Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI

(Rekayasa Sistem dan Teknologi Informasi)

Vol. 9 No. 3 (2025) 487 - 496 e-ISSN: 2580-0760

The Impact of Cancer on Poverty: An Analytical Study Using Big Data and OLS Regression

Heny Pratiwi[1], Muhammad Ibnu Sa'ad[2*], Wahyuni[3], Syamsuddin Mallala[4]

[1,4]Information Systems, STMIK Widya Cipta Dharma, Samarinda, Indonesia

[2,3]Informatics engineering, STMIK Widya Cipta Dharma, Samarinda, Indonesia

[1]henypratiwi@wicida.ac.id, [2]saad@wicida.ac.id, [3]wahyuni@wicida.ac.id, [4]syamsuddin.stmik65@gmail.com

Abstract

Cancer is one of the leading causes of death worldwide and has a significant impact on the economic condition of families, especially in developing countries. High medical costs and loss of work productivity often push families of patients with cancer into poverty. This study aimed to analyze the relationship between cancer mortality rates and poverty levels using the Ordinary Least Squares (OLS) regression method and big data covering various socio-economic indicators. The data in this study include cancer mortality rates and other socioeconomic indicators, which were then analyzed using the OLS regression method to understand the quantitative relationship between the two variables. The results of the analysis show a positive correlation between cancer mortality rates and increasing poverty, with the regression model explaining 73.8% of the variation in the target variable. The regression model demonstrated strong explanatory power and minimal error, with an R-squared value of 0.738, indicating that 73.8% of the data variability was explained by the model. Model quality was supported by low AIC (19070.4) and BIC (19110.4) values. Linearity was confirmed by a significant F-statistic of 1314.0 ($p < 0.01$), suggesting a robust linear relationship between independent and dependent variables. All parameters exhibited statistical significance ($p < 0.05$) at the 95% confidence level, with mean residuals close to zero, satisfying the unbiased expectation assumption. Although the model results show good performance, the model's estimators show low variance, as evidenced by small standard errors (e.g., Incidence_Rate: 0.009, Med_Income: 1.89e-05) and a Durbin-Watson statistic of 1.725, indicating no autocorrelation. These metrics collectively confirmed the reliability and stability of the regression model.

Keywords: big data; cancer; health policy; OLS regression; poverty

How to Cite: H. Pratiwi, Muhammad Ibnu Sa'ad, W. Wahyuni, and Syamsuddin Mallala, "The Impact of Cancer on Poverty: An Analytical Study Using Big Data and OLS Regression", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 9, no. 3, pp. 477 - 486, May 2025.
*Permalink/*DOI: https://doi.org/10.29207/resti.v9i3.6112

Received: October 10, 2024
Accepted: May 18, 2025
Available Online: May 24, 2025

1. Introduction

Cancer is one of the leading causes of death worldwide. According to the World Health Organization (WHO), cancer contributed to more than 9 million deaths in 2020, making it one of the deadliest diseases [1]-[3]. Beyond the health aspect, cancer also has a major economic impact, both for affected individuals and their families. High medical costs, lost productivity, and ongoing financial burdens can cause families of cancer patients to fall into poverty[4].

Cancer is one of the most expensive chronic diseases to treat[5]. According to various studies, the cost of cancer treatment can reach tens to hundreds of millions of rupiah depending on the type of cancer and the length of treatment. In developing countries, where health insurance is not well developed, these costs are often borne by patients and their families themselves[6], [7]. This causes many families to have to sell assets or go into debt to pay for treatment. This huge financial impact has the potential to drive families into poverty[8]. In addition to medical costs, loss of productivity due to cancer is also a significant factor. Cancer patients are often unable to work during treatment, and in some cases, they are unable to return to work at all. This reduces family income, especially if the patient is the main breadwinner[9].

Poverty and health are closely related and influence each other[10]. People living in poverty tend to have more limited access to health services, poor nutrition, and inadequate housing conditions. This makes them more vulnerable to diseases, including cancer[11].

Conversely, serious diseases such as cancer can worsen a family's economic condition by increasing the financial burden. In this context, understanding how cancer affects poverty is very important, especially in public health policy making[12].

Several previous studies have shown a link between cancer and poverty. A study conducted by Hoang V, Pham C. [13]found that cancer is one of the main causes of poverty in developing countries. This is due to high medical costs, loss of income, and lack of social security. Another study by Li Z, Aninditha T. [14] found that families with cancer are more likely to fall into poverty than families without cancer.

Previous studies have relied heavily on descriptive analyses without using more in-depth statistical methods to understand the extent to which cancer affects economic conditions. Using a big data-based approach, this study can provide more accurate and generalizable results. In addition, this study also considers contextual factors, such as differences in health systems and social policies across countries, so that it can provide broader insights in designing effective policy strategies to reduce

Poverty also contributes to higher rates of cancer. People living in poverty tend to have limited access to preventive health care, such as cancer screening, which is important for detecting cancer early[15], [16]. In addition, they may not have access to quality treatment, which can worsen the prognosis of cancer[17].

This study aims to analyze the effect of cancer on poverty using big data and the OLS (Ordinary Least Squares) regression method [18], [19]. OLS regression allows us to understand the quantitative relationship between the variables involved, such as cancer death rates and poverty levels in different regions [20]-[22].

