Comparison of Curve Estimation of the Smoothing Spline Nonparametric Function Path Based on PLS and PWLS In Various Levels of Heteroscedasticity

Adji Achmad Rinaldo Fernandes¹,5,*, Benny Hutahayan², Solimun¹,5, Endang Arisoesilaningsih³, Indah Yanti⁴,5, Ani Budi Astuti¹,5, Nurjannah¹,5, Luthfatul Amaliana¹,5

¹Department of Statistics, Faculty of Mathematics and Natural Sciences, University of Brawijaya, Malang, Indonesia
²Department of Business Administration, Faculty of Administrative Sciences, University of Brawijaya, Malang, Indonesia
³Department of Biology, Faculty of Mathematics and Natural Sciences, University of Brawijaya, Malang, Indonesia
⁴Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Brawijaya, Malang, Indonesia
⁵Kelompok Kajian Unggulan Pemodelan Statistika di Bidang Manajemen, Faculty of Mathematics and Natural Sciences, University of Brawijaya, Malang, Indonesia

*Corresponding author: fernandes@ub.ac.id

Abstract. Linearity assumption that has not been fulfilled in the path analysis should use nonparametric approach. This research uses smoothing spline nonparametric path analysis with generated data where the condition of heteroscedasticity level measured through MAPD statistic will be applied to the data. The conditions are MAPD 0.01 – 0.20; 0.21 – 0.40; 0.41 – 0.60; 0.61 – 0.80; and 0.81 – 1.00. The purpose of this research is to determine the comparison of curve estimation of spline smoothing nonparametric path function on every level of heteroscedasticity category (DM) and without considering the heteroscedasticity (TM). The research results found that relative efficiency value of DM (PWLS) estimator with TM (PLS) that is always more than 1 for every heteroscedasticity level and every observation size. Thus, it was obtained better DM estimator (PWLS approach) comparing to TM (PLS).

Keywords: Nonparametric path, Smoothing Spline, Heteroscedasticity, PLS and PWLS.

1. Introduction
Regression analysis is statistical method used to determine the relationship between predictor variable against response variable [4]. In the regression analysis, if the linearity assumption is fulfilled and it can be described in the form of certain function, linear, quadratic polynomial, cubic polynomial, exponential etc. then a parametric approach can be done. Semi parametric approach is used if only half of the curve shape is known [5]. One of the ways to test this linearity is through regression Specification Error Test (RESET). Nonparametric can be used when the assumption is not fulfilled at
all. In conducting curve form estimation, nonparametric approach on the regression analysis has high flexibility [2].

Although regression analysis application is useful in various researches, it cannot facilitate the data in the form of more complex relationship that contains several response variables with pattern relationship between different variables. Therefore, path analysis is developed to be able to form a model with a relationship between variables described by directions. In the end, the model in the path analysis will be made based on path diagram.

Nonparametric approach can be applied on regression analysis as well as on the path analysis. There are many nonparametric approaches such as smoothing spline, truncated spline, kernel, polynomial, MARS, and Wavelet. Among those examples, smoothing spline has special characteristic namely good flexibility in adjusting. Hence, various forms of curve can be obtained based on different smoothing parameter [13]. Smoothening parameter has the role as controller of roughness or smoothness of the spline curve estimator. Researches on spline have been carried out by several researchers, including [14] who discussed smoothing spline ANOVA for exponential distribution families of Wisconsin epidemiological case studies of diabetic retinopathy patients, [2] on spline models with optimal knots and [8] on spline estimation for multiple response nonparametric regression longitudinal data multi-predictors.

In the analysis of nonparametric paths, the assumption of homoscedasticity should be fulfilled. Normal OLS estimation when studying parametric temporarily produces unbiased and consistent estimator. However, it is inefficient because it does not minimize the variety of errors [11]. Therefore, an alternative is needed to produce the best and efficient estimator. The alternative is Weighted Least Square (WLS) [12]. For handling heteroscedasticity cases, estimation of weighted spline is chosen by the Penalized Weighted Least Square (PWLS) approach. Several studies have been conducted in relation to PWLS among other [10], images of PLWS reconstruction for positron emission tomography and [9] on the comparison of autocorrelation in longitudinal data.

The level of heterogeneity of various systems is certainly different for different data. One of the statistics that measure this is Mean Absolute Percentage Deviation (MAPD). MAPD which has a value between 0 and 1 is a measure of the accuracy of a method to form an appropriate model by calculating the percentage of the average absolute value of data deviation [1]. Value variation of MAPD is observed by considering the heteroscedasticity (DM) or without considering heteroscedasticity (TM). The data to be used is simulation data considering that it is very difficult to get the form of data in accordance with the desired conditions. The data will later be generated according to different levels of heteroscedasticity.

