Henderson, Daniel J.; Souto, Anne-Charlotte

Working Paper

An Introduction to Nonparametric Regression for Labor Economists

IZA Discussion Papers, No. 11914

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Henderson, Daniel J.; Souto, Anne-Charlotte (2018) : An Introduction to Nonparametric Regression for Labor Economists, IZA Discussion Papers, No. 11914, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/193208

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11914

An Introduction to Nonparametric Regression for Labor Economists

Daniel J. Henderson
Anne-Charlotte Souto

OCTOBER 2018

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11914

An Introduction to Nonparametric Regression for Labor Economists

Daniel J. Henderson
University of Alabama and IZA

Anne-Charlotte Souto
University of Alabama

OCTOBER 2018

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

An Introduction to Nonparametric Regression for Labor Economists

In this article we overview nonparametric (spline and kernel) regression methods and illustrate how they may be used in labor economic applications. We focus our attention on issues commonly found in the labor literature such as how to account for endogeneity via instrumental variables in a nonparametric setting. We showcase these methods via data from the Current Population Survey.

JEL Classification: C14, C26, I24, J24, J31

Keywords: endogeneity, kernel, labor, nonparametic, regression, spline

Corresponding author:
Daniel J. Henderson
Department of Economics
Finance and Legal Studies
University of Alabama
Box 870224
Tuscaloosa, AL 35487-0224
United States

E-mail: djhender@cba.ua.edu

1 Introduction

This survey aims to (re-)introduce applied labor economists to nonparametric regression techniques. Specifically, we discuss both spline and kernel regression, in an approachable manner. We present an intuitive discussion of estimation and model selection for said methods. We also address the use of nonparametric methods in the presence of endogeneity, a common issue in the labor literature, but seldom accounted for in applied nonparametric work.

Accounting for endogeneity is well understood in the parametric literature once a suitable instrument is obtained. Standard methods have been around for some time, but these methods do not always transfer in a straightforward manner in the nonparametric setting. This has caused many to shy away from their use, even with the knowledge that this can lead to additional insight (Henderson and Parmeter [2015]).

To showcase these methods, we will look at the relationship between experience, education and earnings. We will begin by ignoring the endogeneity of education and then will discuss how to control for this via a nonparametric control function approach. While nonparametric estimation may seem like a choice, it should be stated that the parametric alternative requires strict functional form assumptions, which if false, likely lead to biased and inconsistent estimators. In practice, the functional relationship between education and earnings as well as between education and its instruments is typically unknown. By using nonparametric regression, we relax these functional form restrictions and are more likely to uncover the causal relationship.

To empirically illustrate these methods, we use individual-level data obtained from the March Current Population Survey (CPS) to highlight each concept discussed. To eliminate additional complications, we primarily focus on a relatively homogeneous sub-group, specifically, working age (20 to 59 years old) males with four-year college degrees.

In what follows, we first slowly introduce the fundamentals of spline and kernel estimators and then discuss how to decide upon various options of each estimator. This should build the foundation for understanding the more advanced topic of handling endogenous regressors. By illustrating these techniques in the context of labor-specific examples, we hope that this helps lead to widespread use of these methods in labor applications.

2 Nonparametric Regression

In a parametric regression model, we assume which functional form best describes the relationship between the response and explanatory variables. If this form is correct, and the remaining Gauss-Markov assumptions hold, we will have unbiased and efficient estimators. However, if these assumptions do not hold, these estimators are likely biased and inconsistent. Nonlinear parametric models exist, but are often complicated to estimate and still require *a priori*

knowledge of the underlying functional form.

Nonparametric regression offers an alternative. The methods discussed here estimate the unknown conditional mean by using a "local" approach. Specifically, the estimators use the data near the point of interest to estimate the function at that point and then use these local estimates to construct the global function. This can be a major advantage over parametric estimators which use all data points to build their estimates (global estimators). In other words, nonparametric estimators can focus on local peculiarities inherent in a data set. Those observations which are more similar to the point of interest carry more weight in the estimation procedure.

This section will introduce two commonly used nonparametric techniques, and will provide the notation and concepts that will be used for the remainder of this review. Specifically, we discuss spline and kernel regression estimation. To help bridge gaps, we make connections to well-known techniques such as ordinary and weighted least-squares.

2.1 Spline Regression

Spline regression can be thought of as an extension of ordinary least-squares (OLS). Consider the basic univariate linear model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ i = 1, 2, \ldots, n, \tag{1}$$

where for a sample of n observations, y is our response variable, x is our explanatory variable, ϵ is our usual error term and we have two parameters: a constant and a slope (α and β, respectively). The right-hand side of (1) can be thought of as a linear combination of 1 and x, we call them the "bases" of the model. One popular way to transform (1) into a nonlinear function is to add higher-order polynomials. A quadratic model would add one extra basis function x^2 to the model, which corresponds to adding the term $\beta_2 x_i^2$ to (1). In matrix form, the number of bases would correspond to the number of columns in the matrix X:

$$y = X\beta + \epsilon, \tag{2}$$

where

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

for the linear case (2 bases), and

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}$$

for the quadratic case (3 bases).

These two cases are illustrated in Figure 1 where x is years of experience and y is the log wage (adjusted for inflation). To highlight a relatively homogeneous group, we restrict our sample to college-educated (16 years of schooling) males working in personal care and service (or related occupations) between 2006 and 2016.[1] For each panel, the solid line represents white males and the dashed line non-whites. Our linear model (i.e., OLS) shows a strong wage gap between whites and non-whites which seems to remain constant (in percentage terms) as the sale workers gain experience (i.e., similar slopes). Adding experience squared to the model (quadratic model) allows us to better capture the well-known nonlinear relationship between log wage and experience. As workers gain experience, we expect their log wage to increase, but at a decreasing rate. The quadratic model (bottom-left panel) shows a large increase in log wages early in a career with a slight downfall towards the end. Also, this model tends to suggest that the wage gap between white and non-white males working in personal care and service varies with experience. Non-white workers appear to have a more constant and slower increase in their predicted log wages.

2.1.1 Linear Spline Bases

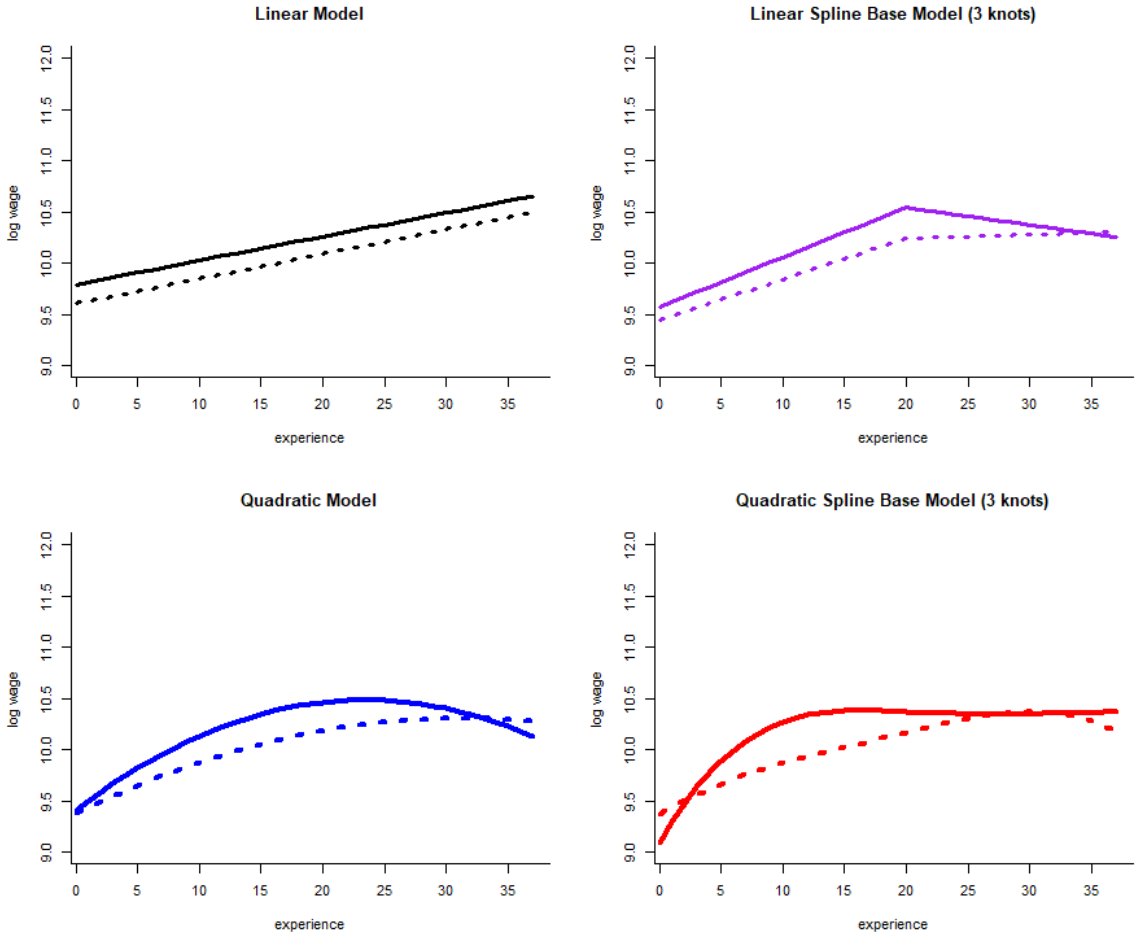
In our example, we could argue that although wages should increase with experience (increase in competence/knowledge), there may be a point where more experience will not increase wages or perhaps even decrease it (slower cognitive ability/decreases in efficiency). Suppose we created a model with the equivalent of two linear regression: one for the first 20 years of experience, and another for the latter years. This would be equivalent of adding the following basis function to our linear model:

$$(x - 20)_+,$$

where the $+$ sign indicates that the function is set to zero for all values of x where $(x - 20)$ is negative. This model is sometimes called the *broken stick* model because of its shape, but more generally is referred to as a linear spline base model with 3 knots. The 3 knots are at 0 (minimum value), 20, and 37 (maximum value) years of experience. Note that the maximum and minimum values of x will always be consider to be knots. For example, the linear model in equation (9) has two knots. Here we arbitrarily fixed the middle knot at 20 years of experience.

[1] Fixing the sample to college educated males allows us to plot these figures in two dimensions.

Figure 1: Log-wages versus Experience for White versus Non-white College-educated Males Working in Personal Care and Service



We will discuss which knots to select and how many to select in Section 3.

