Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2023, 2023:105
https://doi.org/10.28919/cmbn/8165
ISSN: 2052-2541

# PROGNOSTIC RISK FACTORS INDUCING ACUTE HEPATITIS CONTAGION IN JAKARTA, INDONESIA: LINEAR PREDICTIVE MODEL APPLICATION

SELLY ANASTASSIA AMELLIA KHARIS[1,*], ARMAN HAQQI ANNA ZILI[2], AGUSTIANI PUTRI[3]
[1]Department of Mathematics, Universitas Terbuka, South Tangerang 15437, Indonesia
[2]Department of Mathematics, Universitas Indonesia, Depok 16424, Indonesia
[3]Department of Mathematics, Universitas Negeri Malang, Malang 65145, Indonesia



**Abstract.** Hepatitis is a serious global health issue caused by a variety of infectious viruses and noninfectious agents, affecting the liver organ of the human body. In Indonesia, hepatitis has been widely transmitted, including in DKI Jakarta, where the number of acute hepatitis cases is a major concern. In this study, we investigated the number of acute hepatitis cases in 44 sub-districts of DKI Jakarta, Indonesia. Using RStudio, we constructed 15 mathematical models to identify risk factors that induce hepatitis transmission. The models were then subjected to a global Anova test (f-test) to determine the factors that significantly impact the number of acute hepatitis patients. Our analysis found that several risk factors were strongly associated with hepatitis transmission in DKI Jakarta. One of the most significant factors was the number of infants who had been immunized with hepatitis B vaccine (HBO). This finding suggests that increasing the number of infants who receive the vaccine could have a significant impact on reducing the number of acute hepatitis cases in the region. Other factors that were found to be strongly associated with hepatitis transmission included the total population and the number of diabetics in each sub-district. To support our findings, we used the pairs function to observe various scatter plots, which helped to visualize the relationship between the risk factors and the number of acute hepatitis patients. Our study provides important insights into the factors that contribute to hepatitis transmission in DKI Jakarta and highlights the need for effective prevention and control strategies to reduce the burden of this disease. Overall, our findings have important implications for public health policy and practice, not only in Indonesia but globally.

*Corresponding author
E-mail address: selly@ecampus.ut.ac.id
Received August 08, 2023

**Keywords:** acute hepatitis; linear model; infection; risk factors; Jakarta.

**2020 AMS Subject Classification:** 92C60.

## 1. INTRODUCTION

Hepatitis is a disease that is widely dispersed worldwide, including in Indonesia, and is recognized as a serious global health issue [1]. According to WHO report, approximately 1.45 million individuals die each year after suffering from chronic hepatitis [2]. Viral hepatitis causes more deaths than other diseases such as malaria, tuberculosis, and HIV, and the number of cases is still rising since 1990 [3]. Hepatitis virus identified into five varieties, which are A, B, C, D, and E [4]. After Myanmar, Indonesia has the second-highest endemic hepatitis B infection rate in Southeast Asia [5]. In addition, the hepatitis virus has a harmful impact on the morbidity and mortality of individuals during the acute phase [6]. Therefore, it is necessary to preserve Indonesian people from acute hepatitis through concerted efforts from stakeholders, including the government and non-governmental organizations.

Hepatitis is an infection that specifically targets the liver organ in the human body. It is classified as acute when its infection persists for more than six months, while regular hepatitis normally contaminates people for less than six months [7]. People suffering from acute hepatitis not only experience symptoms ranging from mild to medium, but they also endure chronic complications [8]. Yet, most people experiencing hepatitis are unaware of its symptoms, mistaking decreased appetite and exhaustion as normal indications after extra work [9], [10]. If left untreated, acute hepatitis can lead to mortal effects on humans since other diseases caused by it are brain function drop, spleen magnification, and cirrhosis, among others [11]. The process by which humans undergoing acute hepatitis are identified by a medical center. They will undergo a sequence of examinations, including a blood check to identify heart function, and imaging checks such as biopsies, MRI, and ultrasound tests for heart cancer filtration [12].

The risk of transmission can be reduced by addressing the various risk factors associated with the disease [13]. Despite the availability of treatment, only two out of three individuals with acute hepatitis seek treatment in a timely manner [14]. The growth of hepatitis infection is unpredictable, with symptoms that are often insensible to some individuals until they experience them personally and reach the cirrhosis stage [15], [16]. The symptoms of hepatitis infection are observed from several physical manifestations, including feeling unwell, frequent exhaustion, decreased appetite, high fever, limb swelling, constricted blood vessels, fluid accumulation in the abdomen, upper abdominal pain, and jaundice or yellowing of the skin and eyes [17], [18].

The Ministry of Health has announced information regarding 18 suspected cases of chronic hepatitis with unidentified factors in January 2022. These cases were reported in several districts, including Jakarta, Bangka Belitung Island, East Java, West Java, West Sumatera, East Kalimantan and North Sumatera, with most cases occurring in Jakarta [19]. Preventing the transmission of hepatitis is crucial to improving public health. The purpose of this experiment is to develop a model to forecast the transmission of acute hepatitis and to identify the predictor variables that significantly influence the response variables. The model aims to predict the number of acute hepatitis cases in each district of Jakarta, with the goal of reducing the spread of the disease. Additionally, this study aims to analyze the correlation between predictor variables. The model for predicting the number of chronic hepatitis cases in any district of Jakarta has been successful, and the activities undertaken during this research have been beneficial. The study also examines the trends in how predictor variables affect the final data.

## 2. MATERIALS AND METHOD

This study involves data in the form of the number of acute hepatitis sufferers in districts of Jakarta. The data consists of 44 rows and 9 columns. Rows in the data show the number of sub-districts in DKI Jakarta studied, while the column in the data indicates eight predictor variables and one response variable. Eight predictor variables include the number of babies who had been immunized with HBO ($X_1$), the number of health workers ($X_2$), the number of health facilities ($X_3$), the number of residents ($X_4$), the number of residents who had proper sanitation ($X_5$), the number of drinking water facilities which met health standards ($X_6$), the number of diabetics ($X_7$), and the number of HIV sufferers ($X_8$). The combination of predictor variables is analyzed by using a multiple linear regression model in order to identify the influence exerted on chronic hepatitis transmission. The following general equation of multiple linear regression.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \tag{1}$$

Assumptions for multiple linear regression model:

a. $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$ is the deterministic part of model $y_i$ and $\varepsilon_i$ is the probabilistic or stochastic part of model between $y_i$ and are mutually independent.

b. $\varepsilon_i \sim NIID(0, \sigma^2)$.

c. There is no multicollinearity among independent variables.

F test is subsequently used to identify the suitability of the model by looking at how the influence of all the independent variables together on the dependent variable or to examine whether the regression model that we create is fit or unfit. We pose the hypothesis that will be verified on the F test as follows.

$H_0$ : $\beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$ ; k = 1, 2, 3, ….

$H_1$ : At least one $\beta_k \neq 0$ ; k = 1, 2, 3, ….

Decision rules :

Reject $H_0$ if $P_{value} < a = 0.05$, we declare that the model is fit, and the testing can continue.

Accept $H_0$ if $P_{value} > a = 0.05$, we declare that the model is unfit.

The next examination after knowing the F test is in the form of T test application which is beneficial for testing whether these parameters have a significant effect on the model used. The following is the hypothesis tested on the T test.

$H_0$ : $\beta_k = 0$ ; k = 1, 2, 3, ….

$H_1$ : $\beta_k \neq 0$ ; k = 1, 2, 3, ….

Decision rules :

Reject $H_0$ if $P_{value} < a = 0.05$, independent variables significantly effect on dependent variable.

Accept $H_0$ if $P_{value} > a = 0.05$, independent variables have no effect on the dependent variable.

This study also considers Variance Inflation Factor (VIF) that is useful for determining how much multicollinearity is in a set of multiple regression variables. Table 1 showing classification to generate decision-making.

TABLE 1. The classification of VIF value.

| **Value** | **Category** |
|---|---|
| VIF < 1.5 | Excellent |
| 1.5 ≤ VIF < 2 | Good |
| 2 ≤ VIF < 4 | Moderate |
| VIF ≥ 4 | Bad |

## 3. Exploratory Analysis Data

Data pre-processing is taken in order to retrieve some desired variables, as well as determine the interrelationships among variables. Data processing in this study utilizes the RStudio program.

