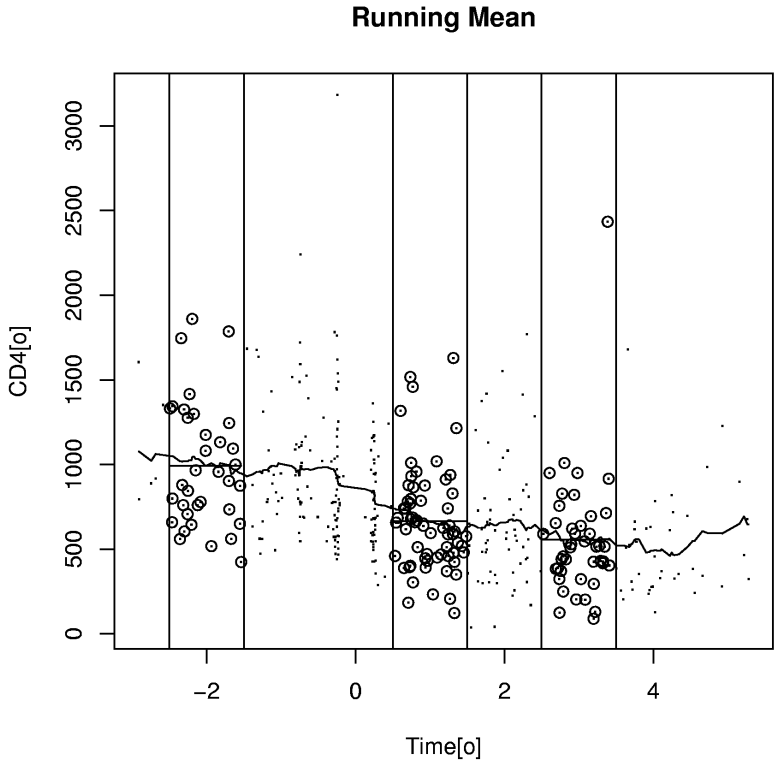
Chapter 6

Kernel Methods

Below is the results of using running mean (K nearest neighbor) to estimate the effect of time to zero conversion on CD4 cell count.

One of the reasons why the running mean (seen in Figure 6.1) 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 one of the main goals of kernel smoothers.

Figure 6.1: Running mean estimate: CD4 cell count since seroconversion for HIV infected men.



6.1 Kernel Smoothers

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

$$\hat{f}(x) = \sum_{i=1}^{n} W_i(x) y_i. \tag{6.1}$$

Most smoothers can be considered to be kernel smoothers in this very general definition. However, 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 $W_i(x)$ via 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

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

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 6.2 for $h = 1$ year.

In Figure 6.3 we can see the weight sequence for the box and Gaussian kernels for three values of x.

Figure 6.2: CD4 cell count since seroconversion for HIV infected men.

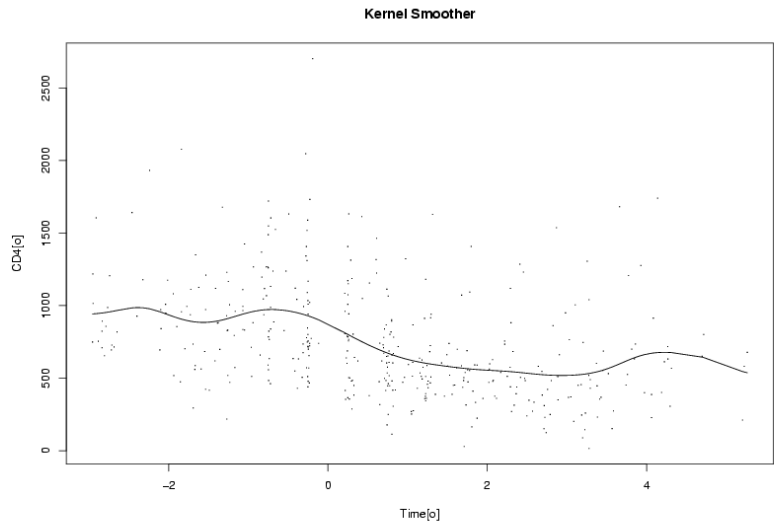
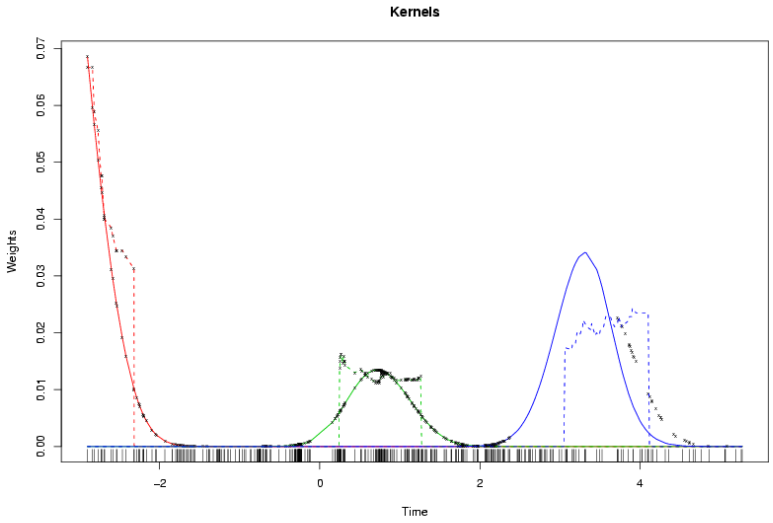


Figure 6.3: CD4 cell count since seroconversion for HIV infected men.



