COMPARISON OF B-SPLINE MODEL AND ITERATED CONDITIONAL MODES (ICM) FOR DATA WITH MEASUREMENT ERROR (ME)

Hartatik, Agus Purnomo
Informatics Engineering Department, Faculty of Mathematics and Natural Sciences , Sebelas Maret University, Surakarta, Indonesia
E-mail: hartatik.uns@gmail.com

Abstract. Direct observation results are often used to review the estimation model. However, actual data observation findings still need to be re-examined, because of measurement error factors (ME). In the regression modeling if X is a random variable with measurement error then the complicated calculation will not loose from application of Computer and Technology. As is the case for a review of the following model estimation, given data (Xi, Yi), then the regression model is

\[ Y_i = g(X_i) + \varepsilon_i \]

where \( X_i \) is the element i from the predictor variables X and \( Y_i \) is the element i of the response variable Y. The variable X is the predictor variables From the findings specific observations usually are constants, but generally found X which is a random variable variable or where Fixed value is not constant. In this case is called the regression model Regression Model with the measurement Errors. Purpose of this research are estimated nonparametric model approach with B-Spline Method to review regression with Measurement Errors are ignored and methods Iterative Conditional Mode (ICM) for review Model regression with measurement error.

Keywords: ICM, Bayesian, B-Spline, regression, measurement error, non-parametric,

1. Introduction

Data from the measurement or observation is often taken as a result of the finding of observations used to estimate the model estimates, and spline regression model. The results of the measurement data is generally still to be assessed, because of measurement error (ME) [1].

Many factors influence the measurement result data, including: tools, data retrieval process is correct, human factors, and nature factors. These factors can cause measurement results contain the measurement error (measurement error). But often, and most of researchers take and consider the data as a fixed value and believe that a data observation are correct, without checking how large measurement errors can be ignored [2]. The regression model when significant measurement error would be different if there is no measurement error. In equation (1) as below
\[ Y = g(X) + \varepsilon \tag{1} \]

where \( X \) is the predictor variables, the response variable \( Y \), \( g \) is a function Certain functions, and \( \varepsilon \) is a error random independent with mean zero and variance \( \sigma^2_\varepsilon \) [3].

The variable \( X \) is the predictor variables from the findings of the observation is usually assumed as a variable Fixed (fixed variables). But in reality, often encountered \( X \) is not fixed but variable random variable \( X \) measured with errors (errors in variables). But this error often overlooked for a Practical Reason and easy of calculation. However, if the error performed repeatedly and variable on the High Risk Level then the consequences would be fatal Against Decision. This cases can be found at epidemiological problems, geology, and survival.

Measurement error (ME) is an error appears when a recorded value exactly the same is not true value in the Measurement Process. Thus relates with this definition, there are 3 variables in Measurement Error Model, are

\[ W = X + U \tag{2} \]

by \( W \) is a variable that states observations called variable replacement (surrogate variable), \( X \) is the predictor variables for unobserved (latent variables), and \( U \) is the independent variable measurement errors are assumed normal with mean 0 and variance \( \sigma^2_u \).

Based on the regression model (1) and (2), there are two approaches were used to estimate the regression curve is Parametric and nonparametric regression methods. The purpose of this study was to perform steps regression model estimation using S-Plus software, namely the B-Spline method, and ICM [4].

2. Research Methods

This research is in the development of science, namely in the field of nonparametric statistics:
3. Comparison of nonparametric Models

3.1 B-SPLINE Model

Nonparametric regression model, if there is a spline of order $m$ and a collection of knots that meet $a < \xi_1 < ... < \xi_k < b$. Can be defined a number of $2m$ knot additional

$\xi_{-(m-1)},...,\xi_1, \xi_0, \xi_{k+1},...,\xi_{k+m}$

Where $\xi_{-(m-1)},...,\xi_0 = a$ dan $\xi_{k+1},...,\xi_{k+m} = b$.

