Nonparametric Estimation of Regression Functions In the Presence of Irrelevant Regressors

Peter Hall, Qi Li, Jeff Racine

1 Introduction

Nonparametric techniques – robust to functional form specification.
“curse of dimensionality” and the number of discrete cells.
The conventional frequency estimation method.
Smooth the discrete variables, Aitchison and Aitken (1976).
Bandwidth selection: Data-driven methods.

Irrelevant covariates (regressors) – should be removed.
Testing procedure – not powerful and not efficient.
The data-driven cross-validation method.
Automatic removal of irrelevant variables.
2 Regression Models with Irrelevant Regressors

\[ X_d^i, \ q \times 1 \ \text{vector of discrete variables.} \]

\[ X_c^i \in R^p \ \text{continuous regressors.} \]

\[ X_{is}^d, \ \text{the } \textit{s} \text{th component of } X_d^i, \]

\[ X_{is}^d \ \text{takes } c_s \geq 2 \ \text{different values.} \]

\[ X_i = (X_d^i, X_c^i), \]

We want to estimate \( E(Y_i|X_i) \)

The first \( p_1 \ (1 \leq p_1 \leq p) \) components of \( X_c \) and the first \( q_1 \ (0 \leq q_1 \leq q) \) components of \( X_d \) are “relevant” regressors.

Let \( \bar{X} \) be relevant components of \( X \) and let \( \bar{X} = X \setminus \{X\} \) denote the irrelevant components of \( X \).

We assume that

\[(\bar{X}, Y) \ \text{is independent of } \bar{X} \quad (2.1)\]

(2.1) implies that

\[ E(Y|X) = E(Y|\bar{X}) , \quad (2.2) \]

So that \( Y = g(X) + \text{error,} \)

\[ Y_i = \bar{g}(\bar{X}_i) + u_i, \quad (2.3) \]

\[ E(u_i|\bar{X}_i) = 0. \]
For $1 \leq s \leq q$, define
\[
I(X_{is}^d, x_s^d, \lambda_s) = \begin{cases} 
1 & \text{if } X_{is}^d = x_s^d, \\
\lambda_s & \text{if } X_{is}^d \neq x_s^d. 
\end{cases} \tag{2.4}
\]
The product kernel for $x^d = (x_1^d, \ldots, x_q^d)$ is given by
\[
K^d(x^d, X_i^d) = \prod_{s=1}^{q} I(X_{is}^d, x_s^d, \lambda_s). 
\]

$0 \leq \lambda_s \leq 1$. When $\lambda_s = 1$,

$K^d(x^d, X_i^d)$ is unrelated to $(x_s^d, X_{is}^d)$

$x_s^d$ is smoothed out.

For the continuous variables $x^c = (x_1^c, \ldots, x_p^c)$, the product kernel is
\[
K^c(x^c, X_i^c) = \prod_{s=1}^{p} \frac{1}{h_s} K \left( \frac{x_s^c - X_{is}^c}{h_s} \right),
\]
where $K$ is a symmetric, univariate density function, $0 < h_s < \infty$.

Note that when $h_s = \infty$, $x_s^c$ is smoothed out.

\[
K(x, X_i) = K^c(x^c, X_i^c) K^d(x^d, X_i^d). 
\]

We estimate $E(Y|X = x)$ by
\[
\hat{g}(x) = \frac{n^{-1} \sum_{i=1}^{n} Y_i K(x, X_i)}{n^{-1} \sum_{i=1}^{n} K(x, X_i)}. \tag{2.5}
\]

We choose $(h, \lambda) = (h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q)$ by minimizing the cross-validation function given by
\[
CV(h, \lambda) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{g}_{-i}(X_i))^2 w(X_i), \tag{2.6}
\]
where \( \hat{g}_{-i}(X_i) = \frac{\sum_{j \neq i} Y_j K(X_i, X_j)}{\sum_{j \neq i} K(X_i, X_j)} \) is the leave-one-out kernel estimator of \( E(Y_i|X_i) \), and \( w(.) \) is a weight function.