6.1.1 Technical Note: 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

$$Y_i = f(X_i) + \varepsilon_i, i = 1, \ldots, n \tag{6.2}$$

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: Optional homework. Hint: Start by proving the fixed design model.

6.2 Local Regression

Local regression is used to model a relation between a predictor variable and response variable. To keep things simple we will consider the fixed design model. We assume a model of the form

$$Y_i = f(x_i) + \varepsilon_i$$

where $f(x)$ is an unknown function and ε_i is an error term, representing random errors in the observations or variability from sources not included in the x_i.

We assume the errors ε_i are IID with mean 0 and finite variance var$(\varepsilon_i) = \sigma^2$.

We make no global assumptions about the function f but assume that locally it can be well approximated with a member of a simple class of parametric function, e.g. a constant or straight line. Taylor's theorem says that any continuous function can be approximated with polynomial.

6.2.1 Techinical note: Taylor's theorem

We are going to show three forms of Taylor's theorem.

- This is the original. Suppose f is a real function on $[a, b]$, $f^{(K-1)}$ is continuous on $[a, b]$, $f^{(K)}(t)$ is bounded for $t \in (a, b)$ then for any distinct points $x_0 < x_1$ in $[a, b]$ there exist a point x between $x_0 < x < x_1$ such that

$$f(x_1) = f(x_0) + \sum_{k=1}^{K-1} \frac{f^{(k)}(x_0)}{k!}(x_1 - x_0)^k + \frac{f^{(K)}(x)}{K!}(x_1 - x_0)^K.$$

  Notice: if we view $f(x_0) + \sum_{k=1}^{K-1} \frac{f^{(k)}(x_0)}{k!}(x_1 - x_0)^k$ as function of x_1, it's a polynomial in the family of polynomials

$$\mathcal{P}_{K+1} = \{f(x) = a_0 + a_1 x + \ldots + a_K x^K, (a_0, \ldots, a_K)' \in \mathbb{R}^{K+1}\}.$$

- Statistician sometimes use what is called Young's form of Taylor's Theorem:

  Let f be such that $f^{(K)}(x_0)$ is bounded for x_0 then

$$f(x) = f(x_0) + \sum_{k=1}^{K} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + o(|x - x_0|^K), \text{ as } |x - x_0| \to 0.$$

  Notice: again the first two term of the right hand side is in $\mathcal{P}_{K+1}$.

- In some of the asymptotic theory presented in this class we are going to use another refinement of Taylor's theorem called Jackson's Inequality:

Suppose f is a real function on $[a, b]$ with K is continuous derivatives then

$$\min_{g \in \mathcal{P}_k} \sup_{x \in [a,b]} |g(x) - f(x)| \leq C \left(\frac{b-a}{2k} \right)^K$$

with $\mathcal{P}_k$ the linear space of polynomials of degree k.

6.2.2 Fitting local polynomials

Local weighter regression, or loess, or lowess, is one of the most popular smoothing procedures. It is a type of kernel smoother. The default algorithm for loess adds an extra step to avoid the negative effect of influential outliers.

We will now define the recipe to obtain a loess smooth for a target covariate x_0.

The first step in loess is to define a weight function (similar to the kernel K we defined for kernel smoothers). For computational and theoretical purposes we will define this weight function so that only values within a *smoothing window* $[x_0 + h(x_0), x_0 - h(x_0)]$ will be considered in the estimate of $f(x_0)$.

Notice: In local regression $h(x_0)$ is called the span or bandwidth. It is like the kernel smoother scale parameter h. As will be seen a bit later, in local regression, the span may depend on the target covariate x_0.

This is easily achieved by considering weight functions that are 0 outside of

$[-1, 1]$. For example Tukey's tri-weight function

$$W(u) = \begin{cases} (1 - |u|^3)^3 & |u| \leq 1 \\ 0 & |u| > 1. \end{cases}$$

The weight sequence is then easily defined by

$$w_i(x_0) = W\left(\frac{x_i - x_0}{h(x)}\right)$$

We define a window by a procedure similar to the k nearest points. We want to include $\alpha \times 100\%$ of the data.

Within the smoothing window, $f(x)$ is approximated by a polynomial. For example, a quadratic approximation

$$f(x) \approx \beta_0 + \beta_1(x - x_0) + \frac{1}{2}\beta_2(x - x_0)^2 \text{ for } x \in [x_0 - h(x_0), x_0 + h(x_0)].$$

For continuous function, Taylor's theorem tells us something about how good an approximation this is.

To obtain the local regression estimate $\hat{f}(x_0)$ we simply find the $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ that minimizes

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^3} \sum_{i=1}^{n} w_i(x_0)[Y_i - \{\beta_0 + \beta_1(x_i - x_0) + \frac{1}{2}\beta_2(x_i - x_0)\}]^2$$

and define $\hat{f}(x_0) = \hat{\beta}_0$.

Notice that the Kernel smoother is a special case of local regression. Proving this is a Homework problem.

6.2.3 Defining the span

In practice, it is quite common to have the x_i irregularly spaced. If we have a fixed span h then one may have local estimates based on many points and others is very few. For this reason we may want to consider a nearest neighbor strategy to define a span for each target covariate x_0.