B-spline order $m$ corresponding to knots $\xi_1,...,\xi_k$ expressed with the following models

$$B_{j,m}(X) = \frac{X - \xi_j}{\xi_{j+m} - \xi_j} B_{j,m-1}(X) + \frac{\xi_{j+m} - X}{\xi_{j+m} - \xi_{j+1}} B_{j+1,m-1}(X), \quad (3)$$

with $B_{j,1}(X) = \begin{cases} 1, X \in [\xi_j, \xi_{j+1}) \\ 0, \text{the others} \end{cases}$.

As example it is assumed that $\hat{g}(X) = \sum_{j=1}^p B_{j,m}(X)\beta_j$, where $B_1(X),...,B_p(X)$, is a basis for a vector spline of degree $m$ with a point knots $\xi_1,...,\xi_k$ and $p=m+k$. So,

$$S(g) = \sum_{i=1}^n \left( Y_i - \hat{g}(X_i) \right)^2 + \alpha \int_a^b \left( g^{(2)}(x) \right)^2 \, dx$$

Penalized Least Square estimator [5]so that could be written by

$$S(g) = (Y - B(X)\beta)\,^T\,(Y - B(X)\beta) + \alpha \beta^T D \beta, \quad (4)$$

with all elements- $ij$ of $D$ is

$$D = \int_a^b \left\{ B_i''(x)B_j''(x) \right\} \, dx.$$  

So that the probe is function $g$

$$\hat{g}(X) = \sum_{j=1}^p B_{j,m}(X)\beta_j, \quad p=m+k \quad (5)$$

with

$$\beta = \left( (B(X))^T B(X) + \alpha D \right)^{-1} (B(X))^T \, Y.$$  

Nonparametric regression model with B-spline method is applied to the data when the measurement errors are ignored and if the measurement error is not negligible [7]. To estimate the model when the measurement error ignored is by regressing between $W$ and $Y$, while for
nonparametric regression curve estimation when the measurement error is negligible average regressing between \( W \) and \( Y \) called Naive Methods.

3.2. ICM Method for Model ME

In estimating function \( g \) in measurement error models with ICM method is required prior information before determining the posterior distribution which then later used for iteration in the ICM method.

i. Selection Prior

Suppose the observation data are assumed normal distribution,

\[
Y \sim N(g(X_i), \sigma_Y^2)
\]

\[
W \sim N(X_i, \sigma_u^2)
\]

Prior to the distribution of \( X \) and \( g \) is used for measurement error models (1) are as follows

\[
X \sim N(\mu_x, \sigma_x^2)
\]  

(6)

Where \( \mu_x \) dan \( \sigma_x^2 \) is a constant determined.

As for the prior distribution \( g \) do approach "partially improper":

\[
p(g) \propto \exp\left(\frac{-\alpha}{2}(g^T \mathbf{K} g)\right),
\]

where for two matrices \( Q \) and \( R \), the matrix \( \mathbf{K} \) is defined as \( \mathbf{K} = QR^{-1}Q^T \)

The measures define a metric \( \mathbf{K} \):

- Define that: \( h_j = t_{i+1} - t_i \), \( i=1, \ldots, n-1 \)
- Determine the matrix \( A \) is the matrix size \( n \times (n-2) \) with \( q_{ij} \) are as follows

\[
q_{j-i,i} = h_{j-i}, q_{j,j} = -h_{j-i}^{-1}, q_{j+1,j} = h_j^{-1},
\]

For \( i=1, \ldots, n \), dan \( j=2, \ldots, n-1 \), and \( q_{ij} = 0 \) for \( |i-j| \geq 2 \).

ii. Writing In The Form Of Matrix:

So that the matrix \( Q \) can be written as follows:
\[
Q = \begin{bmatrix}
q_{12} & q_{13} & q_{14} & \cdots & q_{1,n-1} \\
q_{22} & q_{23} & q_{24} & \cdots & q_{2,n-1} \\
q_{32} & q_{33} & q_{34} & \cdots & q_{3,n-1} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
q_{n2} & q_{n3} & q_{n4} & \cdots & q_{n,n-1}
\end{bmatrix}
\]

\[
\begin{bmatrix}
\frac{1}{x_i - x_j} & 0 & 0 & 0 & \cdots & 0 & 0 \\
-\frac{1}{x_i - x_j} & \frac{1}{x_j - x_i} & 0 & 0 & \cdots & 0 & 0 \\
\frac{1}{x_i - x_j} & -\frac{1}{x_j - x_i} & \frac{1}{x_j - x_i} & 0 & \cdots & 0 & 0 \\
0 & -\frac{1}{x_j - x_i} & -\frac{1}{x_j - x_i} & \frac{1}{x_j - x_i} & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & \frac{1}{x_n - x_{n-1}} & \frac{1}{x_n - x_{n-1}}
\end{bmatrix}
\]

Subsequently determine the matrix \( R \) is a matrix of size \((n-2) \times (n-2)\) with elements for \( i \) and \( j \)

\[r_{ii} = \frac{1}{3}(h_{i-1} + h_i), \quad i = 2, \ldots, n-1\]

from 2 to \(n-1\) is

\[r_{i,i+1} = r_{i+1,i} = \frac{1}{6}(h_i), \quad i = 2, \ldots, n-1\]

and \( r_{ij} = 0 \) for \(|i - j| \geq 2\).

iii. **Writing in the form of Matrix:**

\( R \) can be written in matrix form as follows:

\[
R = \begin{bmatrix}
r_{22} & r_{23} & r_{24} & \cdots & r_{2,n-2} & r_{2,n-1} \\
r_{32} & r_{33} & r_{34} & \cdots & r_{3,n-2} & r_{3,n-1} \\
r_{42} & r_{43} & r_{44} & \cdots & r_{4,n-2} & r_{4,n-1} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
r_{n-1,2} & r_{n-1,3} & r_{n-1,4} & \cdots & r_{n-1,n-2} & r_{n-1,n-1}
\end{bmatrix}
\]
\[
R = \begin{bmatrix}
\frac{1}{3}(h_1 + h_2) & \frac{1}{6}h_2 & 0 & \cdots & 0 & 0 \\
\frac{1}{6}h_2 & \frac{1}{3}(h_1 + h_3) & \frac{1}{6}h_3 & \cdots & 0 & 0 \\
0 & \frac{1}{6}(h_3) & \frac{1}{3}(h_1 + h_4) & \cdots & 0 & 0 \\
0 & 0 & 0 & \cdots & \frac{1}{6}(h_{n-2}) & \frac{1}{3}(h_{n-1} + h_{n-1}) \\
\end{bmatrix}
\]

so matrix \textit{can be write as} \( K = QR^{-1}Q^T \).

Prior to variances in the ICM method is the result of the estimate, is as follows:

\[
\hat{\sigma}^2_{\varepsilon} = \frac{\sum_{i=1}^{n}(m_i - 1)s_i^2}{\sum_{i=1}^{n}(m_i - 1)},
\]

(7)

and a estimate of the error variance \( i \) is

\[
\sigma^2_{\varepsilon} = \frac{\sum_{i=1}^{n}[(Y_i - g(X))]^2}{tr[I - A(\hat{\alpha})]},
\]

(8)

with \( \hat{\alpha} \) estimated by the method of GCV.

\textbf{iv. Formation Density Posterior with ICM method}

Based on the prior (7), here in after defined posterior distribution of parameters in nonparametric regression. The parameters in the model (1) is \( X, g, \sigma^2_x, \sigma^2_\varepsilon, \text{dan} \sigma^2_{u} \). Suppose,

\[
\Theta = \left( X, g, \sigma^2_x, \sigma^2_\varepsilon, \sigma^2_{u} \right)
\]

maka based on the equation (6), the posterior density of \( \Theta \) is
The ICM method approach is used to estimate the parameters \( \theta \). Estimation of parameter \( \theta \) with ICM method of determining the posterior distribution of the above equation, which is similar to determining the Bayesian posterior to the method in which all of the parameters have a prior distribution. But in the ICM method, which is also known as Bayesian some, but not all the parameters have distribution, in this case \( \sigma_2 \sigma^2 \), \( \text{dan} \sigma^2_u \) estimation of data[8].

