Modelling the percentage of poverty based on an open unemployment rate using some nonparametric regression techniques

Y Andriyana¹, J Suprijadi², Y Suparman³, S Winarni⁴, I G N M Jaya⁵

¹,²,³,⁴,⁵ Statistics Department, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran

Corresponding author: y.andriyana@unpad.ac.id

Abstract. Poverty is a basic problem to be solved happened in all countries, including Indonesia. Although the economic development of Indonesia is increasing from time to time, but the population is also highly increasing. The population grows rapidly which is not followed by the same Gross Domestic Product development. The unbalance situation implies the gap of social-economic conditions. In order to reduce the poverty rate, we need to build a model showing the influence of some factors to the poverty rate. Since, the open unemployment rate plays an important role to the poverty rate, then in this study, we will focus only on modelling the open unemployment rate to the poverty in Indonesia. Due to the characteristic of the scatter plot of the data, where the pattern is not easy to model parametrically, we then propose some nonparametric regression techniques. We present some results using nonparametric approach, such as, kernel and splines techniques. For the kernel approach we propose to use Nadaraya-Watson technique and local polynomial approach, and for the spline techniques we propose to apply Smoothing Splines, B-Splines and P-Splines techniques. The estimation curves shows that all nonparametric approaches fit nicely to the data.

1. Introduction
There are several important things influenced the poverty such as macro, sectoral, community, household, or individual characteristics. At macro characteristic, the poverty is commonly caused by regional economic potential, inflation, etc. Regarding to the sectoral characteristic, the poverty may cause by an unemployment rate, education level, or the contribution of the primary sector on the economic development. For the community aspect, it causes by the infrastructure such as clean water, electricity, and transportation. In the individual and household aspect, the poverty can be affected by the number of family member, mean year schooling, and their livelihood.

Focusing on the unemployment rate, there are already some studies to evaluate the influence of its rate to the poverty on an area. There exists a strong correlation between the poverty and the rate of unemployment. Indonesia itself, the unemployment is a huge problem which is not so easy to be overcome. The number of inhabitant grows rapidly implying the need of a fast development of job vacancies. An unbalance situation is appearing from time to time. Moreover, in some situations, the expertise of the employees is not satisfy with the market needs.
Some studies investigating the poverty can be seen in some references, such as, Kakwani and Pernia [8], Perry et al. [10], Warr [12], among others. Special for investigating the influence of unemployment to the poverty, it has been done by, for example, De Fina and Robert [2] clearly mentioned that the unemployment gives an effect to the alternative of poverty measures.

2. Poverty and Unemployment

In this section, we will briefly discuss about some short descriptions and definitions related to poverty and open unemployment rate.

2.1. Poverty

The percentage of poverty in Indonesia was firstly measured by Indonesian Statistical Bureau (BPS) in 1984. At that time, the measurement was taken for period 1976 – 1981 using consumption module of National Economic-Social Survey (SUSENAS). Since then, the percentage of poverty is measured every three years. Since 2003, Indonesian Statistical Bureau publishes its percentage yearly because started form that year, BPS collects the SUSENAS data of consumption panel every February or March.

The concept of poverty is based on the ability to fulfil a basic need, either food or non-food needs. This ability is shown by the poverty line (GK) which is also divided by food and non-food poverty line. The BPS. The poverty line itself is defined as the number of Indonesian Rupiah ( IDR) has to be spent by the people to get 2100 kkal/day per capita for food added by the minimum amount of money spent for non-food aspect, such as, housing, education, and health. Its minimum IDR is obtained from the mean expenditure per day per people. Hence the poor people is defined as the one who has smaller per capita expenses than the poverty line, either for food or non-food aspects.

There are three indicators of poverty, as follows;
(1) Head Count Index (HCI-P0), defined as the percentage of poor people under the poverty line,
(2) Poverty Gap Index (P1), showing the mean of the expenditure gap of each people to the poverty line,
(3) Poverty Severity Index (P2), showing the distribution of expenditure between poor people.

Foster, et al. [5] proposed a formulation to measure the poverty given in expression (1),

\[ P_\alpha = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{z - Y_i}{z} \right)^{\alpha} \]  

(1)

where,
\[ \alpha = 0 \ (HCI-P_0), 1 \ (P_1), 2 \ (P_2) \]
\[ z = \text{Poverty line} \]
\[ Y_i = \text{Mean of per capita expenditure for the people who are below the poverty line} \]
\[ q = \text{The number of people below the poverty line} \]
\[ n = \text{Population} \]

2.2. Unemployment

An unemployment is the number of employee which does not have any job or is actively trying to find a job. The International Labour Organization (ILO) gives the following definition;
(1) Open unemployment, given for the people who enter the age of working but they do not have a job,
(2) Disguised unemployment, one may happened because there are so many employee of a specific field but in fact reducing the employee will not influence the production number. It caused by the expertise of the employee is not satisfy,
(3) Under unemployment, assigned to the employee who are not working optimally because there is no position for them in a specific time.

