Application of Dummy Regression to Estimate the Income of the Working Population in East Lombok

Umam Hidayaturrahman[(1)], Basirun[(2)], Dita Septiana Ayundasari [(3)], Muh. Zulkarnain Alayyubi [(4)]

Program Studi Statistika, Fakultas MIPA, Universitas Hamzanwadi

Jl. TGKH M. Zainuddin Abdul Madjid No. 132 Pancor (83611) Selong-Lombok Timur-NTB
e-mail: hidayaturrohman0809@gmail.com[(1)] basirun@hamzanwadi.ac.id[(2)]
ertinasa.210304007@student.hamzanwadi.ac.id[(3)] efidaa80@gmail.com[(4)]

ABSTRAK

Penelitian ini dilaksanakan di Badan Pusat Statistik Kabupaten Lombok Timur pada periode 7 Oktober hingga 20 November 2024. Tujuan utama dari penelitian ini adalah untuk mengevaluasi berbagai faktor yang berkontribusi terhadap tingkat pendapatan penduduk di wilayah tersebut. Data yang dianalisis berasal dari data sekunder, yakni hasil Survei Angkatan Kerja Nasional (Sakernas). Fokus penelitian ini adalah untuk menilai sejauh mana variabel-variabel seperti jenis kelamin, usia, status pernikahan, tingkat pendidikan terakhir, total jam kerja per minggu, status dalam pekerjaan, klasifikasi bidang kerja, serta jenis pekerjaan memengaruhi penghasilan individu yang bekerja di Lombok Timur. Dari hasil analisis, diperoleh sebuah model regresi yang mampu memprediksi pendapatan masyarakat berdasarkan faktor-faktor tersebut. Model ini memiliki koefisien determinasi sebesar 0,747 atau 74,7%, yang mengindikasikan adanya hubungan yang cukup kuat antara variabel-variabel bebas dan pendapatan sebagai variabel terikat. Selain itu, ditemukan bahwa 52,1% variasi dalam tingkat pendapatan dapat dijelaskan oleh variabel-variabel dalam model, sementara sisanya sebesar 47,9% dipengaruhi oleh faktor lain yang tidak tercakup dalam penelitian ini. Temuan dari penelitian ini memberikan pemahaman lebih mendalam mengenai faktor-faktor penentu pendapatan di Lombok Timur, dan hasilnya berpotensi dijadikan dasar dalam perumusan kebijakan untuk mendorong peningkatan kesejahteraan serta menurunkan tingkat kemiskinan di wilayah tersebut.
Kata kunci: Sakernas, Model Regresi, Pendapatan, Kemiskinan

ABSTRACT

This research was carried out at the Central Bureau of Statistics in East Lombok between October 7 and November 20, 2024, aiming to explore the determinants of individual income within the region. The study utilized secondary data, specifically drawn from the National Labor Force Survey (Sakernas). It investigates how variables such as sex, age, marital status, level of education, working hours, job status, industry sector, and job type contribute to the income levels of workers in East Lombok. The analysis produced a regression model capable of estimating income based on the aforementioned factors. The model achieved a coefficient of determination of 0.747, suggesting a moderately strong relationship between the predictor variables and income as the outcome variable. Furthermore, the results indicate that 52.1% of income variation is explained by the variables included in the model, while the remaining 47.9% is attributable to external influences not captured in this study. Overall, this study offers valuable insight into the key factors shaping income in East Lombok and may serve as a useful reference for policymakers aiming to enhance community welfare and address poverty reduction in the area.
Keywords: *Sakernas, Regression Model, Income, Unemployment*

INTRODUCTION

Indonesia, as a vast archipelagic nation endowed with rich natural resources and a substantial population, faces ongoing social challenges. As recorded in the 2020 Population Census, the population reached 270.20 million. This continual demographic growth is intrinsically tied to a range of socio-economic problems, particularly poverty. Poverty refers to a condition in which individuals are unable to fulfill essential needs—such as food, clothing, and housing—due to a combination of internal and external contributing factors [1]

A key driver of poverty is the high rate of unemployment. Data from the Central Statistics Agency (BPS) indicates that 7.39 million people were unemployed out of a total labor force of 118.19 million. This disparity between workforce expansion and limited job availability intensifies socio-economic inequality. The issue is further compounded by the presence of approximately 360,000 unemployed university graduates [2].