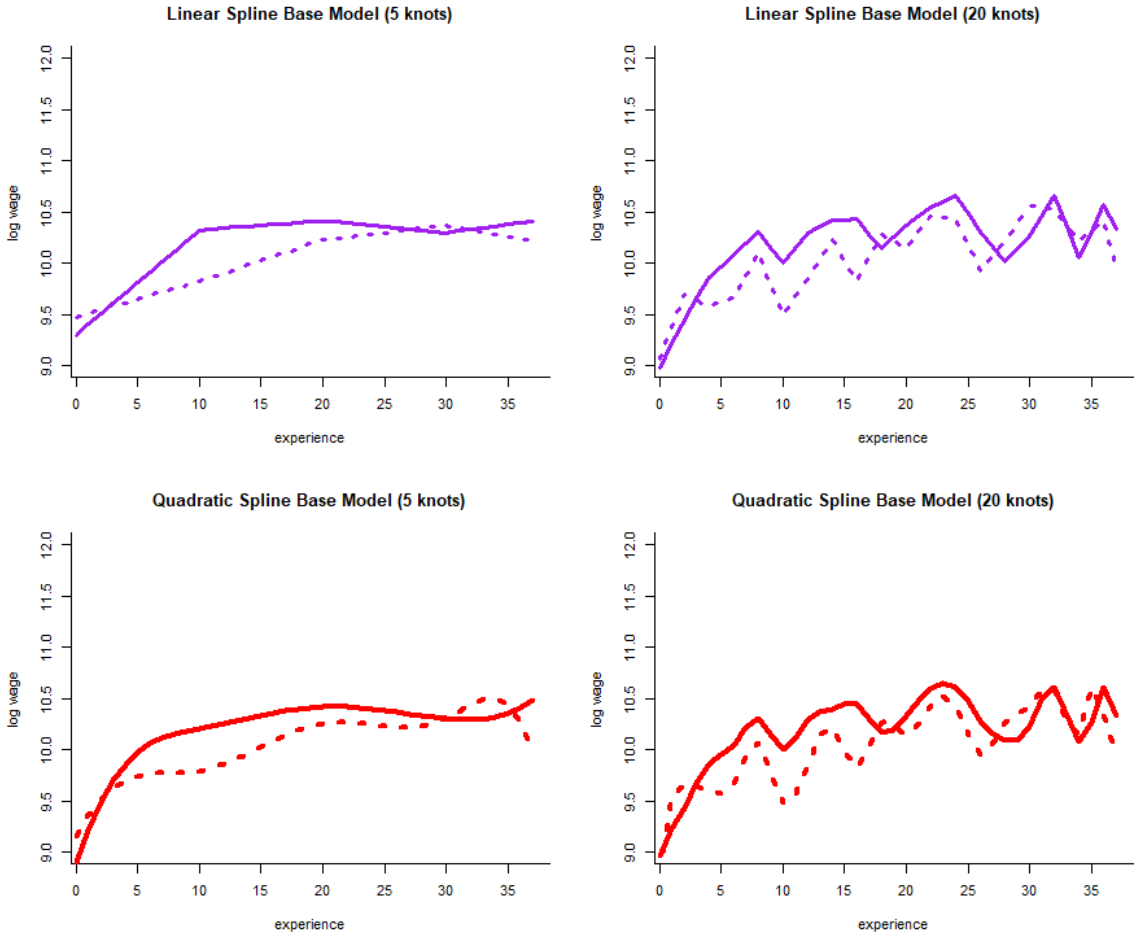
The *broken stick* model with a break at $x = 20$ is written as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i - 20)_+ + \epsilon_i \tag{3}$$

and is illustrated in upper-right panel of Figure 1. We see a similar result to the quadratic model, that is, for white workers, we see a strong increase in wages in the first part of their career followed by a smaller decrease towards the end of their career. That being said, we arbitrarily fixed the middle knot at 20 years of experience. Without strong reasons to do so, it is premature to say anything about when the increase in the log wage stops and when the decrease begins. Noting the aforementioned issue, we also observe the wage gap widen at first with experience, but then converge at higher levels of experience.

Figure 2 illustrates how adding knots at different values can change the results. We present a model with 5 knots at $x = 0, 10, 20, 30, 37$, and a model with 20 knots (every 2 years) at

Figure 2: Log-wage versus Experience for White versus Non-white College-educated Males Working in Personal Care and Service



$x = 0, 2, 4, ..., 34, 36, 37$. In matrix form, equation (2), the X matrix with 5 knots is given as

$$X = \begin{bmatrix} 1 & x_1 & (x_1-10)_+ & (x_1-20)_+ & (x_1-30)_+ & (x_1-37)_+ \\ 1 & x_2 & (x_2-10)_+ & (x_2-20)_+ & (x_1-30)_+ & (x_2-37)_+ \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & x_n & (x_n-10)_+ & (x_n-20)_+ & (x_1-30)_+ & (x_n-37)_+ \end{bmatrix}$$

and with 20 knots,

$$X = \begin{bmatrix} 1 & x_1 & (x_1-2)_+ & ... & (x_1-36)_+ & (x_1-37)_+ \\ 1 & x_2 & (x_2-2)_+ & ... & (x_1-36)_+ & (x_2-37)_+ \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & x_n & (x_n-2)_+ & ... & (x_1-36)_+ & (x_n-37)_+ \end{bmatrix}$$

Adding knots at 10 and 30 years of experience allows the model to account for the commonly seen mid-career flattening period. However, the function is still not very smooth and it is hard to tell from this model when log wages start to flatten out. Adding more knots allows for more flexibility, but this can potentially lead to overfitting. For example, in the linear base model with 20 knots (upper-right panel of Figure 2), the fitted line is appears to be modeling noise.

2.1.2 Quadratic Spline Bases

The linear spline base model is a combination of linear bases. The quadratic spline base model is a combination of quadratic bases. In other words, we simply add the corresponding squared function for each of the linear base functions. Consider our previous broken stick model with a middle knot at $x = 20$, we can transform it into a quadratic spline base model with a knot at $x = 20$ by replacing $(x - 20)_+$ with the following bases:

$$x^2,\ (x - 20)_+^2.$$

This quadratic spline base model is represented by the following equation

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 (x_i - 20)_+^2 + \epsilon_i, \tag{4}$$

and is illustrated in the bottom-right panel of Figure 1. We can see that the quadratic spline base model suggests a slightly different relationship between experience and log wage. The predicted log wage increases more dramatically for the first 5 years of work experience, but flattens out thereafter. The racial gap seems to be small at first, but widens greatly over the first 5 years. Non-white workers appear to slowly catch up over the course of their careers.

One of the main advantages of the quadratic over the linear spline base model is that it does not have any sharp corners (i.e., undefined gradients). It follows that for any number of knots, the resulting function will have continuous first derivatives. This is both a useful and aesthetically pleasing property. Adding more knots (lower-right panel of Figure 2) to the model adds more variability. It appears that for this example, 5 knots would be sufficient.

An important concept in economics (typically of secondary importance in statistics textbooks) is recovery of the gradients. In the linear case, the gradient is simply the estimated coefficient between two particular knots. In the quadratic (or higher-order) case, we use the same method to get the gradient as in a simple quadratic OLS model. The difference is that we calculate it between each knot. That is, to estimate a particular gradient for any type of spline model, we can simply take the partial derivative with respect to the regressor x. In its general

form, our estimated gradient $\widehat{\beta}(x)$ for a particular regressor x is

$$\widehat{\beta}(x) = \frac{\partial \widehat{y}(x)}{\partial x}. \tag{5}$$

For our linear spline base example with 3 knots, this is

$$\widehat{\beta}(x) = \beta_1 + \begin{cases} \beta_2, & \text{if } x \in [20, 37) \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

and for our quadratic spline base example with 3 knots

$$\widehat{\beta}(x) = \beta_1 + 2\beta_2 x_i + \begin{cases} 2\beta_3(x_i - 20), & \text{if } x \in [20, 37) \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

2.1.3 B-Splines

We introduced linear and quadratic spline models with the truncated power basis function. Using the same truncated power functions, those models can be generalized to

$$y = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{\mathbb{k}=1}^{\mathbb{K}} \beta_{p\mathbb{k}} (x - \kappa_{\mathbb{k}})_+^p + \epsilon_i, \tag{8}$$

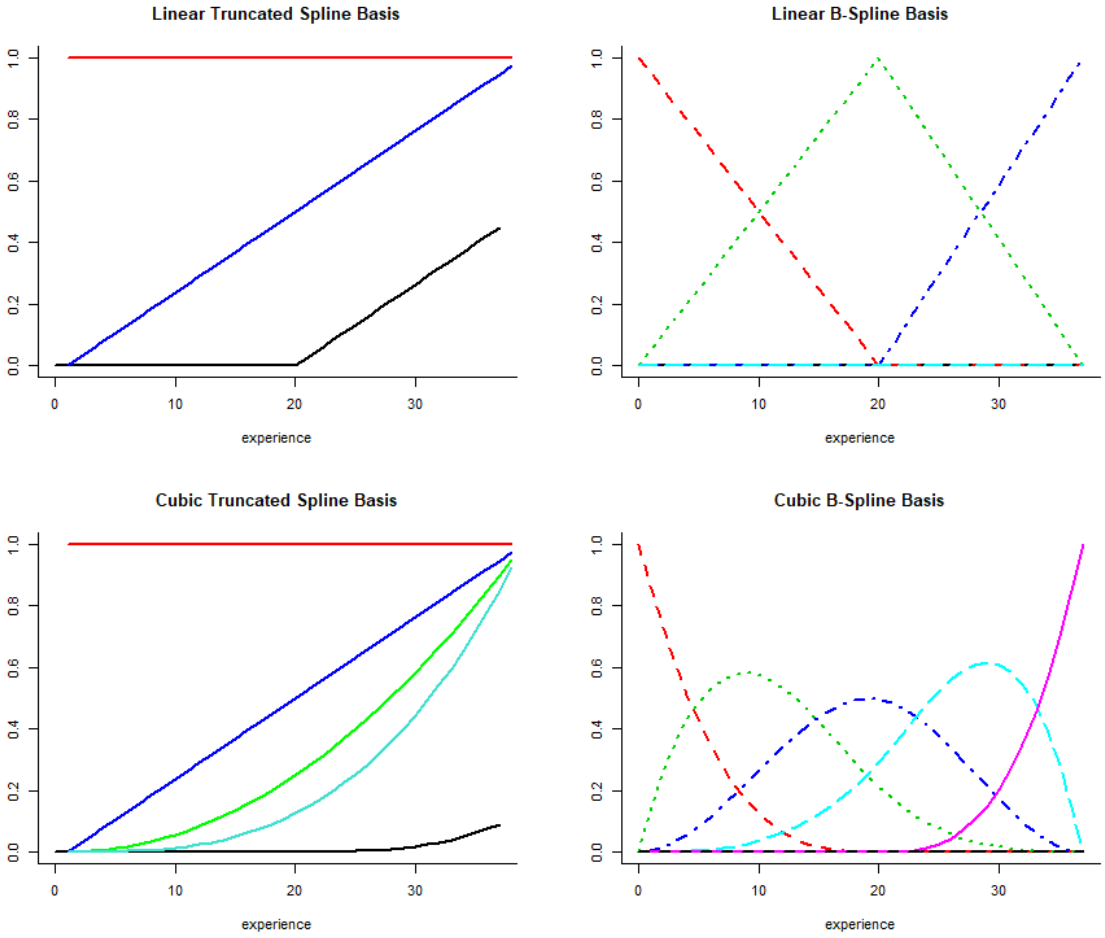
where p is the degree of the power basis (truncated power basis of degree p). This generalizes our model by allowing for (1) other spline models (using p degrees), and (2) other bases for a given spline model (using $\mathbb{K}$ knots). This function has $p - 1$ continuous derivatives and thus higher values of p should lead to "smoother" spline functions. Similar to before, the general form of the gradient is defined as

$$\widehat{\beta}(x) = \beta_1 + \cdots + p\beta_p x^{p-1} + \sum_{\mathbb{k}=1}^{\mathbb{K}} p\beta_{p\mathbb{k}} (x - \kappa_{\mathbb{k}})_+^{p-1}. \tag{9}$$

While this general form seems reasonable, splines computed from the truncated power bases in equation 8 may be numerically unstable. The values in the X-matrix may become very large (for large p), and the columns of the X-matrix may be highly correlated. This problem will only become worse with a higher number of knots. Therefore, (8) is rarely used in practice, but is instead typically transformed into equivalent bases with more stable numerical properties. One of the most popular is the *B-spline* basis.