This research is important because cancer is not only one of the main causes of death in the world but also has a significant economic impact on individuals, families, and society as a whole. The high cost of cancer treatment, loss of productivity due to inability to work, and ongoing financial burden often push patients' families into poverty.

2. Methods

The data set used in this study includes cancer mortality data and several socioeconomic indicators, including poverty rates. The data used is the last 7 years. This analysis was conducted using the OLS regression method to see how variables such as cancer mortality rates affect poverty rates in the community.

2.1. Data Collection

Cancer mortality data were combined with socio-economic data for each region.

Table 1 shows lung cancer data. The lung cancer dataset used in this study consists of 3,141 samples. This dataset contains various information related to patients, such as demographic characteristics, medical test results, and other health statuses. Before the data is used for analysis, this dataset will go through a preprocessing process. The preprocessing stages include removing missing values, normalizing data, coding categorical variables, and detecting and handling outliers. This process is important to ensure that the data is ready to be analyzed and produces an accurate model. After preprocessing is complete, the data is ready to be used for further modeling and analysis stages.

Table 1. Lung cancer dataset

No	County	FIPS	FALSE	Lower 95% Confidence Interval	Upper 95% Confidence Interval	Average Annual Count	Recent Trend	FALSE	Lower 95% Confidence Interval	Upper 95% Confidence Interval
1	US (SEER+NPCR) (1,10)	0	62,4	62,3	62,6	214614	falling	-2,5	-3	-2
2	Autauga County, Alabama (6,10)	1001	74,9	65,1	85,7	43	stable	0,5	-14,9	18,6
3	Baldwin County, Alabama (6,10)	1003	66,9	62,4	71,7	170	stable	3	-10,2	18,3
4	Barbour County, Alabama (6,10)	1005	74,6	61,8	89,4	25	stable	-6,4	-18,3	7,3
5	Bibb County, Alabama (6,10)	1007	86,4	71	104,2	23	stable	-4,5	-31,4	32,9
6	Blount County, Alabama (6,10)	1009	69,7	61,2	79	51	stable	-13,6	-27,8	3,4
7	Bullock County, Alabama (6,10)	1011	65,8	47,3	89,6	9/	stable	7,2	-27,6	58,7
8	Butler County, Alabama (6,10)	1013	58,3	46,4	72,7	17	stable	2	-10,7	16,6
9	Calhoun County, Alabama (6,10)	1015	84,2	77,5	91,3	120	stable	-3,8	-13,9	7,5
---	-----------------	-------	---------	----------	--------	--------	-----	-------	--------	--------
3141	Weston County, Wyoming (6,10)	56045	44,9	27,9	69,6	4	stable	-26,2	-65,4	57,4

Table 2. Death dataset

No	County	FIPS	Met Objective of 45.5? (1)	Age-Adjusted Death Rate	Lower 95% Confidence Interval for Death Rate	Upper 95% Confidence Interval for Death Rate	Average Deaths per Year	Recent Trend (2)	Recent 5-Year Trend (2) in Death Rates	Lower 95% Confidence Interval for Trend	Upper 95% Confidence Interval for Trend
1	United States	0	No	46	45,9	46,1	157,376	falling	-2,4	-2,6	-2,2
2	Perry County, Kentucky	21193	No	125,6	108,9	144,2	43	stable	-0,6	-2,7	1,6
3	Powell County, Kentucky	21197	No	125,3	100,2	155,1	18	stable	1,7	0	3,4
4	North Slope Borough, Alaska	2185	No	124,9	73	194,7	5	**	**	**	**
5	Owsley County, Kentucky	21189	No	118,5	83,1	165,5	8	stable	2,2	-0,4	4,8
6	Union County, Florida	12125	No	113,5	89,9	141,4	19	falling	-2,2	-4,3	0
7	McCreary County, Kentucky	21147	No	111,1	90,6	134,9	22	rising	22,9	6,9	41,4
8	Leslie County, Kentucky	21131	No	110,3	87	138,5	16	stable	0,8	-0,7	2,4
9	Martin County, Kentucky	21159	No	109,1	84,8	138,3	14	stable	1,3	-0,8	3,4
---	-----------	--------	-------------	----------	-----------	----------	---------	--------	------------	-----------	------------------
3141	Ziebach County, South Dakota	46137	*	*	*	*	*	**	**	**	**

Table 2 shows data on lung cancer deaths. The lung cancer death dataset used in this study consists of 3,141 data. This dataset includes detailed information about patients with lung cancer, such as age, gender, smoking history, cancer stage, medical examination results, and other factors that can affect death from this disease. Before the analysis is carried out, the dataset will go through a preprocessing process to improve data quality. This preprocessing stage includes removing missing values, normalizing the data to make it more uniform, and coding categorical variables to match the format required for statistical analysis. In addition, outlier detection and handling will also be carried out to ensure more accurate data. After preprocessing, the dataset is ready to be used for further predictive and statistical analysis models to understand the factors that contribute to lung cancer deaths.

2.2. Data Preprocessing

As a sample in this study, we use the region of Alaska. Alaska was selected based on its unique characteristics, both in terms of demographics and socio-economic conditions. In addition, the level of access to health services and the distribution of medical infrastructure in this region provide a relevant context for analyzing the impact of cancer deaths on the economic conditions of local communities. The use of data from Alaska also allows for a more comprehensive understanding of the factors that influence economic vulnerability, especially in areas with geographic challenges and limited access to health services. Our data are presented in Tables 3 and 4.