Define $\Delta_i(x_0) = |x_0 - x_i|$, let $\Delta_{(i)}(x_0)$ be the ordered values of such distances. One of the arguments in the local regression function `loess()` (available in the modreg library) is the `span`. A span of α means that for each local fit we want to use $\alpha \times 100\%$ of the data.

Let q be equal to αn truncated to an integer. Then we define the span $h(x_0) = \Delta_{(q)}(x_0)$. As α increases the estimate becomes smoother.

In Figures 6.4 – 6.6 we see loess smooths for the CD4 cell count data using spans of 0.05, 0.25, 0.75, and 0.95. The smooth presented in the Figures are fitting a constant, line, and parabola respectively.

Figure 6.4: CD4 cell count since seroconversion for HIV infected men.

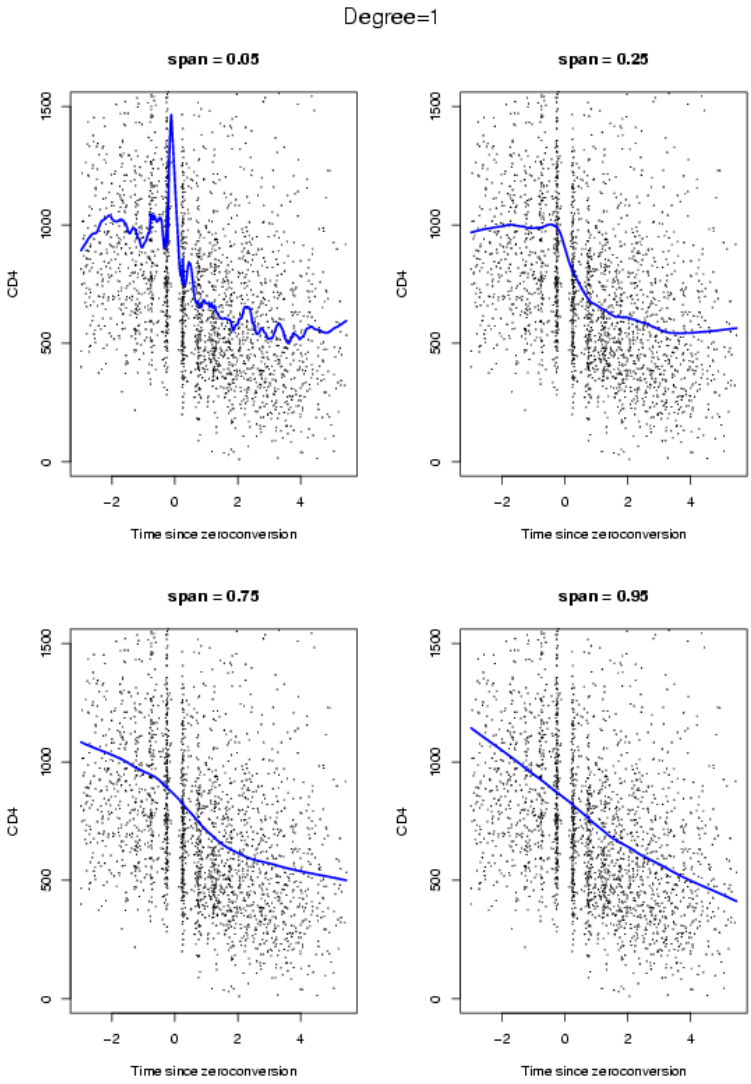


Figure 6.5: CD4 cell count since seroconversion for HIV infected men.

