JURNAL EMACS
[Engineering, MAthematics and Computer Science] Vol.6 No.3 September 2024: 229-235
e-ISSN: 2686-2573
DOI: 10.21512/emacsjournal.v6i3.12094

An Implementation of Ordinal Probit Regression Model on Factor Affecting East Java Human Development Index

Mohammad Dian Purnama

Department of Mathematics, Faculty of Mathematics and Natural Sciences,
University of Surabaya,
Surabaya, Indonesia 60231
mohammaddian.20053@mhs.unesa.ac.id

Correspondence: mohammaddian.20053@mhs.unesa.ac.id

Abstract – *An instrument for measuring human development, the Human Development Index (HDI) looks at how well human development has been achieved in relation to a few fundamental aspects of quality of life. In 2023, East Java's HDI showed an increase in the last three years with the latest value of 73.38. Despite the increase, East Java still has the lowest HDI in Java and Bali. This situation suggests the need for an in-depth analysis of the factors that influence HDI. This study aims to identify factors that contribute to HDI to formulate more appropriate policies in the future. The data used is the HDI of East Java in 2023 with ordinal categories. To analyze the ordinal data, the ordinal probit regression method was applied. The results show that the percentage of poor people has a significant influence on HDI. In addition, the classification accuracy of the model is obtained with a value of 50.5%, which indicates that the accuracy of the model in predicting HDI into the right category reaches 50.5%.*

Keywords: Human Development Index; Ordinal Probit Regression; Regression Model.

I. INTRODUCTION

Success in human development is a key indicator of national progress, in addition to high rates of economic growth. An instrument for measuring human development, the Human Development Index (HDI) looks at how well human development has been achieved in relation to a few fundamental aspects of quality of life. Four main components of data are used to calculate the HDI: average years of schooling, literacy rate, educational participation, and healthy life expectancy, which represent the health sector; average expenditure per capita, which indicates income; and the purchasing power of the community, which is measured by average expenditure per capita (Putri et.al, 2024).

Additionally, the Central Statistical Agency (BPS) (2021) emphasizes the significance of the HDI as a critical indicator in evaluating more general aspects of development, especially in initiatives to raise the standard of living for people. The HDI shows how people can use development's outcomes to get access to resources like money, healthcare, education, and other necessities.

East Java's HDI grew for three years in a row in 2023, with the most recent value being 73.38%. However, according to BPS (2023), East Java continues to be the province in Java and Bali with the lowest HDI. Effective policies are therefore required to raise East Java's HDI. The objective of this research is to offer a more comprehensive view for future policy planning. This study's identification of the variables influencing HDI in East Java is highly pertinent and significant in the given context. To maximize human development, a fuller understanding of the dynamics of that development will be gained by analysis of these components.

Previous HDI study was carried out by Haya (2024), who used a spatial regression approach to model HDI in Papua Province. Purnama & Sofro (2024) also used ordinal logistic regression to study the factors

influencing HDI in the province of East Java. An overview of the intricate and frequently connected correlations between different variables is given by both studies.

The HDI is classified into low, medium, high, and very high categories in addition to existing in nominal form (Prasetyoningrum & Sukmawati, 2018). Ordinal scales are these sets of groups that have a level or order. Ordinal data is one type of category that is frequently used in data analysis to help with comprehension and analysis. Similar to nominal data, ordinal data comprises categorical features, but the degree, order, or rank of the objects change (Sartika, 2010).

An adequate strategy is needed to comprehend the link between response variables—especially those with ordinal data—and predictor factors. Ordinal probit regression is regarded as an appropriate method for evaluating ordinal data in order to meet the study's goals.

II. METHODS

Ordinal probit regression is used in this research process. The purpose of this analysis is to determine the relationship between a response variable (Y) with ordered categories or ordinal values and two or more predictor variables (X).

Descriptive statistical analysis will be performed following the response variable's categorization to produce an overview of the data that will be utilized. The modeling of ordinal probit regression will then be used, and parameter significance testing will come next. There are two phases to this test: partial and simultaneous. Ultimately, the established model will be put into practice. Figure 1 provides an illustration of these experimental steps.