2. Literature Review

2.1. Parametric Path Analysis
Path analysis is an extension of multiple regression analysis that allows more complex models based on path diagrams such as the example in Figure 1 below [3]:

![Figure 1. Simple Path Diagram](image)

The model formed from the path diagram in Figure 1 is:

\[
\begin{align*}
y_{li} &= \beta_{1i} x_i + \epsilon_{li} \\
y_{2i} &= \beta_{12} y_{li} + \epsilon_{2i}
\end{align*}
\]

(1)

Where \( y, \beta, x, \) and \( \epsilon \) are endogenous variables, parameters, predictor variables, and errors respectively.
2.2. Smoothing Spline Nonparametric Path Analysis

Functions formed in the smoothing spline nonparametric path analysis are [6]:

\[ f = Td + Vc \]  \hspace{1cm} (2)

**T Matrix:**

\[
T = \begin{bmatrix}
T_1 & 0 \\
0 & T_2 \end{bmatrix}
\]

\[
T_k = \begin{bmatrix}
\langle \eta_{k1}, \phi_{k1} \rangle & \cdots & \langle \eta_{km}, \phi_{km} \rangle \\
\vdots & \ddots & \vdots \\
\langle \eta_{km}, \phi_{k1} \rangle & \cdots & \langle \eta_{km}, \phi_{km} \rangle
\end{bmatrix}_{n \times m}
\]

\[
\langle \eta_{ki}, \phi_{k0} \rangle = \frac{x_i^{j-1}}{(j-1)!}
\]  \hspace{1cm} (3)

**V Matrix:**

\[
V = \begin{bmatrix}
V_1 & 0 \\
0 & V_2 \end{bmatrix}
\]

\[
V_k = \begin{bmatrix}
\langle \xi_{k1}, \xi_{k1} \rangle & \cdots & \langle \xi_{kn}, \xi_{kn} \rangle \\
\vdots & \ddots & \vdots \\
\langle \xi_{kn}, \xi_{k1} \rangle & \cdots & \langle \xi_{kn}, \xi_{kn} \rangle
\end{bmatrix}_{n \times n}
\]

\[
\langle \xi_{ki}, \xi_{ks} \rangle = \int_a^b \frac{(x_i - u)^{m-1}(x_j - u)^{m-1}}{((m-1)!)^2} du
\]

\[
(x_i - u)^{m-1} = \begin{cases} 
(x_i - u)^{m-1}, & x_i \geq u \\
0, & x_i < u
\end{cases}
\]  \hspace{1cm} (4)

Where \( m \) is the degree of polynomial spline which in this study is \( m = 2 \).

2.3. Penalized Weighted Least Square in Smoothing Spline Nonparametric Path Analysis

The PWLS estimation method is to minimize the function as shown in equation (5) [7]:

\[
\text{Min} \left( M^{-1}(y - f)^T (y - f) + \sum_{k=1}^s \lambda_k \left( \int f^{(m)}(x)^2 \right) dx \right)
\]  \hspace{1cm} (5)

The value of \( M \) in equation (5) above is \( 2n \). Optimization in equation (5) considers the goodness of fit (first syllable) and roughness penalty (second syllable) where \( \lambda_k \) is the smoothing parameter. The smoothing parameter value was determined using the minimized Generalized Cross Validation (GCV) method. The GCV value is shown by equation (6) [8]:

\[
\text{GCV} = \frac{M^{-1}y^T (I - A_j)^T (I - A_j) y}{(M^{-1} \text{trace}(I - A_j))^2}
\]  \hspace{1cm} (6)

The minimum optimization of equation (5) is obtained by the first partial derivative of the coefficient vectors \( d \) and \( c \). Equations (7) and (8) below show the coefficient vector which minimizes the function in equation (5).

**Vector \( d \):**
Vector $\zeta$:

$$
\hat{\zeta} = \hat{\Sigma}^{-1} \left( \mathbf{I} - \mathbf{T} \left( \mathbf{T}^T \hat{\Sigma}^{-1} \mathbf{T} \right)^{-1} \right) \mathbf{T}^T \hat{\Sigma}^{-1} \mathbf{y}
$$

Where:

$$
\mathbf{U} = \Sigma^{-1} \mathbf{V} + M \Lambda
$$

2.4. Level of Heteroscedasticity

The size of the heteroscedasticity level in one of the data is by looking at the MAPD value which stands for Mean Absolute Percentage Deviation. MAPD is obtained from the function of absolute difference in the square of the variance as explained in equation (10) below:

$$
MAPD = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\hat{\sigma}_i^2 - \sigma^2}{\sigma^2} \right|
$$

2.5. Relative Efficiency

The efficiency of the unbiased estimator is the ratio of the minimum range of errors of the estimator. Whereas the relative efficiency is the ratio of various errors of the two estimators compared. Suppose and are two function estimators that follow the general conditions of Cramer-Rao then Relative Efficiency (ER) of and are defined as the ratio of the various errors as follows:

$$
ER(\hat{f}_{DM}, \hat{f}_{TM}) = \frac{KTG_{TM}}{KTG_{DM}}
$$

If $ER(\hat{f}_{DM}, \hat{f}_{TM}) > 1$, then $\hat{f}_{DM}$ estimator is more efficient than $\hat{f}_{TM}$ estimator [15].

