Kernel densities and regression

Richard S.J. Tol[a,b,c,d,e,f], [a] Department of Economics, University of Sussex, Falmer, United Kingdom; [b] Institute for Environmental Studies, Vrije Universiteit, Amsterdam, the Netherlands; [c] Department of Spatial Economics, Vrije Universiteit, Amsterdam, the Netherlands; [d] The Netherlands Tinbergen Institute, Amsterdam, the Netherlands; [e] CESifo, Munich, Germany; and [f] Payne Institute for Public Policy, Colorado School of Mines, Golden, CO, United States

Key concepts

- Kernel density
- Kernel function
- Bandwidth
- Bootstrap: Ordinary, Bayesian, and Smoothed
- Silverman rule
- Adaptive, asymmetrics and higher-order kernels
- Kernel decomposition
- Kernel regression: Nadaraya-Watson and Priestley-Chao
- Smoothing: LOWESS, LOESS and Kriging
- Quantile regression
- Semi-parametric regression: Yatchew and Robinson

Glossary

Kernel function A function that smudges an observation over its neighborhood

Bandwidth A parameter that defines the neighborhood of an observation in a *kernel* function; the bandwidth is the same for all observations unless the kernel is *adaptive*.

Kernel density A probability density function that has no predefined characteristics

Kernel regression A collection of methods to estimate the relationship between variables without specifying a functional form; can be applied to the conditional mean as well as to the conditional quantiles

Semi-parametric regression A collection of methods that combines the flexibility of *kernel regression* for the variable of interest with the computational ease of parametric regression for control variables

Abstract

Kernel densities are a flexible tool to describe data without pre-determining characteristics such as symmetry, unimodality, or kurtosis. Care needs to be taken in the choice of bandwidth and kernel function, although there are rules of thumb to inform this choice. Kernel regression describes the relationship between variables without a pre-determined functional form. Semi-parametric regression offers the same flexibility with lower computational burden.

Keywords

Kernel density; Kernel regression

Introduction

The Normal distribution is the workhorse of statistical analysis. The Central Limit Theorem has that the mean of any sample with finite variance converges to a Normal distribution as the sample grows. Other statistics also converge to Normality. However, samples may be small and the Normal distribution has particular properties that may not describe the data very well: It is unimodal, symmetric, and has thin tails. There are many other probability distributions. As with the Normal distribution, a few parameters describe these completely. Their characteristics, therefore, need not match the data. A *kernel* distribution, on the other hand, describes the data as they are. Kernel distributions have many applications, including in regression analysis. Parametric regression requires that you specify the functional form of the relation between the dependent and explanatory variables; typically, a linear equation is assumed. You also have to assume that the regression residuals follow a particular distribution, typically the Normal one, lest statistical tests are invalid. Kernel regression is less restrictive, allowing for flexible functional form and distributional assumptions.

This chapter discusses kernel densities in Section **Kernel densities** and kernel regression in Section **Kernel regression**. I emphasize intuition and appropriate application over technical detail. There are many books on non-parametric methods in statistics, filled with technicalities skipped here. My preferred book is Takezawa (2005), but the interested reader has a wealth of choice. Journals in mathematical statistics provide even more detail. This is not the right place for that. Neither do I discuss statistical software. Most packages have options for kernel density estimation and kernel regression. This chapter helps to understand software manuals, but is no substitute for them. Instead, I aim to give the reader an understanding of what kernel methods can and cannot do, and why.

Kernel densities

A *kernel density* $f(y)$ is defined as

$$f(y) := \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{y - y_i}{h}\right) \tag{1}$$

where $y_i, i = 1, 2, \ldots, N$ is a sequence of N observations, K is the *kernel function* and h is its *bandwidth*. The kernel function peaks at 0 and integrates to 1.

The intuition is this. The density f of y is determined by the weighted sum of all observations y_i, where the observations closest to y are given the largest weight. A kernel distribution can be interpreted as a mixture distribution, where every observation gets its own component in the mixture. In this interpretation, probabilities are *added*, rather than multiplied. A kernel density is a form of vote-counting, rather than belief-updating.

Fig. 1 illustrates how it works. There are 25 observations of the carbon tax efficacy, the percentage of carbon dioxide emissions reduced from what they otherwise would have been in the first decade of imposing a price on emissions (Tol, 2023). Each observation is assigned a kernel function, a Normal one in this case, which is centered on the observation but with a spread around it. These are the thin lines. The kernel density is then the sum of the kernel functions. The result is a bimodal distribution, with a pronounced mode somewhere in the middle of 24 of the 25 observations and a second mode around the one outlier.

Fig. 1 was constructed in MS Excel. This is not a serious programming language, but I find it an excellent way to prototype code, as Excel allows, indeed almost forces you to inspect every intermediate step in your algorithm. Coding a kernel density estimator is straightforward. It is not necessary—canned routines do this for you—but coding yourself reveals how things work. To construct a kernel density in Excel, you first define a grid: 0.015 to 0.070 in steps of 0.001. For each point on the grid and for each observation, you evaluate the kernel function—making sure that the sum over the grid equals one. You then add up across the observations, and divide by the number of observations.

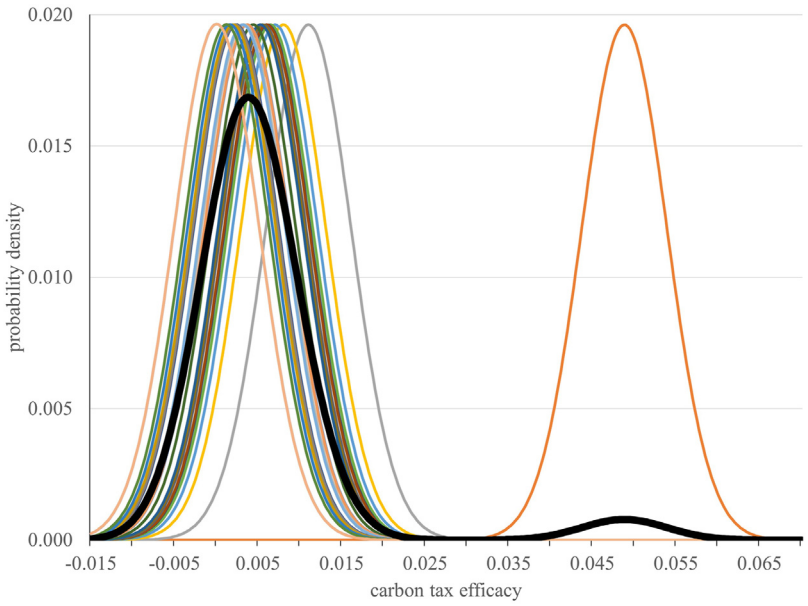

Fig. 1 Construction of a kernel density.
The thin lines show the Normal kernel *functions* for the individual observations of carbon tax efficacy, using the Silverman rule to set the bandwidth. The thick line is the resulting kernel *density*.