In our study, we will only focus on the first criteria, open unemployment.
3. Nonparametric Regression
A parametric regression technique is used when we can easily specify the pattern of a scatter plot between the response and the covariate. For example, we choose a linear modelling if the scatter plot behave in a linear way. However, there are some situations where the scatter plot of the data is very hard to specify parametrically. We then end up with a nonparametric modelling setting [6]. Let \( Y \) is a response variable and the predictor is denoted by \( X \). We have a set of data \((X_i, Y_i), i=1, 2, \ldots, n\). Suppose now we have the following model;

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

where \( m(X) \) is an unknown smooth function to be estimated and \( \varepsilon_i \) is the error term of the model.

3.1. Kernel approach (Nadaraya – Watson Estimator)
In the kernel approach, the unknown smooth function \( m(X) \) is approximated by a kernel function via the following expression;

\[
\hat{m}(X) = \frac{\sum_{i=1}^{n} K_h(X_i-x)Y_i}{\sum_{i=1}^{n} K_h(X_i-x)}
\]  

with \( K_h(\cdot) \) is a rescaled kernel function and \( h \) is a bandwidth that need to be optimized.

3.2. Local polynomial estimator
The local polynomial estimator is used to estimate \( m(X) \) via a polynomial model as follows,

\[
Y_i = \beta_0 + \beta_1(X_i - x) + \beta_2(X_i - x)^2 + \cdots + \beta_p(X_i - x)^p + \varepsilon_i
\]

The least squared estimator of the above expression is given by [4],

\[
\sum_{i=1}^{n} (Y_i - \sum_{k=0}^{p} \beta_k(X_i-x)^kK_h(X_i-x))^2 = \min_{[\beta_0, \ldots, \beta_p]}
\]

Then, the estimator of the coefficient is then given by,

\[
\hat{\beta} = (X^TWX)^{-1}X^TWF
\]

where \( X^TWX \) is invertible, \( W=diag(W_{ii}) \) with \( W_{ii} = K_h(X_i-x) \), \( Y = (Y_1, \ldots, Y_n)^T \) and

\[
X = \begin{pmatrix}
1 & (X_1 - 1) & \cdots & (X_1 - x)^p \\
\vdots & \ddots & \vdots & \vdots \\
1 & (X_n - x) & \cdots & (X_n - x)^p
\end{pmatrix}
\]

3.3. Splines approaches
There are several alternatives of splines approach. The most classical technique is smoothing splines (see for example in [9] and [11]). Recall nonparametric model (2). Now, we estimate the unknown function \( m(X) \) by minimizing the following objective function;

\[
\sum_{i=1}^{n} (Y_i - m(X_i))^2 + \lambda \int (m(X))^2 \, dx
\]
where $0 < \lambda < 1$ is the roughness penalty to control the trade of between the fidelity and penalty part.

Differently, Eilers and Marx [3] proposed to use a P-splines approach. This approach is the development of B-splines where we need to put a penalty term to the adjacent coefficients of the basis B-splines as can be seen in [3] and [7] among others. We firstly need to approximate the function $m(X)$ by a linear combination of the basis B-splines,

$$m(X) \approx \sum_{j=1}^{m} \alpha_j B_j(X; v)$$

(8)

where $v$ is the degree of splines, $m$ relates to the number of knot to be optimized, and

$$B_j(x; v) = \frac{x-t_j}{t_{j+v-1}-t_j} B_j(x; v-1) + \left(1 - \frac{x-t_{j+1}}{t_{j+v}-t_{j+1}}\right)$$

(9)

with

$$B_j(x; 0) = \begin{cases} 1 & \text{if } t_j \leq x \leq t_{j+1} \\ 0 & \text{otherwise} \end{cases}$$

The B-splines objective function is then given by the following expression,

$$\min_\alpha \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{m} \alpha_j B_j(X_i; v)\right)^2$$

(10)

If we are working on P-splines objective function, we need to add a penalty term into the above B-splines objective function. We then have the following P-splines objective function;

$$\min_\alpha \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{m} \alpha_j B_j(X_i; v)\right)^2 + \lambda \sum_{j=d+1}^{m} \left(\Delta^d \alpha_j\right)^2$$

(11)

where $\lambda > 0$ is the smoothing parameter and $\Delta^d$ is a difference operator expressed by

$$\Delta^d \alpha_j = \sum_{t=0}^{d} (-1)^t \binom{d}{t} \alpha_{j-t} \quad \text{for } d \in \mathbb{N}.$$

4. Data Analysis

In this section, we analyse the effect of open unemployment rate to the poverty in Indonesia. The data used in this analysis taken from Indonesia Statistical Bureau (www.bps.go.id). We firstly present the scatter plot of the data which given in Figure 1,

![Figure 1](scatter_plot.png)

**Figure 1.** Scatter plot of the open unemployment rate versus the percentage of poverty
As can be seen in Figure 1, the pattern seems to be a cubic function. However, we need to specify in a more flexible way. Sometimes, it is not easy to specify parametrically. However, there is no specific testing available to check whether or not we need to choose a parametric or nonparametric model. We can simply look at the pattern of the scatter plot. In Figure 1 we see that when the unemployment rate increasing until more or less 5%, the poverty increasing, but it decreases until the unemployment reach 9% and then one increases again which is in general a bit hard to model in a parametric way.

