Science Program Comprehensive Assessment

# Crime rates in the United States
# Multiple Linear Regression Model

Imene Mesli (2242755)
Salma Sadouk (2237888)

Probability and Statistics
201-HTH-05 sect. 00001
Presented to Ivan Ivanov

December 20, 2024

# Table of contents:

# Abstract:

Crime rates refer to the frequency of criminal activities within a specific area over a given period, often measured per 1,000 or 100,000 people. These rates help assess the safety of communities and are influenced by various factors. In this paper, we will focus on analyzing the crime rates in the United States and what factors affect them. In order to do so, we will review the basics of multiple linear regression models and explain the theory behind it. Using the discussing concepts, we aim to build a model that accurately predicts crime rates in the United States based on the selected variables.


# Introduction:

Simple linear regression is a commonly used method to analyze the relationship between two variables. However, many situations involve more than two variables. For instance, the likelihood of crime rates increasing in certain regions is influenced by multiple factors. These factors might be socioeconomic conditions, law enforcement presence, or even the percentage of employment. High crime rates can have significant consequences, including economic decline or reduced quality of life. Analysing these tendencies helps identify the underlying causes and improve public safety. To give some perspective, Venezuela is said to have the highest crime rate in the world with an index of 82.1, which is caused by issues like corruption, economic changes as well as social challenges. While Switzerland and Japan are among the countries with the lowest crime rates since they have effective law enforcement as well as strict gun laws.

This paper will focus on determining these factors that impact the crime rates in the United States with the help of the multiple linear regression model. In fact, in the case where numerous variables are influencing the response variable, multiple linear regression models are used, since they allow the analysis of several predictors simultaneously. The goal is then to create this sort of model that predicts outcomes more accurately by carefully selecting significant variables while avoiding the inadequate ones, and thus ensuring both effectiveness and accuracy. In order to do that, we will have to follow the theory behind this model, find an appropriate dataset and analyze the latter with the help of a Jupyter Notebook provided in class.

# Multiple Linear Regression Model:

Multiple linear regression model that connects a response y-variable to several predictors variables ($x_1$,$x_2$,...,$x_k$) can be expressed with the following equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \dots \ + \beta_k x_k + \varepsilon \qquad (1)$$

The term "linear" in the model indicates that it is linear in its predictors, $x_1$, $x_2$, ..., $x_n$, meaning that each of these predictors is multiplied with its corresponding regression coefficient. These coefficients are unknown model parameters. The "regression" term in the model's name refers to the regression function, being the sum of these coefficients times the predictors ($x_j$). For the first regression coefficient, $\beta_0$, the concept of $\beta_j$ multiplied by an x-variable is still applied, with that predictor variable being the constant function "1".

We suppose that the $\varepsilon$ follows the normal distribution, with a mean of 0 and a standard deviation of $\sigma^2$. It is also referred to as the noise, which is the random error constituent that describes the discrepancy between the responses and the mean value of y. This general form of the linear regression model represents in the k-dimensional plane, a hyperplane, with a $\beta_0$ being its intercept, where a certain number "*k*" of independent predictors estimate the response variable (*y*).

Multiple linear regression models can be enlarged to nonlinear predictors as well. In fact, each *x*-variable can be either a predictor variable or can even be a transformation of predictor variables, for instance a square of predictor variable ($x_1=x^2$), or two regressor variables which are multiplied together ($x_5= x_2x_4$). These nonlinear transformations of regressor variables allow the model to incorporate the nonlinear relationships between the *x*-variable (predictors) with the response variables (*y*), thus making the model more efficient/useful.

To be able to estimate the $\beta$ parameters, values that minimize the sum of squared errors (SSE) from the sample data will be considered. This technique is called *Least Squares Estimation of Parameters*. This method will enable reasonable predictions with this estimation of the regression coefficients since it is not possible to describe the true relationship between the response and the predictors. In fact, since data do not fit in an exactly straight hyperplane, the model must rely on predictions that approximate the scattered distribution of the data This technique can also be called *Line of Best Fit*. To represent the sample estimate of the $\beta$ parameter, it will be denoted by the letter $b$ (e.g. $b_0$ for $\beta_0$, $b_1$ for $\beta_1$,..., $b_k$ for $\beta_k$ ).

The following figure describes the technique of least squares estimation of parameters for a simple linear regression.

