Linear regression model as an approach to analyze spatio temporal data to know the influence factors of the number of crimes in North Sumatera

E N Manurung, Y Widyaningsih and S Aminah
Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Depok 16424, Indonesia
Corresponding author’s email: yekti@sci.ui.ac.id

Abstract. Crimes in Indonesia tend to increase every year, both in variation and in number. Crimes in this study include murder, torture, rape, theft, fraud, and other cases that make people feel insecure. This study aims to know what factors affect the number of criminal acts in the Province of North Sumatera and determine the influence of time and location in 33 districts/cities in North Sumatra. This study uses secondary data collected by POLRI, BPS, and Dinsos in North Sumatera Province in 2013, 2014, 2015 and 2016. The independent variables used are the number of unemployed people (X1), the number of poor people (X2), population density (X3), human development index (X4), the number of places of worship (X5), the economic growth rate (X6) and the number of families with a social psychological problem (X7). The linear regression model is used as an approach to analyze. The results of this study show that the independent variables significantly influence the number of crimes acts in North Sumatera Province. The independent variables are unemployment (X1), the number of poor (X2), population density (X3), and human development index (X4). There is a significant location influence on the occurrence of the number of criminal acts in 33 districts/cities of North Sumatera Province. Based on the descriptive statistics, Sibolga, Tebing Tinggi, Padangsidempuan, Tanjungbalai, Samosir, and Gunungsitoli are the six districts with the highest proportions in criminal actions, where those districts are all cities.

Keywords: Crime, linear regression model, spatio-temporal

1. Introduction
Criminal acts are all forms of unlawful acts and social and religious norms. Crime also harms society economically and psychologically. According to the Badan Pusat Statistika (BPS) data of Regency/City of North Sumatera Province, in 2013–2016, the highest number of crimes occurred in Medan. Meanwhile, the lowest crime rate occurred in Pakpak Bharat Regency. Based on data obtained from the Polisi Republik Indonesia (POLRI), in the period 2013-2016, the number of criminal acts that occurred in Indonesia did not show any decrease and increase. During 2015 there is at least one crime every 1 minute 29 seconds [1].

This study aims to determine the factors that significantly affect the number of criminie acts. The data used is in the form of spatio-temporal data or spatiotemporal data which is a spatial data of regency/city lattice in North Sumatera Province in 2013–2016. Therefore, the model used is the spatio-temporal model, but using a linear regression approach. While the variables used in the modeling are the number
of unemployed (X1), the number of poor (X2), population density (X3), human development index (X4), the number of places of worship (X5), the economic growth rate (X6) and the number of families with a social psychological problem (X7).

2. Methodology

2.1. Data
This study uses secondary data consists of observations of 33 districts/cities collected by POLRI (Polisi Republik Indonesia), BPS (Badan Pusat Statistik) and Dinsos (Dinas Sosial) in North Sumatera Province in 2013, 2014, 2015 and 2016.

2.2. Multiple regression linear model
Multiple linear regression model for \( k \) independent variables is the following,

\[
y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i
\]  
(1)

or

\[
y_i = \beta_0 + \sum_{k=1}^{k} \beta_k x_{ki} + \varepsilon_i \quad , i = 1, ..., n.
\]

equation 1 in the matrix form is

\[
Y = X\beta + \varepsilon
\]

with an assumption that \( \varepsilon_i \sim N(0, \sigma^2) \) [2].

2.3. Diagnostic of the assumption of multiple linear regression model
The following are assumption tests of multiple linear regression model.

2.3.1. Normality test. The Normality test is the most important form of classical assumption testing. In this study, the data should approach the normal distribution. Normally distributed assumption can be determined through normal probability plots [3].

2.3.2. Multicollinearity test. The Multicollinearity test aims to identify whether the independent variables in the regression model are correlated. Variance Inflation Factor (VIF) is a value to detect the symptoms of multicollinearity of independent variables with the following decision criteria: if VIF value is less than 10, there is no multicollinearity between two independent variables on the regression model; if VIF value is greater than 10, there is multicollinearity between two independent variables on the regression model [3].