### 3.1 Pre-processing

Data pre-processing completely involves four activities. The first activity is in the form of completing the blanks in the table with the average value. We conduct this action to fill the empty data that could possibly induce errors when data processing and calculation. Subsequently, we stipulate a method in order to complete the blanks with the average point of these variables. This method was selected with the argument that filling in empty values with the average score will not significantly affect the dataset. In the dataset, there are 2 blank data in $X_6$ column. Hence, it will be filled with the average score of $X_6$ column, which is 38. It will simplify further data processing and minimize calculation errors.

FIGURE 1. Fill in blank data with an average score.

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Y | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 |
| 2 | 0 | 11236 | 671 | 223 | 155275 | 126289 | 118 | 7,164 | 195 |
| 3 | 18 | 898 | 1977 | 85 | 22069 | 18553 | 6 | 4,877 | 113 |
| 4 | 282 | 11128 | 1407 | 201 | 116091 | 107434 | 57 | 8054 | 446 |
| 5 | 4 | 3220 | 3214 | 134 | 41895 | 39590 | 28 | 6,977 | 238 |
| 6 | 6 | 8476 | 1094 | 236 | 135134 | 84344 | 38 | 12,219 | 241 |
| 7 | 3 | 5353 | 712 | 143 | 72349 | 67867 | 11 | 6,806 | 143 |
| 8 | 8 | 6169 | 1127 | 125 | 52590 | 22023 | 44 | 8,254 | 155 |
| 9 | 194 | 7953 | 1833 | 184 | 91862 | 70445 | 112 | 11,752 | 270 |
| 10 | 28 | 888 | 1947 | 128 | 21434 | 16555 | 23 | 1,157 | 121 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 11 | 4 | 4613 | 1283 | 143 | 73317 | 51904 | 85 | 6227 | 328 |
| 12 | 31 | 4951 | 1101 | 232 | 89245 | 86427 | 34 | 10,723 | 256 |
| 13 | 95 | 6253 | 1085 | 143 | 77225 | 45707 | 31 | 6,475 | 249 |
| 14 | 0 | 1254 | 309 | 87 | 38886 | 26046 | 15 | 6,453 | 128 |
| 15 | 40 | 8727 | 1578 | 184 | 92358 | 63262 | 60 | 10630 | 321 |
| 16 | 2 | 2267 | 2624 | 185 | 43569 | 35547 | 35 | 6,412 | 206 |
| 17 | 4 | 4672 | 1573 | 194 | 72563 | 32629 | 15 | 9,419 | 204 |
| 18 | 240 | 7705 | 1753 | 165 | 134271 | 76948 | 112 | 13894 | 351 |
| 19 | 2 | 2724 | 1045 | 136 | 51259 | 45964 | 38 | 4,388 | 179 |

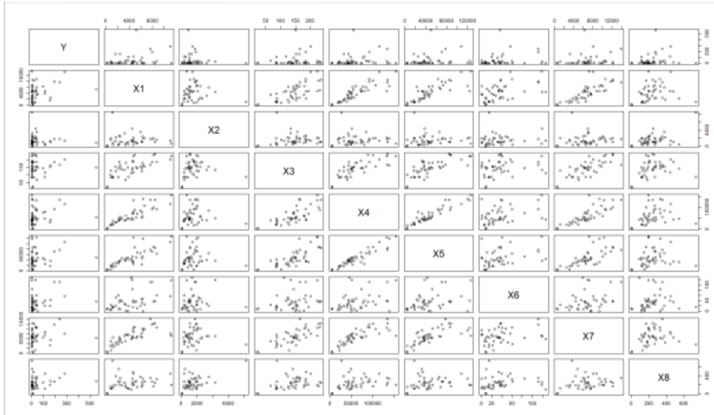
We subsequently inspect the correlation among the data provided. We process data correlation in order to identify the correlation between the predictor variables and the target variable. After recognizing the correlation between the target variable and the predictor variables or even among the target variables, we notice that the predictor variables ($X_i$) have no significant effect on the target variable (Y). In addition, this investigation finds a high correlation among variables $X_1$, $X_2$, $X_3$, and among variables $X_1$, $X_4$, $X_5$ and $X_7$ indicating a powerful relationship among several predictor variables (independent variables). This affair requires to be suspected as a case that have multicollinearity in the dataset owned. A multicollinearity occurs when there is a strong relationship between two or more independent variables in a multiple regression model. A multicollinearity must be avoided since it increases the possibility of rounding errors in estimating $\beta$ and standard errors. As a result, the regression output will be confusing and tend to be wrong. Further checking will be conducted in order to recognize and remove one or more correlated independent variables.

FIGURE 2. Correlation among variables.

| | Y | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 |
|---|---|---|---|---|---|---|---|---|---|
| Y | 1.0000000 | 0.27781848 | -0.110514286 | 0.13934663 | 0.15810051 | 0.153048828 | 0.19145698 | 0.17233918 | 0.15727708 |
| X1 | 0.2778185 | 1.00000000 | -0.039071523 | 0.70040829 | 0.86504771 | 0.806342153 | 0.50435192 | 0.77112298 | 0.19434867 |
| X2 | -0.1105143 | -0.03907152 | 1.000000000 | 0.08395658 | -0.01293943 | 0.006976645 | 0.07503674 | 0.05512551 | 0.51357287 |
| X3 | 0.1393466 | 0.70040829 | 0.083956581 | 1.00000000 | 0.75283330 | 0.743399441 | 0.26984060 | 0.73806598 | 0.24395661 |
| X4 | 0.1581005 | 0.86504771 | -0.012939433 | 0.75283330 | 1.00000000 | 0.928374397 | 0.37258892 | 0.80543754 | 0.22613557 |
| X5 | 0.1530488 | 0.80634215 | 0.006976645 | 0.74339944 | 0.92837440 | 1.000000000 | 0.33067517 | 0.70442970 | 0.14345184 |
| X6 | 0.1914570 | 0.50435192 | 0.075036735 | 0.26984060 | 0.37258892 | 0.330675168 | 1.00000000 | 0.34477471 | 0.08983855 |
| X7 | 0.1723392 | 0.77112298 | 0.055125510 | 0.73806598 | 0.80543754 | 0.704429701 | 0.34477471 | 1.00000000 | 0.17788261 |
| X8 | 0.1572771 | 0.19434867 | 0.513572875 | 0.24395661 | 0.22613557 | 0.143451841 | 0.08983855 | 0.17788261 | 1.00000000 |

We decide to conduct the next activity in the form of testing the description of the pairs among the predictor variables and the predictor with the target. The pairs data is used to describe the relationship between the predictor variable and the target variable. After recognizing the relationship between the target variable and the predictor or even among the target variables, we discover that the relations between the predictor variables ($X_i$) tends to be weak towards the target variable (Y). If you pay more attention, there is a strong relationship among predictor variables of $X_1$, $X_4$, $X_5$, and $X_7$. The data discovered reinforces the notion that there is multicollinearity in the dataset owned.

FIGURE 3. Several pairs among variables as predictors and between predictors with targets.




In the final section, we eliminate several outliers from the dataset. After having several considerations, we determine to release one data from Kembangan District because the value of Y variable is too far compared to Y value in other data related to the predictor variable data whose numbers are similar.

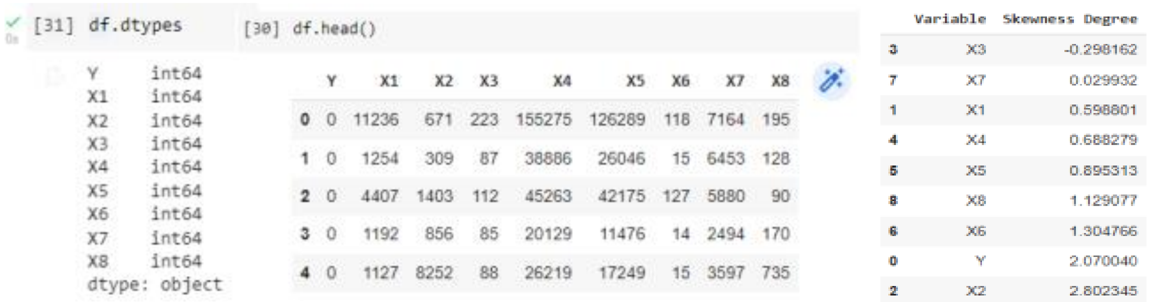
Figure 4: Data outlier and boxplot.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 42 | Duren Sawit | 194 | 7953 | 1833 | 184 | 91862 | 70445 | 112 | 11.752 | 270 |
| 43 | Kebon Jeruk | 240 | 7705 | 1753 | 165 | 134271 | 76948 | 112 | 13894 | 351 |
| 44 | Cengkareng | 282 | 11128 | 1407 | 201 | 116091 | 107434 | 57 | 8054 | 446 |
| 45 | Kembangan | 553 | 5260 | 909 | 151 | 54497 | 49677 | 36 | 6257 | 272 |
| 46 | **Total** | **2349** | **191147** | **69922** | **6354** | **2800316** | **2115299** | **1599** | **300424** | **10799** |

### 3.2 Dataset Visualization and Variable Analysis in General

In the initial stage, we must learn about several numeric variables since they are components in the model formulated, both the response variables and the predictor variables. Subsequently, the next stage asks us to describe data according to the table provided and visualize each variable. For instance, the graph of $X_3$ shapes the right skewed corresponding to a negative degree of skewness, while the rest are left skewed linking with a positive degree of skewness. The section of visualization is conducted after removing some outliers from the data.