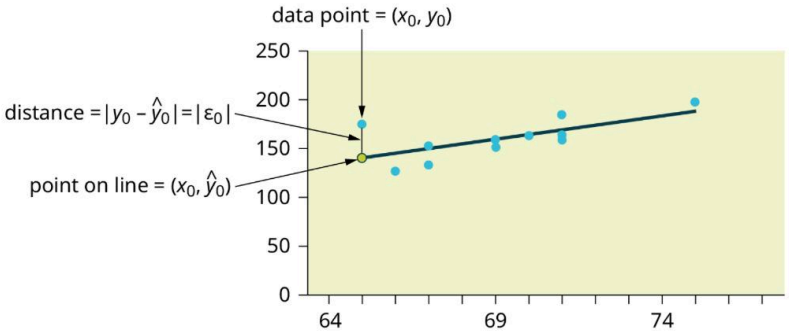


***Figure 1.*** *Method of Least Squares.* (OpenStax).

This model aims to find the "best fit line" for a set of data points by making sure the total vertical distance between the actual data points ($y_i$) and the predicted points on the line ($\hat{y}_i$) is as small as possible. The model works to minimize the sum of the SSE so the line fits the data as closely as possible. When two predictors are being considered, this model will have the form of a plane. As for more than two predictors, it will yield a hyperplane, which will be the model used for this project. It will estimate the model's coefficient values.

The term $\left|y_i - \widehat{y}_i\right| = \left|\varepsilon_i\right|$, (where i = 1, 2, 3, ..., n for each data point) describes the error, also known as the residual. Its absolute value measures the vertical distance between the two values of y, the actual data point and the estimated one. If the observed data point is situated above the line, the residual will then be positive and the prediction will be an underestimation of the actual data point. On the other hand, if the observed value is under the line, the residual is negative and the prediction is an overestimation of the actual value.

The function related to the Least Squares concept with the residuals is as follows:

$$\varepsilon_i = y_i\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}\right) \tag{2}$$

where i = 1, 2, …, n and j = 1, 2, …, k in $x_{ij}$ which represent the i-th measurement and given that n > k observations. Each observed data point is written as ($x_{i1}$, $x_i$, …, $x_{ik}$, $y_i$), and all these data points follow the model outlined with the first equation.

Let the Sum of Squared Errors (SSE) function be:

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij}\right)^2 \tag{3}$$

This quantity is the one that should be minimized by reducing the L with respect to the β values ($\beta_0, \beta_1, \ldots, \beta_k$). This will allow to form a stronger prediction equation which is as follows:

$$\frac{\partial L}{\partial \beta_i} = -2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{1j}\right) x_{ij} = 0 \tag{4}$$

In order to solve for the regression coefficients ($\beta_i$), it is best to write the stronger prediction equation using matrix notation. This notation is also recommended when there are a large number of predictors since it is more efficient. Knowing this, it is possible to rewrite Equation 1 in the following form:

$$y = X\beta + \varepsilon \tag{5}$$

The matrix notation would be as follows:

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{6}$$

The solutions of the least squares equations will be as follows:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{7}$$

The symbol of the hat on the top of the beta signifies the estimated values found in the data. When it comes to the fitted model predictions, the equation would be as follows:

$$\hat{y} = X\hat{\beta} \tag{8}$$

and its residuals would be expressed in the subsequent vector:

$$e = y - \hat{y} \tag{9}$$

As mentioned previously, the noise term ($\varepsilon$) is normally distributed with a mean of 0 and a variance of $\sigma^2$ . It is also independent and identically distributed. Knowing this, Equation 7 is an unbiased estimator for $\beta$, meaning that it is possible to demonstrate that the value of $\hat{\beta}$ is

expected to be centered on β through a series of mathematical operations and thus be written as follows:

$$E(\hat{\beta}) = \beta \tag{10}$$

As for $\sigma^2$ , when the three criterias of the noise term are satisfied, the unbiased estimator would be the subsequent equation:

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SSE}{n-p} \tag{11}$$

where p = k + 1, representing the number of parameters in the model and n denoting the sample size.

# Analysis of the regression

Many different statistics can be used to analyze a regression model. In this multiple linear regression, a large number of predictors are individually assessed for their significance and their contribution to explaining the crime rate. Hypothesis testing on the different regression coefficients using t-tests, as well as the coefficient of multiple determination, $R^2_{adj}$, are used to examine the model's reliability and evaluate the relevance of the predictors in explaining the variation in crime rates.