This can be relatively difficult to present and code, but luckily there exist regression packages to easily transform the X-matrix into the more numerically stable version. Formally, we can

Figure 3: Truncated and B-Spline Corresponding Bases with Knots at 0, 20, and 37 Years of Experience



compute the equivalence as

$$X_b = X_t L_p,$$

where X_t is a matrix of the bases (explanatory variables) used in (8) and L_p is a squared invertible matrix. The most commonly used transformation in the linear case is

$$B(x)_j = \begin{cases} \frac{x-\kappa_j}{\kappa_{j+1}-\kappa_j}, & \text{if } x \in [\kappa_j, \kappa_{j+1}) \\ \frac{\kappa_{j+2}-x}{\kappa_{j+2}-\kappa_{j+1}}, & \text{if } x \in [\kappa_{j+1}, \kappa_{j+2}) \\ 0, & \text{otherwise} \end{cases}$$

for $j = -1, 2, 3, ..., \mathbb{K}$.

To better illustrate this, consider our *broken stick* example from Figure 1: the linear-spline with one middle knot at 20 years of experience. The corresponding bases for this model are 1, x and, $(x-20)_+$ and are shown in the upper-left panel of Figure 3. The B-spline transformation

of the second knot (20 years of experience) for this example is

$$B(x)_{j=2} = \begin{cases} \frac{x-0}{20-0}, & \text{if } x \in [0, 20) \\ \frac{37-x}{37-20}, & \text{if } x \in [20, 37) \\ 0, & \text{otherwise} \end{cases}$$

The corresponding bases of this transformation are shown in the upper-right panel of Figure 3. $B(x)_{j=2}$ corresponds to the inverse V-shaped function which equals 1 when experience equals 20. The other two functions can be computed similarly using $j = -1$, and 3. Adding a higher degree to our model will change the shape of our basis functions. The two bottom panels of Figure 3 show the equivalent truncated spline basis and B-spline basis for the cubic case ($p = 3$).

While other basis functions exist (for example, radial basis functions), practitioners may prefer B-splines as they are both numerically more stable and relatively easy to compute. Both `R` and Stata packages are available. We used the now defunct $bs(\cdot)$ function in the *splines* package[2] in `R`. The *bspline* module is available in Stata for B-splines.

2.2 Kernel Regression

Instead of assuming that the relationship between y and x come from a polynomial family, we can state that the conditional mean is an unspecified smooth function $m(\cdot)$ and our model will be given as

$$y_i = m(x_i) + \epsilon_i, \; i = 1, 2, \ldots, n, \tag{10}$$

where the remaining variables are described as before. In much the same way spline regression can be thought of as an extension of OLS, kernel regression can be seen as an extension of WLS. That is, we are still minimizing a weighted residual sum of squares, but now we will weight observations by how close they are to the point of interest (i.e., a "local" sample). With spline regression, our local sample is defined as all the points included between two knots, where each point within that sample is weighted equally. Kernel regression goes a step further by estimating each point using a weighted local sample that is centered around the point of interest. The local sample is weighted using a kernel function, which possess several useful properties.

A kernel function defines a weight for each observation within a (typically) symmetric predetermined bandwidth. Unlike an OLS regression which makes no distinction of where the data are located when estimating the conditional expectation, kernel regression will estimate the point of interest using data within a bandwidth.

[2]See https://stat.ethz.ch/R-manual/R-devel/library/splines/html/bs.html and the seemingly equivalent *bSpline*$(\cdot)$ function in the *splines2* package.

Before introducing the kernel estimators, let us first derive a kernel function. Consider x our point of interest; we can write an indicator function such that data fall within a range h (our bandwidth) around x:

$$n_x = \sum_{i=1}^{n} 1\left\{x - \frac{h}{2} \leq x_i \leq x + \frac{h}{2}\right\},$$

The corresponding probability of falling in this box (centered on x) is thus n_x/n. This indicator function can be rewritten as

$$n_x = \sum_{i=1}^{n} \left(\frac{1}{2}\right) 1\left\{\left|\frac{x_i - x}{h}\right| \leq 1\right\}. \tag{11}$$

This function is better known as a uniform kernel and is more commonly written as

$$k(\psi) = \begin{cases} 1/2, & \text{if } |\psi| \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

where we have written $k(\psi)$ for convenience, where ψ is defined as $(x_i - x)/h$ and represents how "local" the observation x_i is relative to x. Though very simple and intuitive, the uniform kernel is not smooth. It is discontinuous at -1 and 1 (when the weight switches from $1/2$ to zero) and has a derivative of 0 everywhere except at theses two points (where it is undefined).

Table 1: Commonly used Second-order Kernel Functions

Kernel	$k(\psi)$	$\kappa_2(k)$
Uniform ($s = 0$)	$\frac{1}{2}\mathbf{1}\{\lvert\psi\rvert \leq 1\}$	$1/3$
Epanechnikov ($s = 1$)	$\frac{3}{4}(1-\psi^2)\mathbf{1}\{\lvert\psi\rvert \leq 1\}$	$1/5$
Biweight ($s = 2$)	$\frac{15}{16}(1-\psi^2)^2\mathbf{1}\{\lvert\psi\rvert \leq 1\}$	$1/7$
Triweight ($s = 3$)	$\frac{35}{32}(1-\psi^2)^3\mathbf{1}\{\lvert\psi\rvert \leq 1\}$	$1/9$
Gaussian ($s = \infty$)	$\frac{1}{\sqrt{2\pi}}e^{-(1/2)\psi^2}$	1

This kernel is rarely used, but it does possess some basic properties that we typically require

of kernel functions. More formally, if we let the moments of the kernel be defined as

$$\kappa_j(k) = \int_{-\infty}^{\infty} \psi^j k(\psi) d\psi, \tag{12}$$

these properties are

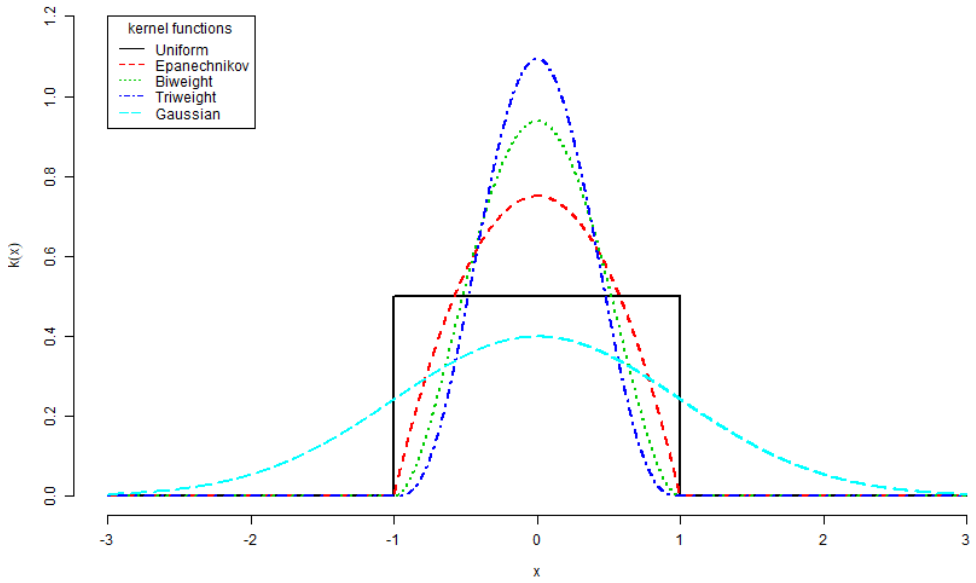
1. $\kappa_0(k) = 1$ ($k(\psi)$ integrates to one),
2. $\kappa_1(k) = 0$ ($k(\psi)$ is symmetric), and
3. $\kappa_2(k) < \infty$ ($k(\psi)$ has a finite variance).

These are known as second-order kernels. In addition to the uniform kernel, several commonly known kernel functions can be found in Table 1 (with their second-moments) and Figure 4. Each of them are derived from the general polynomial family:

$$k_s(\psi) = \frac{(2s+1)!!}{2^{s+1}s!}(1-\psi^2)^s \mathbf{1}\{|\psi| \leq 1\}, \tag{13}$$

where !! is the double factorial. The most commonly used kernel function in econometrics is the Gaussian kernel as it has derivatives of all orders. The most commonly used kernel function in statistics is the Epanechnikov kernel function as it has many desirable properties with respect to mean squared error. We will discuss how to choose the kernel function and smoothing parameter (h) in Section 3.

Figure 4: Commonly used Second-order Kernel Functions



2.2.1 Local-Constant Least-Squares

The classic kernel regression estimator is the local-constant least-squares (LCLS) estimator (also known as the Nadaraya-Watson kernel regression estimator, see Nadaraya [1964] and Watson [1964]). While it has fallen out of fashion recently, it is useful as a teaching tool and still useful in many situations (e.g., binary left-hand-side variables).

To begin, recall how we construct the OLS estimator. Our objective function is

$$\min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - x_i\beta)^2,$$

which leads to the slope and intercept estimators, $\widehat{\beta}$ and $\widehat{\alpha}$.

Suppose instead of a linear function of x, we simply regress y on a constant (a). Our objective function becomes

$$\min_{a} \sum_{i=1}^{n} [y_i - a]^2,$$

which leads to the estimator $\widehat{a} = (1/n)\sum_{i=1}^{n} y_i = \bar{y}$. A weighted least-squares version of this objective function can be written as

$$\min_{a} \sum_{i=1}^{n} [y_i - a]^2 W(x_i),$$

where $W(x_i)$ is the weighting function, unique to the point x_i. If we replace the weighting function with a kernel function, minimizing this objective function yields the LCLS estimator

$$\widehat{a} = \widehat{m}(x) = \frac{\sum_{i=1}^{n} y_i k\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)}. \tag{14}$$

This estimator represents a local average. Essentially, we regress y locally, on a constant, weighting observations via their distance to x.