Inference

A kernel distribution is an *estimate* of the population distribution. Estimates should come with confidence intervals. That is difficult in this case. A kernel distribution is the weighted sum of kernel functions, rather than its product (for which we have a lot of theory). Note that summation in **Eq. (1)** is *not* the theoretically well-understood summation of randomly distributed variables—we do not add random variables, but smudged individual realizations y_i of the same variable y.

If theory leaves you at sea, there is always the *bootstrap*. In classical statistics, observations are assumed to be random draws from the population about which we make inferences. The bootstrap leverages that insight, but instead of drawing from the unknown population, it draws, with replacement, from the known sample. That is, the bootstrap creates B sets of N pseudo-observations where each set consists of most of the actual observations, some omitted, some duplicated, a few triplicated, and so on. The statistics of the bootstrap converge to the population statistics.

Fig. 2 shows the 90% confidence interval of the kernel distribution of the carbon tax efficacy, already shown in **Fig. 1**. The confidence interval is relatively narrow for most of the kernel distribution, but wider for the outlier. This is intuitive: If that one observation is not included in the pseudo-sample, the secondary mode disappears. However, if the outlier is pseudo-observed twice, the main part of the kernel distribution is less pronounced, which is why the confidence interval around the primary mode is asymmetric.

Note that **Fig. 2** is based on the *Bayesian bootstrap* (Rubin, 1981). Sample statistics are a weighted function of the observations, where all weights are equal. For instance, the sample mean is the sum of the observations divided by N. The statistics of every pseudo-sample in the bootstrap are also weighted functions of the observations, with most weights set to $1/N$, some to 0, some to $2/N$, a few to $3/N$, and so on. The bootstrap statistics are then an (unweighted) function of the pseudo-sample statistics. However, instead of inverse *integer* weights (and a large B, the number of pseudo-samples), we can also use *real* weights, with expectation 1 and adding to N—and a smaller B.

Besides using the bootstrap for inference on kernel distributions (and kernel regression; see below), we can also use kernel distributions in bootstrapping. This is known as the *smooth* bootstrap. Returning to the example, we have 25 observations of carbon tax efficacy. Ordered from large to small, carbon tax efficacy can be 4.9, 1.1, 0.86, 0.71, 0.65, …(unit is tonne of carbon dioxide per dollar). This is peculiar. Surely, carbon tax efficacy could just as well be 0.75 or 0.69, but a standard bootstrap will never return these values. If we draw from the kernel density of the observations[1] rather than directly from the observations, we do see plausible intermediate values. This is particularly influential if there are few observations (as in this case). Do not use the smooth bootstrap for inference on the kernel distribution.

[1]Drawing from the kernel density is not hard. First, create the kernel distribution $F(Y<X)$ by aggregating over the kernel density $F(Y<X)=\sum_{x<X} f(y)$. Second, draw a random number U from a Uniform (0,1) distribution. Third, find X for which $F(Y<X)=U$. Then X is a random draw from your kernel density f.

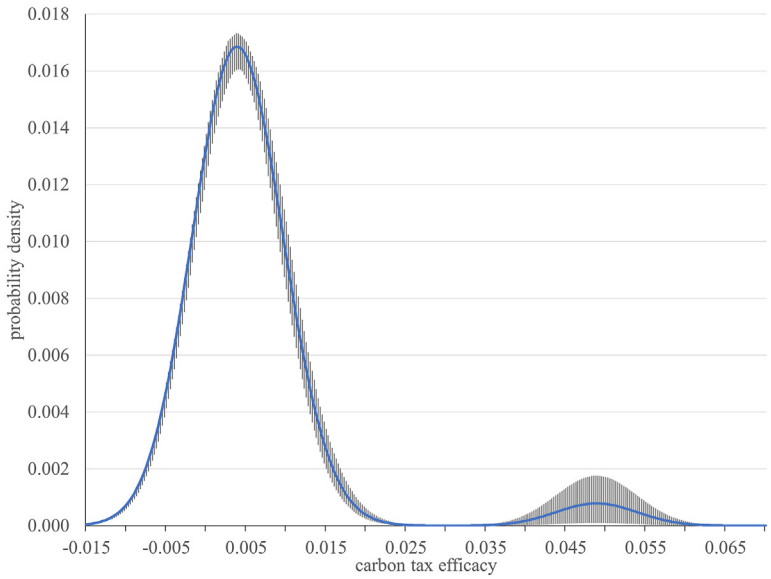

Fig. 2 Uncertainty about a kernel density.
The solid line shows the kernel density of **Fig. 1**, the vertical bars its 90% confidence interval.

Bandwidth

Although kernel methods are often described as "non-parametric", they are not quite. Two important choices are made in **Eq. (1)**. First, bandwidth h is key. **Fig. 1** uses the so-called *Silverman rule*. (See below.) **Fig. 3** shows two alternatives: 3 times the Silverman rule and 0.1 times the Silverman rule. If we use a wider bandwidth, the kernel functions are more diffuse, and so is the kernel density. The second mode disappears. If we use a smaller bandwidth, the kernel functions are sharper. The kernel density now has 5 modes. If we reduce the bandwidth further, every observation has its own mode. Indeed, if the bandwidth approaches zero, the kernel density approaches the histogram.