Hypothesis Testing on Individual Regression Coefficients

Hypothesis Testing is used in a multiple regression model to help determine whether the predictors significantly contribute in predicting or explaining the model, as in the crime rates. To determine whether each variable is a useful predictor, the following test could be done:

$$H_o : \beta_j = 0 \qquad \text{vs} \qquad H_1 : \beta_j \neq 0, \tag{12}$$

where the null hypothesis ($H_o$) refers to the case in which $x_j$ and $y$ are not linearly related. If $H_o$ is rejected, then it could be interpreted that the variable in question, $x_j$, significantly contributes to the model. Failing to reject the hypothesis would imply that $x_j$ is not a relevant predictor for the model, given that other predictors are present. The regressor $x_j$ can then be deleted from the model.

The t-statistic is:

$$t_0 = \frac{\widehat{\beta}_j - 0}{\sqrt{\sigma^2 c_{jj}}} = \frac{\widehat{\beta}_j - 0}{se(\widehat{\beta}_j)} \; ; (n - p)\, df, \tag{13}$$

where $c_{jj}$ is the $j^{th}$ diagonal element of $(X^T X)^{-1}$.

Each coefficient $\beta_j$ in the model will have its p-value calculated on the statistical software, Python, using the t-statistic and the significance level $\alpha = 0.05$. We fail to reject the hypothesis when the computed p-value is superior to α, implying that the predictor is not significant enough to conclude it has an effect on the dependent variable, and that the variance observed may be due to chance. It does not provide strong evidence of a meaningful correlation and can therefore be removed from the model. When the p-value is inferior to α, it

can be concluded that the predictor is significant enough in the model. The null hypothesis is then rejected, and the alternative hypothesis can be accepted.

This t-test is useful for determining which predictors significantly contribute to explaining the variance in the dependent variable, helping identify the most important predictors to build an effective model. For accurate results, the predictors with high p-values should be removed one by one using backwards elimination, as the significance of each predictor might depend on the presence of others.

$R^2$ and $R^2_{adj}$

In linear regression, the coefficient of determination, $R^2$, is a statistical measure that indicates the proportion of variation in $y$ that can be explained in a model by its predictors. It measures how much of the information of the dataset can be captured by the model.

The general formula for $R^2$ is:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \tag{14}$$

where:

$$SS_E = \sum_{i=1}^{n} \varepsilon_i^2 \tag{15}$$

$$SS_T = \sum_{i=1}^{n} y_i^2 - \frac{1}{n} \left( \sum y_i \right)^2 \tag{16}$$

$$SS_R = SS_T - SS_E \tag{17}$$

$SS_E$ refers to the error sum of squares evaluates the regression model accuracy by measuring the difference between the observed and the predicted values. $SS_T$, or the total sum of squares, measures the total variability in the dependent variable, and the regression sum of squares ($SS_R$) measures how well the regression model fits the dataset. An $R^2$ of 1 implies that all of the predictors are deterministic and that the model explains all variability, while an $R^2$ of 0 suggests that the predictors explain none of the variability observed in $y$, and that no model can be made out of those variables.

Although the $R^2$ increases every time a variable is added, it does not take into consideration whether the added variable actually improves the multiple linear regression model, making it harder to interpret properly. Adjusted $R^2$ is more relevant to use for the analysis of a multiple linear regression model, as it helps identify which predictors should be included or excluded. It is a useful statistic for preventing overfitting by including predictors that are not very relevant to the model:

$$R^2_{adj} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} \tag{18}$$

To be able to build the confidence and prediction intervals in multiple linear regression, it is needed to take into consideration once again the criterias of the noise term, being that it is independent, identically and normally distributed and has a mean of 0 and variance of $\sigma^2$ . In addition, let ( n - p ) represent the number of degrees of freedom when it comes to the following intervals.

1. Confidence Intervals on the Regression Coefficients

Considering Equation 7, a 100(1 - **α**) % of this confidence interval for β would be as follows:

$$\widehat{\beta_j} - t_{\alpha/2}\sqrt{\widehat{\sigma^2}c_{jj}} \leq \beta_i \leq \widehat{\beta_j} + t_{\alpha/2}\sqrt{\widehat{\sigma^2}c_{jj}} \; ; \; (n-p)df \tag{19}$$

2. Confidence Interval on the Mean Response:

A 100(1 - **α**) % of this confidence interval at the point $x_{01}, \ldots, x_{0k}$ is as follows:

$$\widehat{\mu}_{y|x_0} - t_{\alpha/2}\sqrt{\widehat{\sigma^2}x_0^T(X^TX)^{-1}x_0} \leq \mu_{y|x_0} \leq \widehat{\mu}_{y|x_0} + t_{\alpha/2}\sqrt{\widehat{\sigma^2}x_0^T(X^TX)^{-1}x_0} \tag{20}$$