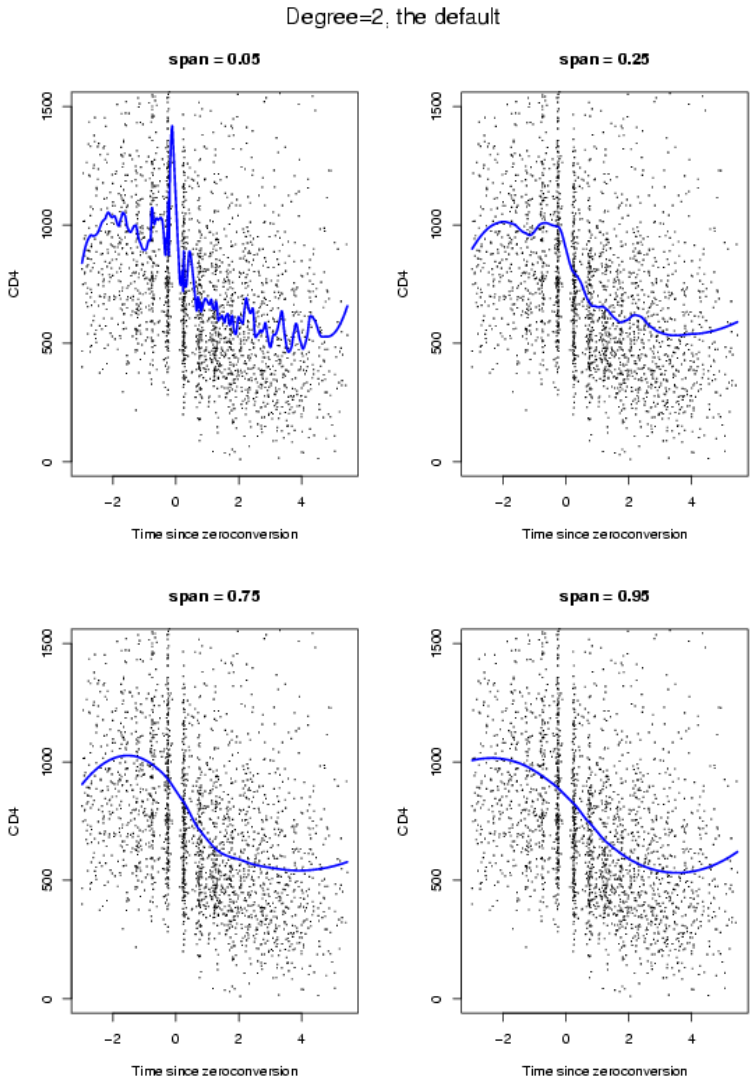
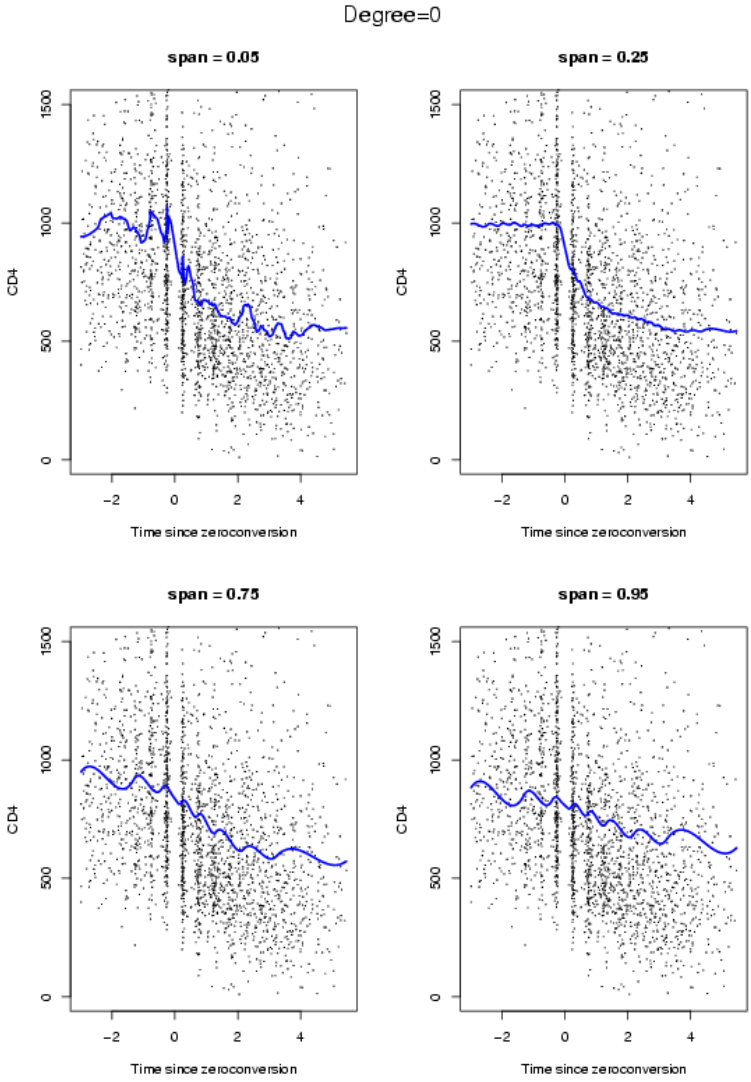


Figure 6.6: CD4 cell count since seroconversion for HIV infected men.



6.2.4 Symmetric errors and Robust fitting

If the errors have a symmetric distribution (with long tails), or if there appears to be outliers we can use robust loess.

We begin with the estimate described above $\hat{f}(x)$. The residuals

$$\hat{\varepsilon}_i = y_i - \hat{f}(x_i)$$

are computed.

Let

$$B(u;b) = \begin{cases} \{1-(u/b)^2\}^2 & |u| < b \\ 0 & |u| \geq b \end{cases}$$

be the bisquare weight function. Let m = median($|\hat{\varepsilon}_i|$). The robust weights are

$$r_i = B(\hat{\varepsilon}_i; 6m)$$

The local regression is repeated but with new weights $r_i w_i(x)$. The robust estimate is the result of repeating the procedure several times.

If we believe the variance var$(\varepsilon_i) = a_i\sigma^2$ we could also use this double-weight procedure with $r_i = 1/a_i$.

6.2.5 Multivariate Local Regression

Because Taylor's theorems also applies to multidimensional functions it is relatively straight forward to extend local regression to cases where we have more

than one covariate. For example if we have a regression model for two covariates

$$Y_i = f(x_{i1}, x_{i2}) + \varepsilon_i$$

with $f(x, y)$ unknown. Around a target point $\mathbf{x}_0 = (x_{01}, x_{02})$ a local quadratic approximation is now

$$f(x_1, x_2) \approx \beta_0 + \beta_1(x_1 - x_{01}) + \beta_2(x_2 - x_{02}) + \beta_3(x_1 - x_{01})(x_2 - x_{02}) + \frac{1}{2}\beta_4(x_1 - x_{01})^2 + \frac{1}{2}\beta_5(x_2 - x_{02})^2$$

Once we define a distance, between a point $\mathbf{x}$ and $\mathbf{x}_0$, and a span h we can define define waits as in the previous sections:

$$w_i(\mathbf{x}_0) = W\left(\frac{||\mathbf{x}_i, \mathbf{x}_0||}{h}\right).$$

It makes sense to re-scale x_1 and x_2 so we smooth the same way in both directions. This can be done through the distance function, for example by defining a distance for the space $\mathbb{R}^d$ with

$$||\mathbf{x}||^2 = \sum_{j=1}^{d} (x_j / v_j)^2$$

with v_j a scale for dimension j. A natural choice for these v_j are the standard deviation of the covariates.

