International Journal of Scientific and Innovative Mathematical Research(IJSIMR)
Volume 8, Issue 2, 2020, PP 10-21
ISSN No. (Online) 2454-9444
DOI: http://dx.doi.org/10.20431/2347-3142.0802002
www.arcjournals.org

Estimating Regression Models and Density Functions with Classical Parametric and Nonparametric Methods: A Practical Overview

Ming-Lu Wu*

Division of Business Management, United International College Beijing Normal University - Hong Kong Baptist University 2000 Jintong Road, Tangjiawan, Zhuhai, Guangdong 519085, P. R. China

Corresponding Author: *Ming-Lu Wu, *Division of Business Management, United International College Beijing Normal University - Hong Kong Baptist University 2000 Jintong Road, Tangjiawan, Zhuhai, Guangdong 519085, P. R. China*

Abstract: *This paper overviews some classical parametric and nonparametric approaches to estimating regression models concisely for practitioners' easy understanding and applications. Parametric approaches, mainly the least squares method, are first clearly described for cases when the regression model is of known form with unknown parameters. A number of nonparametric approaches, such as the kernel, series, spline and sieve methods, are then systematically presented for cases when the regression model is of unknown form. Nonparametric methods for estimating density functions are also outlined, which is important by itself and also helpful in estimating regression models.*

Keywords: *Density functions; kernel estimates; least squares; linear splines; method of sieves; orthogonal series; regression models.*

1. INTRODUCTION

Suppose that we are interested in a variable y and have collected its data on a sample of n subjects: y_1, …, y_n. Then we can obtain its frequency distribution and descriptive statistics such as central tendency (e.g., mean and median) and variation (e.g., overall range and standard deviation). Such basic information can help us understand the variable and can even be used to forecast y's future or expected values, e.g., using its mean (and standard deviation). In case of time-series sample data, we can also exploit y's time trend and forecast its future values by building, e.g., auto-regression and moving-average models.

In addition to the above single-variable analysis, we are often more interested in multi-variable causal analysis that may include some or all of the following steps. First, those variables that have impacts in y should be identified. Since there are usually many factors affecting y directly or indirectly with different strengths, attention is often restricted to a few major variables that have noticeable impacts in y. To be simple while without loss of much generality, we assume that there is only a single affecting variable x and we have also collected its sample data on the same n subjects: $x_1, \ldots x_n$.

Second, the relationship between y and x should be examined and established, usually in a functional form. Since x is only one of the many factors affecting y, such relationship is usually characterized by a major part through a function g(x) reflecting the effect of x in y *and* an error term ε reflecting the effects of the omitted factors (and measurement errors):

$$y = g(x) + \varepsilon \qquad (1)$$

Hence for the observed subjects we have

$$y_i = g(x_i) + \varepsilon_i, \;\; 1 \le i \le n \qquad (2)$$

Eq.(1) or (2) is a general regression or curve-fitting model where g(x) will be referred to as a regression or fitting function in this paper.

Third, the concrete form of the function g(x) should be estimated using the available sample data on y and x. This may include two cases. The first case is when the functional form of g(x) is known but with some unknown parameters. For example, based on theory and experience, g(x) may be taken as linear ($a + bx$), polynomial (e.g., $a + bx + cx^2$), exponential (e.g., $a + be^{cx}$), or logistic (e.g., $a/(1 + be^{cx})$), with a, b and c being unknown parameters. This is the common case where the objective is to estimate the unknown parameters using the sample data. The second case is when the functional form of g(x) is unknown, and our objective is to estimate this unknown function using observational data.

Fourth, the goodness of model-fitting should be evaluated, which may also include two cases. The first case is when the probability density function for error ε, h(ε), is known but with some unknown parameters. Usually ε can be assumed to be normal with

$h(\varepsilon) = [\exp(-0.5\varepsilon^2/\sigma^2)]/[\sigma(2\pi)^{0.5}]$,

where σ is ε's standard deviation, but may also follow other distributions (e.g., lognormal). In this common case the objective is to estimate the unknown parameters in g(x) as well as in h(ε) using observational data. With ε's and hence y's distribution estimated, we can then evaluate the accuracy of the parameter estimates and the goodness of the overall model-fitting. The second case is when the functional form of h(ε) is unknown, and our objective is to estimate this unknown density using observational data and then to evaluate the model-fitting.

It should be noticed that, without estimating h(ε) or without any information on the error distribution, it is also possible to evaluate the goodness of model-fitting, e.g., through re-sampling techniques such as bootstrapping and jackknifing or through studying the large sample properties of the model estimates.

Finally, the values of y corresponding to new values of x should be forecast and the accuracy of such forecast should be evaluated. With g(x) estimated, such forecast is straightforward. With h(ε) estimated, evaluating the forecast accuracy is also not difficult. Even without estimating h(ε), it is still possible to evaluate the forecast accuracy through bootstrapping or jackknifing methods.

The estimation of model (1) can also be considered from another perspective. For two random variables X and Y with joint density $h_{X,Y}(x,y)$, the marginal density of X can be obtained by integrating $h_{X,Y}(x,y)$ with respect to y as $h_X(x) = \int h_{X,Y}(x,y)dy$ and the conditional density of Y given X is given as $h_{Y|X}(y|x) = h_{X,Y}(x,y)/h_X(x)$. Given a realization or observation of X: X = x, there may be many corresponding Y values whose mean is the so-called conditional expectation of Y calculated as: $u(x) = E[Y|X=x] = \int yh_{Y|X}(y|x)dy = \int yh_{X,Y}(x,y)dy/h_X(x)$. Denoting $\varepsilon = Y - u(x)$, then we have $Y = u(x) + \varepsilon$ where $E[\varepsilon|X = x] = 0$, and for any pair of observation (x_i, y_i) on (X, Y) we have $y_i = u(x_i) + \varepsilon_i$. This is the same as the above regression model (1) or (2). Thus, estimating g(x) in model (1) is essentially an estimation of the conditional expectation. In other words, model (1) can also be estimated with the help of the relevant density estimates.

