Stat 526 Term 2, 1998-99
Regression for Exponential Families

March 16, 1999

1 Review of Exponential Families

1.1 When we don't have a regression problem

Recall that Y has a distribution from an exponential family with natural parameter η if Y has a density/probability mass function of the form

$$f_Y(y) = \exp[y\eta - b(\eta) + d(y)].$$

Then $\mathrm{E}(Y) = b'(\eta)$ and $\mathrm{var}(Y) = b''(\eta)$. Typically, η is unrestricted, that is η can be any real number. Often $\mathrm{E}(Y)$ is restricted, e.g. to be positive.

If Y_i, $i = 1, \cdots, n$, are iid then the maximum likelihood estimate of η maximizes

$$\ell(\eta) = \sum [Y_i \eta - b(\eta)],$$

that is, $\hat{\eta}$ solves

$$\ell'(\eta) = \sum [Y_i - b'(\eta)] = 0,$$

or

$$\hat{\eta} = (b')^{-1}(\sum Y_i/n)$$

where $(b')^{-1}$ is the inverse of b'. Note that b'' is positive, if it is indeed to be a variance, so b' is increasing and thus invertible. The observed Fisher information is $\mathcal{I}_n^o(\eta) = -\ell''(\eta) = nb''(\eta)$ and the Fisher information is $\mathcal{I}_n(\eta) = \mathrm{E}(\mathcal{I}_n^o(\eta))$. For exponential families with natural parameterization, the observed Fisher information and the Fisher information are equal. The asymptotic variance of $\hat{\eta}$ is $\mathcal{I}_n^{-1}(\eta) = 1/(nb''(\eta)) = 1/(n\text{var}(Y_i))$. This can be estimated by $\mathcal{I}_n^{-1}(\hat{\eta})$ or by the less model-based estimate $1/\sum(Y_i - \bar{Y})^2$.

We can use $\hat{\eta}$ to find the maximum likelihood estimate of $\mathrm{E}(Y) \equiv \mu$: $\hat{\mu} = b'(\hat{\eta}) = \sum Y_i/n$. The standard error of $\hat{\mu}$ is then $\sqrt{b''(\hat{\eta})/n}$.

Althought it's not necessary in this simple ase, we can use the delta method to find the standard error of $\hat{\mu}$ from the standard error of $\hat{\eta}$: the standard error of $\hat{\mu}$ is $|b''(\hat{\eta})|$ times the standard error of $\hat{\eta}$, that is $\sqrt{b''(\hat{\eta})/n}$.

Example 1. Consider Y Poisson with expected value λ. Then

$$P\{Y = y\} = \lambda^y e^{-\lambda}/y! = \exp(y \ln \lambda - \lambda - \ln y!).$$

So $\eta = \ln \lambda$, $b(\eta) = \exp(\eta)$, $b'(\eta) = \exp(\eta)$, $(b')^{-1}(\lambda) = \ln \lambda$ and $b''(\eta) = \exp(\eta)$. The maximum likelihood estimate of η satisfies $\hat{\eta} = \ln(\sum Y_i/n)$, and so $\hat{\lambda} = \sum Y_i/n$. The standard error of $\hat{\eta}$ is $1/\sqrt{n \exp(\hat{\eta})} = 1/\sqrt{n\hat{\lambda}}$. The standard error of $\hat{\lambda}$ is $\sqrt{\hat{\lambda}/n}$.

1.2 Dispersion Parameter

Often we want to incorporate a (possibly unknown) *dispersion parameter* into the density/probability mass function of an exponential family. We say that Y is from an exponential family with natural parameter η and dispersion parameter ϕ if

$$f_Y(y) = \exp[(y\eta - b(\eta))/\phi + d(y, \phi)].$$

Then $\mathrm{E}(Y) = b'(\eta)$ and $\text{var}(Y) = \phi b''(\eta)$.

If Y_i, $i = 1, \cdots, n$, are iid as Y then the maximum likelihood estimate of η will still be the value that maximizes

$$\ell(\eta) = \sum_{i=1}^{n} [Y_i\eta - b(\eta)].$$

Example 2. Consider Y normally distributed with mean μ and variance σ^2. Then Y is from an exponential family with natural parameter $\eta = \mu$, $b(\eta) = \eta^2/2$, and with dispersion parameter $\phi = \sigma^2$.

Example 3. Consider Y, integer valued, with probability mass function

$$P\{Y = y\} = \exp([y \ln \lambda - \lambda]/\phi + d(\phi, y)). \tag{1}$$

If $\phi = 1$, then Y has a Poisson distribution. If $\phi > 1$, then Y has an *overdispersed* Poisson distribution. If $\phi < 1$, then Y has an *underdispersed* Poisson distribution. The natural parameter is $\eta = \ln\lambda$, $\mathrm{E}(Y) = \lambda$ and $\mathrm{var}(Y) = \phi\lambda$. When should one incorporate a dispersion parameter for count data? You should do this if you have Y_i iid, but with the sample mean very different from the sample variance.

Unfortunately, I have been unable to find the exact form of Y's probability mass function. One way the overdispersed Poisson is explained is as follows. Fix L and suppose that λ_i, $i = 1, \cdots, L$, are iid positive numbers with expected value λ/L and variance τ^2/L. Given $\lambda_1, \cdots, \lambda_L$, Y_i is Poisson with mean λ_i, independent of the other Y_j's. Then $Y = Y_1 + \cdots + Y_L$ is called an overdispersed Poisson. It's not clear to me that Y has probability mass function given by (1). However, note that $\mathrm{E}(Y_i|\lambda_i) = \lambda_i$ so $\mathrm{E}(Y_i) = \mathrm{E}(\lambda_i) = \lambda/\mathrm{L}$ and $\mathrm{E}(Y) = \lambda$. Also, $\mathrm{var}(Y_i|\lambda_i) = \lambda_i$ so $\mathrm{var}(Y_i) = \mathrm{E}[\mathrm{var}(Y_i|\lambda_i)] + \mathrm{var}[\mathrm{E}(Y_i|\lambda_i)]$ $= \lambda/L + \tau^2/L$, and thus $\mathrm{var}(Y) = \lambda + \tau^2 = \phi\mathrm{E}(Y)$ where $\phi = (\lambda + \tau^2)/\lambda$.

Another way the overdispersed Poisson arises is to let $Z_i, i \geq 1$, be Poisson with mean λ, and let N be Poisson with mean η, N independent of the Z_i's. Take $Y = \sum_1^{N+1} Z_i$. Then $\mathrm{E}(Y|N) = (N+1)\lambda$ so $\mathrm{E}(Y) = (\eta+1)\lambda$. Also, $\mathrm{var}(Y|N) = (N+1)\lambda$ so $\mathrm{var}(Y) = \mathrm{E}[\mathrm{var}(Y|N)] + \mathrm{var}[\mathrm{E}(Y|N)] = (\eta+1)\lambda + \lambda^2\eta$, and thus $\mathrm{var}(Y) = \phi\mathrm{E}(Y)$ where $\phi = 1 + \lambda\eta/(\eta+1)$. Again, it's not clear to me that Y has probability mass function given by (1).

