6
Nonparametric regression

Frequently in econometric analysis of accounting data, one is concerned with departures from standard parametric model probability assignments. Semi- and nonparametric methods provide an alternative means to characterize data and assess parametric model robustness or logical consistency. Here, we focus on regression. That is, we examine the conditional relation between Y and X. The most flexible fit of this conditional relation is nonparametric regression where flexible fit refers to the degree of distributional or structural form restrictions imposed on the data in estimating the relationship.

6.1 Nonparametric (kernel) regression

Nonparametric regression is motivated by at least the following four objectives: (1) it provides a versatile method for exploring a general relation between variables, (2) it give predictions without reference to a fixed parametric model, (3) it provides a tool for identifying spurious observations, and (4) it provides a method for 'fixing' missing values or interpolating between regressor values (see Hardle [1990, p 6-7]).

A nonparametric (kernel) regression can be represented as follows (Hardle [1990]).

$$E\left[Y|X\right] = m\left(X\right)$$

where $m\left(X\right) = \dfrac{n^{-1}h^{-d}\sum\limits_{i=1}^{n} K\left(\frac{X-x_i}{h}\right)y_i}{n^{-1}h^{-d}\sum\limits_{i=1}^{n} K\left(\frac{X-x_i}{h}\right)}$, y_i $\left(x_i\right)$ is the ith observation for Y $\left(X\right)$, n is the number of observations, d is the dimension (number of regressors) of X,

$K(\cdot)$ is any well-defined (multivariate) kernel, and h is the smoothing parameter or bandwidth (see *GCV* below for bandwidth estimation). Notice as is the case with linear regression each predictor is constructed by regressor-based weights of each observed value of the response variable $M(h)Y$ where

$$M(h) = \begin{bmatrix} \frac{K\left(\frac{X_1 - x_1}{h}\right)}{\sum\limits_{i=1}^{n} K\left(\frac{X_i - x_1}{h}\right)} & \frac{K\left(\frac{X_2 - x_1}{h}\right)}{\sum\limits_{i=1}^{n} K\left(\frac{X_i - x_1}{h}\right)} & \cdots & \frac{K\left(\frac{X_n - x_1}{h}\right)}{\sum\limits_{i=1}^{n} K\left(\frac{X_i - x_1}{h}\right)} \\ \frac{K\left(\frac{X_1 - x_2}{h}\right)}{\sum\limits_{i=1}^{n} K\left(\frac{X_i - x_2}{h}\right)} & \frac{K\left(\frac{X_2 - x_2}{h}\right)}{\sum\limits_{i=1}^{n} K\left(\frac{X_i - x_2}{h}\right)} & \cdots & \frac{K\left(\frac{X_n - x_2}{h}\right)}{\sum\limits_{i=1}^{n} K\left(\frac{X_i - x_2}{h}\right)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{K\left(\frac{X_1 - x_n}{h}\right)}{\sum\limits_{i=1}^{n} K\left(\frac{X_i - x_n}{h}\right)} & \frac{K\left(\frac{X_2 - x_n}{h}\right)}{\sum\limits_{i=1}^{n} K\left(\frac{X_i - x_n}{h}\right)} & \cdots & \frac{K\left(\frac{X_n - x_n}{h}\right)}{\sum\limits_{i=1}^{n} K\left(\frac{X_i - x_n}{h}\right)} \end{bmatrix}$$

To fix ideas, compare this with liner regression. For linear regression, the predictions are $\widehat{Y} = P_X Y$, where

$$P_X = X\left(X^T X\right)^{-1} X^T$$

the projection matrix (into the columns of X), again a linear combination (based on the regressors) of the response variable.

A multivariate kernel is constructed, row by row, by computing the product of marginal densities for each variable in the matrix of regressors X.[1] That is, $h^{-d} K\left(\frac{X - x_i}{h}\right) = \prod\limits_{j=1}^{d} h^{-1} K\left(\frac{x_j - x_{ji}}{h}\right)$, where x_j is the jth column vector in the regressors matrix. Typically, we employ leave-one-out kernels. That is, the current observation is excluded in the kernel construction to avoid overfitting — the principal diagonal in $M(h)$ is zeroes. Since nonparametric regression simply exploits the explanatory variables to devise a weighting scheme for Y, assigning no weight to the current observation of Y is an intuitively appealing means of avoiding overfitting.