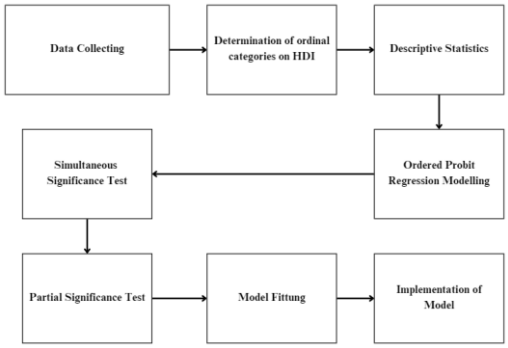

Figure 1. Steps of experiment

2.1 Data Acquisition

Secondary data from the Central Bureau of Statistics (BPS) of East Java Province's release page were used in this study. The 38 districts and cities in East Java Province for the year 2023 are included in the data analysis. The selection of 2023 is predicated on the most recent information, which encompasses data from 2023, for the predictor variables released by BPS East Java in 2024. The Human Development Index (HDI) is the response variable (Y) used in this study. It has been converted into ordinal data. The Central Bureau of Statistics has classified the HDI in the following Table I.

Table I. HDI Categories

Scale	Category
Moderate	$60 \leq \text{HDI} < 70$
High	$70 \leq \text{HDI} < 80$
Very High	$\text{HDI} \geq 80$

The predictor variables (x) used are economic growth (X_1), percentage of poor people (X_2), labor force participation rate (X_3), and gross regional domestic product (GRDP) per capita (X_4).

2.2 Descriptive Statistic

One way to examine data is through descriptive statistical analysis, which involves explaining the obtained data. The objective is to present a broad overview of the data using variables like mean, standard deviation, lowest and maximum values, and so on. Descriptive statistics facilitate the conversion of data into more comprehensible information and offer insight into the correlation between the study's answer variables (Anugrahayu & Azmi, 2023).

2.3 Ordinal Probit Regression

Regresi probit ordinal is a statistical technique used to create models from response variables with ordinal sifts, or variables with discrete categories within a given range. One type of non-linear regression that can be used to analyze the relationship between a response variable (nominal or ordinal with two categories) and/or a response variable (polynominal or ordinal with three categories) and one or more predictor variables is called a probit regression (Ruspriyanti & Sofro, 2018).

In this approach, the ordinal part of the response variable is determined by the cumulative energy that is derived by probit function. In an ordinal data context, the variable respons estimasi ke-i (Y_i) has an observational nilai (y_i), which is equivalent to r_i for categorical data display (Riadi & Kartikasari, 2020).

Ordinal probit regression modeling begins by considering the following model in equation 1:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad (1)$$

$\boldsymbol{\beta}$ is the coefficient parameter vector with $\boldsymbol{\beta} = [\beta_0, \beta_1, \dots, \beta_p]$, where p is the number of predictor variables, and $\boldsymbol{Y}$ is the discrete response variable. With $\boldsymbol{X} = [X_1, X_2, \dots, X_p]$, $\boldsymbol{\beta}$ is the predictor variable matrix, and $\boldsymbol{\varepsilon}$ is the error vector, which is presumed to be standard normally distributed N (0,1) (Febyanti, 2022).

Probit regression classifies $\boldsymbol{Y}$ in a binary fashion by providing a limit or threshold (α). This means that $Y \leq \alpha$ is classified as $Y = 0$, $Y \leq \alpha$ is classified as $Y = 1$, and $\alpha_{j-1} \leq Y \leq \alpha$ is classified as $Y = j$. As a result, the following is the model of ordinal probit regression obtained:

$$P(Y = 1) = \boldsymbol{\phi}[\boldsymbol{\alpha_1} - (\boldsymbol{\beta'X})] \quad (2)$$
$$P(Y = 2) = \boldsymbol{\phi}[\boldsymbol{\alpha_2} - (\boldsymbol{\beta'X})] - \boldsymbol{\phi}[\boldsymbol{\alpha_1} - (\boldsymbol{\beta'X})] \quad (3)$$
$$P(Y = i) = \boldsymbol{\phi}[\boldsymbol{\alpha_i} - (\boldsymbol{\beta'X})] - \boldsymbol{\phi}[\boldsymbol{\alpha_{i-1}} - (\boldsymbol{\beta'X})] \quad (4)$$
$$P(Y = k) = 1 - \boldsymbol{\phi}[\boldsymbol{\alpha_{c-1}} - (\boldsymbol{\beta'X})] \quad (5)$$

