Kernel Smoothing Function and Choosing Bandwidth for Non-Parametric Regression Methods

Murat Kayri and Gürol Zırhlıoğlu
Yuzuncu Yil University, Faculty of Education, Van/Turkey

Abstract: It is possible to get detailed results between independent and dependent variables in non-parametric regression methods. By using non-parametric regression methods, the estimation of curve may have complexity. This is a crucial problem to read and interpret the regression curve. To solve this problem, some smoothing functions were developed. Kernel function is one of the smoothing techniques. Kernel function choices the optimal bandwidth value \((h)\) to get a readable curve. Also the distribution type is important for Kernel smoothing function. The Kernel function obtains the different curves according to distribution type (Gaussian, Uniform, Quartic, etc). To show Kernel smoothing process, internet dependency levels of intermediate students were evaluated. The relation between internet dependency point (dependent variable) and daily hours of internet use (independent variable) were examined and scatter plot with curve was obtained. In this study, in terms of different bandwidth values, some curves have been obtained and the differences of these curves have been argued.

Keywords: curve estimation, bandwidth, non-parametric distribution

INTRODUCTION

It is known that regression model is too crucial process for estimation of future. Essentially, regression methods are gathered in three main topics. These are parametric, non-parametric and semi-parametric regression methods (Härdle, Müller, Sperlich & Werwatz, 2004; Kaki, 2004). In estimation of curve methods, parametric methods are not always capable of obtaining information sufficiently. For all that, it is difficult to get detailed information about population with using non-parametric methods (Härdle ve Tsybakov, 1997). Although parametric and linear equation \((y = f(x))\) can make possible to conclude any estimation for each person, non-linear function can’t conclude regression equation for each one. Generalized Linear Model (GLM) was developed to estimate parameters robustly because of parametric linear regression method can’t explain variance sources efficiently. But, GLM is vulnerable to explain a non-linear situation which can not interpret attitude of individual in distribution curve (Kaki 2004; Muller, 2001). In case of interpreting any one’s situation in population, non-parametric methods can be used as an alternative to GLM. It must also considered that non-parametric methods are not enough powerful to analyze plenty of discontinuous variables. And it will be difficult to get a regression curve (Pinkse, 1994). In literature, curse of dimensionality is too crucial problem and it leads unreadable curve for population. Some smoothing methods are used to solve the unreadable curve. By this way, it will be possible to reduce curse of dimensionality. Especially, using semi-parametric methods can obtain the most clarity curves with using partial linear models. Also, it can rescue the complexity by using Kernel smoothing method (Pinkse, 1994; Muller, 2001). The other smoothing method is Spline methods which has different process from Kernel. Kernel smoothing method is so appropriate for distribution types such as Gaussian, uniform and Quartic. Spline smoothing is preferred for two or more level polynomial cases. Purpose of Kernel and Spline methods are to control the error of variance. Controlling variance error can make easy the readable of curves.

\(^1\) This study was presented in X.Biostatistic Congress as a summary oral paper.
In this study, Kernel which is one of the smoothing methods was investigated with its mathematical features. It was also aimed to show Kernel’s capabilities on reading and understanding distribution curves. The choice of bandwidth was also underlined to get optimal distribution.

**MATERIAL AND METHOD**

To show Kernel smoothing process, internet dependency levels of intermediate students’ were evaluated. Data set was obtained for determining the regression curve. This data set was consisted by Gunuc & Kayri (2009) to develop an Internet Addiction Scale. The aim of using this data set is not identify the dependency level. The single aim of using this data set is just to present how Kernel function is capable of reducing the curve complexity according as choosing appropriate bandwidth. The relation between internet dependency point (dependent variable) and daily hours of internet use (independent variable) were examined and scatter plot with curve was obtained. In this study, the bandwidth (h) value was chosen as 0.01, 0.10, 0.50 and 1.00. According to these values, the differences of regression curves were compared with each other. Also with same bandwidth value, Normal, Uniform and Epanechnikov distribution curves were compared. SPSS was used for all analysis.

**Kernel Method**

Kernel function estimates choosing of origin point a bin grid on curve. To make estimation, it calculates \( f(x) \) function:

\[
\hat{f}(x) = \frac{1}{2hn} \sum_{i=1}^{n} \left\{ X_i \in [x-h, x+h] \right\}
\]  

(1)

Here, 2hn interval length and \([x-h, x+h]\) are handled for distribution type (Härdle, Müller, Sperlich & Werwatz, 2004). Kernel function has alteration according to distribution type. If distribution is uniform Kernel function will change. Taking account of uniform distribution, Kernel function is:

\[
K(u) = \frac{1}{2} I(|u| \leq 1)
\]

(2)

Here, \(u=(X-X_i)/h\) and \(I\) is an identity matrix. In case of examining equation 2, it will be seen that \(1/2\) weighted additions will be added for each observed values in uniform distribution. Kernel smoothing process for a uniform distribution will carry out a probability density function. This density function can be showed in equation 3.

\[
\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{X-X_i}{h} \right) = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} I \left( \left| \frac{X-X_i}{h} \right| \leq 1 \right)
\]

(3)

The different types of Kernel smoothing functions were summarized in Table 1 (Härdle, Müller, Sperlich & Werwatz, 2004). In general, a probability density function of Kernel can be calculated in equation 4 (Härdle, Müller, Sperlich & Werwatz, 2004):

\[
\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_n(X-X_i)
\]

(4)

Here, \(K_n(\bullet) = \frac{1}{h} K(\bullet/h)\). \(K(\bullet)\) is one of Kernel function which has been showed in Table 1.
Table 1. Kernel smoothing function related to the distribution types (Härdle, Müller, Sperlich & Werwatz, 2004).