While equation 14 gives us the fit, economists are typically interested in the marginal effects (i.e., gradients). To estimate a particular gradient, we simply take the partial derivative of $\widehat{m}(x)$ with respect to the regressor of interest, x. Our estimated gradient $\widehat{\beta}(x)$ is thus

$$\widehat{\beta}(x) = \frac{\left(\sum_{i=1}^{n} y_i \frac{\partial k\left(\frac{x_i - x}{h}\right)}{\partial x}\right)\left(\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)\right) - \left(\sum_{i=1}^{n} y_i k\left(\frac{x_i - x}{h}\right)\right)\left(\sum_{i=1}^{n} \frac{\partial k\left(\frac{x_i - x}{h}\right)}{\partial x}\right)}{\left(\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)\right)^2}, \tag{15}$$

where, for example, $\frac{\partial k\left(\frac{x_i - x}{h}\right)}{\partial x} = \left(\frac{x_i - x}{h^2}\right) k\left(\frac{x_i - x}{h}\right)$ for the Gaussian kernel. Higher-order derivatives

can be derived in a similar manner.

2.2.2 Local-Linear Least-Squares

While the LCLS estimator is intuitive, it suffers from biases near the boundary of the support of the data. As an alternative, most applied researchers use the local-linear least-squares (LLLS) estimator. The LLLS estimator locally fits a line as opposed to a constant.

The local-linear estimator is obtained by taking a first-order Taylor approximation of equation (10) via

$$y_i \approx m(x) + (x_i - x)\beta(x) + \epsilon_i,$$

where $\beta(x)$ is the gradient. Similar to the LCLS case, by labeling $m(x)$ and $\beta(x)$ as the parameters a and b, we get the following minimization problem

$$\min_{a,b} \sum_{i=1}^{n} [y_i - a - (x_i - x)b]^2 k\left(\frac{x_i - x}{h}\right),$$

which, in matrix notation (with q regressors) is

$$\min_{\delta} (y - X\delta)' K(x)(y - X\delta),$$

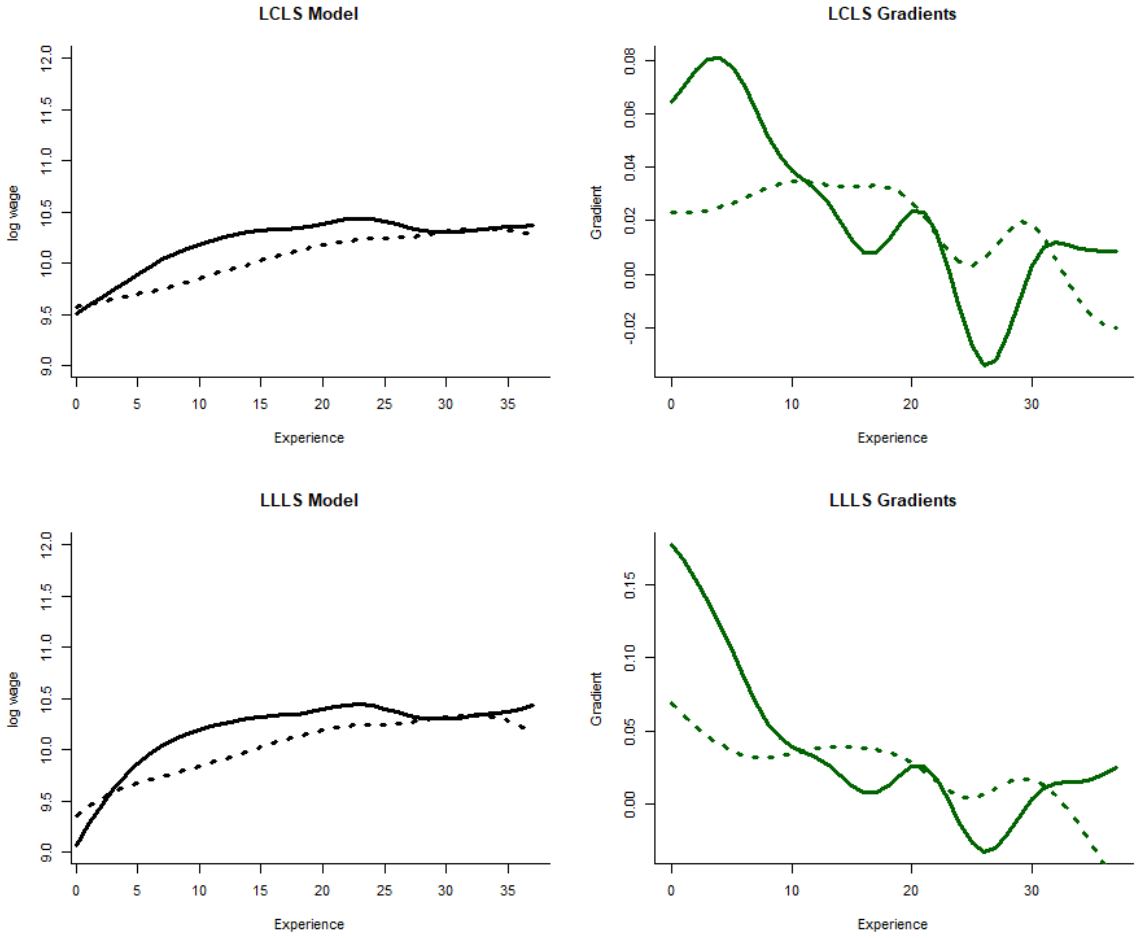
where $\delta = (a, b)'$, X is a $n \times (q+1)$ matrix with its ith row equal to $(1, (x_i - x))$ and $K(x)$ is a $n \times n$ diagonal matrix with its ith element equal to $k\left(\frac{x_i - x}{h}\right)$. This leads to the LLLS estimators of the conditional expectation ($\widehat{m}(x)$) and gradient ($\widehat{\beta}(x)$) as

$$\widehat{\delta}(x) = \begin{pmatrix} \widehat{m}(x) \\ \widehat{\beta}(x) \end{pmatrix} = (X'K(x)X)^{-1} X'K(x)y,$$

Notice that we can obtain the OLS estimator by replacing $K(x)$ by an identity matrix (giving all observations equal weight, i.e., bandwidth tending towards infinity), the weighted least-squares (WLS) estimator by replacing it with some other weighting function, and the generalized least-squares (GLS) estimator by replacing it with the inverse of the variance-covariance matrix of the errors (Ω).

Figure 5 gives both the LCLS and LLLS estimates for white (solid line) and non-white (dashed line) college-educated males working in personal care and service. The gradients for each level of experience are also shown. Compared to the LCLS model, the LLLS model captures a stronger increase in log wage during the first 5 years of work experience with gradients ranging from 0.10 to 0.17. If taken literally, after only a year of working in personal care and service, white college-educated males wages increases by almost 17% on average while non-white college-educated males' wages increases by about 7%. The LCLS model, while showing a similar overall

Figure 5: Log-wage versus Experience for White versus Non-white College-educated Males Working in Personal Care and Service



shape, shows a much slower increase in those first few years of work experience with less than 4% increases in wages for non-whites and 5% to 8% increases for whites. Both models suggest that while white workers have much higher percent increases in their wages in the first few years, those year-to-year percent increases in their wages fall below non-white workers after 10 years of experience.

2.2.3 Local-Polynomial Least-Squares

The derivation of the LLLS estimator can be generalized to include higher-order expansions. The resulted family of estimators are called local-polynomial least-squares (LPLS) estimators. For the general case, if we are interested in the pth-order Taylor expansion, and we assume that the $(p+1)$th derivative of the conditional mean at the point x exists, we can write our equation as

$$y_i \approx m(x) + (x_i - x)\frac{\partial m(x)}{\partial x} + (x_i - x)^2 \frac{\partial^2 m(x)}{\partial x^2}\frac{1}{2!} + ... + (x_i - x)^p \frac{\partial^p m(x)}{\partial x^p}\frac{1}{p!} + \epsilon_i.$$

Replacing the parameters by $(a_0, ..., a_p)$, our kernel weighted least-squares problem can be written as

$$\min_{a_0,...,a_p} \sum_{i=1}^{n} [y_i - a_0 - (x_i - x)a_1 - (x_i - x)^2 a_2 - ... - (x_i - x)^p a_p]^2 k\left(\frac{x_i - x}{h}\right).$$

In matrix notation, our objective function becomes

$$\min_{\delta} (y - X\delta)' K(x)(y - X\delta)$$

where the only difference from the LLLS case ($p = 1$) is that the ith row of X is defined as $[1, (x_i - x), (x_i - x)^2, ..., (x_i - x)^p]$ and $\delta = (a_0, a_1, ..., a_p)'$. Minimizing the objective function leads to the local-polynomial least-square estimator

$$\widehat{\delta}(x) = \left(\widehat{m}(x), \frac{\partial \widehat{m}(x)}{\partial x}, \frac{\partial^2 \widehat{m}(x)}{\partial x^2}, ..., \frac{\partial^p \widehat{m}(x)}{\partial x^p}\right)' = (X' K(x) X)^{-1} X' K(x) y$$

The first question then becomes, how many expansions should we take? More expansions lead to less bias, but increased variability. This becomes a bigger problem when the number of covariates (q) is large and the sample size (n) is small. One promising data driven method to determine the number of expansions is considered in Hall and Racine [2015].

As is the case for splines, there exist options to employ these methods in popular software packages. In `R` we recommend the *np* package (Hayfield and Racine [2008]) and in Stata we recommend the *npregress* command.

3 Model Selection

For both spline and kernel regression, many seemingly arbitrary choices can greatly influence fit. The typical trade-off is between bias and variance. We want to make selections such that we avoid overfitting or underfitting. In this section, we first discuss penalty selection, knot selection, and degree selection in spline models; and then, kernel and bandwidth selection in kernel models.

3.1 Spline Penalty and Knot Selection

In Section 2.1, we saw that the fit is influenced by both our choice of degree of the piecewise polynomials, and by the number and locations of knots we include. However, in spline models, there is a third, more direct way, to influence fit: add an explicit penalty. In short, we want to select the degree of the piecewise polynomials, the knot locations, and the smoothing parameter

λ (penalty) which best capture the underlying shape of our data. Though we will briefly discuss the selection of all three, it is easy to show that the choices of degree and knots are much less crucial than the choice of λ, the smoothing parameter (we will see a similar result for kernel regression). That is, when using a high enough number of knots and degrees, the "smoothness" of our fit can be controlled by λ. Hence, we will focus most of our discussion on the choice of λ when the degree and number of knots are fixed. Although there exist several ways to select our parameters in a data-driven manner, we will concentrate on one of the most commonly used approaches: cross-validation (CV).