Marginal effects are used, according to Greene (2000), to interpret the ordinal probit regression model generated by equations (2) to (5). According to Ratnasari (2012), the marginal effect indicates the degree to which each significant predictor variable influences the probability of each category on the response variable.

$$\frac{\partial P(Y=1)}{\partial X} = -\boldsymbol{\beta}\{\boldsymbol{\phi}[\boldsymbol{\alpha_1} - (\boldsymbol{\beta'X})]\} \quad (6)$$
$$\frac{\partial P(Y=i)}{\partial \boldsymbol{X}} = \boldsymbol{\beta}\{\boldsymbol{\phi}[\boldsymbol{\alpha_i} - (\boldsymbol{\beta'X})] - \boldsymbol{\phi}[\boldsymbol{\alpha_{i-1}} - (\boldsymbol{\beta'X})]\} \quad (7)$$
$$\frac{\partial P(Y=k)}{\partial \boldsymbol{X}} = \boldsymbol{\beta}\{\boldsymbol{1} - \boldsymbol{\phi}[\boldsymbol{\alpha_{c-1}} - (\boldsymbol{\beta'X})]\} \quad (8)$$

2.4 Estimation of Parameters

Maximum Likelihood Estimation (MLE) is one technique used to estimate the parameters in the ordinal probit regression equation. This strategy maximizes the likelihood function to estimate the parameter β. The likelihood function's equation is as follows:

$$L(\beta) = \prod_{i=1}^{n} [p_1(x_i)]^{y_{1i}} [p_2(x_i)]^{y_{2i}} \dots [p_c(x_i)]^{y_{ci}} \quad (9)$$

Then the ln likelihood is performed, namely

$$ln\, L(.) = \sum_{i=1}^{n} \sum_{k=1}^{c} y_{ki} \ln p_k(x_i) \quad (10)$$

The next step is to derive the In-likelihood for ₩, which is:

$$\frac{\partial\, ln\, L(.)}{\partial \beta} = \frac{\partial}{\partial \beta} \sum_{i=1}^{n} \sum_{k=1}^{c} y_{ki} \ln p_k(x_i) \quad (11)$$
$$= \frac{\partial}{\partial \beta} \sum_{i=1}^{n} \sum_{k=1}^{c} y_{ki} \frac{1}{p_k(x_i)} \frac{\partial p_k(x_i)}{\partial \beta} \quad (12)$$

Based on the results of estimating the β parameter using the Maximum Likelihood Estimation (MLE) method, the function obtained is implicit, so the parameter estimation cannot be directly obtained. To obtain the parameter estimate, an iterative approach is used through the Newton-Raphson method. The steps to be taken in this method begin with determining the initial value of $\beta^{(0)}$, then calculating the initial values of $g^{(0)}$ and $H^{(0)}$ which depend on $\beta^{(0)}$. After that, the iteration starts from t=0, with the calculation of $\beta^{(t+1)} = \beta^{(t)} - \left[H^{(t)}\right]^{-1} g^{(t)}$ at each iteration step using the given formula. This process continues until a convergent value is reached, which is when $\left|\left|\beta^{(t+1)} - \beta^{(t)}\right|\right| \leq \varepsilon$. If convergence has not been achieved, the iteration continues by recalculating $g^{(t)}$ and $H^{(t)}$ until the convergence condition is met.

2.5 Simultaneous Parameter Significance Test

To determine whether all the parameter estimates in the model are significant overall, simultaneous significance testing is done. This test's objectives are to evaluate the suitability and significance of the employed model as well as the concurrent impact of the model's predictor variables. This test employs the chi-square test (Hosmer et al., 2013). The following is the hypothesis for concurrently testing the parameter coefficients' significance.

$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ or there are no predictor variables that have a significant effect on the model.

