*Hint:* See Appendix B.2.

3.4 In this exercise, generalized cross validation is used to compare conditionally parametric fits with bivariate smooth fits for the ethanol dataset.

a) Make a GCV plot for the model `NOx~E+C`, with `scale=0`. Use smoothing parameters ranging from 0.25 to 0.8.

b) Repeat for the conditionally parametric model `NOx~E+cpar(C)`. Use both the conditionally quadratic (the default) and conditionally linear, by setting `deg=1`. Compare the results.

c) `scale=0` is equivalent to `scale=c(0.204,3.932)` (the sample standard deviations). Compute the GCV plot for other scale parameters, such as `scale=c(0.204,8)`. The conditionally parametric fit is obtained as the second component tends to infinity.

3.5 This exercise compares asymptotic and finite sample approximations to the local regression variance.

a) Generate a sample with $n = 50$, with $x_i$ sampled i.i.d. from the standard normal distribution. Also generate a sample $Y_i \sim N(0,1)$ (the mean function doesn't matter for this exercise).

b) Compute a local linear fit, with constant bandwidth $h = 1$. Plot the standard deviation $\|l(x)\|$ using the LOCFIT command `plot(fit,what="nlx")`. Compute and plot the asymptotic approximation (2.39). Note that

$$\int W(v)^2 dv / (\int W(v) dv)^2 = 175/247$$

for the tricube weight function. Remember the square root!

c) Repeat using a nearest neighbor bandwidth with $\alpha = 0.7$. When computing the asymptotic variance, approximate the nearest neighbor bandwidth by $h(x) \approx \alpha/(2f(x))$.

d) Repeat this exercise using two predictor variables, with both components i.i.d. $N(0,1)$.

# 4

# Local Likelihood Estimation

Generalized linear models (McCullagh and Nelder 1989) provide a generalization of linear regression to likelihood models, for example, when the responses are binary or Poisson count data. Fitting of smooth likelihood models dates to Henderson (1924b), who fitted penalized likelihood models to binary data. This paper, although rarely referred to in modern literature, is particularly noteworthy as it was one of the earliest works on likelihood based regression models.

In this chapter a local likelihood approach is used. This was first proposed in Brillinger (1977) and studied in detail by Tibshirani (1984), Tibshirani and Hastie (1987) and Staniswalis (1989) among others. The local likelihood model is described in Section 4.1. Section 4.2 discusses fitting with LOCFIT. Section 4.3 introduces diagnostic procedures for local likelihood models, including residuals and model assessment criteria. Section 4.4 presents some theoretical results for local likelihood, including existence of the estimates and approximations to the bias and variance.

## 4.1 The Local Likelihood Model

The likelihood regression model assumes response variables have a density

$$Y_i \sim f(y, \theta_i)$$

where $\theta_i = \theta(x_i)$ is a function of the covariates $x_i$. Examples include the exponential distribution with mean $\theta$,

$$f(y,\theta) = \frac{1}{\theta} e^{-y/\theta} I_{[0,\infty)}(y)$$

and the discrete Bernoulli distribution with parameter $p$,

$$f(0,p) = 1-p; \qquad f(1,p) = p.$$

Let $l(y,\theta) = \log(f(y,\theta))$. The global log-likelihood of a parameter vector $\theta = (\theta(x_1),\ldots,\theta(x_n))$ is

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(Y_i, \theta(x_i)). \tag{4.1}$$

A generalized linear model assumes $\theta(x)$ has a parametric linear form; for example, $\theta(x) = a_0 + a_1 x$. The local likelihood model no longer assumes a parametric form but fits a polynomial model locally within a smoothing window. The local polynomial log-likelihood is

$$\mathcal{L}_x(a) = \sum_{i=1}^{n} w_i(x) l(Y_i, \langle a, A(x_i - x)\rangle). \tag{4.2}$$

Maximizing over the parameter $a$ leads to the local likelihood estimate.

**Definition 4.1 (Local Likelihood Estimate)** Let $\hat{a}$ be the maximizer of the local likelihood (4.2). The local likelihood estimate of $\theta(x)$ is

$$\hat{\theta}(x) = \langle \hat{a}, A(0)\rangle = \hat{a}_0.$$

**Example 4.1.** (Local Logistic Regression.) Consider the Bernoulli regression model, where

$$P(Y_i = 1) = p(x_i); \qquad P(Y_i = 0) = 1 - p(x_i).$$

The log-likelihood is

$$\begin{aligned} l(Y_i, p(x_i)) &= Y_i \log(p(x_i)) + (1 - Y_i)\log(1 - p(x_i)) \\ &= Y_i \log\left(\frac{p(x_i)}{1-p(x_i)}\right) + \log(1 - p(x_i)). \end{aligned}$$

A local polynomial approximation could be used for $p(x_i)$. But this isn't necessarily a good idea, since $0 \le p(x_i) \le 1$, while polynomials have no such constraints. Instead, the interval $(0,1)$ is mapped to $(-\infty,\infty)$ using the logistic link function

$$\theta(x) = \log\left(\frac{p(x)}{1-p(x)}\right).$$



Correspondingly, the local polynomial log-likelihood is

$$\mathcal{L}_x(a) = \sum_{i=1}^{n} w_i(x)\left(Y_i\,\langle a, A(x_i - x)\rangle - \log(1 + e^{\langle a, A(x_i - x)\rangle})\right).$$

The local polynomial estimate is $\hat{\theta}(x) = \hat{a}_0$. To estimate $p(x)$, the link function is inverted:

$$\hat{p}(x) = \frac{e^{\hat{\theta}(x)}}{1 + e^{\hat{\theta}(x)}}.$$

**Definition 4.2 (Link Function)** Suppose $f(y,\theta)$ is a parametric family of distributions, with mean

$$\mu = \mu(\theta) = E_\theta(Y).$$

Suppose further that $\mu(\theta)$ is 1-1. The link function is the inverse mapping of this relation; that is, the function $g(\cdot)$ satisfying

$$\theta = g(\mu).$$

The local likelihood estimate of $\mu(x)$ is

$$\hat{\mu}(x) = g^{-1}(\hat{\theta}(x)).$$

In parametric regression models, the choice of link function is largely dictated by the data. If the true mean is log-linear, one has to use the log link. With local regression models, one does not assume the model is globally correct, so the choice of link can be driven more by convenience. One compelling requirement, used to motivate the logistic link in Example 4.1, is that the parameter space for $\theta(x)$ be $(-\infty,\infty)$. For non-negative parameters, the log link is often a natural choice. Another requirement is that $l(y,\theta)$ be concave. This helps ensure stability of the local likelihood algorithm; see Section 4.4.

The variance stabilizing link satisfies

$$-\mathrm{E}\frac{\partial^2}{\partial\theta^2} l(Y,\theta)$$

is constant, independent of the parameter $\theta$. When the link satisfies this property, $\mathrm{var}(\hat{\theta}(x))$ is also independent of $\theta(x)$, at least asymptotically (see Section 4.4). This property is used for confidence interval construction in Section 9.2.3.

Another link, the canonical link, has some attractive theoretical properties. An exponential family of distributions has densities of the form

$$f(y,\mu) = \exp(\tau(\mu)y - \psi(\mu))f_0(y).$$

The canonical link is $\theta = \tau(\mu)$. When a local polynomial model is used for $\theta(x)$, the local likelihood (and hence $\hat{\theta}(x)$) $\mathcal{L}_x(a)$ depends on the data only through $\sum_{i=1}^n w_i(x)A(x_i - x)Y_i$. This locally sufficient statistic simplifies theoretical calculations.

## 4.2 Local Likelihood with LOCFIT

LOCFIT supports local likelihood regression with a variety of families and link functions, as summarized in Table 4.1. By default, a Gaussian family is assumed; this is the standard local regression discussed in Chapter 2.

| | Link Function | | | | | |
|---|---|---|---|---|---|---|
| | ident | log | logit | inverse | sqrt | arcsin |
| Gaussian | d,c,v | | | | | |
| Binomial | y | | d,c | | | v |
| Poisson | y | d,c | | | v | |
| Gamma | y | d,v | | c | | |
| Geometric | y | d | | | | |
| Von Mises | d,v | | | | | |
| Cauchy | d,v | | | | | |
| Huber | d,v | | | | | |

TABLE 4.1. Supported local likelihood families and link functions: default link (d), canonical link (c), variance stabilizing link (v) and other supported links (y).

**Example 4.2.** The mine dataset consists of a single response; the number of fractures in the upper seam of coal mines. There are four predictor variables. Fitting log-linear Poisson models, Myers (1990) showed that one predictor variable (percentage of extraction from the lower seam) was highly significant, while two other predictors had some importance. Here, we use the single predictor variable `extrp` and fit using a local log-linear model. The variable selection problem is considered later.

```
> fit <- locfit(frac~extrp, data=mine, family="poisson",
+   deg=1, alpha=0.6)
> plot(fit, band="g", get.data=T)
```

The Poisson family is specified by the `family` argument. The default link is the log link (Table 4.1); the `plot()` method automatically back-transforms to display the estimated mean (Figure 4.1). The plot also shows approximate 95% pointwise confidence intervals for the mean.

The plot shows the mean initially increases, then levels off for `extrp` > 80. The confidence intervals suggest the leveling off is a real feature; the bands do not cover any curve of the form $e^{a+bx}$, and thus a log-linear model would appear inadequate for this dataset.



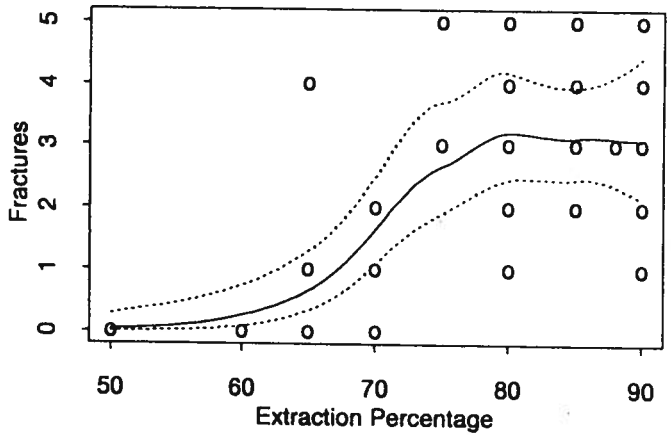


FIGURE 4.1. Mine fracture dataset: local Poisson regression.

**Example 4.3.** Mortality data of the type considered in Figure 1.1 is one example of binomial data; the observed mortality rates for each age are the number of deaths divided by the number of patients. Unfortunately, the original source for this dataset did not give the number of patients. Here, we use a second mortality dataset, from Henderson and Sheppard (1919) for which this information is available. The number of trials at each age is given as the `weights` argument to the `locfit()` call:

```
> fit <- locfit(deaths~age, weights=n, family="binomial",
+   data=morths, alpha=0.5)
> plot(fit, band="g", get.data = T)
```

Figure 4.2 displays the fit, with 95% pointwise confidence intervals. The data has been smoothed using local quadratic logistic regression, with nearest neighbor span of 0.5. This shows a gradual increasing trend, with some wild behavior at the right boundary. One must be careful when interpreting this plot because there are large differences in the weights. For ages between 70 and 80, there are as many as 150 at-risk patients, but just one for `age=99`. Likewise, there are just six patients for ages 55 and 56; this (as well as the usual boundary variability) leads to the wide confidence intervals at the left boundary.

We now define the families supported by LOCFIT. Each family is specified using the mean parameter $\mu(x_i)$. Also included is a weight parameter $n_i$, which for most families can be interpreted as a prior weight or the number of replications for each observation.

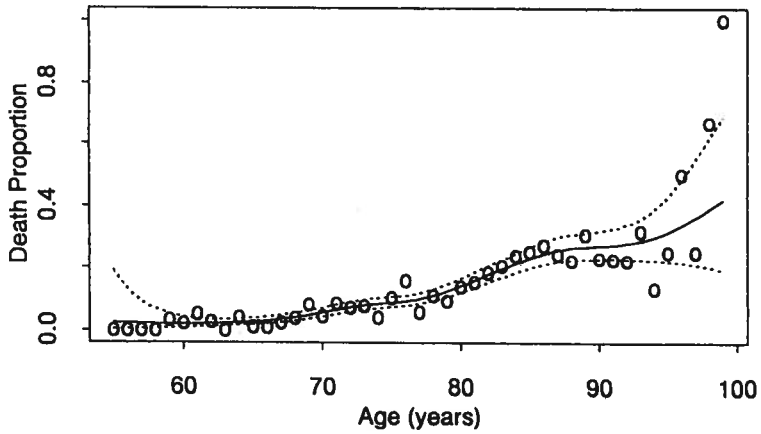


FIGURE 4.2. Local logistic regression for mortality data of Henderson and Sheppard.