So, what bandwidth should you use? I think you should rely on your judgment. The bandwidth in a way reflects your confidence in the data. Use a narrow bandwidth if you are fairly certain that your data are accurate. You may have information about measurement uncertainty. The sample standard deviation is another guide. It is a measure, not of the accuracy of the observations, but of the range of values in the population: A different sample would have yielded different observations.

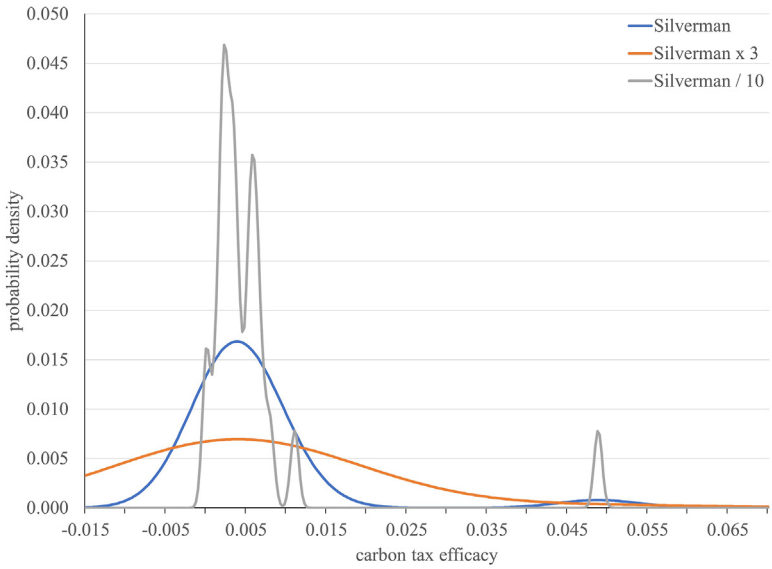

Fig. 3 The choice of bandwidth.
The kernel density of carbon tax efficacy for three alternative bandwidths.

Software packages for kernel density estimation offer a default bandwidth, typically the Silverman rule—and that is the one used in Fig. 1. The rule, due to Silverman (1986), follows from minimizing the Mean Integrated Square Error (MISE):

$$\min_h MISE := E\int_\gamma \left[f(\gamma) - \widehat{f}(\gamma|h)\right]^2 \mathrm{d}\gamma \tag{2}$$

where $\widehat{f}(\gamma|h)$ is the target distribution. MISE is thus the squared distance between the kernel distribution and what you would like the distribution to be.

If the kernel function is the Normal distribution and the target is Normal too, then $h^S = 1.06\sigma n^{-0.2}$, where n is the number of observations and σ is either the sample standard deviation or the interquartile range divided by 1.349, whichever is smaller.

This is, in a way, cheating. You are using a kernel distribution because you did not want to use a parametric distribution. So why would you want your kernel distribution to be as close to a parametric distribution as possible? The advantage of the Silverman rule is that you outsource your decision on bandwidth to a grizzled old statistician.

Adaptive kernels

I assume above that the bandwidth is constant. This is not needed. You can vary the bandwidth over the domain of γ, but this is numerically unstable. It is better to vary the bandwidth by observation. One rule (Abramson, 1982) is

$$h_i \propto \left(\sqrt{f(\gamma_i)}\right)^{-1} \tag{3}$$

This rule uses the kernel function to assess the credibility of observation i. That is, you assign a narrow bandwidth to observations in the middle of the kernel distribution, and a wide bandwidth to outliers.

Abramson's rule is solved iteratively. You start with a constant bandwidth h, find h_i, make sure that $h = \sum_i h_i$, re-estimate the kernel distribution, find the new h_i, and so on until the kernel distribution no longer changes.

Fig. 4 illustrates this, using the data on carbon tax efficacy again. The adaptive kernel converges rapidly. More importantly, the prominent outlier is barely visible after one iteration and not at all after two. Fig. 4 also shows the kernel density on a log scale. The density of the outlier falls by two orders of magnitude per iteration. It is still there, just very unlikely.

Fig. 4 highlights both the power and danger of adaptive kernels. You can make outliers vanish as if they disappeared into the Salvadorean Gulag! You should only do this if you are confident that the outliers are indeed that. Sometimes, "strange" data points hold more information than regular ones. Recall that the first observation of the hole in the ozone layer was dismissed as an outlier due to sensor malfunction.

Kernel functions

Eq. (1) also requires a kernel function K. Typically, a kernel function.

- integrates to one;
- is non-negative;
- is smooth;
- is symmetric; and
- has zero mean.

Strictly, a kernel function only needs to integrate to one.

Many functions meet the above criteria. The standard Normal distribution[2] is a popular choice $K(u) = \frac{1}{\sqrt{2\pi}}e - 0.5u^2$. The Epanechnikov kernel is widely used too:

$$K(u) = \frac{3}{4}\left(1 - u^2\right) \; |u| \leq 1 \tag{4}$$

The Epanechnikov kernel is efficient in the sense that it minimizes the mean integrated squared distance to the true distribution, regardless of what that distribution may be (Epanechnikov, 1969, see Tsybakov (2009) for pushback).

The Epanechnikov kernel has finite support. This implies that the associated kernel distribution assigns zero probability to anything larger (smaller) than the largest (smallest) observations plus (minus) the bandwidth. That may be appropriate for some data but not for others. A Normal kernel assigns a small but positive probability to things outside the observed range. Again, that may be the correct assumption in some situations but not in others. As above, I think you should consider the nature of your data, rather than follow convention, when choosing your kernel.

For instance, it is reasonable to assume that the highest age is slightly above the maximum observed age. An Epanechnikov kernel reflects this. A Normal kernel instead has a small but positive chance of encountering a 33,000-year-old warrior from

[2]The Normal distribution is also known as the Gaussian distribution, after Carl Friedrich Gauss, particularly when used as a kernel function. Pierre-Simon Laplace demonstrated the key features of the Normal distribution long before Gauss did, but he has a different, less prominent distribution named after him.