Among the various issues or steps in estimating the general model (1) as outlined above, choosing an appropriate estimation method from many alternatives may be the most influential one, or at least an important starting point, especially for practitioners. The current paper is just so oriented to provide a practical review of and operational guide to some statistical methods available in the classical literature for estimating model (1) or more general regression models. For that purpose we first in Section 2 outline the popular least squares (LS) method in linear regression, followed by a description of the nonlinear regression procedures, to estimate model (1) when g(x) is of known form with unknown parameters. Then in Section 3 we turn to the case when g(x) in model (1) is of unknown form and present a number of nonparametric methods to estimate the unknown g(x), including the kernel method, series method, penalized LS method, spline method, and the method of sieves. Since density estimation is important by itself and helpful in estimating model (1) as indicated above, we describe several (nonparametric) methods to estimate the unknown densities in Section 4. Some concluding remarks are offered finally in Section 5.

2. Parametric Methods for Estimating Regression Models

2.1. Linear Regression

The general regression model (2), $y_i = g(x_i) + \varepsilon_i$, $1 \le i \le n$, is mainly characterized by the regression function g(x), which can be of either known form with a few unknown parameters or unknown form. Let us first consider the parametric case. Except for the trivial case when g(x) is a constant which can easily be estimated as y's sample mean, the simplest and most common case is a linear regression model with g(x) being linear (i.e., g(x) = a + bx):

$$y_i = a + bx_i + \varepsilon_i,\ \ 1 \le i \le n \tag{3}$$

Naturally, appropriate estimates for a and b in model (3) should be found in such a way that the errors $\varepsilon_1, \ldots, \varepsilon_n$ are as small as possible, i.e, the observational data points $(x_1, y_1), \ldots, (x_n, y_n)$ are close to the straight line y = a + bx as much as possible. A straightforward and actually more reliable realization of this intuitive idea is the least absolute deviations (LAD) method which estimates a and b by minimizing the sum of the absolute errors (Bloomfield and Steiger 1983; Harris 1950):

$$\mathit{Min}\ \sum_{1\le i\le n}|\varepsilon_i| = \sum_{1\le i\le n}|y_i - (a + bx_i)|$$

However, computational difficulties in obtaining the LAD estimates make it seldom adopted by practitioners. In contrast, the computationally easy LS approach to estimating a and b by minimizing the sum of the squared errors can be performed in many computer packages and even in some calculators and hence is much more popular in practice:

$$\mathit{Min}\ \sum_{1\le i\le n}(\varepsilon_i)^2 = \sum_{1\le i\le n}[y_i - (a + bx_i)]^2 \tag{4}$$

Denoting the sample means of x and y as $\bar{x} = \sum_{1\le i\le n} x_i/n$ and $\bar{y} = \sum_{1\le i\le n} y_i/n$, then the LS estimates for a and b can easily be solved out from minimization problem (4), a simple quadratic programming, as:

$$\hat{b} = \sum_{1\le i\le n}(x_i - \bar{x})(y_i - \bar{y})/\sum_{1\le i\le n}(x_i - \bar{x})^2 = \sum_{1\le i\le n}[(x_i - \bar{x})/\sum_{1\le k\le n}(x_k - \bar{x})^2]y_i$$

$$\hat{a} = \bar{y} - \hat{b}\,\bar{x} = \sum_{1\le i\le n}[1/n - \bar{x}(x_i - \bar{x})/\sum_{1\le k\le n}(x_k - \bar{x})^2]y_i$$

Clearly, both $\hat{a}$ and $\hat{b}$ are weighted sums of $y_1, \ldots, y_n$. The model-fitted value of y_i corresponding to x_i can then be calculated as:

$$\hat{y}_i = \hat{a} + \hat{b}\,x_i = (\bar{y} - \hat{b}\,\bar{x}) + \hat{b}\,x_i = \bar{y} + (x_i - \bar{x})\hat{b}$$

$$= \sum_{1\le j\le n} y_j/n + (x_i - \bar{x})\sum_{1\le j\le n}[(x_j - \bar{x})/\sum_{1\le k\le n}(x_k - \bar{x})^2]y_j$$

$$= \sum_{1\le j\le n}[1/n + (x_i - \bar{x})(x_j - \bar{x})/\sum_{1\le k\le n}(x_k - \bar{x})^2]y_j \equiv \sum_{1\le j\le n}[w_j(x_i)]y_j \tag{5}$$

which is a weighted *average* of $y_1, \ldots, y_n$, where the *normalized* weight

$$w_j(x_i) \equiv 1/n + (x_i - \bar{x})(x_j - \bar{x})/\sum_{1\le k\le n}(x_k - \bar{x})^2$$

is linear in x_i, dependent on the distance between x_i and the sample mean $\bar{x}$, and especially satisfies the *normalization* condition of $\sum_{1\le j\le n}[w_j(x_i)] = 1$ (for any x_i).

Without any conditions imposed, we can evaluate the overall LS fitting by examining the so-called R-square — that portion of the sample variance of y explained by the model:

$$R^2 = \sum_{1\le i\le n}(\hat{y}_i - \bar{\hat{y}})^2/\sum_{1\le i\le n}(y_i - \bar{y})^2 \tag{6}$$

where $\bar{\hat{y}} = \sum_{1\le i\le n} \hat{y}_i/n$ is the sample mean of the fitted $\hat{y}$.

Under some general conditions, the LS estimates are unbiased, consistent, and *best* linear unbiased estimators. If we further assume that the errors are normally distributed, then $y_1, \ldots, y_n$ and hence the LS parameter estimates are also normally distributed, and therefore their accuracies can be evaluated. If the errors are normally distributed, we can also use the maximum likelihood (ML) method to obtain estimates for the parameters and evaluate their accuracies. For more about the LS as well as the ML estimation methods, interested readers may consult appropriate textbooks on statistics or econometrics, such as Davidson and MacKinnon (2004), Davison (2003), Gelman and Nolan (2002) and Middleton (2004).

2.2. Nonlinear Regression

If g(x) in model (1) is nonlinear with some unknown parameters, the parameters can also be estimated using sample data and the LS method. One simple case is when g(x) is polynomial: $g(x) = a_0 + a_1x + a_2x^2 + \ldots + a_mx^m$, then we can use the LS method similarly as in the linear regression case to get estimates for the parameters a_0, a_1, a_2, …, a_m. Another simple case is when our model is exponential: $y = ae^{bx}\varepsilon$, which can first be converted into a linear one: $\ln(y) = \ln(a) + bx + \ln(\varepsilon)$, and then be estimated by the LS method.

