

Empirical likelihood and smoothing

This chapter adapts empirical likelihood to curve estimation problems such as density estimation and nonparametric regression. Kernel methods are an attractive choice for this, because they lead easily to estimating equations. We also investigate some regression splines.

5.1 Kernel estimates

Figure 5.1 shows diastolic blood pressure, in millimeters of mercury, plotted against age in years, for some men in New Zealand. There are 7532 data points. The data come from two sources: the Auckland Heart & Health study (Jackson, Yee, Priest, Shaw & Beaglehole 1995) and another study called the Fletcher-Challenge study. The original data were integer valued; the plotted points have $U(-0.5, 0.5)$ random variables added to them to show them better. Blood pressures that are multiples of 10 are more common than others, due to rounding. The rounding stops at around age 60. This point is significant and we return to it later.

Superimposed on the data is a smooth curve taken as a local average of blood pressure. Letting X_i denote age and Y_i denote blood pressure, the curve is

$$\hat{\mu}(x) = \frac{\sum_{i=1}^n K_h(X_i - x)Y_i}{\sum_{i=1}^n K_h(X_i - x)}, \quad (5.1)$$

the Nadaraya-Watson estimator, where

$$K_h(z) = \frac{1}{\sqrt{2\pi}h} \exp\left(-\frac{z^2}{2h^2}\right).$$

The bandwidth h used in Figure 5.1 is 5 years. This local average may be thought of as an estimate of $\mu(x) = E(Y | X = x)$. The denominator in (5.1) can be awkward, and so it is convenient to write $\hat{\mu}(x)$ in terms of the estimating equation

$$0 = \frac{1}{n} \sum_{i=1}^n K_h(X_i - x)(Y_i - \hat{\mu}(x)). \quad (5.2)$$

More general kernel estimates may be formed through

$$K_h(z) = \frac{1}{h} K\left(\frac{z}{h}\right)$$

Men's diastolic blood pressure

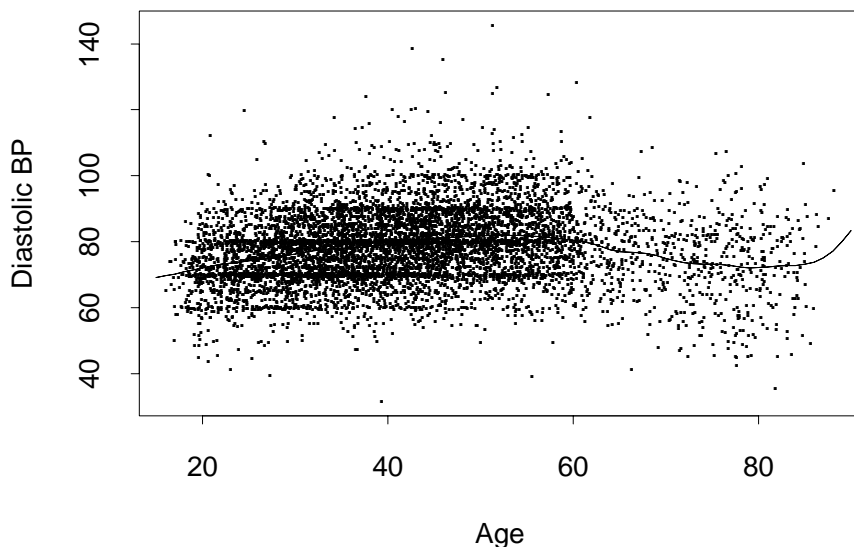


Figure 5.1 *Diastolic blood pressure is plotted versus age. There are 7532 points from men in New Zealand.*

where a kernel function K of integer order $r \geq 1$ satisfies

$$\begin{aligned}\int_{-\infty}^{\infty} K(z) dz &= 1, \\ \int_{-\infty}^{\infty} z^j K(z) dz &= 0, \quad 1 \leq j < r, \\ \int_{-\infty}^{\infty} z^r K(z) dz &\neq 0.\end{aligned}$$

A common choice is for K to be a symmetric probability density function, so that $r = 2$. Higher order kernels take some negative values, typically in “side lobes”, and this can improve accuracy if μ is smooth enough. The benefits of higher order kernels can also be attained by replacing the local average estimator (5.1) by a local linear or local polynomial model. See [Exercise 5.2](#).

Now suppose that $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}^q$, and that independent identically distributed pairs (X_i, Y_i) are observed. An estimate $\hat{\mu}(x)$ of $\mu(x) = E(Y | X = x)$ may still be defined by (5.2), except that the kernel must obviously have a

p -dimensional argument, and the natural scaling is

$$K_h(z) = \frac{1}{h^p} K\left(\frac{z}{h}\right).$$

The order of K may be defined by analogy to the one-dimensional case. Due to a curse of dimensionality, the estimate $\hat{\mu}$ rapidly loses its effectiveness as p increases.

Kernel methods are also widely used to estimate probability density functions. Suppose $X_i \in \mathbb{R}^p$ are independent with a common density f . Then the kernel density estimate of f is

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n K_h(X_i - x).$$

As for kernel regression, kernel density estimation is really only practical for small p .