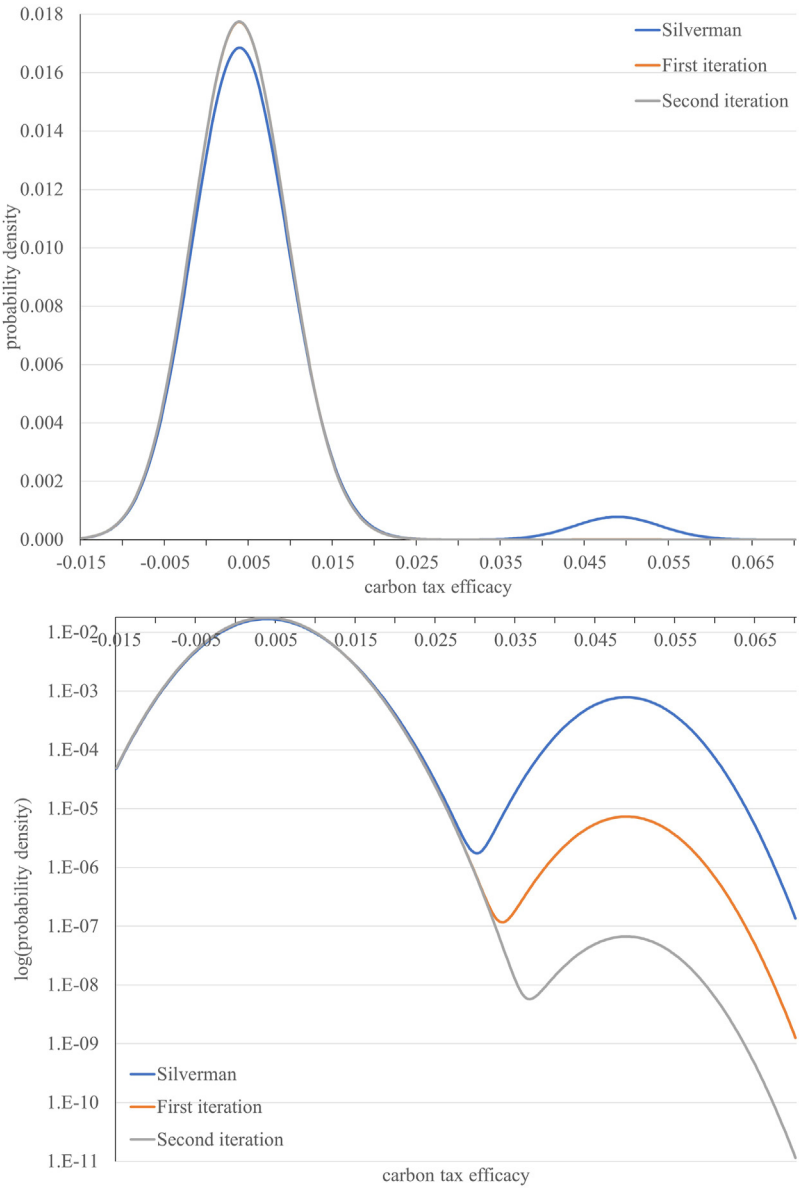

Fig. 4 An adaptive kernel.

In the zeroth iteration, the kernel density of carbon tax efficacy is set using the Silverman rule for the bandwidth. In the first (second) iteration, the bandwidth is proportional to the square root of the kernel density in the zeroth (first) iteration. The top panel shows the density, the bottom panel the log-density.

Lemuria. On the other hand, income data rarely include the super-rich. Assuming that the highest income in the *population* is close to the highest income in the *sample* is wrong. An Epanechnikov kernel is inappropriate in this case, a Normal kernel less so.

Fig. 5 illustrates this using the carbon tax efficacy data. The kernel density based on Epanechnikov kernels is sharper than the one using Normal kernels.

Transformations and asymmetric kernels

Income data is usually transformed with the natural logarithm. The Epanechnikov kernel then assumes that the maximum income in the population is a multiple of the maximum income in the sample, while the Normal kernel assumes a heavy tail.[3] As with the choice of bandwidth and kernel function, you should choose the transformation that is most suited to your data and your study.

[3]The logarithmic transformation assumes that income cannot be negative, which is true for wage earners. Incomes can be negative for self-employment and coerced labour.

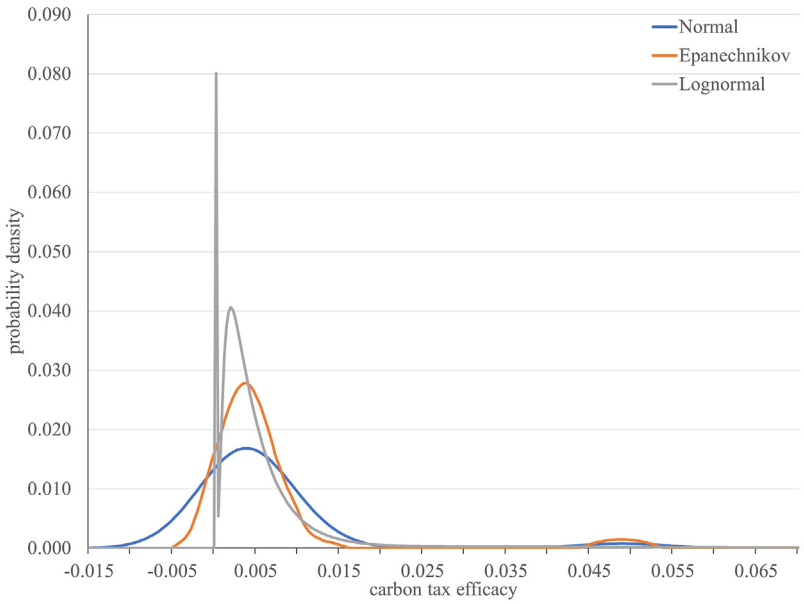

Fig. 5 Kernel functions.
The kernel *density* for three alternative kernel *functions*: Normal, Epanechnikov, and Lognormal.