H_1 : There is at least one $\beta_j \neq 0$ with $j=1,2,\ldots,p$ or there is at least one predictor variable that has a significant effect.

test statistic formula:

$$G = -2 \ln \left[\frac{L_1}{L_2}\right] \quad (13)$$

With L_1 is likelihood function without independent variables and L_2 is likelihood function with independent variable. The G test statistic is based on the Chi-square distribution, in which the number of model parameters is represented by the degrees of freedom (db). If the G test statistic value is more than $\chi^2(\alpha, db)$ or if the p-value is less than α, with a significance level (α) of 0.05 or 5%, it is decided to reject H_0.

2.6 Partial Parameter Significance Test

The Wald test is performed when the likelihood ratio test shows results that reject H_0. This test aims to determine the effect of parameter β individually, with the hypothesis formulated as follows:

$H_0 = \beta_j = 0$

$H_1 = \beta_j \neq 0, j = 1,2,\ldots,p$

test statistic formula:

$$W_{count} = \frac{\hat{\beta}}{SE(\hat{\beta})} \quad (14)$$

It is necessary to reject H_0 if $W_{Count} > Z_\alpha$ or if the p-value is less than α, with α = 0.05 or 5% (Sofro et.al, 2019).

2.6 Pseudo R^2 McFadden

Pseudo R^2 McFadden was used to gauge the quality of the model. It is a commonly used criterion for selecting the best model when dealing with binary response variables. The two log-likelihood values that serve as the basis for this measurement are expressed in the following equation:

$$R^2_{Mc} = 1 - \frac{\ln L_M}{\ln L_0} \quad (15)$$

where R^2_{Mc} is the McFadden's coefficient of determination, L_M is the likelihood estimate for the model, and L_0 is the likelihood function for the model without predictors (Febyanti, 2022).

III. RESULTS AND DISCUSSION

Based on the results of the classical assumption test carried out, the following results were obtained:

3.1 Descriptive Statistic

In this study, the response variable is the human development index (HDI). The response variable data is categorical data with three ordinal categories category 1 states a moderate level, category 2 states a high level and category 3 states a very high level. The characteristics of the response variable can be seen in Figure 2 as follows:

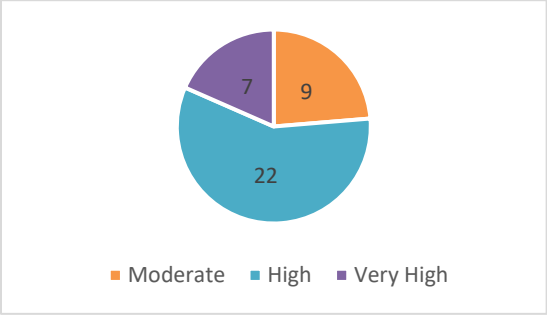

Figure 2. Characteristics of HDI

Descriptive data description on the predictor variable (X) can be seen in Table II as follows:

Table II. Calculation result of significance t-test

Variable	Average	Minimum	Maximum
x_1	4.71	1.20	6.19
x_2	10.29	3.31	21.76
x_3	73.16	66.89	81.64
x_4	70645.06	23842.1	541112.5

3.3 Ordinal Probit Regression Model

The following is displayed in Table III as the outcome of parameter estimate for ordinal logistic regression using Maximum Likelihood estimate (MLE).

Table III. Calculation result of significance t-test

Parameter	Est	Std Error	Wald	P-Value
α_1	-7.677	5.508	1.943	0.163
α_2	-4.323	5.393	0.643	0.423
β_1	-0.444	0.315	1.985	0.159
β_2	-0.348	0.105	10.909	0.001
β_3	-0.023	0.064	0.125	0.724
β_4	-0.000	0.000	3.282	0.070

Based on Table 3, the following model of regression probit ordinal can be developed:

$$P(Y = 1) = \phi[-7.677 - (C)]$$
$$P(Y = 2) = \phi[-4.323 - (C)] - \phi[-7.677 - (C)]$$
$$P(Y = 3) = 1 - \phi[-4.323 - (C)]$$

Where C is a probit function with the following equation.

$$C = -0.444X_1 - .0.348X_2 - 0.023X_3 - 0.000X_4$$

The number of categories in the response variable (Y) is reflected in the generated probit regression model equations. For the response variable categories (Y), the six equations are ordinal probit regression models with Y = 1 denoting the lowest category and Y = 3 denoting the highest. The marginal effects will be computed using the resulting regression model, aiding in the model's interpretation.
The probability value of a region falling into each HDI category is then calculated based on the three opportunity models that have been obtained. For instance, data from Ponorogo Regency, where X_1 through X_4 have values of 5.14, 9.53, 26313.90, and 75.88, respectively, are used in the calculation. After that, the values will be entered into the constructed ordinal probit regression model.