The **Gaussian** family has densities

$$f_{Y_i}(y) = \frac{\sqrt{n_i}}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{n_i}{2\sigma^2}(Y_i - \mu(x_i))^2\right),$$

and the local likelihood criterion is equivalent to the local least squares criterion. Thus, `family="gauss"` produces the local regression estimate, *but assumes* $\sigma^2 = 1$. This distinction is important when constructing confidence intervals; the usual family for local regression is the quasi family `family="qgauss"`. For more discussion of this distinction, see the discussion of quasi-likelihood in Section 4.3.4.

The **binomial** family has probability mass function

$$P(Y_i = y) = \binom{n_i}{y} \mu(x_i)^y (1 - \mu(x_i))^{n_i - y}; y = 0, 1, \ldots, n_i. \tag{4.3}$$

The Bernoulli distribution ($n_i = 1$) represents the outcome of a single trial with success probability $\mu(x_i)$. The binomial distribution counts the number of successes in $n_i$ independent trials.

The **Poisson** family is used to model count data. The distribution has the mass function

$$P(Y_i = y) = \frac{(n_i\mu(x_i))^y}{y!} e^{-n_i\mu(x_i)}; y = 0, 1, 2, \ldots. \tag{4.4}$$



The **exponential** and **gamma** families (`family="gamma"`) are often used to model survival times. The gamma density function is

$$f_{Y_i}(y) = \frac{\mu(x_i)^{-n_i} y^{n_i - 1}}{\Gamma(n_i)} e^{-y/\mu(x_i)}; y \geq 0. \tag{4.5}$$

The special case $n_i = 1$ is the exponential distribution.

The **geometric** and **negative binomial** families (`family="geom"`) can be regarded as discrete analogs of the exponential and gamma distributions. The negative binomial distribution has mass function

$$P(Y_i = y) = \binom{n_i + y - 1}{n_i - 1} \frac{\mu(x_i)^y}{(1 + \mu(x_i))^{n_i + y}}; y = 0, 1, \ldots. \tag{4.6}$$

The geometric distribution is the special case $n_i = 1$. If one observes a sequence of Bernoulli trials with success probability $p(x_i) = \mu(x_i)/(1+\mu(x_i))$, the geometric distribution models the number of successes observed before a single failure. The negative binomial distribution models the number of successes until $n_i$ failures are observed.

The **von Mises** family (`family="circ"`) has densities

$$f_{Y_i}(y) = \frac{1}{I(n_i)} e^{n_i \cos(y - \mu(x_i))}; -\pi \leq y \leq \pi,$$

where $I(n_i)$ is a normalizing constant. This distribution is frequently used to model datasets where the responses are angular or measured on a circle. Regression models for $\mu(x)$ were introduced by Gould (1969). Fisher (1993) is an extensive resource for statistical methods for circular data.

Numerically, the von Mises family can be difficult to fit, since the log-likelihood has multiple local maxima. If $\hat{\mu}(x)$ is a local likelihood estimate, so is $\hat{\mu}(x) + 2\pi$. More serious problems are caused by adding a linear term. If the $x_i$ are uniform random variables (and hence irrational), some number theoretic arguments show one can come arbitrarily close to interpolation, simply by choosing a linear function with a carefully chosen large slope.

This is related to the barber's pole problem discussed by Gould (1969) and in more detail by Johnson and Wehrly (1978) and Fisher and Lee (1992), who discuss various ways of restricting $\hat{\mu}(x)$ to $[-\pi, \pi]$. None of the solutions seem entirely satisfactory, since $\mu(x)$ may genuinely have multiple circles over the range of the data. For practical purposes, the identifiability problems shouldn't create too much difficulty, unless the data is close to uniform. It also helps if the origin is chosen as a favored direction, so the estimate shouldn't skip from $-\pi$ to $\pi$.

The **Cauchy** and **Huber** families are intended mainly for local robust regression. A full description is given in Section 6.4.

## 4.3 Diagnostics for Local Likelihood

This section discusses diagnostic and model selection issues for local likelihood. Largely, the techniques are natural extensions of the local regression methodology discussed in Section 2.3. Work devoted to diagnostic issues for local likelihood includes Firth, Glosup and Hinkley (1991) and Staniswalis and Severini (1991). The methods are generally similar to techniques used in parametric generalized linear models by McCullagh and Nelder (1989).

### *4.3.1 Deviance*

In Chapter 2, we developed diagnostic methods based on the residuals $Y_i - \hat{\mu}(x_i)$, and the residual sum of squares. For local likelihood models, these tools are less natural. For example, for the gamma family (4.5), $\mu(x)$ is a scale parameter. In this case, it is more natural to consider diagnostics based on the ratio $Y_i/\hat{\mu}(x_i)$ rather than the difference $Y_i - \hat{\mu}(x_i)$.

The natural predictor of a future observation at a point $x$ is $g^{-1}(\hat{\theta}(x))$ where $g(\cdot)$ is the link function. One possible loss function is the **deviance**, for a single observation $(x, Y)$, defined by

$$D(Y, \hat{\theta}(x)) = 2\left(\sup_{\theta} l(Y,\theta) - l(Y, \hat{\theta}(x))\right).$$

It is easily seen that $D(Y,\hat{\theta}) \geq 0$, and $D(Y,\hat{\theta}) = 0$ if $Y = g^{-1}(\hat{\theta})$. Since it is based on the likelihood, the deviance provides a measure of the evidence an observation $Y$ provides *against* $\hat{\theta}(x)$ being the true value of $\theta(x)$. With a Gaussian likelihood and $\sigma = 1$, the deviance is simply the squared residual.

The total deviance is defined as

$$\sum_{i=1}^{n} D(Y_i, \hat{\theta}(x_i)). \tag{4.7}$$

This generalizes the residual sum of squares for a regression model.

**Example 4.4.** Let $Y_i$ be an observation from the gamma family with parameters $n_i$ (known) and $\mu_i$ (unknown). The log-likelihood is

$$l(Y_i, \mu_i) = -n_i \log(\mu_i) + (n_i - 1)\log(Y_i) - \frac{Y_i}{\mu_i} - \log(\Gamma(n_i)).$$

For fixed $Y_i$ and $n_i$, this is maximized at $\mu_i = Y_i/n_i$. Thus, the deviance for an estimate $\hat{\mu}_i$ is

$$D(Y_i, \hat{\mu}_i) = 2\left(-n_i \log(\frac{Y_i}{n_i\hat{\mu}_i}) + \frac{Y_i}{\hat{\mu}_i} - n_i\right).$$

As expected, this depends on $Y_i$ and $\hat{\mu}_i$ only through the ratio $Y_i/\hat{\mu}_i$. Using the Taylor series approximation $\log(x) \approx x - 1 - (x-1)^2/2$ yields

$$D(Y_i, \hat{\mu}_i) \approx \frac{1}{n_i\hat{\mu}_i^2}(Y_i - n_i\hat{\mu}_i)^2.$$



The variance of $Y_i$ is $n_i\mu_i^2$. Thus, the deviance is approximately $(Y_i - \mathrm{E}(Y_i))^2/\mathrm{var}(Y_i)$. As $n_i \to \infty$, one has the limiting distribution

$$D(Y_i, \hat{\mu}_i) \Rightarrow \chi^2_1, \tag{4.8}$$

provided $\hat{\mu}_i$ is consistent. This limiting distribution can be generalized to other likelihoods.

### *4.3.2 Residuals for Local Likelihood*

In the case of generalized linear models, a number of suitable extensions of the definition of residuals are discussed in McCullagh and Nelder (1989, section 2.4) and Hastie and Pregibon (1992, page 205). Four possible definitions are:

- Deviance residual
$$r_i = \mathrm{sign}(Y_i - \hat{\mu}_i)D(Y_i, \hat{\theta}_i)^{1/2};$$
- Pearson residual
$$r_i = \frac{Y_i - \hat{\mu}_i}{\sqrt{V_i}};$$
- Response residual
$$r_i = Y_i - \hat{\mu}_i;$$
- Likelihood derivative
$$r_i = \frac{\partial}{\partial\theta} l(Y_i, \hat{\theta}_i),$$

where $\hat{\theta}_i = \hat{\theta}(x_i)$, $\hat{\mu}_i = \hat{\mu}(x_i)$ and $V_i = \mathrm{var}(Y_i)$. For the sample residuals, these are estimated using the fitted values.

For the Gaussian likelihood, all four definitions produce the residuals $Y_i - \mu_i$. For other likelihoods, the definitions do not coincide, and all have slightly different interpretations. The Pearson residuals all have variance 1, and under the assumption $n_i \to \infty$, the residuals are asymptotically $N(0,1)$. Using (4.8), the deviance residuals have a similar property.

**Example 4.5.** We compute residuals for the mortality data of Henderson and Sheppard used in Example 4.3. The residuals are found using LOCFIT's `residuals()` function. The type of residual is specified by the `type` argument; the default is the deviance residuals:

```
> for(ty in c("deviance", "pearson", "response", "ldot")) {
+   res <- residuals(fit, type=ty)
+   plot(morths$age, res, main=ty, type="b")
+   abline(h = 0, lty = 2)
+ }
```

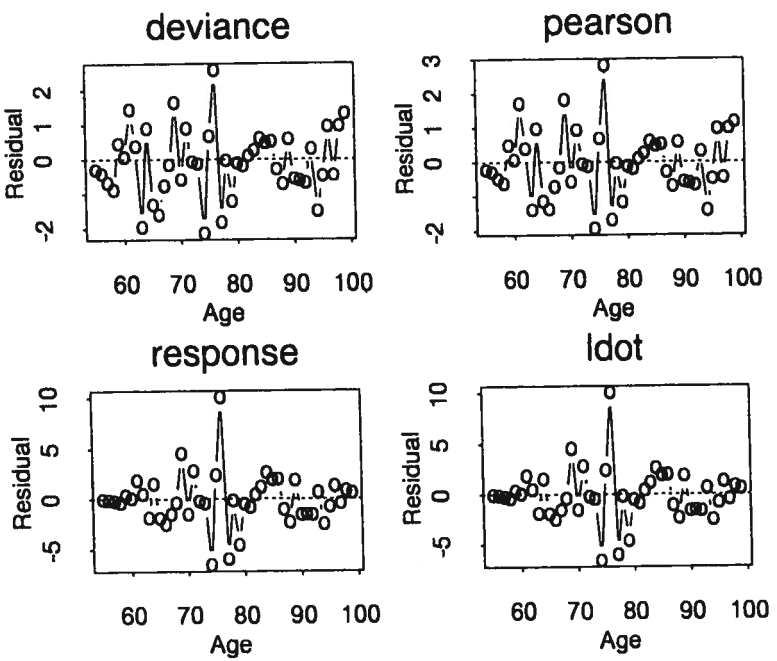


FIGURE 4.3. Residual plots for the mortality data of Henderson and Sheppard.

Figure 4.3 shows four sets of residuals plotted against age. Given the small sample sizes, there is little benefit to smoothing the residual plots, so points are simply joined by lines. No strong patterns appear in the residual plots. Both the deviance and Pearson residuals are mainly in the interval $[-2, 2]$, which indicates that the binomial model adequately models the variability of this dataset.

### *4.3.3 Cross Validation and AIC*

To help guide the choice of local likelihood model, we need extensions of the cross validation and CP methods introduced in Chapter 2. It is natural to consider methods based directly on the likelihood or deviance functions.

**Definition 4.3** The **likelihood (or deviance) cross validation** criterion is defined by substituting the leave-$x_i$-out estimate $\hat{\theta}_{-i}(x_i)$ in the total deviance (4.7);

$$\text{LCV}(\hat{\theta}) = \sum_{i=1}^{n} D(Y_i, \hat{\theta}_{-i}(x_i))$$



$$= C - 2\sum_{i=1}^{n} l(Y_i, \hat{\theta}_{-i}(x_i)) \tag{4.9}$$

where $C$ depends on the observations $Y_i$, but not the estimate $\hat{\theta}(x)$ and hence not the bandwidth or local polynomial degree.

Computation of the $n$ leave-$x_i$-out estimates can be expensive. An alternative to deletion methods is the method of infinitesimal perturbations, developed in Cook (1977) for linear models, and Pregibon (1981) for logistic regression models. The technique underlies Theorem 2.2, which relates the deletion estimate $\hat{\mu}_{-i}(x_i)$ with the estimate $\hat{\mu}(x_i)$ and the influence function $\text{infl}(x_i)$.