5.2 Bias and variance

This section illustrates the trade-off between bias and variance for kernel estimates, using a slightly simplified kernel regression estimator with $p = q = 1$.

For $p = 1$, kernel methods put relatively large weight only on the $O(nh)$ nearest observations to x_0 . These neighbors are at distance $O(h)$ from x_0 . If we choose a small h , then the observations are more nearly identically distributed, reducing bias, while for large h there are more of them, reducing variance. This sort of trade-off is not restricted to kernel methods, but is ubiquitous in curve estimation problems. As $n \rightarrow \infty$ we should have $h \rightarrow 0$ and $nh \rightarrow \infty$ to account for bias and variance, respectively.

Suppose for simplicity, that instead of using (5.2), we define

$$\hat{\mu}(x_0) = \frac{1}{n} \sum_{i=1}^n K_h(X_i - x_0) Y_i.$$

Then $E(\hat{\mu}(x_0)) = \tilde{\mu}(x_0)$, where

$$\tilde{\mu}(x_0) = E(K_h(X - x_0)Y) = \int_{-\infty}^{\infty} K_h(x - x_0)\mu(x)f_X(x)dx,$$

where f_X is the density of X . Assuming some smoothness of μ and small h we may write

$$\begin{aligned} \tilde{\mu}(x_0) &\doteq \mu(x_0) + \frac{\mu^{(r)}(x_0)}{r!} \int_{-\infty}^{\infty} K_h(x - x_0)(x - x_0)^r f_X(x)dx \\ &\doteq \mu(x_0) + \frac{h^r \mu^{(r)}(x_0)}{r!} f_X(x_0) \int K(z)z^r dz. \end{aligned}$$

This bias of order h^r in kernel estimates raises difficulties for confidence interval construction below.

To study the variance of $\hat{\mu}(x_0)$, introduce $\sigma^2(x) = \text{Var}(Y | X = x)$. Now

$$\begin{aligned} \text{Var}(\hat{\mu}(x_0)) &= \frac{1}{n} E \left((K_h(X - x_0)Y - \tilde{\mu}(x_0))^2 \right) \\ &= \frac{1}{n} \int_{-\infty}^{\infty} \left[(K_h(x - x_0)\mu(x) - \tilde{\mu}(x_0))^2 \right. \\ &\quad \left. + K_h(x - x_0)^2 \sigma^2(x) \right] f_X(x) dx \\ &\doteq \frac{1}{nh} \int_{-\infty}^{\infty} K(z)^2 [\mu(x_0 + hz)^2 + \sigma^2(x_0 + hz)] f(x_0 + hz) dz \\ &\doteq \frac{1}{nh} (\mu(x_0)^2 + \sigma^2(x_0)) f(x_0) \int_{-\infty}^{\infty} K(z)^2 dz. \end{aligned} \quad (5.3)$$

We see that the variance is of order $(nh)^{-1}$ instead of the rate n^{-1} familiar for finite dimensional parametric vectors. The appearance of $\mu^2(x_0)$ in (5.3) arises because the kernel weights are not made to sum to one. In practice, we fix this by constraining the reweighted mean of $K_h(X_i - x_0)$ to be 1, or by defining $\hat{\mu}$ through the estimating equation (5.2). Then, a more complicated derivation leads to the same asymptotic orders for bias and variance.

The above derivation shows that for $p = 1$, the bias is of order h^r and the variance is of order $(nh)^{-1}$, so the mean squared error is of order $h^{2r} + n^{-1}h^{-1}$. This order is minimized by taking $h \propto n^{-1/(2r+1)}$. The resulting mean squared error is of order $n^{-r/(2r+1)}$. Symmetric densities K have order $r = 2$. Then the optimal rate for h is $n^{-2/5}$, and therefore the mean squared error decreases as $n^{-4/5}$ compared to n^{-1} for vector parameters. The variance decreases as $n^{-4/5}$ so that the kernel method has an effective sample size of order $n^{4/5}$.

5.3 EL for kernel smooths

In practice, a kernel method requires a choice of h . There is a large literature on choosing h from the data. Some references are given in Chapter 5.8. We will study empirical likelihood for fixed sequences $h = h(n)$ in order to gain insight into its behavior. In practice, when h is determined from the data, bootstrap calibration of the profile empirical log likelihood should be used, with h being determined in each bootstrap replication. We begin by considering pointwise inferences at a single value x . Confidence bands with uniform coverage over a set of x values are considered in Chapter 5.6.

The profile empirical likelihood ratio function for kernel regressions based on (5.2) is

$$\mathcal{R}_x(\mu) = \max \left\{ \prod_{i=1}^n n w_i \mid \sum_{i=1}^n w_i Z_{in}(x, \mu) = 0, w_i \geq 0, \sum_{i=1}^n w_i = 1 \right\}, \quad (5.4)$$

where

$$Z_{in}(x, \mu) = K_h(X_i - x)(Y_i - \mu)$$

and $h = h(n)$. There are two difficulties in using this statistic for inferences. The first difficulty, which we return to below, is that due to bias in kernel estimation, the expected value of Z_{in} is zero not at $\mu = E(Y | X = x)$ but at some other value $\tilde{\mu}(x)$.