For general nonlinear function of known form, $g(x) \equiv g(\boldsymbol{\theta}, x)$, where $\boldsymbol{\theta}$ is a column vector of the unknown parameters, we have the general nonlinear model: $y_i = g(\boldsymbol{\theta}, x_i) + \varepsilon_i$, $1 \le i \le n$. We can still apply the LS-type method to get the parameter estimate, $\hat{\boldsymbol{\theta}}$, by minimizing the sum of the squared errors as in the LS Eq.(4) for linear regression:

$$\mathit{Min}\ \sum_{1\le i\le n}(\varepsilon_i)^2 = \sum_{1\le i\le n}[y_i - g(\boldsymbol{\theta}, x_i)]^2$$

For example, we can first linearize $g(\boldsymbol{\theta}, x_i)$ at an appropriately-determined initial estimate $\boldsymbol{\theta}_0$: $g(\boldsymbol{\theta}, x_i) \approx g(\boldsymbol{\theta}_0, x_i) + (\partial g(\boldsymbol{\theta}_0, x_i)/\partial\boldsymbol{\theta})^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$, and then get a new estimate $\boldsymbol{\theta}_1$ by minimizing

$$\sum_{1\le i\le n}\{y_i - [g(\boldsymbol{\theta}_0, x_i) + (\partial g(\boldsymbol{\theta}_0, x_i)/\partial\boldsymbol{\theta})^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0)]\}^2$$

Repeating this process until two consecutive estimates $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_{t+1}$ are sufficiently close we can obtain the final parameter estimate $\hat{\boldsymbol{\theta}} \equiv \boldsymbol{\theta}_{t+1}$. This is the popular Gauss-Newton iterative method. Another more effective approach is to directly apply the popular Newton-Raphson nonlinear optimization method to the objective function $\sum_{1\le i\le n}[y_i - g(\boldsymbol{\theta}, x_i)]^2$ to get the parameter estimate $\hat{\boldsymbol{\theta}}$.

As in linear regression, without any conditions imposed on the nonlinear regression model we can still evaluate its overall LS-fitting by examining the R^2 as given by Eq.(6) with $\hat{y}_i = g(\hat{\boldsymbol{\theta}}, x_i)$ being the model-fitted value of y_i and $\bar{\hat{y}} = \sum_{1\le i\le n}\hat{y}_i/n$ being the sample mean of the fitted $\hat{y}$. However, no matter whether the error distribution is of known or unknown form in a nonlinear regression model, it is not easy to get the distributions for its parameter estimates and hence not easy to evaluate their accuracies. For more about estimating nonlinear regression models, interested readers may consult appropriate textbooks on statistics or econometrics, such as Davidson and MacKinnon (2004), Davison (2003), Gelman and Nolan (2002) and Middleton (2004) as cited in Section 2.1.

3. Nonparametric Methods for Estimating Regression Models

3.1. Local Regression Method

For regression model (2), if the functional form of $g(x)$ is unknown, then we are facing with the so-called nonparametric estimation of $g(x)$ using sample data (x_1, y_1), …, (x_n, y_n). Consider a simple case when X is a *discrete* random variable, and let one of its possible values be x^*. Then we can check how many X-observations $x_1,\dots,x_n$ equal to x^*. To be simple, suppose the first n^* X-observations equal to x^*: $x_i = x^*$, $1 \le i \le n^*$. Then a sensible estimate for the corresponding $y^* = g(x^*) + \varepsilon^*$ is the average of the first n^* Y-observations: $\hat{y}^* = \hat{g}(x^*) = \sum_{1\le i\le n^*} y_i/n^*$, which is a consistent estimate if $n^* \to \infty$ as $n \to \infty$. Clearly,

$$n^* = \sum_{1\le j\le n^*} I(x_i = x^*) = \sum_{1\le i\le n} I(x_i = x^*) = \sum_{1\le i\le n} I(-0.5\omega \le x_i - x^* \le 0.5\omega)$$
$$= \sum_{1\le i\le n} I(-0.5 \le (x_i - x^*)/\omega \le 0.5)$$

where $I(\Omega)$ is an indicator or characteristic function equal to 1 if condition Ω holds and zero otherwise, and ω is a bandwidth smaller than the distances between any pair of the distinct X-observations. Hence,

$$\hat{y}^* = \hat{g}(x^*) = \sum_{1\le i\le n^*} y_i/n^*$$
$$= \sum_{1\le i\le n} y_i\, I(-0.5 \le (x_i - x^*)/\omega \le 0.5)/\sum_{1\le j\le n} I(-0.5 \le (x_j - x^*)/\omega \le 0.5)$$
$$\equiv \sum_{1\le i\le n} [w_i(x^*)]y_i \qquad (7)$$

i.e., the estimated $\hat{y}^*$ or $\hat{g}(x^*)$ corresponding to x^* equals to the weighted *average* of all the observed Y-values, where the normalized weight:

$$w_i(x^*) = I(-0.5 \le (x_i - x^*)/\omega \le 0.5)/\sum_{1\le j\le n} I(-0.5 \le (x_j - x^*)/\omega \le 0.5)$$

depends on the distance between x_i and x^*, leading to the name of local regression estimate.

The above estimate (7) is very similar to the so called k^{th}-nearest neighbor (k-NN) estimates as initially proposed by Lofstgaarden and Quesenberry (1965), Royall (1966) and Watson (1964), in which

$w_i(x^*) = k^{-1}I_{ki}(x^*)$

where $I_{ki}(x^*) = 1$ if x_i is one of the k nearest observations to x* and 0 otherwise. It is also noticed that the local regression estimate (7) as well as the k-NN estimate are consistent with the LS fitted y-values as in Eq.(5) that are weighted averages of the observed $y_1, \ldots, y_n$, a reasonable property for such fits.

3.2. Kernel Method

If X is a *continuous* random variable, considering its any exact value x* as above is not meaningful. Another problem with the simple g(x) estimate in Eq.(7) is its discontinuities at points $x_i \pm 0.5\omega$. A natural expansion is to estimate $y^* = g(x^*) + \varepsilon^*$ as the average of such y_i's that correspond to those x_i's in a small interval around x*. This can be done by replacing the above indicator function $I(-0.5 \le (x_i - x^*)/\omega \le 0.5)$ by a general kernel function $K((x_i - x^*)/\omega)$:

$$\hat{y}^* = \hat{g}(x^*) = \sum_{1\le i\le n} y_i K((x_i - x^*)/\omega)/\sum_{1\le j\le n} K((x_j - x^*)/\omega) \qquad (8)$$

which is called the Nadaraya-Watson kernel estimate (Nadaraya 1964, 1965; Watson 1964), where a kernel K(x) is a continuous function such that $K(x) \ge 0$ for any x, $K(-\infty) = K(+\infty) = 0$ and $\int K(x)dx = 1$, and the bandwidth ω controls the kernel estimate's smoothness. Commonly-used kernels include the Bartlett (1963) or Epanechnikov (1969) kernel with $K(x) = 0.75(1 - x^2)I(-1 \le x \le 1)$ and the normal kernel with $K(x) = (2\pi)^{-0.5}\exp(-0.5x^2)$.