In the local likelihood setting, the simplification of Theorem 2.2 no longer holds. Instead, approximations must be developed; details are provided in Section 4.4.3 and Exercise 4.6. First, we identify an influence function such that

$$\hat{\theta}_{-i}(x_i) \approx \hat{\theta}(x_i) - \text{infl}(x_i)\dot{l}(Y_i, \hat{\theta}(x_i)). \tag{4.10}$$

We use $\dot{l}(y, \theta)$ and $\ddot{l}(y, \theta)$ to denote the first and second partial derivatives of $l(y, \theta)$ with respect to $\theta$. Substituting (4.10) into the deviance and using a one-term Taylor series gives

$$D(Y_i, \hat{\theta}_{-i}(x_i)) \approx D(Y_i, \hat{\theta}_i(x_i)) + 2\text{infl}(x_i)\dot{l}(Y_i, \hat{\theta}(x_i))^2.$$

Summing this over all observations gives an approximation to the likelihood cross validation statistic (4.9). Since $\text{E}(\dot{l}(Y, \theta)^2) = -\text{E}(\ddot{l}(Y, \theta))$, the fitted degrees of freedom are defined as

$$\nu_1 = \sum_{i=1}^{n} \text{infl}(x_i)\text{E}(-\ddot{l}(Y_i, \hat{\theta}(x_i))).$$

This leads to a generalization of the Akaike information criterion (Akaike, 1973, 1974) to local likelihood models.

**Definition 4.4** The **Akaike information criterion** (AIC) for local likelihood is

$$\text{AIC}(\hat{\theta}) = \sum_{i=1}^{n} D(Y_i, \hat{\theta}(x_i)) + 2\nu_1 \tag{4.11}$$

where $\nu_1$ is the degrees of freedom for the local likelihood fit.

**Example 4.6.** We apply the AIC statistic to the mine dataset, sing a variety of nearest neighbor bandwidths:

```
> a <- seq(0.4, 1, by=0.05)
> plot(aicplot(frac~extrp, data=mine, family="poisson",
+   deg=1, alpha=a))
```

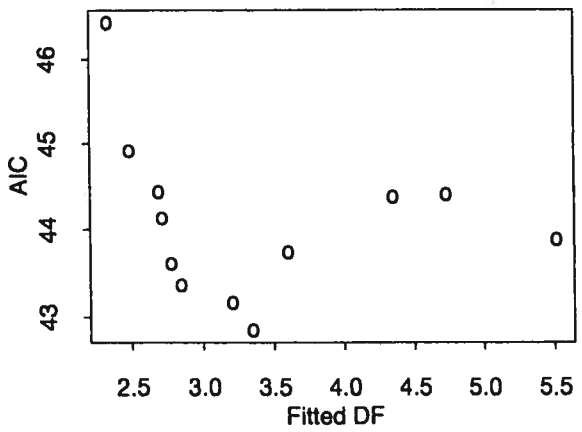


FIGURE 4.4. Akaike information criterion for the mine dataset.

The `aicplot()` function is similar to `gcvplot()` (Section 3.4.2). Figure 4.4 shows the AIC plot. The minimum AIC occurs at about 2.9 degrees of freedom ($\alpha = 0.6$). Larger smoothing parameters (i.e., smaller degrees of freedom) result in inferior fits. This provides evidence that the parametric log-linear model is inadequate for this dataset, and the curvature in Figure 4.1 is real.

### *4.3.4 Overdispersion*

If a likelihood model correctly models a dataset, the Pearson residuals defined in Section 4.3.2 should have mean 0 and variance 1. The deviance residuals are similar, using the approximation of Example 4.4. If the residuals exhibit a nonzero mean (for example, several successive residuals have the same sign), this indicates that the data is oversmoothed, and smaller bandwidths should be used.

Overdispersion occurs when the residuals have variance larger than 1. For example, the Poisson distribution has the property $\text{var}(Y_i) = \text{E}(Y_i)$. But count data often exhibit more variability than this relation can explain. The mean can still be estimated using Poisson regression, but the variance of $\hat{\mu}(x)$ may be severely underestimated.

There are several ways to handle overdispersed data. One method is through a variance stabilizing transformation, where one finds a function $g(Y)$ such that the transformed data $g(Y_i)$ has approximately constant variance. A local regression model is then fitted to the transformed data. The most commonly used family of transformations is the Box-Cox, or



power, family (Box and Cox 1964). A more sophisticated implementation is the ATS (average, transformation and smoothing) method of Cleveland, Mallows and McRae (1993), which includes a presmoothing step prior to the transformation.

Another technique is to find a family of distributions that better fits the data. For example, the negative binomial distribution (4.6) has mean $w\mu$ and variance $w\mu(1+\mu)$; in this case, the variance is always larger than the mean. One then estimates the shape parameter $w$ and fits the corresponding negative binomial model. An example using this approach is provided in Section 7.3.1.

A cleaner solution is quasi-likelihood, introduced by Wedderburn (1974); see also chapter 9 of Wedderburn (1974) and the recent book by Heyde (1997). Fan, Heckman and Wand (1995) discuss the local quasi-likelihood method. In quasi-likelihood models, one assumes a relation between the mean and variance of the observations:

$$\text{var}(Y_i) = \sigma^2 V(\mu_i)$$

where $V(\mu)$ is a known function and $\sigma^2$ is an unknown dispersion parameter. For example, under a Poisson model, one has $\text{var}(Y_i) = \mu_i$, so the quasi-Poisson model takes $V(\mu) = \mu$. Table 4.2 summarizes the variance relationships for the common families supported in LOCFIT. In `locfit()` calls, the quasi-family is obtained, for example, with the `family="qpoisson"` argument.

| Family | Variance $\sigma^2 V(\mu)$ |
|---|---|
| quasi-Gaussian | $\sigma^2$ |
| quasi-binomial | $\sigma^2\mu(1-\mu)$ |
| quasi-Poisson | $\sigma^2\mu$ |
| quasi-gamma | $\sigma^2\mu^2$ |
| quasi-geometric | $\sigma^2\mu(\mu+1)$ |

TABLE 4.2. Quasi-likelihood families and their variance functions.

Note that fitting a quasi-likelihood model is identical to fitting the corresponding likelihood model. The difference is in variance estimation: While the likelihood families assume the dispersion parameter is $\sigma^2 = 1$, the quasi-likelihood families estimate the dispersion parameter. The estimate used by LOCFIT is

$$\hat{\sigma}^2 = \frac{n}{n - 2\nu_1 + \nu_2} \frac{\sum_{i=1}^n \dot{l}(Y_i, \hat{\theta}(x_i))^2}{\sum_{i=1}^n \ddot{l}(Y_i, \hat{\theta}(x_i))}.$$

## 4.4 Theory for Local Likelihood Estimation

This section addresses some of the theoretical issues concerning local likelihood. Our emphasis is on results that have immediate practical consequences. First, we look at the motivation for maximizing the local likelihood. Then, we turn to important computational concerns and related issues such as existence and uniqueness. Finally, approximate representations for the estimate are derived; this leads to bias and variance approximations, and definitions of degrees of freedom.

### 4.4.1 *Why Maximize the Local Likelihood?*

The log-likelihood $\mathcal{L}(\theta)$, for fixed $\theta$, is a random variable, dependent on the observations $Y_1, \ldots, Y_n$. The mean $E(\mathcal{L}(\theta))$ is a function of the parameter vector $\theta$, and this mean function is maximized at the true parameter vector $\theta$. For any parameter vector $\theta^*$, Exercise 4.2 shows that

$$\mathrm{E}\left(\mathcal{L}(\theta^*)\right) \leq \mathrm{E}\left(\mathcal{L}(\theta)\right). \tag{4.12}$$

This motivates maximum likelihood: parameter values $\theta$ for which $\mathcal{L}(\theta)$ are the most likely values of $\theta$, given the observed data. Thus, among a class of candidate parameter vectors, we select the one that maximizes the empirical log-likelihood.

This maximum likelihood property extends to the local log-likelihood:

$$\mathrm{E}\left(\sum_{i=1}^{n} w_i(x) l(Y_i, \theta_i^*)\right) \leq \mathrm{E}\left(\sum_{i=1}^{n} w_i(x) l(Y_i, \theta_i)\right) \tag{4.13}$$

with equality if and only if $\theta_i^* = \theta_i$ for all $i$ with $w_i(x) > 0$. The local likelihood estimate considers candidate classes of the form $\theta_i^* = \langle a, A(x_i - x)\rangle$ and maximizes over this class of candidates.

### 4.4.2 *Local Likelihood Equations*

Assuming the likelihood is nicely behaved, the parameter vector $\hat{a}$ is a solution of the local likelihood equations

$$\sum_{i=1}^{n} w_i(x) A(x_i - x) \dot{l}(Y_i, \langle a, A(x_i - x)\rangle) = 0, \tag{4.14}$$

obtained by differentiating (4.2). In matrix notation, the local likelihood equations can be written

$$\mathbf{X}^T \mathbf{W} \dot{l}(Y, \mathbf{X}a) = 0 \tag{4.15}$$



where, as before, $\mathbf{X}$ is the design matrix and $\mathbf{W}$ is the diagon entries $w_i(x)$.

For most likelihoods, the local likelihood equations (4.14) closed form solution, and must be solved by iterative metho to two questions:

1. Does the maximizer $\hat{a}$ exist?
2. Is the maximizer $\hat{a}$ unique?

The following theorem addresses these questions for concave

**Theorem 4.1** Suppose the log-likelihood $l(y, \theta)$ is defined fc interval $(a, b)$ ($a = -\infty$ and $b = \infty$ are permitted); $l(y, \theta)$ ha derivative with respect to $\theta$ and $l(y, \theta) \to -\infty$ as $\theta \downarrow a$ or $\theta$ suppose $\mathbf{WX}$ has full column rank. Then the maximizer $\hat{a}$ isfies the local likelihood equations (4.14). If in addition $l(y,$ the solution of (4.14) is unique.

**Proof:** Let $a^{(j)}$ be a sequence of parameter estimates suc

$$\lim_{j \to \infty} \mathcal{L}_x(a^{(j)}) = \sup_a \mathcal{L}_x(a).$$

If $a^{(j)}$ has a limit point $a^*$, then by continuity, $\mathcal{L}_x(a^*) = \sup$ $a^* = \hat{a}$. Otherwise, $\|a^{(j)}\| \to \infty$; since $\mathbf{WX}$ has full rank $\theta_i^{(j)} = \langle a^{(j)}, A(x_i - x)\rangle \to \pm\infty$ for some $i$ with $w_i(x) > 0$. B is bounded above, this contradicts (4.16).

Since the parameter space is open, $\hat{a}$ lies in the interior, solution of the local likelihood equations. Differentiating (4. Jacobian matrix $-\mathbf{J}_1(\mathbf{X}a)$, where

$$\begin{aligned} \mathbf{J}_j(\theta) &= -\sum_{i=1}^{n} W\left(\frac{x_i - x}{h}\right)^j A(x_i - x) A(x_i - x)^T \ddot{l} \\ &= \mathbf{X}^T \mathbf{W}^j \mathbf{V} \mathbf{X} \end{aligned}$$

and $\mathbf{V}$ is a diagonal matrix with elements $-(\ddot{l}(Y_i, \theta_i)$. Th $\ddot{l}(Y_i, \theta_i)$ implies $\mathbf{J}_1(\theta)$ is positive definite; strictly so since rank. This implies uniqueness of $\hat{a}$.

Theorem 4.1 gives a number of conditions on the choice zation that help ensure the local likelihood estimation is well fortunately the conditions are rather restrictive; particular families. Fortunately, modifying the results for specific fam straightforward. Exercises 4.3 and 4.4 study the Poisson and ilies more closely.

### 4.4.3 Bias, Variance and Influence

Because of the nonlinear definition of $\hat{a}$, it is not possible to derive exact means and variances of $\hat{a}$; indeed, these often don't exist because of singularities that occur with small probabilities. For example, in the binomial family, there is always a positive probability that all responses are 0, in which case the local likelihood estimate does not exist. We still need distributional approximations for the local likelihood estimate, and to make headway we need approximations to the estimate itself. We should emphasize the approximations derived here depend on the *design points* $x_1, \ldots, x_n$, and not on an asymptotic design density. This is quite different from previous results in Fan, Heckman and Wand (1995) and Fan and Gijbels (1996, pages 196-197).

The results proceed in three parts. First, Theorem 4.2 establishes consistency of the local likelihood estimate. Theorem 4.3 establishes the asymptotic representation of the estimate, from which variance approximations can be derived. Theorem 4.4 derives a bias approximation using derivatives of $\theta(x)$.