3.1.1 Penalty Selection using Cross Validation

There are several ways to impose a penalty, but here we focus on a method that avoids extreme values (and hence too much variability). In a univariate setting using a linear spline, this penalty is

$$\sum \beta_{1\mathbb{k}}^2 < C,$$

where $\beta_{1\mathbb{k}}^2$ is the coefficient on the $\mathbb{k}$th knot[3]. In matrix form, our constrained objective function can thus be written as

$$\min_{\beta} || \ y - X\beta \ ||^2 \ \text{ s.t. } \beta' D \beta \leq C,$$

and leads to Lagrangian[4]

$$\mathcal{L}(\beta, \lambda) = \min_{\beta,\lambda} || \ y - X\beta \ ||^2 + \lambda^2 \beta' D \beta, \tag{16}$$

where D is a $(\mathbb{K}+2) \times (\mathbb{K}+2)$ matrix with diagonal $(0, 0, 1_1, ..., 1_{\mathbb{K}})$. Note that consistency will require that λ tends towards zero as the sample size (n) tends towards infinity.

The second term of (16) is called a roughness penalty because it penalizes through the value of the smoothing parameter (λ) the curvature of our estimated function. This type of regression is referred to as a penalized spline (p-spline) regression and yields the following solution and fitted values:

$$\widehat{\beta}_\lambda = (X'X + \lambda^2 D)^{-1} X'y$$

$$\widehat{y} = X(X'X + \lambda^2 D)^{-1} X'y.$$

To generalize these results to the pth degree spline model (equation (8)), we replace λ^2 by λ^{2p}[5] and transform the D-matrix: $D = \text{diag}(0_{p+1}, 1_{\mathbb{K}})$. A penalized B-spline (PB-spline) would

[3] The matrix of the coefficients being $\beta = [\beta_0, \beta_1, \beta_{11}, ..., \beta_{1\mathbb{K}}]'$

[4] Note that the last term $-\lambda^2 C$ disappears as it does not influence the solution

[5] We raise λ to the power of $2p$ because the way we add bases. Intuitively, raising λ to the power of $2p$ can be explained by the following example: if we transform X into αX for any $\alpha > 0$, we want to have the equivalent transformation done on the smoothing parameter $\lambda \to \alpha\lambda$ to get the same fit.

simply include the transformation done to X (i.e., the square invertible matrix L_p) into the penalty term as well:

$$\widehat{y} = X_B(X_B'X_B + \lambda^{2p}L_p'DL_p')^{-1}X_B'y.$$

As $\lambda^{2p} \to \infty$ (infinite smoothing), the curvature penalty becomes predominant and the estimate converges to OLS. As $\lambda^{2p} \to 0$, the curvature penalty becomes insignificant. In this case, the function will become rougher (we will see a similar result with the bandwidth parameter for a LLLS regression). Figure 6 illustrates this effect using linear p-spline estimates for college-educated males working in personal care and service. The knots have been fixed at every five years of experience (0,5,10,...). As the penalty (λ) increases, it is clear that the fit becomes smoother and converges to an OLS estimate.

Figure 6: Log-wage versus Experience for College-educated Males Working in Personal Care and Service with Different Penalty (λ) Factors

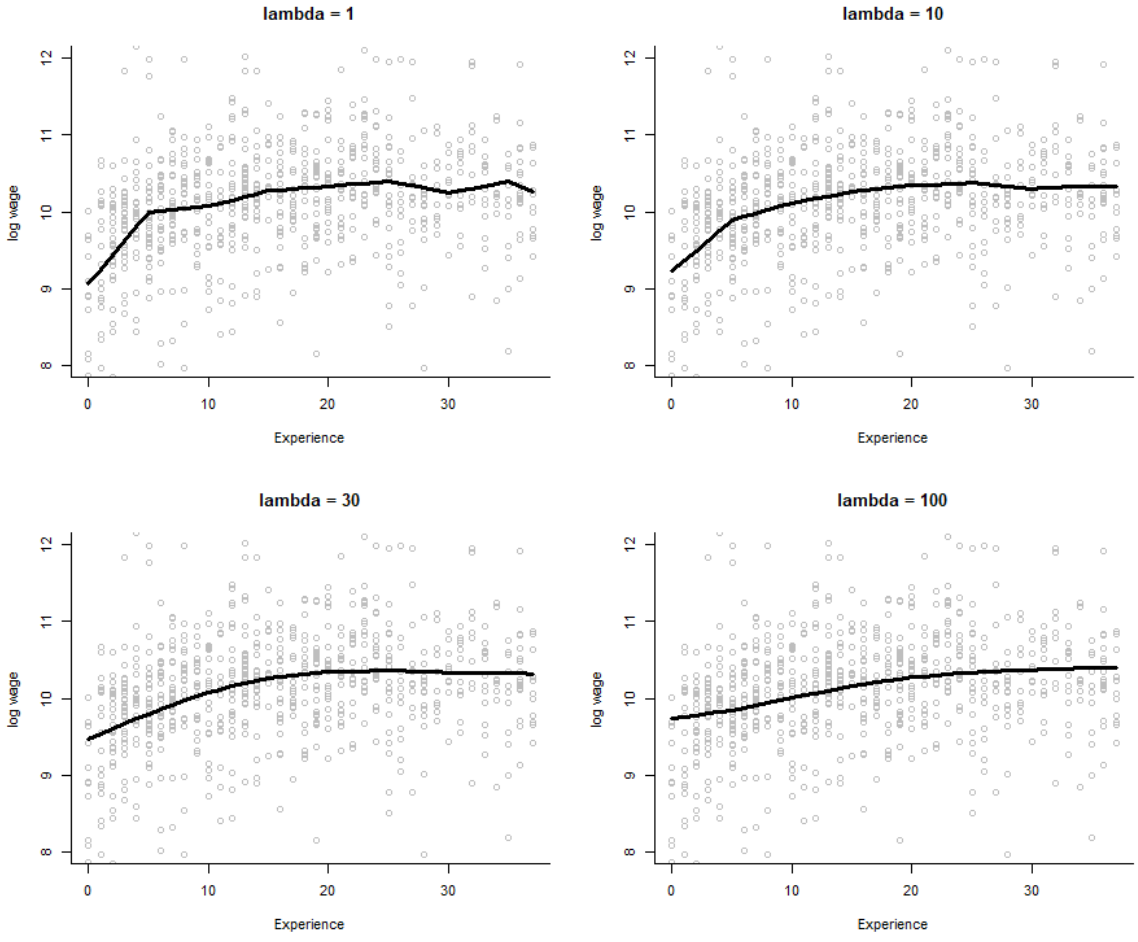


Figure 6 shows an intuitive fit of the data for a value of λ around 10. However, using a more systematic method to select λ would lead to less subjective and more comparable results. If we let $\widehat{m}(x_i;\lambda)$ be our nonparametric regression estimate at the point x with smoothing parameter

λ, we can write a residual sum of squares objective function as

$$RSS(\lambda) = \sum_{i=1}^{n} \left[y_i - \widehat{m}(x_i; \lambda) \right]^2 . \tag{17}$$

The problem with this approach, is that $\widehat{m}(x_i; \lambda)$ uses y_i as well as the other observations to predict y_i. This objective function is minimized when $\lambda = 0$. This problem can be avoided by using a leave-one-out estimator. Least-Squares Cross-Validation (LSCV) is the technique whereby we minimize equation 17, where the fit is replaced by a leave-one-out estimator

$$CV(\lambda) = \sum_{i=1}^{n} \left[y_i - \widehat{m}_{-i}(x_i; \lambda) \right]^2 , \tag{18}$$

where $\widehat{m}_{-i}(\cdot)$ is our leave-one-out estimator, and is defined as our original nonparametric regression estimator $\widehat{m}(\cdot)$ applied to the data, but with the point (x_i, y_i) omitted. We will thus choose a smoothing parameter $\widehat{\lambda}_{CV}$ that will minimize $CV(\lambda)$ over $\lambda \geq 0$.

Using the same number of knots, the top panel of Figure 7 shows the corresponding CV and RSS curves at different values of λ. We can see that the RSS curve is strictly increasing as theory predicts and would choose a lambda of zero. The CV curve, on the other hand, decreases at first and reaches a minimum when $\lambda = 7$. The resulting fit (bottom panel of Figure 7) is smoother than what the RSS criterion would provide.[6]

3.1.2 Knots and Degree Selection

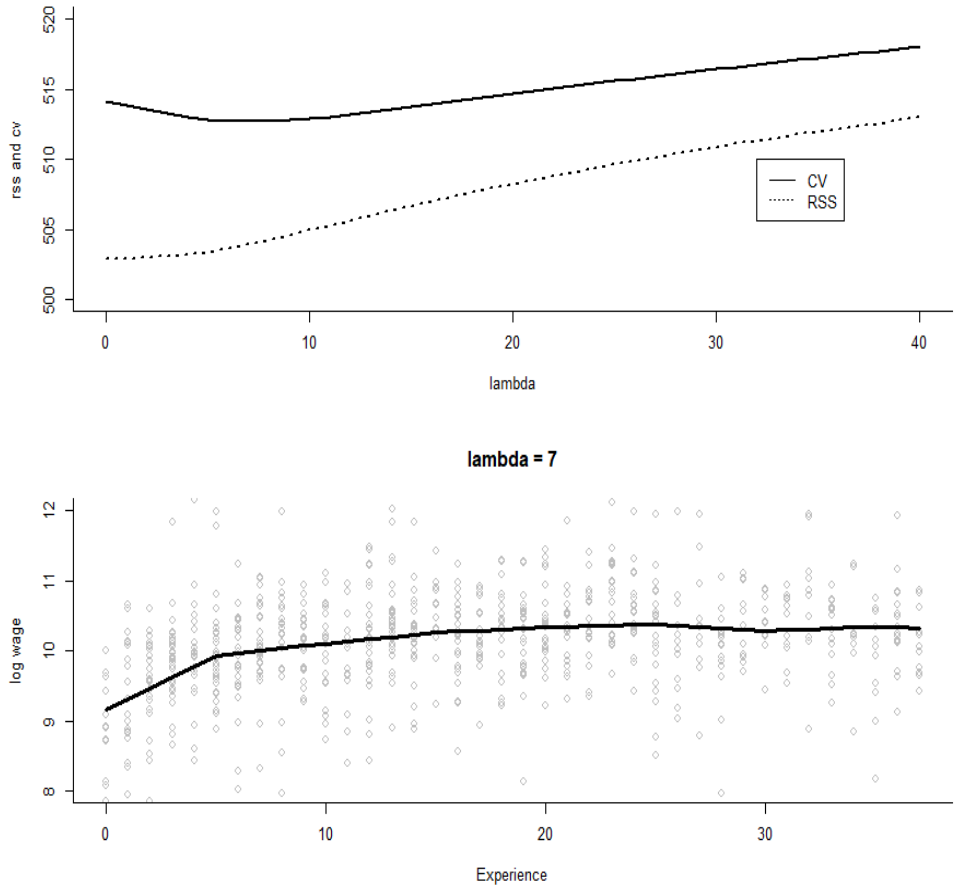
Using an "optimal" lambda and CV criterion, we can compare p-spline models that use different numbers (and location) of knots and different bases (degrees). From experimenting with the number of knots and degrees, the literature finds that (1) adding more knots only improves the fit for a small number of knots (2) when using many knots, the minimum CV for linear and quadratic fits become indistinguishable. In general, we suggest using quadratic or cubic basis functions.