3. Research Methodology

3.1. Data Sources

The data used in this study are simulation data that meet the following criteria:

1. Data is cross-section.
2. Scale ratio and there is no measurement model in it.
3. The level of heteroscedasticity was applied to data based on MAPD. These levels are MAPD 0.01–0.20; 0.21–0.40; 0.41–0.60; 0.61–0.80; and 0.81–1.00.
4. The observation size is 3 conditions, namely $n = 100$, 50, and 25.
5. Maximum iteration of 100 times and 30 replications.

3.2. Steps

The steps of this study are as follows:

1. Set the design time point for variable $x$ with the following formula:

$$
x_i = \frac{2i - 1}{2n}
$$

2. Obtain values of $f(x_i)$ and $f(y_{ni})$ based on the logit function

$$
f(x_i) = \frac{\exp(50 - 100x_i)}{1 + \exp(50 - 100x_i)}
$$
\[ f(y_i) = \frac{\exp(20 - 75y_i)}{1 + \exp(20 - 75y_i)} \]  

(14)

3. Generating MAPD values through uniform distribution \((a_i, b_i)\) where the limits of \(a_i\) and \(b_i\) are the limits of the level of heteroscedasticity

4. Equates the level of heteroscedasticity in errors of endogenous variables \(y_1\) and \(y_2\).

5. Generating random errors from multivariate normal distributions \(\epsilon \sim N_{2n}(0, \Sigma)\).

6. Establish EV (Error Variance) of 0.01 as a substituent \(\sigma^2\).

   a. Matrix \(\Sigma_{TM}\) is obtained by considering the existence of heteroscedasticity conditions, namely 
      \[ \text{diag}_i(\Sigma_{DM}) \sim U(c_i, d_i) \] 
      \(c_i\) and \(d_i\) are the lower and upper limits which are formulated as:
      \[ c_i = \sigma^2 \left(1 - \text{MAPD}\right) \quad \text{and} \quad d_i = \sigma^2 \left(1 + \text{MAPD}\right) \] 
   
   b. Matrix \(\Sigma_{TM}\) is a diagonal matrix whose diagonal element is EV.

7. Form a value from \(y_1\) and \(y_2\) based \(y_{1i} = f(x_i) + \epsilon_{1i}\) and \(y_{2i} = f(y_{1i}) + \epsilon_{2i}\).

8. Implement a nonparametric smoothing spline PWLS (DM) and PLS (TM) approach

9. Calculate the relative efficiency value (EV) of the DM and TM estimating functions through the coefficient of determination

4. Result and Discussion

4.1. Comparison of Estimates of Smoothing Spline Nonparametric Path Functions

The following is an example of a comparison of the estimated curve of the smoothing spline nonparametric path function at the level of moderate heteroscedasticity (MAPD 0.41–0.60):

![Figure 2. Path Curve on Moderate Level of Heteroscedasticity](image)
Figure 2 in addition shows the estimation of the smoothing spline path curve at the level of moderate heteroscedasticity (MAPD 0.41–0.60). There is little difference in the shape of the estimation curve between the DM and TM approaches. The DM estimation curve looks closer to the data pattern comparing to the TM with not so visible difference.

The equation of the function of the path estimator using the DM smoothing spline nonparametric approach is shown in the following equation:

\[
f(x_i) = 1.0589 - 2.2066x_i - 144.519\left( \frac{1}{2}(x_i + x_i) + \frac{1}{3} \right) + ... - 1225.46\left( x_{100}x_i - \frac{1}{2}(x_{100} + x_i) + \frac{1}{3} \right)
\]

\[f(y_{i,j}) = 0.4852 - 2.6159y_{i,j} - 2444.82\left( \frac{1}{2}(y_{i,j} + y_{i,j}) + \frac{1}{3} \right) + ... + 364.58\left( y_{100i,j} - \frac{1}{2}(y_{100i,j} + y_{i,j}) + \frac{1}{3} \right)
\]

Whereas in the TM approach, the equations obtained are as follows:

\[
f(x_i) = 1.0159 - 0.0664x_i + 29.7593\left( x_i x_i - \frac{1}{2}(x_i + x_i) + \frac{1}{3} \right) + ...
\]

\[f(y_{i,j}) = 0.9956 + 0.0666y_{i,j} - 15.353\left( y_{i,j} y_{i,j} - \frac{1}{2}(y_{i,j} + y_{i,j}) + \frac{1}{3} \right) + ...
\]

The goodness of the model measured by the coefficient of determination \(R^2\) is obtained by the PWLS (DM) and PLS (TM) approaches which are 95.35% and 92.49%. The \(R^2\) value of the higher PWLS approach proves that the estimation with the DM approach heteroscedasticity is able to give better model.