Notice: We have not talked about k-nearest neighbors. As we will see in Chapter VII the *curse of dimensionality* will make this hard.

6.2.6 Example

We look at part of the data obtained from a study by Socket et. al. (1987) on the factors affecting patterns of insulin-dependent diabetes mellitus in children. The objective was to investigate the dependence of the level of serum C-peptide on various other factors in order to understand the patterns of residual insulin secretion. The response measurement is the logarithm of C-peptide concentration (pmol/ml) at diagnosis, and the predictors are age and base deficit, a measure of acidity. In Figure 6.7 we show a loess two dimensional smooth. Notice that the effect of age is clearly non-linear.

6.3 Linear Smoothers: Influence, Variance, and Degrees of Freedom

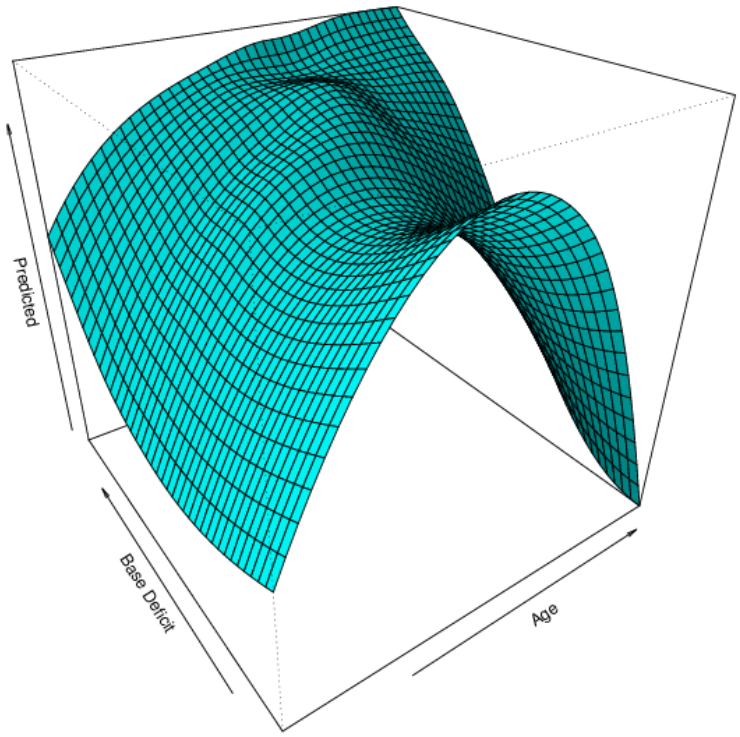
All the smoothers presented in this course are linear smoothers. This means that we can think of them as version of Kernel smoothers because every estimate $\hat{f}(x)$ is a linear combination of the data Y thus we can write it in the form of equation (6.1).

If we forget about estimating f at every possible x and consider only the observerd (or design) points $x_1, \ldots, x_n$, we can write equation (6.1) as

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

Here $\mathbf{f} = \{f(x_1), \ldots, f(x_n)\}$ and $\mathbf{S}$ is defined by the algorithm we are using.

Figure 6.7: Loess fit for predicting C.Peptide from Base.deficit and Age.



Question: What is $\mathbf{S}$ for linear regression? How about for the kernel smoother defined above?

How can we characterize the amount of smoothing being performed? The smoothing parameters provide a characterization, but it is not ideal because it does not permit us to compare between different smoothers and for smoothers like loess it does not take into account the shape of the weight function nor the degree of the polynomial being fit.

We now use the connections between smoothing and multivariate linear regression (they are both linear smoothers) to characterize pointwise criteria that characterize the amount of smoothing at a single point and global criteria that characterize the global amount of smoothing.

We will define variance reduction, influence, and degrees of freedom for linear smoothers.

The variance of the interpolation estimate is $\text{var}[Y_1] = \sigma^2$. The variance of our smooth estimate is

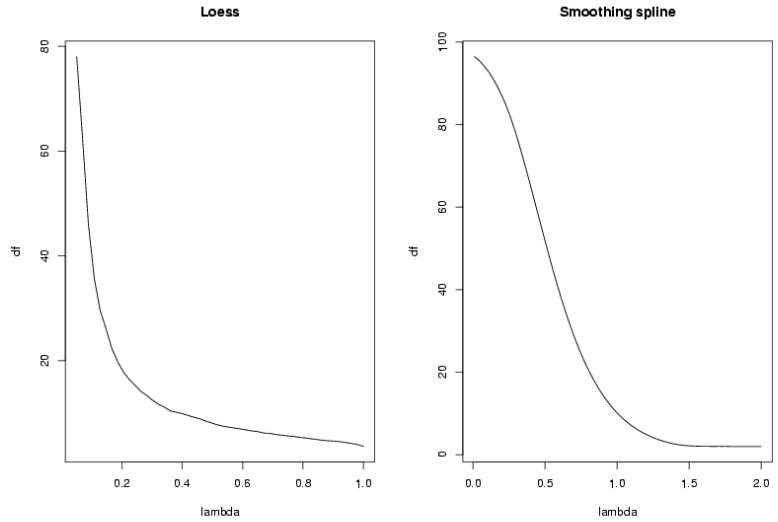
$$\text{var}[\hat{f}(x)] = \sigma^2 \sum_{i=1}^{n} W_i^2(x)$$

so we define $\sum_{i=1}^{n} W_i^2(x)$ as the variance reduction. Under mild conditions one can show that this is less than 1.