I suspect that (1) can arise under more than one situation, so that the probability mass function of an over-dispersed Poisson is not actually well-defined. However, any random variable with probabiliy mass function of the form (1) has $\mathrm{E}(Y) = b'(\eta)$ and $\mathrm{var}(Y) = \phi b'(\eta)$. Since $\mathrm{E}(Y)$ is usually the parameter of interest, this ambiguity concerning the probability mass function is, in some sense, unimportant. And this moves us into the realm of quasi-likelihood!

1.3 The regression model: parametric case

We fix a particular exponential family with natural parameter η and then consider η to actually be a function. Specifically we assume that Y_i, $i = 1, \cdots, n$, are independent and that, given x_i, Y_i is from an exponential family with natural parameter $\eta(x_i)$ and with dispersion parameter $\phi = 1$. Let $m(x) = \mathrm{E}(Y|x) = b'(\eta(x))$. We can model $\eta(x) = \sum_1^p \beta_j\eta_j(x)$ for some known functions η_j and for unknown β_j's. We can estimate the β_j's via maximizing the likelihood, then estimate $\eta(x)$ by $\hat{\eta}(x) \equiv \sum_j \hat{\beta}_j \eta_j(x)$ and $m(x)$ by $\hat{m}(x) = b'(\hat{\eta}(x))$. The asymptotic covariance matrix of the $\hat{\beta}_j$'s is equal to the inverse of the Fisher Information matrix. Thus, we can calculate the asymptotic variance of $\hat{\eta}(x)$ and then, via the delta method, calculate the asymptotic variance of $\hat{m}(x)$.

The maximum likelihood estimates of the β_j's maximize

$$\ell(\beta_1, \cdots, \beta_p) = \sum_{i=1}^{n} \left[Y_i \sum_{j=1}^{p} \beta_j \eta_j(x_i) - b\left(\sum_{j=1}^{p} \beta_j \eta_j(x_i) \right) \right]$$

that is, the $\hat{\beta}_j$'s solve

$$0 = \frac{\partial \ell}{\partial \beta_J} = \sum_{i=1}^{n} \eta_J(x_i) \left[Y_i - b'\left(\sum_{j=1}^{p} \beta_j \eta_j(x_i) \right) \right] \quad J = 1, \cdots, p. \tag{2}$$

The observed Fisher information matrix $\mathcal{I}_n^o(\beta_1, \cdots, \beta_p)$ has IJth entry equal to

$$\mathcal{I}_{IJ}^o = \frac{\partial^2 \ell}{\partial \beta_I \partial \beta_J} = \sum_{i=1}^{n} \eta_I(x_i) \eta_J(x_i) b''\left(\sum_{j=1}^{p} \beta_j \eta_j(x_i) \right)$$

and this can be estimated by $\mathcal{I}_n^o(\hat{\beta}_1, \cdots, \hat{\beta}_p)$. The Fisher information matrix, $\mathcal{I}_n(\beta_1, \cdots, \beta_p)$, is the expected value of $\mathcal{I}_n^o(\beta_1, \cdots, \beta_p)$. For exponential families with natural parameterization, the observed information matrix and the information matrix are the same.

Usually, the p equations in (2) must be solved numerically, by iteratively reweighted least squares, a sort of Newton-Raphson method described in section 2.

Why don't we model $\mathrm{E}(Y)$ instead of η? That is, why don't we write $m(x) = \sum_{j=1}^{p} \beta_j m_j(x)$? Of course, the appropriate model depends on the data, and choosing an appropriate model should be the main concern. However, parameterizing in terms of η has a mathematical advantage. Usually, there are no restrictions on the value of $\eta(x)$. Often $m(x)$ is restricted in some way. For instance, $m(x)$ must be positive for the Poisson model. Therefore, if we parameterize in terms of m we must place restrictions on the β_j's and maximize a restricted likelihood. This is not a nice numerical problem, and calculations might be slow. However, not enforcing the restrictions on the β_j's could lead to values of the $\hat{\beta}_j$'s that are not permitted.

Example 1 continued. Suppose the Y_i's are independent Poisson random variables with the natural parameter for Y_i being $\eta(x_i) = \beta_1 + \beta_2 x_i$. Then the MLE's of β_1 and β_2 maximize

$$\ell(\beta_1, \beta_2) = \sum \Big[Y_i(\beta_1 + \beta_2 x_i) - \exp(\beta_1 + \beta_2 x_i) \Big],$$

that is, the MLE's satisfy the two equations

$$\frac{\partial \ell}{\partial \beta_1} = \sum \left[Y_i - \exp(\beta_1 + \beta_2 x_i) \right] = 0$$

and

$$\frac{\partial \ell}{\partial \beta_2} = \sum x_i \left[Y_i - \exp(\beta_1 + \beta_2 x_i) \right] = 0.$$

The Fisher information matrix is equal to

$$\mathcal{I}_n(\beta_1, \beta_2) = \begin{bmatrix} \sum \exp(\beta_1 + \beta_2 x_i) & \sum x_i \exp(\beta_1 + \beta_2 x_i) \\ \sum x_i \exp(\beta_1 + \beta_2 x_i) & \sum x_i^2 \exp(\beta_1 + \beta_2 x_i) \end{bmatrix}$$

which can be estimated by $\mathcal{I}_n(\hat{\beta}_1, \hat{\beta}_2)$. The asymptotic covariance matrix of $(\hat{\beta}_1, \hat{\beta}_2)$ is $\mathcal{I}^{-1}$, the asymptotic variance of $\hat{\eta}(x) \equiv \hat{\beta}_1 + \hat{\beta}_2 x$ is $(1, x)\mathcal{I}^{-1}(1, x)^t$ and the asymptotic variance of $\hat{m}(x) \equiv b'(\hat{\eta}(x))$ is, by the delta method, $[b''(\hat{\eta}(x))]^2 (1, x)\mathcal{I}^{-1}(1, x)^t$.

2 The Newton Raphson Algorithm and Iteratively Reweighted Least Squares

Section 2.1 describes how to numerically solve a system of p equations via the Newton Raphson algorithm. Section 2.2 describes the application of the algorithm to the solution of equation (2). Section 2.3 relates the algorithm to a weighted least squares problem.