**Theorem 4.2** Suppose $l(y, \theta)$ is concave, bounded and twice differentiable for all $y$. Then for either random or regular designs,

$$\mathbf{H}\hat{a} \xrightarrow{p} \begin{pmatrix} \theta(x) \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

aas $h \to 0$ and $nh^d \to \infty$. Here, $\mathbf{H}$ is a diagonal matrix of powers of $h$; $\mathbf{H}A(v/h) = A(v)$.

*Remark:* This result implies consistency of $\hat{\theta}(x) = \hat{a}_0$. It does not imply the remaining elements of $\hat{a}$ converge to 0.

**Proof:** Applying the weak law of large numbers and using the continuity of $\theta(x)$ one obtains, for any fixed vector $a$,

$$\begin{aligned} & \frac{1}{nh^d} \sum_{i=1}^{n} W\left(\frac{x_i - x}{h}\right) l(Y_i, \left\langle a, A(\frac{x_i - x}{h}) \right\rangle) \\ & \xrightarrow{p} \frac{f(x)}{h^d} \int\int W\left(\frac{u - x}{h}\right) l(y, \left\langle a, A(\frac{u - x}{h}) \right\rangle) e^{l(y, \theta(x))} dy du \\ & = \int\int W(v) l(y, \langle a, A(v) \rangle) e^{l(y, \theta(x))} dy dv \end{aligned}$$

where $f(x)$ is the design density. The left-hand side is maximized at $a = \mathbf{H}\hat{a}$, while an argument similar to (4.13) shows the right-hand side is maximized at $(\theta(x), 0, \ldots, 0)^T$. The theorem follows using convexity of the likelihood. □



The components of the vector $\hat{a}$ should estimate the coefficient of the Taylor series expansion of $\theta(\cdot)$ expanded around the fitting As a first step in obtaining an asymptotic representation, we loo discrepancy $\hat{a} - \tilde{a}$. This leads to the following result.

**Theorem 4.3** Under the conditions of Theorem 4.2,

$$\mathbf{H}(\hat{a} - \tilde{a}) = \mathbf{H}\mathbf{J}_1^{-1}\mathbf{X}^T\mathbf{W}\dot{l}(Y, \mathbf{X}\tilde{a}) + o((nh^d)^{-1/2}).$$

**Proof:** Expanding the local likelihood equations yields

$$\begin{aligned} 0 &= \mathbf{H}^{-1}\mathbf{X}^T\mathbf{W}\dot{l}(Y, \mathbf{X}\hat{a}) \\ &= \mathbf{H}^{-1}\mathbf{X}^T\mathbf{W}\dot{l}(Y, \mathbf{X}\tilde{a}) - \mathbf{H}^{-1}\mathbf{J}_1(\hat{a} - \tilde{a}) + o(nh^d\mathbf{H}(\hat{a} - \tilde{a} \end{aligned}$$

and hence

$$\mathbf{H}(\hat{a} - \tilde{a}) = \mathbf{H}\mathbf{J}_1^{-1}\mathbf{X}^T\mathbf{W}\dot{l}(Y, \mathbf{X}\tilde{a}) + o(\mathbf{H}(\hat{a} - \tilde{a})).$$

The result follows since $\mathbf{H}\mathbf{J}_1^{-1}\mathbf{X}^T\mathbf{W}\dot{l}(Y, \mathbf{X}\tilde{a})$ has size $O_p((nh^d)^{-1/}$

In Theorem 4.3, the first row of the matrix $\mathbf{J}_1^{-1}\mathbf{X}^T\mathbf{W}$ plays a rol to the weight diagram in local regression. The influence function defined to be the $i$th component of this weight diagram:

$$\text{infl}(x) = W(0)e_1^T\mathbf{J}_1^{-1}e_1.$$

This measures the sensitivity of the estimate $\hat{\theta}(x_i)$ to changes in A rather more subtle interpretation of the influence function is tl $x_i$-out cross validation approximation

$$\hat{\theta}_{-i}(x_i) = \hat{\theta}(x_i) - \text{infl}(x_i)\dot{l}(Y_i, \hat{\theta}_i);$$

see Exercise 4.6. One also obtains an approximate variance of $\hat{\theta}($ Theorem 4.3:

$$\text{vari}(x) = e_1^T\mathbf{J}_1^{-1}\mathbf{J}_2\mathbf{J}_1^{-1}e_1.$$

The fitted degrees of freedom for a local likelihood model are defi

$$\begin{aligned} \nu_1 &= \sum_{i=1}^{n} \text{infl}(x_i)v_i \\ \nu_2 &= \sum_{i=1}^{n} \text{vari}(x_i)v_i \end{aligned}$$

where $v_i = -\ddot{l}(Y_i, \theta(x_i))$. One may prefer to use $\mathrm{E}(v_i)$ in place (4.20) and the matrices $\mathbf{J}_j$, since the expected values are nonran necessarily positive, even when the log-likelihood is not concave essentially the question of observed versus expected Fisher inform parametric models, and makes little difference asymptotically.

The final step in the asymptotic representation is to identify the bias of the local likelihood estimate. This can be expressed using higher order derivatives of $\theta(x)$. The result is stated for one dimensional $x_i$; the multivariate result involves terms for all partial derivatives.

**Theorem 4.4** The first term of the bias expansion is

$$\begin{aligned}&\mathrm{E}(\mathbf{H}\mathbf{J}_1^{-1}\mathbf{X}^T\mathbf{W}\dot{l}(Y,\mathbf{X}\tilde{a}))\\ &= \frac{\theta^{(p+1)}(x)}{(p+1)!}\mathbf{H}\mathbf{J}_1^{-1}\sum_{i=1}^{n} w_i(x)(x_i-x)^{p+1}A(x_i-x)v_i + o(h^{p+1}).\end{aligned}$$

For $p \geq 1$, the second term involving $\theta^{(p+2)}$ is similar.

**Proof:** Let $\tilde{\theta}_i = \langle \tilde{a}, A(x_i - x)\rangle$. Then

$$\theta(x_i) = \tilde{\theta}_i + \frac{(x_i-x)^{p+1}}{(p+1)!}\theta^{(p+1)}(x) + O(h^{p+2})$$

uniformly on the smoothing window, and

$$\begin{aligned}\dot{l}(Y_i,\tilde{\theta}_i) &= \dot{l}(Y_i,\theta(x_i)) + (\tilde{\theta}_i - \theta(x_i))\ddot{l}(Y_i,\theta(x_i)) + O((\theta(x_i)-\tilde{\theta}_i)^2)\\ &= \dot{l}(Y_i,\theta(x_i)) - \frac{(x_i-x)^{p+1}}{(p+1)!}\theta^{(p+1)}(x)\ddot{l}(Y_i,\theta(x_i)) + O(h^{p+2}).\end{aligned}$$

Substituting into Theorem 4.3 and remembering $\mathrm{E}(\dot{l}(Y_i,\theta(x_i))) = 0$ completes the proof. □

We remark that the careful theoretical analysis of local likelihood is important. Many statistical software packages include functions for fitting generalized linear models: the `glm()` function in S-Plus, and similar functions in other packages. These functions usually allow weights for each observation, so local likelihood models can be fitted by calling `glm()` repeatedly, with a new set of weights for each fitting point. This implementation was used by Bowman and Azzalini (1997) and the associated software.

This approach produces correct estimates but incorrect inferences. The problem is that `glm()` interprets weights as a sample size; for example, the $n_i$ in (4.3). This appears as a multiplier for the $\mathbf{V}$ matrix in the Jacobian (4.17), rather than as the required $\mathbf{W}$. In particular, this implies the matrix $\mathbf{J}_2$ is computed incorrectly, and the standard errors are not correct, even asymptotically.

## 4.5 Exercises

4.1 This exercise uses the Henderson and Shepherd mortality dataset, from Example 4.3.



a) Compute a local quadratic fit, using the arcsin link. Plo and confidence intervals. Compare with Figure 4.2. Exp narrower confidence intervals near the left boundary.

b) Compute and compare AIC and LCV plots for both the and arcsin links. Use both local quadratic and local linear Which fits appear best? Does a global linear model (wit link function) appear satisfactory?

4.2 a) Prove for any $a$, $b$,

$$\log(a) \leq \log(b) + \frac{a-b}{b}.$$

b) Suppose a random variable $Y$ has density $g(y)$, and let $g$ any other density. Show that

$$\mathrm{E}(\log g^*(Y)) \leq \mathrm{E}(\log g(Y))$$

with equality if and only if $g = g^*$ almost everywhere.

c) Prove (4.12) and (4.13).

4.3 For the Poisson family, the conditions of Theorem 4.1 are not s when $Y_i = 0$ for some $i$, since $l(0,\mu) = -\mu$ is monotone.

a) Using the canonical link $\theta = \log(\mu)$, show the result of T 4.1 still holds, with the additional requirement that $\mathbf{W}$ full rank after deleting rows corresponding to $Y_i = 0$.

b) Show the existence extends to the identity and square ro Provide an example to show the estimate might not sat local likelihood equations.

4.4 For the Bernoulli family, the situation is even worse, since th hood is monotone for all observations. Using local linear fitti the logistic link, show the local likelihood estimate exists if a if no $\gamma \neq 0$ and $c$ exists for which

$$\begin{aligned}\langle \gamma, x_i\rangle &\leq c \quad \forall \quad i \text{ with } w_i(x) > 0, Y_i = 0\\ \langle \gamma, x_i\rangle &\geq c \quad \forall \quad i \text{ with } w_i(x) > 0, Y_i = 1;\end{aligned}$$

that is, no hyperplane separates the observations with $Y_i =$ those with $Y_i = 1$.

4.5 Consider Bernoulli trials $(x_i, Y_i)$ with $Y_i \in \{0,1\}$ and repli values. The dataset can be smoothed directly using logistic re or replicated $x$ values pooled to form a new dataset $(x_j^*, n_j, Y_j$ $n_j$ as the `weights` argument.

a) If the same bandwidths are used for each dataset, show the same estimate results. Also show the influence function is the same for each dataset.

b) Show the likelihood cross validation scores for the two datasets are unequal, so that minimizing $\text{LCV}(\hat{\theta})$ may yield two different answers. Show $\text{AIC}(\hat{\theta})$ is the same, up to an additive constant (independent of $\hat{\theta}$).

4.6 This exercise develops the method of infinitesimal perturbations and derives the approximation (4.10). Consider the local likelihood estimate at a point $x = x_i$ and the modified local likelihood equations

$$\mathbf{X}^T\mathbf{W}\dot{l}(Y, \mathbf{X}a) - \lambda W(0)e_1\dot{l}(Y_i, \langle a, A(0)\rangle) = 0$$

where $\lambda$ is a parameter and the solution is $\hat{a}(\lambda)$.

a) Show $\hat{a}(0)$ is the full local likelihood parameter estimate, while $\hat{a}(1)$ is the leave-$x_i$-out parameter estimate.

b) Show

$$\left.\frac{\partial\hat{a}(\lambda)}{\partial\lambda}\right|_{\lambda=0} = \mathbf{J}^{-1}e_1W(0)\dot{l}(Y_i, \hat{\theta}(x_i)).$$

c) Conclude, to a first order approximation, that

$$\hat{\theta}_{-i}(x_i) \approx \hat{\theta}(x_i) - \text{infl}(x_i)\dot{l}(Y_i, \hat{\theta}(x_i)),$$

and hence

$$\text{LCV}(\hat{\theta}) \approx \sum_{i=1}^{n} D(Y_i, \hat{\theta}(x_i)) + 2\sum_{i=1}^{n}\text{infl}(x_i)\dot{l}(Y_i, \hat{\theta}(x_i))^2.$$

# 5

# Density Estimation

Suppose observations $X_1, \ldots, X_n$ have an unknown density $f(x)$.
sity estimation problem is to estimate $f(x)$.

The histogram is a density estimate, where the $x$ space is div
bins, and counts of the data are provided for each bin. This is a si
intuitive approach, but it has problems for continuous data. H
choose the bins, and where should they be placed? A discrete h
may smooth out important features in the data.