Among the provinces experiencing persistent poverty is West Nusa Tenggara (NTB), despite its wealth of natural assets. In particular, East Lombok District continues to report a high poverty incidence. According to BPS East Lombok data from 2013, around 20.7% of the district's residents were living below the poverty line. A key contributor to this is the relatively low regional minimum wage (UMR) compared to other provinces, alongside factors such as gender, age, educational attainment, employment type, and economic sector, all of which significantly affect income levels in the area.

To examine the determinants of income, this study employs the dummy regression technique, which enables the incorporation of qualitative variables into quantitative analysis. Dummy variables act as proxies for categorical factors assumed to influence continuous outcomes—in this case, individual income [3]. The strength of this approach lies in its enhanced predictive accuracy and its ability to support more informed decision-making compared to conventional multiple regression models.

In light of these considerations, this research aims to thoroughly investigate the factors affecting income among East Lombok residents. By applying dummy regression, the study will identify key qualitative attributes—such as gender, age, educational background, employment status, and industry classification—that exert a significant impact on earnings. The findings are intended to inform policymakers in designing targeted interventions to alleviate poverty in the region

METHODS

Regression Analysis

Regression analysis is a widely utilized statistical method for examining the association between a dependent variable and one or more explanatory variables [4]. As noted by [5], this technique provides a mathematical framework for detecting and interpreting patterns in variable relationships, with the principal aim of forecasting or estimating the outcome variable based on known values of the predictors. Through this approach, researchers are able to assess the strength and structure of the relationship between variables of interest.

Essentially, regression analysis serves as a tool for modeling how one or more independent variables relate to a dependent variable. It facilitates insights into potential causal links among the variables. Nevertheless, the presence of a statistical relationship does not inherently imply a causal effect, and further investigation is needed to substantiate causality claims [6]. Muhyi et al. [7] suggest that a model represents a simplified abstraction of a complex real-world system. Given the intricacy of natural phenomena and the limitations of human perception and analytical tools, modeling enables the reduction of complexity into manageable forms—often based on past knowledge or assumptions about inter-variable relationships within a system.

The general form of a simple linear regression model, involving only one independent variable, is expressed as follows [8]:

$$Y_i = \beta_0 + \beta_1 x + \varepsilon \quad (1)$$

In this case :

Y_i : response variable

β_0 : arameter, which is the y-intercept

β_1 : arameter, which is the slope of the line

ε : random error component

x : independent variable (predictor of y)

To ensure accurate and reliable estimation of regression parameters, several key assumptions must be satisfied in simple regression analysis [9]. These include: linearity of the relationship, absence of autocorrelation, homoscedasticity (constant variance of errors), normal distribution of residuals, and no multicollinearity among variables.

The Kolmogorov-Smirnov (K-S) test is a non-parametric tool used to compare the cumulative distributions of two independent datasets, in order to assess whether they originate from the same underlying distribution. It also serves to evaluate the conformity of a sample distribution to a specified theoretical distribution, such as the normal distribution. One advantage of the K-S test lies in its simplicity and objectivity, especially when compared to graphical normality assessments [10]. This test works by identifying the maximum difference between the empirical cumulative distribution function of the observed data and the cumulative distribution function of the reference distribution. This difference, referred to as the Maximum Deviation (D), is evaluated in relation to the sample size (N). By comparing the computed D value to critical values in the K-S distribution table, researchers can determine whether the deviation is statistically significant [11].