<table>
<thead>
<tr>
<th>Kernel</th>
<th>$K(u)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>$\frac{1}{2} I(</td>
</tr>
<tr>
<td>Triangle</td>
<td>$(1-</td>
</tr>
<tr>
<td>Epanechnikov</td>
<td>$\frac{3}{4}(1-u^2)I(</td>
</tr>
<tr>
<td>Quartic (Biweight)</td>
<td>$\frac{15}{16}(1-u^2)^2I(</td>
</tr>
<tr>
<td>Triweight</td>
<td>$\frac{35}{32}(1-u^2)^3I(</td>
</tr>
<tr>
<td>Gaussian</td>
<td>$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$</td>
</tr>
<tr>
<td>Cosine</td>
<td>$\frac{\pi}{4} \cos\left(\frac{\pi}{2} u\right)I(</td>
</tr>
</tbody>
</table>

Value of $h$ shows the bandwidth of curve. Choosing of optimal bandwidth is too important topic for Kernel smoothing parameter (Rice, 1984; Sheather & Jones, 1991; Hall & Marron, 1991). According to bandwidth ($h$) value, the distribution curve shows changeability. Kernel function is capable of obtaining more readable and more interpretable curve if choosing of distribution (Gaussian, Quartic, Uniform) is appropriate. The main aim of optimal choosing of bandwidth is to reduce the mean square error (Doksum, Peterson & Samarov, 2000; Kaki, 2004). Choosing of optimal bandwidth can be obtained by two methods. These methods are “Plug-in” and “Cross-Validation” (Jones, Marron & Sheather, 1996). It is known that common method of bandwidth choosing is Cross-Validation (Rice & Silverman, 1991). The purposes of Cross-Validation method is to reduce mean square error and to examine transformation of smoothing function in a reliability border. For having more sensitive examination, Generalized Cross-Validation was developed. Generalized Cross-Validation method was showed in equation 5.

$$GCV(h) = n^{-1} \frac{RRS(h)}{\{1-n^{-1}trA(h)\}^2}$$  \hspace{1cm} (5)

In equation 5, $RRS$ shows sum of error, $n$ show sample size. It is known that $RRS$ is calculated as equation 6:

$$RRS = \sum_{i=1}^{n}(Y_i - m(x_i))^2$$  \hspace{1cm} (6)

In equation 6, $m(x_i)$ shows the smoothing function and $Y_i$ means the observation values.

RESULTS

Readable and interpretable regression curves are obtained with using different bandwidth values and Kernel density estimation was showed in Figure 1.
In case of scrutinizing Figure 1, if h value has a low bandwidth value, the curve is unreadable and the regression curve is so roughly. In the event of increasing h value, the distribution has more readable and smooth. But choosing inappropriate high bandwidth value, the variance of population will be hidden. Because of all these reasons, Kernel density function smooth at an optimal value of bandwidth. Contrast to high level value of h, the small value can produce curse of dimensionality. So that choosing of bandwidth value is considered too crucial. Considering distribution type, Cross-Validation algorithm can determine optimal bandwidth value. According to distribution type such as Gaussian, Uniform, Epanechnikov and Quartic, it will be obtained different regression curve. Each distribution has different curve and mathematical equation which has been showed in Table 1. In this study, It was also proved that Uniform, Normal and Epanechnikov distribution types have different curves. And this result has been showed in Figure 2.
Figure 2. Normal, Epanichnikov and Uniform curves with same bandwidth (h = 1.00).

Although bandwidth value is same for each distribution the regression curves are obtained different. In researches, the type of distribution must be considered and it must be determined before analyzing or getting any curve.

DISCUSSION

Although parametric regression methods are insufficient for parameter estimation non-parametric methods are detailed information about population unnecessarily. So that regression curves which belong to non-parametric methods are too roughly and it makes it difficult the readability and interpretability of curves (Pinkse, 1994; Müller, 2001). This problem is called as “curse of dimensionality”. Solving this problem depends on some smoothing methods which enhance the readability of curves (Härdle, Müller, Sperlich & Werwatz, 2004). Besides, determining distribution type of population and choosing of bandwidth must be taken account of process. It must be considered that smoothing function varies according to distribution type. Kernel function is one of the smoothing methods. Kernel function is capable of creating optimal curves according to appropriate distribution types. In this study, curse dimensionality was focused and Kernel function was scrutinized to fix this problem.

REFERENCES