4.2. Relative Efficiency of DM and TM Functions

Relative efficiency (ER) is obtained by dividing the value of the alleged variety in the DM and TM estimating functions. The following is a plot of the value of the relative efficiency of the DM function estimator against TM in each observation measure:
Based on Figure 3, it is found that the relative efficiency value of DM to TM is always more than 1 for all levels of diversity and all measures of observation. Thus, it is found that DM predictors or PWLS approaches are better at estimating data patterns compared to using TM estimators or PLS approaches.

5. Conclusion
Smoothing spline nonparametric estimators with DM approach are proved to be better at estimating data patterns at all levels of heteroscedasticity. It can be proven through better DM estimation curve pattern which has a relatively small distance to the actual point of the data. In addition, the average value of the goodness of the model of the coefficient of determination shows higher value. It means that the diversity of data is better explained in the DM estimator compared to the TM estimator.

Calculation of the value of the relative efficiency of DM to TM always results in an ER value of more than 1 for all levels of diversity and all measures of observation. Thus, it is found that DM predictors or PWLS approaches are better at estimating data patterns compared to using TM estimators or PLS approaches.

Based on the evaluation of the research that has been done, the suggestions that can be given for further research are analysing path diagram that can be able to coordinate direct and indirect influence simultaneously, combining the level of heteroscedasticity in each endogenous variable, simulation for varying levels of diversity and minimize the tolerance limit in the iteration of determining smoothing and GCV parameter values.

References
[1] Brandimarte, P. (2011), Quantitative Methods an Introduction for Business Management, New Jersey: John Wiley and Sons, Inc.
[2] Budiantara, I.N. (2006), “Model Spline dengan knot optimal”. Jurnal Ilmu Dasar Vol 7, No 2, hal.77-85.
[3] Dillon, W.R. dan Goldstein, M. (1984), Multivariate Analysis Methods and Application. New York: John Wiley and Sons
[4] Draper, N. R., and Smith, H. (1998), Applied Regression Analysis 3rd Edition. New York: John Wiley & Sons, Inc.
[5] Eubank, R.L. (1999), Nonparametric Regression and Spline Smoothing 2nd Edition, New York: Marcel Decker Inc.
[6] Fernandes, A.A.R., Budiantara, I.N., I.,Otok, B.W.,Suhartono. (2014), “Spline Estimator for Bi-responses Nonparametric Regression Model for Longitudinal Data”, Applied Mathematical Sciences, Vol 8 No 114, pp 5653-5665.
[7] Fernandes, A.A.R., Budiantara, I.N., I.,Otok, B.W., Suhartono. (2014), “Reproducing Kernel Hilbert space for penalized regression multi-predictors: Case in longitudinal data”,
International Journal of Mathematical Analysis, Vol 8 No 40, pp 1951-1961.

[8] Fernandes, A. A. R., Budiantara, I.N., Otok, B.W. dan Suhartono. (2015), “Spline Estimator for Bi-Responses and Multi-Predictors Nonparametric Regression Model in Case of Longitudinal Data”, Journal of Mathematics and Statistics, Vol 11, No 2, 2015, hal. 61-69.

[9] Fernandes, A. A. R., Janssen, P., Sa’adah, U., Solimun, Efendi, A., Nurjannah, Amaliana, L. (2017), “Comparison of Spline Estimator at Various Levels of Autocorrelation in Smoothing Spline Nonparametric Regression for Longitudinal Data”, Communications in Statistics – Theory and Methods, Vol 46, No 24, pp 12401-12424.

[10] Fessler, J.A. (1994), “Penalized Weighted Least-Squares Image Reconstruction for Positron Emission Tomography”, IEEE Transactions on Medical Imaging, Vol. 13, No. 2, hal. 290-300.

[11] Gujarati, D. (2004), Basic Econometrics 4th Edition. New York: The McGraw-Hill Companies.

[12] Kutner, M.H., Nachtsheim, C.J., Neter, J.& Li W. (2005), Applied Linear Statistical Models 5th Edition. Boston: McGraw-Hill International.

[13] Takezawa, K. (2006), Introduction to Nonparametric Regression. USA:John Wiley and Sons Inc.

[14] Wahba, G., et al. (1995), “Smoothing Spline ANOVA for Exponential Families with Application to The Wincosin Epidemiology Study of Diabetic Renitopathy”, The Annals of Statistics Vol.23(6):1865-1895.

[15] Zacks, S. (1981), Parametric Statistical Inference: Basic Theory and Modern Approaches, Pergamonn Press Inc., New York.