$P(Y = 1) = 0.19$
$P(Y = 2) = 0.80$
$P(Y = 3) = 0.10$

It is known that there is a greater chance of falling into category 2 than the other categories, with a probability value of 0.80, based on the HDI calculated for Ponorogo Regency. As a result, Ponorogo Regency in East Java can be classified as high or in HDI category 2.

3.4 Simultaneous Parameter Significance Test

The link between the predictor variables and the response variable is concurrently displayed using a simultaneous test, which is the next step in the ordinal logistic regression model. The G test statistic is used for this test, and the following hypothesis is applied:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ or there are no predictor variables that have a significant effect on the model.
H_1 : There is at least one $\beta_j \neq 0$ with *j=1,2,...,p* or there is at least one predictor variable that has a significant effect.

Table IV. Calculation result of significance t-test

Parameter	-2 Log Likelihood	Chi-Square	df	P-Value
Final	36.441	37.217	4	0.000

In Table IV, the χ^2 value of 36.441 is known to be more than the $\chi^2_{(3;0.05)}$ value of 9.488 based on the likelihood ratio test findings in the above table, or if based on the p-value, it is known that the p-value is 0.000, which is smaller than α (0.05). This indicates that at least one independent variable significantly affects HDI in East Java, rejecting the hypothesis H_0.

3.5 Partial Parameter Significance Test

By dividing the estimated value by the standard error, or as the Wald test statistic illustrates, the Wald test value in the partial parameter significance test is equal to the Z table value. It is necessary to reject H_0 if $W_{Count} > Z_\alpha$ or if the p-value is less than α, with α = 0.05 or 5%. The value of just the percentage of the impoverished population variable (X_2) has a p-value < 0.05, according to table 5's p-value. Thus, it may be stated that in East Java in 2023, the only variable that significantly affects HDI is the percentage of impoverished people (X_2).

3.6 Implementation of Model

Furthermore, the marginal effects displayed in equations (6) through (8) will provide the basis

for interpretation. This marginal effect shows how much the chance of a location falling into each HDI category is affected by the addition of one unit in the predictor variable. The following is the general model for the marginal effects of economic growth (x_1), percentage of poor people (x_2), labor force participation rate (x_3), and gross regional domestic product (GRDP) per capita (x_4). An example equation for the marginal impact of economic growth on the HDI in Ponorogo Regency is shown below.

$$\frac{\partial P(Y=1)}{\partial X_1} = 0.444\,\{\phi[-7.677-(C)]\}$$
$$\frac{\partial P(Y=2)}{\partial X_1} = -0.444\,\{\phi[-4.323-(C)]-\phi[-7.677-(C)]\}$$
$$\frac{\partial P(Y=3)}{\partial X_1} = -0.444\,\{1-\phi[-4.323-(C)]\}$$

The findings of calculating the marginal impact value in the first data for each predictor variable in the house price category are shown in Table V below for further information.

Table V. Calculation result of significance t-test

Variable	$\frac{\partial P(Y=1)}{\partial X_j}$	$\frac{\partial P(Y=2)}{\partial X_j}$	$\frac{\partial P(Y=3)}{\partial X_j}$
X_1	0.084	-0.355	-0.044
X_2	0.066	-0.278	-0.034
X_3	0.004	-0.018	-0.002
X_4	0.000	-0.000	-0.000

Table 5 indicates that the HDI tends to move toward the moderate range if economic growth (x_1), percentage of poor people (x_2), labor force participation rate (x_3), and gross regional domestic product (GRDP) per capita (x_4) picks up speed. Furthermore, equation (15) is used to calculate the classification accuracy value. The findings indicate that 50.5% accuracy in HDI classification is achieved. This indicates that the model can predict the HDI's classification into the correct category with a 50.5% accuracy rate.