2.3.3. Autocorrelation test. In general, the autocorrelation test is used to detect whether there is a correlation between the residuals in the current period (t) and the residuals in the previous period (t–s). The autocorrelation symptoms can be detected using the statistical Durbin-Watson test with the following decision criteria. If the value of D–W is less than -2, it means there is a positive autocorrelation in residuals; if the value of D–W is in the interval (-2, +2), it means there is no autocorrelation; and if the value of D–W is larger than +2, it means there is a negative autocorrelation in residuals. The hypothesis for this test is \( H_0 \) there is no autocorrelation, and \( H_1 \) there is autocorrelation. If the significance level \( \alpha = 0.05 \), then \( H_0 \) is rejected if p-value < \( \alpha \).
2.3.4. Heteroscedasticity Test. Heteroscedasticity test is to determine whether the variance of residuals is homogeneity. The plot of residuals vs predicted values of y shows the distribution of residuals. If the plot spreads randomly resembling figure 1a, then it can be concluded the assumption of homoscedasticity is met. If the plot results in a funnel resembling figure 1b, it appears that the variance increase as the predicted value increases; therefore, it can be concluded that the assumption of homoscedasticity is not met [3].

2.4. Spatial weighted matrix
The spatial weighted matrix $W$ is a matrix whose elements express the relationship between the observed regions. $w_{ij}$ is an element of the matrix $W$ on the ith row of the jth column, which denotes the relationship between region i and j for $i, j = 1, 2, ..., n$. The matrix form $W$ is as follows [4].

$$
W = \begin{bmatrix}
w_{11} & w_{12} & \cdots & w_{1n} \\
w_{21} & w_{22} & \cdots & w_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
w_{n1} & w_{n2} & \cdots & w_{nn}
\end{bmatrix}
$$

The $w_{ij}$ value is based on the contiguity between the neighboring regions. The value of $w_{ij}$ is a binary code. The weighted formula with binary code is as follows: $w_{ij}$ equals one if region i and j are adjacent, 0 otherwise.

There are several types of boundary borders namely, Rook, Bishop, and Queen contiguity. In this study, the Queen contiguity is used to define a weighted matrix: $w_{ij}$ equals one if regions i and j are adjacent, and 0 otherwise.

2.5. Moran’s I test
Moran’s I test is a spatial statistical test to determine whether spatial dependency occurs in a variable that has location information. Spatial autocorrelation is the correlation between variables by itself based on location.

Suppose there are values of variable $X_i$ and $X_j$ where $i \neq j$, $i = 1, 2, ..., n$, $j = 1, 2, ..., n$ with $n$ observations, the formula of Moran’s I is:

$$
I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}(X_i - \bar{X})^2} \quad (-1 < I < 1)
$$

with $\bar{X}$ is the average of the variable $X$ and $w_{ij}$ is an element of a weighted matrix. The expected value of Moran's I is [5]:

![Figure 1. The plot of heteroscedasticity test](image-url)
The hypothesis of Moran’s I test statistics is:

- \( H_0 : I = 0 \) (There is no spatial autocorrelation)
- \( H_1 : I \neq 0 \) (There is spatial autocorrelation)

Test statistic:

\[
Z = I - \frac{E(I)}{\sqrt{\text{Var}(I)}}
\]

Decision criteria: rejected \( H_0 \) if \( |Z| > Z_{a/2} \).