FIGURE 5. Variable type analysis and skewness illustration.


[31] df.dtypes

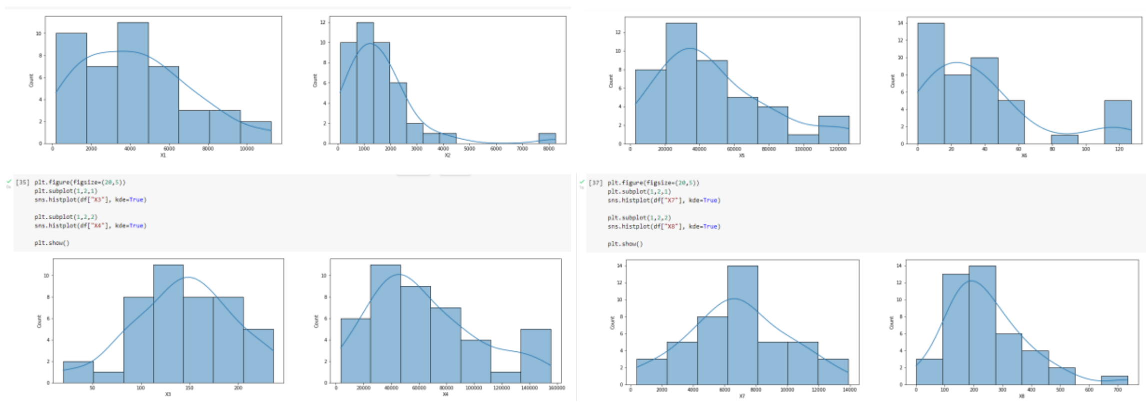
| | |
|---|---|
| Y | int64 |
| X1 | int64 |
| X2 | int64 |
| X3 | int64 |
| X4 | int64 |
| X5 | int64 |
| X6 | int64 |
| X7 | int64 |
| X8 | int64 |
| dtype: object | |

[30] df.head()

| | Y | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 11236 | 671 | 223 | 155275 | 126289 | 118 | 7164 | 195 |
| 1 | 0 | 1254 | 309 | 87 | 38886 | 26046 | 15 | 6453 | 128 |
| 2 | 0 | 4407 | 1403 | 112 | 45263 | 42175 | 127 | 5880 | 90 |
| 3 | 0 | 1192 | 856 | 85 | 20129 | 11476 | 14 | 2494 | 170 |
| 4 | 0 | 1127 | 8252 | 88 | 26219 | 17249 | 15 | 3597 | 735 |

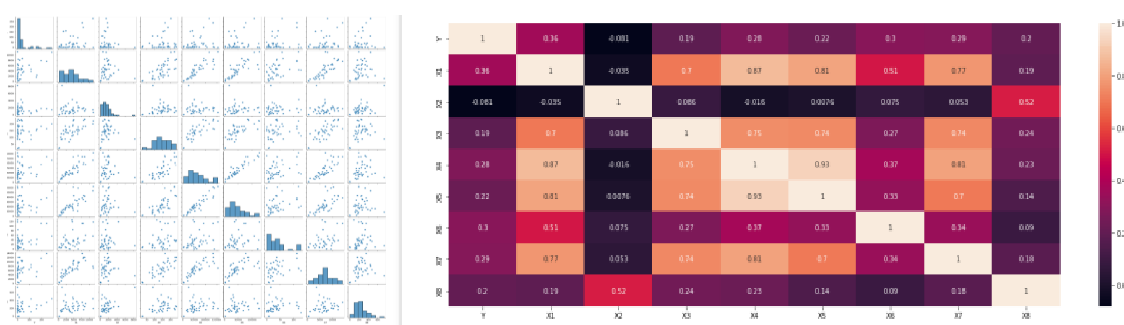
| | Variable | Skewness Degree |
|---|---|---|
| 3 | X3 | -0.298162 |
| 7 | X7 | 0.029932 |
| 1 | X1 | 0.598801 |
| 4 | X4 | 0.688279 |
| 5 | X5 | 0.895313 |
| 8 | X8 | 1.129077 |
| 6 | X6 | 1.304766 |
| 0 | Y | 2.070040 |
| 2 | X2 | 2.802345 |

FIGURE 6. Data visualization after removing outliers.



We continue to visualize the relationship among numeric variables along with a heatmap of the numeric variables.

FIGURE 7. Visualization of the relationship among numeric variables and its heatmaps.



The output indicates that the predictor variables ($X_i$) have no significant effect on the target variable (Y). In addition, we also notice a high correlation among $X_1$, $X_3$, $X_4$, $X_5$ and $X_7$. This claim demonstrates a strong relationship among these predictor variables or independent variables.

FIGURE 8. Correlation results.

```
       Y               X1             X2              X3               X4               X5               X6               X7              X8
 Min.   :  0.00   Min.   :  180   Min.   :  84   Min.   : 20.0   Min.   :  3362   Min.   :  2980   Min.   :  0.00   Min.   :  395   Min.   :  0.0
 1st Qu.:  2.00   1st Qu.: 2217   1st Qu.: 836   1st Qu.:115.0   1st Qu.: 39548   1st Qu.: 26663   1st Qu.: 14.50   1st Qu.: 4833   1st Qu.:151.0
 Median :  8.00   Median : 4407   Median :1310   Median :143.0   Median : 51259   Median : 39590   Median : 31.00   Median : 6806   Median :206.0
 Mean   : 41.77   Mean   : 4323   Mean   :1605   Mean   :144.3   Mean   : 63856   Mean   : 48038   Mean   : 38.12   Mean   : 6841   Mean   :244.8
 3rd Qu.: 41.00   3rd Qu.: 6211   3rd Qu.:1972   3rd Qu.:182.0   3rd Qu.: 86844   3rd Qu.: 66663   3rd Qu.: 46.00   3rd Qu.: 9060   3rd Qu.:324.0
 Max.   :282.00   Max.   :11236   Max.   :8252   Max.   :236.0   Max.   :155275   Max.   :126289   Max.   :127.00   Max.   :13894   Max.   :735.0

> cor(Data_Hepatitis_DKI)
            Y          X1          X2         X3          X4         X5         X6        X7         X8
Y   1.0000000  0.36355059 -0.08052090 0.18842106  0.28176607 0.22280617 0.30044903 0.2929739 0.20483426
X1  0.3635506  1.00000000 -0.03534764 0.70043854  0.86863832 0.80700410 0.50549271 0.7738418 0.19319492
X2 -0.0805209 -0.03534764  1.00000000 0.08574524 -0.01584176 0.00759826 0.07454474 0.0531493 0.51756065
X3  0.1884211  0.70043854  0.08574524 1.00000000  0.75424050 0.74341338 0.27009075 0.7390623 0.24352690
X4  0.2817661  0.86863832 -0.01584176 0.75424050  1.00000000 0.92933537 0.37251583 0.8052669 0.22748079
X5  0.2228062  0.80700410  0.00759826 0.74341338  0.92933537 1.00000000 0.33077265 0.7049502 0.14328833
X6  0.3004490  0.50549271  0.07454474 0.27009075  0.37251583 0.33077265 1.00000000 0.3446637 0.09015492
X7  0.2929739  0.77384181  0.05314930 0.73906230  0.80526695 0.70495016 0.34466375 1.0000000 0.17885549
X8  0.2048343  0.19319492  0.51756065 0.24352690  0.22748079 0.14328833 0.09015492 0.1788555 1.00000000
```

## 4. MODELLING AND MODEL SELECTION

After pre-processing data by eradicating outliers in the previous segment, data continue to process in R. An illustration of the process is provided in the following explanation.