If kernels are allowed to be discontinuous, then the indicator function I(Ω) in Section 3.1 is obviously a such kernel, called the uniform or naive kernel with $K(x) = I(-0.5 \le x \le 0.5)$. Hence, the kernel estimate of g(x) in Eq.(8) is a direct generalization over the simple local regression estimate in Eq.(7).

3.3. Series Method

Series estimation method was initially proposed by Čencov (1962) using orthogonal Fourier series. Without loss of generality, suppose random variable X takes values on the unit interval [0, 1], implying that the conditional expectation g(x) is defined on [0, 1]. Then g(x) can be expressed as a Fourier series $\sum_{0\le j\le\infty} a_j\varphi_j(x)$, where the coefficients $a_j = \int_{[0,1]} g(x)\varphi_j(x)dx$, $0 \le j \le \infty$, and the orthogonal sequence $\{\varphi_j(x)\}_{0\le j\le\infty}$ is taken as: $\varphi_0(x) = 1$, $\varphi_j(x) = 2^{0.5}\cos(j+1)\pi x$ when j is odd, and $\varphi_j(x) = 2^{0.5}\sin(\pi jx)$ when j is even.

Given sample observations $(x_1, y_1), \ldots, (x_n, y_n)$ on (X, Y), we need to estimate the coefficient

$$a_j = \int_{[0,1]} g(x)\varphi_j(x)dx$$

in order to estimate $g(x) = \sum_{0\le j\le\infty} a_j\varphi_j(x)$. To be simple, suppose $x_1, \ldots, x_n$ are distinct from each other and are ranked as $0 \le x_1 < x_2 < \ldots < x_n \le 1$, then we can divide the interval [0, 1] into n small non-overlapping intervals $A_1, A_2, \ldots, A_n$ so that $x_i \in A_i$ and, for any $x \in A_i$, $g(x) \approx g(x_i) \approx y_i$, which leads to:

$$a_j = \int_{[0,1]} g(x)\varphi_j(x)dx = \sum_{1\le i\le n}\int_{A_i} g(x)\varphi_j(x)dx \approx \sum_{1\le i\le n} y_i\int_{A_i} \varphi_j(x)dx$$

That is, we get an estimate for a_j as $\hat{a}_j = \sum_{1\le i\le n} y_i\int_{A_i} \varphi_j(x)dx$, and hence an orthogonal series estimate for $g(x) = \sum_{0\le j\le\infty} a_j\varphi_j(x)$ is given by:

$$\hat{y} = \hat{g}_m(x) = \sum_{0\le j\le m} \hat{a}_j\varphi_j(x) = \sum_{0\le j\le m}[\sum_{1\le i\le n} y_i\int_{A_i} \varphi_j(x)dx]\varphi_j(x)$$

$$= \sum_{1\le i\le n} y_i[\sum_{0\le j\le m} \varphi_j(x)\int_{A_i} \varphi_j(x)dx] \qquad (9)$$

where the cutoff point m in the infinite sum determines the degree of smoothing in the estimate, corresponding to the bandwidth ω in the kernel estimates. It is noticed that, since $\hat{a}_j$ is a weighted sum of $y_1, \ldots, y_n$, the series estimate in Eq.(9) is too and hence consistent in this sense with the kernel estimate (8) and the LS estimate (5).

3.4. Linear Spline Method

A linear spline is a piece-wise continuous linear function with a number of linear functions joined together at some points called knots. For model (2): $y_i = g(x_i) + \varepsilon_i$, $1 \le i \le n$, we can take $\hat{g}(x)$ as a linear spline with the n observations $(x_1, y_1), \ldots, (x_n, y_n)$ as knots to estimate g(x). The so-obtained

$\hat{g}(x)$ is a perfect match with the sample data, but this perfect match is practically useless. A sensible way is to find a linear spline $\hat{g}(x)$ with fewer knots but still fitting the sample data satisfactorily. For this purpose, assume without loss of generality that $g(x)$ is defined over a limited range $[p, q]$. Then we can use an (m+1)-knot spline $\hat{g}_m(x)$ to approximate $g(x)$. The m+1 knots, $p = t_0 < t_1 < \ldots < t_m = q$, are usually assumed to be equally spaced over $[p, q]$ with width or mesh size $\omega = (q-p)/m$, i.e., $t_j = p + j\omega$, $0 \le j \le m$. Then

$$\hat{g}_m(x) = [(t_j - x)/\omega]g(t_{j-1}) + [(x - t_{j-1})/\omega]g(t_j), \quad x \in [t_{j-1}, t_j],\ 1 \le j \le m$$

More clearly,

$$\begin{aligned}\hat{g}_m(x) &= [(t_1 - x)/\omega]g(t_0) + [(x - t_0)/\omega]g(t_1), && \text{if } x \in [t_0, t_1]\\ &= [(t_2 - x)/\omega]g(t_1) + [(x - t_1)/\omega]g(t_2), && \text{if } x \in [t_1, t_2]\\ &\ldots\ldots\ldots\ldots\\ &= [(t_m - x)/\omega]g(t_{m-1}) + [(x - t_{m-1})/\omega]g(t_m), \text{ if } x \in [t_{m-1}, t_m]\end{aligned}$$

That is, over each sub-interval $[t_{j-1}, t_j]$, $1 \le j \le m$, $\hat{g}_m(x)$ is a linear function linking or passing through the two points: $(t_{j-1}, g(t_{j-1}))$ and $(t_j, g(t_j))$. Hence, if the mesh size ω or each sub-interval is sufficiently small, $\hat{g}_m(x)$ can approximate $g(x)$ quite closely.