Nonparametric (kernel) regression is the most flexible model that we employ and forms the basis for many other kernel density estimators. While nonparametric regression models provide a very flexible fit of the relation between Y and X, this does not come at zero cost. In particular, it is more difficult to succinctly describe this relation, especially when X is a high dimension matrix. Also, nonparametric regressions typically do not achieve parametric rates of convergence (i.e., they converge more slowly than square root n).[2] Next, we turn to models that retain

[1] As we typically estimate one bandwidth for all regressors, the variables are first scaled by their estimated standard deviation.

[2] It can be shown that optimal rates of convergence for nonparametric models are $n - r$, $0 < r < 1/2$. More specifically, $r = (\rho + \beta - k)/(2[\rho + \beta] - d)$, where ρ is the number of times the smoothing

some of the flexibility of nonparametric regression but enhance interpretability (i.e., semiparametric models).

6.2 Semiparametric regression models

6.2.1 *Partial linear regression*

Frequently, we are concerned about the relation between Y and X but troubled that the analysis is plagued by omitted, correlated variables. One difficulty is that we do not know the functional form of the relation between our variables of interest and these other control variables. That is, we envision a *DGP* where

$$E\left[Y \mid X, Z\right] = X\beta + \theta\left(Z\right)$$

Provided that we can observe these control variables, Robinson [1988] suggests a two-stage approach analogous to *FWL* (see chapter 3) which is called partial linear regression. Partial linear regression models nonparametrically fit the relation between the dependent variable, Y, and the control variables, Z, and also the experimental regressors of interest, X, and the control variables Z. The residuals from each nonparametric regression are retained, $e_Y = Y - E\left[Y|Z\right]$ and $e_X = X - E\left[X|Z\right]$, in standard double residual regression fashion.

Next, we simply employ no-intercept *OLS* regression of the dependent variable residuals on the regressor residuals, $e_Y = e_X\beta$. The parameter estimator for β fully captures the influence of the otherwise omitted, control variables and is accordingly, asymptotically consistent. Of course, we now have parameters to succinctly describe the relation between Y and X conditional on Z. Robinson demonstrates that this estimator converges at the parametric (square-root n) rate.

6.2.2 *Single-index regression*

The partial linear model discussed above imposes distributional restrictions on the relation between Y and X in the second stage. One (semiparametric) approach for relaxing this restriction and retaining ease of interpretability is single-index regression. Single-index regression follows from the idea that the average derivative of a general function with respect to the regressor is proportional to the parameters of the index. Suppose the DGP is

$$E\left[Y|X\right] = G\left(X\beta\right)$$

then define $\delta = \partial E\left[Y|X\right]/\partial X = dG/d\left(X\beta\right)\beta = \gamma\beta$. Thus, the derivative with respect to X is proportional to β for all X, and likewise the average derivative $E\left[dG/d\left(X\beta\right)\right]\beta = \gamma\beta$, for $\gamma \neq 0$, is proportional to β.

function is differentiable, k is the order of the derivative of the particular estimate of interest ($k \leq \rho$), β is the characteristic or exponent for the smoothness class, and d is the order of the regressors (Hardle [1990, p. 93]).

Our applications employ the density-weighted average derivative single-index model of Powell, Stock, and Stoker [1989].[3] That is,

$$\hat{\delta} = -2n^{-1} \sum_{i=1}^{n} \frac{\partial \hat{f}_i \left(X_i \right)}{\partial X} Y_i$$

exploiting the U statistic structure (see Hoeffding [1948])

$$= -2 \left[n \left(n-1 \right) \right]^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} h^{-(d+1)} K' \left(\frac{X_i - X_j}{h} \right) \left(Y_i - Y_j \right)$$

For a Gaussian kernel, K, notice that $K\prime \left(u \right) = -uK \left(u \right)$. Thus,

$$\hat{\delta} = 2 \left[n \left(n-1 \right) \right]^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} h^{-(d+1)} K \left(\frac{X_i - X_j}{h} \right) \left(\frac{X_i - X_j}{h} \right) \left(Y_i - Y_j \right)$$

where $K \left(u \right) = \left(2\pi \right)^{-1/2} exp \left\{ -\frac{u^2}{2} \right\}$. The asymptotic covariance matrix for the parameters $\Sigma_{\hat{\delta}}$is estimated as

$$\widehat{\Sigma}_{\hat{\delta}} = 4n^{-1} \sum_{i=1}^{n} \hat{r} \left(Z_i \right) \hat{r} \left(Z_i \right)^T - 4\widehat{\hat{\delta}\hat{\delta}}^T$$

where

$$\hat{r} \left(Z_i \right) = \left(-n-1 \right)^{-1} \sum_{\substack{j=1 \\ i \neq j}}^{n} h^{-(d+1)} K \left(\frac{X_i - X_j}{h} \right) \left(\frac{X_i - X_j}{h} \right) \left(Y_i - Y_j \right)$$