This chapter studies an adaptation of the local likelihood metho
sity estimation. Section 5.1 derives the estimate. Section 5.2 des
implementation, using the LOCFIT software. Section 5.3 introdu
nostic methods such as residual plots and AIC. The more technic
5.4 studies theoretical properties for the local likelihood estimate

## 5.1 Local Likelihood Density Estimation

An extension of local likelihood methods to the density estimatio
is described in Loader (1996b) and Hjort and Jones (1996). Con
log-likelihood function

$$\mathcal{L}(f) = \sum_{i=1}^{n}\log(f(X_i)) - n(\int_{\mathcal{X}} f(u)du - 1)$$

where $\mathcal{X}$ is the domain of the density. The definition (5.1) of
likelihood is unusual, with the added a penalty term $n(\int_{\mathcal{X}} f(u)du$

is a density, the penalty is 0, so (5.1) coincides with the usual log-likelihood in this case. The reason for adding the penalty to (5.1) is that $\mathcal{L}(f)$ can be treated as a likelihood for any non-negative function $f$ without imposing the constraint $\int f(x)dx = 1$. A more complete justification is given in Section 5.4.

A localized version of the log-likelihood is

$$\mathcal{L}_x(f) = \sum_{j=1}^{n} W\left(\frac{X_j - x}{h}\right) \log(f(X_j)) - n\int_{\mathcal{X}} W\left(\frac{u-x}{h}\right) f(u)du. \quad (5.2)$$

We consider a local polynomial approximation for $\log(f(u))$; $\log(f(u)) \approx \langle a, A(u-x)\rangle$ in a neighborhood of $x$. The local likelihood becomes

$$\mathcal{L}_x(a) = \sum_{j=1}^{n} W\left(\frac{X_j - x}{h}\right) \langle a, A(X_j - x)\rangle$$
$$-n\int_{\mathcal{X}} W\left(\frac{u-x}{h}\right) \exp(\langle a, A(u-x)\rangle)du. \quad (5.3)$$

**Definition 5.1** Let $\hat{a} = (\hat{a}_0, \ldots, \hat{a}_p)^T$ be the maximizer of the local log-likelihood (5.3). The **local likelihood density estimate** is defined as

$$\hat{f}(x) = \exp(\langle \hat{a}, A(0)\rangle) = \exp(\hat{a}_0). \quad (5.4)$$

Under fairly general conditions, the local parameter vector $\hat{a}$ is the solution of the system of local likelihood equations obtained by differentiating (5.3):

$$\frac{1}{n}\sum_{j=1}^{n} A(X_j - x)w_j(x)$$
$$= \int_{\mathcal{X}} A(u-x)W\left(\frac{u-x}{h}\right) e^{\langle \hat{a}, A(u-x)\rangle}du \quad (5.5)$$

where $w_j(x) = W((X_j - x)/h)$. These equations have a simple and intuitive interpretation. The left-hand side of (5.5) is simply a vector of localized sample moments up to order $p$, while the right-hand side is localized population moments using the log-polynomial density approximation. The local likelihood estimate simply matches localized sample moments with localized population moments.

**Example 5.1.** (Local Constant Fitting). When the local constant polynomial ($p = 0$) is used, (5.5) consists of the single equation

$$\frac{1}{n}\sum_{j=1}^{n} w_j(x) = \int_{\mathcal{X}} W\left(\frac{u-x}{h}\right) \exp(\hat{a}_0)du,$$



yielding the closed form for the density estimate

$$\hat{f}(x) = \exp(\hat{a}_0) = \frac{1}{nh\int W(v)dv}\sum_{j=1}^{n} w_j(x).$$

This is the kernel density estimate considered by Rosenblatt (1956), (1958) and Parzen (1962).

The kernel density estimate has been widely studied; see, for c the books by Prakasa Rao (1983), Silverman (1986), Scott (1992) ar and Jones (1995). Being based on a local constant approximation, i from the same problems as local constant regression, such as trim peaks. An additional problem occurs in the tails, since increasin widths for data sparsity can lead to severe bias. This problem wa tigated more fully by Loader (1996b), where relative efficiencies c and local log-polynomial methods were compared.

### 5.1.1 *Higher Order Kernels*

The system of equations (5.5) defining the local likelihood estim the simple moment-matching interpretation noted previously. The l matching equations can also be used with other local approxima the density. The identity link $f(u) \approx \langle a, A(u-x)\rangle$ gives the systen

$$\frac{1}{n}\sum_{j=1}^{n} A(X_j - x)w_j(x) = \int_{\mathcal{X}} A(u-x)W\left(\frac{u-x}{h}\right)\langle \hat{a}, A(u-x)\rangle du$$

with the density estimate being $\hat{f}(x) = \hat{a}_0$. Since (5.7) is a linear of equations, one can solve explicitly for $\hat{a}$ and $\hat{f}(x)$. Local approxi estimates of this type were considered in Sergeev (1979).

Some manipulation shows the solution of (5.7) can be written

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} W^*\left(\frac{X_i - x}{h}\right)$$

where $W^*(v) = \langle \beta, A(v)\rangle W(v)$ for an appropriate coefficient vector kernel $W^*(v)$ satisfies the moment conditions,

$$\int W^*(v)dv = 1$$
$$\int v^j W^*(v)dv = 0, j = 1, \ldots, p.$$

Weight functions satisfying these moment conditions are known as of order $p+1$, and were introduced by Parzen (1962). The mot

is bias reduction: If the bias of $\hat{f}(x)$ is expanded in a Taylor series, the moment conditions (5.8) ensure that the low order terms are zero. The close connection between density estimation using higher order kernels and local polynomial fitting was investigated by Lejeune and Sarda (1992).

For practical purposes, the higher order kernel estimates tend to be less satisfactory than the local likelihood approach based on (5.5). The reason is that (5.5) applies a local polynomial approximation for $\log(f(x))$ rather than $f(x)$ itself. Since $f(x)$ must be non-negative, the polynomial approximation for $\log(f(x))$ is usually better, particularly in the tails of densities.

### 5.1.2 *Poisson Process Rate Estimation*

A problem closely related to density estimation is estimating the intensity function for a point process. If $X_1, \ldots, X_N$ are the random points of a point process, the corresponding counting process is

$$Z(A) = \sum_{i=1}^{N} I(X_i \in A)$$

for any set $A$. The intensity function, $\lambda(x)$, defines the mean of $Z(A)$:

$$\mathrm{E}(Z(A)) = \int_A \lambda(x)dx. \tag{5.9}$$

A simple example of a point process is the nonhomogeneous Poisson process, where $Z(A)$ has a Poisson distribution with mean (5.9). For this process, the log-likelihood function is

$$\mathcal{L}(\lambda, N) = \sum_{i=1}^{N} \log \lambda(X_i) - \int_{\mathcal{X}} \lambda(x)dx.$$

See, for example, Cox and Lewis (1966). This differs from the likelihood (5.1) for density estimation in only one important respect: the dropping of the factor $N$ in front of the integral. The localization of the likelihood and derivation of the local likelihood equations follow similarly, and the implementation of the estimation procedure is almost identical.

### 5.1.3 *Discrete Data*

In practice, all datasets are discrete. For the types of measurements usually modeled as coming from continuous distributions, this discreteness is often at a very fine level and can be ignored. With more heavily rounded data, the discreteness becomes important, and it must be modeled using a discrete



probability mass function rather than a continuous density. Smoothability estimates of a mass function have been widely studied using methods; see, for example, Dickey (1968), Aitchison and Aitken Titterington (1980) and Simonoff (1987, 1995, 1996). The last of the considers local likelihood approaches.

A local log-likelihood for the mass function is obtained by replac integrals in (5.1) and (5.2) by sums over the mass points. Assume th points $X_1, \ldots, X_n$ are integer valued, and consider the $(j, Y_j)$ pairs, $Y_j$ is the number of observations equal to $j$. The total number of vations is $n = \sum_{-\infty}^{\infty} Y_j$. The corresponding probabilities to be est are $p(j) = P(X_1 = j)$. Using a local polynomial model for $\log(p(j$ neighborhood of a fitting point $x$, the discrete version of the local like (5.2) is

$$\begin{aligned}\mathcal{L}_x(a) &= \sum_{j=-\infty}^{\infty} W\left(\frac{j-x}{h}\right) \langle a, A(j-x)\rangle Y_j \\ &\quad -n \sum_{j=-\infty}^{\infty} W\left(\frac{j-x}{h}\right) e^{\langle a, A(j-x)\rangle}.\end{aligned}$$

This is the local likelihood (4.2), with $l(y, \mu) = y\log(\mu) - n\mu$. Exce the factor $n$, this is the Poisson log-likelihood. Thus, estimating a function is almost equivalent to a local Poisson regression. Note th on the right-hand side of (5.11) is *not* restricted to values of $j$ with $Y$

Although the close relation between discrete probability estimation, son regression and density estimation is apparent, there are importan oretical differences. The raw probability $Y_j/n$ is a $\sqrt{n}$-consistent est of $p(j)$, and, *given a sufficiently large sample*, this will be the best estimate. Thus, the large sample behavior of the continuous densit discrete probability estimates are quite different.

Discreteness also has a major impact on bandwidth selection. This will be discussed more later, but the important point is that *discret tributions do not have densities.* Thus, if a selector designed for conti data is blindly applied to discrete data, problems *should* result, as th lector will prefer densities that place a spike at each data point. Sele have to be adapted specifically to discrete data, and the result $h = 0$ is, use the raw probabilities) has to be considered a legitimate answe

## 5.2 Density Estimation in LOCFIT

In LOCFIT, density estimation corresponds to `family="density"`. family becomes the default when no left-hand side is specified in the m formula. Using `family="rate"` gives the Poisson process rate estimat

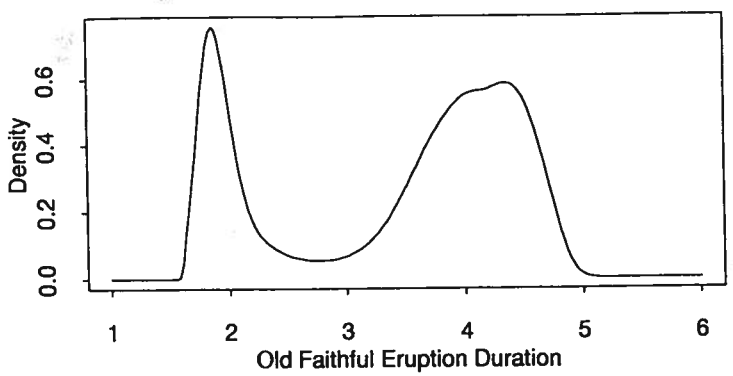


FIGURE 5.1. Density estimation for the Old Faithful geyser dataset.

**Example 5.2.** The Old Faithful geyser dataset, as given by Weisberg (1985) and Scott (1992), contains the durations of 107 eruptions. The density is estimated using a mixed smoothing parameter with a fixed component of 0.8 and nearest neighbor span of 0.1:

```
> fit <- locfit(~geyser,alpha=c(0.1,0.80),flim=c(1,6))
> plot(fit, mpv=200, xlab="Old Faithful Eruption Duration",
+   ylab="Density", get.data=T)
```

The fit is shown in Figure 5.1. This clearly shows two peaks in the data: a sharp peak around two minutes and a broader peak around 4 minutes. Note the `flim=c(1,6)` argument given to the `locfit()` call; this specifies fitting limits slightly outside the range of the data, thus allowing us to see the tails of the density. The `get.data=T` option causes the data points to be displayed as a 'rug' along the bottom of the plot, rather than the scatter plot used in the regression setting.

**Example 5.3.** The high order kernels discussed in Section 5.1.1 can be fitted using `link="ident"`. We use the fourth order kernel (local quadratic) estimate for the Old Faithful dataset:

```
> fit <- locfit(~geyser, alpha=c(0.1,0.6), flim=c(1,6),
+   link="ident")
> plot(fit, mpv=200, xlab="Old Faithful Eruption Duration",
+   ylab="Density", get.data=T)
```

The resulting fit in Figure 5.2 seems less satisfactory than that obtained previously in Figure 5.1: The estimate is not constrained to be positive,



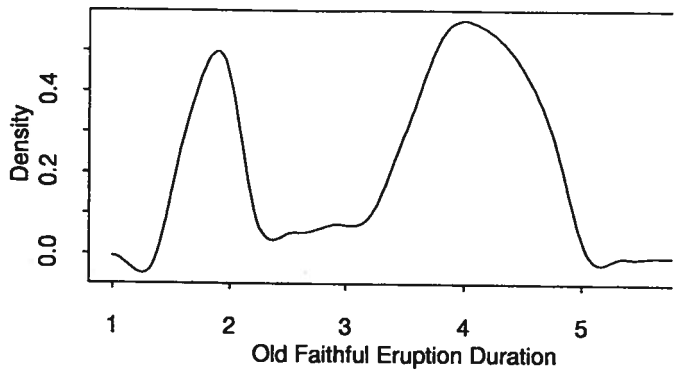


FIGURE 5.2. Local quadratic (fourth order kernel) fit to the Old Faithfi dataset.

and the method seems to oversmooth the left peak, despite the smaller bandwidth.