Thus conditional on \( \sigma^2_X, \sigma^2_u, \text{dan} \sigma^2_u \), the posterior distribution with ICM method is

\[
p(\theta|Y, W) \propto p(\theta)p(Y|\theta)p(W|\theta)
\]

\[
\propto p(X, w, \sigma^2_X, \sigma^2_u)p(Y|g, X, \sigma^2_\epsilon)p(W|X, \sigma^2_u)
\]

\[
\propto p(Y|g, X, \sigma^2_\epsilon)p(W|X, \sigma^2_u)p(X)p(g)p(\sigma^2_X)p(\sigma^2_u)p(\sigma^2_\epsilon)p(\alpha)
\]

ICM approach as described, by finding the posterior mode (9). In the ICM method parameters are updated (update) in each iteration. Unlike the fully Bayesian method, the method of Bayesian posterior of each parameter estimated from the full condition[9].

**Algoritma ICM**

1. **Initialize** \( \sigma^2_X, \sigma^2_\epsilon, \text{dan} \sigma^2_u \), and \( g^{(0)} \)

   \( g^{(0)} \) is estimated using Naïve method, which is regressing between the value of Y with an average of W. The initial value of \( \sigma^2_X, \sigma^2_\epsilon, \text{and} \sigma^2_u \) derived from the equation (6), (7), (8).

2. Based on one condition, determine the data conditional \( g^{(i-1)} \) maximizing (9)

   Value of \( X^{(i)} \) which gives the maximum value at (4:17) can not be solved analytically. Therefore, in this case, the method used to maximize the uniform grid (9).

3. Determining vector \( g^{(i)} \) conditional on \( X^{(i)} \) using Spline.
Based on the equation (9), with conditional on \( X^{(i)} \), maximizing the posterior density (9) as well as minimizing

\[
\left[ \frac{1}{2\hat{\sigma}^2_e} \sum_{i=1}^{n} (Y_i - g(X_i))^2 + \frac{\hat{\alpha}^T}{2\hat{\sigma}^2_e} g^T Kg \right].
\]

equation (8) is an equation smoothing spline. To that end, estimate \( g \) can be used B-Spline, with conditional on \( X^{(i)} \).

4. iteration \( i\rightarrow i+1 \) is repeated until a certain \( i \).

Simulations conducted using S-Plus for data model that represents the process in statistical data model is exponential and trigonometric[10]. Then performed simulations for each model 1 and model 2, with variations in the value of \( n = 25, 50, 100, 250, \sigma^2_e = 0.01^2, 0.1^2, 1^2, \), \( \sigma^2_u = 0.01^2, 0.1^2, 0.5^2 \) and \( 1^2 \), \( \mu_x = 0 \), dan \( \sigma^2_x = 1 \).

**Model 1:** Exponent Function

\[ g(x) = \exp(x), \]

where \( n=250, \sigma^2_e = 0.01^2, \sigma^2_u = 1^2, \mu_x = 0 \), and \( \sigma^2_x = 1 \).

And from the output S-plus, obtained spline regression curve and the MSE of the above simulation, as shown in Table 1 and Figure 1.

| Method  | MSE    |
|---------|--------|
| Non ME[1]   | 19.90525 |
| Non ME[2]   | 25.0184  |
| Naïve       | 12.80861 |
| ICM         | 4.231545 |

**Table 1.** Values MSE Exponential Function
Figure 1. Curve Estimation for Exponential Function.

Model 2: Trigonometric function

\[ g(x) = \frac{\sin(\pi x / 2)}{1 + 2x^2 \sin g(x) + 1}, \]

\[ n=250, \sigma_e^2 = 0.01^2, \sigma_u^2 = 0.1^2, \mu = 0, \text{ and } \sigma_x^2 = 1. \]

Based on the results of the output S-plus, spline regression curve obtained from the above simulation is as follows:

