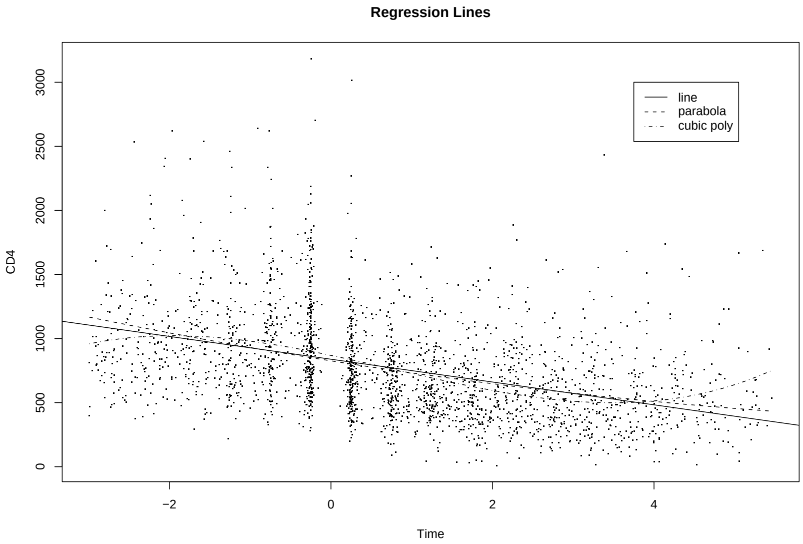
2.1 Parametric smoother

These are what you have seen already. We force a function defined by “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 $S[\mathbf{y}|\mathbf{x}](x_0) = (1\; x_0)\; (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ with $\mathbf{X}$ a design matrix containing a column of 1’s and $\mathbf{x}$ (`cbind(1,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.



2.2 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

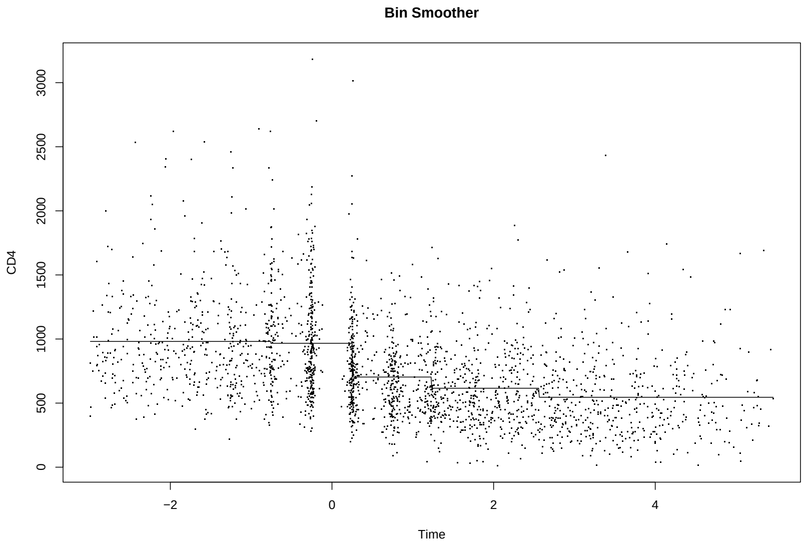
$$R_k = \{i; c_k \leq x_i < c_{k+1}\}; k = 0, \ldots, K$$

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

$$s(x_0) = \text{ave}_{i \in R_k}\{y_i\} \text{ if } x_0 \in R_k$$

Notice that the bin smoother will have discontinuities.

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



2.3 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*

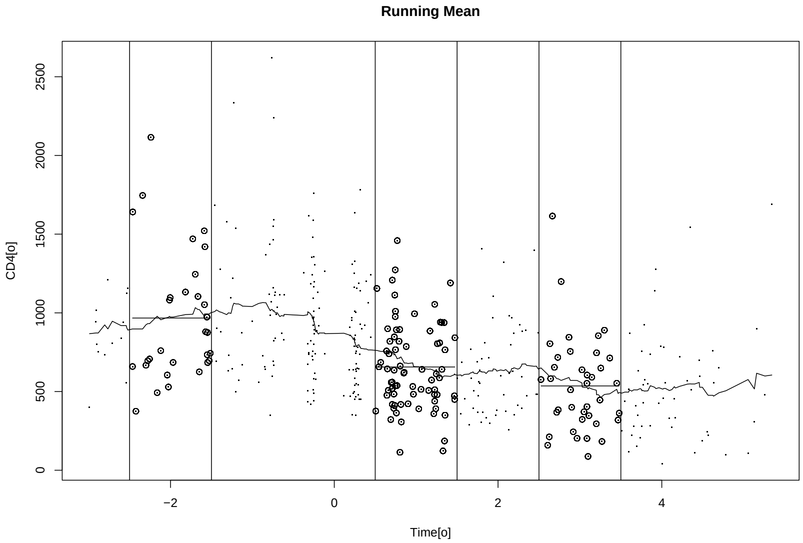
$$N^S(x_i) = \{\max(i-k,1),\dots,i-1,i,i+1,\min(i+k,n)\}$$

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?

2.4 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.

Figure 2.5: CD4 cell count since seroconversion for HIV infected men.