3. Prediction Interval for Future Observation:

A 100(1 - **α**) % of this prediction interval of the response, *Y*, at $x_{01}, \ldots, x_{0k}$ is as follows:

$$\widehat{y}_0 - t_{\alpha/2}\sqrt{\widehat{\sigma}^2(1 + x_0^T(X^TX)^{-1}x_0)} \leq y_0 \leq \widehat{y}_0 + t_{\alpha/2}\sqrt{\widehat{\sigma}^2(1 + x_0^T(X^TX)^{-1}x_0)} \tag{21}$$

There are multiple ways to select the variables that will be needed for this project and build the model. In fact, there is first stepwise regression which is, in simple terms, adding or removing predictors based on the F-test at every step. There is also forward selection which is essentially predictors that are added one at a time. Finally, there is backwards elimination which will be the one used in our model.

This way of selecting our predictors and building our model consists of starting with all the variables possible and eliminating one by one the ones that are not significant enough. In order to remove these variables and keep only the meaningful ones, we will be using the p-values of the hypothesis testing of individual regression coefficients and the values of $R^2$. In fact, with the help of Jupyter Notebook and Google Colab, all the regressors possessing a high p-value will be eliminated. An acceptable value for the p-value would be less than 0.05. For the value of $R^2$, it should be around 0.5 to 1. This process will end once all the predictors present in the model are significant and thus none can be eliminated.

Another tool that will be used to eliminate some predictors will be using the concept of multicollinearity. When this notion is present in a model, it causes a decrease in the precision in the estimations of the regression coefficients. In fact, we can expect that some of the predictors might be dependent on one another, or resemble one another greatly. To explain the with equations, if there is dependency, the matrix related to it ($X^TX$) will act as a singular matrix and thus does not have an inverse $(X^TX)^{-1}$. However, in order to estimate the values from the regression coefficients, an inverse matrix is needed since it will be used in Equation 7 ( $\hat{\beta} = (X^T X)^{-1} X^T y$ ). This can be analyzed with the help of heat maps which demonstrate the highly correlated predictors with the very much lighter and very much darker colour. In addition, multicollinearity can also be assessed using the condition number, which quantifies the sensitivity of the regression coefficients to small changes in the data. A high condition

number would indicate a higher degree of multicollinearity and greater instability in the estimations. In other words, considering that our baseline will be 100 for a condition number, a higher value than that would mean that some predictors are too dependent on each other and thus doesn't produce good estimates, but if it is lower, than the predictors are significant and independent and the model is accurate and efficient enough.

# Model Building:

With all of this taken into consideration, the dataset that will be used in order to analyze the crime rates in the United States was collected by the UC Irvine studies on communities and crime done in 2009. In the start there will be 127 predictors which will be analyzed in relation to the dependent variable, *y*, representing crime rates, to determine their significance as well as their impact on the model. A multiple linear regression model will be built on Google Colab and the Jupyter Notebook associated with the study's dataset.

# Results

The purpose of this project was to build a model that predicts the factors affecting the crime rates in the United States. Starting with a large set of 127 variables covering various domains such as demographic or socioeconomic, the predictors are all initially assumed to account for variations in the model, and are used to build the initial multiple linear regression model. A t-test backwards elimination with a significance level $\alpha = 0.05$ was used to identify the most useful predictors from the statistically insignificant ones, bringing them down to 39. For example, the predictor 'householdsize' displayed the highest p-value of 0.999 and had to be

removed first due to its lack of relevance in the model, followed by 'PctTeen2Par' with a p-value of 0.983, etc.

However, the high condition number of $3.48 \cdot 10^3$ suggests strong multicollinearity between some of those 39 remaining predictors. In an attempt to build a better model, a correlation matrix was generated to identify redundant predictors, using the method of backwards elimination, starting with the variables with the strongest and more frequent correlations. This process, along with t-testing, allowed to bring down the condition number and build a more accurate and stable final model with 10 predictors, achieving an adjusted $R^2$ of 0.652 and a final condition number of 25.6. These values suggest that the selected predictors explain approximately 65.2% of the variance in crime rates and that the multicollinearity in the model has been significantly reduced. This also indicates an increased confidence in the estimated coefficients and the predictive power of the model, as the variables are now less likely to be strongly influenced by some variation that could happen in the data. All predictors are relatively independent of each other, allowing for a more stable estimation of their individual effects on crime rates, which can be done by analysing their coefficients β:

**racepctblack** ($\beta_1 = 0.1984$): For a one-unit increase in the percentage of the population of Black people, the crime rate is expected to increase by 0.1984 units. This coefficient suggests a positive correlation between the percentage of Black people and crime rates in this model.