The above estimator is proportional to the index parameters. Powell, et al also proposed a properly-scaled instrumental variable version of the density-weighted average derivative. We refer to this estimator as $\hat{d} = \hat{\delta}_X^{-1} \hat{\delta}$, where

$$\begin{aligned} \hat{\delta}_X &= -2n^{-1} \sum_{i=1}^{n} \frac{\partial \hat{f}_i \left(X_i \right)}{\partial X} X_i^T \\ &= \frac{2 \sum\limits_{i=1}^{n-1} \sum\limits_{j=i+1}^{n} h^{-(d+1)} K \left(\frac{X_i - X_j}{h} \right) \left(\frac{X_i - X_j}{h} \right) \left(X_i - X_j \right)^T}{n \left(n-1 \right)} \end{aligned}$$

[3] Powell et al's description of the asymptotic properties of their average derivative estimator exploits a 'leave-one-out' approach, as discussed for nonparametric regression above. This estimator also achieves the parametric (square-root n) rate of convergence.

rescales $\hat{\delta}$. The asymptotic covariance estimator for the parameters $\Sigma_{\hat{d}}$ is estimated as $\hat{\Sigma}_{\hat{d}} = 4n^{-1} \sum\limits_{i=1}^{n} \hat{r}_d \left(Z_i\right) \hat{r}_d \left(Z_i\right)^T$, where

$$\hat{r}_d \left(Z_i\right) = \hat{\delta}_x^{-1} \frac{\sum\limits_{\substack{j=1 \\ i \neq j}}^{n} h^{-(d+1)} K \left(\frac{X_i - X_j}{h}\right) \left(\frac{X_i - X_j}{h}\right) \left(\widehat{U}_i - \widehat{U}_j\right)}{-n-1}$$

$$\widehat{U}_i = Y_i - X_i \hat{d}$$

The optimal bandwidth is estimated similarly to that described for nonparametric regression. First, $\hat{d}$ (and its covariance matrix) is estimated (for various bandwidths). Then, the bandwidth that produces minimum mean squared error is identified from the leave-one out nonparametric regression of Y on the index $X\widehat{d}$ (the analog to regressing Y on X in fully nonparametric regression). This yields a readily interpretable, flexibly fit set of index parameters, the counterpart to the slope parameter in *OLS* (linear) regression.

6.2.3 Partial index regression models

Now, we put together the last two sets of ideas. That is, nonparametric estimates for potentially omitted, correlated (control) variables as in the partial linear model are combined with single index model parameter estimates for the experimental regressors. That is, we envision a *DGP* where

$$E\left[Y \mid X, Z\right] = G\left(X\beta\right) + \theta\left(Z\right)$$

Following Stoker [1991], these are called partial index models. As with partial linear models, the relation between Y and Z (the control variables) and the relation between X and Z are estimated via nonparametric regression. As before, separate bandwidths are employed for the regression of Y on Z and X on Z. Again, residuals are computed, e_Y and e_X. Now, single index regression of e_Y on e_X completes the partial index regression. Notice, that a third round of bandwidth selection is involved in the second stage.

6.3 Specification testing against a general nonparametric benchmark

Specification or logical consistency testing lies at the heart of econometric analysis. Borrowing from conditional moment tests (Ruud [1984], Newey [1985], Pagan and Vella [1989]) and the U statistic structure employed by Powell et al, Zheng [1996] proposed a specification test of any parametric model $f\left(X, \theta\right)$ against a general nonparametric benchmark $g\left(X\right)$.

Let $\varepsilon_i \equiv Y_i - f(X_i, \theta)$ and $p(\bullet)$ denote the density function of X_i. The null hypothesis is that the parametric model is correct (adequate for summarizing the data)

$$H_0 : \Pr E[Y_i|X_i] = f(X_i, \theta_0) = 1 \text{ for some } \theta_0 \in \Theta$$

where $\theta_0 = \arg\min_{\theta\in\Theta} E[Y_i - f(X_i, \theta_0)]^2$. The alternative is the null is false, but there is no specific alternative model

$$H_0 : \Pr E[Y_i|X_i] = f(X_i, \theta) < 1 \text{ for all } \theta \in \Theta$$

The idea is under the null, $E[\varepsilon_i|X_i] = 0$. Therefore, we have

$$E[\varepsilon_i E[\varepsilon_i|X_i] p(X_i)] = 0$$

while under the alternative we have

$$E[\varepsilon_i E[\varepsilon_i|X_i] p(X_i)] = E\left[\{E[\varepsilon_i|X_i]\}^2 p(X_i)\right]$$

since $E[\varepsilon_i|X_i] = g(X_i) - f(X_i, \theta)$

$$E[\varepsilon_i E[\varepsilon_i|X_i] p(X_i)] = E\left[\{g(X_i) - f(X_i, \theta)\}^2 p(X_i)\right] > 0$$