### 4.1 First Regression Model

The first regression model is the primary model for this dataset, as it serves as the basis for selecting predictor variables and identifying multicollinearity. However, the p-values of the predictor variables $X_1$ to $X_8$ in this main effect model are larger than 0.05 and supported by a small R square, indicating that the model cannot accurately explain the variation in the response variable Y. The small R squared value also supports this conclusion.

FIGURE 9. The first regression output.

```
Call:
lm(formula = df$Y ~ df$X1 + df$X2 + df$X3 + df$X4 + df$X5 + df$X6 +
    df$X7 + df$X8)

Residuals:
   Min     1Q Median     3Q    Max
-82.18 -45.18 -13.85  19.15 172.93

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.162e+00  3.509e+01  -0.062    0.951
df$X1        7.444e-03  8.452e-03   0.881    0.385
df$X2       -1.170e-02  9.356e-03  -1.251    0.219
df$X3       -1.926e-01  3.425e-01  -0.562    0.578
df$X4       -3.625e-04  1.018e-03  -0.356    0.724
df$X5        6.356e-05  9.839e-04   0.065    0.949
df$X6        3.378e-01  3.567e-01   0.947    0.350
df$X7        4.496e-03  6.420e-03   0.700    0.489
df$X8        1.421e-01  9.491e-02   1.497    0.144

Residual standard error: 67.17 on 34 degrees of freedom
Multiple R-squared:  0.2243,	Adjusted R-squared:  0.04173
F-statistic: 1.229 on 8 and 34 DF,  p-value: 0.3127

> vif(lm1)
    df$X1     df$X2     df$X3     df$X4     df$X5     df$X6     df$X7     df$X8
 5.296659  1.557640  2.931092 14.233947  9.203800  1.419178  3.848155  1.673738
```

TABLE 2. The classification of VIF values among predictor variables in the first regression.

| Code | Description | VIF Value | Category |
|---|---|---|---|
| $X_1$ | The number of babies who had been immunized with HBO | 5.26659 | Bad |
| $X_2$ | The number of health workers | 1.557640 | Good |
| $X_3$ | The number of health facilities | 2.931092 | Moderate |
| $X_4$ | The number of residents | 14.233947 | Bad |
| $X_5$ | The number of residents who had proper sanitation | 9.203800 | Bad |
| $X_6$ | The number of drinking water facilities who have met health standards | 1.419178 | Excellent |
| $X_7$ | The number of diabetics | 3.848155 | Moderate |
| $X_8$ | The number of HIV sufferers | 1.673738 | Good |

The figure 9 shows that the predictor variables (Xi) have no significant effect on the target variable (Y). Moreover, the output indicates a high correlation between $X_1$, $X_3$, $X_4$, $X_5$, and $X_7$, indicating strong relationships among predictor variables. Additionally, the VIF values for $X_1$, $X_3$, $X_4$, $X_5$, and $X_7$ are close to 4, indicating multicollinearity. This issue will be further investigated by creating hypotheses and conducting modeling experiments to determine the best variable and remove one or more correlated independent variables to create the desired model. In conclusion, this model indicates the existence of significant multicollinearity among its variables.

### 4.2 Second Regression Model

The second regression model generates output in the form of the numbers displayed in Figure 10, then we can compile a model using the fit model equation as follows.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4$$

$$\hat{y} = -7.952 + 5.114 * 10^{-3} X_1 - 1.107 * 10^{-2} X_2 - 3.708 * 10^{-1} X_3 + 1.276 * 10^{-1} X_4 \quad (2)$$

Notation

$\hat{y}$ = Target variable prediction for the number of chronic hepatitis patients.

$\hat{\beta}_0$ = Intersection points on the y-axis.

$\hat{\beta}_1$ = Predictor variable for the number of infants who have been immunized with HBO.

$\hat{\beta}_2$ = Predictor variable for the number of health workers.

$\hat{\beta}_3$ = Predictor variable for the number of drinking water facilities who have met health standards.

$\hat{\beta}_4$ = Predictor variable for the number of HIV sufferers

FIGURE 10. The second regression output.

```
Call:
lm(formula = df$Y ~ df$X1 + df$X2 + df$X6 + df$X8)

Residuals:
    Min      1Q  Median      3Q     Max
-110.72  -40.82  -11.87   21.03  170.56

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.951763  23.985494  -0.332    0.742
df$X1        0.005114   0.004228   1.209    0.234
df$X2       -0.011069   0.008579  -1.290    0.205
df$X6        0.370807   0.335392   1.106    0.276
df$X8        0.127617   0.085026   1.501    0.142

Residual standard error: 64.34 on 38 degrees of freedom
Multiple R-squared:  0.2045,    Adjusted R-squared:  0.1207
F-statistic: 2.441 on 4 and 38 DF,  p-value: 0.06331

> vif(lm2)
   df$X1    df$X2    df$X6    df$X8
1.444824 1.427358 1.367316 1.463953
```

We get some information from the output presented in Figure 10. Firstly, the outcome of F test indicates that $P_{value} > a$ with 0.6331 > 0.05, we deduce that the second model is unfit for predicting the number of chronic hepatitis patients. Furthermore, the outcome of T test indicates that there are no factors that influence the model. Figure 10 also reports another information that we can interpret, adjusted-$R^2 = 0.1207$ which indicates that 12.07% of the sample variation y can be described through the model owned. VIF values of predictor variables are generated as follows.

TABLE 3. The classification of VIF values among predictor variables in the second regression.

| Code | Description | VIF Value | Category |
|---|---|---|---|
| $X_1$ | The number of babies who had been immunized with HBO | 1.444824 | Excellent |
| $X_2$ | The number of health workers | 1.427358 | Excellent |
| $X_6$ | The number of drinking water facilities who have met health standards | 1.367316 | Excellent |
| $X_8$ | The number of HIV sufferers | 1.463953 | Excellent |

We conclude that there is no significant multicollinearity for several variables in this model.

### 4.3 Third Regression Model

The third regression model releases outcome in the form of the numbers presented in the Figure 11, then a model using the fit model equation we can create below.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$$

$$\hat{y} = 3.311 - 1.319 * 10^{-2} X_1 - 5.784 * 10^{-1} X_2 - 1.535 * 10^{-1} X_3 \tag{3}$$

Notation

$\hat{\beta}_1$ = Predictor variable for the number of health workers.

$\hat{\beta}_2$ = Predictor variable for the number of drinking water facilities who have met health standards.

$\hat{\beta}_3$ = Predictor variable for the number of HIV sufferers.

FIGURE 11. The third regression output.

```
Call:
lm(formula = df$Y ~ df$X2 + df$X6 + df$X8)

Residuals:
   Min     1Q Median     3Q    Max
-92.65 -28.62 -15.95  17.19 195.81

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.311402  22.234589   0.149   0.8824
df$X2       -0.013197   0.008447  -1.562   0.1263
df$X6        0.578382   0.289857   1.995   0.0530 .
df$X8        0.153546   0.082766   1.855   0.0711 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 64.72 on 39 degrees of freedom
Multiple R-squared:  0.1738,	Adjusted R-squared:  0.1103
F-statistic: 2.735 on 3 and 39 DF,  p-value: 0.05654

> vif(lm3)
   df$X2    df$X6    df$X8
1.367340 1.009275 1.370884
```

TABLE 4. The classification of VIF values among predictor variables in the third regression.

| Code | Description | VIF Value | Category |
|---|---|---|---|
| $X_2$ | The number of health workers | 1.367340 | Excellent |
| $X_6$ | The number of drinking water facilities who have met health standards | 1.009275 | Excellent |
| $X_8$ | The number of HIV sufferers | 1.370884 | Excellent |

We obtain information from the outcome provided in Figure 11. Firstly, the output of F test shows that $P_{value} > a$ with 0.05654 > 0.05, then we can resume that this third model is unfit for predicting the number of acute hepatitis sufferers. In addition, the outcome of T test reveals that there are no factors that impact the model. Figure 11 also reveals another information that we can construe, adjusted-$R^2 = 0.1103$ which presents that 11.03% of the sample variation y can be described through the model owned. Furthermore, we attain VIF values among predictor variables in Table 4. All evidence demonstrates that there is no significant multicollinearity for several variables in this model.