2.1 The Newton Raphson Algorithm

The Newton Raphson algorithm is a numerical technique for solving p equations $F_1(\beta_1, \cdots, \beta_p) = 0$, $\cdots$, $F_p(\beta_1, \cdots, \beta_p) = 0$ for $\beta_1, \cdots, \beta_p$. It is most easily described using matrix notation. Let $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^t$, let $\mathbf{F}(\boldsymbol{\beta})$ be the p-vector $(F_1(\boldsymbol{\beta}), \cdots, F_p(\boldsymbol{\beta}))^t$ and let $\mathbf{F}'(\boldsymbol{\beta})$ denote the p by p matrix with

$$[\mathbf{F}'(\boldsymbol{\beta})]_{IJ} = \left.\frac{\partial F_I}{\partial \beta_J}\right|_{\boldsymbol{\beta}}.$$

We want to solve $\mathbf{F}(\boldsymbol{\beta}) = 0$ for $\boldsymbol{\beta}$. We start with an initial value $\boldsymbol{\beta}^0$ and let $\boldsymbol{\beta}^1 = \boldsymbol{\beta}^0 - [\mathbf{F}'(\boldsymbol{\beta}^0)]^{-1}\mathbf{F}(\boldsymbol{\beta}^0)$. We iterate: given $\boldsymbol{\beta}^L$ we let

$$\boldsymbol{\beta}^{L+1} = \boldsymbol{\beta}^L - [\mathbf{F}'(\boldsymbol{\beta}^L)]^{-1}\mathbf{F}(\boldsymbol{\beta}^L). \tag{3}$$

We stop iterating when $\|\boldsymbol{\beta}^{L+1} - \boldsymbol{\beta}^L\|$ is small.

The rationale behind this is that, by a Taylor series expansion about $\boldsymbol{\beta}^L$,

$$\mathbf{F}(\boldsymbol{\beta}^{L+1}) - \mathbf{F}(\boldsymbol{\beta}^L) \sim \mathbf{F}'(\boldsymbol{\beta}^L)(\boldsymbol{\beta}^{L+1} - \boldsymbol{\beta}^L).$$

Pretending that $\boldsymbol{\beta}^{L+1}$ is a solution, that is, that $\mathbf{F}(\boldsymbol{\beta}^{L+1}) = 0$, yields $-\mathbf{F}(\boldsymbol{\beta}^L) \sim \mathbf{F}'(\boldsymbol{\beta}^L)(\boldsymbol{\beta}^{L+1} - \boldsymbol{\beta}^L)$, which gives the iterative formula (3).

By making graphs of the iterative procedure in the case $p = 1$, we can easily see why the algorithm would work, and we can also see what convergence problems might arise. In particular, note that $\mathbf{F}'(\boldsymbol{\beta}^L)$ must be invertible.

Usually, code that uses a Newton-Raphson type algorithm will have two parameters: maxit, the maximum number of iterations (so the code doesn't run forever), and tolerance, usually equal to the "stop" value of $\|\boldsymbol{\beta}^{L+1} - \boldsymbol{\beta}^L\|$.

2.2 The Newton-Raphson Method Applied to Maximizing the Log Likelihood

Suppose that we have a parametric model with unknown parameters $\beta_1, \cdots, \beta_p$ and the maximum likelihood estimates of the β_j's are found by solving the p equations $\partial\ell/\partial\beta_j = 0$. Let $F_J = \partial\ell/\partial\beta_J$. Then $\mathbf{F}'(\boldsymbol{\beta}) = -\mathcal{I}_n^o(\boldsymbol{\beta})$, the observed Fisher Information matrix, and the iterative formula from section 2.1 is

$$\boldsymbol{\beta}^{L+1} = \boldsymbol{\beta}^L + [\mathcal{I}_n^o(\boldsymbol{\beta}^L)]^{-1}\left.\frac{\partial\ell}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}^L}. \tag{4}$$

The initial value of $\boldsymbol{\beta}$ is usually taken as the MLE for some very simple submodel. For instance, for the exponential family regression models of section

1.2, we could take $\boldsymbol{\beta}^0 = (\hat{\beta}_1^0, 0, \cdots, 0)^t$, where $\hat{\beta}_1^0$ is the MLE of η assuming that $\eta(x) \equiv$ a constant. That is, $\hat{\beta}_1^0$ is the $\hat{\eta}$ as found in section 1.1.

Typically the observed information matrix will have Y_i's in it. If, in the iterative algorithm, we use the information matrix instead of the observed information matrix, the resulting iterative method is called *Fisher scoring*. Either method can be used for likelihoods that are not based on an exponential family. Note that, for exponential families with natural parameterization, the observed information matrix is equal to the information matrix.

Example 1, continued. Suppose that Y_i is Poisson with natural parameter $\eta(x_i) = \beta_1 + \beta_2 x_i$. Then we can take $\beta_2^0 = 0$ and $\beta_1^0 = \ln(\bar{Y})$. Our next estimates of β_1 and β_2 are

$$\begin{pmatrix} \hat{\beta}_1^1 \\ \hat{\beta}_2^1 \end{pmatrix} = \begin{pmatrix} \ln \bar{Y} \\ 0 \end{pmatrix} + \begin{bmatrix} n\bar{Y} & n\bar{x}\bar{Y} \\ n\bar{x}\bar{Y} & \bar{Y}\sum x_i^2 \end{bmatrix}^{-1} \begin{pmatrix} 0 \\ \sum x_i(Y_i - \bar{Y}) \end{pmatrix}.$$

2.3 Maximum Likelihood as Iteratively Reweighted Least Squares: Using Pseudo-data

Suppose that we are carrying out the Newton Raphson algorithm to find the MLE's in the exponential family regression model of section 1.2. We can re-write (4) as the solution to a weighted least squares problem, regressing pseudo-data z_i's on the x_i's. Specifically suppose that we have $\boldsymbol{\beta}^L$ and thus we have $\hat{\eta}^L(x)$ and $\hat{m}^L(x) = b'(\hat{\eta}^L(x))$. Let

$$z_i = \hat{\eta}^L(x_i) + \frac{Y_i - \hat{m}^L(x_i)}{b''(\hat{\eta}^L(x_i))}$$

and consider the $\hat{\alpha}_j$'s that minimize

$$\sum_i \left(z_i - \sum_{j=1}^p \alpha_j \eta_j(x_i) \right)^2 b''(\hat{\eta}^L(x_i)). \tag{5}$$

This is a weighted least squares problem with weights $b''(\hat{\eta}^L(x_i))$. The minimization is easily carried out in Splus with, e.g., *lm*. To further see the weighted least squares connection, note that, if we pretend that $\hat{\eta}^L(\cdot)$ and $\hat{m}^L(\cdot)$ are non-random and are the true functions η and m, then

$$\mathrm{E}(z_i) = \hat{\eta}^L(x_i)$$

and

$$\mathrm{var}(z_i) = \frac{1}{b''(\hat{\eta}^L(x_i))}.$$

Thus, if the z_i's were normally distributed, then (5) would be the negative log-likelihood of $\boldsymbol{\alpha}$.

Below, we'll see that $\hat{\boldsymbol{\alpha}}$, the minimizer of (5) is equal to $\boldsymbol{\beta}^{L+1}$, that is, that $\hat{\eta}^{L+1}(x_i) = \hat{z}_i$, gotten by performing the weighted local least squares regression of the z_i's on the x_i's.