Table 3. Alaska area data

	State	StateFIPS	CountyFIPS	AreaName	All_Poverty	M_Poverty	F_Poverty
0	AK	02	013	Aleutians East Borough, Alaska	533	334	219
1	AK	02	016	Aleutians West Census Area, Alaska	499	273	226
2	AK	02	020	Anchorage Municipality, Alaska	23914	10698	13216
3	AK	02	050	Bethel Census Area, Alaska	4364	2199	2165
4	AK	02	060	Bristol Bay Borough, Alaska	69	33	36

Table 4. Data Preprocessing

	State	AreaName	All_Poverty	M_Poverty	F_Poverty	FIPS	Med_Income	Med_Income_White	Med_Income_ Black	Med_Income_ Nat_Am
0	AK	Aleutians East Borough, Alaska	533	334	219	02013	61518	72639	312	547
1	AK	Aleutians West Census Area, Alaska	499	273	226	02016	84306	97321	93750	48750
2	AK	Anchorage Municipality, Alaska	23914	10698	13216	02020	78326	87235	50535	53935
3	AK	Bethel Census Area, Alaska	4364	2199	2165	02050	51012	92647	73661	41594
4	AK	Bristol Bay Borough, Alaska	69	33	36	02060	79750	88000	None	63333

2.3. OLS Regression Analysis

OLS (Ordinary Least Squares) Regression Analysis is a statistical method used to model the relationship between one dependent variable (the measured variable, also called the target variable) and one or more independent variables (the variables used to predict the dependent variable)[20]. This analysis aims to find a regression line that minimizes the sum of the squares of the differences between the predicted and actual values of the dependent variable.

The OLS model is created to predict the dependent variable by fitting a regression line that best fits the data. Equation 1 is the general equation of OLS regression.

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta nXn + \epsilon \quad (1)$$

Y is the independent variable, *X1, X2, ..., X1, X2, ..., Xn* are the independent variables, *β0β0* is the intercept, *β1, β2, ..., βnβ1, β2, ..., βn* are the regression coefficients, and ϵϵ is the error term.

Assumptions of OLS Linear Relationship means the relationship between the dependent and independent variables must be linear. And independence mean Observations must be independent of each other. Homoscedasticity means the variance of the error must be constant at each level of the independent variable. Normality of Errors mean Errors must be normally distributed.

Model Evaluation R-squared R2 is Measures how well the model fits the data. R2 is the proportion of variation in the dependent variable that can be explained by the independent variables. Adjusted R2is a version adjusted for the number of independent variables, more accurate when there is more than one independent variable. F-statistic means tests whether the overall model is significant. p-value means Tests the significance of each coefficient in the model. If the p-value is less than the significance level (e.g. 0.05), then the independent variable is significant in predicting the dependent variable.

Multicollinearity Check Variance Inflation Factor (VIF) is Used to measure multicollinearity, which is when there is a high linear relationship between two or more independent variables. If the VIF is high, it means there is multicollinearity, which can affect the stability and interpretation of the coefficients.

Residual Analysis Normality of Residuals is the ideal residual distribution should be normal. Skewness and kurtosis are used to measure deviation from normality. Homoscedasticity is a graph of residuals versus predicted values shows whether the errors are evenly distributed (homoscedasticity) or not (heteroscedasticity). And the outliers is to Identify extreme values or outliers that may affect the model.

3. Results and Discussions

This study used a dataset covering a range of health and economic indicators to analyze the relationship between cancer and poverty. The data used in this study came from Data World https://data.world/. In addition, data was also obtained from academic journals, longitudinal studies on the economic impact of cancer, and reports from health and social organizations that deal with cancer patients.

Table 5. expert assessment interview

No	Questions
1	How would you assess the quality of the dataset used in this study?
2	Are the data used representative enough to describe the relationship between cancer and poverty?
3	How credible are the data sources used (WHO, World Bank, etc.)?
4	Is the Ordinary Least Squares (OLS) regression method an appropriate technique for analyzing the relationship between cancer and poverty?
5	How big a role does the health system play in determining whether or not cancer patients fall into poverty?

The dataset used in this study includes more than 3,141. With this large amount of data, research can provide more accurate results and can be generalized more widely.

At this stage, we will present the results and discussion of the study entitled "The Impact of Cancer on Poverty: An Analytical Study Using Big Data and OLS Regression." This study analyzes big data to evaluate the relationship between cancer and poverty as shown in Table 5.