However, a Normal kernel applied to log-transformed data is equivalent to a Lognormal kernel applied to the original data. I stated above that a typical kernel is symmetric. The Lognormal distribution is not. This reveals that the symmetry assumption is one of convenience. There is no mathematical necessity.

One convenient aspect of a symmetric kernel is that you do not need to worry about interpreting the observed values. The kernel function gives a spread around each observation; the observation is the mean, mode, and median of the kernel function.

If you use an asymmetric kernel, you need to decide whether your observations are best thought of means, modes, or medians. The equivalence between a Normal kernel for log-transformed data and a Lognormal kernel for the original data only holds if the observations are interpreted as the *medians* of their kernel functions. In addition, the bandwidth would need to vary with the observations. Data transformations are rarely innocent.

Fig. 5 shows alternative kernel densities of the carbon tax efficacy. Normal and Epanechnikov kernels are discussed above. Fig. 5 also includes a Lognormal kernel, interpreting the observations as the median. A logarithmic transformation is appropriate in this case. A negative efficacy implies a perverse increase in emissions in response to carbon pricing. The Lognormal kernel function leads to a trimodal kernel density, as two observations are close to zero. These observations are blurred by a symmetric kernel function but enhanced by an asymmetric one.

As with the choice of bandwidth, the kernel function requires careful consideration.

Higher-order kernels

We abandoned the convention that a kernel function must be symmetric in the previous section—and so the convention that it has a zero mean: A kernel function should be centered at zero, but this is an ambiguous statement with an asymmetric distribution.

We here ditch another convention: A kernel function must be positive. A higher-order kernel function integrates to one, but it can assume negative values.

Higher-order kernels are also known as bias-reducing kernels. That sounds nice, but they are dangerous. They should not be used by novices. The only reason I mention them here is because some software packages offer the kernel *order* as a parameter, and an unsuspecting user may mistake a higher order for a better choice.

A zeroth-order kernel is always positive. A second-order kernel is positive in the center and negative in the tails. A fourth-order kernel is positive in the center and the tails but negative in between. And so on.[4]

In order to understand higher-order kernel functions, recall the intuition behind a kernel density. We want to know the probability density of y, having observed $y_1, y_2, \ldots y_n$. The probability density of y is the weighted sum of a spread around all observations, where the weights relate to the inverse distance between y and $y_1, y_2, \ldots y_n$, so that weights are smaller when the distance is larger. Essentially, we make y similar to observations that are close to y and largely ignore observations that are far away.

[4]Odd-order kernels would oscillate more in one-half of the distribution than in the other. For example, a first-order kernel would be negative in either the left or right tail, but not in both. I have thankfully not come across any of these.

With a higher-order kernel function, some weights are *negative*. With a second-order kernel, we want y to be similar to observations that are close but *dissimilar* to observations that are far away.

Let that sink in. We do not discard distant observations as irrelevant. Instead, we steer clear of them.

If you understand your data and your estimator very well, then you can use higher-order kernel functions to reduce bias. Do not use higher-order kernels simply because it is a parameter you can choose in a piece of software.

Kernel decomposition

If the only actual requirement is that a kernel function integrates to one, then any density function can serve as a kernel function—including a kernel density.

This allows us to decompose a kernel density of a population into its components for subpopulations. For

$$f(y) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{y-y_i}{h}\right) =$$

$$\frac{1}{Nh}\left(\sum_{i=1}^{N_1} K\left(\frac{y-y_i}{h}\right) + \sum_{i=N_1+1}^{N} K\left(\frac{y-y_i}{h}\right)\right) =$$

$$\frac{N_1}{N}\frac{1}{hN_1}\sum_{i=1}^{N_1} K\left(\frac{y-y_i}{h}\right) + \frac{N-N_1}{N}\frac{1}{h(N-N_1)}\sum_{i=N_1+1}^{N} K\left(\frac{y-y_i}{h}\right) \tag{5}$$

More generally,

$$f(y) = \sum_j \frac{N_j}{N}\frac{1}{hN_j}\sum_{i=1}^{N_j} K\left(\frac{y-y_i}{h}\right) \tag{6}$$

for $\sum_j N_j = N$.

Of course, we do not need to assume the same bandwidth. If you are confident that your subpopulations are systematically different—e.g., measurement error is different—then you can take this a step further and use

$$f(y) := \sum_j \frac{N_j}{N}\frac{1}{h_j N_j}\sum_{i=1}^{N_j} K\left(\frac{y-y_i}{h_j}\right) \tag{7}$$

In this case, the weighted sum of the kernel densities of the subpopulations is no longer equal to the kernel density of the whole population—and this may be exactly what you wanted.

Kernel regression, discussed below, is a handy tool to plot the relationship between two continuous variables. It does not work for the relationship between a continuous and a discrete variable. Kernel decomposition does.

Fig. 6 illustrates this, using data from a survey on patience (Falk et al., 2018, 2023). Falk's index combines the answers to a qualitative and quantitative question on patience for almost 80,000 people in 76 countries. The kernel density shows a pronounced mode at -0.6, a secondary mode at -1.3, and a flat right tail of more patient respondents. This highlights the descriptive power of kernel densities: No parametric probability distribution looks anything like this.

Fig. 6 decomposes the kernel density into the responses by men and women. Differences are subtle but visible. This is confirmed in **Table 1**, which has the probability mass of the five quintiles of the kernel density of the whole sample and its two subsamples. Men are less impatient: The male probability mass in the bottom quintile is two percent lower than the female mass; it is two percent higher in the top quintile. This difference is statistically significant, mostly because there are so many observations, according to Pearson's test for the equality of proportions: $\chi^2_4 = 114.92$, $p = 0.0000$.

Kernel regression

Kernel regression is also known as non-parametric regression. This is a misnomer. As we saw above, the construction of a kernel density relies on a parameter, the bandwidth, and a functional form, the kernel function. In so-called non-parametric methods, the parameters are one level deeper. It is better to call this kernel regression, also to emphasize its intimate ties to kernel density estimation.