### 4.4 Fourth Regression Model

The fourth regression model generates output in the form of the numbers displayed in Figure 12, then we can compose a model using the fit model equation as follows.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4$$

$$\hat{y} = -5.991 - 1.299 * 10^{-2} X_1 + 8.384 * 10^{-2} X_2 + 5.468 * 10^{-1} X_3 + 1.457 * 10^{-1} X_4 \quad (4)$$

Notation

$\hat{\beta}_1$ = Predictor variable for the number of health workers.

$\hat{\beta}_2$ = Predictor variable for the number of health facilities.

$\hat{\beta}_3$ = Predictor variable for the number of drinking water facilities who have met health standards.

$\hat{\beta}_4$ = Predictor variable for the number of HIV sufferers.

FIGURE 12. The fourth regression outcome.

```
Call:
lm(formula = df$Y ~ df$X2 + df$X3 + df$X6 + df$X8)

Residuals:
   Min     1Q Median     3Q    Max 
-96.92 -28.50 -14.70  16.65 193.26 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.990500  32.235171  -0.186   0.8536
df$X2       -0.012993   0.008554  -1.519   0.1370
df$X3        0.083844   0.208268   0.403   0.6895
df$X6        0.546760   0.303367   1.802   0.0794 .
df$X8        0.145728   0.085893   1.697   0.0979 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 65.43 on 38 degrees of freedom
Multiple R-squared:  0.1773,	Adjusted R-squared:  0.09074 
F-statistic: 2.048 on 4 and 38 DF,  p-value: 0.1071

> vif(lm4)
   df$X2    df$X3    df$X6    df$X8 
1.372111 1.142122 1.081796 1.444721 
```

TABLE 5. The category of VIF values among predictor variables in the fourth regression model.

| Code | Description | VIF Value | Category |
|---|---|---|---|
| $X_2$ | The number of health workers | 1.372111 | Excellent |
| $X_3$ | The number of health facilities | 1.142122 | Excellent |
| $X_6$ | The number of drinking water facilities who have met health standards | 1.081796 | Excellent |
| $X_8$ | The number of HIV sufferers | 1.444721 | Excellent |

We get some information from the output presented in Figure 12. Firstly, the result of F test demonstrates that $P_{value} > a$ with 0.1071 > 0.05, then we can conclude that this fourth model is unfit for predicting the number of acute hepatitis patients. Moreover, the outcome of T test reveals that there are no factors that impact the model. Figure 12 also reports another information that we can interpret, adjusted-$R^2 = 0.09074$ which indicates that 9.074% of the sample variation y can be illustrated through the model owned. In addition, we obtain some VIF values as follows. In conclusion, this model indicates the absence of multicollinearity among its variables.

### 4.5 Fifth Regression Model

The fifth regression model produces output in the form of the numbers presented in Figure 13, then we can formulate a model using the fit model equation as follows.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4$$

$$\hat{y} = -2.747 - 1.205 * 10^{-2} X_1 - 1.458 * 10^{-1} X_2 + 7.237 * 10^{-3} X_3 + 1.445 * 10^{-1} X_4 \quad (5)$$

Notation

$\hat{\beta}_1$ = Predictor variable for the number of health workers.

$\hat{\beta}_2$ = Predictor variable for the number of health facilities.

$\hat{\beta}_3$ = Predictor variable for the number of diabetics.

$\hat{\beta}_4$ = Predictor variable for the number of HIV sufferers.

FIGURE 13. Fifth regression model.

```
Call:
lm(formula = df$Y ~ df$X2 + df$X3 + df$X7 + df$X8)

Residuals:
   Min     1Q Median     3Q    Max
-82.86 -38.40 -17.22  12.77 208.26

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.746693  32.501781  -0.085    0.933
df$X2       -0.012053   0.008647  -1.394    0.171
df$X3       -0.145807   0.297026  -0.491    0.626
df$X7        0.007237   0.004789   1.511    0.139
df$X8        0.144536   0.086918   1.663    0.105

Residual standard error: 66.21 on 38 degrees of freedom
Multiple R-squared:  0.1576,    Adjusted R-squared:  0.06896
F-statistic: 1.778 on 4 and 38 DF,  p-value: 0.1535

> vif(lm5)
   df$X2    df$X3    df$X7    df$X8
1.369486 2.268684 2.204300 1.444790
```

We obtain some information from the result presented in Figure 13. Firstly, the outcome of F test indicates that $P_{value} > a$ with 0.1535 > 0.05, then we can conclude that this fifth model is unfit for predicting the number of sufferers of acute hepatitis. Furthermore, the outcome of T test reveals that there are no factors that influence the model. Figure 13 also reports another information that we can interpret, adjusted-$R^2 = 0.06896$ which indicates that 6.896% of the sample variation y can be described through the model owned. In addition, we have some VIF values in Table 6.

TABLE 6. The category of VIF values among predictor variables in the fifth regression model.

| Code | Description | VIF Value | Category |
|---|---|---|---|
| $X_2$ | The number of health workers | 1.369486 | Excellent |
| $X_3$ | The number of health facilities | 2.268684 | Moderate |
| $X_7$ | The number of diabetics | 2.204300 | Moderate |
| $X_8$ | The number of HIV sufferers | 1.444790 | Excellent |

We deduce that there is no significant multicollinearity among several variables in this model.

### 4.6 Sixth Regression Model

The sixth regression model generates outcome in the form of the numbers provided in Figure 14, then a model utilizing the fit model equation we can compile as follows.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4$$

$$\hat{y} = -15.36 - 1.262 * 10^{-2} X_1 + 4.627 * 10^{-1} X_2 + 3.797 * 10^{-3} X_3 + 1.379 * 10^{-1} X_4 \quad (6)$$

Notation

$\hat{\beta}_1$ = Predictor variable for the number of health workers.

$\hat{\beta}_2$ = Predictor variable for the number of drinking water facilities who have met health standards.

$\hat{\beta}_3$ = Predictor variable for the number of diabetics.

$\hat{\beta}_4$ = Predictor variable for the number of HIV sufferers.

FIGURE 14. The sixth regression result.

```
Call:
lm(formula = df$Y ~ df$X2 + df$X6 + df$X7 + df$X8)

Residuals:
   Min     1Q Median     3Q    Max
-84.87 -36.94 -15.12  20.80 196.64

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.362826  27.757812  -0.553    0.583
df$X2        -0.012618   0.008436  -1.496    0.143
df$X6         0.462665   0.306932   1.507    0.140
df$X7         0.003797   0.003398   1.117    0.271
df$X8         0.137948   0.083675   1.649    0.107

Residual standard error: 64.52 on 38 degrees of freedom
Multiple R-squared:  0.2001,    Adjusted R-squared:  0.1159
F-statistic: 2.377 on 4 and 38 DF,  p-value: 0.06902

> vif(lm6)
   df$X2    df$X6    df$X7    df$X8
1.372511 1.138905 1.168291 1.410119
```

We have some information from the output presented in Figure 14. Firstly, the outcome of F test indicates that $P_{value} > a$ with 0.06902 > 0.05, then we can conclude that this sixth model is unfit for predicting the number of acute hepatitis patients. In addition, the result of T test reveals that there are no factors that impact the model. Figure 14 also reports another information that we can interpret, adjusted-$R^2 = 0.1159$ which indicates that 11.59% of the sample variation y can be described through the model owned. The following VIF values among predictor variables given.

TABLE 7. The classification of VIF values among predictor variables in the sixth regression.

| Code | Description | VIF Value | Category |
|---|---|---|---|
| $X_2$ | The number of health workers | 1.372511 | Excellent |
| $X_6$ | The number of drinking water facilities who have met health standards | 1.138905 | Excellent |
| $X_7$ | The number of diabetics | 1.168291 | Excellent |
| $X_8$ | The number of HIV sufferers | 1.410119 | Excellent |

All evidence demonstrates that there is no significant multicollinearity for variables in this model.

### 4.7 Seventh Regression Model

The seventh regression model releases output in the form of the numbers provided in Figure 15, then we can formulate a model using the fit model equation as follows.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2^3 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4$$

$$\hat{y} = 2.189 - 1.311 * 10^{-2} X_1 + 3.950 * 10^{-7} X_2^3 + 5.695 * 10^{-1} X_3 + 1.523 * 10^{-1} X_4 \quad (7)$$

Notation

$\hat{\beta}_1 =$ Predictor variable for the number of health workers.