Dummy Variabel

In the context of multiple regression analysis, independent variables are generally numerical in nature. However, in practical research settings, not all predictors are quantitative. Researchers frequently encounter explanatory variables that are categorical. To address this, a technique known as dummy regression—utilizing dummy variables—has been developed. These variables are employed to convert qualitative attributes (such as gender, ethnicity, religion, policy interventions, or contextual variations) into a numerical format. Essentially, dummy variables are used to represent categorical data that are believed to influence a continuous outcome variable [3].

A crucial aspect of regression involving categorical predictors is the coding system used for these variables. In such cases, mutually exclusive classification is fundamental—each observation must belong to only one category. For instance, an individual cannot simultaneously be classified as both male and female, nor can someone be coded as both a civil servant and an entrepreneur, even if they engage in both roles. This exclusivity underpins the structure of dummy coding.

Using binary coding (0 and 1), dummy variables are always dichotomous. A value of 1 indicates that the respondent belongs to a specific category, while a value of 0 indicates otherwise. This approach ensures that for each category, an individual is either classified as included (1) or excluded (0). The binary coding system can be conceptually likened to an on/off switch: a code of 1 turns a category “on” (indicating the presence of a trait or group membership), whereas a code of 0 switches it “off” (indicating its absence).

Dummy variables are also known as binary, proxy, dichotomous, or categorical variables. By design, a dummy variable assumes the value D = 1 for one group and D = 0 for the comparison group. Several general model forms involving dummy variables are as follows:

I. Y = a + bX + c D1 (**Intercept Dummy Model**) (2)

II. Y = a + bX + c (D1X) (**Slope Dummy Model)** (3)

III. Y = a + bX + c (D1X) + d D1 (**Combination**)` (4)

Each model reflects a different way dummy variables can modify either the intercept, the slope, or both, depending on category membership. These models enhance the flexibility of regression analysis when incorporating qualitative information into quantitative frameworks.

RESULT AND DISCUSSION

Descriptive Statistic

Table 1. Summary of Descriptive Statistics of Wage Variables in East Lombok

Wage	Range	Minimum	Maximum	Mean
Lombok Timur	4.380.000	120.000	4.500.000	1.084.252

Based on the data presented in the output above, the maximum monthly income of the people in East Lombok reaches Rp 4,500,000.00, while the minimum recorded income is Rp 120,000.00 per month. The average income of the East Lombok community is Rp 1,084,252.00 per month, which is lower than the Regional Minimum Wage (UMR) of West Nusa Tenggara Province, which stands at Rp 1,120,000.00 per month. Furthermore, when compared to the average income of Indonesians in general, which is Rp 2,236,045.00 per month, the income of the East Lombok community is also below the national average. This indicates an economic disparity in the region.

Dummy Regression Analysis for Estimating Wage Variables

Before conducting the dummy regression analysis, the first step was to assign levels to the categories of each independent variable with categorical data.

1. Gender

The dummy coding for the Gender variable was as follows:

Table 2. Dummy coding for Gender

Gender	DJK1
Male (1)	0

Female (2)	1

In each gender variable presented in the table above, there are two levels, consisting of one dummy variable (DJK1) and one level designated as the reference category (male). Each level is assigned a code (0,1), as explained below:

$$DJK1 = \begin{cases} 1, \text{if the level was Female} \\ 0, \text{if the level was not Female} \end{cases} \quad (5)$$

This means that when the category is "Female," the dummy variable (DJK1) is coded as 1, and when the category is not "Female" (i.e., Male), it is coded as 0.