Define \( \sigma^2(\bar{x}) = E(u_i^2|\bar{X}_i = \bar{x}) \) and let \( S^c \) denote the support of \( w \). We also assume that

\[
g, f \text{ and } \sigma^2 \text{ have two continuous derivatives; } w \text{ is continuous, nonnegative and has compact support; } f \text{ is bounded away from zero for } x = (x^c, x^d) \in S. \quad (2.7)
\]

We impose the following conditions on the bandwidth and kernel functions. Define

\[
H = (\prod_{s=1}^{p_1} h_s) \prod_{s=p_1+1}^{p} \min(h_s, 1). \quad (2.8)
\]

Letting \( 0 < \epsilon < 1/(p + 4) \) and for some constant \( c > 0 \),

\[
n^\epsilon - 1 \leq H \leq n^{-\epsilon}; \quad n^{-c} < h_s < n^c \text{ for all } s = 1, \ldots, p;
\]

the kernel \( K \) is a symmetric, compactly supported, Hölder-continuous probability density; and \( K(0) > K(\delta) \) for all \( \delta > 0 \). \( (2.9) \)

The above conditions basically ask that each \( h_s \) does not converge to zero, or to infinity, too fast, and that \( nh_1 \ldots h_{p_1} \to \infty \).

We expect that, as \( n \to \infty \), the smoothing parameters associated with the relevant regressors will converge to zero, while those associated with the irrelevant regressors will not. It would be convenient to further assume that \( h_s \to 0 \) for \( s = 1, \ldots, p_1 \), and that \( \lambda_s \to 0 \) for \( s = 1, \ldots, q_1 \), however, the relevant components are known a priori.
Therefore, we shall assume the following condition holds. Defining
\[ \bar{\mu}_g(x) = E[\hat{g}(x)\hat{f}(x)]/E[\hat{f}(x)] \]
we assume that
\[ \int_{\text{supp } w} [\bar{\mu}_g(x) - \bar{g}(\bar{x})]^2 \bar{w}(\bar{x})\bar{f}(\bar{x})d\bar{x}, \]
as a function of \( h_1, \ldots, h_{p_1}, \) and \( \lambda_1, \ldots, \lambda_{q_1} \) vanishes
if and only if all of the smoothing parameters vanish, (2.10)
where \( \bar{w} \) is a weight function defined in (2.16) below.

Define an indicator function
\[ I_s(v^d, x^d) = I(v^d_s \neq x^d_s) \prod_{t \neq s, t=1}^q I(v^d_t = x^d_t). \] (2.11)
Note that \( I_s(v^d, x^d) = 1 \) if and only if \( v^d \) and \( x^d \) differ in their \( s \)th component only.

The leading term of \( CV \) is
\[ \int \frac{k^{p_1} \sigma^2(\bar{x})}{n h_1 \ldots h_{p_1}} w(x)\tilde{R}(x)\tilde{f}(\bar{x})dx \] (2.12)
\[ + \int \left( \sum_{s=1}^{q_1} \lambda_s \sum_{\bar{v}^d} [I_s(\bar{v}^d, \bar{x}^d) \{ \bar{g}(\bar{x}^e, \bar{v}^d) - \bar{g}(\bar{x}) \} \bar{f}(\bar{x}^e, \bar{v}^d)] \right) \]
\[ + \frac{1}{2} \kappa_2 \sum_{s=1}^{p_1} h_s^2 \left\{ \bar{g}_{ss}(\bar{x})\tilde{f}(\bar{x}) + 2\tilde{f}_s(\bar{x})\bar{g}_s(\bar{x}) \right\} \tilde{f}(\bar{x})^{-1}\bar{w}(\bar{x})d\bar{x}, \] (2.13)

\(^1\)Note that \( \bar{\mu}_g(\bar{x}) \) does not depend on \( \bar{x} \), nor does it depend on \( (h_{p_1+1}, \ldots, p, \lambda_{q_1+1}, \ldots, \lambda_q) \), because the components in the numerator related to the irrelevant variables cancel with those in the denominator.
with $\kappa = \int K(v)^2 dv$, $\kappa_2 = \int K(v)v^2 dv$, and where $\tilde{R}(x) = \tilde{R}(x, h_{p_1+1}, \ldots, h_p, \lambda_{q_1+1}, \ldots, \lambda_q)$ is given by

$$\tilde{R}(x) = \frac{\nu_2(x)}{\nu_1(x)},$$

(2.14)

where for $j = 1, 2$,

$$\nu_j(x) = E \left( \prod_{s=p_1+1}^p h_s^{-1} K \left( \frac{X_{is}^c - x_s^c}{h_s} \right) \prod_{s=q_1+1}^q \lambda_s I(x_{is}^d \neq x_s^d) \right)^j,$$

(2.15)

and

$$\bar{w}(\bar{x}) = \int \tilde{f}(\bar{x}) w(x) d\bar{x},$$

(2.16)

where, again, we define $\int d\bar{x} = \sum_{\bar{x}^d} \int d\bar{x}^c$.