The sample analog of $E[\varepsilon_i E[\varepsilon_i|X_i] p(X_i)]$ is used to form a test statistic. In particular, kernel estimators of the components are employed. A kernel estimator of the density function p is

$$\hat{p}(x_i) = (-n-1)^{-1} \sum_{\substack{j=1 \\ i\neq j}}^{n} h^{-d} K\left(\frac{X_i - X_j}{h}\right)$$

and a kernel estimator of the regression function $E[\varepsilon_i|X_i]$ is

$$E[\varepsilon_i|X_i] = (-n-1)^{-1} \sum_{\substack{j=1 \\ i\neq j}}^{n} h^{-d} \frac{K\left(\frac{X_i - X_j}{h}\right)\varepsilon_i}{\hat{p}(X_i)}$$

The sample analog to $E[\varepsilon_i E[\varepsilon_i|X_i] p(X_i)]$ is completed by replacing ε_i with $e_i \equiv Y_i - f\left(X_i, \hat{\theta}\right)$ and we have

$$V_n \equiv (-n-1)^{-1} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i\neq j}}^{n} h^{-d} K\left(\frac{X_i - X_j}{h}\right) e_i e_j$$

Under the null, Zheng shows that the statistic $nh^{d/2}V_n$ is consistent asymptotic normal (*CAN*; see appendix) with mean zero and variance Σ. Also, the variance can be consistently estimated by

$$\hat{\Sigma} = 2(n(-n-1))^{-1} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i\neq j}}^{n} h^{-d} K^2\left(\frac{X_i - X_j}{h}\right) e_i^2 e_j^2$$

Consequently, a standardized test statistic is

$$T_n \equiv \sqrt{\frac{n-1}{n}} \frac{nh^{d/2} V_n}{\sqrt{\hat{\Sigma}}}$$

$$= \frac{\sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} h^{-d} K\left(\frac{X_i - X_j}{h}\right) e_i e_j}{\left\{ 2 \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} h^{-d} K^2 \left(\frac{X_i - X_j}{h}\right) e_i^2 e_j^2 \right\}^{1/2}}$$

Since V_n is *CAN* under the null, the standardized test statistic converges in distribution to a standard normal, $Tn \xrightarrow{d} N(0,1)$ (see the appendix for discussion of convergence in distribution).

6.4 Locally linear regression

Another local method, *locally linear regression*, produces smaller bias (especially at the boundaries of X) and no greater variance than regular kernel regression.[4] Hence, it produces smaller *MSE*.

Regular kernel regression solves

$$\min_{g} \sum_{i=1}^{n} (y_i - g)^2 h^{-d} K\left(\frac{X - x_i}{h}\right)$$

while locally linear regression solves

$$\min_{g,\beta} \sum_{i=1}^{n} \left(y_i - g - (X - x_i)^T \beta\right)^2 h^{-d} K\left(\frac{X - x_i}{h}\right)$$

Effectively, kernel regression is a constrained version of locally linear regression with $\beta = 0$. Both are regressor-based weighted averages of Y.

Newey [2007] shows the asymptotic *MSE* for locally linear regression is

$$MSE_{LLR} = \frac{1}{nh} \nu_0 \frac{\sigma^2(X)}{f_0(X)} + \frac{h^4}{4} g_0^{''}(X) \mu_2^2$$

while for kernel regression we have

$$MSE_{KR} = \frac{1}{nh} \nu_0 \frac{\sigma^2(X)}{f_0(X)} + \frac{h^4}{4} \left[g_0^{''}(X) + 2 g_0^{'}(X) \frac{f_0^{'}(X)}{f_0(X)} \right] \mu_2^2$$

[4] This section draws heavily from Newey [2007].

where $f_0(X)$ is the density function for $X = [x_1, \ldots, x_n]^T$ with variance $\sigma^2(X)$, $g_0(X) = E[Y|X]$,

$$u = \frac{X - X_i}{h} \mu_2 = \int K(u) u^2 du, \nu_0 = \int K(u)^2 du$$

and kernel regression bias is

$$bias_{KR} = \left(\frac{1}{2} g_0^{''}(X) + g_0^{'}(X) \frac{f_0^{'}(X)}{f_0(X)} \right) \mu_2 h^2$$

Hence, locally linear regression has smaller bias and smaller *MSE* everywhere.