2. Marital Status

Below is the dummy variable for marital status. The dummy coding for the marital status variable will be displayed in the table below:

Tabel 3. Dummy Coding for Marital Status

Marital Status	**Dstat_Kawin1**	**Dstat_kawin2**	**Dstat_kawin3**
Unmarried (1)	0	0	0
Married (2)	1	0	0
Divorced (3)	0	1	0
Widowed (4)	0	0	1

For marital status, as shown in the table above, there were four categories: unmarried, married, divorced, and widowed. Since there were four categories, three dummy variables were created: Dstat_kawin1, Dstat_kawin2, and Dstat_kawin3, with one category serving as the reference category, which in this case was unmarried. Each category was assigned a code (0,1) as follows:

$$Dstat\_kawin1 = \begin{cases} 1, \text{if the level was Married} \\ 0, \text{if the level was not Married} \end{cases} \quad (6)$$

In dummy Dstat_kawin1, the code 1 was assigned to all samples that were married, and code 0 was assigned to all samples that were not married.

$$Dstat\_kawin2 = \begin{cases} 1, \text{if the level was Divorced} \\ 0, \text{if the level was not Divorced} \end{cases} \quad (7)$$

In dummy Dstat_kawin2, the code 1 was assigned to all samples that were divorced, and code 0 was assigned to all samples that were not divorced.

$$Dstatus\_kawin3 = \begin{cases} 1, \text{if the level was Widowed} \\ 0, \text{if the level was not Widowed} \end{cases} \quad (8)$$

In dummy Dstat_kawin3, the code 1 was assigned to all samples that were widowed, and code 0 was assigned to all samples that were not widowed.

Additionally, for other variables such as Higher Education, Employment Status, Job Classification, and Type of Occupation, the same coding process was applied.

Normality Test

One-Sample Kolmogorov-Smirnov Test

		Unstandardized Residual
N		105
Normal Parameters[a]	Mean	.0000000
	Std. Deviation	2.93021631E2
Most Extreme Differences	Absolute	.056
	Positive	.056
	Negative	-.039
Kolmogorov-Smirnov Z		.572
Asymp. Sig. (2-tailed)		.899

a. Test distribution is Normal.

Figure 1. Kolmogorov-Smirnov normality test results for the data distribution

Referring to Figure 1, the Kolmogorov-Smirnov test for normality produced a p-value of 0.899, which exceeds the significance threshold of 0.05. Consequently, it can be inferred that the assumption of normality is satisfied, indicating that the data are normally distributed.

Heteroscedasticity Test

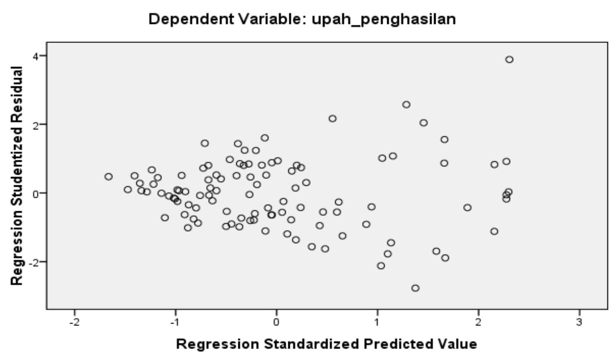

Figure 2. Results of the heteroscedasticity testn the regression residuals

Based on the scatter plot presented above, it can be observed that the data points are randomly and evenly dispersed around the zero line on the Y-axis, both above and below it. This random distribution of points suggests the absence of any discernible pattern or systematic variance in the residuals, indicating that heteroscedasticity is not present in the regression model. Consequently, this supports the validity of the model for prediction purposes, as the constant variance assumption is met.

Autocorelation Test

A robust regression model should be free from autocorrelation among residuals, meaning that the residuals must be independent of each other. One common method to detect the presence or absence of autocorrelation is the Run test, a non-parametric statistical procedure. The Run test evaluates whether the sequence of residuals exhibits randomness or shows systematic patterns indicating correlation. When residuals are uncorrelated, they are considered random, satisfying a key assumption of regression analysis. The Run test helps to statistically verify this randomness by analyzing the sequence of residuals for runs, or consecutive observations of similar signs. Below are the results of the Run test applied to the residual data, which provide insight into the presence or absence of autocorrelation in the model [9].