To express the linear spline $\hat{g}_m(x)$ more effectively, it is helpful to define m+1 triangle-type base functions as follows (Prenter 1976):

- For j = 0 (right triangle with base width of ω and height of 1):

  $$\begin{aligned}B_{m,0}(x) &= (t_1 - x)/\omega, && \text{if } x \in [t_0, t_1]\\ &= 0, && \text{otherwise}\end{aligned}$$

- For j = 1, …, m-1 (symmetric triangle centered at t_j with base width of 2ω and height of 1):

  $$\begin{aligned}B_{m,j}(x) &= (x - t_{j-1})/\omega, && \text{if } x \in [t_{j-1}, t_j]\\ &= (t_{j+1} - x)/\omega, && \text{if } x \in [t_j, t_{j+1}]\\ &= 0, && \text{otherwise}\end{aligned}$$

- For j = m (right triangle with base width of ω and height of 1):

  $$\begin{aligned}B_{m,m}(x) &= (x - t_{m-1})/\omega, && \text{if } x \in [t_{m-1}, t_m],\\ &= 0, && \text{otherwise}\end{aligned}$$

Then the linear spline $\hat{g}_m(x)$ can be written as a weighted sum of these base functions with $g(t_j)$ as weights:

$$\hat{g}_m(x) = \textstyle\sum_{0\le j\le m} g(t_j)B_{m,j}(x) \qquad (10)$$

It can be proved (Prenter 1976) that such $\hat{g}_m(x)$ is a good approximation to $g(x)$ as $m \uparrow \infty$ (i.e., there are more and more knots or sub-intervals) or equivalently as $\omega \downarrow 0$ (i.e., each sub-interval is narrower and narrower):

$$|\hat{g}_m(x) - g(x)| \le (\| g^{(2)}(x) \|_\infty /4)\omega^2 \qquad (11)$$

Hence, we can use the linear spline $\hat{g}_m(x)$ as defined in Eq.(10) to fit the unknown $g(x)$ with the n data points $(x_1, y_1), \ldots, (x_n, y_n)$ by minimizing the sum of the squared fitting-errors, still an LS-type approach:

$$\mathit{Min}\ \textstyle\sum_{0\le i\le n}[y_i - \hat{g}_m(x_i)]^2 = \sum_{0\le i\le n}[y_i - \sum_{0\le j\le m} g(t_j)B_{m,j}(x_i)]^2 \qquad (12)$$

As long as we choose no more than n knots (i.e., $m \le n-1$), we can find from Eq.(12) the optimal $g(x)$ values at these knots: $g(t_j)$, $0 \le j \le m+1$. Connecting these m+1 knots by straight lines we get

$$\hat{g}_m(x) = \textstyle\sum_{0\le j\le m} g(t_j)B_{m,j}(x)$$

as in Eq.(10), the optimal linear spline estimate for g(x). If our sample size n is large, then we can choose more knots (i.e., a bigger m or a smaller ω), which can increase the accuracy of such linear spline estimation according to Eq.(11).

3.5. Penalized Ls Method

For our model (2): $y_i = g(x_i) + \varepsilon_i$, $1 \le i \le n$, where g(x) is an unknown function which in this Section is *assumed* to be mth-order differentiable. The traditional LS method to obtain the optimal g(x) by minimizing $\sum_{1\le i\le n} (\varepsilon_i)^2 \equiv \sum_{1\le i\le n} [y_i - g(x_i)]^2$ does not work here, since any m^th-order differentiable function passing through the n sample points (x_1, y_1), …, (x_n, y_n) is such an optimal solution with $\sum_{1\le i\le n}[y_i - g(x_i)]^2$ reaching the minimum possible value of 0 and there are infinitely many such functions. For example, for any real number r, it is possible to find a polynomial with n coefficients: $\hat{g}(x) = x^r(a_0 + a_1x + a_2x^2 + \ldots + a_{n-1}x^{n-1})$, which is m^th-order differentiable and passes through all the n sample points.

To resolve this problem, one possible way is to introduce a penalty function into the objective function to make the minimization problem meaningfully solvable, i.e., to get the optimal g(x) by solving the following modified minimization problem:

$$Min \sum_{1\le i\le n}[y_i - g(x_i)]^2 + \lambda\int[g^{(m)}(x)]^2dx \qquad (13)$$

where $g^{(m)}(x)$ is the m^th-order derivative of g(x) and $\lambda > 0$ is a smoothing parameter. In Eq. (13), minimizing the original LS term $\sum_{1\le i\le n}[y_i - g(x_i)]^2$ is as before to guarantee a good fit of g(x) to the sample data, while minimizing the penalty term $\int[g^{(m)}(x)]^2dx$ is to require the smoothness of g(x) in the sense of having a small m^th-order derivative. It turns out that the optimal g(x), $\hat{g}(x)$, is a polynomial spline of order 2m-1 with possible knots at the sample data points (see Schoenberg (1964) for the case of m = 1, Reinsch (1967) for the case of m = 2, and Kimeldorf and Wahba (1970a and 1970b) for the general case). It is also noticed that λ controls the degree of smoothness of g(x): when $\lambda \to 0$ the original LS term dominates and $\hat{g}(x)$ tends to be an interpolating function of the sample data, fitting all the n data points exactly and making $\sum_{1\le i\le n}[y_i - g(x_i)]^2 = 0$; and when $\lambda \to \infty$ the penalty term dominates and $\hat{g}(x)$ tends to be the LS polynomial of order m-1 (since $\int[g^{(m)}(x)]^2dx$ tends to be zero if and only if g(x) is a polynomial of order m-1) passing through the sample data points.

3.6. Method of Sieves

As mentioned above, when we use the LS method to estimate the unknown g(x) in model (2): $y_i = g(x_i) + \varepsilon_i$, $1 \le i \le n$, there are infinitely many functions which minimizes $\sum_{1\le i\le n}(\varepsilon_i)^2 \equiv \sum_{1\le i\le n}[y_i - g(x_i)]^2$. This is mainly due to the fact that, unlike traditional linear regression with only a few unknown parameters, here the "parameter" space S = {g(x) | g(x) is continuous or differentiable over [a, b]} is too "big" with infinite dimensions. So another possible approach to resolving this problem is to limit the infinite-dimensional functional space S to a series of increasingly bigger (but still finite-dimensional) functional spaces, $S_0 \subset S_1 \subset S_2 \subset S_3 \subset \ldots$, which more and more approaches S. Here the functional-space series $\{S_m\}_{0\le m\le\infty}$ is called a "sieve". Within each finite-dimensional space S_m, we can get the optimal solution $\hat{g}_m(x)$ by minimizing $\sum_{1\le i\le n}[y_i - g(x_i)]^2$ subject to $g(x) \in S_m$. And the limit of $\hat{g}_m(x)$ is what we want. This is the so-called "method of sieves" initially suggested by Grenander (1981). In practical applications, depending on our sample size, usually a moderate m such as m = 5, 10 or 20 may be enough, i.e., we may take $\hat{g}_5(x)$, $\hat{g}_{10}(x)$ or $\hat{g}_{20}(x)$ as the final solution.