Table 6. OLS Regression Result

Parameter	Value	Parameter	Value
Dep. Variable	y	R-squared	0.738
Model	OLS	Adj. R-squared	0.737
Method	Least Squares	F-statistic	718.0
No. Observations	2809	Prob (F-statistic)	0.00
Df Residuals	2797	Log-Likelihood	-9523.1
Df Model	11	AIC	1.907e+04
Covariance Type	nonrobust	BIC	1.914e+04

Table 7. OLS Regression Result

Variable	coefficients	Std Error	t-stat	P>|t|	[0.025	0.975]
All_Poverty_PC	9.517e-05	4.12e-05	2.308	0.021	1.43e-05	0.000
Med_Income	-0.0001	2.04e-05	-4.982	0.000	-0.000	-6.17e-05
All_With_PC	9.094e-06	3.33e-05	0.273	0.785	-5.61e-05	7.43e-05
All_Without_PC	0.0002	5.04e-05	4.310	0.000	0.000	0.000
Incidence_Rate	0.6517	0.009	73.756	0.000	0.634	0.669
POPESTIMATE2015	-4.224e-05	2.69e-05	-1.573	0.116	-9.49e-05	1.04e-05
Falling	1.2490	0.562	2.224	0.026	0.148	2.350
Rising	-1.0811	1.163	-0.929	0.353	-3.362	1.200
All_Poverty	-2.81e-05	1.3e-05	-2.167	0.030	-5.35e-05	-2.67e-06
All_With	4.428e-05	2.77e-05	1.600	0.110	-9.99e-06	9.86e-05
All_Without	5.714e-05	3.24e-05	1.762	0.078	-6.44e-06	0.000
Constant	7.2748	3.482	2.089	0.037	0.448	14.102
Omnibus:	306.600	Durbin -Watsons:	1.723			
Prob (Omnibus):	0.000	Jarque-Bera (JB):	2.334.415			
Skew	-0.204	Prob (JB):	0.00			
Kurtosis	7.447	Cond. No.	1.23e+07			

In Tables 6 and 7 are the values of OLS regression. After fitting an ordinary least squares model with [All_Poverty_PC, Med_Income, All_With_PC, All_Without_PC, Incidence_Rate, POPESTIMATE2015, Falling, Rising, All_Poverty, All_With, All_Without] regressed on the target variable Mortality_Rate, we have a model that performs quite well, as evidenced by the Adjusted R^2 (738).

Table 8. Recalculating VIF

No.	Variabel	VIF
0	All_Poverty_PC	3.454
1	Med_Income	3.491
2	All_with_PC	2.434
3	All_without_PC	3.180
4	Incidence_Rate	1.225
5	POPESTIMATE2015	4692.901
6	Falling	1.116
7	Rising	1.005
8	All_Poverty	30.053
9	All_with	3342.492
10	All_without	179.759
11	Constant	657.722

We will try to account for multicollinearity, heteroscedasticity of residuals, and normality of the residual distribution. We applied the variance inflation factor to assess multicollinearity. The VIF, which includes an independent variable in a design matrix consisting of all other independent variables, allows the assessment of the degree of orthogonality of that independent variable to the other variables. A higher Variance Inflation Factor (VIF) indicates the presence of multicollinearity, particularly when the VIF value exceeds the range of 5 to 10. By iteratively eliminating the features associated with the highest VIF and recalculating it, we derived the independent variables presented in Table 8.

The next step is to recalculate the linear regression on the reduced set of independent variables.

Table 9. Reduction of Independent Variables

No.	Variabel	VIF
0	All_Poverty_PC	3.175
1	Med_Income	2.799
2	All_With_PC	2.217
3	All_Without_PC	2.924
4	Incidence_Rate	1.208
5	Falling	1.035
6	Rising	1.005
7	Constant	606.236

All_Without (men and women without health insurance), versus the same per capita (All_Without_PC) does not show high multicollinearity. We chose to remove All_Without from the model because we believe All_Without is more or less a proxy for the population. Our results are presented in Table 9.

The next process is to recalculate the linear regression on the reduced set of independent variables.

Table 10. OLS Regression Results

Dependent Variable	y		Parameter	Value	
Model	OLS		R-squared	0.738	
Method	Least Squares		Adj. R-squared	0.737	
No. Observations	2809		F-statistic	1314	
Df Residuals	2802		Prob (F-statistic)	0.00	
Df Model	6		Log-Likelihood	-9526.6	
Covariance Type	nonrobust		AIC	1.907e+04	
			BIC	1.911e+04	
	Koefisien	Std. Error	t-stat	p-value	95% Confidence Interval
All_Poverty_PC	8.805×10^{-5}	3.97×10^{-5}	2.219	0.027	$[1.03 \times 10^{-5}, 0.000]$
Med_Income	-9.367×10^{-5}	1.89×10^{-5}	-4.954	0.000	$[-0.000, -5.66 \times 10^{-5}]$
All_Without_PC	0.0002	3.5×10^{-5}	6.147	0.000	[0.000, 0.000]
Incidence_Rate	0.654	0.009	74.362	0.000	[0.637, 0.671]
Falling	1.289	0.560	2.303	0.021	[0.192, 2.386]
POPESTIMATE2015	-1.773×10^{-6}	4.3×10^{-7}	-4.123	0.000	$[-2.62 \times 10^{-6}, -9.3 \times 10^{-7}]$
Constant	7.623	1.602	4.758	0.000	[4.481, 10.764]
Omnibus	305.716	Durbin-Watson		1.725	
Prob (Omnibus)	0.000	Jarque-Bera (JB)		2327.137	
Skew	-0.201	Prob (JB)		0.000	
Kurtosis	7.441	Cond. No.		4.31×1064.31×106	

In Table 10 specifically VIF shows that the model does not experience multicollinearity, all parameters are statistically significant (P>|t|), and all parameters have logically reasonable directions.