The second difficulty is in applying an ELT to Z_{in} . If we consider $X_i = x_i$ to be fixed predictors, then the Z_{in} for $i = 1, \dots, n$ and a given value of n , are not identically distributed. For IID (X_i, Y_i) pairs, the distribution of Z_{in} changes with n and has a variance diverging to infinity. The difficulty in applying empirical likelihood may be resolved through the triangular array ELT, [Theorem 4.1](#) of Chapter 4.3.

[Theorem 4.1](#) requires a common mean for the Z_{in} , a convex hull condition, and two conditions on the variances of Z_{in} . Suppose that $(X_i, Y_i) \in \mathbb{R}^2$ are an IID sample. Then Z_{1n}, \dots, Z_{nn} have a common mean 0, but at a point $\tilde{\mu}(x)$ differing from μ by a bias of order h^r . The convex hull condition is satisfied as soon as at least two suitable data points appear: one has $Y_i > \tilde{\mu}(x)$ and $K_h(X_i - x) > 0$, while the other has $Y_i < \tilde{\mu}(x)$ and again $K_h(X_i - x) > 0$. For compactly supported kernels there are at least $O(nh)$ points with $K_h(X_i - x) > 0$ and for other kernels there may be n such points. The convex hull condition is very quickly satisfied in this case.

Easy calculations show that $V_n = \text{Var}(Z_{in}) = O(h^{-1})$ as $n \rightarrow \infty$ and $h = h(n) \rightarrow 0$. In this case V_n is its own smallest and largest eigenvalue. The ratio of largest to smallest eigenvalues is thus constant with n and so does not raise difficulties. Next, we require a limit of zero for

$$\frac{E((Z_{in} - E(Z_{in}))^4)}{nV_n^2}. \tag{5.5}$$

Mild moment assumptions give a rate of $O(h^{-3})$ for the numerator in (5.5) and then the ratio itself is $O((nh)^{-1})$. We already needed $nh \rightarrow \infty$ to control the variance of the estimate. Thus the triangular array ELT applies to kernel smoothing under weak conditions.

The fact that $E(\sum_{i=1}^n Z_{in}) = 0$ not at $\mu(x)$, but at $\tilde{\mu}(x)$, is more problematic. There are several approaches to dealing with this problem, none of them completely satisfactory.

The first approach is based on undersmoothing. The value of h is taken small enough that the bias in $\hat{\mu}$ is negligible compared to its standard deviation. Then the error $\mu - \hat{\mu}$ is primarily due to sampling fluctuations and not bias, making empirical likelihood inferences on $\tilde{\mu}$ relevant to μ . The disadvantages of this approach are that the undersmoothed estimate $\hat{\mu}$ is less accurate than it would be with the usual choice of h , and that the curve $\hat{\mu}$ that results is more wiggly than the usual one. In applications where we are most concerned about $\mu(x_0)$ for one or a small number of specific x_0 values, this roughness is less important. Undersmoothing

is well suited to problems where getting a reliable confidence region for $\mu(x)$ is more important than getting the best possible point estimate of $\mu(x)$.

The second approach is to accept the empirical likelihood inferences as confidence statements about $\tilde{\mu}$, a smoothed version of μ . The disadvantage here is that $\tilde{\mu}$ is not very interpretable and the error $\tilde{\mu} - \mu$ may be as large as, or larger than, the diameter of the confidence region. As n and h change, $\tilde{\mu}$ changes too, and so it is not even a feature of the joint (X, Y) distribution. There are some settings in which this approach is adequate. Sometimes the curve $\hat{\mu}$ is used for qualitative interpretation, and not strictly as an estimate of μ . In settings like this, one might prefer an oversmoothed estimate $\hat{\mu}$, using a larger value of h than the usual one. Empirical likelihood confidence regions for $\tilde{\mu}$ can be used to assess how much sampling fluctuation might contribute to features in $\hat{\mu}$.

A third approach is to compute an estimate of the bias $\tilde{\mu}(x) - \mu(x)$, and subtract this from the estimate $\hat{\mu}(x)$ and from the boundary of the confidence set. The bias can be estimated by using a kernel of higher order. The disadvantage of this approach is that the point estimate of $\mu(x)$ is essentially produced by a higher order kernel, while the confidence set around it is constructed for the original kernel.

5.4 Blood pressure trajectories

The complete blood pressure data, after removing 63 incomplete cases, has the age and blood pressures (systolic and diastolic) of 7532 men and 2934 women. As people age, their blood pressures tend to increase, though the two blood pressures increase in different ways, and the pattern is different for men and women. [Figure 5.2](#) shows the trajectories taken by the average blood pressure measurements for both men and women. A Gaussian kernel with a bandwidth of $h = 5$ years was used for both trajectories. Here we have age $X \in \mathbb{R}$ and blood pressure $Y \in \mathbb{R}^2$, so the conditional means μ_M and μ_F , for men and women, respectively, are space curves. [Figure 5.3](#) replots the same kernel smooths in perspective to show them as space curves.