Kernel regression has two key advantages. First, we do not need to assume that the error terms follow a Normal or indeed any particular distribution. Second, we do not need to assume a functional form for the relationship between the dependent variable y and the explanatory variables X. In this sense, kernel regression is truly non-parametric.

The ordinary least squares regression is typically written as

$$y_i = X_i\beta + \varepsilon_i \; \varepsilon_i \sim N(0, \sigma^2) \tag{8}$$

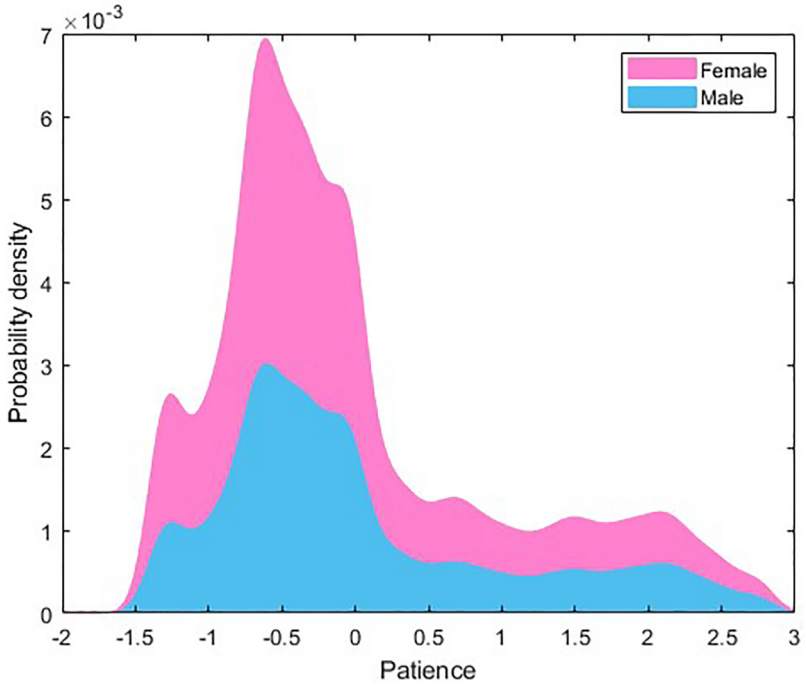

Fig. 6 Patience by gender.
The graph shows the kernel density of Falk's index of patience, decomposed by self-reported gender. Note that the survey only records binary gender.

This is the same as

$$y_i \sim N\left(X_i\beta, \sigma^2\right) \tag{9}$$

This implies

$$\mathrm{E}[y_i|X_i] = \int y f_{Y|X}(y|X_i)\mathrm{d}y \tag{10}$$

That is, our forecast for y_i is the expectation of the conditional distribution of y given X.

Kernel regression uses this. Recall that

$$f_{Y|X}(y|x) = \frac{f_{Y,X}(y,x)}{f_X(x)} \tag{11}$$

So, for a simple regression, we need to construct the bivariate kernel distribution of the dependent and explanatory variables, integrate out to the marginal kernel distribution of the explanatory variable, divide, and take the expectation of the resulting conditional distribution.

This sounds intimidating but it is not. Kernel distributions are constructed numerically on discrete grids for y and x, so we are really just taking column and row sums in a matrix. This is readily coded.

Nadaraya-Watson regression

There is no need to code yourself, though, as any respectable statistical software package has routines for kernel regression. Many rely on computational-time-saving tricks like the following, due to Nadaraya (1964) and Watson (1964).

We want to know the conditional expectation of y given X:

Table 1 Patience by gender.

Patience	*−0.7700*	*−0.4700*	*−0.1100*	*0.8000*	*3.0000*
All	0.1913	0.1984	0.2026	0.2060	0.2016
Male	0.1835	0.1876	0.2059	0.2098	0.2131
Female	0.2031	0.2010	0.2013	0.2023	0.1923

$$\mathrm{E}[y|x] = \int y f_{Y|X}(y|x)\mathrm{d}y = \int y \frac{f_{Y,X}(y,x)}{f_X(x)}\mathrm{d}y$$

$$= \int y \frac{\frac{1}{Nh_yh_x}\sum_{i=1}^{N} K_y\left(\frac{y-y_i}{h_y}\right)K_x\left(\frac{x-x_i}{x_y}\right)}{\frac{1}{Nh_x}\sum_{i=1}^{N} K_x\left(\frac{x-x_i}{x_y}\right)}\mathrm{d}y$$

$$= \frac{\sum_{i=1}^{N}\left[K_x\left(\frac{x-x_i}{x_y}\right)\int yK_y\left(\frac{y-y_i}{h_y}\right)\mathrm{d}y\right]}{h_y\sum_{i=1}^{N} K_x\left(\frac{x-x_i}{x_y}\right)} \tag{12}$$

Define $z_i = \frac{y-y_i}{h_y}$. $\frac{\mathrm{d}y}{\mathrm{d}z_i} = h_y$. By definition, $\int_z K_y(z_i)\mathrm{d}z_i = 1$ and $\int_z z_i K_y(z_i)\mathrm{d}z_i = 1$. Then

$$\frac{1}{h_y}\int_y yK_y\left(\frac{y-y_i}{h_y}\right)\mathrm{d}y = \frac{1}{h_y}\int_{z_i}(h_yz_i+y)K_y(z_i)h_y\mathrm{d}z_i = y_i$$

so that

$$\mathrm{E}[y|x] = \frac{\sum_{i=1}^{N}\left[K_x\left(\frac{x-x_i}{x_y}\right)y_i\right]}{\sum_{i=1}^{N} K_x\left(\frac{x-x_i}{x_y}\right)} = \sum_{i=1}^{N} W_i(x)y_i \tag{13}$$

where

$$W_i(x) = \frac{K_x\left(\frac{x-x_i}{x_y}\right)}{\sum_{i=1}^{N} K_x\left(\frac{x-x_i}{x_y}\right)}$$

That is, the expectation of y given x is the weighted sum of all observed y_i where the weights are the relative distance of x to x_i.