Irrelevant variable $\bar{x}$ appears in $\tilde{R}$. By Hölder’s inequality, $\tilde{R}(x) \geq 1$ for all choices of $x$, $h_{p_1+1}, \ldots, h_p$, and $\lambda_{q_1+1}, \ldots, \lambda_q$. Also, $\tilde{R} \to 1$ as $h_s \to \infty$ ($p_1 + 1 \leq s \leq p$) and $\lambda_s \to 1$ ($q_1 + 1 \leq s \leq q$).

One needs to select $h_s$ ($s = p_1 + 1, \ldots, p$) and $\lambda_s$ ($s = q_1 + 1, \ldots, q$) to minimize $\tilde{R}$.

The only smoothing parameter values for which $\tilde{R}(x, h_{p_1+1}, \ldots, h_p, \lambda_{q_1+1}, \ldots, \lambda_q) = 1$ are $h_s = \infty$ for $p_1 + 1 \leq s \leq p$, and $\lambda_s = 1$ for $q_1 + 1 \leq s \leq q$.

Define $Z_n(x) = \prod_{s=p_1+1}^p K \left( \frac{x_s-X_{is}^d}{h_s} \right) \prod_{s=q_1+1}^q \lambda_s I(x_s^d \neq X_{is}^d)$. If at least one $h_s$ is finite (for $p_1 + 1 \leq s \leq p$), or one $\lambda_s < 1$ (for $q_1 + 1 \leq s \leq q$), then by $K(0) > K(\delta)$ for all $\delta > 0$, we know that $\text{Var}(Z_n) = E[Z_n^2] - [E(Z_n)]^2 > 0$ so that $R = E[Z_n^2]/[E(Z_n)]^2 > 1$. 

6
Only when all \( h_s = \infty \) and all \( \lambda_s = 1 \), we have \( Z_n \equiv K(0)^{p-p_1} \) and \( \text{Var}(Z_n) = 0 \). \( \tilde{R} = 1 \) only in this case.

Smoothing parameters related to irrelevant regressors must all converge to their upper extremities, so that \( \tilde{R}(x) \to 1 \) as \( n \to \infty \) for all \( x \in \mathcal{S} \).

Replacing \( \tilde{R}(x) \) by 1, the first term of (2.12) becomes

\[
\int \frac{\kappa^p \sigma^2(\bar{x})}{nh_1 \ldots h_{p_1}} \bar{w}(x)d\bar{x},
\]

(2.17)

\( a_s = h_sn^{1/(q_1+4)} \) and \( b_s = \lambda_sn^{2/(q_1+4)} \), then (2.12) (with (2.17) as its first term since \( \tilde{R}(x) \to 1 \)) becomes \( n^{-4/(q_1+4)} \tilde{\chi}(a_1, \ldots, a_{p_1}, b_1, \ldots, b_{q_1}) \), where

\[
\tilde{\chi}(a, b) = \int \frac{\kappa^p \sigma^2(\bar{x})}{a_1 \ldots a_{p_1}} \bar{w}(\bar{x})d\bar{x} + \int \left( \sum_{s=1}^{q_1} b_s \sum_{\vartheta^d} I_s(\bar{\vartheta}^d, \bar{x}^d) \left( \bar{g}(\bar{x}^c, \bar{\vartheta}^d) - \bar{g}(\bar{x}) \right) \bar{f}(\bar{x}^c, \bar{\vartheta}^d) \right)^2 \bar{f}(\bar{x})^{-1} \bar{w}(\bar{x})d\bar{x}.
\]

(2.18)

Let \( a_0^0, \ldots, a_{p_1}^0, b_1^0, \ldots, b_{q_1}^0 \) denote values of \( a_1, \ldots, a_{p_1}, b_1, \ldots, b_{q_1} \) that minimize \( \tilde{\chi} \) subject to each of them being nonnegative. We require that

Each \( a_s^0 \) is positive and each \( b_s^0 \) nonnegative, all are finite and uniquely defined.