The changes in mean blood pressure with age tend to be small compared to the fluctuations between people at a single age. A relatively small shift in the blood pressure of an entire population can, however, be a very significant public health issue.

Both average blood pressures increase for men, until some point around 50 to 55 years of age. Then the systolic blood pressure keeps on increasing, while the diastolic blood pressure starts to decrease.

Women's average blood pressures also increase with age, until about age 55. Then the diastolic blood pressure stays roughly constant while the systolic increases then decreases. The final decrease is for the highest ages, and is estimated with a smaller sample.

The women's blood pressure curve starts lower than the men's. Their diastolic blood pressure increases more slowly, especially during their 30's and 40's and

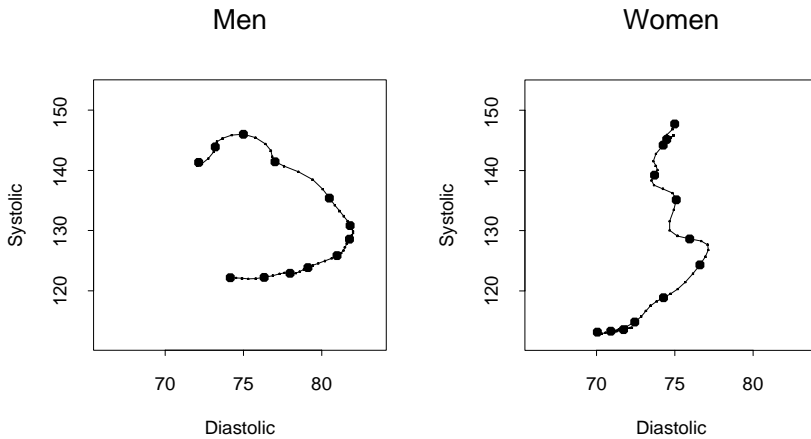


Figure 5.2 Shown are the trajectories taken by the kernel smoothed systolic and diastolic blood pressures of men and women in New Zealand. There is a point for each integer age from 25 to 80 inclusive. The large points are for ages that are multiples of 5 years. In both cases an age of 25 years corresponds to the lowest systolic blood pressures.

does not rise to the levels reached by men's average diastolic blood pressure. The rapid decrease in average diastolic blood pressure for men brings that average towards the women's average by about age 70.

Some of the reasons for these patterns are well understood. For example, estrogen protects younger women from hypertension. Other reasons are more complicated. The men's decrease in diastolic blood pressure could be due to increased mortality among those with higher blood pressure, or to increased use of medication to reduce blood pressure. Notice that men's systolic blood pressure does not show the same pattern. Chapter 5.7 revisits the pattern in men's diastolic blood pressure.

The bandwidth for Figures 5.2 and 5.3 is 5 years. Figure 5.4 shows pointwise empirical likelihood confidence regions for μ_M and μ_F both at age 40. The ellipses are narrower for men, because there are more men in the data set. In both cases a positive correlation between systolic and diastolic blood pressure is evident.

5.5 Conditional quantiles

Average blood pressure is perhaps less interesting than extreme blood pressure. The α -quantile of $Y \in \mathbb{R}$ conditional on $X = x_0 \in \mathbb{R}^p$ may be estimated by the

Mean blood pressure trajectories

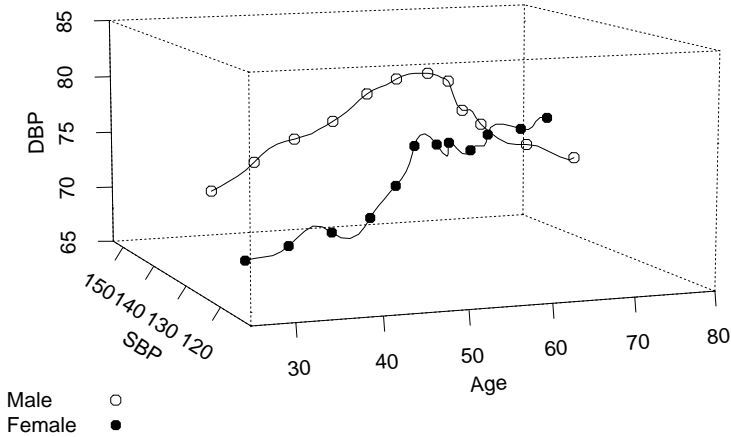


Figure 5.3 The blood pressure trajectories from Figure 5.2 are replotted in a perspective plot to depict them as space curves. The circles show the men's trajectory, the disks show the women's trajectory.

solution $Q^\alpha(x_0)$ of the estimating equation

$$0 = \frac{1}{n} \sum_{i=1}^n K_h(X_i - x)(1_{Y_i \leq Q^\alpha(x)} - \alpha). \quad (5.6)$$

Table 5.1 on page 120 shows estimated conditional 90th percentiles of women's systolic blood pressure at ages from 25 to 80 by steps of 5 years. A Gaussian kernel with a bandwidth of 5 years was used to smooth the data. From Figure 5.2 it appears that women's average systolic blood pressure starts to move sharply upwards at around age 40 to 45. The estimates in Table 5.1 show that large increases in the 90th percentile may start earlier, between the ages of 35 and 40. Figure 5.5 shows the empirical likelihood function for $Q^{0.90}(x_0)$ at ages x_0 from 30 to 80 by steps of 10 years.