To see that $\hat{\boldsymbol{\alpha}} = \boldsymbol{\beta}^{L+1}$ write the z_i's and (5) in matrix notation: let $\mathbf{z} = (z_1, \cdots, z_n)$, $\boldsymbol{\eta}^L = (\hat{\eta}^L(x_1), \cdots, \hat{\eta}^L(x_n))^t$, $\mathbf{m}^L = (\hat{m}^L(x_1), \cdots, \hat{m}^L(x_n))^t$, $\mathbf{Y} = (Y_1, \cdots, Y_n)^t$, $\mathbf{X}$ the n by p design matrix with $\mathbf{X}_{ij} = \eta_j(x_i)$ and $\mathbf{B}$ the p by p diagonal matrix with $\mathbf{B}_{ii} = b''(\hat{\eta}^L(x_i))$. Then

$$\mathbf{z} = \boldsymbol{\eta}^L + \mathbf{B}^{-1}(\mathbf{Y} - \mathbf{m}^L) = \mathbf{X}\boldsymbol{\beta}^L + \mathbf{B}^{-1}(\mathbf{Y} - \mathbf{m}^L)$$

and (5) becomes

$$(\mathbf{z} - \mathbf{X}\boldsymbol{\alpha})'\mathbf{B}(\mathbf{z} - \mathbf{X}\boldsymbol{\alpha})$$

which has minimizer

$$\hat{\boldsymbol{\alpha}} = (\mathbf{X}^t\mathbf{B}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{B}\mathbf{z} = \boldsymbol{\beta}^L + (\mathbf{X}^t\mathbf{B}\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{Y} - \mathbf{m}^L).$$

We easily see that, for this exponential family model, $\partial\ell/\partial\boldsymbol{\beta} = \mathbf{X}^t(\mathbf{Y} - \mathbf{m}^L)$ and $\mathcal{I}_n(\boldsymbol{\beta}^{L+1}) = \mathbf{X}^t\mathbf{B}\mathbf{X}$. Therefore $\hat{\boldsymbol{\alpha}} = \boldsymbol{\beta}^{L+1}$.

Example 1 continued. Suppose that we take $\hat{\beta}_1^0 = \ln \bar{Y}$ and $\hat{\beta}_2^0 = 0$. Then $\hat{\beta}_1^1$ and $\hat{\beta}_2^1$ solve the following weighted least squares problem. Let $z_i = \ln \bar{Y} + (Y_i - \bar{Y})/\bar{Y}$ and find $\hat{\alpha}_1 = \hat{\beta}_1^1$ and $\hat{\alpha}_2 = \hat{\beta}_2^1$ to minimize $\sum(z_i - \alpha_1 - \alpha_2 x_i)^2 \bar{Y}$.

3 Local Polynomial Estimation

Consider the exponential family regression model of section 1.2, but now suppose that we only assume that $\eta(x)$ is smooth. To estimate $\eta(x)$, we use a local likelihood, near x approximating η by a polynomial of degree p. Specifically, we find $\hat{\beta}_j$, $j = 1, \cdots, p+1$, that maximize the local log likelihood:

$$\sum_{i=1}^{n} \left[Y_i \sum_{j=1}^{p+1} \beta_j (x_i - x)^{j-1} - b \left(\sum_{j=1}^{p+1} \beta_j (x_i - x)^{j-1} \right) \right] K\left(\frac{x_i - x}{h}\right). \tag{6}$$

This maximization is done numerically, by iteratively reweighted least squares. Note that the second partial derivative of the log likelihood with respect to β_I and β_J is $-\sum_i b''(\sum_j \beta_j (x_i - x)^{j-1})(x_i - x)^{I+J-2} K((x_i - x)/h)$, which, for positive K and for most x_i's, forms a negative definite matrix. So any critical point of the log likelihood is unique and is a maximizer.

We let $\hat{\eta}(x) = \hat{\beta}_1$ and

$$\widehat{\eta^{(j-1)}}(x) = (j-1)!\hat{\beta}_j.$$

The estimate of $m(x)$ is $b'(\hat{\eta}(x)) = b'(\hat{\beta}_1)$. The estimate of $m'(x) = b''(\eta(x))\eta'(x)$ is $b''(\hat{\eta}(x))\widehat{\eta'}(x) = b''(\hat{\beta}_1)\hat{\beta}_2$. The estimate of $m''(x) = b'''(\eta(x))[\eta'(x)]^2 + b''(\eta(x))\eta''(x)$ is $b'''(\hat{\beta}_1)[\hat{\beta}_2]^2 + 2b''(\hat{\beta}_1)\hat{\beta}_3$. Estimates of higher order derivatives of m are calculated similarly.

Exercise 1. The calculations in this exercise will mimic the development of the parametric case in section 2. Suppose that the Y_i's are Bernouilli with $\mathrm{P}\{Y_i = 1|x_i\} = p(x_i)$. (a.) What is the natural parameter $\eta(x_i)$? What is the log likelihood? Fix x and consider estimating $\eta(x)$ by a local linear fit, finding β_1 and β_2 to minimize

$$\sum_{i=1}^{n} \Big[Y_i(\beta_1 + \beta_2(x_i - x)) - b\left(\beta_1 + \beta_2(x_i - x)\right) \Big] K\left(\frac{x_i - x}{h}\right).$$

This can be done iteratively. (b.) Find the formula for an initial estimate of (β_1, β_2). (c.) For the Newton Raphson algorithm, what is the formula analogous to (4) for $\beta_1^{L+1}, \beta_2^{L+1}$, given β_1^L, β_2^L? (d.) How can the formula in (c.) be gotten from local polynomial fitting? Explain/prove your answer.

4 Effective Number of Parameters

Recall that, if $\hat{\eta}$ is linear in the data, that is, if $(\hat{\eta}(x_1), \cdots, \hat{\eta}(x_n))' = S(Y_1, \cdots, Y_n)'$ for some matrix S that doesn't depend on the Y_i's, we can define the effective number of parameters used by $\hat{\eta}$ as the trace of S.

For exponential models that aren't normal, we don't have $\hat{\eta}$ linear in the data. However for most smoothing methods, at each stage of a modified iteratively reweighted least squares algorithm, we have: $(\hat{\eta}^{L+1}(x_1), \cdots, \hat{\eta}^{L+1}(x_n))' = S^L(\hat{\eta}^L(x_1), \cdots, \hat{\eta}^L(x_n))'$. Here S^L typically depends on the Y_i's, but we ignore this fact. We take the effective number of parameters to be the trace of this S^L matrix, at the time when the iteration has stopped.

5 Asymptotics

This is modified from Fan, Heckman and Wand (*JASA*, 1995), who studied quasi-likelihood estimation in the random design case. The proof is given in section 7.

Theorem. Suppose that

a. K is a kernel with $K(u) = 0$ for $|u| > 1$ and with K' Lipschitz;

b. the x_i's have a regular design with design density f, f Lipschitz;

c. $\eta^{(p+1)}$ exists and is continuous on [0,1];

d. $h \to 0$, nh^{2p+3} bounded, and $nh^2 \to \infty$.