**Example 5.4.** Izenman and Sommer (1988) and Sheather (1992 a dataset on measurements of the thickness of 486 postage stamp 1872 Hidalgo issue of Mexico. The thicknesses are recorded to the 0.001 millemetres. This discreteness is coarse enough to matter, as when bandwidth selectors are applied to this problem (Exercise local qudratic density estimate is computed using the Poisson re model:

```
> fit <- locfit(count~thick, weights=rep(0.486,76),
+   data=stamp, family = "poisson", alpha = c(0, 0.004))
> plot(fit, m=200, get.data=T)
```

The critical point is the `weights` argument. Setting `weights=rep` effectively divides the Poisson regression by n, leading to estimatio mass function. The probability of a bin centered at a point $x_i$ is imately $n\Delta f(x_i)$ where $\Delta$ is the size of the bin and $f(x)$ is the Comparing with the Poisson family (4.4), we set the weight $n_i = n$ the mean $\mu(x_i) = f(x_i)$. In this example, $n = 486$ and $\Delta = 0.001$.

Figure 5.3 shows the resulting multimodal estimate. The explana the multimodality, provided by Izenman and Sommer (1988), is that number of different types of paper were used to print this stamp.

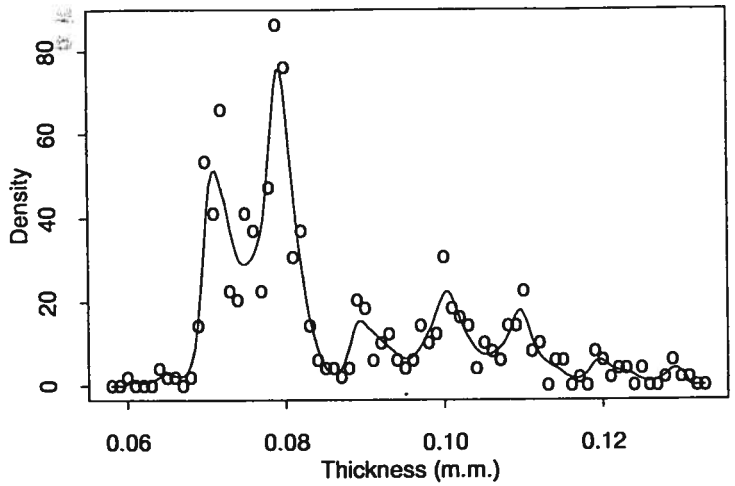


FIGURE 5.3. Postage stamp data. Density estimate using local Poisson regression for discrete data.

### 5.2.1 *Multivariate Density Examples*

Multivariate density estimation requires multiple predictor variables in the model formula, similar to the regression case in Section 3.5. In this section, some examples are presented.

**Example 5.5.** (Multivariate Density Estimation). The `trimod` dataset is a bivariate dataset with 225 observations from a trimodal distribution. Each of the three components is a bivariate standard normal distribution, with centers at $(3\sqrt{3}/2, 0), (-3\sqrt{3}/2, 3)$ and $(-3\sqrt{3}/2, -3)$. The true peak height is about $1/(6\pi) = 0.053$.

The multivariate density is estimated by specifying multiple terms on the right-hand side of the formula. Here, we fit a local log-quadratic model, with a 35% nearest neighbor bandwidth:

```
> fit.trim <- locfit(~x0+x1, data=trimod, alpha=0.35)
> plot(fit.trim, type="persp")
```

Figure 5.4 shows the fit.

A common density estimation problem is to estimate the smallest region containing a fixed probability mass. At first, constructing such a region may appear to require tricky numerical integration of the density estimate. However, a trick to estimate the contour level is to order the fitted values at the data points, and use the corresponding empirical level.



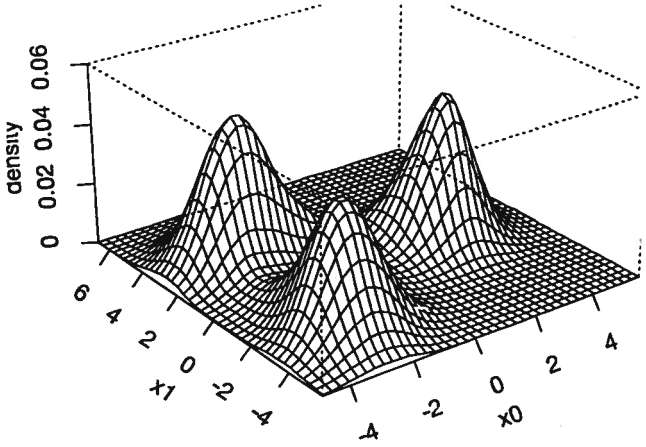


FIGURE 5.4. Bivariate density estimation.

**Example 5.6.** (Probability Contours). We compute 95% and 50 contours for the trimodal sample used in Example 5.5. First, use to compute fitted values at the data points. Then, produce a cont with the appropriate empirical contour levels:

```
> emp <- sort(fitted(fit.trim))
> plot(fit.trim, vband=F, v=emp[floor(c(0.05,0.5)*225)]
> points(trimod$x0, trimod$x1, col=2, cex=0.5)
```

Figure 5.5 shows the result. The 50% contour defines three separate and the 95% contour has a small hole in the middle.

## 5.3 Diagnostics for Density Estimation

Does the density estimate fit the data? The question of diagnostic as important for density estimation as it is for regression. But ar the question is much more difficult. The source of the problem is There is no natural definition for residuals for a density estimate, saturated model. In Section 5.3.1 some possible definitions of resid considered, along with graphical displays for detecting lack of fit. goodness of fit criteria based on the likelihood are considered in 5.3.2 and squared error methods in Section 5.3.3.

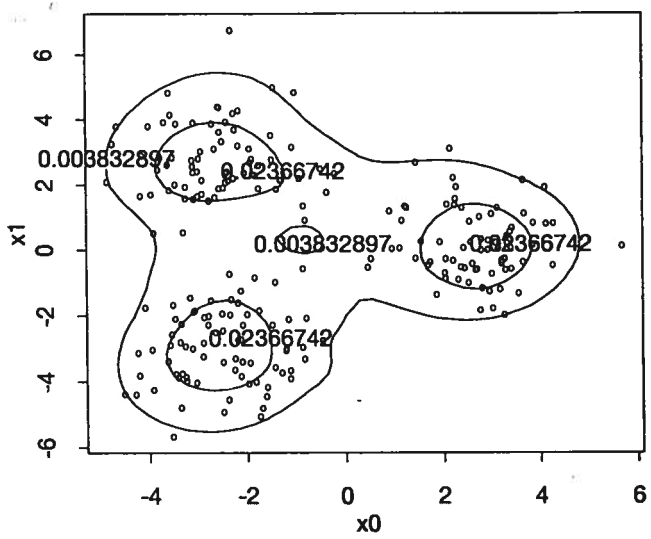


FIGURE 5.5. Probability contour plots: 50% and 95% mass contours for a trimodal example.

### 5.3.1 *Residuals for Density Estimation*

There are a number of ways to construct residual type diagnostics for density estimation. Perhaps the most obvious is to compare the integral of the density estimate,

$$\hat{F}(x) = \int_{-\infty}^{x} \hat{f}(u)du,$$

with the empirical distribution function

$$\hat{F}_{\text{emp}}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x).$$

**Example 5.7.** Figure 5.6 shows the empirical distribution function and the integral of a local density estimate. The smoothing parameter for the density estimate is $\alpha = (0.1, 1.2)$, which is larger than that used in Figure 5.1:

```
> fit <- locfit(~geyser, alpha=c(0.1,1.2),
+   flim=c(1,6), renorm=T)
> x <- seq(1, 6, by=0.01)
> z <- predict(fit, x)
```



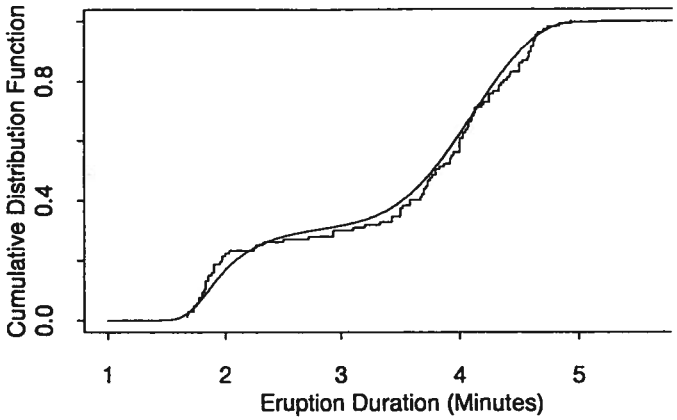


FIGURE 5.6. Empirical distribution function (step curve) and integrated estimate (smooth curve) for the Old Faithful dataset.

```
> plot(x, 0.01*cumsum(z), type="l")
> lines(sort(geyser), (1:107)/107, type="s")
```

The `renorm=T` argument rescales the density estimate so that it in to 1.

In Figure 5.6, the empirical distribution function is steeper than t mate between 1.8 and 2, which indicates that the peak has been t The flatness of the empirical distribution function between 2 and cates that the estimate has overfilled the valley.

The P-P and Q-Q plots are based $\hat{F}$ and $\hat{F}_{\text{emp}}$. The P-P (or p ity) plot uses the result that $F(X_i)$ behave like a sample from a distribution. If $X_{(i)}$ is the $i$th order statistic, then $\text{E}(F(X_{(i)})) = i/$ Thus, a plot of $\hat{F}(X_{(i)})$ against $i/(n+1)$ should be close to a strai large departures from a straight line indicate lack of fit. The Q-Q tile) plot transforms back to the observation scale, ploting $X_{(i)}$ $\hat{F}^{-1}(i/(n+1))$.

An alternative residual diagnostic for density estimation is to be a small bandwidth and look at the change in the estimate as the of smoothing is increased; can this change be attributed to noise, it indicate lack of fit? The simplest implementation of this idea i gin with a histogram, computed at a small bandwidth. Then, t histogram counts and smooth them using local Poisson regressior scribed in Section 5.1.3 and Example 5.4. One can then compute r for the Poisson model, as discussed in Section 4.3.2.

**Example 5.8.** We construct residual plots for the Old Faithful geyser dataset. First, a raw histogram of the data is constructed using a bin width of 0.05:

```
> geyser.round <- data.frame(duration=seq(1.05,5.95,by=0.05),
+   count=as.numeric(table(cut(geyser,
+     breaks=seq(1.025,5.975,length=100)))))
```

Note that care is required to ensure zeros are retained. The fit and residual plots can now be constructed:

```
> fit <- locfit(count~duration, data=geyser.round,
+   weights=rep(107*0.05,99), alpha=c(0.1,1.2),
+   family="poisson")
> plot(fit, get.data = T)
> res <- residuals(fit)
> fitr <- locfit.raw(geyser.round$duration, res, alpha=0.1)
> plot(geyser.round$duration, res)
> lines(fitr)
```

Figure 5.7 shows the fits and smoothed residual plots for three different smoothing parameters. As the smoothing parameter decreases, the fit shows the left peak getting sharper and the trough for $2 \leq$ duration $\leq 3.5$ getting deeper. The residual plots also show this: In the top residual plot, there is a pronounced peak and five successive positive residuals, around duration $= 1.8$. The residuals also show some evidence of the trough being filled in, even at smallest smoothing parameter.

### 5.3.2 *Influence, Cross Validation and AIC*

The likelihood cross validation criterion for density estimation is

$$\mathrm{LCV}(\hat{f}) = \sum_{i=1}^{n} \log \hat{f}_{-i}(X_i) - n\left(\int_{\mathcal{X}} \hat{f}(u)du - 1\right) \tag{5.12}$$

where $\hat{f}_{-i}(X_i)$ denotes the density estimate at $X_i$ when this observation is deleted from the dataset. This criterion was first proposed for the kernel density estimate (5.6) by Habbema, Hermans and Van Der Broek (1974) and Duin (1976).