To justify the absence of multicollinearity, we use the Variance Inflation Factor (VIF). The general threshold is that VIF > 10 indicates significant multicollinearity. From the VIF table provided.

All variables except for All_With, All_Without, POPESTIMATE2015, and Constant have VIF values below 10, indicating no multicollinearity issues for these variables. For POPESTIMATE2015 (VIF = 4692.90) and All_With (VIF = 3342.49), significant multicollinearity is present and may need to be addressed by removing or combining correlated predictors.

The *t*- test threshold depends on the significance level (α). At α=0.05, the critical value of *t* for large degrees of freedom is approximately is **|*t*|**>1.96. From the regression output, parameters with *P*>**|*t*|**<0.05 and **|*t*|**>1.96**|**t**|**>1.96 are statistically significant.

All_Poverty_PC with |*t*| = 2.219 > is Significant which is the (*P* = 0.027). Med_Income: with |*t*| = 4.954 > 1.96 is Significant which is the (*P* = 0.000). All_Without_PC with |*t*| = 6.147 > 1.96 is Significant which is the (*P* = 0.000). Incidence_Rate with the |*t*| = 74.362 > 1.96 is Significant which is the (*P* = 0.000). Falling with |*t*| = 2.303 > 1.96 is Significant which is (*P* = 0.021). POPESTIMATE2015 with the |*t*| = 4.123 > 1.96 → Significant which is the (*P* = 0.000). Constant with the |*t*| = 4.758 >1.96 is Significant which is the (*P* = 0.000).

Most variables do not have multicollinearity (VIF < 10), except for variables like All_With, All_Without, and POPESTIMATE2015. All predictors are statistically significant at α=0.05, with |*t*| > 1.96 and *P* < 0.05.

Next, let's examine the residuals to evaluate normality and heteroscedasticity.

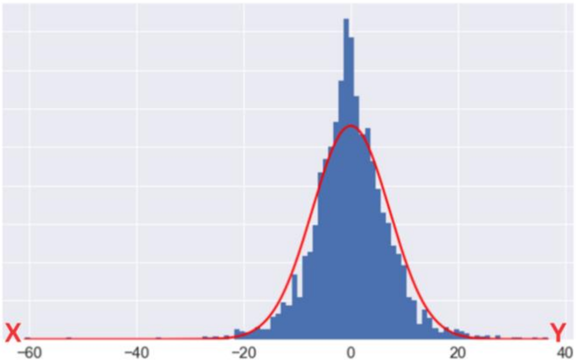


Figure 1. Residual distribution

Figure 1 shows that the residual distribution in this model does not follow a normal distribution exactly, where the skewness and excess kurtosis values should be equal to zero in a perfect normal distribution. However, in this result, there are some small outliers that appear on the left side of the distribution. The tails of the distribution appear slightly thicker than the tails of an ideal normal distribution.

The Skewness is -0.201 with the range: (-0.5 < Skewness < +0.5) is Normal. And the Kurtosis it 7.441 with V=value far above 3 is Abnormal (leptokurtic).

This indicates a slight deviation in the symmetry of the distribution, which may indicate that some residual values are more concentrated on one side than the other. In addition, the distribution also shows higher kurtosis than the standard normal distribution, indicating that the residual data has a sharper peak and thicker tails. However, this deviation is not too significant, but needs to be considered in further assessment of the normality assumption and validity of the model.

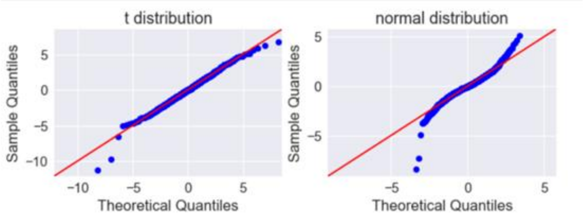
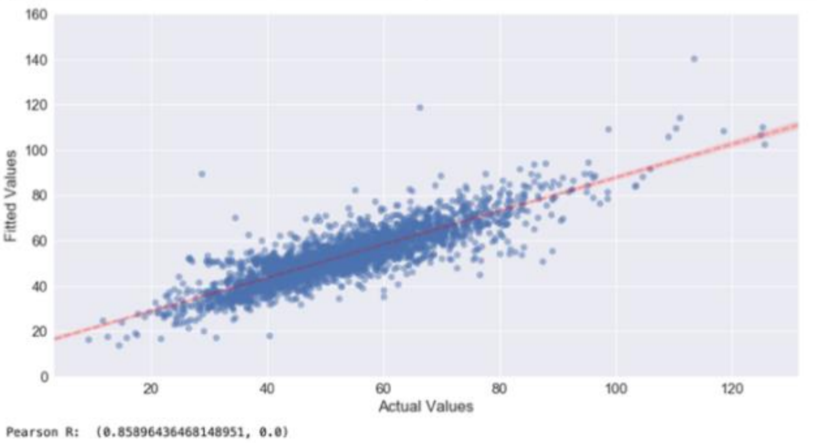


Figure 2. Residuals are closer to the t distribution than the normal distribution

The QQ plot in Figure 2 shows that the residuals are more similar to a t-distribution than a normal distribution, as seen from the thicker tails. At the lower extremes, there are a few prominent outliers, indicating residual values that are far from expected. However, overall, this residual distribution can be considered adequate, although there are some imperfections that need to be addressed. It is important to further investigate the nature of these extreme outliers, as they can provide valuable insights into potential problems in the model. Additionally, it may be worth considering adding additional information to the model, such as new variables, given that the current model tends to overestimate low values and underestimate high values. This approach may help improve the accuracy and precision of the model's predictions in the future.