### 2.6. Linear regression model as an approach to analyze spatio-temporal data

Related to the model selection, the dependent variable in this study is the number of crimes, which is assumed to have a normal distribution; therefore, linear regression is an appropriate model. Beside covariates of every district, the data in this study consists of time and location information. To analyze the data with these components simultaneously, spatio-temporal modeling can be used as a multiple linear regression approach by inserting spatial element \( \textbf{WY} \). Matrix \( \textbf{W} \) is a spatial weighted matrix that states the inter-area linkages [6]. Interconnection between times will be expressed using dummy variables in this model. The following is the multiple linear regression model as an approach to analyze spatio-temporal data.

\[
Y_{it} = \alpha_1 D_{i,t-1} + \alpha_2 D_{i,t-2} + \cdots + \alpha_m D_{i,t-m} + \rho \textbf{W} Y_{it} + \beta_0 + \beta_1 X_{pit} + \cdots + \beta_p X_{pit} + \cdots + \beta_k X_{kit} \tag{2}
\]

- \( Y_{it} \) is the value of the dependent variable at location-i and time-t
- \( D_{i,t-r} \) is the value of the dummy variable at location-i and time \((t - r)\)
- \( \rho, \beta_0, \beta_1, \ldots, \beta_k \) are parameters of the model
- \( \alpha_1, \ldots, \alpha_m \) coefficient for dummy variable to \((t-r)\)
- \( \textbf{W} \) is a weighted spatial \( N \times N \) matrix
- \( X_{pit} \) is the value of the independent variable in the i-location, t-time; \( p = 1, \ldots, k; i = 1, \ldots, N; t = 1, \ldots, T \)
- \( m \) is the number of observation periods less than one
- \( k \) is the number of independent variables.

### 3. Results and discussion

The results are divided into two parts, descriptive statistics and modeling analysis.

#### 3.1. Descriptive statistics

Table 1 and figure 2 show the six districts with highest proportion of crimes based on the population, which are Sibolga, Tebing Tinggi, Padangsidempuan, Tanjungbalai, Samosir and Gunungsitoli.

#### 3.2. Linear regression analysis

For the first step, multiple linear regression is used to analyze the data and to test the condition of the data. The method of parameter estimation of linear regression model is the Ordinary Least Square (OLS) method. Table 2 is a summary of the result of parameter estimation along with partial testing; and VIF value of each variable: the number of unemployed, the number of poor, population density, human development index, the number of places of worship, the economic growth rate and the number of families with a social psychological problem.
### Table 1. The six districts with the highest proportion of crimes.

| Districts/City    | 2013  | 2014  | 2015  | 2016  |
|------------------|-------|-------|-------|-------|
| Sibolga          | 0.062 | 0.140 | 0.149 | 0.003 |
| Tebing Tinggi    | 0.050 | 0.148 | 0.141 | 0.005 |
| Padangsidempanan| 0.039 | 0.048 | 0.048 | 0.004 |
| Tanjungbalai     | 0.032 | 0.088 | 0.092 | 0.003 |
| Samosir          | 0.031 | 0.045 | 0.045 | 0.002 |
| Gunungsitoli     | 0.030 | 0.045 | 0.093 | 0.003 |

**Figure 2.** The six districts with the highest proportion of crimes.

### Table 2. Parameter estimate of regression and VIF.

| Variables                            | Estimate | Std. Error | t value | Pr(>|t|) | VIF  |
|--------------------------------------|----------|------------|---------|---------|------|
| (Intercept)                          | -3020    | 1330       | -2.275  | 0.0246  | 2.3849 |
| Number of unemployed                 | 0.0301   | 0.00702    | 4.289   | 0.0000358 | 5.1040 |
| The number of poor                   | 0.0380   | 0.00344    | 11.066  | < 2e-16 | 3.3498 |
| Population density                   | 0.305    | 0.0535     | 5.696   | 0.000000084 | 2.1581 |
| Human development index              | 32.4     | 18.9       | 1.713   | 0.0892  | 1.7635 |
| Number of places of worship          | -0.287   | 0.111      | -2.595  | 0.0106  | 2.3684 |
| Economic growth rate                 | 21.4     | 61.9       | 0.346   | 0.7297  | 1.1355 |
| Number of families with a social psychological problem | 0.246 | 0.485 | 0.507 | 0.6132 | 1.3373 |