$\hat{\beta}_2 =$ Predictor variable for the number of health facilities.

$\hat{\beta}_3 =$ Predictor variable for the number of drinking water facilities who have met health standards.

$\hat{\beta}_4 =$ Predictor variable for the number of HIV sufferers.

FIGURE 15. The seventh regression outcome.

```
Call:
lm(formula = df$Y ~ df$X2 + I(df$X3^3) + df$X6 + df$X8)

Residuals:
   Min     1Q Median     3Q    Max 
-94.68 -27.94 -14.79  17.40 194.65 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)  2.189e+00  2.401e+01   0.091   0.9278  
df$X2       -1.311e-02  8.581e-03  -1.527   0.1349  
I(df$X3^3)   3.950e-07  2.926e-06   0.135   0.8933  
df$X6        5.695e-01  3.008e-01   1.894   0.0659 .
df$X8        1.523e-01  8.433e-02   1.806   0.0788 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 65.55 on 38 degrees of freedom
Multiple R-squared:  0.1742,	Adjusted R-squared:  0.0873 
F-statistic: 2.004 on 4 and 38 DF,  p-value: 0.1135

> vif(lm7)
     df$X2 I(df$X3^3)      df$X6      df$X8 
  1.375575   1.064856   1.059388   1.387280 
```

We obtain some data from the outcome displayed in Figure 15. Firstly, the output of F test demonstrates that $P_{value} > a$ with 0.1135 > 0.05, then we deduce that this seventh model is unfit for predicting the number of chronic hepatitis. Furthermore, the result of T test reveals that there are no factors that impact the model. Figure 15 also reveals another information that we can see, adjusted-$R^2 = 0.0873$ which means that 8.73% of the sample variation y can be described through the model owned. In addition, we obtain some VIF values among predictor variables below.

TABLE 8. The category of VIF values among predictor variables in the seventh regression.

| Code | Description | VIF Value | Category |
|---|---|---|---|
| $X_2$ | The number of health workers | 1.375575 | Excellent |
| $X_3$ | The number of health facilities | 1.064856 | Excellent |
| $X_6$ | The number of drinking water facilities who have met health standards | 1.059388 | Excellent |
| $X_8$ | The number of HIV sufferers | 1.387280 | Excellent |

We resume that there is no significant multicollinearity among some variables in this model.

### 4.8 Eighth Regression Model

The eighth regression model produces outcome in the form of the numbers presented in Figure 16, then we can collate a model using the fit model equation as follows.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1^4 + \hat{\beta}_2 X_2^4 + \hat{\beta}_3 X_3^4$$

$$\hat{y} = 19.64 + 1.102 * 10^{-14} X_1^4 - 2.387 * 10^{-19} X_2^4 + 3.898 * 10^{-15} X_3^4 \tag{8}$$

Notation

$\hat{\beta}_1$ = Predictor variable for the number of babies who had been immunized with HBO.

$\hat{\beta}_2$ = Predictor variable for the number of residents.

$\hat{\beta}_3$ = Predictor variable for the number of diabetics.

FIGURE 16. The eighth regression output.

```
Call:
lm(formula = df$Y ~ I(df$X1^4) + I(df$X4^4) + I(df$X7^4))

Residuals:
    Min      1Q  Median      3Q     Max
-112.53  -26.66  -18.95   18.55  137.23

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.964e+01  1.128e+01   1.741  0.08950 .
I(df$X1^4)   1.102e-14  3.634e-15   3.033  0.00430 **
I(df$X4^4)  -2.387e-19  1.088e-19  -2.194  0.03427 *
I(df$X7^4)   3.898e-15  1.386e-15   2.813  0.00764 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59.38 on 39 degrees of freedom
Multiple R-squared:  0.3047,	Adjusted R-squared:  0.2512
F-statistic: 5.696 on 3 and 39 DF,  p-value: 0.002472

> vif(lm8)
I(df$X1^4) I(df$X4^4) I(df$X7^4)
  1.887306   2.328508   1.359060
```

TABLE 9. The classification of VIF values among predictor variables in the eighth regression.

| Code | Description | VIF Value | Category |
|---|---|---|---|
| $X_1$ | The number of babies who had been immunized with HBO | 1.887306 | Good |
| $X_4$ | The number of residents | 2.328508 | Moderate |
| $X_7$ | The number of diabetics | 1.359060 | Excellent |

We get some information from the output presented in Figure 16. Firstly, the outcome of F test indicates that $P_{value} < \alpha$ with 0.002472 < 0.05, then we conclude that this eighth model is fit for predicting the number of sufferers of acute hepatitis and testing can continue. Furthermore, the outcome of T test reveals that there are factors that significantly impact the model. Figure 16 also reports another data that we can analyze, adjusted-$R^2 = 0.2512$ which indicates that 25.12% of the sample variation y can be described through the model owned. In conclusion, this model indicates the presence of significant multicollinearity among its variables.

### 4.9 Ninth Regression Model

The ninth regression model releases outcome in the form of the numbers provided in Figure 17, then we can formulate a model utilizing the fit model equation as follows.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1^{10} + \hat{\beta}_2 X_2^{10} + \hat{\beta}_3 X_3^{10}$$

$$\hat{y} = 30.03 + 7.994 * 10^{-39} X_1^{10} - 3.212 * 10^{-50} X_2^{10} + 9.306 * 10^{-40} X_3^{10} \qquad (9)$$

Notation

$\hat{\beta}_1$ = Predictor variable for the number of babies who had been immunized with HBO.

$\hat{\beta}_2$ = Predictor variable for the number of residents.

$\hat{\beta}_3$ = Predictor variable for the number of diabetics.

FIGURE 17. The ninth regression outcome.

```
Call:
lm(formula = df$Y ~ I(df$X1^10) + I(df$X4^10) + I(df$X7^10))

Residuals:
   Min     1Q Median     3Q    Max
-86.84 -29.04 -24.03  13.86 129.94

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.003e+01  8.259e+00   3.637 0.000797 ***
I(df$X1^10)   7.994e-39  1.649e-39   4.847 2.03e-05 ***
I(df$X4^10)  -3.212e-50  8.006e-51  -4.012 0.000264 ***
I(df$X7^10)   9.306e-40  1.864e-40   4.993 1.28e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50.12 on 39 degrees of freedom
Multiple R-squared:  0.5045,	Adjusted R-squared:  0.4664
F-statistic: 13.24 on 3 and 39 DF,  p-value: 4.169e-06

> vif(lm9)
I(df$X1^10) I(df$X4^10) I(df$X7^10)
   1.925004    2.001282    1.093223
```

We can interpret some data from the output given in Figure 17. Firstly, the outcome of F test indicates that $P_{value} < a$ with $4.169 * 10^{-6} < 0.05$, we conclude that this ninth model is fit for predicting the number of chronic hepatitis patients and testing can be continued. Furthermore, the outcome of T test reveals that there are factors that significantly impact the model. Figure 17 also reports another information that we can see, adjusted-$R^2 = 0.4664$ which indicates that 46.64% of the sample variation y can be described through the model owned.

TABLE 10. The category of VIF values among predictor variables in the ninth regression model.

| Code | Description | VIF Value | Category |
|---|---|---|---|
| $X_1$ | The number of babies who had been immunized with HBO | 1.925004 | Good |
| $X_4$ | The number of residents | 2.001282 | Moderate |
| $X_7$ | The number of diabetics | 1.093223 | Excellent |

All evidence indicates that there is significant multicollinearity among variables in this model.

**4.10 Tenth Regression Model**

The tenth regression model provides result in the form of the numbers displayed in Figure 18, then a model using the fit model equation we can compose as follows.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1^5 + \hat{\beta}_2 X_2^5 + \hat{\beta}_3 X_3^5$$

$$\hat{y} = 22.43 + 1.084 * 10^{-10} X_1^5 - 1.815 * 10^{-24} X_2^5 + 3.394 * 10^{-19} X_3^5 \quad (10)$$

Notation

$\hat{\beta}_1 =$ Predictor variable for the number of babies who had been immunized with HBO.

$\hat{\beta}_2 =$ Predictor variable for the number of residents.

$\hat{\beta}_3 =$ Predictor variable for the number of diabetics.