(2.19)

**Theorem 2.1** Assume conditions (2.7), (2.9), (2.10), and (2.19) hold, and let \( \hat{h}_1, \ldots, \hat{h}_p, \hat{\lambda}_1, \ldots, \hat{\lambda}_q \) denote the smoothing parameters that minimize CV.
Then

\[ n^{1/(p_1+4)} \hat{h}_s \to a^0_s \quad \text{in probability for } 1 \leq s \leq p_1, \]

\[ P(\hat{h}_s > C) \to 1 \quad \text{for } p_1 + 1 \leq s \leq p \quad \text{and for all } C > 0, \]

\[ n^{2/(p_1+4)} \hat{\lambda}_s \to b^0_s \quad \text{in probability for } 1 \leq s \leq q_1, \]

\[ \hat{\lambda}_s \to 1 \quad \text{in probability for } q_1 + 1 \leq s \leq q, \]

and \( n^{4/(p_1+4)} CV(\hat{h}_1, \ldots, \hat{h}_{p}, \hat{\lambda}_1, \ldots, \hat{\lambda}_q) \to \inf \bar{\chi} \) in probability.

The asymptotic normality result for \( \hat{g}(x) \).

**Theorem 2.2** Under the same conditions as in Theorem 2.1, for \( x = (x^c, x^d) \in S = S^c \times S^d \), then

\[
\left( n^{1/2}(\hat{h}_1 \ldots \hat{h}_{p_1}) \right)^{1/2} \left[ \hat{g}(x) - \tilde{g}(\bar{x}) - \sum_{s=1}^{q_1} B_{1s}(\bar{x}) \hat{\lambda}_s^2 - \sum_{s=1}^{p_1} B_{2s}(\bar{x}) \hat{h}_s^2 \right] \xrightarrow{d} N(0, \Omega(\bar{x})),
\]

(2.21)

where \( B_{1s}(\bar{x}) = \sum_{\bar{v}^d} \bar{I}_s(\bar{v}^d, \bar{x}) \left( \tilde{g}(\bar{x}^c, \bar{v}^d) - \tilde{g}(\bar{x}) \right) \bar{f}(\bar{x}^c, \bar{v}^d) \bar{f}(\bar{x})^{-1} \), \( B_{2s}(\bar{x}) = \frac{\kappa_2}{2} \left( \tilde{g}_{ss}(\bar{x}) + 2 \frac{\tilde{f}(\bar{x}) \tilde{g}_s(\bar{x})}{\tilde{f}(\bar{x})} \right) \), and \( \Omega(\bar{x}) = \kappa^{p_1} \sigma^2(\bar{x}) / \bar{f}(\bar{x}) \) are the asymptotic bias and variance terms respectively.
The Presence of Ordered Discrete Regressors

If $x^d_s$ takes $c_s$ different values, $\{0, 1, \ldots, c_s - 1\}$, and is ordered,

We suggest the use

$$K^d(x^d_s, v^d_s) = \lambda^{[x^d_s - v^d_s]}.$$  \hspace{1cm} (2.22)

$\lambda_s \in [0, 1]$.

Aitchison and Aitken (1976,p.29) suggested using

$$K^d(x^d_s, v^d_s, \lambda_s) = \binom{c_s}{t} \lambda_s^t (1 - \lambda_s)^{c_s - t}$$

where $|x^d_s - v^d_s| = t \ (0 \leq t \leq c_s)$, and $\binom{c_s}{t} = c_s!/[t!(c_s - t)!]$.

It can not smooth out irrelevant regressors,
Cross-Validation and Estimation of Conditional Densities

Peter Hall, Jeff Racine, Qi Li

Let $f(x, y)$ denote the joint pdf of $(X, Y)$, and $m(x)$ the marginal pdf $X$, and $g(y|x) = f(x, y)/m(x)$.