6.5 Generalized cross-validation (*GCV*)

The bandwidth h is frequently chosen via generalized cross validation (*GCV*) (Craven and Wahba [1979]). *GCV* utilizes principles developed in ridge regression for addressing computational instability problems in a regression context.

$$GCV(h) = \frac{n^{-1} ||y - \hat{m}(h)||^2}{[1 - n^{-1} tr(M(h))]^2}$$

where $\hat{m}(h) = M(h) Y$ is the nonparametric regression of Y on X given bandwidth h, $||\cdot||^2$ is the squared norm or vector inner product, and $tr(\cdot)$ is the trace of the matrix.

Since the properties of this statistic are data specific and convergence at a uniform rate cannot be assured, we evaluate a dense grid of values for h to numerically find the minimum MSE. Optimal bandwidths are determined by trading off a 'good approximation' to the regression function (reduction in bias) and a 'good reduction' of observational noise (reduction in noise). The former (latter) is increasing (decreasing) in the bandwidth (Hardle [1990, p. 29-30, 149]).

For leave-one-out nonparametric regression estimator, *GCV* chooses the bandwidth h that minimizes the mean squared errors

$$\min_h n^{-1} ||Y - \hat{m}_{-t}(h)||^2$$

That is, the penalty function in *GCV* is avoided (as $tr(M_{-t}(h)) = 0$, the denominator is 1) and *GCV* effectively chooses the bandwidth to minimize the model mean square error.

$$M_{-t}\left(h\right)=\begin{bmatrix} \frac{0}{\sum\limits_{i=2}^{n} K\left(\frac{x_i-x_1}{h}\right)} & \frac{K\left(\frac{x_2-x_1}{h}\right)}{\sum\limits_{i=2}^{n} K\left(\frac{x_i-x_1}{h}\right)} & \cdots & \frac{K\left(\frac{x_n-x_1}{h}\right)}{\sum\limits_{i=2}^{n} K\left(\frac{x_i-x_1}{h}\right)} \\ \frac{K\left(\frac{x_1-x_2}{h}\right)}{\sum\limits_{i=1,3}^{n} K\left(\frac{x_i-x_2}{h}\right)} & \frac{0}{\sum\limits_{i=1,3}^{n} K\left(\frac{x_i-x_2}{h}\right)} & \cdots & \frac{K\left(\frac{x_n-x_2}{h}\right)}{\sum\limits_{i=1,3}^{n} K\left(\frac{x_i-x_2}{h}\right)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{K\left(\frac{x_1-x_n}{h}\right)}{\sum\limits_{i=1}^{n-1} K\left(\frac{x_i-x_n}{h}\right)} & \frac{K\left(\frac{x_2-x_n}{h}\right)}{\sum\limits_{i=1}^{n-1} K\left(\frac{x_i-x_n}{h}\right)} & \cdots & \frac{0}{\sum\limits_{i=1}^{n-1} K\left(\frac{x_i-x_n}{h}\right)} \end{bmatrix}$$

As usual, the mean squared error is composed of squared bias and variance.

$$\begin{aligned} MSE\left(\hat{\theta}\right) &= E\left[\left(\hat{\theta}-\theta\right)^2\right] \\ &= E\left[\hat{\theta}^2\right]-2E\left[\hat{\theta}\right]\theta+\theta^2 \\ &= \left\{E\left[\hat{\theta}^2\right]-E\left[\hat{\theta}\right]^2\right\}+\left\{E\left[\hat{\theta}\right]^2-2E\left[\hat{\theta}\right]\theta+\theta^2\right\} \end{aligned}$$

The leading term is the variance of $\hat{\theta}$ and the trailing term is the squared bias.

6.6 Additional reading

There is a burgeoning literature on nonparametric regression and its semiparametric cousins. Hardle [1990] and Stoker [1991] offer eloquent overviews. Newey and Powell [2003] discuss instrumental variable estimation of nonparametric models. Powell et al's average derivative estimator assumes the regressors are continuous. Horowitz and Hardle [1996] proposed a semiparametric model that accommodates some discrete as well as continuous regressors.

When estimating causal effects in a selection setting, the above semiparametric methods are lacking as the intercept is suppressed by nonparametric regression. Andrews and Schafgans [1998] suggested a semiparametric selection model to remedy this deficiency. Variations on these ideas are discussed in later chapters.