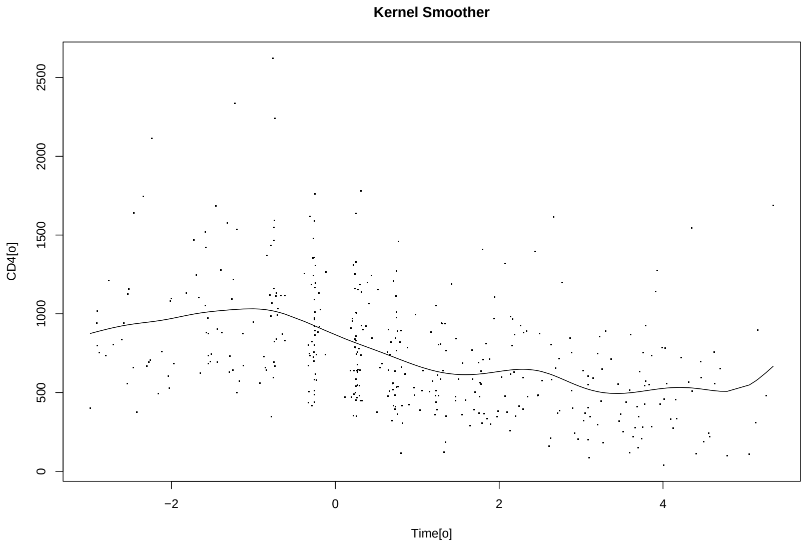
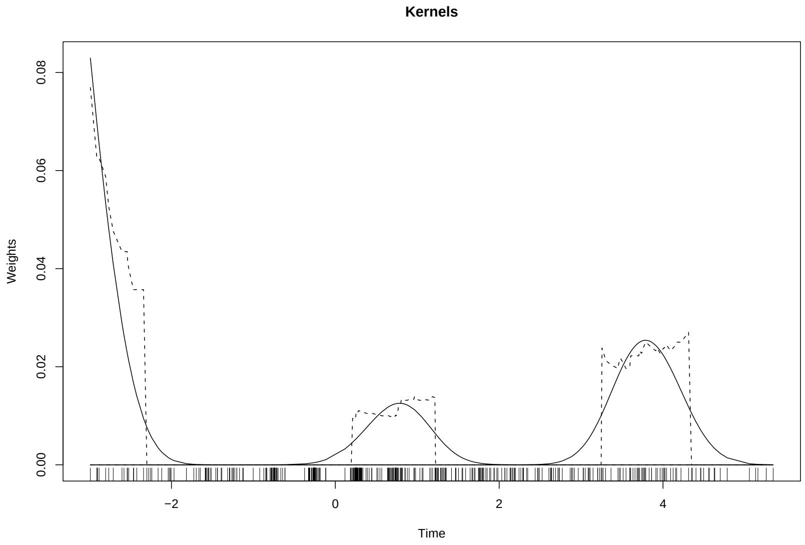


Figure 2.6: CD4 cell count since seroconversion for HIV infected men.



2.4.1 An Asymptotic result

For the asymptotic theory presneted here we will assume the stochastic design model with a one-dimensional covariate.

For the first time in this Chapter we will set down a specific stochastic model. Assume we have n IID observations of the random variables (X, Y) and that

$$Y_i = f(X_i) + \epsilon_i, i = 1, \dots, n \tag{2.1}$$

where X has marginal distribution $f_X(x)$ and the ϵ_i IID errors independent of the X. A common extra assumption is that the errors are normally distributed. We are now going to let n go to infinity... What does that mean?

For each n we define an estimate for $f(x)$ using the kernel smoother with scale parameter h_n.

Theorem 1 *Under the following assumptions*

1. $\int |K(u)| \, du < \infty$
2. $\lim_{|u| \to \infty} uK(u) = 0$
3. $E(Y^2) \leq \infty$
4. $n \to \infty, h_n \to 0, nh_n \to \infty$

Then, at every point of continuity of $f(x)$ and $f_X(x)$ we have

$$\frac{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) y_i}{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)} \to f(x) \text{ in probability.}$$

Proof: Homework. Hint: Start by proving the fixed design model.

2.5 Linear smoothers

Most of the smoother presented here are linear smoothers which means that the fit at any point x_0 can be written as

$$s(x) = \sum_{j=1}^{n} S_j(x) y_j.$$

In practice we usually have the model

$$Y_i = f(X_i) + \epsilon_i$$

and we have observations $\{(x_i, y_i)\}$. Many times it is the vector $\mathbf{f} = \{f(x_1), \dots, f(x_n)\}'$ we are after. In this case the vector of estimates $\hat{\mathbf{f}} = \{\hat{f}(x_1), \dots, \hat{f}(x_n)\}'$ can be written as

$$\hat{\mathbf{f}} = \mathbf{S}\mathbf{y}$$

with $\mathbf{S}$ a matrix with the i,j-th entry $S_j(x_i)$. We will call $\hat{\mathbf{f}}$ the *smooth*.

This makes it easy to figure out things like the variance of $\hat{\mathbf{f}}$ since

$$\text{var}[\mathbf{S}\mathbf{y}] = \mathbf{S}\text{var}[\mathbf{y}]\mathbf{S}'$$

which in the case of IID data is $\sigma^2\mathbf{S}\mathbf{S}'$.