Now, we need to model in a nonparametric way as mentioned in the previous section. We propose five techniques with the following parameters; (a) Nadaraya-Watson uses Normal kernel with the bandwidth is optimized by AIC criteria, (b) Local polynomial uses polynomial of degree 2 and normal kernel, (c) Smoothing splines uses a fixed number of knots which is equal to 5, (d) B-splines uses fixed not points which is equal to 5 and uses quadratic splines, and (e) P-splines uses the same parameter as in B-splines where the smoothing parameter, $\lambda$, is optimized using SIC. The results are presented in Figure 2.
Figure 2. Estimated curves using (a) Nadaraya-Watson estimation, (b) Local Polynomial, (c) Smoothing splines, (d) B-Splines, (e) P-splines, and (f) Comparing all techniques.

From Figure 2, we see that the estimation curves have the characteristics which are grouped into the left ((a), (c), and (e)) and the right panel ((b) and (d)). The same behaviour shown in Nadaraya-Watson, Smoothing splines and (P-splines) techniques where the curves are more wiggly than the other two techniques (Local polynomial and B-splines). The underfitting situation is given in B-splines objective function. This is in fact caused by the property of the B-splines itself that the choice of the number of knot is very sensitive in determining the wiggliness of the curve. However, these all techniques may not really optimal since we do not do any optimization on some parts, for example, in all splines techniques, we fix the number of knots which is equal to 5. There are several techniques to get an optimal number of knots, such as SIC, BIC, CV, L-curve, etc. Noted that the use of such goodness of fit techniques (SIC or BIC) is not attended to compare the quality of the estimators, but as mentioned before that those techniques are used to get an optimal number of knots in case of splines approach (or optimal smoothing parameter for P-Splines approach). Similarly, they are also used to get an optimal bandwidth in case of Kernel and Local Polynomial approximations. Hence, the quality of each techniques are measured via the Root Mean Squared Errors (RMSE) presented in Table 1;

Table 1. Root Mean Squared Errors (RMSE) of the proposed techniques.

| Methods                | RMSE       |
|------------------------|------------|
| Nadaraya – Watson      | 5.740706   |
| Local Polynomial       | 11.32663   |
| Smoothing Splines      | 5.521042   |
| B-Splines              | 5.827160   |
| P-Splines              | 5.027192   |

According to Table 1, we can see that the P-Splines technique gives a smallest RMSE which means that the performance of this technique is better than others. However, in a modelling context, the smallest RMSE is not the only goal to have a good curve, some studies focussed also on the smoothing quality
of the curve. In some situations, we face with the fact that the curve tends to over-smooth or under-smooth. Therefore, in this study, we present all possible estimation curves and we can freely choose one of them based on the need of the researcher itself. As a suggestion, if we want to have good curves with respect to its squared errors, we then choose the P-Splines technique. On the other hand, if we want to have a smoother curve, then B-Splines techniques is able to answer the problem.

5. Acknowledgement

This work is supported by the Basic Research Grant of the Directorate General of the Higher Education, Indonesia (PDUPT, SP DIPA-042.06.1.401516/2018).

References

[1] Badan Pusat Statistik 2013 Keadaan Angkatan Kerja di Indonesia Agustus 2013 Jakarta
[2] De Fina and Robert H 2002 The Impact Of Unemployment On Alternatif Poverty Measures Working Paper Departement Reseach Federal Reserve Bank Of Filadelfia
[3] Eilers P HC and Marx B D 1996 Flexible smoothing using B-splines and penalized likelihood Statistical Science 11(2): 89-121
[4] Fan J and Gijbels I 1995 Local Polynomial Modelling and Its Applications. Chapmann and Hall.
[5] Foster J, Joel G; Erik T 1984 A class of decomposable poverty measures Econometrica 3. 52 (3): 761–766
[6] Hardle W 1990 Applied Nonparametric Regression Cambridge University Press New York.
[7] Hudoyo L P, Andriyana Y, and Handoko B 2017 Modeling relationship between mean years of schooling and household expenditure at Central Sulawesi using constrained B-splines (COBS) in quantile regression AIP Conference Proceedings 1827(1):020021
[8] Kakwani N and Pernia EM 2000 What is pro-poor growth? Asian Development Review 18:1, 1-16
[9] Mulyani S, Andriyana Y, and Sudartianto 2017 Modeling the human development index and the percentage of poor people using quantile smoothing splines AIP Conference Proceedings 1827(1):020007
[10] Perry G E, Arias O S, Lopez J H, Maloney W F, Serven L 2006. Poverty Reduction and Growth Virtuous and Vicious Circles. New York: World Bank
[11] Wahba G 1990 Spline Models For Observation Data SIAM Pensylvania
[12] Warr P 2006 Poverty and growth in southeast asia ASEAN Economic Bulletin 23: 279 – 302