**pctUrban** ($\beta_2 = 0.0422$): For a one-unit increase in the percentage of population living in urban areas, the crime rate is expected to increase by 0.0422 units, suggesting a weak positive correlation between urbanization and crime rates.

**PctEmploy** ($\beta_3$= -0.0754): For a one-unit increase in the percentage of population that is employed, the crime rate is expected to decrease by 0.0754 units, suggesting a weak negative correlation between employment rates and crime rates.

**TotalPctDiv** ($\beta_4$= 0.2446): For a one-unit increase in the total percentage of divorced couples, the crime rate is expected to increase by 0.2446 units, indicating a positive correlation between divorce rates and crime rates.

**PctWorkMom** ($\beta_5$= -0.0658): For a one-unit increase in the percentage of working mothers, the crime rate is expected to decrease by 0.0658, suggesting a weak negative correlation between the variable and crime rates.

**PctIlleg** ($\beta_6$= 0.3051): For a one-unit increase in the percentage of illegitimate children, the crime rate is expected to increase by 0.3051, indicating a strong positive correlation to the crime rates, although it can be either a cause or an effect to the increase of crime rates.

**PctPersDenseHous** ($\beta_7$= 0.1959): For a one-unit increase in the percentage of persons living in dense housing, the crime rate is expected to increase by 0.1959, suggesting a positive correlation between population density and crime rates.

**HousVacant** ($\beta_8$= 0.1083): For a one-unit increase in the number of vacant houses, the crime rate is expected to increase by 0.1083 units, indicating a positive relationship between vacant houses and crime rates.

**PctHousOccup** ($\beta_9$= -0.0736): For a one-unit increase in the percentage of housing occupied, the crime rate is expected to decrease by 0.0736 units, suggesting a weak negative correlation between housing occupancy and crime rates.

**NumStreet** ($\beta_{10}$= 0.1858): For a one-unit increase in the number of streets, the crime rate is expected to increase by 0.1858 units, suggesting a positive correlation between the number of streets and crime rates.

The values of the coefficients are valid as long as all other factors are held constant. As it can be observed, none of the 10 predictors contradict another.

Let $x_1$= racepctblack, $x_2$= pctUrban, $x_3$= PctEmploy, $x_4$= TotalPctDiv, $x_5$= PctWorkMom, $x_6$= PctIlleg, $x_7$= PctPersDenseHous, $x_8$= HousVacant, $x_9$= PctHousOccup, $x_{10}$ = NumStreet.

**Linear model:**

$$y = 0.1984\,x_1 + 0.0422\,x_2 - 0.0754\,x_3 + 0.2446\,x_4 - 0.0658\,x_5 + 0.3051\,x_6 + 0.1959\,x_7 + 0.1083\,x_8 - 0.0736\,x_9 + 0.1858\,x_{10}$$

# Conclusion

Although this model dives into the relationship between some socioeconomic factors and crime rates, it is important to remember that correlation does not necessarily imply causation. For instance, PctIlleg, the predictor which shows the strongest positive correlation to crime rate, with a coefficient of 0.3051, could either contribute to or be a result of the increase of crime rates. Further studies should be done in order to establish cause-to-effect relationships and understand how each of these factors affect or influence crimes. Overall, the model resulted in an adjusted $R^2$of 0.652, explaining 65.2% of the variance in crime rates, and a final condition number of 25.6, proving that multicollinearity between predictors was significantly reduced.

# References


Ivanov, I.T. (2024). §27: Linear Regression. Probability and Statistics, 201-HTH-05, Lectures.

Openstax. Chapter 12: Linear Regression and Correlation, *Introduction Statistics 2e.* https://openstax.org/books/introductory-statistics-2e/pages/12-introduction.

PennState. 5.3: Multiple Linear Regression Model. https://online.stat.psu.edu/stat501/lesson/5/5.3.

Smalheiser, N.R. (2017). Correlation and Other Concepts You Should Know: Multiple Linear Regression Analysis, *Data Literacy*. https://www.sciencedirect.com/topics/medicine-and-dentistry/multiple-linear-regression-analysis.

World Population Review. Crime Rate by Country 2024. https://worldpopulationreview.com/country-rankings/crime-rate-by-country.

PennState. 1.5: The Coefficient of Determination, $R^2$. https://online.stat.psu.edu/stat501/lesson/1/1.5.

Mahata, Akashdeep. "Understanding Sum of Squares: SST, SSR, SSE - Akashdeep Mahata - Medium." *Medium*, 1 Dec. 2024, https://akashdeepmahata.medium.com/understanding-sum-of-squares-sst-ssr-sse-1dd8c6d2b6e4.