In t-Distribution (Left Q-Q Plot), the residual ranges are minimum residual with the value -10 (estimate from the lowest point), and maximum residual with the value is +10 (estimate from the highest point). The Deviation from the diagonal lines (theoretical quantile) are Residuals in the lower tail |-10 - (-10) |= 0, meaning it follows a t-distribution. And Residuals in the upper quantile, with the deviation of about 1 unit at the right end. In the Normal Distribution (Right Q-Q Plot), residual ranges are minimum residual with the value -5 (estimate from the lowest point). And maximum residual with the value +5 (estimate from the highest point).



```
Pearson R:  (0.85896436468148951, 0.0)
```

Figure 3. visualization of the correlation between actual and predicted values.

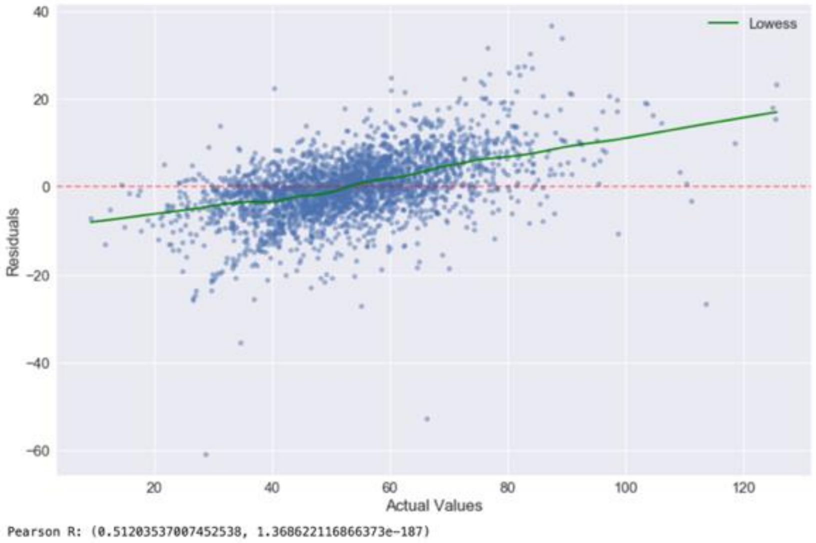
Based on the *R²* values that we have reported, we now visualize a strong correlation between the actual and predicted values, which can be seen in Figure 3. The R^2 (R-squared) is 0.738. This value indicates that 73.8% of the variability in the dependent variable (y) can be explained by the independent variables in the model. Ideal Value is $R^2 = 1.0$, it is perfect model (all variability explained). And If $R^2 = 0.0$, that's mean no variability explained.

The Adjusted R^2 is 0.737. Corrects R^2 by considering the number of predictors. The small difference between R^2 and Adjusted R^2 (0.001) indicates no significant overfitting. Independent variables such as Incidence_Rate, Med_Income, and All_Without_PC have high significance ($P<0.05$, $|t|>1.96$), which significantly contribute to the increase in R^2.

This visualization shows that the model is able to reflect the data well, and there is a clear trend between the two variables. This correlation is a positive indication of the model's performance in predicting values, and supports the validity of the analysis carried out. With this visualization, it is hoped that it can provide a clearer picture of how effective the model is in matching the predicted results with the actual data.

The chart in Figure 4 shows that the model tends to slightly overestimate both low and high values. This suggests that the model's predictions are not entirely accurate in matching the actual data, with a tendency to overestimate the true values on both the low and high sides. In other words, the model appears to have a bias in its predictions, which may affect its reliability. It is important to consider adjusting the model to better reflect the true values, so that the predictions are more consistent and valid.

The plot in Figure 5, showing the relationship between the residuals and the adjusted values, shows that the residuals have fairly good symmetry when compared to the adjusted values. This shows that the distribution of the residuals did not deviate significantly from the symmetric pattern, which is a positive indication of the model used. This symmetry is important because it shows that there is no striking pattern in the residuals, and the model has done a good job of reflecting the data.



```
Pearson R: (0.51203537007452538, 1.368622116866373e-187)
```