Because

$$\sum_{i=1}^{n} \text{var}[\hat{f}(x_i)] = \text{tr}(\mathbf{S}\mathbf{S}')\sigma^2,$$

Figure 6.8: Degrees of freedom for loess and smoothing splines as functions of the smoothing parameter. We define smoothing splines in a later lecture.



the total variance reduction from $\sum_{i=1}^{n} \text{var}[Y_i]$ is $\text{tr}(\mathbf{SS}')/n$.

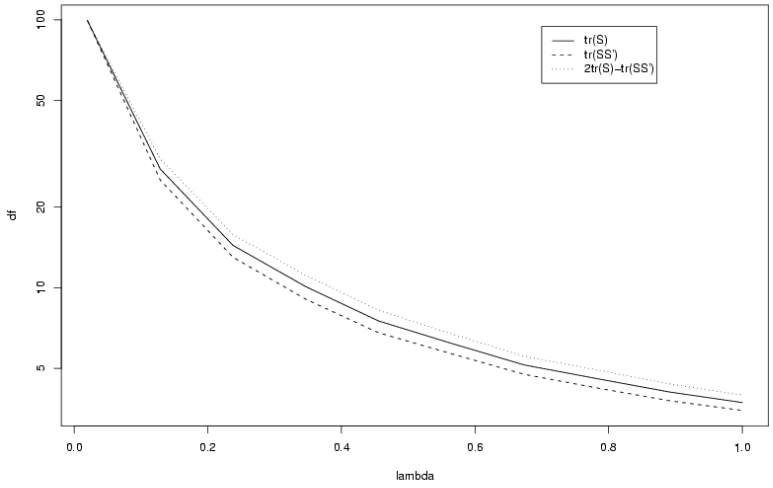
In linear regression the variance reduction is related to the degrees of freedom, or number of parameters. For linear regression, $\sum_{i=1}^{n} \text{var}[\hat{f}(x_i)] = p\sigma^2$. One widely used definition of degrees of freedoms for smoothers is $df = \text{tr}(\mathbf{SS}')$.

The sensitivity of the fitted value, say $\hat{f}(x_i)$, to the data point y_i can be measured by $W_i(x_i)/\sum_{i=1}^{n} W_n(x_i)$ or $\mathbf{S}_{ii}$ (remember the denominator is usually 1).

The total influence or sensitivity is $\sum_{i=1}^{n} W_i(x_i) = \text{tr}(\mathbf{S})$.

In linear regression $\text{tr}(\mathbf{S}) = p$ is also equivalent to the degrees of freedom. This is also used as a definition of degrees of freedom.

Figure 6.9: Comparison of three definition of degrees of freedom



Finally we notice that

$$\text{E}[(\mathbf{y} - \hat{\mathbf{f}})'(\mathbf{y} - \hat{\mathbf{f}})] = \{n - 2\text{tr}(\mathbf{S}) + \text{tr}(\mathbf{S}\mathbf{S}')\}\sigma^2$$

In the linear regression case this is $(n-p)\sigma^2$. We therefore denote $n - 2\text{tr}(\mathbf{S}) + \text{tr}(\mathbf{S}\mathbf{S}')$ as the residual degrees of freedom. A third definition of degrees of freedom of a smoother is then $2\text{tr}(\mathbf{S}) - \text{tr}(\mathbf{S}\mathbf{S}')$.

Under relatively mild assumptions we can show that

$$1 \leq \text{tr}(\mathbf{S}\mathbf{S}') \leq \text{tr}(\mathbf{S}) \leq 2\text{tr}(\mathbf{S}) - \text{tr}(\mathbf{S}\mathbf{S}') \leq n$$

6.4 Local Likelihood

The model $Y = f(X) + \epsilon$ does not always make sense because of the additive error term. An simple example is count data. Kernel methods are available for such situations. The idea is to model the likelihood of Y as a smooth function of X.

Suppose we have independent observation s $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ that are the realization of a response random variable Y given a $P \times 1$ covariate vector $\mathbf{x}$ which we consider to be known. Given the covariate $\mathbf{x}$, the response variable Y follows a parametric distribution $Y \sim g(y; \theta)$ where θ is a function of $\mathbf{x}$. We are interested in estimating θ using the observed data.

The log-likelihood function can be written as

$$l(\theta_1, \ldots, \theta_n) = \sum_{i=1}^{n} \log g(y_i; \theta_i) \tag{6.3}$$

where $\theta_i = s(\mathbf{x}_i)$. A standard modeling procedure would assume a parsimonious form for the θ_is, say $\theta_i = \mathbf{x}_i'\boldsymbol{\beta}$, $\boldsymbol{\beta}$ a $P \times 1$ parameter vector. In this case the log-likelihood $l(\theta_1, \ldots, \theta_n)$ would be a function of the parameter $\boldsymbol{\beta}$ that could be estimated by maximum likelihood, that is by finding the $\hat{\boldsymbol{\beta}}$ that maximizes $l(\theta_1, \ldots, \theta_n)$.