5.6 Simultaneous inference

For some purposes, we seek a confidence region for the whole function $\mu(x)$ over x in a domain $\mathcal{D} \subset \mathbb{R}^p$ of interest. Examples for \mathcal{D} include finite sets of $k \geq 1$ points, intervals when $p = 1$, and hyper-rectangles, spheres, or balls when $p > 1$. In principle the choice for \mathcal{D} is very open, but in practice it is necessary to be able

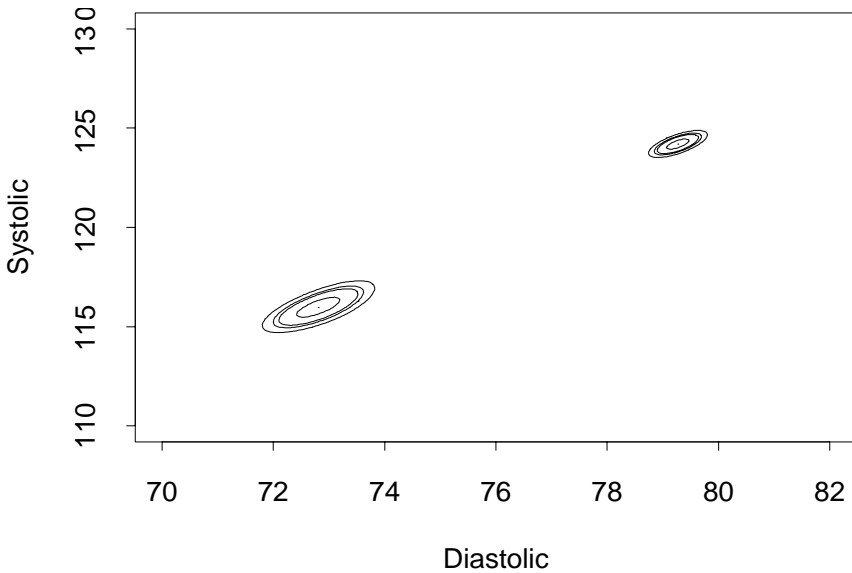


Figure 5.4 Systolic blood pressure is plotted versus diastolic. The ellipses shown give confidence regions for the mean blood pressure of 40-year-old men and 40-year-old women. The ellipses for men are at higher blood pressures than those for women. The contours correspond to confidence levels of 50, 90, 95, and 99 percent, based on a $\chi_{(2)}^2$ calibration.

to optimize over \mathcal{D} . The confidence region is the set of functions μ given by

$$\left\{ \mu : \mathcal{D} \rightarrow \mathbb{R}^q \mid \sup_{x \in \mathcal{D}} -2 \log (\mathcal{R}_x(\mu(x))) \leq C \right\},$$

where \mathcal{R}_x is given by (5.4). For $p = 1$ and \mathcal{D} an interval, this is a confidence band or confidence tube depending on whether $q = 1$ or $q > 1$. For $p = 2$ and $q = 1$ and \mathcal{D} a rectangle, this is a confidence sandwich. Clearly the threshold C should be larger than the $1 - \alpha$ point of a $\chi_{(q)}^2$ distribution.

There is some extreme value theory for choosing C , but in practice, it is probably better to use bootstrap calibration. Suppose that $p = 1$, $\mathcal{D} = [a, b]$ is an interval, and h is given. Let $\hat{\mu}(x)$ be the kernel estimate (5.2). Let (X_i^b, Y_i^b) be independent samples from the EDF \hat{F} of (X, Y) pairs, for $i = 1, \dots, n$, and $b = 1, \dots, B$. Let

$$C^b = \sup_{x \in \mathcal{D}} -2 \log (\mathcal{R}_x^b(\hat{\mu}(x))),$$

Age	Estimate	Lower	Upper
25	126.20	125.0	127.9
30	127.80	126.0	130.0
35	130.50	130.0	134.0
40	135.55	132.5	137.0
45	139.55	137.0	140.0
50	145.50	142.5	147.5
55	151.50	149.0	155.0
60	157.45	155.0	160.9
65	164.45	160.0	169.0
70	169.70	165.0	172.0
75	171.45	169.5	176.0
80	173.50	170.0	181.0

Table 5.1 Shown are kernel-based estimates of the 90th percentile of women's systolic blood pressure at ages separated by 5-year intervals. The lower and upper 95% confidence bounds for these quantiles are based on the empirical likelihood with a $\chi^2_{(1)}$ calibration.

where $\mathcal{R}_x^b(\mu)$ is

$$\max \left\{ \prod_{i=1}^n n w_i \mid \sum_{i=1}^n w_i K_h(X_i^b - x) (Y_i^b - \mu) = 0, w_i \geq 0, \sum_{i=1}^n w_i = 1 \right\},$$

and define the order statistics $C^{(1)} \leq C^{(2)} \leq \dots \leq C^{(B)}$. Then the order statistic $C^{((1-\alpha)B)}$ provides bootstrap calibration at the level $1 - \alpha$.