Let us assume that y_i is a characteristic of person i. The prediction y is then the weighted average of all observations y_i, where the weights are determined by the similarity of the observables x_i to the target x. Suppose you want to know the height of a 9-year-old but only know the heights of kids aged 5, 7, 11 and 13. You take the weighted average, with high weights for the 7 and 11 years of age, and low weights for 5 and 13 years. Kernel regression and propensity score matching are thus conceptually alike, although details differ.

Priestley-Chao regression

Priestley and Chao (1972) take a different route. They consider that a linear interpolation between y_1 and y_2 would take the form

$$y = f(x) = y_1 + \frac{x_2 - x}{x_2 - x_1}(y_2 - y_1) = \frac{(x - x_1)y_1 + (x_2 - x)y_2}{x_2 - x_1} \quad x_1 \le x < x_2$$

This is a weighted sum of two observations.

Note that we interpolate using a covariate. For instance, you observe the heights of a 7-year-old and an 11-year-old and want to estimate the height of a 9-year-old—halfway between the two observed heights. Priestley-Chao interpolation takes the weighted average of two *shifted* observations. Nadaraya-Watson takes the weighted average of the actual observations, rather than shifted ones.

We typically have many more than two observations. It would be strange to restrict the basis for interpolation to the cases just above and just below. The Priestley-Chao estimator uses all information:

$$y(x) = \frac{1}{h}\sum_{i=1}^{N-1} K\left(\frac{x - x_i}{h}\right)(x_{i+1} - x_i)y_i \tag{14}$$

That is, the prediction for y given x is the weighted sum of all observed y_i shifted by the distance between x and x_i, with larger weights for observations that are closer and so shifted less.

Smoothing

The relationship between x and y assumes no particular functional form in kernel regression. Brook Taylor showed that any function can be approximated as

$$y = \gamma_0 + \frac{\partial f}{\partial x}(x-\gamma_0) + \frac{\partial^2 f}{2\partial x^2}(x-\gamma_0)^2 + \frac{\partial^3 f}{6\partial x^3}(x-\gamma_0)^3 + \ldots$$

$$= a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots \tag{15}$$

The Nadaraya-Watson estimator can be interpreted as a *zero*-order Taylor expansion: It is the weighted average of all observations. The Priestley-Chao estimator is the weighted average of linear interpolations between pairs of observations, which is a *first*-order Taylor expansion.

Local polynomial regression builds on this insight. You can use quadratic, cubic, or quartic interpolation and take the kernel-weighted average of triplets, quadruplets, or quintuplets of observations. This is known as *Locally Weighted Scatterplot Smoothing* or LOWESS (Cleveland and Devlin, 1988).

In software applications of LOWESS, you need to set four parameters: The kernel function, its bandwidth, the order p, and the locality L, the size of the subset of observations included. For $p = 0$ and $L = N$, LOWESS is Nadaraya-Watson. For $p = 1$ and $L = N$, LOWESS is Priestley-Chao. For $p > 2$, you may want to set $L < N$ to reduce computational time. This has little or no effect for kernel functions with a finite support or for narrow bandwidths.

Locally Estimated Scatterplot Smoothing or LOESS generalizes further. It uses two or more explanatory variables. Instead of conditioning y on x as in LOWESS, LOESS conditions y on x and z and perhaps on w and v as well (Cleveland and Devlin, 1988).

If you impose that the estimator must be the Best Linear Unbiased Prediction, LOESS is known as *Kriging* (Robinson, 1991). Like the Silverman rule, Kriging feels a bit like cheating in that you force a non-parametric method to be as close as possible to a parametric one. Kriging is particularly popular for spatial applications.

Quantile regression

Above, we focus on the conditional expectation of y given x. To get there, we found the entire distribution of y conditional on x. We can therefore just as readily derive the conditional mode of y on x, the conditional median y and every other percentile, and the conditional variance and every other higher moment. There are two things to consider here. First, confidence in the central tendency of a distribution requires fewer observations than confidence in its higher moments or tails. Second, the Silverman rule for the choice of bandwidth ensures that the *expectation* of the kernel distribution is well-behaved, but does not guarantee this for other characteristics. There is a theoretical literature on the optimal bandwidths for different statistics that I do not discuss here as there are many technicalities but few insights.

Fig. 7 returns to Falk's index of patience to illustrate kernel quantile regression. The conditional expectation of patience given age is almost flat between 15 and 65 years of age, and then steadily declines as older people are more impatient. The median shows the same pattern at a lower level. Recall that Fig. 6 revealed a right-skewed distribution. The interquartile range shows a different pattern. The 25th percentile shows a step change around 35 years. The 75th percentile displays an increase in patience starting in the early 30s—family formation?—and a sharp decrease starting in the early 60s—retirement?

Semi-parametric regression

The derivation in section Inference was for simple regression. Multiple regression is more appropriate in almost every situation. The principle is the same: Construct the multivariate kernel distribution of y and X, and the multivariate kernel distribution of X. Derive the conditional kernel distribution of y on all its explanatory variables X, and integrate over that to find the conditional distribution.

The problem with this strategy is that the multivariate kernel distribution can have many dimensions. Tens of explanatory variables are not uncommon. The computational burden is large. Suppose we want to know the relationship between y and x, controlling for Z. This is

$$f_{Y|X}(y|x) = \int \frac{f_{Y,X,Z}(y,x,Z)}{f_X(x)} \mathrm{d}Z \tag{16}$$

That is, we first compute the distribution of y and Z conditional on x, and then the marginal distribution of y (still conditional on x). If there are N control variables, you need to integrate in an N-dimensional space. This is expensive.

People therefore use semi-parametric regression, which uses kernel or non-parametric regression for the variable of interest and parametric regression for the control variables.

The general form of semi-parametric regression is

$$y = X\beta + m(z) + \varepsilon \tag{17}$$

where z is the variable of interest and X are the control variables.