So the key for applying the method of sieves is to construct appropriate sieves $\{S_m\}_{0\le m\le\infty}$. One example is based on the Fourier series expansions as discussed in Section 3.3:

$$S_m = \{g(x) \mid g(x) = \textstyle\sum_{0\le j\le m} a_j\varphi_j(x)\}$$

Another example comes from the splines as discussed in Section 3.4:

$$S_m = \{g(x) \mid g(x) = \textstyle\sum_{0\le j\le m} g(t_j)B_{m,j}(x)\}$$

Yet a further example corresponds to the penalized LS estimation in Section 3.5:

$$S_m = \{g(x) \mid g(x) \text{ is } (m+1)^{th}\text{-order differentiable with } \int[g^{(m+1)}(x)]^2dx < \infty\}$$

For more about the method of sieves, interested readers may consult, e.g., Geman and Hwang (1982) and Grenander (1981).

4. Estimating Density Functions

Estimating the probability density function h(x) of a random variable X given its sample observations $x_1, \ldots, x_n$ is of theoretical and practical values. If h(x) is of known form with some unknown parameters, then we can use the popular ML method to estimate these parameters and hence to clearly determine X's distribution. The moment method can also be applied to solve the same problem by matching X's theoretical moments with its sample ones. If the form of h(x) is unknown, then we need to apply nonparametric methods to estimate it, much the same way as we estimate the unknown regression function g(x) in Section III with certain differences since h(x) is a density function, not that arbitrary as g(x).

On the other hand, density estimation can also help estimate the regression function g(x), as mentioned in the introduction Section. Here a noticeable fact is that g(x) is the conditional expectation of Y given X = x, which depends on the joint density of X and Y. Hence, if we can estimate joint densities, we can then estimate the corresponding regression functions.

4.1. Local Histogram Estimate of Density

Consider a random variable X with realizations $x_1, \ldots, x_n$. If X is *discrete*, then its density h(x) equals to the probability of X = x. A common and consistent estimate for h(x) is the empirical histogram whose function form is:

$$\hat{h}(x) = \{\text{number of } x_1,\ldots,x_n \text{ equal to } x\}/n = n^{-1}\sum_{1\le i\le n} I(x_i = x)$$

If X is *continuous*, then $h(x)\omega$ can be approximated as the probability of $X \in [x - 0.5\omega, x + 0.5\omega]$ for small ω values, and hence can also be estimated by the empirical histogram as:

$$\hat{h}(x)\omega = \{\text{number of } x_1,\ldots,x_n \text{ within } [x - 0.5\omega, x + 0.5\omega]\}/n$$

$$= n^{-1}\sum_{1\le i\le n} I(x - 0.5\omega \le x_i \le x + 0.5\omega) = n^{-1}\sum_{1\le i\le n} I(-0.5 \le (x_i - x)/\omega \le 0.5)$$

from which h(x) can be estimated as:

$$\hat{h}(x) = (n\omega)^{-1}\sum_{1\le i\le n} I(-0.5 \le (x_i - x)/\omega \le 0.5)$$

Clearly, $\hat{h}(x)$ has discontinuities at $x_i \pm 0.5\omega$, and a lot of observations are needed to make it sufficiently smooth and close to the true density h(x).

4.2. Kernel Estimate of Density

To make the above histogram-type density estimates smooth, a sensible way is to replace the indicator function I(x) by a continuous kernel function K(x):

$$\hat{h}(x) = (n\omega)^{-1}\sum_{1\le i\le n} K((x_i - x)/\omega) \tag{14}$$

which is also called the Rosenblatt (1956, 1969) kernel estimate. Under general conditions, the kernel density estimate is asymptotic unbiased and consistent, and the bandwidth ω can be optimally chosen to be proportional to $n^{-0.2}$ (see, e.g., Silverman 1986).

As mentioned in Section 3.2, kernel function K(x) can be selected as the Bartlett or Epanechnikov kernel or the normal kernel, among other choices. Interested readers may consult Härdle (1990) and Silverman (1986) for more discussions about kernels, but it should be noted that the specific nature of the selected kernel is usually *not* critical to the performance of the density estimate (Pagan and Ullah 1999).

4.3. Estimating Joint and Marginal Densities and Conditional Expectations

For two random variables X and Y, their joint density $h_2(x, y)$ can similarly be estimated using sample observations $(x_1, y_1), \ldots, (x_n, y_n)$ as the following two-dimensional histogram:

$$\hat{h}_2(x, y) = (n\omega)^{-2}\sum_{1\le i\le n} I(x - \omega/2 \le x_i \le x + \omega/2) \times I(y - \omega/2 \le y_i \le y + \omega/2)$$

$$= (n\omega)^{-2}\sum_{1\le i\le n} I(-1/2 \le (x_i - x)/\omega \le 1/2) \times I(-1/2 \le (y_i - y)/\omega \le 1/2)$$

or more generally using a two-dimensional kernel as:

$\hat{h}_2(x, y) = (n\omega)^{-2}\sum_{1\le i\le n} K_2((x_i - x)/\omega, (y_i - y)/\omega)$

It is clear that $K_1(x) = \int K_2(x, y)dy$ is a one-dimensional kernel, and the kernel estimate of the marginal density $h_1(x)$ is

$\hat{h}_1(x) = \int\hat{h}_2(x, y)dy = (n\omega)^{-2}\sum_{1\le i\le n} \int K_2((x_i - x)/\omega, (y_i - y)/\omega)dy$

$= (n\omega)^{-1}\sum_{1\le i\le n} \int K_2((x_i - x)/\omega, y)dy = (n\omega)^{-1}\sum_{1\le i\le n} K_1((x_i - x)/\omega)$

which is the same as the previous kernel density estimate in Eq.(14).