The estimation of linear regression model is:

\[ \hat{Y} = -3.020 + 0.0301 X_1 + 0.0380 X_2 + 0.305 X_3 + 32.4 X_4 - 0.287 X_5 + 21.4 X_6 + 0.246 X_7. \]

Based on the table 2, it can be seen that number of unemployed, the number of poor, population density, human development index, the number of places of worship, are significant effects. The values of Variance Inflation Factor (VIF) in table 1 are less than 10 for all independent variables so that it can be concluded that there is no multicollinearity.
Figure 3, plot of $\hat{Y}$ vs error, shows that the assumption of error is not met, variances of error are not uniform. The spots appear to be non-randomly distributed but form a certain pattern that is quite obvious. This means there is heteroscedasticity on regression model error, so the regression model is not recommended for modelling the number of crimes acts in North Sumatra Province.

Figure 4, QQ plot regression, shows that, the errors are not normally distributed. The Quantile error plot is not entirely located on a straight line.

The assumption of independence of the error or absence of autocorrelation is tested by Durbin-Watson (DW) test which gave the result that p-value is equal to 0.02491 so that the null hypothesis is rejected which means autocorrelation occurs.

![Figure 3. Plot of $\hat{Y}$ vs error](image1.png)

![Figure 4. QQ plot regression](image2.png)
3.3. Autocorrelation spatial
The number of crime acts in North Sumatra Province is represented as a spatial data. Event in a location tends depend on the event in the neighborhood. Moran’s I test is used to test the spatial dependency or to detect the autocorrelation. The hypothesis test of Moran’s I is:
\[ H_0 : I = 0 \quad \text{(There is no autocorrelation between locations)} \]
\[ H_1 : I \neq 0 \quad \text{(There is autocorrelation between locations)} \]
The output of Moran’s I using Geoda program is as follows table 3.

3.4. Linear regression model as an approach to analyze the spatio-temporal data
The model in equation 2 is used to analyze the data. Table 4 is a summary of the output of the linear regression model as an approach to analyze spatio-temporal data. Regression modeling aims to know whether time and place factors affect the occurrence of crime in 33 districts/cities of North Sumatra Province, also to know which variables affect the number of crimes.

The following is the obtained multiple linear regression model:

\[ \hat{y} = -5030 - 218D_{2013} - 42.8D_{2014} + 236D_{2015} - 0.1277WY + 0.0379X_{1it} + 0.0364X_{2it} + 0.2133X_{3it} + 0.630X_{4it} - 0.0995X_{5it} + 60.5X_{6it} + 0.189X_{7it} \]

The value of the determination coefficient, \( R^2 = 0.8967 \) and \( R^2 \) (adjusted) = 0.8872. This means that the 89 % variation of the number of criminal acts can be explained by the variance of the seven independent variables while the remaining 11 % is explained by variables outside the model. A significantly negative coefficient value indicates that if a district/city with a criminal case is surrounded by an area with a reduced number of criminal acts, the number of criminal acts in each of the surrounding areas is reduced by the coefficient \( p \) multiplied by the average of the area that has the number of crimes around it.

| Table 3. The results of Moran’s I test |
|---------------------------------------|
| Moran’s I | Value | Z | P-value | Decision |
|------------|-------|---|---------|----------|
| 0.4568 | 4.2462 | 0.00002 | Rejected \( H_0 \) |