Let A, C, and $\mathbf{v}$ be as in the local least squares case: $A_{ij} = \nu_{i+j-2}(K, x, h)$, $C_{ij} = \nu_{i+j-2}(K^2, x, h)$, and $\mathbf{v}_i = \nu_{i+p}(K, x, h)$. Then, for $p - i$ even,

$$\mathrm{E}((i-1)!\hat{\beta}_i) = \eta^{(i-1)}(x) + (i-1)!\frac{\eta^{(p+1)}(x)}{(p+1)!}[A^{-1}\mathbf{v}]_i h^{p+2-i} + o(h^{p+2-i})$$

and

$$\mathrm{var}((i-1)!\hat{\beta}_i) = ((i-1)!)^2 \frac{1}{f(x)b''(\eta(x))}[A^{-1}CA^{-1}]_{ii}\frac{1}{nh^{2i-1}} + o\left(\frac{1}{nh^{2i-1}}\right).$$

Thus, using the kernel $K_{r,p}$ from the local polynomial handout, for $p-r$ odd,

$$\mathrm{E}\left(\widehat{\eta^{(r)}}(x)\right) = \eta^{(r)}(x) + \frac{\eta^{(p+1)}(x)}{(p+1)!}h^{p-r+1}\nu_{p+1}(K_{r,p}, x, h) + o(h^{p-r+1})$$

and

$$\mathrm{var}\left(\widehat{\eta^{(r)}}(x)\right) = \frac{1}{f(x)b''(\eta(x))}\frac{1}{nh^{2r+1}}\nu_0(K_{r,p}^2, x, h) + o\left(\frac{1}{nh^{2r+1}}\right).$$

Comments. If $p - r$ is even one can show that the bias in the interior of the interval is of a larger order of magnitude than the bias at the boundary. Therefore we take p so that $p - r$ is odd.

Optimal bandwidth: Assume that $p - r$ is odd. Then we can use the expressions for the asymptotic bias and variance of $\widehat{\eta^{(r)}}(x)$ to find $h_{opt,r,p}$, the bandwidth that minimizes the asymptotic MSE or the asymptotic integrated MSE. The AMSE (or integrated AMSE) is of the order $h^{2p-2r+2} + 1/(nh^{2r+1})$ and $h_{opt,r,p}$ is of order $n^{-1/(2p+3)}$.

Exercise 2. Suppose that the Y_i's are independent Bernouili's with $\mathrm{P}\{Y_i = 1\} = p(x_i)$. Find the asymptotic bias and variance of the local linear estimate of $\eta(x)$, the natural parameter. Find the asymptotic bias and variance of the resulting estimate of $p(x)$. Suppose that the x_i's are equally spaced on $[0, 1]$ and that the conditions of the Theorem are satisfied.

6 Generalized Additive Models: Model Checking

6.1 The Deviance of a Model

In usual parametric regression, one can test if certain parameters in a model are equal to zero using a χ^2 test or an F test (the latter really only holds for normal data).

Suppose that Y_i, $i = 1, \cdots, n$, are independent with the log likelihood of Y_i equal to $l(\eta(\mathbf{x}_i), Y_i)$. So the full likelihood under η is

$$\ell(\eta, \mathbf{Y}) = \sum_{i=1}^{n} l(\eta(\mathbf{x}_i), Y_i).$$

Suppose there are no restrictions on η. This is called the *saturated model* and the MLE's of the $\eta(\mathbf{x}_i)$'s under this model are denoted $\eta^S(\mathbf{x}_i)$. If the $\mathbf{x}_i$'s are distinct, then typically the $\eta^S(\mathbf{x}_i)$'s are based on an interpolation of the data.

Example. Suppose the Y_i's are independent normals with $\mathrm{E}(Y_i) = \eta(\mathbf{x}_i)$ and variance σ^2. Then

$$\ell(\eta, \mathbf{Y}) = -\frac{n}{2} \ln[2\pi\sigma^2] - \frac{1}{2\sigma^2} \sum_{1}^{n} \left(Y_i - \eta(\mathbf{x}_i)\right)^2 ,$$

and, if the $\mathbf{x}_i$'s are distinct,

$$\eta^S(\mathbf{x}_i) = Y_i \quad \text{and} \quad \ell(\eta^S, \mathbf{Y}) = -\frac{n}{2}\ln[2\pi\sigma^2].$$

If the Y_i's are from an exponential family with natural parameter function $\eta(\cdot)$, corresponding mean function $\mu(\cdot)$, and dispersion parameter ϕ, then

$$\ell(\eta, \mathbf{Y}) = \sum_{i=1}^{n}\left[\frac{Y_i\eta(\mathbf{x}_i) - b(\eta(\mathbf{x}_i))}{\phi} + d(Y_i, \phi)\right]$$

and, if the $\mathbf{x}_i$'s are distinct, $\eta^S(\mathbf{x}_i)$ satisfies

$$0 = Y_i - b'(\eta^S(\mathbf{x}_i)) = Y_i - \mu^S(\mathbf{x}_i) \Rightarrow \mu^S(\mathbf{x}_i) = Y_i.$$

The *deviance* for a proposed η^M is defined as

$$D(\mathbf{Y}, \eta^M) = 2\phi[\ell(\eta^S, \mathbf{Y}) - \ell(\eta^M, \mathbf{Y})].$$

Notice the deviance is always non-negative. A small deviance means that the model η^M fits the data well.

If the Y_i's are normal with variance σ^2, then the dispersion parameter is σ^2, if the $\mathbf{x}_i$'s are distinct, then $\eta^S(\mathbf{x}_i) = Y_i$, $\ell(\eta^S, \mathbf{Y}) = 0$ and so

$$D(\mathbf{Y}, \eta^M) = \sum_{i=1}^{n}\left(Y_i - \eta^M(\mathbf{x}_i)\right)^2.$$

Suppose we're comparing two models for η, say η^1 and η^2, and that η^1 is a submodel of η^2. Then we usually compare η^1 and η^2 by comparing their deviances:

$$\left[D(\mathbf{Y}, \eta^1) - D(\mathbf{Y}, \eta^2)\right]/\phi = 2[\ell(\eta^2, \mathbf{Y}) - \ell(\eta^1, \mathbf{Y})].$$

If this change in deviance is small, the fit of the two η's is about the same.

6.2 Parametric Models

Consider the parametric model of η: $\eta(x) = \eta(x; \beta_1, \cdots, \beta_{d+p})$ with the β_j's unknown. We can test

$$H_o : \beta_{p+j} = 0, \quad j = 1, \cdots, d$$

via a generalized log-likelihood ratio test. Let $\hat{\beta}_j^o$, $j = 1, \cdots, p$, be the MLE's under the null model and $\hat{\beta}_j^F$, $j = 1, \cdots, p+d$, the MLE's under the full model. Then the generalized log-likelihood ratio statistic is

$$\begin{aligned} T_n &= 2\Big[\ell(\eta(\cdot; \hat{\beta}_1^o, \cdots, \hat{\beta}_p^{p+d}), \mathbf{Y}) - \ell(\eta(\cdot; \hat{\beta}_1^o, \cdots, \hat{\beta}_p^o), \mathbf{Y})\Big] \\ &= \Big[D(\mathbf{Y}, \eta(\cdot; \hat{\beta}_1^o, \cdots, \hat{\beta}_p^o)) - D(\mathbf{Y}, \eta(\cdot; \hat{\beta}_1^F, \cdots, \hat{\beta}_{p+d}^F))\Big] / \phi. \end{aligned}$$

Under conditions on ℓ, under H_o T_n is asymptotically chi-squared with d degrees of freedom and we reject H_o if T_n is large. If ϕ is unknown, we can replace it with a consistent es timate of ϕ, and still use this chi-squared test.