We can analyze information from the output provided in Figure 18. Firstly, the outcome of F test indicates that $P_{value} < a$ with $0.0006104 < 0.05$, we resume that this tenth model is fit for predicting the number of chronic hepatitis patients and testing can continue. Furthermore, the outcome of T test reveals that there are factors that significantly impact the model. Figure 18 provides another information that we can process, adjusted-$R^2 = 0.3054$ which means that 30.54% of the sample variation y can be described through the model owned.

FIGURE 18. The tenth regression result.

```
Call:
lm(formula = df$Y ~ I(df$X1^5) + I(df$X4^5) + I(df$X7^5))

Residuals:
    Min      1Q  Median      3Q     Max 
-116.61  -25.81  -20.51   19.14  136.04 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)  2.243e+01  1.032e+01   2.173  0.03593 * 
I(df$X1^5)   1.084e-18  3.129e-19   3.463  0.00131 **
I(df$X4^5)  -1.815e-24  6.934e-25  -2.618  0.01253 * 
I(df$X7^5)   3.394e-19  9.872e-20   3.438  0.00141 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57.19 on 39 degrees of freedom
Multiple R-squared:  0.355,	Adjusted R-squared:  0.3054 
F-statistic: 7.154 on 3 and 39 DF,  p-value: 0.0006104

> vif(lm10)
I(df$X1^5) I(df$X4^5) I(df$X7^5) 
  1.806495   2.151061   1.260807 
```

TABLE 11. The classification of VIF values among predictor variables in the tenth regression.

| Code | Description | VIF Value | Category |
|---|---|---|---|
| $X_1$ | The number of babies who had been immunized with HBO | 1.806495 | Good |
| $X_4$ | The number of residents | 2.151061 | Moderate |
| $X_7$ | The number of diabetics | 1.260807 | Excellent |

We conclude that there is significant multicollinearity among several variables in this model.

**4.11 Eleventh Regression Model**

The eleventh regression model generates output in the form of the numbers presented in Figure 19, then we can arrange the following model applying the fit model equation as follows.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1^6 + \hat{\beta}_2 X_2^6 + \hat{\beta}_3 X_3^6$$

$$\hat{y} = 24.63 + 1.049 * 10^{-22} X_1^6 - 1.341 * 10^{-29} X_2^6 + 2.802 * 10^{-23} X_3^6 \quad (11)$$

Notation

$\hat{\beta}_1$ = Predictor variable for the number of babies who had been immunized with HBO.

$\hat{\beta}_2$ = Predictor variable for the number of residents.

$\hat{\beta}_3$ = Predictor variable for the number of diabetics.

FIGURE 19. The eleventh regression output.

```
Call:
lm(formula = df$Y ~ I(df$X1^6) + I(df$X4^6) + I(df$X7^6))

Residuals:
    Min      1Q  Median      3Q     Max
-115.91  -25.70  -22.63   19.00  134.66

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.463e+01  9.645e+00   2.554 0.014680 *
I(df$X1^6)   1.049e-22  2.720e-23   3.856 0.000420 ***
I(df$X4^6)  -1.341e-29  4.464e-30  -3.004 0.004635 **
I(df$X7^6)   2.802e-23  7.077e-24   3.959 0.000309 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55.16 on 39 degrees of freedom
Multiple R-squared:  0.3999,    Adjusted R-squared:  0.3537
F-statistic: 8.662 on 3 and 39 DF,  p-value: 0.0001576

> vif(lm11)
I(df$X1^6) I(df$X4^6) I(df$X7^6)
  1.793815   2.057532   1.204993
```

We get some information from the output presented in Figure 19. Firstly, the outcome of F test indicates that $P_{value} < \alpha$ with 0.0001576 < 0.05, we can conclude that the eleventh model is fit for predicting the number of sufferers of acute hepatitis and testing can continue. Furthermore, the outcome of T test reveals that there are factors that significantly impact the model. Figure 19 also reports another information that we can see, adjusted-$R^2 = 0.3537$ which indicates that 35.37% of the sample variation y can be described through the model owned.

TABLE 12. The classification of VIF values among predictor variables in the eleventh model.

| Code | Description | VIF Value | Category |
|---|---|---|---|
| $X_1$ | The number of babies who had been immunized with HBO | 1.793815 | Good |
| $X_4$ | The number of residents | 2.057532 | Moderate |
| $X_7$ | The number of diabetics | 1.204993 | Excellent |

In conclusion, this model shows the presence of significant multicollinearity among its variables.

### 4.12 Twelfth Regression Model

The twelfth regression model releases outcome in the form of the numbers displayed in Figure 20, then we can compile a model using the fit model equation as follows.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1^7 + \hat{\beta}_2 X_2^7 + \hat{\beta}_3 X_3^7$$

$$\hat{y} = 26.42 + 9.986 * 10^{-27} X_1^7 - 9.659 * 10^{-35} X_2^7 + 2.218 * 10^{-27} X_3^7 \quad (12)$$

Notation

$\hat{\beta}_1$ = Predictor variable for the number of babies who had been immunized with HBO.

$\hat{\beta}_2$ = Predictor variable for the number of residents.

$\hat{\beta}_3$ = Predictor variable for the number of diabetics.

FIGURE 20. The twelfth regression outcome.

```
Call:
lm(formula = df$Y ~ I(df$X1^7) + I(df$X4^7) + I(df$X7^7))

Residuals:
    Min      1Q  Median      3Q     Max 
-111.39  -26.84  -24.22   18.45  133.27 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.642e+01  9.136e+00   2.892 0.006237 ** 
I(df$X1^7)   9.986e-27  2.380e-27   4.196 0.000152 ***
I(df$X4^7)  -9.659e-35  2.891e-35  -3.341 0.001847 ** 
I(df$X7^7)   2.218e-27  5.073e-28   4.372 8.86e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 53.41 on 39 degrees of freedom
Multiple R-squared:  0.4374,	Adjusted R-squared:  0.3941 
F-statistic: 10.11 on 3 and 39 DF,  p-value: 4.659e-05

> vif(lm12)
I(df$X1^7) I(df$X4^7) I(df$X7^7) 
  1.812351   2.010525   1.167802 
```

We have some information from the output presented in Figure 20. Firstly, the outcome of F test indicates that $P_{value} < \alpha$ with $4.659 * 10^{-5} < 0.05$, we can conclude that the twelfth model is fit for predicting the number of sufferers of acute hepatitis and testing can continue. Furthermore, the outcome of T test reveals that there are factors that significantly impact the model. Figure 20 also reveals another information that we can analyze, adjusted-$R^2 = 0.3941$ which means that 39.41% of the sample variation y can be described through the model owned.

TABLE 13. The classification of VIF values among predictor variables in the twelfth regression.

| Code | Description | VIF Value | Category |
|---|---|---|---|
| $X_1$ | The number of babies who had been immunized with HBO | 1.812351 | Good |
| $X_4$ | The number of residents | 2.010525 | Moderate |
| $X_7$ | The number of diabetics | 1.167802 | Excellent |

We deduce that there is significant multicollinearity among several variables in this model.

### 4.13 Thirteenth Regression Model

The thirteenth regression model produces outcome in the form of the numbers given in Figure 21, then we can compose a model using the fit model equation as follows.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1^8 + \hat{\beta}_2 X_2^8 + \hat{\beta}_3 X_3^8$$

$$\hat{y} = 27.89 + 9.371 * 10^{-31} X_1^8 - 6.805 * 10^{-40} X_2^8 + 1.699 * 10^{-31} X_3^8 \quad (13)$$

Notation

$\hat{\beta}_1 =$ Predictor variable for the number of babies who had been immunized with HBO.

$\hat{\beta}_2 =$ Predictor variable for the number of residents.

$\hat{\beta}_3 =$ Predictor variable for the number of diabetics.