IV. CONCLUSION

Based on the results and discussion, the probit regression model can be obtained as follows:

$$P(Y=1) = \phi[-7.677-(C)]$$
$$P(Y=2) = \phi[-4.323-(C)]-\phi[-7.677-(C)]$$
$$P(Y=3) = 1-\phi[-4.323-(C)]$$

Where C is a probit function with the following equation.

$$C = -0.444X_1 - .0.348X_2 - 0.023X_3 - 0.000X_4$$

Based on the results of the simultaneous parameter significance test, it may be concluded that one predictor variable significantly affects HDI in East Java if the p-value of 0.000 is less than α (0.05) or if the χ^2 value of 36.441 is more than the $\chi^2_{(3;0.05)}$ value of 9.488. Only the variable representing the percentage of the poor population (x_2) has a p-value less than 0.05, according to the partial parameter significance test. Thus, it may be stated that in East Java in 2023, the only variable that significantly affects HDI is the percentage of impoverished people (x_2).

REFERENCES

Anugrahayu, M., & Azmi, U. (2023). Stock Portfolio Optimization Using Mean-Variance and Mean Absolute Deviation Model Based On K-Medoids Clustering by Dynamic Time Warping. *Jurnal Matematika, Statistika dan Komputasi*, *20(1)*, 164-183.

Badan Pusat Statistika Indonesia. (2021). Indeks Pembangunan Manusia Indonesia. https://www.bps.go.id/id/publication/2021/04/30/8e777ce2d7570ced44197a37/indeks-pembangunan-manusia-2020.html

Badan Pusat Statistika Indonesia. (2023). Indeks Pembangunan Manusia menurut Provinsi. https://sumsel.bps.go.id/indicator/26/593/1/-metode-baru-indeks-pembangunan-manusia-menurut-provinsi-.html

Febyanti, F. (2022). Pemodelan faktor-faktor yang mempengaruhi harga rumah di Jabodetabek menggunakan metode regresi probit. *Jurnal Riset Statistika*, *2*(1), 50-56.

Greene, W. H. (2000). Econometric analysis 4th edition. *International edition, New Jersey: Prentice Hall*, 201-215.

Haya, A. (2024). PEMODELAN INDEKS PEMBANGUNAN MANUSIA DI PROVINSI PAPUA TAHUN 2022 MENGGUNAKAN ANALISIS REGRESI SPASIAL. *Media Edukasi Data Ilmiah dan Analisis (MEDIAN), 7(01),* 60-71.

Hosmer Jr, D.W., Lemeshow, S., dan Sturdivant, R. X., 2013. *Applied logistic regression, vol. 398*. John Wiley & Sons.

Prasetyoningrum, A. K., & Sukmawati, U. S. (2018). Analisis pengaruh Indeks Pembangunan Manusia (IPM), pertumbuhan ekonomi dan pengangguran terhadap kemiskinan di Indonesia. *Equilibrium: Jurnal Ekonomi Syariah, 6(2),* 217-240.

Purnama, M. D. (2024). Pemodelan Faktor-Faktor yang Mempengaruhi Indeks Pembangunan Manusia Jawa Timur dengan Regresi Logistik Ordinal. *MATHunesa: Jurnal Ilmiah Matematika, 12(3),* 654-661.

Putri, M. R., Ridla, M. A., & Azise, N. (2024). PENGARUH TINGKAT KEMISKINAN TINGKAT PENGANGGURAN UPAH MINIMUM KABUPATEN/KOTA DAN LAJU PERTUMBUHAN EKONOMI TERHADAP INDEKS PEMBANGUNAN MANUSIA DI PROVINSI RIAU. *Jurnal Cybernetic Inovatif, 8(6)*.

Ratnasari, V. (2012). Estimasi Parameter dan Uji Signifikansi Model Probit Bivariat. *Surabaya: Institut Teknologi Sepuluh Nopember*.

Riadi, R. A., dan Kartikasari, M. D., 2020. Implementasi k-means clustering dan regresi logistik ordinal terhadap kinerja cabang pt. x. *PROSIDING SENDIKA, 6(1)*.

Sartika, E., 2010. Pengolahan data berskala ordinal. *Sigma-Mu, 2(1),* 60–69.

Sofro, A., Oktaviarina, A., & Maulana, D. A. (2019). *Metode Statistika*. Surabaya, Indonesia: Unesa Press.