For the New Zealand men's blood pressure data, 1000 bootstrap values of C^b were computed using for \mathcal{D} a grid of ages ranging from 20 to 80 inclusive by steps of 5 years. The distribution of $-2 \log(\mathcal{R}_x^b(\hat{\mu}(x)))$ fits the $\chi^2_{(2)}$ very closely at each x from 20 to 80, as might be expected for such a large sample. For simultaneous coverage over a set of ages, we are interested in the distribution of C^b , a maximum of correlated random variables, each with nearly the $\chi^2_{(2)}$ distribution. In this instance, the 95th percentile of C^b was 10.41. Using this threshold, we can produce a confidence tube for the mean blood pressure trajectories, with simultaneous coverage over $x_0 \in \mathcal{D}$ of 95%. That tube is displayed in Figures 5.6 and 5.7.

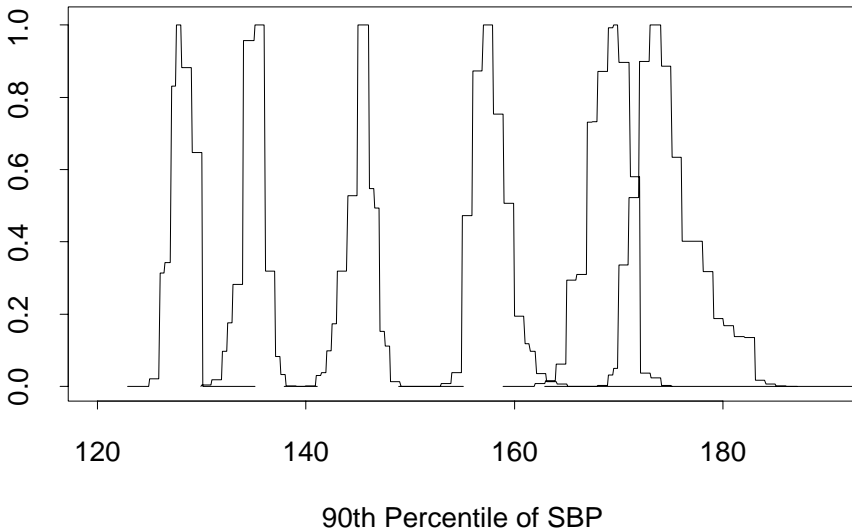


Figure 5.5 The horizontal axis represents the 90th percentile of women's systolic blood pressure. The vertical axis is the empirical likelihood ratio. The curves, from left to right, are for ages 30 through 80 by steps of 10 years. A Gaussian kernel with bandwidth of 5 years was used. The sample is very sparse near 80 years of age, and this accounts for the greater uncertainty regarding the blood pressure there.

5.7 An additive model

One particularly interesting feature in the blood pressure data is the eventual decline in men's diastolic blood pressure at greater ages. This decline starts to set in at around the same age where the blood pressures in [Figure 5.1](#) stop showing rounding to multiples of 10. The data come from two studies done on different populations and with different measurement methods, as described in Chapter 5.8. The populations have different age ranges, but there is some overlap. A graphical exploration of the age range where the studies overlap indicates that the male diastolic blood pressures at a given age tend to run higher in the Fletcher study, which had younger subjects. Thus some or all of the decline in blood pressure could be an artifact of age differences between the men in the two studies.

It would be interesting to know whether the decline is solely attributable to the study difference or not. To handle this, we formulate a model in which the mean diastolic blood pressure has the form $\mu(x, z) = s(x) + z\beta$ where x is age, $z = 1$ for the Fletcher-Challenge study, $z = 0$ for the Auckland Heart & Health study, β is a scalar coefficient, and s is a smooth function. Then if $s(x)$ shows the eventual decline with age, we can be more certain that it is real.

A convenient way to encode the smooth function $s(x)$ is through a cubic spline.

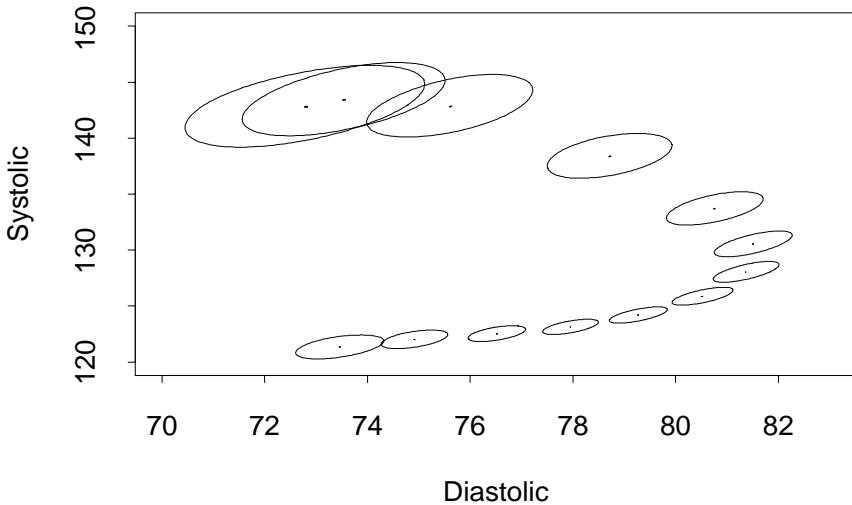


Figure 5.6 Systolic blood pressure is plotted versus diastolic. The ellipses shown give simultaneous 95% confidence regions for the kernel smoothed mean blood pressure of men ranging in age from 20 to 80 years at intervals of 5 years. Increasing age corresponds to a counter-clockwise movement from the lower left. At extreme ages there are fewer data points, and consequently larger confidence regions.