Suppose that the Y_i's are independent normals with $\mathrm{E}(Y_i) = \eta(\mathbf{x}_i)$ and with common variance σ^2 and that our full model is

$$\eta(\mathbf{x}) = \sum_{j=1}^{p+d} \beta_j \phi_j(\mathbf{x})$$

where the ϕ_j's are known. Then

$$T_n = \frac{1}{\sigma^2}(RSS^o - RSS^F)$$

where

$$RSS^o = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \hat{\beta}_j^o \phi_j(\mathbf{x}_i) \right)^2$$

and

$$RSS^F = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p+d} \hat{\beta}_j^F \phi_j(\mathbf{x}_i) \right)^2 .$$

Under H_o, T_n is asymptotically χ^2 with d degrees of freedom. If σ^2 is unknown, we can use a consistent estimate in T_n. Recall that we can also use an F test, which is exact for normal data.

6.3 In Smooth Models

We use the ideas from the parametric χ^2 test and F test to check fits in generalized additive models. Note: there's no theory to support this! It's simply exploratory, but works well in practice. We should only compare two models that are nested (e.g. one is a submodel of the other). For instance $y \sim x_1 + lo(x_2)$ is a submodel of $y \sim lo(x_1) + lo(x_2)$. This latter model is approximately (!) a submodel of $y \sim lo(x_1, x_2)$, if we adjust the amount of smoothing appropriately. However $y \sim x_1$ isn't a submodel of $y \sim x_2$. We calculate deviances of the various models of interest and use effective number of parameters to calculate degrees of freedom.

Effective Number of Parameters

Recall that, when we have just one independent variable, we define the effective number of parameters as the trace of S^L, where, at the final stage of the iteratively reweighted least squares algorithm, $(\hat\eta^{L+1}(x_1), \cdots, \hat\eta^{L+1}(x_n))' = S^L(\hat\eta^L(x_1), \cdots, \hat\eta^L(x_n))'$. Since *gam* reports the intercept and the slope parameter separately from the smooth, in the case of one regressor, *gam* would say that the smoothed part of $\hat\eta$ uses $\mathrm{tr} S^L - 2$ degrees of freedom.

However things are a bit complicated with *gam* when there is more than one regressor. In the backfitting step, each independent variable has an associated hat matrix, S_j. The estimate $\hat\eta$ and the "final" hat matrix S^* depend in a complicated way on these individual hat matrices. However, *gam* ignores this complication, and still reports degrees of freedom in terms of the individual hat matrices. I think that's what *gam* does!

Standard Errors

First consider the normal model case. Suppose that in the back-fitting algorithm, we're using linear regression/smoothers, so there exist hat matrices that define the fixed point: S_j^*, $j = 1, \cdots, d$ with $\hat{\mathbf{m}}_j = S_j \mathbf{Y}$. (These are NOT the hatmatrices S_j used in the individual smooths.) Thus, for $\mathrm{var}(Y_i) = \sigma^2$, the covariance matrix of $\hat{\mathbf{m}}_j$ is

$$\mathrm{cov}(\hat{\mathbf{m}}_j) = \sigma^2 S_j^* S_j^{*t}$$

and

$$\mathrm{var}(\hat m_j(x_{ij})) = \sigma^2 [S_j^* S_j^{*t}]_{ii}.$$

For iteratively reweighted least squares, we are estimating $\eta(\mathbf{x})$ by $m(\mathbf{x}) = \alpha + \sum_1^d m_j(x_j)$. To find standard errors of $\hat m_j(x_{ij})$ we replace Y_i in the above with z_i, the pseudo-data at the last iteration in the iteratively reweighted least squares algorithm. So $\hat{\mathbf{m}}_j = S_j^* \mathbf{z}$ and, since we've pretended that, in the back-fitting algorithm, $\mathrm{var}(z_i) = 1/b''(\hat{\eta}(\mathbf{x}_i)) = 1/b''(\hat{m}(\mathbf{x}_i))$,

$$\mathrm{cov}(\hat{\mathbf{m}}_j) = S_j^* W S_j^{*t}$$

and

$$\mathrm{var}(\hat{m}_j(x_{ij})) = [S_j^* W S_j^{*t}]_{ii}$$

where W is an n by n diagonal matrix with $W_{ii} = 1/b''(\hat{m}(\mathbf{x}_i))$.

The matrices S_j^* take a long time to compute, so *gam* uses an appoximation (not S_j as with the effective number of parameters).

7 Proof of the Theorem

To study the asymptotic properties of $\hat{\beta}_j$, we want to re-write (6) so that it's close to being quadratic in the β_j's and then maximize the quadratic expression. We'll use a Taylor series expansion in a modified form of (6) to get an expression that is roughly quadratic. We'll make use of the following lemma (see Pollard, *Econometric Theory*, 1991, and Fan, Heckman and Wand, *JASA*, 1995).

Lemma. Suppose that L_n is a sequence of (possibly random) functions defined on Θ, an open convex subset of $\Re^{p+1}$, and that, for each n, $\hat{\boldsymbol{\theta}}_n$ the maximizer of L_n exists. Suppose also that

$$L_n(\boldsymbol{\theta}) = U_n^t \boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}^t W_n \boldsymbol{\theta} + f_n(\boldsymbol{\theta})$$

for each fixed $\boldsymbol{\theta}$. Here $\mathbf{U}_n$ and W_n do not depend on $\boldsymbol{\theta}$, U_n is stochastically bounded, W_n is non-random, and $W_n \to F$. Suppose that $\alpha_n \to 0$, $f_n(\boldsymbol{\theta}) = o(\alpha_n)$, $f_n'(\boldsymbol{\theta}) = o(\alpha_n)$, and $f_n''(\boldsymbol{\theta}) = o(\alpha_n)$ for each fixed $\boldsymbol{\theta}$. Then

$$\hat{\boldsymbol{\theta}}_n = W_n^{-1}\mathbf{U}_n + o(\alpha_n).$$

Definition. $\mathbf{U}_n$ is *stochastically bounded* if

$$\lim_{C\to\infty} \limsup_{n\to\infty} P\{\|\mathbf{U}_n\| > C\} = 0.$$

If $\mathbf{U}_n$ converges in distribution to $\mathbf{U}$, finite, then $\mathbf{U}_n$ is stochastically bounded. If $\limsup |\mathrm{E}(Z_n)| < \mathrm{E}_0 < \infty$ and $\limsup \operatorname{var}(Z_n) < V_0 < \infty$, then Z_n is stochastically bounded, by the following argument.

$$P\{|Z_n| > C\} \le \frac{\mathrm{E}(Z_n^2)}{C^2} = \frac{\mathrm{E}^2(Z_n) + \operatorname{var}(Z_n)}{C^2} < \frac{\mathrm{E}_0^2 + V_0}{C^2} \to 0$$

as $C \to \infty$.