Though there exist more formal criterion to select the number and location of knots, Ruppert et al. [2003] provide simple solutions which often work well. Their default choice of $\mathbb{K}$ is

$$\mathbb{K} = \min\{(1/4) \times \text{ number of unique } x_i, 35\},$$

[6] Note that to compute our CV statistics, we transformed equation 18 to avoid the high computational cost of calculating n versions of $\widehat{m}_{-i}(x_i; \lambda)$ (i.e., the order-n^2 algorithm) using fast order-n [Hutchinson and De Hoog, 1985].

Figure 7: Objective Functions for Choosing Penalty Factors for Linear P-splines for College-educated Males Working in Personal Care and Service



where $\mathbb{K}$ is the number of knots. For knot locations they suggest

$$\kappa_{\mathbb{K}} = \left(\frac{\mathbb{k}+1}{\mathbb{K}+2}\right)\text{th sample quantile of the unique } x_i$$

for $\mathbb{k} = 1, ..., \mathbb{K}$.

Eilers and Marx [1996, 2010] argue that equally spaced knots are always preferred. Eilers and Marx [2010] present an example where equally spaced knots outperform quantile spaced knots. The best type of knot spacings is still under debate and both methods are still commonly used.[7]

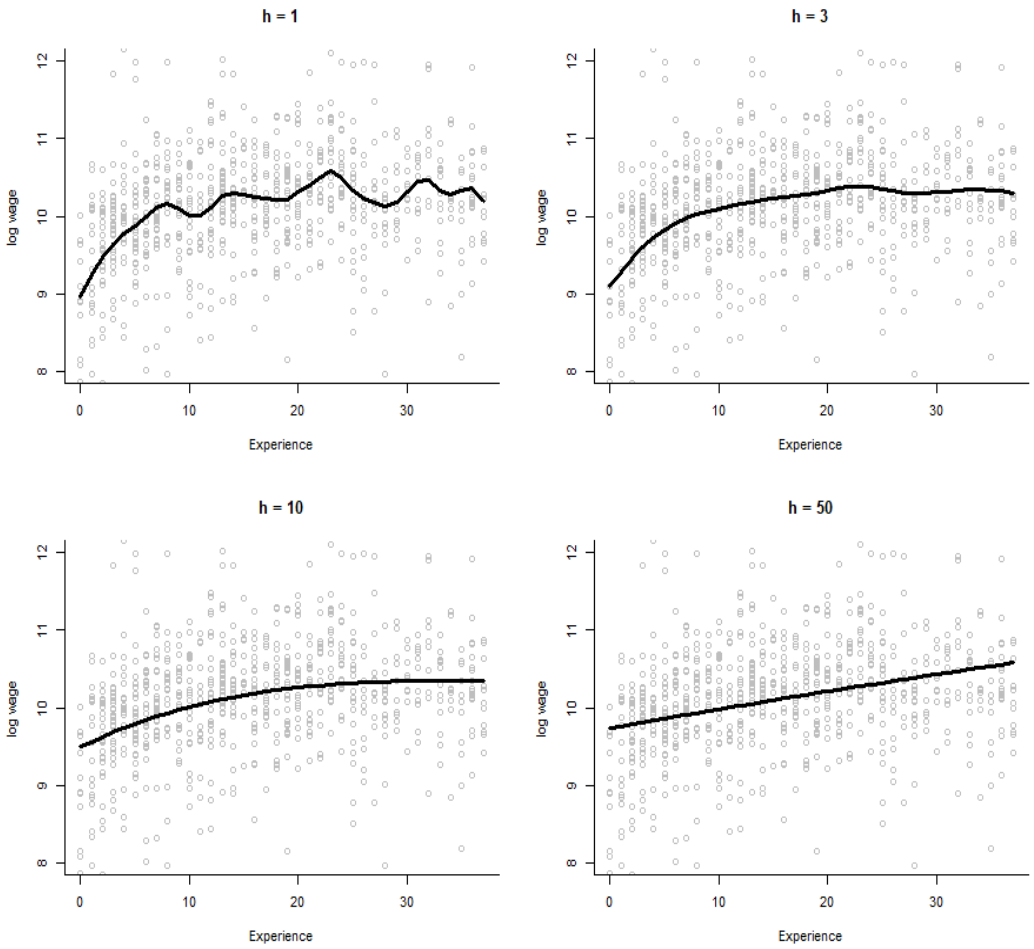
[7] Montoya et al. [2014] use a simulation to test the performance of different knot selection methods with equidistant knots in a p-spline model. Specifically, they compare the methods presented in Ruppert et al. [2003] with the myopic algorithm knot selection method, and the full search algorithm knot selection method. Their results show that the default choice method performs just as well or better than the other methods when using different commonly used smoothing parameter selection methods.

While knots' location and degree selection usually have little effect on the fit when using a "sufficiently" large amount of knots, they may become important when dealing with more complex problems. For example, when trying to smooth regression functions with strong varying local variability or with sparse data. In these cases, using a more complex algorithm to make your selection may be more appropriate.

3.2 Kernel and Bandwidth Selection

Choosing a kernel function is similar to choosing the degree of the piecewise polynomials in spline models, and choosing the size of the bandwidth (h) is similar to choosing the number and location of knots. There exist equivalents to having a direct penalty (λ) incorporated in a kernel model, but those are rarely used in applied kernel estimation. We will therefore focus our discussion on kernel and bandwidth selection.

Figure 8: Log-wage versus Experience for College-educated Males Working in Personal Care and Service when Varying the Bandwidth (h) Parameter



Similar to adding more knots or decreasing the penalty λ in a spline model, decreasing the

bandwidth will lead to less bias, but more variance. Figure 8 illustrates this effect using LLLS and a Gaussian kernel for college-educated males working in personal care and service. As the size of the bandwidth (h) increases, the fit becomes smoother and converges to OLS.

Choice of bandwidth and kernel can be chosen via the asymptotic mean square error (AMSE) criterion (or more specifically, via asymptotic mean integrated square error). In practice, the fit will be more sensitive to a change in bandwidth than a change in the kernel function. Reducing the bandwidth (h), leads to a decrease in the bias at the expense of increasing the variance. In practice, as the sample size (n) tends to infinity, we need to reduce the bandwidth (h) slowly enough so that the amount of "local" information (nh) also tends to infinity. In short, consistency requires that

$$\text{as } n \to \infty, \text{ we need } h \to 0 \text{ and } nh \to \infty.$$

The bandwidth is therefore not just some parameter to set, but requires careful consideration. While many may be uncomfortable with an estimator that depends so heavily on the choice of a parameter, remember that this is no worse than pre-selecting a parametric functional form to fit your data.

3.2.1 Cross-Validation Bandwidth Selection

In practice, there exist several methods to obtain the "optimal bandwidth" which differ in the way they calculate asymptotic mean square-error (or asymptotic mean intergrated square-error). Three typical approaches to bandwidth selection are: (1) reference rules-of-thumb, (2) plug-in methods and (3) cross-validation methods. Each has its distinct strengths and weaknesses in practice, but in this survey we will focus on the data driven method: cross-validation.[8] Henderson and Parmeter [2015] provide more details on each of these methods.

LSCV is perhaps the most popular tool for cross-validation in the literature. This criterion is the same as the one described in Section 3.1.1 to select the penalty parameter in spline regression. That is, we use a leave-one-out estimator

$$CV(h) = \sum_{i=1}^{n} \left[y_i - \widehat{m}_{-i}(x_i) \right]^2, \tag{19}$$

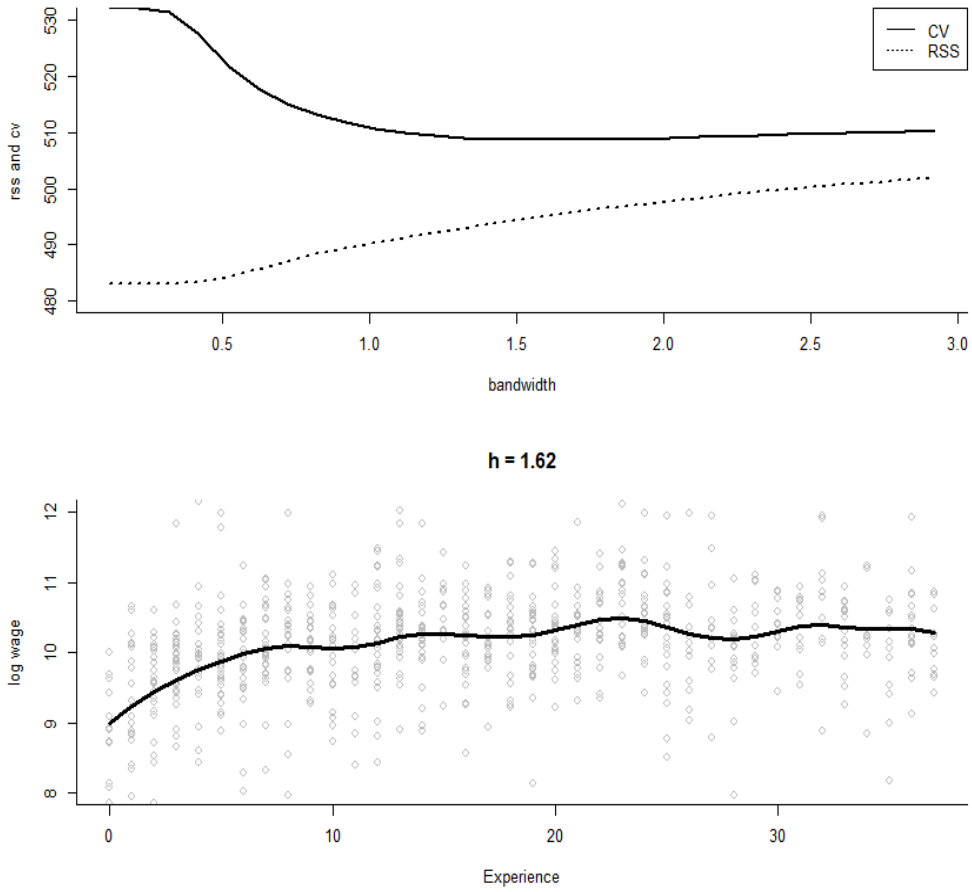
whereby we minimize the objective function with respect to h instead of λ and the (LCLS)

[8] While there is no theoretical justification for doing so, it is common to use rule-of-thumb methods designed for density estimation as a form of exploratory analysis. In fact, we used a rule-of-thumb to compute the bandwidth in our previous examples (Section 2). In its general form, the bandwidth (designed for Gaussian densities with a Gaussian kernel) is $h_{rot} = 1.06\sigma_x^2 n^{-1/5}$. For the remainder of the article, we will use bandwidths selected via cross-validation.

leave-one-out estimator is defined as

$$\widehat{m}_{-i}(x_i) = \frac{\sum_{\substack{j=1 \\ j \neq i}}^{n} y_j K_h(x_j, x_i)}{\sum_{\substack{j=1 \\ j \neq i}}^{n} K_h(x_j, x_i)}.$$

Figure 9: Objective Functions for Choosing Bandwidths for Kernel Estimators for College-educated Males Working in Personal Care and Service



In the top panel of Figure 9, we show an analogous figure to that presented in Section 3.1.1. It shows the corresponding CV and RSS curves for different bandwidths. When failing to use the leave-one-out estimator, the RSS curve is strictly increasing (i.e., the optimal bandwidth is zero). Using the leave-one-out estimator, the objective function is minimized at $h = 1.62$. The resulting fit, (bottom panel of Figure 9) shows more variation than the linear p-spline (Figure 7). This is not surprising as the linear p-spline forces a linear fit between each knot. The two graphs would have looked more similar if we had used a cubic p-spline, allowing for curvature

between knots.