This is a function that is piecewise cubic between points called knots, and has two continuous derivatives at the knots. For this example, knots were placed at ages 30, 40, 50, 60, and 70. The truncated power basis for $s(x)$ takes the form

$$s(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^5 \beta_{3+j} [x - 10(2 + j)]_+^3, \quad (5.7)$$

where $[x - t]_+ = x - t$ if $x \geq t$ and is 0 if $x \leq t$. This basis can be very badly conditioned numerically, and so another basis for the same family of curves was constructed using the S-PLUSfunction `bs()`. This B-spline basis is much more stable numerically, but the individual functions in it are not as interpretable as those in (5.7).

Consider the additive model

$$E(Y_i | X_i = x_i, Z_i = z_i) = \beta_0 + \sum_{j=1}^8 \beta_j \phi_j(x_i) + \beta_9 z_i,$$

where ϕ_j are the spline basis functions, X_i is age, Y_i is the diastolic blood pressure, and Z_i is the study indicator variable described above. The least squares estimate of the study effect is $\hat{\beta}_9 = 6.35$ and the least squares estimate of the age

Mean blood pressure confidence tube

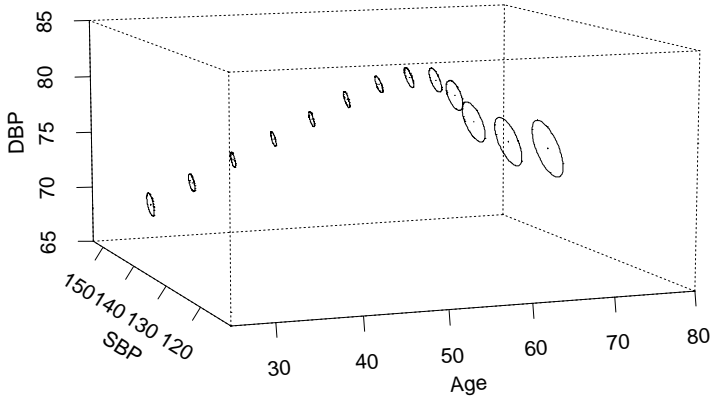


Figure 5.7 The simultaneous confidence ellipses from Figure 5.6 are plotted in the same coordinate space as used in Figure 5.3 to display mean blood pressure trajectories. They outline the shape of a 95% confidence tube for the smoothed path of mean blood pressure with age.

effect

$$\gamma \equiv s(75) - s(50) = \sum_{j=1}^8 \beta_j (\phi_j(75) - \phi_j(50))$$

is

$$\hat{\gamma} = \sum_{j=1}^8 \hat{\beta}_j (\phi_j(75) - \phi_j(50)) = 3.52.$$

In Figure 5.2, DBP appears to decrease by about 10 mmHg from age 50 to age 75. The data for 75-year-olds are almost entirely from the Auckland Heart & Health study, while most of the data for 50-year-olds are from the Fletcher-Challenge study. Therefore at least a large part of the 10-mmHg decline can be attributed to the difference between the two studies. We expect $\hat{\gamma} + \hat{\beta}_9$ should be close to 10 but it need not match exactly because kernel and spline smoothers are slightly different, and because the kernel smooth at age 50 mixes data from both studies.

The empirical log likelihood for γ has $-2 \log(\mathcal{R}(0)) = 15.48$, providing very strong evidence that the age effect γ , comparing 75- and 50-year-old men, is not 0. A 95% confidence interval for the age effect γ , using a $\chi^2_{(1)}$ calibration, extends from 1.77 to 5.26.

5.8 Bibliographic notes

The blood pressure data, kindly supplied by Thomas Yee of the University of Auckland, had two sources. One source was a study on the Fletcher-Challenge company for which most subjects were between 20 and 60 years of age. The other source was the Auckland Heart & Health study (Jackson et al. 1995), which sampled people from the electoral rolls. Their ages were evenly distributed between 35 and 85 years. In the Fletcher-Challenge study, blood pressure was measured with a standard Hg sphygmomanometer and recorded by a nurse. In the Auckland Heart & Health study, blood pressure was measured with a Hawsley random sphygmomanometer for which the nurse was not aware of the blood pressure value.

The smoothing presented here used a bandwidth of 5 years in all the blood pressure examples. For such a large data set, this is likely to be oversmoothing. It might be better to use a small bandwidth, or better yet, a local linear smooth as described in Chapter 5.9. The analysis of women's systolic blood pressure did not include an indicator variable for the studies. Such an indicator might improve the analysis, though there only seemed to be a small study effect for women's systolic blood pressure.