We estimate $g(y|x) = f(x, y)/m(x)$ by

$$
\hat{g}(y|x) = \frac{\hat{f}(x, y)}{\hat{m}(x)}.
$$

(2.23)

Using the notation $\int dz = \sum_z d z^c$, a weighted integrated square difference between $\hat{g}(\cdot)$ and $g(\cdot)$ is

$$
I_n = \int [\hat{g}(y|x) - g(y|x)]^2 m(x) \, dz
= \int [\hat{g}(y|x)]^2 m(x) \, dz - 2 \int \hat{g}(y|x) g(y|x) m(x) \, dz + \int [g(y|x)]^2 m(x) \, dz
\equiv I_{1n} - 2I_{2n} + I_{3n}.
$$

(2.24)

$I_{3n}$ is independent of $(h, \lambda)$.

Define

$$
\hat{G}(x) = \int [\hat{f}(x, y)]^2 dy = n^{-2} \sum_i \sum_j K_{X_i, x} K_{X_j, x} \int K_{Y_i, y} K_{Y_j, y} dy
= n^{-2} \sum_i \sum_j K_{X_i, x} K_{X_j, x} K_{Y_i, Y_j}^{(2)},
$$

(2.25)

where $K_{Y_i, Y_j}^{(2)} = \int K_{Y_i, y} K_{Y_j, y^c} dy \equiv \sum_{y^c} \int K_{Y_i, y} K_{Y_j, y^c} dy^c$ is the second order convolution kernel, $K_{Y_i, y} = W_{Y_i, y} L_{Y_i, y}$.
Using Eq. (2.25), we have
\[
I_{1n} = \int \int [\hat{g}(y|x)]^2 m(x) \, dz = \int \frac{\int [\hat{f}(x,y)]^2 dy}{\hat{m}(x)^2} m(x) \, dx
\]
\[
= \int \frac{\hat{G}(x)}{\hat{m}(x)^2} m(x) \, dx = E_X \left[ \frac{\hat{G}(X)}{\hat{m}(X)^2} \right],
\]
where $E_X(\cdot)$ denotes the expectation with respect to $X$ only.

Also,
\[
I_{2n} = \int \hat{g}(y|x) g(y|x) m(x) \, dz = \int \hat{g}(y|x) f(x,y) \, dx \, dy
\]
\[
= \int \left[ \frac{\hat{f}(x,y)}{\hat{m}(x)} \right] f(x,y) \, dx = E_Z \left[ \frac{\hat{f}(Z)}{\hat{m}(X)} \right],
\]
where $E_Z$ denotes the expectation with respect to $Z$ only.

Choosing $m(x)$ as the weighting function gives $I_{1n}$ and $I_{2n}$ some simple forms.

\[
I_{1n} - 2I_{2n} = E_X \left\{ \frac{\hat{G}(X)}{\hat{m}(X)^2} \right\} - 2E_Z \left[ \frac{\hat{f}(X,Y)}{\hat{m}(X)} \right],
\]
(2.28)

Use the leave-one-out estimators for $\hat{f}(X, Y_l)$ and $\hat{m}(X_l)$, also use a leave-one-out estimator for $G(X_l)$:

\[
\hat{G}_{-l}(X_l) = n^{-2} \sum_{i \neq l} \sum_{j \neq l} K_{X_i, X_l} K_{X_j, X_l} K_{Y_i, Y_j}^{(2)}.
\]
(2.29)

\[
CV(h, \lambda) \overset{\text{def}}{=} \frac{1}{n} \sum_{l=1}^{n} \frac{\hat{G}_{-l}(X_l)}{\hat{m}_{-l}(X_l)^2} - \frac{2}{n} \sum_{l=1}^{n} \frac{\hat{f}_{-l}(X_l, Y_l)}{\hat{m}_{-l}(X_l)}.
\]
(2.30)

We will choose $(\lambda, h)$ to minimize $CV(h, \lambda)$. 

11
Theorem 2.3 Assume some regularity conditions, and let \( \hat{h}_1, \ldots, \hat{h}_p, \hat{\lambda}_1, \ldots, \hat{\lambda}_q \) denote the smoothing parameters that minimize CV. Then

\[
\begin{align*}
n^{1/(p_1+5)} \hat{h}_s & \to a_s^0 \quad \text{in probability for } 0 \leq s \leq p_1, \\
P(\hat{h}_s > C) & \to 1 \quad \text{for } p_1 + 1 \leq s \leq p \text{ and for all } C > 0, \\
n^{2/(p_1+5)} \hat{\lambda}_s & \to b_s^0 \quad \text{in probability for } 1 \leq s \leq q_1, \\
\hat{\lambda}_s & \to 1 \quad \text{in probability for } q_1 + 1 \leq s \leq q,
\end{align*}
\]

(2.31)

and \( n^{4/(p_1+5)} \text{CV}(\hat{h}_0, \hat{h}_1, \ldots, \hat{h}_p, \hat{\lambda}_1, \ldots, \hat{\lambda}_q) \to \inf \chi_p(.) \) in probability.