3.2.2 Kernel Function Selection

Kernel selection is typically considered to be of secondary performance as it is believed that it makes minor differences in practice. The optimal kernel function, in the AMISE sense, is the Epanechnikov kernel function. However, as stated previously, it may not be useful in some situations as it does not posses more than one derivative. Gaussian kernels are often used in economics as they possess derivatives of any order, but there are losses in efficiency. In the univariate density case, the loss in efficiency is around 5%. However, Table 3.2 of Henderson and Parmeter [2015] shows that this loss in efficiency increases with the dimension of the data (at least in the density estimation case). In practice, it may make sense to see if the results of a study are sensitive to the choice of kernel.

3.3 Splines versus Kernels

In these single-dimension cases, our spline and kernel estimates are more or less identical. Spline regressions have the advantage that they are much faster to compute. While it is uncommon to have an economic problem with a single covariate, if that were the case, we likely would suggest splines.

In a multiple variable setting, the difference between the two methods are more pronounced. The computation time for kernels increases exponentially with the number of dimensions. The additional computational time required for splines is minor. On the other hand, kernels handle interactions and discrete regressors (see Ma et al. [2015] for using discrete kernels with splines) well (both common features in economic data). It is also relatively easier to extract gradients with kernel methods.

In reality there are camps: those who use kernels and those who use splines. However, the better estimator probably depends upon the problem at hand. Both should be considered in practice.

4 Instrumental Variables

Nonparametric methods are not immune to the problem of endogeneity. A first thought about how to handle this issue would be to use some type of nonparametric two-stage least-squares procedure. However, this is not feasible as there exists an ill-posed inverse problem (to be discussed below). It turns out that this problem can be avoided by using a control function approach much like that in the parametric literature (e.g., see Cameron and Trivedi [2010]).

To motivate this problem, consider a common omitted-variable problem in labor economics: ability in the basic compensation model. A correctly specified wage equation could be described as:

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 z_1 + \beta_3 abil + \epsilon \tag{20}$$

where *educ* is years of education, *abil* is ability, and z_1 is a vector of other relevant characteristics (e.g., experience, gender, race, marital status). However, in applied work, ability (*abil*) cannot be directly measured/observed.

If we ignore ability (*abil*), it will become part of the error term

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 z_1 + u, \tag{21}$$

where $u = \epsilon + \beta_3 abil$ and *abil* is correlated with both u and with *educ*. Our resulting estimated return to education (β_1) will be biased and inconsistent. We can resolve this problem if we can find an instrumental variable (IV) which is uncorrelated with u (and so uncorrelated with ability), but correlated with *educ*. Several IVs have been considered in the literature for this particular model,[9] each with their own strengths and weaknesses, but for the purpose of this illustration, we will use spouse's wage. That is, we assume that spouse's wage is correlated with education, but not with ability.

In the parametric setting, the control function (CF) approach to IVs is a two-step procedure. In the first step, we regress the endogenous variable on the exogenous vector z:

$$educ = \gamma_0 + \gamma_1 z + v,$$

where $z = (z_1, spwage)$ and *spwage* is the spouse's wage, and obtain the reduced form residuals $\widehat{v}$. In the second step, we add $\widehat{v}$ to equation 21 and regress

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 z_1 + \beta_3 \widehat{v} + u.$$

By directly controlling for v, *educ* is no longer viewed as endogenous.

4.1 The Ill-Posed Inverse Problem and Control Function Approach

Let us first go back and consider the general nonparametric regression setting

$$y = m(x) + u, \tag{22}$$

[9] Those IVs include, but are not limited to: minimum school-leaving age, quarter of birth, school costs, proximity to schools, loan policies, school reforms, spouse's and parents' education/income.

where $E[u|x] \neq 0$, but there exists a variable z such that $E[u|z] = 0$. For the moment, assume that x and z are scalars.

Using the condition

$$E[u|z] = [y - m(x)|z] = 0,$$

yields the conditional expectation

$$E[y|z] = [m(x)|z] = \int m(x) f(x|z) dx. \tag{23}$$

Although we can estimate both the conditional mean of y given z ($E[y|z]$) as well as the conditional density of x given z ($f(x|z)$), we cannot recover $m(x)$ by inverting the relationship. That is, even though the integral in equation 23 is continuous in $m(x)$, inverting it to isolate and estimate $m(x)$ does not represent a continuous mapping (discontinuous). This is the so-called ill-posed inverse problem and it is a major issue when using instrumental variables in nonparametric econometrics.

Luckily, we can avoid this problem by placing further restrictions on the model (analogous to additional moment restrictions in a parametric model). Here we consider a control function approach. Similar to the parametric case above, we consider the triangular framework

$$x = g(z) + v, \tag{24}$$

$$y = m(x) + u \tag{25}$$

with the conditions $E[v|z] = 0$ and $E[u|z, v] = E[u|v]$. The first condition implies that z is a valid instrument for x and the second allows us to estimate $m(x)$ and avoid the ill-posed inverse problem. It does so by restricting u to depend on x only through v. More formally,

$$\begin{aligned} E[y|x, v] &= m(x) + E[u|x - v, v] \\ &= m(x) + E[u|z, v] \\ &= m(x) + E[u|v] \\ &= m(x) + r(v), \end{aligned}$$

and hence

$$m(x) = E[y|x, v] - r(v), \tag{26}$$

where both $E[y|x, v]$ and $r(v)$ can be estimated nonparametrically. In short, we control for v through the nonparametric estimation of the function $r(v)$.

4.2 Spline Regression with Instruments

Now that we have the basic framework, we can discuss nonparametric estimation in practice. Consider our previous compensation model, but without functional form assumptions:

$$log(wage) = m(educ, z_1) + u \tag{27}$$

where ability (*abil*) is unobserved. It is known that *abil* will be correlated with both the error u and the regressor *educ* (i.e., $E[u|\text{educ}] \neq 0$). Similar to the parametric setting, if we have an instrument z such that

$$E[u|z] = [log(wage) - m(educ, z_1)|z] = 0,$$

we can avoid the bias due to endogeneity. We again define $z = (z_1, spwage)$, where our excluded instrument, *spwage*, is the spouse's wage. This yields

$$E[log(wage)|z] = [m(educ, z_1)|z].$$

Our problem can now be written via the triangular system attributable to Newey et al. [1999]:

$$educ = g(z) + v \tag{28}$$

$$log(wage) = m(educ, z_1) + u, \tag{29}$$

where $E[v|z] = 0$ and $E[u|z, v] = E[u|v]$. Similar to the parametric case, we first estimate the residuals from the reduced-form equation (i.e., $\widehat{v}$). We then include the reduced-form residuals nonparametrically as an additional explanatory variable:

$$log(wage) = w(educ, z_1, \widehat{v}) + u, \tag{30}$$

where $w\left(educ, z_1, \widehat{v}\right) \equiv m\left(educ, z_1\right) + r\left(\widehat{v}\right)$ and $\widehat{m}\left(educ, z_1\right)$ can be recovered by only extracting those terms that depend upon *educ* and z_1. Note that we need to use splines that do not allow for interactions between *educ* or z_1 and u (interactions between *educ* and z_1 are allowed).

In what follows, we will use cubic B-splines (the default for most `R` packages) with $\mathbb{K} = \min\{(1/4) \times \text{ number of unique } x_i, 35\}$ equi-quantile knots (see Section 3) in both stages. Figure 10 shows the fitted results for the first stage. In the left panel, we see that as individuals' experience increases, the level of education slowly decreases, with a significant drop after 35 years of work experience. In the right panel, we observe a quadratic relationship between education level and spouse's wage. That is, the higher the spouse's wage, the higher the individual's education, but for individuals whose spouse have a high level of income, the relationship becomes

negative.

Figure 10: First-stage Estimates for Education for College-educated Males Working in Personal Care and Service versus Experience and Spousal Income

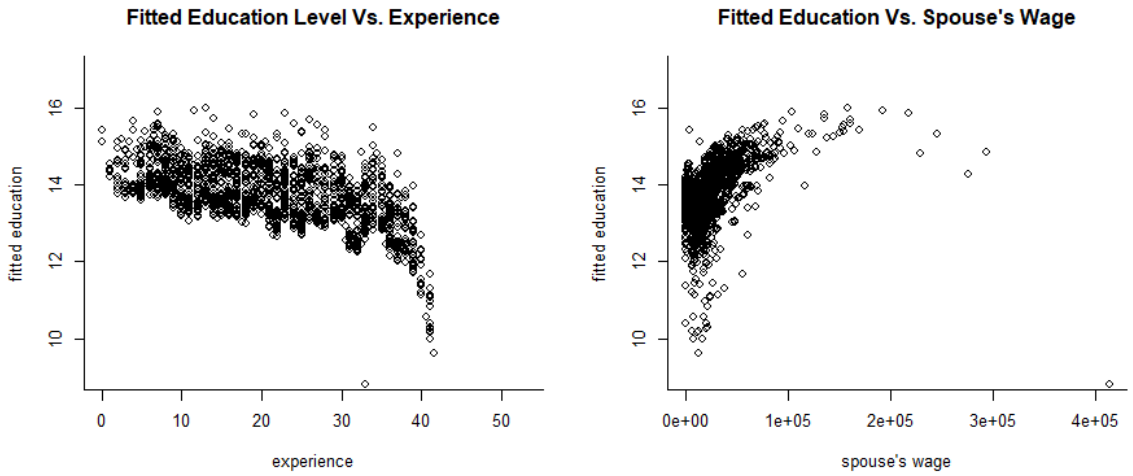
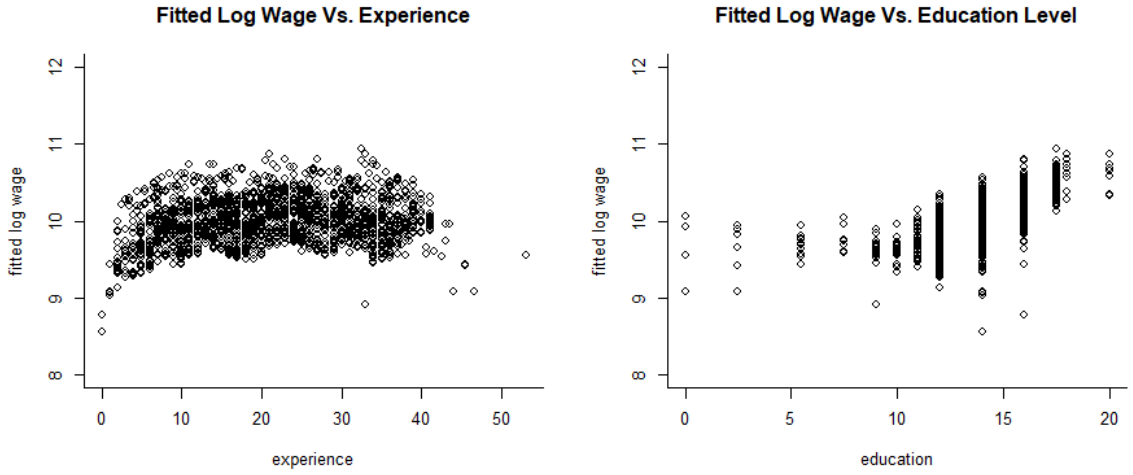