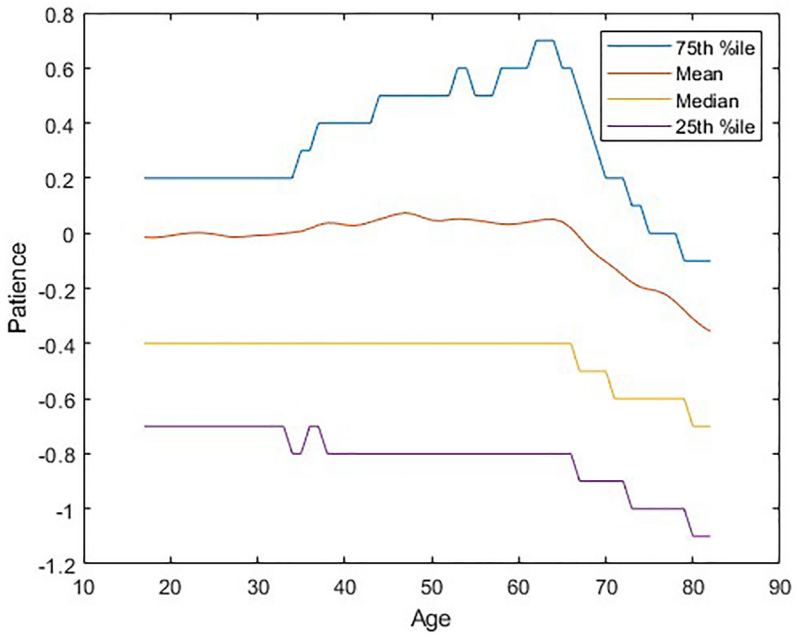

Fig. 7 Patience by age.
The graphs show selected characteristics—mean, median, interquartile range—of the kernel distribution of Falk's index of patience, conditional on the age of survey respondents.

There are two ways to estimate this. The first is due to Yatchew (1998). First regress y on X to find $\widehat{\beta} = (X'X)^{-1}Xy$ and then use kernel regression for $y - X\widehat{\beta}$ on z. This estimator is consistent if $m(z)$ is smooth, but there is a loss of efficiency and small-sample bias if z and X are correlated.

Earlier, Robinson (1988) proposed a three-stage estimator. First, regress non-parametrically

$$y = m'(z) + \varepsilon'$$

$$x_i = n_i(z) + \varepsilon'' \tag{18}$$

Then, regress parameterically

$$y - \widehat{m(z)} = (X - \widehat{n(z)})\beta + \varepsilon''' \tag{19}$$

Finally, regress non-parametrically

$$y - X\widehat{\beta} = m(z) + \varepsilon \tag{20}$$

This procedure takes away the multicollinearity bias between z and X in small samples but, like all multi-step estimators, it is not efficient.

Conclusion

I discuss kernel density estimation and kernel regression. These methods are often referred to as "non-parametric" but this is a bit of a misnomer for kernel densities as the user needs to make two key choices: bandwidth and kernel function. A kernel density is the sum of smudged observations. The bandwidth determines the extent of the smudge, the kernel its shape and support. Although there is some theory to guide these choices, the nature of the data at hand should be just as important a consideration.

Kernel regression is more deserving of the non-parametric moniker. A kernel density still needs to be specified, but the data and the data alone determine the relationship between the dependent and explanatory variables. Kernel regression readily generalizes from the mean to the quantiles. Unfortunately, multiple kernel regression is computationally intensive, so it may be better to use parametric regression for the control variables and kernel regression for the variable of interest.

The table shows the probability mass in the quintiles of the kernel density of Falk's index of patience, decomposed by self-reported gender. The top row shows the upper bound of each quintile. The bottom rows show the probability mass for the entire sample and the two subsamples. Note that the survey only records binary gender.

References

Abramson, I.S., 1982. On bandwidth variation in kernel estimates-a square root law. Ann. Stat. 10 (4), 1217—1223. http://www.jstor.org/stable/2240724.
Cleveland, W.S., Devlin, S.J., 1988. Locally weighted regression: an approach to regression analysis by local fitting. J. Am. Stat. Assoc. 83 (403), 596—610.
Epanechnikov, V.A., 1969. Non-parametric estimation of a multivariate probability density. Theor. Probab. Appl. 14 (1), 153—158. https://doi.org/10.1137/1114019.
Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., Sunde, U., 2018. Global evidence on economic preferences. Q. J. Econ. 133 (4), 1645—1692.
Falk, A., Becker, A., Dohmen, T., Huffman, D., Sunde, U., 2023. The preference survey module: A validated instrument for measuring risk, time, and social preferences. Manag. Sci. 69 (4), 1935—1950.
Nadaraya, E.A., 1964. On estimating regression. Theor. Probab. Appl. 9 (1), 141—142. https://doi.org/10.1137/1109020.
Priestley, M.B., Chao, M.T., 1972. Non-parametric function fitting. J. Roy. Stat. Soc. B 34 (3), 385—392. https://doi.org/10.1111/j.2517-6161.1972.tb00916.x.
Robinson, P.M., 1988. Root-n-consistent semiparametric regression. Econometrica 56 (4), 931—954. http://www.jstor.org/stable/1912705.
Robinson, G., 1991. That blup is a good thing: the estimation of random effects. Stat. Sci. 6 (1), 15—32.
Rubin, D.B., 1981. The bayesian bootstrap. Ann. Stat. 9 (1), 130—134. http://www.jstor.org/stable/2240875.
Silverman, B., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall, London.
Takezawa, K., 2005. Introduction to Nonparametric Regression. John Wiley and Sons, Hoboken.
Tol, R.S.J., 2023. The fiscal implications of stringent climate policy. Econ. Anal. Pol. 80 (C), 495—504.
Tsybakov, A.B., 2009. Introduction to Nonparametric Estimation. Springer Nature, Berlin.
Watson, G.S., 1964. Smooth regression analysis. Sankhya: The Indian J. Stat., Series A 26 (4), 359—372. http://www.jstor.org/stable/25049340.
Yatchew, A., 1998. Nonparametric regression techniques in economics. J. Econ. Lit. 36 (2), 669—721. http://www.jstor.org/stable/2565120.