The asymptotic normality result.

Theorem 2.4 Under some regularity conditions, we have

\[
\begin{align*}
\sqrt{n}h^p(\hat{g}(y|x) - g(y|x) - \hat{h}^2B_1(z) - \hat{\lambda}B_2(z)) & \to N(0, \Omega(z)) \quad \text{in distribution},
\end{align*}
\]

where \( B_1(z) = (1/2)(1/m(z))tr[\nabla^2 f(z)][\int w(v)v^2dv], \)

\[
B_2(z) = (1/m(z))\sum_{z^c, z^d} [f(z^c, z^d) - f(z^c, z^d)], \quad \text{and} \quad \Omega(z) = [f(z)/m^2(x)][\int W^2(v)dv] (\nabla^2 \text{ is with respect to } z^c).
\]
In the conditional probability paper, we define the kernel function for discrete variables via for \(1 \leq s \leq q\),

\[
l(X_{is}^d, x_s^d, \lambda_s) = \begin{cases} 
1 - \lambda_s & \text{if } X_{is}^d = x_s^d, \\
\frac{\lambda_s}{c_s - 1} & \text{if } X_{is}^d \neq x_s^d.
\end{cases}
\] (2.32)

\(\lambda_s \in [0, (c_s - 1)/c_s]\).

If \(x_s^d \in \{0, 1\}\), a dummy variable with \(c_s = 2\). Then

\[
l(X_{is}^d, x_s^d, \lambda_s) = \begin{cases} 
1 - \lambda_2 & \text{if } X_{is}^d = x_s^d, \\
\lambda_s & \text{if } X_{is}^d \neq x_s^d.
\end{cases}
\] (2.33)

\(\lambda_s \in [0, 1/2]\).

The reason for this is that the kernel weights add up to one.

\[
\sum_{x_s^d \in S^d} L(x_s^d, X_i^d, \lambda) = 1.
\]

It is needed in unconditional density estimation.

This is not needed in regression estimation, or the conditional covariates.

The relationship between the two \(\lambda\)'s are

Relative weight is \(\lambda/1 = \lambda\) earlier.

Now the relative weight is \(\lambda/[(c - 1)(1 - \lambda)]\).

For example, if \(c = 2\) and \(\lambda = 0.2\) in the second expression, then the equivalent \(\lambda\) in the first expression should be

\[
\lambda = 0.2/[(1)(1 - 0.2)] = 0.2/0.8 = 0.25.
\]
3 Applications

Two empirical applications.

Out-of-sample predictive performance.

3.1 Count Survival Data - Veteran’s Administration Lung Cancer Trial

The data set is from Kalbfleisch and Prentice (1980, pg 223-224), which models survival in days of cancer patients with six discrete explanatory variables being treatment type, cell type, Karnofsky score, months from diagnosis, age in years, and prior therapy.

The dataset contains 137 observations, and the number of cells greatly exceed the number of observations.

Model the expected survival in days. The estimation sample \( n_1 = 132 \), and the prediction sample is \( n_2 = 5 \).

The out-of-sample \( MSE = n_2^{-1} \sum_{i=1}^{n_2} (y_i - \hat{y}_i)^2 \).

Randomly split the sample into two sub-samples of size \( n_1 \) and \( n_2 \) 100 times.

<table>
<thead>
<tr>
<th>Model</th>
<th>Median MSE</th>
<th>Mean MSE</th>
<th>Mean MSE / Mean MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NONP</td>
<td>9,770.5</td>
<td>17,614.9</td>
<td>100%</td>
</tr>
<tr>
<td>OLS</td>
<td>13,831.7</td>
<td>28,422.6</td>
<td>62%</td>
</tr>
<tr>
<td>POISSON</td>
<td>15,319.8</td>
<td>29,052.9</td>
<td>60%</td>
</tr>
</tbody>
</table>
Median values of the CV smoothing parameters over the 100 sample splits.