As in Section 4.3.3, the likelihood cross validation score can be approximated using the method of infinitesimal perturbations. This leads to

$$\log \hat{f}_{-i}(X_i) \approx \log \hat{f}(X_i) - \frac{W(0)}{n} e_1^T \mathbf{M}_1^{-1} e_1 + \frac{1}{n} \tag{5.13}$$

where

$$\mathbf{M}_j = \int_{\mathcal{X}} W\left(\frac{u-x}{h}\right)^j A(u-x)A(u-x)^T e^{\langle \hat{a}, A(u-x)\rangle} du.$$



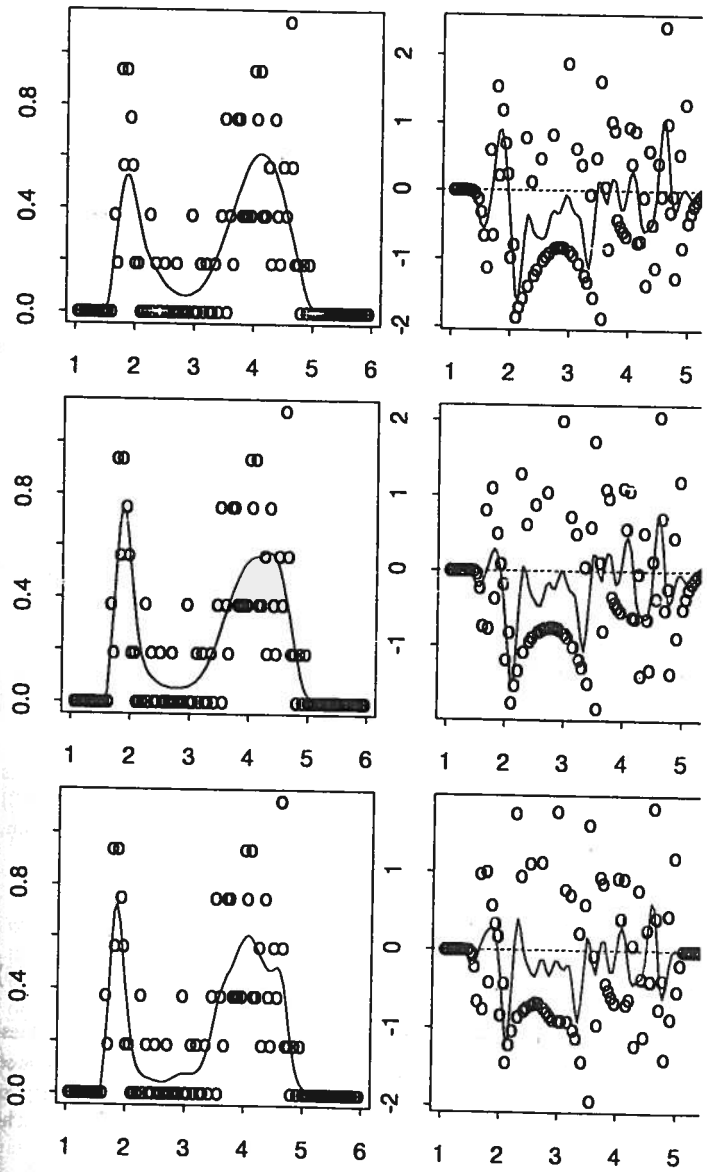


FIGURE 5.7. Fits and smoothed residual plots for geyser data: $\alpha = (0.1,$ (top), $\alpha = (0.1, 0.8)$ (middle) and $\alpha = (0.1, 0.5)$ (bottom).

The influence function for density estimation is defined as

$$\text{infl}(x) = n^{-1}W(0)e_1^T\mathbf{M}_1^{-1}e_1; \tag{5.14}$$

the dependence on $x$ is through the matrix $\mathbf{M}_1$. Then

$$\sum_{i=1}^{n}\log\hat{f}_{-i}(X_i) \approx \sum_{i=1}^{n}\log\hat{f}(X_i) - \sum_{i=1}^{n}\text{infl}(X_i) + 1.$$

Summing over the observations leads to the Akaike information criterion for density estimation:

$$\text{AIC}(\hat{f}) = -2\sum_{i=1}^{n}\log\hat{f}(X_i) + 2\sum_{i=1}^{n}\text{infl}(X_i) + 2n\left(\int_{\mathcal{X}}\hat{f}(u)du - 1\right). \tag{5.15}$$

The factor of $-2$ is introduced here to be consistent with our definition of AIC for local likelihood regression. The quantity

$$\nu_1 = \sum_{i=1}^{n}\text{infl}(X_i)$$

is one definition of the degrees of freedom for a density estimation fit, extending the regression $\nu_1$ defined by (2.16). Correspondingly, we can extend the $\nu_2$ definition to

$$\nu_2 = \sum_{i=1}^{n}\text{vari}(X_i)$$

where $\text{vari}(x) = n^{-1}e_1^T\mathbf{M}_1^{-1}\mathbf{M}_2\mathbf{M}_1^{-1}e_1$.

### 5.3.3 *Squared Error Methods*

An entirely different method of cross validation, known as least squares cross validation, was developed for density estimation by Rudemo (1982) and Bowman (1984). This method does not target the likelihood function, but rather the integrated squared error;

$$\begin{aligned}\text{ISE}(\hat{f}, f) &= \int_{-\infty}^{\infty}(\hat{f}(x) - f(x))^2dx \\ &= \int_{-\infty}^{\infty}\hat{f}(x)^2dx - 2\int_{-\infty}^{\infty}\hat{f}(x)f(x)dx + \int_{-\infty}^{\infty}f(x)^2dx\end{aligned} \tag{5.16}$$

The third term on the right-hand side of (5.16) does not depend on the estimate $\hat{f}(x)$. If the object is to choose $\hat{f}$ to minimize the integrated squared error, then the final term can be ignored. The first term, $\int_{-\infty}^{\infty}\hat{f}(x)^2dx$, depends only on the density estimate and can be evaluated numerically. The central term can be expressed as

$$\int_{-\infty}^{\infty}\hat{f}(x)f(x)dx = \text{E}(\hat{f}(X))$$



where $X$ is a random variable with density $f(\cdot)$ and is independen original sample. This can be estimated by leave-one-out cross valic

$$\hat{\text{E}}(\hat{f}(X)) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}_{-i}(X_i).$$

This leads to the following definition.

**Definition 5.2** The **least squares cross validation** criterion fo sity estimate $\hat{f}(x)$ is

$$\text{LSCV}(\hat{f}) = \int_{-\infty}^{\infty}\hat{f}(x)^2dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(X_i).$$

As usual, the cross validation component can be approximated us influence function. Using (5.13) and (5.14), we have

$$\hat{f}_{-i}(X_i) \approx \hat{f}(X_i)\exp(n^{-1})\exp(-\text{infl}(X_i)) \approx \frac{n}{n-1}\hat{f}(X_i)(1 - \text{infl}(.$$

Thus, the LSCV criterion can be approximated by

$$\text{LSCV}(\hat{f}) \approx \int_{-\infty}^{\infty}\hat{f}(x)^2dx - \frac{2}{n-1}\sum_{i=1}^{n}\hat{f}(X_i)(1 - \text{infl}(X_i)).$$

This is exact for local constant fitting.

### 5.3.4 *Implementation*

The `aicplot()` and `lcvplot()` functions introduced in Section 4.3 be used directly for density estimation. By default, these ignore the i term in (5.15). To renormalize the density estimate so that $\int\hat{f}(x)d$ add the `renorm=T` argument.

The likelihood criteria must be applied rather carefully, since th considerable attention to the tail of densities. But any density estima perform poorly in the tails and choice of bandwidth is largely an as tion. For example, should a single outlier represent its own little pea density, or should it represent a long tail?

Schuster and Gregory (1981) note that LCV, when used to se constant bandwidth estimate, always selects a bandwidth *larger* th smallest separation between data points, and thus produces extremel results for long tailed distributions. AIC also exhibits anomolous be at small bandwidths.

This is not a criticism of AIC or LCV, but simply a recognitio constant bandwidth estimates are poor in tails. The solution comes i

parts. First, ensure that larger bandwidths are used in the tails; for example, by using a nonzero nearest neighbor component in LOCFIT's two-component specification. Second, compare the criteria with the fitted degrees of freedom, and look over a sensible range.

A second problem is caused by ties in the data. This effect has been mostly studied with the LSCV criterion and local constant estimation (Silverman 1986; Sheather 1992). The main result is that if there are too many ties in the data, $\text{LSCV}(\hat{f}_h) \to -\infty$ as $h \to 0$. But again LSCV should not be criticized for this behavior. A sample from a continuous density does not have ties. By selecting $h = 0$, LSCV is simply trying to reproduce the raw data histogram. But problems where this occurs should be treated as discrete, and the LSCV criterion modified accordingly (Exercise 5.4).

**Example 5.9.** In Figure 5.8 we compute the AIC criterion for local constant, local linear and local quadratic density estimates for the Old Faithful dataset. A typical call to `aicplot()` is:

```
> a0 <- cbind(0.05, c(0.17,seq(0.2,0.7,by=0.05)))
> plot(aicplot(~geyser, alpha=a0, deg=0, renorm=T,
+   flim=c(1,6), ev="grid", mg=51), pch="0")
```

To control tail behavior, the nearest neighbor component of the smoothing parameter is fixed at $\alpha = 0.05$ for local constant and local log-linear fitting, and $\alpha = 0.1$ for local log-quadratic. The constant component $h$ of the smoothing parameter is changed from fit to fit. Corresponding computation of the LSCV criterion is shown on the right of Figure 5.8.

We use the fitted degrees of freedom $\nu_2$ as the $x$-axis. Both criteria, and each local polynomial degree (0, 1 and 2), show similar patterns. Fewer than five degrees of freedom is inadequate, while for more than five degrees of freedom the criteria are indecisive. Local log-quadratic fitting is better than local log-linear and local constant.

For local quadratic fitting, six degrees of freedom corresponds to the smoothing parameter $(0.1, 0.9)$, and twelve degrees of freedom corresponds to $(0.1, 0.4)$. The AIC criterion relates to what was shown in the fits and residual plots in Figure 5.7. The largest smoothing parameter, $(0.1, 1.2)$ was too large, with little to choose between the smaller parameters.

While all the curves in Figure 5.8 show a similar pattern, the location of the minimum varies substantially. This emphasizes the importance of looking at the whole cross validtaion curve, rather than just the minimum.

If the bandwidths are decreased further, most of the criteria will downturn again, as discreteness and tails of the data take over. But by ploting the criteria against degrees of freedom, as in Figure 5.8, we obtain a sensible view of the data. Fits above 14 degrees of freedom are rarely useful for datasets of 107 observations.



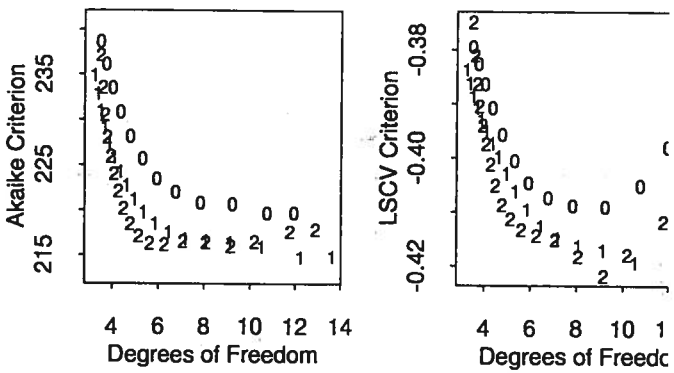


FIGURE 5.8. Akaike's criterion (left) and least squares cross validatior for the Old Faithful dataset. Values for local constant fitting (0), loc fitting (1) and local quadratic fitting (2).

## 5.4 Some Theory for Density Estimation

This section derives basic theoretical properties for the local lik density estimate and develops an approximate distribution theo results are similar to the corresponding results for local likelihood re models in Section 4.4, so only the main ideas are sketched here.

### 5.4.1 *Motivation for the Likelihood*

The attractiveness of maximum likelihood estimation stems from (4 the density estimation notation this can be written as

$$E_f \mathcal{L}(f_1) \le E_f \mathcal{L}(f),$$

with equality only when $f_1 = f$ almost everywhere. With the e definition of the likelihood (5.1), this property holds for *all* non-n functions $f_1$; we do not require $f_1$ to be a density. One consequ this extension is that maximum likelihood estimation can be per with multiplicative parameters. For example, fitting the family $f(x; C\exp(-(x-\mu)^2/2)$ by maximum likelihood gives $\hat{C} = (2\pi)^{-1/2}$.

The property (5.18) extends to the local log-likelihood;

$$E_f \mathcal{L}(f_1, x) \le E_f \mathcal{L}(f, x)$$

with equality when $f(u) = f_1(u)$ on the support of $W((u-x)/h$ suggests estimating $f(x)$ by maximizing (5.2) over a suitable class tions.