| Table 4. The output of the linear regression model as an approach to analyze spatio-temporal data. |
|-----------------------------------------------|
| Variables | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|-----------|--------|---------|
| (Intercept) | -5030 | 1330 | -3.784 | 0.000242 |
| Di2013 | -218 | 208 | -1.045 | 0.297899 |
| Di2014 | -42.8 | 202 | -0.212 | 0.832239 |
| Di2015 | 236 | 207 | 1.137 | 0.257783 |
| WY | -0.127 | 0.0259 | -4.914 | 2.86E-06 |
| Number of unemployed | 0.0379 | 0.00744 | 5.097 | 1.30E-06 |
| The number of poor | 0.0364 | 0.00338 | 10.785 | < 2e-16 |
| Population density | 0.213 | 0.0524 | 4.055 | 8.96E-05 |
| Human development index | 0.630 | 18.9 | 3.325 | 0.001172 |
| Number of places of worship | -0.0995 | 0.0109 | -0.916 | 0.36167 |
| Economic growth rate | 60.5 | 59.0 | 1.027 | 0.306631 |
| Number of families with a social psychological problem | 0.189 | 0.446 | 0.424 | 0.672014 |
Interpretation of $\rho$ (the coefficient of WY): The number of criminals in a district is affected by the average of the number of criminals of the districts around it times the value of $\rho$. In other words, the number of criminals in a district is affected by $\rho W Y$.

The coefficient value of unemployment factor ($X_{1, it}$) in 2013, 2014, 2015 and 2016 is 0.0379 which means that if there is a number of unemployment in regional/city t-year 100 people, will have an impact on the chance of finding crimes in region/city to-i t-year as many as 4 people. Coefficient value of the factor of the number of poor people ($X_{2, it}$) of 0.0364 which means that if there are number of poor people in the district/city to-i year t 100 people, will have an impact on the chance of finding criminal cases in to-i district/city t-year for up to 4 people, so that the government of North Sumatra Province pays more attention to the welfare of the population. The coefficient value of the population density factor ($X_{3, it}$) of 0.213 means that if there is an increase in population density in the to-i region/city t-year 10 as much as 10 inhabitants/km$^2$ will affect on the chances of finding criminal cases in to-i region/city t-year for up to 1 person.

4. Conclusion
According to the result of the regression analysis in this study, the number of unemployed, the number of poor, population density, and human development index are a statistically significant effect to the number of criminal acts in North Sumatra Province.

The number of criminals in a district is affected by the average of the number of criminals of the districts around it. The time of the incident did not affect the occurrence of the number of criminal acts in North Sumatra Province; in other words, there is no difference in the number of criminals from year to year.

Based on the references, economic growth rate and (or) number of families with a social psychological problem are strongly related to the number of crimes in a particular area. But, the results of this study, the economic growth rate does not affect the number of crimes. This condition can be caused technically, that is modeling should be improved, by change the role of variables. In other words, the study can be continued with the economic growth rate as the dependent variable and the number of crimes as the independent variable. This idea is based on the report from the sociology and economic study that the high number of criminals affects the investor as the actor of economic growth to continue their business.

Acknowledgments
This work was financially supported by Universitas Indonesia under research grant PITTA 2018 with grant contract number 2337/UN2.R3.1/HKP.05.00/ 2018.

References
[1] Badan Pusat Statistik 2016 *Statistik Kriminal 2016* (Jakarta: Badan Pusat Statistik)
[2] Rencher A C and Christensen W F 2012 *Methods of Multivariate Analysis* (Canada: John Wiley & Sons)
[3] Montgomery D C, Peck E A and Vining G G 2012 *Introduction to Linear Regression Analysis* (New York: John Wiley & Sons)
[4] Fotheringham A S, Rogerson P 2009 *The SAGE Handbook of Spatial Analysis* (Los Angeles: SAGE Publications)
[5] Lee J and Wong D W S 2001 *Statistical Analysis with Arcview GIS* (New York: John Wiley Sons)
[6] Millo G and Piras G 2012 *J. Stat. Softw.* 47 1-38
[7] Purnomo S 2016 *Pengaruh Kriminalitas Terhadap Dunia Ekonomi* available at http://www.medanbisnisdaily.com/news/read/2016/06/25/242283/pengaruh-kriminalitas-terhadap-dunia-ekonomi