The local likelihood approach is based on a more general assumption, namely that $s(\mathbf{x})$ is a "smooth" function of the covariate $\mathbf{x}$. Without more restrictive assumptions, the maximum likelihood estimate of $\boldsymbol{\theta} = \{s(\mathbf{x}_1), \ldots, s(\mathbf{x}_n)\}$ is no longer useful because of over-fitting. Notice for example that for the case of

regression with all the $\mathbf{x}_i$s distinct the maximum likelihood estimate would simply reproduce the data.

Suppose we are interested in estimating only $\theta_0 = \theta(\mathbf{x}_0)$ for a fixed covariate value $\mathbf{x}_0$. The local likelihood estimation approach is to assume that there is some neighborhood N_0 of covariates that are "close" enough to $\mathbf{x}_0$ such that the data $\{(\mathbf{x}_i, y_i); \mathbf{x}_i \in N_0\}$ contain information about θ_0 through some *link function* η of the form

$$\theta_0 = s(\mathbf{x}_0) \equiv \eta(\mathbf{x}_0, \boldsymbol{\beta}) \text{ and} \tag{6.4}$$

$$\theta_i = s(\mathbf{x}_i) \approx \eta(\mathbf{x}_i, \boldsymbol{\beta}), \text{ for } \mathbf{x}_i \in N_0. \tag{6.5}$$

Notice that we are abusing notation here since we are considering a different $\boldsymbol{\beta}$ for every $\mathbf{x}_0$. Throughout the work we will be acting as if θ_0 is the only parameter of interest and therefore not indexing variables that depend on the choice of $\mathbf{x}_0$.

The local likelihood estimate of θ_0 is obtained by assuming that, for data in N_0, the true distribution of the data, $g(y_i; \theta_i)$ is approximated by

$$f(y_i; \mathbf{x}_i, \boldsymbol{\beta}) \equiv g(y_i; \eta(\mathbf{x}_i, \boldsymbol{\beta})), \tag{6.6}$$

finding the $\hat{\boldsymbol{\beta}}$ maximizes the local log-likelihood equation

$$l_0(\boldsymbol{\beta}) = \sum_{\mathbf{x}_i \in N_0} w_i \log f(y_i; \boldsymbol{\beta}), \tag{6.7}$$

and then using Equation (6.4) to obtain the local likelihood estimate $\hat{\theta}_0$. Here w_i is a weight coefficient related to the "distance" between $\mathbf{x}_0$ and $\mathbf{x}_i$. In order to obtain a useful estimate of θ_0, we need $\boldsymbol{\beta}$ to be of "small" enough dimension so that we fit a parsimonious model within N_0.

Hastie and Tibshirani (1987) discuss the case where the covariate $\mathbf{x}$ is a real valued scalar and the link function is linear

$$\eta(x_i, \boldsymbol{\beta}) = \beta_0 + x_i\beta_1$$

Notice that in this case, the assumption being made is that the parameter function $s(x_i)$ is approximately linear within "small" neighborhoods of x_0, i.e. locally linear.

Staniswalis (1989) presents a similar approach. In this case the covariate $\mathbf{x}$ is allowed to be a vector, and the link function is a constant

$$\eta(\mathbf{x}_i, \beta) = \beta$$

The assumption being made here is that the parameter function $s(x_i)$ is locally constant.

If we assumes a density function of the form

$$\log g(y_i; \theta_i) = C + (y_i - \theta_i)^2/\phi \tag{6.8}$$

where K and ϕ are constants that do not depend on the θ_is, local regression may be considered a special case of local likelihood estimation.

Notice that in this case the local likelihood estimate is going to be equivalent to the estimate obtained by minimizing a sum of squares equation. The approach in Cleveland (1979) and Cleveland and Devlin (1988) is to consider a real valued covariate and the polynomial link function

$$\eta(\mathbf{x}_i, \boldsymbol{\beta}) = \sum_{j=0}^{d} x_i^j \beta_j.$$

In general, the approach of local likelihood estimation, including the three abovementioned examples, is to assume that for "small" neighborhoods around $\mathbf{x}_0$, the distribution of the data is approximated by a distribution that depends on a constant parameter $\boldsymbol{\beta}(\mathbf{x}_0)$, i.e. we have locally parsimonious models. This allows us to use the usual estimation technique of maximum likelihood. However, in the local version of maximum likelihood we often have an a priori belief that points "closer" to $\mathbf{x}_0$ contain more information about θ_0, which suggest a weighted approach.

The asymptotic theory presented in, for example, Staniswalis (1989) and Loader (1986)is developed under the assumption that as the size (or radius) of some neighborhood of the covariate of interest $\mathbf{x}_0$ tends to 0, the difference between the true and approximating distributions within such neighborhood becomes negligible. Furthermore, we assume that despite the fact that the neighborhoods become arbitrarily small, the number of data points in the neighborhood somehow tends to ∞. The idea is that, asymptotically, the behavior of the data within a given neighborhood, is like the one assumed in classical asymptotic theory for non-IID data: The small window size assure that the difference between the true and approximating models is negligible and the large number of independent observations is available to estimate a parameter of fixed dimension that completely specifies the joint distribution. This concept motivates the approach taken in the following sections to derive a model selection criteria.

6.5 Kernel Density Estimation

Suppose we have a random sample $x_1, \ldots, x_n$ drawn from a probability density $f_X(x)$ and we wish to estimate f_X at a point x_0.

A natural local estimate is the histogram:

$$\hat{f}_X(x_0) = \frac{\#x_i \in \mathcal{N}(x_0)}{N\lambda}$$

where $\mathcal{N}(x_0$ is a small metric neighborhood around x_0 of width λ.