### 5.4.2 *Existence and Uniqueness*

Let $\mathcal{C}$ (dependent on the fitting point $x$, the weight function $W$ and the degree of local polynomial $p$) be the parameter space:

$$\mathcal{C} = \{a = (a_0, \ldots, a_p) : \int_{\mathcal{X}} W\left(\frac{u-x}{h}\right) \exp(\langle a, A(u-x)\rangle) du < \infty\}. \tag{5.19}$$

In many cases the set $\mathcal{C}$ is open; for example, if the weight function is bounded and has compact support, $\mathcal{C} = \mathcal{R}^d$. In this case, the parameter vector $\hat{a}$ (if it exists) must lie in the interior of $\mathcal{C}$, and it is a solution of the local likelihood equations (5.5).

The Jacobian of the local likelihood (5.3) is

$$J(a) = -\int_{\mathcal{X}} A(u-x)A(u-x)^T W\left(\frac{u-x}{h}\right) \exp(\langle a, A(u-x)\rangle) du.$$

For non-negative weight functions $W$, this is strictly negative definite. This implies that the local likelihood is concave, and the local likelihood estimate, if it exists, is unique. The following theorem gives precise conditions for existence.

**Theorem 5.1** Suppose the parameter space (5.19) is open. The local likelihood density estimate exists if and only if there exists no parameter vector $a_0 \neq 0$ such that

$$\begin{aligned} \langle a_0, A(X_i - x)\rangle = 0 \quad \forall \quad i : w_i(x) > 0 \\ \langle a_0, A(u-x)\rangle \leq 0 \quad \forall \quad u : W\left(\frac{u-x}{h}\right) > 0. \end{aligned}$$

**Proof:** Suppose such an $a_0$ exists. Then

$$\mathcal{L}_x(\lambda e_1 + c a_0) = \lambda \sum_{i=1}^n w_i(x) - n \int W\left(\frac{u-x}{h}\right) e^{\lambda + c\langle a_0, A(u-x)\rangle} du.$$

Clearly

$$\begin{aligned} \lim_{c\to\infty} \mathcal{L}_x(\lambda e_1 + c a_0) &= \lambda \sum_{i=1}^n w_i(x) \\ \lim_{\lambda\to\infty} \lim_{c\to\infty} \mathcal{L}_x(\lambda e_1 + c a_0) &= \infty; \end{aligned}$$

the likelihood is unbounded and the estimate does not exist.

Conversely, suppose no such $a_0$ exists. Write

$$\sup_a \mathcal{L}(a, x) = \sup_{a:\|a\|=1} \sup_\lambda \mathcal{L}(\lambda a, x); \tag{5.20}$$



we need to show both these suprema are actually achieved. For fixed $\|a\| = 1$, we claim (Exercise 5.3)

$$\begin{aligned} \mathcal{L}_x(\lambda a) &= \lambda \sum_{i=1}^n w_i(x) \langle a, A(X_i - x)\rangle \\ &\quad -n \int_{\mathcal{X}} W\left(\frac{u-x}{h}\right) A(u-x) e^{\lambda\langle a, A(u-x)\rangle} du \end{aligned}$$

is a concave function of $\lambda$ and tends to $-\infty$ as $\lambda \to \pm\infty$ (or when $\lambda$ te the boundaries of the parameter space $\mathcal{C}$, when this is bounded). Th inner supremum of (5.20) must be achieved; let the maximizer be $\lambda =$ Concavity of $\mathcal{L}(a, x)$ implies $\lambda(a)$ must be continuous on the surface unit sphere, and hence the outer supremum is achieved by compactn

What does Theorem 5.1 mean in practical terms? For existence density estimate, we must be unable to find a polynomial (other tha trivial solution, a constant) that attains its maximum at every poi being used in the fit. This generalizes the separating hyperplane the for local logistic regression (Exercise 4.4). The local linear estimate provided at least one observation has nonzero weight, since a linear fun is monotone. A quadratic polynomial may have a single maximum, s local quadratic estimate exists provided two distinct observations re nonzero weight.

### 5.4.3 *Asymptotic Representation*

The main result of this section is an approximate decomposition o local likelihood estimate as the sum of a deterministic bias' compc and a random component. The result is obtained by linearizing the likelihood equations, similarly to the techniques used for local likeli regression in Section 4.4. The following notation is needed:

- $g(x) = \log(f(x))$, and $\tilde{g}$ is the vector of Taylor series coefficient to order $p$.
- $\mathbf{M}_j = \int W(\frac{u-x}{h})^j A(u-x)A(u-x)^T f(u)du; j = 1, 2.$
- $b_p = h^{-(p+1)} \int (u-x)^{p+1} W(\frac{u-x}{h})^j A(u-x) f(u)du.$
- $S_n$ is the left-hand side of the local likelihood equations;

$$S_n = \sum_{i=1}^n w_i(x) A(X_i - x). \tag{5}$$

The decomposition of the local likelihood estimate is, as $n \to \infty$, $h = h$ 0 and $nh_n \to \infty$:

$$\mathbf{H}(\hat{a} - \tilde{g}) = \frac{h^{p+1} g^{(p+1)}(x)}{(p+1)!} \mathbf{H}\mathbf{M}_1^{-1} b_p$$

$$+\frac{1}{n}\mathbf{H}\mathbf{M}_1^{-1}(S_n - \mathrm{E}(S_n)) + o(h^{p+1} + (nh^d)^{-1/2}) \quad (5.23)$$

The first term represents a systematic bias component, and the second term is a random variance component. The bias component as stated is for one dimension; the $d$-dimensional result requires all partial derivatives of $g(x)$ of order $p+1$. The covariance matrix of $S_n$ is evaluated in Exercise 5.1. A central limit theorem (Loader 1996b) shows asymptotic normality of $S_n$, and hence of $\hat{a}$. The normal approximation for $\hat{a}$ has the covariance matrix

$$\frac{1}{n}\mathbf{M}_1^{-1}\mathbf{M}_2\mathbf{M}_1^{-1}.$$

By the delta method, the asymptotic variance of $\hat{f}(x)$ is $f(x)^2$ times the $(1,1)$ element of this matrix.

**Example 5.10.** For the local log-linear density estimate ($p = 1$), one obtains

$$\mathbf{M}_j \approx f(x)\begin{pmatrix} \int W(v)^j dv & 0 \\ 0 & h^2 \int v^2 W(v)^j dv \end{pmatrix}.$$

This yields the variance and bias approximations

$$\begin{aligned} \mathrm{E}(\hat{a}_0) - g(x) &\approx \frac{h^2}{2} g''(x) \frac{\int v^2 W(v) dv}{\int W(v) dv} \\ \mathrm{var}(\hat{a}_0) &\approx \frac{1}{nhf(x)} \frac{\int W(v)^2 dv}{(\int W(v) dv)^2}. \end{aligned}$$

These can be transformed using the delta method to obtain approximate biases and variances for $\hat{f}(x)$:

$$\begin{aligned} \mathrm{E}(\hat{f}(x)) - f(x) &\approx \frac{h^2}{2} f(x) g''(x) \frac{\int v^2 W(v) dv}{\int W(v) dv} \\ \mathrm{var}(\hat{f}(x)) &\approx \frac{f(x)}{nh} \frac{\int W(v)^2 dv}{(\int W(v) dv)^2}. \end{aligned}$$

## 5.5 Exercises

5.1 Consider $S_n$ defined by (5.22), where $X_1, \ldots, X_n$ are independent identically distributed random variables with density $f(x)$.

a) Show

$$\begin{aligned} \mathrm{E}(S_n) &= n \int W\left(\frac{u-x}{h}\right) A(u-x) f(u) du \\ \mathrm{cov}(S_n) &= n\mathbf{M}_2 - \frac{1}{n}\mathrm{E}(S_n)\mathrm{E}(S_n)^T. \end{aligned}$$



Derive a similar expression for the covariance matrix co

b) Suppose the density is continuous at $x$ with $f(x) > 0$, $h = h_n \to 0$ and $nh \to \infty$. Let $\mathbf{H}$ be as defined in (2.37)

$$\frac{1}{\sqrt{n}}\mathrm{cov}(\mathbf{H}^{-1}S_n) = f(x) \int W(v)^2 A(v) A(v)^T dv + o($$

in particular, the covariance term involving $\mathrm{E}(S_n)$ is asy ically negligible. Evaluate $n^{-1}E(\mathbf{H}^{-1}S_n)$ using a Taylo for $f(x)$, retaining terms up to $o(h^2)$.

c) Using Chebycheff's inequality show, on a componentwis $P(|\mathbf{H}^{-1}(S_n - \mathrm{E}(S_n))| \geq \epsilon) \to 0$ for all $\epsilon > 0$. Hence, sho

$$\mathbf{H}^{-1}S_n \to f(x) \int W(v) A(v) dv$$

in probability.

d) Using (5.24) and the local likelihood equations, show

$$\mathbf{H}\hat{a} \to \begin{pmatrix} \log f(x) \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

in probability and that the local likelihood density estin consistent.

5.2 Consider local log-quadratic density estimation in $d$ dimensions, the Gaussian weight function.

a) Write down the local likelihood equations. Express the hand side in terms of the multivariate integrals

$$\int W\left(\frac{u}{h}\right) e^{a + b^T u + u^T \mathbf{C} u} du;$$

$$\int u W\left(\frac{u}{h}\right) e^{a + b^T u + u^T \mathbf{C} u} du;$$

$$\int u u^T W\left(\frac{u}{h}\right) e^{a + b^T u + u^T \mathbf{C} u} du.$$

Here, $b$ is a vector in $\mathcal{R}^d$ and $\mathbf{C}$ is a symmetric $d \times d$ mat

b) Show

$$\begin{aligned} &\int W\left(\frac{u}{h}\right) e^{a + b^T u + u^T \mathbf{C} u} du \\ &= (2\pi)^{d/2} \exp(a + \frac{1}{2} b^T \mathbf{M}^{-1} b) \det(\mathbf{M})^{-1/2} \end{aligned}$$

where $\mathbf{M} = h^{-2} I - 2\mathbf{C}$. Derive closed forms for (5.25) and (5

c) Provide a closed form solution for the density estimate. What condition is necessary for existence of the estimate? Is the parameter space open?

5.3 Consider the log-likelihood $\mathcal{L}_x(\lambda a, x)$ with fixed $a$, $\|a\| = 1$. Suppose $a$ does *not* satisfy the conditions of the vector $a_0$ in Theorem 5.1. That is, either $\langle a, A(X_i - x)\rangle \neq 0$ for some $i$ with $w_i(x) > 0$ or $\langle a, A(u - x)\rangle$ has both positive and negative regions on the support of $W((u-x)/h)$. Show that $\mathcal{L}_x(\lambda a) \to -\infty$ as $\lambda \to \pm\infty$.

5.4 Izenman and Sommer (1988) and Sheather (1992) have fitted kernel density estimates to the postage stamp data (Example 5.4) using the Gaussian kernel and standard deviation about 0.0013. In LOCFIT terms, this is a constant bandwidth of $2.5 \times 0.0013 = 0.00325$.

a) Evaluate and plot this fit. Compare with the local log-quadratic fit (Figure 5.3) and the data. Is the kernel estimate adequate for modeling the peaks?

b) Develop an LSCV algorithm for discrete Poisson regression for kernel density estimation. Use the loss function $\sum_{i=1}^{n}(\hat{p}_i - p_i)^2$ where $p_i$ is the probability of the $i$th bin. The cross validation should use leave-one-observation-out; not leave-one-bin-out. Consider the behavior of LSCV($h$) at small bandwidths. In particular, show it has a finite limit as $h \to 0$ (Bonus: Use the influence function; don't restrict to deg=0).

c) Write an S function to evaluate the discrete LSCV criterion using a LOCFIT fit. Apply this function to the postage stamp data. Compare with the results of Sheather (1992).

*Remark.* The point of this exercise is that discrete data does not have densities, and this is particularly important for model selection when small bandwidths are used.

# 6

# Flexible Local Regression

In this chapter we look at the flexibility that can be obtained by cha the components of local regression: the coefficients, the fitting criteri the weight functions. The specific problems studied include:

- Higher order coefficients and local slopes (section 6.1).
- Periodic and seasonal smoothing (Section 6.2).
- One-sided smoothing and discontinuous function estimation (S 6.3).
- Robust local regression (Section 6.4).

## 6.1 Derivative Estimation

Derivatives are of natural interest in many settings. At the most basic the derivative $\mu'(x)$ measures the effect of the independent variable the mean response. In particular, $\mu'(x) = 0$ implies the covariate is h no effect.

As emphasized in Section 6.1.1, the problem of derivative estimati plagued by identifiability and interpretation difficulties. To make any headway, one must be willing to *assume* that if the local polynomial fit data within the smoothing window, then the local slope provides a approximation to the derivative. This leads to the following local estimate.