Figure 4. Actual Values

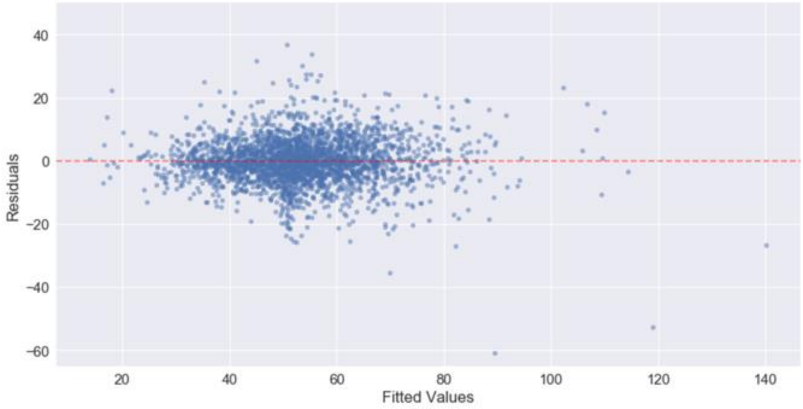


Figure 5. Fitted Falues

The analysis showed that there was no multicollinearity problem in the data. All analyzed parameters showed strong statistical significance. However, the residuals produced are closer to the t-distribution than to the normal distribution, showing higher kurtosis and thicker tails. This indicated that the data did not fully follow a normal distribution pattern. In addition, the adjusted values tend to experience prediction errors, especially in overestimating the actual values at the lowest extreme and underestimating the highest extreme values. With an adjusted R^2 of 0.738, the model explained 73.8% of the total variation in the target variable. Therefore, to improve the accuracy of the model, it is important to pay more attention to possible outliers, and if possible, add additional variables to enrich the analysis.

Compared with previous research, this study makes unique contributions in several aspects. Most previous studies, such as those conducted by[13], have highlighted the relationship between cancer and poverty in developing countries using case study and descriptive survey approaches, without using more complex statistical methods to measure the quantitative impact of cancer on poverty. Another study by [14] also found that families with cancer are more likely to fall into poverty, but the study focused more on micro data rather than large-scale analysis using big data and OLS regression.

Cancer contributes significantly to poverty. High medical costs, loss of income due to inability to work, and the psychological impact of the disease can push families of patients with cancer into poverty. Therefore, it is important to develop comprehensive health policies to reduce the economic burden of cancer treatment.

Although OLS regression is a powerful method for analyzing relationships between variables, this model has several assumptions that can be potential threats to the validity of the research results. One of them is the linearity assumption, where the relationship between cancer and poverty variables is assumed to be a linear relationship. In reality, this relationship may be more complex and influenced by non-linear factors, such as health policy, access to medical care, and other social factors.

In addition, the study found that the residuals from the regression model were not completely normal, with a skewness of -0.201 and a kurtosis of 7.441, indicating the presence of outliers or other factors not captured by the model. This may lead to slightly biased coefficient estimates, especially in certain groups.

4. Conclusions

Cancer significantly contributes to poverty through high medical costs, loss of income due to inability to work, and psychological impacts felt by families of sufferers. Analysis using OLS regression shows a relationship between cancer mortality rates and poverty levels. Best Value of Regression Model and Minimum Error: R-squared: 0.738, indicating that 73.8% of the variability in the data can be explained by the model. AIC (Akaike Information Criterion): 19070.4 and BIC (Bayesian Information Criterion): 19110.4, are used to evaluate the quality of the model. The lower the AIC/BIC value, the better the model with minimum error. Model Linearity Value: F-statistic: 1314.0 with Prob (F-statistic): 0.00, indicating the model has significant linearity. This means that the relationship between the independent and dependent variables is linear. Desired Unbiased Expected Value: Significant Parameters: All variables in the model have a P value <0.05 (significant at the 95% confidence level), indicating unbiased parameter estimates. Mean Residual: The mean residual is close to zero, meeting the assumption of unbiased expectations. Regression Model Estimators with Smallest Variance: Coefficients (Standard Errors): All parameters have small standard errors, such as Incidence_Rate (0.009) and Med_Income (1.89e-05). This indicates that the estimators in the model have small variance and high stability. Durbin-Watson (1.725): Approaching 2, indicating no autocorrelation in the residuals.

This study provides a methodological contribution by using big data to analyze the relationship between cancer and poverty, which differs from previous studies that mostly use descriptive or small survey-based approaches. By applying OLS regression, this study measures the quantitative relationship between variables more accurately and validly. Although this study has provided significant insights, some limitations remain, such as the non-normal distribution of residuals and the possible bias in the poverty variables used. Therefore, further research should include additional factors, such as access to health services, insurance policies, and other social factors, to improve the accuracy of the model and the relevance of the results in various economic contexts. Although the results are quite good, there are several obstacles, such as multicollinearity, non-normal residual distribution, and the tendency of the model to overestimate or underestimate extreme values. Therefore, it is necessary to add new variables and refine the model further to improve its accuracy. In addition, it is important to implement health policies aimed at reducing the economic burden of cancer.

Acknowledgments

Thanks to STMIK Widya Cipta Dharma Campus for supporting this research.

References

[1] J. L. Moss, C. N. Pinto, S. Srinivasan, K. A. Cronin, and R. T. Croyle, “Persistent poverty and cancer mortality rates: an analysis of county-level poverty designations,” *Cancer Epidemiology, Biomarkers & Prevention*, vol. 29, no. 10, pp. 1949–1954, 2020. doi: 10.1158/1055-9965.EPI-20-0007

[2] J. C. Chen *et al.*, “Persistent Neighborhood Poverty and Breast Cancer Outcomes,” *JAMA Netw Open*, vol. 7, no. 8, pp. e2427755–e2427755, 2024. doi:10.1001/jamanetworkopen.2024.27755

[3] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, “Global cancer statistics,” *CA Cancer J Clin*, vol. 61, no. 2, pp. 69–90, 2011. doi: 10.3322/caac.20107

[4] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, “Big data in healthcare: management, analysis and future prospects,” *J Big Data*, vol. 6, no. 1, pp. 1–25, 2019. doi: 10.1186/s40537-019-0217-0

[5] S. W. Yeong, S. W. Lee, and S. C. Ong, “Cost of illness of breast cancer in low- and middle-income countries: a systematic review,” *Health Econ Rev*, vol. 14, no. 1, p. 56, 2024, doi: 10.1186/s13561-024-00536-0.