Runs Test

	Unstandardized Residual
Test Value[a]	-3410.28425
Cases < Test Value	52
Cases >= Test Value	53
Total Cases	105
Number of Runs	60
Z	1.276
Asymp. Sig. (2-tailed)	.202

a. Median

Figure 3. Results of the runs test for autocorrelation detection

Based on the results presented in the table above, the runs test produced a significance value of 0.202, which is higher than the conventional threshold of 0.05. This indicates that there is insufficient evidence to reject the null hypothesis of randomness in the residuals. Consequently, it can be concluded that no autocorrelation exists within the residual data, implying that the residuals are independent and the regression model meets this important assumption

Multicollinearity Test

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	1.602E6	530401.500		3.021	.003		
	Umur	19356.508	8672.345	.156	2.232	.028	.939	1.065
	Jam_kerja	26910.320	7750.981	.275	3.472	.001	.732	1.366
	DKBJI1	-1.610E6	588703.764	-.203	-2.734	.007	.835	1.197
	DKBJI2	-1.179E6	398498.223	-.250	-2.957	.004	.646	1.549
	DKBJI3	-2.468E6	347564.729	-.689	-7.102	.000	.490	2.040
	DKBJI4	-2.506E6	528842.399	-.363	-4.738	.000	.784	1.276
	DKBJI5	-2.593E6	304500.724	-.932	-8.516	.000	.385	2.600
	DKBJI6	-2.607E6	322211.982	-.817	-8.092	.000	.452	2.210

a. Dependent Variable: Upah

Figure 4. Results of the multicollinearity test based on tolerance and variance inflation factor (VIF) values

The output from the SPSS Coefficients table provides key indicators to detect multicollinearity among the independent variables, specifically through the Tolerance and

Variance Inflation Factor (VIF) statistics. According to established criteria, a Tolerance value below 0.10 suggests the presence of multicollinearity, while a VIF value exceeding 10 similarly indicates multicollinearity concerns [9]. Reviewing the results presented above, it is evident that all independent variables have Tolerance values well above the 0.10 threshold and none of the VIF values surpass the critical value of 10. These findings confirm that multicollinearity is not present in the regression model, ensuring that the estimates of the independent variables are reliable and not distorted by intercorrelations.

Multiple Regression Analysis

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.747a	.557	.521	918500.630

a. Predictors: (Constant), DKBJI6, DKBJI1, DKBJI4, DKBJI2, Jam_kerja, Umur, DKBJI3, DKBJI5

Figure 5. Results of the multiple regression analysis

Based on the output presented above, the regression model demonstrated an accuracy reflected by an R-value of 0.747, or 74.7%. This value represents the strength of the association between the independent variables and the dependent variable. Additionally, the Adjusted R Square value was reported as 0.521, indicating that the independent variables included in the model collectively explain 52.1% of the variance in the dependent variable. The remaining 47.9% of the variation is attributed to other factors not captured within this regression model.

Coefficientsa

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.602E6	530401.500		3.021	.003
	Umur	19356.508	8672.345	.156	2.232	.028
	Jam_kerja	26910.320	7750.981	.275	3.472	.001
	DKBJI1	-1.610E6	588703.764	-.203	-2.734	.007
	DKBJI2	-1.179E6	398498.223	-.250	-2.957	.004
	DKBJI3	-2.468E6	347564.729	-.689	-7.102	.000
	DKBJI4	-2.506E6	528842.399	-.363	-4.738	.000
	DKBJI5	-2.593E6	304500.724	-.932	-8.516	.000
	DKBJI6	-2.607E6	322211.982	-.817	-8.092	.000

a. Dependent Variable: Upah

Figure 6. Results of the dummy regression model

Based on the output above, it was obtained that the formed dummy regression model was:

Y = 1.602.488 + 19.356,51X1 + 26.910,32X2 - 1.609.782DKBJI1 - 1.178.512DKBJI2 – 2.468.488DKBJI3 – 2.505.716DKBJI4 – 2.593.132 DKBJI5 – 2.607.300DKBJI6