Comment. Note that the critical point of $\mathbf{U}_n^t \boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}^t W_n \boldsymbol{\theta}$ satisfies $\mathbf{U}_n - W_n \boldsymbol{\theta}^* = 0$, that is $\boldsymbol{\theta}^* = W_n^{-1}\mathbf{U}_n$.

Back to the proof of the Theorem.

Let

$$\beta_j^* = \frac{\eta^{(j-1)}(x)}{(j-1)!}.$$

If the $\hat{\beta}_j$'s maximize (6), then they maximize

$$\sum_{i=1}^n Y_i \sum_{j=1}^{p+1} (\beta_j - \beta_j^*)(x_i - x)^{j-1} K\left(\frac{x_i - x}{h}\right)$$

$$-\sum_{i=1}^n \left[b\left(\sum_{j=1}^{p+1} \beta_j (x_i - x)^{j-1}\right) - b\left(\sum_{j=1}^{p+1} \beta_j^* (x_i - x)^{j-1}\right)\right] K\left(\frac{x_i - x}{h}\right)$$

$$= \sum_{i=1}^n Y_i \sum_{j=1}^{p+1} (\beta_j - \beta_j^*)(x_i - x)^{j-1} K\left(\frac{x_i - x}{h}\right)$$

$$-\sum_{i=1}^n b'(\bar{\eta}_i) \sum_{j=1}^{p+1} (\beta_j - \beta_j^*)(x_i - x)^{j-1} K\left(\frac{x_i - x}{h}\right)$$

$$-\frac{1}{2}\sum_{i=1}^n b''(\bar{\eta}_i) \left[\sum_{j=1}^{p+1} (\beta_j - \beta_j^*)(x_i - x)^{j-1}\right]^2 K\left(\frac{x_i - x}{h}\right)$$

$$+e_n$$

where

$$\bar{\eta}_i = \sum_{j=1}^{p+1} \beta_j^* (x_i - x)^{j-1} = \sum_{j=0}^{p} \frac{\eta^{(j)}(x)}{j!}(x_i - x)^j.$$

and

$$e_n = \sum_{i=1}^{n} b'(\bar{\eta}_i) \sum_{j=1}^{p+1} (\beta_j - \beta_j^*)(x_i - x)^{j-1} K\left(\frac{x_i - x}{h}\right)$$

$$+\frac{1}{2} \sum_{i=1}^{n} b''(\bar{\eta}_i) \left[\sum_{j=1}^{p+1} (\beta_j - \beta_j^*)(x_i - x)^{j-1}\right]^2 K\left(\frac{x_i - x}{h}\right)$$

$$-\left[b\left(\sum_{j=1}^{p+1} \beta_j (x_i - x)^{j-1}\right) - b(\bar{\eta}_i)\right].$$

Rewrite this using a normalization for derivative estimation: let

$$\theta_j = \sqrt{nh}\; h^{j-1}(\beta_j - \beta_j^*) \quad j = 1, \cdots, p+1.$$

So we want to maximize

$$L_n(\boldsymbol{\theta}) = \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} [Y_i - b'(\bar{\eta}_i)] \left[\sum_{j=1}^{p+1} \theta_j \left(\frac{x_i - x}{h}\right)^{j-1}\right] K\left(\frac{x_i - x}{h}\right)$$

$$-\frac{1}{2nh} \sum_{i=1}^{n} b''(\bar{\eta}_i) \left[\sum_{j=1}^{p+1} \theta_j \left(\frac{x_i - x}{h}\right)^{j-1}\right]^2 K\left(\frac{x_i - x}{h}\right) + f_n(\boldsymbol{\theta})$$

where

$$f_n(\boldsymbol{\theta}) = \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} b'(\bar{\eta}_i) \sum_{j=1}^{p+1} \theta_j \left(\frac{x_i - x}{h}\right)^{j-1} K\left(\frac{x_i - x}{h}\right)$$

$$+\frac{1}{2nh} \sum_{i=1}^{n} b''(\bar{\eta}_i) \left[\sum_{j=1}^{p+1} \theta_j \left(\frac{x_i - x}{h}\right)^{j-1}\right]^2 K\left(\frac{x_i - x}{h}\right)$$

$$-\sum_{i=1}^{n} \left[b\left(\bar{\eta}_i + \frac{1}{\sqrt{nh}} \sum_{j=1}^{p+1} \theta_j \left(\frac{x_i - x}{h}\right)^{j-1}\right) - b(\bar{\eta}_i)\right] K\left(\frac{x_i - x}{h}\right).$$

We'll show that $\hat{\boldsymbol{\theta}}_n$, the maximizer of L_n, satisfies

$$\mathrm{E}(\hat{\boldsymbol{\theta}}_n) = \sqrt{nh} h^{p+1} \frac{\eta^{(p+1)}}{(p+1)!} A^{-1} \mathbf{v} + o(\sqrt{nh} h^{p+1})$$

and

$$\mathrm{cov}(\hat{\boldsymbol{\theta}}_n) = \frac{1}{f(x)b''(\eta(x))} A^{-1}CA^{-1}.$$

This gives us

$$\mathrm{E}(\hat{\beta}_i - \beta_i^*) = h^{p+2-i}\frac{\eta^{(p+1)}(x)}{(p+1)!}(A^{-1}\mathbf{v})_i + o(h^{p+2-i})$$

and

$$\mathrm{var}(\hat{\beta}_i) = \frac{1}{nh^{2i-1}}\frac{1}{f(x)b''(\eta(x))}[A^{-1}CA^{-1}]_{ii} + o\left(\frac{1}{nh^{2i-1}}\right).$$

The first term in $L_n(\boldsymbol{\theta})$ is equal to

$$\sum_{j=1}^{p+1}\theta_j\left[\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}(Y_i - b'(\bar{\eta}_i))\left(\frac{x_i - x}{h}\right)^{j-1} K\left(\frac{x_i-x}{h}\right)\right] \equiv \sum_{j=1}^{p+1}\theta_j[\mathbf{U}_n]_j = \boldsymbol{\theta}'\mathbf{U}_n.$$

The second term in $L_n(\boldsymbol{\theta})$ is equal to

$$-\frac{1}{2}\sum_{j=1}^{p+1}\sum_{k=1}^{p+1}\theta_j\theta_k\left[\frac{1}{nh}\sum_{i=1}^{n}b''(\bar{\eta}_i)\left(\frac{x_i-x}{h}\right)^{j+k-2} K\left(\frac{x_i-x}{h}\right)\right] \equiv -\frac{1}{2}\sum_{j=1}^{p+1}\sum_{k=1}^{p+1}\theta_j\theta_k[W_n]_{jk}$$

$$= -\frac{1}{2}\boldsymbol{\theta}'W_n\boldsymbol{\theta}.$$

Therefore

$$L_n(\boldsymbol{\theta}) = \mathbf{U}_n^t\boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}^t W_n\boldsymbol{\theta} + f_n(\boldsymbol{\theta}).$$

To apply the lemma, first consider $\mathbf{U}_n$. I'll show that $[\mathbf{U}_n]_j$ is stochastically bounded by studying its mean and variance.