Table 2: Summary of Median Smoothing Parameters [number of categories parentheses]

<table>
<thead>
<tr>
<th></th>
<th>( \lambda_1 ) [2]</th>
<th>( \lambda_2 ) [4]</th>
<th>( \lambda_3 ) [12]</th>
<th>( \lambda_4 ) [28]</th>
<th>( \lambda_5 ) [40]</th>
<th>( \lambda_6 ) [2]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.934</td>
<td>0.056</td>
<td>0.004</td>
<td>0.157</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

The ‘relevant’ explanatory variables for predicting survival are cancer cell-type (\( \lambda_2 \)), Karnofsky score\(^2\) (\( \lambda_3 \)), and months from diagnosis (\( \lambda_4 \)). The remaining regressors are smoothed out by cross-validation.

\(^2\)The Karnofsky score measures patient performance of activities of daily living. The score has proven useful not only to follow the course of the illness (usually progressive deficit and ultimately death), but also a prognosticator: patients with the highest (best) Karnofsky scores at the time of tumor diagnosis have the best survival and quality of life over the course of their illness.
3.2 Count Data - Fair’s 1977 Extramarital Affairs Data

We consider Fair’s (1978) dataset which models how often an individual engaged in extramarital affairs during the past year. This often cited paper with a complete description of the data is located at http://fairmodel.econ.yale.edu/rayfair/pdf/1978A200.PDF. The dataset contains 601 observations and has eight discrete explanatory variables given by sex, age, number of years married, have children or not, how religious, level of education, occupation, and how they rate their marriage.

Estimation sample \( n_1 = 500 \), prediction sample \( n_2 = 101 \). The goal is to model the expected number of affairs.

We split the sample into two sub-sample of size \( n_1 \) and \( n_2 \) to compute the out-of-sample squared prediction errors, and repeat this 100 times.

Median and mean out-of-sample squared prediction errors by different estimation methods.

<table>
<thead>
<tr>
<th>Model</th>
<th>Median MSE</th>
<th>Mean MSE</th>
<th>( \text{MSPE} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>NONP</td>
<td>9.97</td>
<td>10.18</td>
<td>100%</td>
</tr>
<tr>
<td>OLS</td>
<td>11.90</td>
<td>12.39</td>
<td>82%</td>
</tr>
<tr>
<td>PROBIT</td>
<td>14.66</td>
<td>14.74</td>
<td>69%</td>
</tr>
<tr>
<td>TOBIT</td>
<td>12.81</td>
<td>13.04</td>
<td>78%</td>
</tr>
<tr>
<td>POISSON</td>
<td>12.88</td>
<td>12.99</td>
<td>78%</td>
</tr>
</tbody>
</table>

NP MSPE is 69% to 82% of those by various Para. methods.
Median values of CV smoothing parameters over the 100 sample splits.

Table 4: Summary of Median Smoothing Parameters [number of categories in parentheses]

<table>
<thead>
<tr>
<th>$\lambda_1$</th>
<th>2</th>
<th>$\lambda_2$</th>
<th>9</th>
<th>$\lambda_3$</th>
<th>8</th>
<th>$\lambda_4$</th>
<th>2</th>
<th>$\lambda_5$</th>
<th>5</th>
<th>$\lambda_6$</th>
<th>7</th>
<th>$\lambda_7$</th>
<th>7</th>
<th>$\lambda_8$</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.416</td>
<td>0.064</td>
<td>1.000</td>
<td>0.998</td>
<td>0.507</td>
<td>0.999</td>
<td>0.282</td>
<td>0.004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The most ‘relevant’ explanatory variables are age ($\lambda_2$) and how a person rates their marriage ($\lambda_8$), while ‘irrelevant’ variables are number of years married ($\lambda_3$), having children or not ($\lambda_4$), and level of education ($\lambda_6$) appear to be ‘irrelevant’ for prediction of number of affairs, and the remaining variables appear to have some relevance again from a cross-validation perspective.

In comparison, using a Tobit model, Fair obtained a t-statistic of 3.63 for the number of years married, while our nonparametric method removes this variable from the resulting estimate. The out-of-sample prediction results suggest that the Tobit model is likely to be misspecified.