The Nadaraya-Watson estimator was proposed, independently, by Nadaraya (1965) and Watson (1964). Sufficient conditions for the bias and variance expansions of kernel estimates may be found in Härdle (1990). Basic references on smoothing include Hastie & Tibshirani (1990) and Fan & Gijbels (1996). Specialized accounts of bandwidth estimation appear in Härdle & Marron (1985) and Härdle, Hall & Marron (1988).

Stone (1977) describes conditional M -estimators defined by local reweighting. Owen (1987) describes these local weights as distributions on the space of X values and gives conditions for consistency and asymptotic normality of estimates in terms of convergence of these distributions to δ_x . The monograph Loader (1999) on local likelihood, considers local versions of statistical methods in depth.

Hall & Owen (1993) consider empirical likelihood confidence bands for kernel density estimates. They obtain an asymptotic calibration for the critical likelihood based on extreme value theory. They also propose bootstrap calibration for this problem. Relatively minor changes are required to translate empirical likelihood results from densities to regressions, and vice versa. Hall & Owen (1993) display confidence bands for the probability density function applied to the Old Faithful geyser data.

Chen (1996) studies coverage levels of undersmoothed kernel density estimates. In his aerial transect sampling problems the population density of blue-fin tuna depends on a density function at the origin, corresponding to a zero distance between the school of fish and the spotting plane. He shows that empirical likelihood is effective at forming a confidence interval for the desired quantity $f(0)$, and that Bartlett correction is possible, though even more undersmoothing is required. Zhang (1998) shows that there is no asymptotic benefit from global side

constraints in kernel density estimation. Chen (1997) imposes side constraints (zero mean and skewness) on kernel density estimates of tuna densities estimated from aerial surveys. He shows that there is no asymptotic benefit to imposing those side constraints, but finds an improvement in finite samples.

Fan & Gijbels (1996) is a comprehensive reference on local polynomial smoothing. Chen & Qin (2000) consider empirical likelihood confidence intervals for local linear kernel smoothing. For an undersmoothed estimator, they get a $\chi^2_{(1)}$ limit with a coverage error of $O(nh^5 + h^2 + (nh)^{-1})$. They remark that the use of empirical likelihood produced the same rate of convergence for coverage error at the endpoints as in the middle of the predictor range. An alternative approach (Chen & Qin 2001) using first and second moments has coverage error $O(nh^5 + h^2 + (nh)^{-1})$ in the interior and $O(nh^5 + h + (nh)^{-1})$ near the boundary.

The approach taken here to smoothing started by considering problems in which the curve could be analyzed pointwise, producing at $x_0 \in \mathbb{R}^p$ a confidence region for $\mu(x_0) \in \mathbb{R}^q$. If we were interested in the joint behavior of $\mu(x_1), \dots, \mu(x_k)$ at k points, then we could maximize the empirical likelihood while forcing all of the reweighted values to be 0 simultaneously. We would expect an asymptotic χ^2 distribution with kq degrees of freedom, and we would expect, at least for modest k , that this test should have better power than one based on the supremum of k pointwise empirical likelihoods. But suppose we want to test a hypothesis such as $\mu(x) = 0$ for all $x \in \mathbb{R}^p$. We cannot expect to reweight the data and get $\hat{\mu}(x) = 0$ at infinitely many points x . The approach in Chapter 5.6 is based on the supremum over x of $R_x(0)$. It is reasonable to expect that a better method exists. Sieve techniques letting $k \rightarrow \infty$ with n , as described in Chapter 9.10, may help.

5.9 Exercises

Exercise 5.1 Let $\tilde{\mu}(x)$ satisfy $E(K_h(X - x)(Y - \tilde{\mu}(x))) = 0$. Suppose that $K_h(z) = K(z/h)/h$ where K is a symmetric probability density. Give an informal argument that the bias $\tilde{\mu}(x) - \mu(x)$ is

$$\frac{h^2}{2f(x)} [\mu''(x)f(x) + 2\mu'(x)f'(x) + \mu''(x)f(x)] \int z^2 K(z) dz + O(h^4)$$

as $h \rightarrow 0$. Here X has probability density function f and the expected value of Y given $X = x$ is $\mu(x)$.

Exercise 5.2 Local linear and local polynomial regression are effective ways to smooth data, because they adapt to the local spacing of the X_i . For local linear regression, let $\theta_0 = \theta_0(x)$ and $\theta_1 = \theta_1(x)$ minimize

$$\sum_{i=1}^n K_h \left(\frac{X_i - x_0}{h} \right) \left(Y_i - \theta_0 - \theta_1(X_i - x_0) \right)^2,$$

where $X_i, Y_i, x_0 \in \mathbb{R}$. The smooth value at x is $\theta_0(x)$. Write estimating equations for θ_0 and θ_1 . Extend the local linear estimating equations to local polynomial estimating equations. Extend the local linear estimating equations to $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}^q$.

Exercise 5.3 Formulate estimating equations for an α quantile of Y that is locally linear in X . Here $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$.