Since the conditional expectation of Y given X is $u(x) = E(Y|X=x) = \int yh_2(x,y)dy/h_1(x)$, a kernel estimate for it is proposed as follows (Nadaraya 1964, 1965; and Watson 1964):

$\hat{u}(x) = \int y\hat{h}_2(x, y)dy/\hat{h}_1(x)$

$= (n\omega)^{-2}\sum_{1\le i\le n} \int yK_2((x_i - x)/\omega, (y_i - y)/\omega)dy/[(n\omega)^{-1}\sum_{1\le j\le n} K_1((x_j - x)/\omega)]$

If the two-dimensional kernel $K_2(x, y)$ is symmetric (on y), then the above expression can be reduced to (by making transformation $z = (y_i - y)/\omega$ in the numerator integral):

$\hat{u}(x) = \sum_{1\le i\le n} y_iK_1((x_i - x)/\omega)/\sum_{1\le j\le n} K_1((x_j - x)/\omega)$

which, a weighted average of the observed $y_1, \ldots, y_n$, is just a formal derivation of the previously obtained kernel estimate for the regression function as in Eq.(8).

4.4. Series Estimate of Density

Without loss of generality, now suppose X is a random variable with density h(x) on the unit interval [0, 1]. Then h(x) can be expressed as a Fourier series $\sum_{0\le j\le\infty} c_j\varphi_j(x)$ as in Section 3.3, where the coefficients

$c_j = \int_{[0,1]} h(x)\varphi_j(x)dx = E[\varphi_j(X)], 0 \le j \le \infty$

and the orthogonal sequence $\{\varphi_j(x)\}_{0\le j\le\infty}$ is taken as: $\varphi_0(x) = 1$, $\varphi_j(x) = 2^{0.5}\cos(j+1)\pi x$ when j is odd, and $\varphi_j(x) = 2^{0.5}\sin(\pi jx)$ when j is even.

Given sample observations $x_1, \ldots, x_n$ on X, an obvious estimate for the coefficient $c_j = E[\varphi_j(X)]$ is $\hat{c}_j = n^{-1}\sum_{1\le i\le n}\varphi_j(x_i)$. Hence, as first suggested by Čencov (1962), an orthogonal series estimate for density $h(x) = \sum_{0\le j\le\infty} c_j\varphi_j(x)$ can be given by:

$\hat{h}_m(x) = \sum_{0\le j\le m} \hat{c}_j\varphi_j(x) = n^{-1}\sum_{0\le j\le m}[\sum_{1\le i\le n}\varphi_j(x_i)]\varphi_j(x)$

where as before the cutoff point m in the infinite sum determines the degree of smoothing in the estimate, corresponding to the bandwidth ω in the kernel-type estimates.

4.5. Penalized Likelihood Estimate of Density

By conventional definition, the likelihood of the unknown density h(x) is $L(h) \equiv L(h|x_1,\ldots,x_n) \equiv \prod_{1\le i\le n} h(x_i)$, and $\log(L(h)) = \sum_{1\le i\le n} \log(h(x_i))$, which obviously has no maximum over the universe of all densities. As in the penalized LS estimate of the unknown regression function g(x) discussed in Section 3.5, here we can also get the penalized ML estimate of the unknown density function h(x) by adding to log(L(h)) a term S(h) that represents a smoothness requirement for h(x). That is, we can get the estimate $\hat{h}(x)$ of h(x) by solving the following maximization problem:

$$Max\ \sum_{i\le i\le n}\log[h(x_i)] - \lambda S(h) \qquad (15)$$

over all densities h(x) that satisfy $h(x) \ge 0$ for any x, $\int h(x)dx = 1$, and $S(h) < \infty$, where $\lambda > 0$ is a smoothness-control parameter.

Two good smoothness-control measures are suggested by Good and Gaskins (1971, 1980) as $S_1(h) = \int\{[h^{(1)}(x)]^2/h(x)\}dx$ and $S_2(h) = \int[h^{(2)}(x)]^2dx$. Other forms of S(h) can also be adopted (see, e.g., Silverman 1982, 1984). Interestingly, it is showed (Silverman 1984) that the above penalized likelihood estimate can essentially be regarded as a special kernel estimate with variable bandwidths. It is also noticed that, since Eq.(15) cannot produce an explicit expression for $\hat{h}(x)$, certain computational procedures have emerged to solve the maximization problem (see, e.g., Good and Gaskins 1971, 1980).

4.6. Sieve Estimate of Density

Instead of adding a penalizing term in the likelihood as in Eq.(15), we can also use the method of sieves to directly maximize L(h) or log(L(h)) subject to h(x) $\in S_m$, where the sieve $\{S_m\}_{0\le m\le\infty}$ is a series of increasingly bigger functional spaces. A simple sieve comes from the empirical histograms as discussed in Section 4.1:

$S_m = \{h(x) \mid h(x)$ is a density and is constant when $j\text{-}1 \le (m+1)x < j$, $j = 0, \pm1, \pm2, \ldots\}$

Another example is based on the Fourier series expansions as discussed in Section 4.4:

$S_m = \{h(x) \mid h(x) = \sum_{0\le j\le m} c_j\varphi_j(x)\}$

Yet a further example corresponds to the normal kernels in Section 4.2:

$S_m = \{h(x) \mid h(x) = \int m(2\pi)^{-0.5}\exp(-0.5m^2(x-y)^2)dH(y)$, H is an arbitrary distribution$\}$

For more about the method of sieves for estimating density functions and regression models as well, interested readers may consult, e.g., Geman and Hwang (1982) and Grenander (1981).

5. Concluding Remarks

This paper provides a practical overview of some classical statistical methods to estimate the general regression model: $y = g(x) + \varepsilon$, or its observational form: $y_i = g(x_i) + \varepsilon_i$, $1 \le i \le n$. We first briefly present the parametric approaches when g(x) is of known form with unknown parameters, notably the LS method for linear and nonlinear regression models. Then we discuss the case when the functional form of g(x) is unknown and describe a number of nonparametric methods for estimating g(x). We also outline some nonparametric methods for estimating the unknown probability densities, which is theoretically and practically important by itself and useful in estimating regression models. The methods reviewed in this paper are of course not exhaustive, but certainly many popular methods have been covered. Our review is mainly application-oriented with much attention paid to simple while still complete descriptions of the various estimation methods for practitioners' easy understanding. Also, we provide concise and sufficient operational or procedural details for each method reviewed, so that practitioners can easily apply the appropriate method(s) to their specific estimation problems through straightforward computations or computerized programming.