This estimate is bumpy and a kernel version usually works better:

$$\hat{f}_X(x_0) = \frac{1}{N\lambda} \sum_{i=1} K_\lambda(x_0, x_1)$$

Notice this is just like the scatter-smoother expect all $y = 1$. This intuitive because we are counting occurences of xs.

6.5.1 Kernel Density Classification

If we are able to estimate the densities of the covariates within each class, then we can try to estimate Bayes rule directly:

$$\Pr(G = j|X = x_0) = \frac{\hat{\pi}_j \hat{f}_j(x_0)}{\sum_{k=1}^{J} \hat{\pi}_k \hat{f}_k(x_0)}$$

The problem with this classifier is that if the x are multivariate the density estimation becomes unstable. Furthemore, we do not really need to know the densities to form a classifier. Knowing the likelihood ratios between all classes is enough. The *Naive Bayes Classifier* uses this fact to form a succesful classifier.

6.5.2 Naive Bayes Classifier

This classifier assumes, not only that the outcomes are independent but, that the covariates are independent. This implies that

$$f_j(X) = \prod_{k=1}^{p} f_{jk}(X_k)$$

This implies that we can write

$$\begin{aligned} \log \frac{\Pr(G = l|X)}{\Pr(G = J|X)} &= \log \frac{\pi_l f_l(X)}{\pi_j f_j(X)} \\ &= \log \frac{\pi_l \prod_{k=1}^{p} f_{lk}(X)}{\pi_j \prod_{k=1}^{p} f_{jk}(X)} \\ &= \log \frac{\pi_l}{\pi_j} + \log \frac{\prod_{k=1}^{p} f_{lk}(X)}{\prod_{k=1}^{p} f_{jk}(X)} \end{aligned}$$

$$\equiv \quad \alpha_l + \sum_{k=1}^{p} g_{lk}(X_k)$$

This has the form of a generalized additive model (which we will discuss latter) and can be fitted stabely.

6.5.3 Mixture Models for Density Estimators

The mixture model is a useful tool for density estimators and can be viewed as a kind of kernel mehtod. The Gaussian mixture model has the form

$$f(x) = \sum_{i=1}^{M} \alpha_m \phi(x; \mu_m, \mathbf{\Sigma}_m)$$

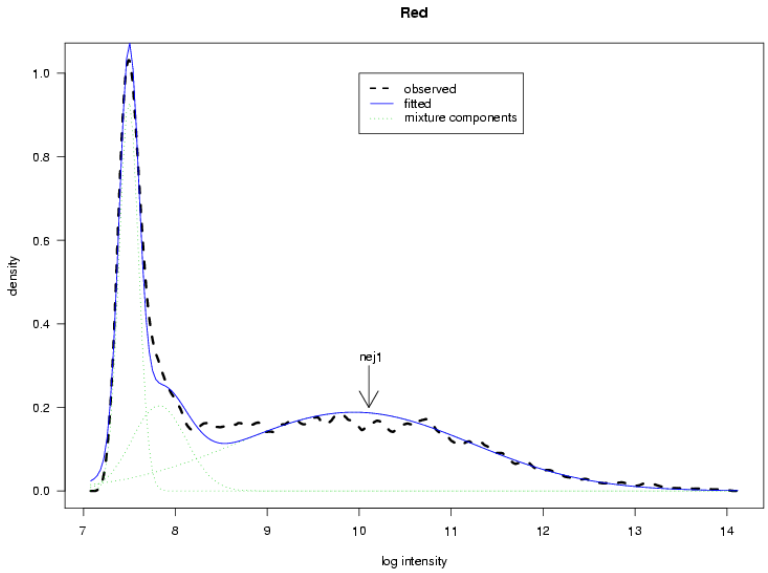
with the *mixing proportions* α_m adding to 1, $\sum_{m=1}^{M} \alpha_m = 1$. Here ϕ represents the Gaussian density.

These turn out to be quite flexible and not too many parameters are needed to get good fits.

To estiamte the parameters we use the EM algorithm which is not very fast, but is very stable.

Figure 6.10 provides an example.

Figure 6.10: Loess fit for predicting C.Peptide from Base.deficit and Age.



Bibliography

[1] Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. (1990). Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6:3–33.

[2] Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83:596–610.

[3] Cleveland, W. S., Grosse, E., and Shyu, W. M. (1993). Local regression models. In Chambers, J. M. and Hastie, T. J., editors, *Statistical Models in S*, chapter 8, pages 309–376. Chapman & Hall, New York.

[4] Loader, C. R. (1999), *Local Regression and Likelihood*, New York: Springer.

[5] Socket, E.B., Daneman, D. Clarson, C., and Ehrich, R.M. (1987). Factors affecting and patterns of residual insulin secretion during the first year of type I (insulin dependent) diabetes mellitus in children. *Diabetes* 30, 453–459.

[6] Loader, C. R. (1996), "Local likelihood density estimation," *The Annals of Statistics*, 24, 1602–1618.

[7] Loader, C. R. (1999), *Local Regression and Likelihood*, New York: Springer.

[8] Staniswalis, J. G. (1989), "The kernel estimate of a regression function in likelihood-based models," *Journal of the American Statistical Association*, 84, 276–283.

[9] Tibshirani, R. and Hastie, T. (1987), "Local likelihood estimation," *Journal of the American Statistical Association*, 82, 559–567.