Figure 2 Estimation Curves for Trigonometry Functions
Table 2. Value MSE for Trigonometry Functions

| method  | MSE   |
|---------|-------|
| Non ME[1] | 0.158436 |
| Non ME[2] | 0.195413 |
| Naïve   | 0.147683 |
| ICM     | 0.030937 |

Based on simulation results for both models above, Figures 1 and 2 show the regression curves of the three methods, the regression curve of the data by ignoring any measurement error (with B-Spline), and then to estimate the model by not ignoring ME (naïve Method, and also with ICM). Judging from the pictures, the neglect of measurement error (model non-ME) effect on the regression curve. It is seen that the regression curve when the measurement error is ignored much of the actual curve. It appears also from Figure 1 and 2, with the regression curve.

ICM is the closest to the actual curve, g (x). In addition, with Naïve Method, also provide an estimate of the regression curve is closer to the curve g (x) compared with the curve when the measurement error is ignored. It provides information that Measurement Errors That fact should not be overlooked, but the measurement data should be re-examined before being used to estimate further review.

Likewise, seen from the results of MSE for both the simulation model, the neglect of measurement error (model non-ME) effect on the results of MSE. It is seen that the estimated regression curve when measurement errors are ignored far greater than the value of its MSE ME models with Methode Naive or ICM. It appears also from Tables 1 and 2, the estimated regression curve for the simulated data with ICM method provides the smallest MSE value than the B-Spline and Naïve methodes.

4. Conclusion

a. Based on the above discussion, it can be concluded that:

The greater $\sigma_{\hat{U}}$, indicating that the MSE differences between non ME and ICM getting bigger. Based on simulation results, MSE non ME (ignoring measurement error) can achieve four times greater than when there is a measurement error models were estimated by ICM, both for $n=25, 50, 100, \text{and 250}$. As for Naïve Method also shows the value of MSE is better than if the measurement error is negligible.
b. Based on simulation results also showed that the ICM method provides a better value than the Naïve Method.

5. References

[1] Bound, John, Brown, C, and Mathiowetz, N. (2001). “Measurement Error in Survey Data.” In Handbook of Econometrics, vol. 5, edited by James J. Heckman, chap. 59. Amsterdam, The Netherlands: Elsevier, 3705–3843.

[2] Blattman, Christopher, Jamison, T.C, Palicz, T.K, Rodrigues, K, and Sheridan, M. (2015). “Measuring the Measurement Error: A Method to Qualitatively Validate Survey Data.” NBER Working Paper Series # 21447.

[3] Fan, J. and Truong, Y. K. (1993), “Nonparametric regression with errors in variables”, Annals of Statistics, 21, 1900–25.

[4] Fuller, W. A and Hidiroglou, M. A (1976), “Regression Estimation After Correcting for Attenuation”, Journal of the American Statistical Association, 73, 99-169.

[5] Amemiya, Y and Fuller, W. A (1984), “Estimation for Multivariate Errors in Variable Model with Estimated Error Covariance Matrix”, Annals of Statistics, Annals of Statistics, 12, 497-509.

[6] Carroll, R. J., Maca, J. D. and Ruppert, D. (1999), “Nonparametric regression with errors in covariates”, Biometrika, 86, 541–554.

[7] Box, G. E. P. and Tiao, G. (1973), Bayesian Inference in Statistical Analysis, Addison–Wesley, London.

[8] Berry, S. A., Carroll, R. J. and Ruppert, D. (2002), “Bayesian smoothing and regression splines for measurement error problems”, Journal of the American Statistical Association, 9, 160–169.

[9] Besag, J. (1986). “On the Statistical Analysis of Dirty Pictures” (with Discussion), Journal of the Royal Statistical Society, Series B, 48, 259-279

[10] Hartatik (2012), Aplikasi S-Plus Dalam Analisa Regresi Guna Estimasi Model Regresi Untuk Data Dengan Kesalahan Pengukuran Sebagai Dasar Dalam Pengambilan Keputusan, Proceeding SNATI 2011, Yogyakarta.