CONCLUSION

Based on the results of the analysis, West Nusa Tenggara Province ranks as the province with the second-lowest average minimum wage in Indonesia. In East Lombok specifically, the average monthly income of residents is approximately Rp 1,084,252.00, which falls below the provincial Regional Minimum Wage (UMR) of Rp 1,120,000.00 per month. The study identified several key variables influencing the income of employed individuals in East Lombok, notably age, working hours, and job classification. The constructed dummy regression model revealed a statistically significant relationship between these independent variables and income, as expressed in the following regression equation: $Y = 1{,}602{,}488 + 19{,}356.51X_1 + 26{,}910.32X_2 - 1{,}609{,}782DKBJI_1 - 1{,}178{,}512DKBJI_2 - 2{,}468{,}488DKBJI_3 - 2{,}505{,}716DKBJI_4 - 2{,}593{,}132DKBJI_5 - 2{,}607{,}300DKBJI_6$. The coefficient of determination (R^2) of 0.747 (74.7%) confirms that the model has a strong explanatory power, indicating a substantial relationship between the independent variables and income. Additionally, the analysis shows that 52.1% of the income variation is accounted for by the variables included in the model, while 47.9% of the variation is likely influenced by other external factors not examined within the scope of this study.

REFRENCES

[1] Binti, M. T. (2017). Analisa Pengaruh Pertumbuhan Ekonomi Terhadap Penurunan Tingkat Kemiskinan di Kalimantan Tengah. *AlKALAM: JURNAL KOMUNIKASI, BISNIS DAN MANAJEMEN*, *3*(2), 69-78.

[2] Badan Pusat Statistik.____.*Survei Angkatan Kerja Nasional*. http://sirusa.bps.go.id/index.php?r=istilah/view&id=699. Diunduh Tanggal 10 November 2024 pukul 9.44

[3] Syahbania, M. 2017. *Variabel Dummy*. http://ethasyahbania.blogspot.com /2011/01/variabel-dummy.html. Diunduh Tanggal 10 Oktober 2024 pukul 19.21.

[4] Basri, H. (2019). Pemodelan Regresi Berganda Untuk Data Dalam Studi Kecerdasan Emosional. *DIDAKTIKA: Jurnal Kependidikan*, *12*(2), 103-116.

[5] Padilah, T. N., & Adam, R. I. (2019). Analisis regresi linier berganda dalam estimasi produktivitas tanaman padi di Kabupaten Karawang. *FIBONACCI: Jurnal Pendidikan Matematika Dan Matematika*, *5*(2), 117-128.

[6] Widodo, E. (2016). Analisis Arah Kausalitas (Causal Ordering). *Journal of Indonesian Economy and Business (JIEB)*, *31*(1), 75-82.

[7] Muhyi, M., et al. (2018). *Qualitative Research Methods, Quantitative Research Methods, and Mixed Methods in Education*.

[8] Padilah, D., & Wulandari, S. (2018). *Multiple Linear Regression Analysis for Estimating Rice Crop Productivity in Karawang Regency*.

[9] Nurdin, I., Sugiman, S., & Sunarmi, S. (2018). Penerapan Kombinasi Metode Ridge Regression (RR) dan Metode Generalized Least Square (GLS) untuk Mengatasi Masalah Multikolinearitas dan Autokorelasi. *Indonesian Journal of Mathematics and Natural Sciences*, *41*(1), 58-68.

[10] Quraisy, A. (2020). Normalitas data menggunakan uji kolmogorov-smirnov dan saphiro-wilk: studi kasus penghasilan orang tua mahasiswa prodi pendidikan matematika unismuh makassar. *J-HEST Journal of Health Education Economics Science and Technology*, *3*(1), 7-

[11] Ahadi, G. D., & Zain, N. N. L. E. (2023). *Pemeriksaan Uji Kenormalan dengan Kolmogorov-Smirnov, Anderson-Darling dan Shapiro-Wilk. Eigen Mathematics Journal, 6 (1), 11–19*.