Figure 11: Second-stage Estimates for Log-wage for College-educated Males Working in Personal Care and Service versus Experience and Education Level



The fitted plots from the second stage are given in Figure 11. Controlling for education,[10] men's log wage seems to increase in the first few years on the job, stabilizes mid-career, and then decreases towards the end of their career in personal care and service. Education seems to affect log wage positively only after 10-12 years of schooling (high-school level).

[10]Recall that in our previous examples, the level of education is fixed at 16 years – college degree.

4.3 Kernel Regression with Instruments

Estimation via kernels is relatively straightforward given what we have learned above. We again use a control function approach, but with local-polynomial estimators. Kernel estimation of this model was introduced by Su and Ullah [2008] and is outlined in detail in Henderson and Parmeter [2015]. In short, the first stage requires running a local-p_1th order polynomial regression of the endogenous regressor on z, obtaining the residuals and then running a local-p_2th order polynomial regression of y on the endogenous regressor, the included exogenous regressors and the residuals from the first stage.

More formally, our first stage regression model for our example is

$$educ = g(z) + v, \tag{31}$$

and the residuals from this stage are used in the second stage regression

$$log(wage) = w(educ, z_1, \widehat{v}) + u, \tag{32}$$

where $w(educ, z_1, \widehat{v}) \equiv m(educ, z_1) + r(\widehat{v})$.

In spline regression, we simply took the estimated components not related to $\widehat{v}$ from $\widehat{w}(\cdot)$ in order to obtain the conditional expectation $\widehat{m}(educ, z_1)$. However, disentangling the residuals is a bit more difficult in kernel regression. While it is feasible to estimate additively separable models, we follow Su and Ullah [2008] and remove them via counterfactual estimates in conjunction with the zero mean assumption on the errors. Under the assumption that $E(u) = 0$, we recover the conditional mean estimate via

$$\widehat{m}(educ, z_1) = \frac{1}{n}\sum_{i=1}^{n} \widehat{w}(educ, z_1, \widehat{v}_i), \tag{33}$$

where $\widehat{w}(educ, z_1, \widehat{v}_i)$ is the counterfactual estimator of the unknown function using bandwidths from the local-p_2 order polynomial regression in the second step (derivatives can be obtained similarly, but summing over the counterfactual derivatives of $\widehat{w}(\cdot)$).[11]

Bandwidth selection and order of the polynomials (p_1 and p_2) are a little more complicated. Here we will give a brief discussion, but suggest the serious user consult Chapter 10 (which includes a discussion of weak instruments) of Henderson and Parmeter [2015].

Bandwidth selection is important in both stages. In the first-stage, v is not observed and we want to make sure that the estimation of it does not impact the second-stage. If we observe

[11] Multiple endogenous regressors can be handled by running separate first stage regressions and putting the residuals from each of those regressions into the second stage regression and finally summing over i to obtain the conditional mean estimates.

the conditions in Su and Ullah [2008], it allows for the following cross-validation criterion

$$CV\left(h_2\right) = \min_{h_2} \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \widehat{m}_{-i}\left(educ_i, z_{1i}\right)\right]^2, \tag{34}$$

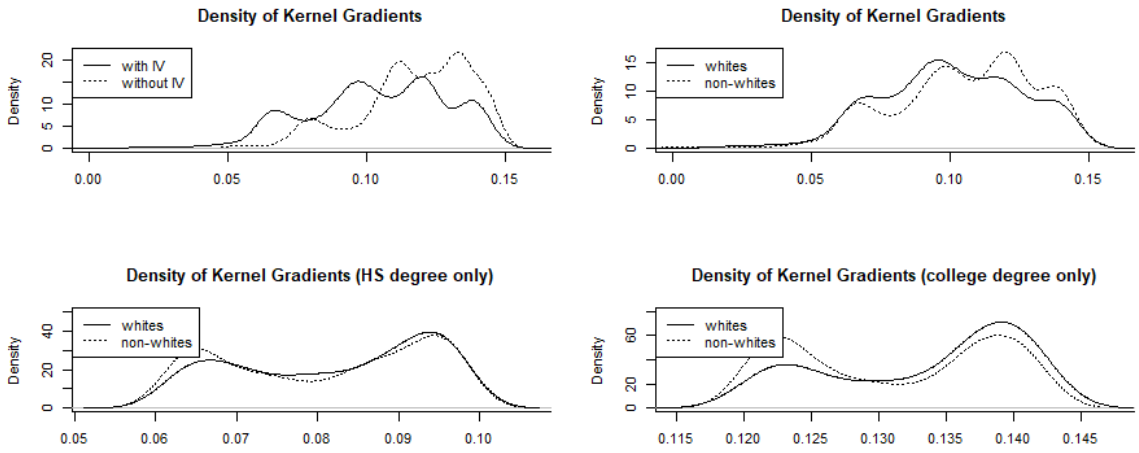
and the first-stage bandwidths can be constructed as

$$\widehat{h}_1 = \widehat{h}_2 n^{-\gamma}, \tag{35}$$

where the acceptable values for γ depend upon the order of the polynomials in each stage.[12]

Henderson and Parmeter [2015] give the admissible combinations of polynomial orders for the Su and Ullah [2008] estimator with a single endogenous variable and a single excluded instrument. In practice, they suggest using a local-cubic estimator in the first stage (local-linear in the first stage is never viable) and a local-linear estimator in the second stage for a just-identified model with a single endogenous regressor. For other cases, the conditions of Assumption A5 in Su and Ullah [2008] need to be checked.

Figure 12: Second-step Gradients of Log-wage with Respect to Education for College-educated Males Working in Personal Care and Service



Using the methods outlined above, Figure 12 shows the impact of controlling for endogeneity. The upper-left panel gives density plots for the gradient estimates across the sample for returns to education both with and without using instrumental variables. Most college-educated men working in personal care and service have a wage increase of about 5 to 15% for each additional year they spend in school. However, the distribution is skewed to the left suggesting that a few

[12]The acceptable range for γ is between $\left(2\left(p_2+1\right)+q_1+1\right)^{-1} \max\left[\frac{p_2+1}{p_1+1}, \frac{p_2+3}{2(p_1+1)}\right]$ and $\left(2\left(p_2+1\right)+q_1+1\right)^{-1} \frac{p_2+q_1}{q_1+q_2}$, where q_1 and q_2 represent the number of elements in the first and second stage regressions, respectively.

men have seen their investment in education yield no returns or even negative returns (for a similar result in a nonparametric setting see Henderson et al. [2011]).

Comparing those results with the gradients without instruments, we clearly see that failing to control for endogeneity would overestimate the returns to education (as expected). That is, the distribution of gradients without using IV is more concentrated around 10 to 15% returns and fewer low returns.

To try to swing back to the examples from before, the upper-right panel of Figure 12 gives densities of gradient estimates controlling for endogeneity for whites versus non-whites. The figure seems to suggest that non-whites have higher rates of return to education. This is commonly found in the literature, but is often attributed to lower average years of education. To try to compare apples to apples, in the bottom two panels we plot the densities of returns to education for fixed levels of education (high school and college, respectively). Here we see while the general shape is similar, whites tend to get more mass on the higher returns and less mass on lower returns, especially for college graduates.[13]

References

Adrian Colin Cameron and Pravin K Trivedi. *Microeconometrics using Stata*, volume 2. Stata press College Station, TX, 2010.

Paul HC Eilers and Brian D Marx. Flexible smoothing with b-splines and penalties. *Statistical Science*, pages 89–102, 1996.

Paul HC Eilers and Brian D Marx. Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(6):637–653, 2010.

Peter Hall and Jeffrey S Racine. Infinite-order cross-validated local polynomial regression. *Journal of Econometrics*, 185:510–525, 2015.

Tristen Hayfield and Jeffrey S Racine. Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27:1–32, 2008.

Daniel J Henderson and Christopher F Parmeter. *Applied Nonparametric Econometrics*. Cambridge University Press, 2015.

Daniel J. Henderson, Solomon W. Polachek, and Le Wang. Heterogeneity in schooling rates of return. *Economics of Education Review*, 30:1202–1214, 2011.

[13] We could combine spline and kernel methods to obtain an IV estimator as in Ozabaci et al. [2014]. Using this combination allows for a lower computational burden and oracally efficient estimates.

Michael F Hutchinson and FR De Hoog. Smoothing noisy data with spline functions. *Numerische Mathematik*, 47(1):99–106, 1985.

Shujie Ma, Jeffrey S Racine, and Lijian Yang. Spline regression in the presence of categorical predictors. *Journal of Applied Econometrics*, 30:705–717, 2015.

Eduardo L Montoya, Nehemias Ulloa, and Victoria Miller. A simulation study comparing knot selection methods with equally spaced knots in a penalized regression spline. *International Journal of Statistics and Probability*, 3(3):96, 2014.

Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1): 141–142, 1964.

Whitney K Newey, James L Powell, and Francis Vella. Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, 67(3):565–603, 1999.

Deniz Ozabaci, Daniel J Henderson, and Liangjun Su. Additive nonparametric regression in the presence of endogenous regressors. *Journal of Business & Economic Statistics*, 32(4):555–575, 2014.

David Ruppert, Matt P Wand, and Raymond J Carroll. *Semiparametric Regression*. Number 12. Cambridge University Press, 2003.

Liangjun Su and Aman Ullah. Local polynomial estimation of nonparametric simultaneous equations models. *Journal of Econometrics*, 144:193–218, 2008.

Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.