$$\mathrm{E}\left([\mathbf{U}_n]_j\right) = \frac{1}{\sqrt{nh}}\sum_i [b'(\eta(x_i)) - b(\bar{\eta}_i]\left(\frac{x_i-x}{h}\right)^{j-1} K\left(\frac{x_i-x}{h}\right).$$

But

$$b'(\eta(x_i)) - b'(\bar{\eta}_i) = b''(\eta(x))\left(\eta(x_i) - \bar{\eta}_i\right) + o\left(\eta(x_i) - \bar{\eta}_i\right)^2$$

and, for $|x_i - x| \le h$,

$$\eta(x_i) - \bar{\eta}_i = \eta(x_i) - \sum_{l=0}^{p}\frac{\eta^{(l)}(x)}{l!}(x_i - x)^l = \frac{\eta^{(p+1)}(x)}{(p+1)!}(x_i - x)^{p+1} + o(h^{p+1}).$$

Therefore

$$
\begin{aligned}
\mathrm{E}\left([\mathbf{U}_n]_j\right) &= \frac{1}{\sqrt{nh}}\sum_i \left(\frac{x_i-x}{h}\right)^{j-1} K\left(\frac{x_i-x}{h}\right) b''(\eta(x))\left(\frac{\eta^{(p+1)}(x)}{(p+1)!}(x_i-x)^{p+1}+o(h^{p+1})\right) \\
&= b''(\eta(x))\frac{\eta^{(p+1)}(x)}{(p+1)!}\sqrt{nh}h^{p+1}\frac{1}{nh}\sum_i \left[\left(\frac{x_i-x}{h}\right)^{p+j}+o(1)\right] K\left(\frac{x_i-x}{h}\right) \\
&= b''(\eta(x))\frac{\eta^{(p+1)}(x)}{(p+1)!}\sqrt{nh}h^{p+1}\left[f(x)\nu_{p+j}(K)+O\left(\frac{1}{nh^2}\right)+o(1)\right] \\
&= b''(\eta(x))\frac{\eta^{(p+1)}(x)}{(p+1)!}\sqrt{nh}h^{p+1}f(x)\nu_{p+j}(K)+o(\sqrt{nh}h^{p+1}) \\
&= O(1)
\end{aligned}
$$

since nh^{2p+3} is bounded. Therefore

$$
\mathrm{E}(\mathbf{U}_n) = \sqrt{nh}h^{p+1}b''(\eta(x))\frac{\eta^{(p+1)}(x)}{(p+1)!}f(x)\mathbf{v}+o(\sqrt{nh}h^{p+1}). \tag{7}
$$

.

I'll analyze the covariance between $[\mathbf{U}_n]_i$ and $[\mathbf{U}_n]_j$, since this will be used later.

$$
\begin{aligned}
&\mathrm{cov}\left([\mathbf{U}_n]_i,[\mathbf{U}_n]_j\right) \\
=\ &\mathrm{cov}\Bigg(\frac{1}{\sqrt{nh}}\sum_k (Y_k-b'(\bar\eta_k))\left(\frac{x_k-x}{h}\right)^{i-1}K\left(\frac{x_k-x}{h}\right), \\
&\quad \frac{1}{\sqrt{nh}}\sum_l (Y_l-b'(\bar\eta_l))\left(\frac{x_l-x}{h}\right)^{j-1}K\left(\frac{x_l-x}{h}\right)\Bigg) \\
=\ &\frac{1}{nh}\sum_{k,l}\left(\frac{x_k-x}{h}\right)^{i-1}\left(\frac{x_l-x}{h}\right)^{j-1}K\left(\frac{x_k-x}{h}\right)K\left(\frac{x_l-x}{h}\right)\mathrm{cov}(Y_k,Y_l) \\
=\ &\frac{1}{nh}\sum_k\left(\frac{x_k-x}{h}\right)^{i+j-2}K^2\left(\frac{x_k-x}{h}\right)b''(\eta(x_k)) \\
=\ &b''(\eta(x))f(x)\nu_{i+j-2}(K^2)+O\left(\frac{1}{nh^2}\right)+o(1).
\end{aligned}
$$

So the variance of $[\mathbf{U}_n]_i$ is bounded and therefore $\mathbf{U}_n$ is stochastically bounded. Furthermore

$$
\mathrm{cov}(\mathbf{U}_n) = b''(\eta(x))f(x)C+o(1). \tag{8}
$$

Now consider W_n.

$$\begin{aligned}[W_n]_{ij} &= \frac{1}{nh}\sum_k \left(\frac{x_k - x}{h}\right)^{i+j-2} b''(\bar{\eta}_k) K\left(\frac{x_k - x}{h}\right) \\ &= \frac{1}{h}\int_u \left(\frac{u-x}{h}\right)^{i+j-2} K\left(\frac{u-x}{h}\right) g(u)\ du \ + \ O\left(\frac{1}{nh^2}\right)\end{aligned}$$

where

$$g(u) = f(u)\ b''\left(\sum_j \beta_j^*(u-x)^{j-1}\right).$$

Substituting $s = (u-x)/h$ yields

$$\begin{aligned}[W_n]_{ij} &= \int_s s^{i+j-2} K(s) g(x+sh)\ ds \ + \ O\left(\frac{1}{nh^2}\right) \\ &= g(x)\nu_{i+j-2}(K) \ + \ o(1) \ + \ O\left(\frac{1}{nh^2}\right) \\ &= f(x) b''(\eta(x)) A_{ij} + o(1).\end{aligned}$$

We omit the proof that $f_n(\boldsymbol{\theta}) = o(h^{p+1})$, $f_n'(\boldsymbol{\theta}) = o(h^{p+1})$, and $f_n''(\boldsymbol{\theta}) = o(h^{p+1})$.

By the lemma,

$$\hat{\boldsymbol{\theta}}_n = W_n^{-1}\mathbf{U}_n + o(h^{p+1}).$$

So

$$\begin{aligned}\mathrm{E}(\hat{\boldsymbol{\theta}}_n) &= W_n^{-1}\mathrm{E}(\mathbf{U}_n) + o(h^{p+1}) \\ &= [f(x)b''(\eta(x))A + o(1)]^{-1}\left(\sqrt{nh}h^{p+1}\frac{\eta^{(p+1)}}{(p+1)!}\mathbf{v} + o(\sqrt{nh}h^{p+1})\right) \\ &= \frac{1}{f(x)b''(\eta(x))}A^{-1}\left(\sqrt{nh}h^{p+1}\frac{\eta^{(p+1)}}{(p+1)!}\mathbf{v}\right) + o(\sqrt{nh}h^{p+1})\end{aligned}$$

and

$$\begin{aligned}\mathrm{cov}(\hat{\boldsymbol{\theta}}_n) &= W_n^{-1}\mathrm{cov}(\mathbf{U}_n)W_n^{-1} + o(h^{p+1}) \\ &= \frac{1}{f(x)b''(\eta(x))}A^{-1}CA^{-1} + o(1).\end{aligned}$$