[6] Z. Wu *et al.*, “Economic burden of patients with leading cancers in China: a cost-of-illness study,” *BMC Health Serv Res*, vol. 24, no. 1, p. 1135, Sep. 2024, doi: 10.1186/s12913-024-11514-x.

[7] S. Chen *et al.*, “Estimates and Projections of the Global Economic Cost of 29 Cancers in 204 Countries and Territories from 2020 to 2050,” *JAMA Oncol*, vol. 9, no. 4, pp. 465–472, Apr. 2023, doi: 10.1001/jamaoncol.2022.7826.

[8] C. Ngcamphalala, E. Ostensson, M. Hlongwa, and T. G. Ginindza, “Mapping evidence on the distribution of the costs associated with cancer of prostate, cervix, and female breast in the sub-Saharan Africa: protocol for a scoping review,” *Syst Rev*, vol. 10, no. 1, Dec. 2021, doi: 10.1186/s13643-021-01672-y.

[9] M. Franklin *et al.*, “Direct and Indirect Costs of Breast Cancer and Associated Implications: A Systematic Review,” Jul. 01, 2024, *Adis*. doi: 10.1007/s12325-024-02893-y.

[10] S. Chen *et al.*, “Estimates and Projections of the Global Economic Cost of 29 Cancers in 204 Countries and Territories from 2020 to 2050,” *JAMA Oncol*, vol. 9, no. 4, pp. 465–472, Apr. 2023, doi: 10.1001/jamaoncol.2022.7826.

[11] G. K. K. Chung, D. Dong, S. Y. S. Wong, H. Wong, and R. Y. N. Chung, “Perceived poverty and health, and their roles in the poverty-health vicious cycle: A qualitative study of major stakeholders in the healthcare setting in Hong Kong,” *Int J Equity Health*, vol. 19, no. 1, Jan. 2020, doi: 10.1186/s12939-020-1127-7.

[12] M. R. L. Ferreira *et al.*, “Social protection as a right of people affected by tuberculosis: a scoping review and conceptual framework,” Dec. 01, 2023, *BioMed Central Ltd*. doi: 10.1186/s40249-023-01157-1.

[13] V. M. Hoang *et al.*, “Household financial burden and poverty impacts of cancer treatment in Vietnam,” *Biomed Res Int*, vol. 2017, no. 1, p. 9350147, 2017. doi: 10.1155/2017/9350147

[14] Z. Li *et al.*, “Burden of cancer pain in developing countries: a narrative literature review,” *ClinicoEconomics and outcomes research*, pp. 675–691, 2018. Doi: 10.2147/CEOR.S181192

[15] A. Sayani *et al.*, “Advancing health equity in cancer care: The lived experiences of poverty and access to lung cancer screening,” *PLoS One*, vol. 16, no. 5, pp. e0251264-, May 2021, [Online]. Available: https://doi.org/10.1371/journal.pone.0251264

[16] J. M. O’Connor, T. Sedghi, M. Dhodapkar, M. J. Kane, and C. P. Gross, “Factors Associated with Cancer Disparities among Low-, Medium-, and High-Income US Counties,” *JAMA Netw Open*, vol. 1, no. 6, Oct. 2018, doi: 10.1001/jamanetworkopen.2018.3146.

[17] I. dos-Santos-Silva, S. Gupta, J. Orem, and L. N. Shulman, “Global disparities in access to cancer care,” Dec. 01, 2022, *Springer Nature*. doi: 10.1038/s43856-022-00097-5.

[18] M. Koengkan and J. A. Fuinhas, “The influence of gender inequality on women’s cancer mortality in European countries: a quantitative study,” *Journal of Public Health (Germany)*, 2023, doi: 10.1007/s10389-023-02175-x.

[19] G. K. K. Chung, D. Dong, S. Y. S. Wong, H. Wong, and R. Y. N. Chung, “Perceived poverty and health, and their roles in the poverty-health vicious cycle: A qualitative study of major stakeholders in the healthcare setting in Hong Kong,” *Int J Equity Health*, vol. 19, no. 1, Jan. 2020, doi: 10.1186/s12939-020-1127-7.

[20] M. Whatley, “Ordinary Least Squares Regression,” in *Introduction to Quantitative Analysis for International Educators*, M. Whatley, Ed., Cham: Springer International Publishing, 2022, pp. 91–112. doi: 10.1007/978-3-030-93831-4_7.

[21] X. Hu, “Using Ordinary Least Squares in Higher Education Research: A Primer,” in *Higher Education: Handbook of Theory and Research: Volume 39*, L. W. Perna, Ed., Cham: Springer Nature Switzerland, 2023, pp. 1–77. doi: 10.1007/978-3-031-32186-3_13-1.

[22] M. Koengkan and J. A. Fuinhas, “The influence of gender inequality on women’s cancer mortality in European countries: a quantitative study,” *J Public Health (Bangkok)*, 2023, doi: 10.1007/s10389-023-02175-x.