We can without difficulty calculate the R^2 for each parametric or nonparametric regression approach as in the traditional LS method to evaluate the goodness-of-fit of an estimated model, and we can also easily make forecasts based on the estimated models. However, in most cases this paper does not involve the properties of the estimates produced by each method reviewed, such as unbiasedness, efficiency, and consistency, mainly due to their mathematical complexities that are difficult to be appropriately handled in our brief review paper. Interested readers may consult such publications as our listed References to examine the statistical properties of the relevant methods' estimates.

For simplicity, this paper focuses on the case of only one explanatory variable. However, the parametric and nonparametric methods reviewed can be extended to the case of multiple explanatory variables without great difficulty. It is noticed that, when multiple explanatory variables are presented, the model may not be fully parametric or nonparametric but can instead be partly parametric and partly nonparametric, i.e., semi-parametric. For example, if we have two explanatory variables x and z, then the regression model becomes $y = g(x, z) + \varepsilon$, for which three cases are possible. The first case is that g(x, z) is of known form with unknown parameters, e.g., $g(x, z) = a + bx + cz$, then the model is fully parametric and the parametric methods as reviewed in Section II can easily be extended to estimate the parameters involved. The second case is when g(x, z) is of unknown form, then the model is fully nonparametric and the nonparametric methods as reviewed in Sections III and IV can be appropriately extended to estimate g(x, z). The third case is when g(x, z) is semi-parametric, e.g., $g(x, z) = a + bx + f(z)$ where a and b are parameters and f(z) is of unknown form, then we need the so-called semi-parametric methods to estimate both the parameters and the unknown function(s) involved. A number of semi-parametric methods are available to estimate various types of semi-parametric models, and interested readers may consult, e.g., Horowitz (1998) and Ruppert et al. (2003), to examine and apply such estimation methods.

ACKNOWLEDGEMENT

This research was supported by a research grant (R201832) from United International College, Beijing Normal University-Hong Kong Baptist University.

REFERENCES

[1] Bartlett, M. S. 1963. Statistical estimation of density functions. *Sankhya* (Series A) 25: 245-254.

[2] Bloomfield, P., and W. L. Steiger. 1983. *Least Absolute Deviations: Theory, Applications, and Algorithms*. Boston: Birkhäuser.

[3] Čencov, N. N. 1962. Evaluation of an unknown distribution density from observations. *Soviet Mathematics* 3: 1559-1562.

[4] Davidson, R., and J. G. MacKinnon. 2004. *Econometric Theory and Methods*. New York: Oxford University Press.

[5] Davison, A. C. 2003. *Statistical Models*. New York: Cambridge University Press.

[6] Epanechnikov, V. A. 1969. Nonparametric estimates of a multivariate probability density. *Theory of Probability and Its Applications* 14: 153-158.

[7] Gelman, A., and Nolan, D. 2002. *Teaching Statistics: A Bag of Tricks*. New York: Oxford University Press.

[8] Geman, S., and C. Hwang. 1982. Nonparametric maximum likelihood estimation by the method of sieves, *Annals of Statistics* 10: 401-414.

[9] Good, J. J., and R. A. Gaskins. 1971. Nonparametric roughness penalties for probability densities. *Biometrika* 58: 255-277.

[10] Good, J. J., and R. A. Gaskins. 1980. Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association* 75: 42-73

[11] Grenander, U. 1981. *Abstract Inference*. New York: Wiley.

[12] Härdle, W. 1990. *Applied Nonparametric Regression*. New York: Cambridge University Press.

[13] Harris, T. 1950. Regression using minimum absolute deviations. *American Statistician* 4: 14-15.

[14] Horowitz, J. L. 1998. *Semiparametric Methods in Econometrics*. New York: Springer-Verlag.

[15] Kimeldorf, G. S., and G. Wahba. 1970a. A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Annals of Mathematical Statistics* 41: 495-502.

[16] Kimeldorf, G. S., and G. Wahba. 1970b, Spline functions and stochastic process. *Sankhya* (Series A) 32: 173-180.

[17] Lofstgaarden, D. O., and C. P. Quesenberry. 1965. A nonparametric estimate of a multivariate density function. *Annals of Mathematical Statistics* 36: 1049-1051.

[18] Middleton, M. R. 2004. *Data Analysis Using Microsoft Excel: Updated for Office XP* (3rd ed.). Belmont, CA: Thomson.

[19] Nadaraya, É. A. 1964. On estimating regression. *Theory of Probability and Its Applications* 9: 141-142.

[20] Nadaraya, É. A. 1965. On nonparametric estimation of density functions and regression curves. *Theory of Probability and Its Applications* 10: 186-190.

[21] Pagan, A., and A. Ullah. 1999. *Nonparametric Econometrics*. New York: Cambridge University Press

[22] Prenter, M. V. 1976. *Splines and Variational Methods*. New York: Wiley.

[23] Reinsch, H. 1967. Smoothing by spline functions. *Numerische Mathematik* 10: 177-183.

[24] Rosenblatt, M. 1956. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics* 27: 642-669.

[25] Rosenblatt, M. 1969. Conditional probability density of regression estimators. In *Multivariate Analysis* II. Edited by P. R. Krishnaiah. New York: Academic Press, pp. 25-31.

[26] Royall, R. M. 1966. *A Class of Nonparametric Estimates of a Smooth Regression Function*. Ph.D. Thesis: Stanford University.

[27] Ruppert, D., M. P. Wand, and R. J. Carroll. 2003. *Semiparametric Regression*. New York: Cambridge University Press.

[28] Schoenberg, I. J. 1964. Spline functions and the problem of graduation. *Proceedings of the National Academy of Sciences of the United States of America* 52: 947-950.

[29] Silverman, B. W. 1982. On the estimation of a probability density function by the maximum penalized likelihood method. *Annals of Statistics* 10: 795-810.

[30] Silverman, B. W. 1984. Spline smoothing: the equivalent variable kernel model. *Annals of Statistics* 12: 898-916.

[31] Silverman, B. W. 1986. *Density Estimation for Statistical and Data Analysis*. New York: Chapman and Hall.

[32] Watson, G. S. 1964. Smooth regression analysis. *Sankhya* (Series A) 26: 359-372.

Citation: *Ming-Lu Wu, (2020). Estimating Regression Models and Density Functions with Classical Parametric and Nonparametric Methods: A Practical Overview. International Journal of Scientific and Innovative Mathematical Research (IJSIMR), 8(2), pp. 10-21. http://dx.doi.org/ 10.20431/2347 - 3142.0802002*