FIGURE 21. The thirteenth regression outcome.

```
Call:
lm(formula = df$Y ~ I(df$X1^8) + I(df$X4^8) + I(df$X7^8))

Residuals:
    Min      1Q  Median      3Q     Max 
-104.28  -27.89  -24.18   17.70  131.98 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.789e+01  8.753e+00   3.186 0.002838 ** 
I(df$X1^8)   9.371e-31  2.094e-31   4.475 6.47e-05 ***
I(df$X4^8)  -6.805e-40  1.879e-40  -3.621 0.000834 ***
I(df$X7^8)   1.699e-31  3.634e-32   4.675 3.47e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 51.99 on 39 degrees of freedom
Multiple R-squared:  0.4669,	Adjusted R-squared:  0.4259 
F-statistic: 11.38 on 3 and 39 DF,  p-value: 1.679e-05

> vif(lm13)
I(df$X1^8) I(df$X4^8) I(df$X7^8) 
  1.844887   1.991319   1.138924 
```

We can analyze some information from the output given in Figure 21. Firstly, the output of F test indicates that $P_{value} < \alpha$ with $1.679 * 10^{-5} < 0.05$, we can conclude that thirteenth model is fit for predicting the number of sufferers of acute hepatitis and testing can continue. Furthermore, the outcome of T test indicates that there are factors that significantly impact the model. Figure 21 also gives another information that we can interpret, adjusted-$R^2 = 0.4259$ which means that 42.59% of the sample variation y can be described through the model owned.

TABLE 14. The category of VIF values among predictor variables in the thirteenth regression.

| Code | Description | VIF Value | Category |
|---|---|---|---|
| $X_1$ | The number of babies who had been immunized with HBO | 1.844887 | Good |
| $X_4$ | The number of residents | 1.991319 | Good |
| $X_7$ | The number of diabetics | 1.138924 | Excellent |

All evidence indicates that there is significant multicollinearity among variables in this model.

### 4.14 Fourteenth Regression Model

The fourteenth regression model generates output in the form of the numbers displayed in Figure 22, then we can draw up a model utilizing the fit model equation as follows.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1^9 + \hat{\beta}_2 X_2^9 + \hat{\beta}_3 X_3^9$$

$$\hat{y} = 29.08 + 8.692 * 10^{-35} X_1^9 - 4.708 * 10^{-45} X_2^9 + 1.269 * 10^{-35} X_3^9 \quad (14)$$

Notation

$\hat{\beta}_1$ = Predictor variable for the number of babies who had been immunized with HBO.

$\hat{\beta}_2$ = Predictor variable for the number of residents.

$\hat{\beta}_3$ = Predictor variable for the number of diabetics.

FIGURE 22. The fourteenth regression output.

```
Call:
lm(formula = df$Y ~ I(df$X1^9) + I(df$X4^9) + I(df$X7^9))

Residuals:
   Min     1Q Median     3Q    Max 
-95.77 -29.08 -24.45  16.91 130.86 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.908e+01  8.468e+00   3.434 0.001422 ** 
I(df$X1^9)   8.692e-35  1.854e-35   4.689 3.32e-05 ***
I(df$X4^9)  -4.708e-45  1.225e-45  -3.843 0.000436 ***
I(df$X7^9)   1.269e-35  2.602e-36   4.877 1.84e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50.91 on 39 degrees of freedom
Multiple R-squared:  0.4888,	Adjusted R-squared:  0.4495 
F-statistic: 12.43 on 3 and 39 DF,  p-value: 7.546e-06

> vif(lm14)
I(df$X1^9) I(df$X4^9) I(df$X7^9) 
  1.883576   1.990090   1.114454 
```

We have some information from the output presented in Figure 22. Firstly, the outcome of F test indicates that $P_{value} < \alpha$ with $7.546 * 10^{-6} < 0.05$, we can conclude that fourteenth model is fit for predicting the number of sufferers of acute hepatitis and testing can continue. Furthermore, the outcome of T test reveals that there are factors that significantly impact the model. Figure 22 also reports another information that we can analyze, adjusted-$R^2 = 0.4495$ which means that 44.95% of the sample variation y can be described through the model owned.

TABLE 15. The classification of VIF values among predictor variables in the fourteenth model.

| Code | Description | VIF Value | Category |
|---|---|---|---|
| $X_1$ | The number of babies who had been immunized with HBO | 1.883576 | Good |
| $X_4$ | The number of residents | 1.990090 | Good |
| $X_7$ | The number of diabetics | 1.114454 | Excellent |

We conclude that there is significant multicollinearity for some variables in this model.

**4.15 Fifteenth Regression Model**

The fifteenth regression model releases output in the form of the numbers provided in Figure 23, then we can compile a model using the fit model equation as follows.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1^1 + \hat{\beta}_2 X_2^4 + \hat{\beta}_3 X_3^7$$

$$\hat{y} = 18.46 + 8.981 * 10^{-3} X_1^1 - 1.389 * 10^{-19} X_2^4 + 1.453 * 10^{-27} X_3^7 \quad (15)$$

Notation

$\hat{\beta}_1 =$ The number of babies who had been immunized with HBO.

$\hat{\beta}_2 =$ The number of residents.

$\hat{\beta}_3 =$ Predictor variable for the number of diabetics.

FIGURE 23. The fifteenth regression outcome.

```
Call:
lm(formula = df$Y ~ I(df$X1^1) + I(df$X4^4) + I(df$X7^7))

Residuals:
    Min      1Q  Median      3Q     Max
-101.38  -29.88  -13.40   11.95  202.24

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.846e+00  1.854e+01   0.100   0.9212
I(df$X1^1)   8.981e-03  4.634e-03   1.938   0.0599 .
I(df$X4^4)  -1.389e-19  1.021e-19  -1.360   0.1817
I(df$X7^7)   1.453e-27  6.246e-28   2.326   0.0253 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 61.68 on 39 degrees of freedom
Multiple R-squared:  0.2497,	Adjusted R-squared:  0.192
F-statistic: 4.327 on 3 and 39 DF,  p-value: 0.01

> vif(lm15)
I(df$X1^1) I(df$X4^4) I(df$X7^7)
  1.888713   1.901249   1.327234
```

We can analyze information from the output displayed in Figure 23. Firstly, the output of F test indicates that $P_{value} < \alpha$ with 0.01 < 0.05, then we can conclude that fifteenth model is fit for predicting the number of acute hepatitis patients. Furthermore, the outcome of T test indicates that there are factors that significantly impact the model. Figure 23 also reports another information that we can see, adjusted-$R^2 = 0.192$ which indicates that 19.2% of the sample variation y can be described through the model owned.

TABLE 16. The category of VIF values among predictor variables in the fifteenth regression.

| Code | Description | VIF Value | Category |
|---|---|---|---|
| $X_1$ | The number of babies who had been immunized with HBO | 1.888713 | Good |
| $X_4$ | The number of residents | 1.901249 | Good |
| $X_7$ | The number of diabetics | 1.327234 | Excellent |

This model indicates the existence of significant multicollinearity among its variables.

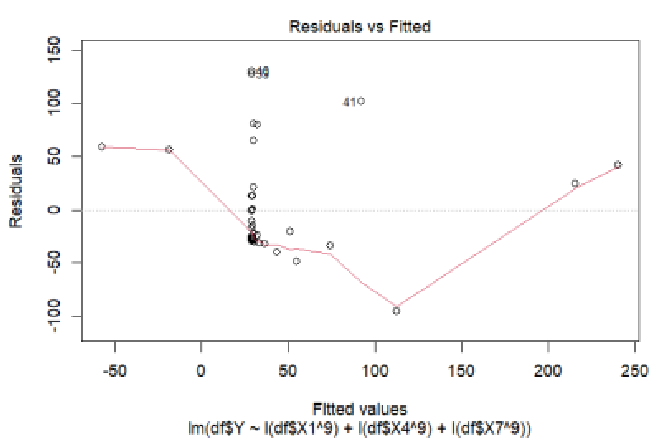
**4.16 Regression Model**

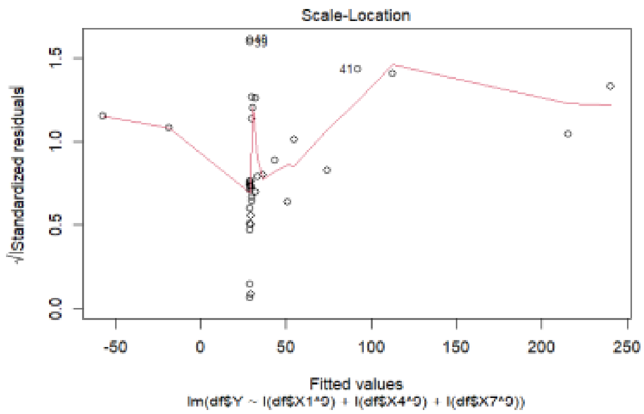
After examining all the models, the 14th regression model is selected as the best model when compared to other models. This is apparent from the following plot, that the data in model 14 is nearby to the estimation outcome of the model.

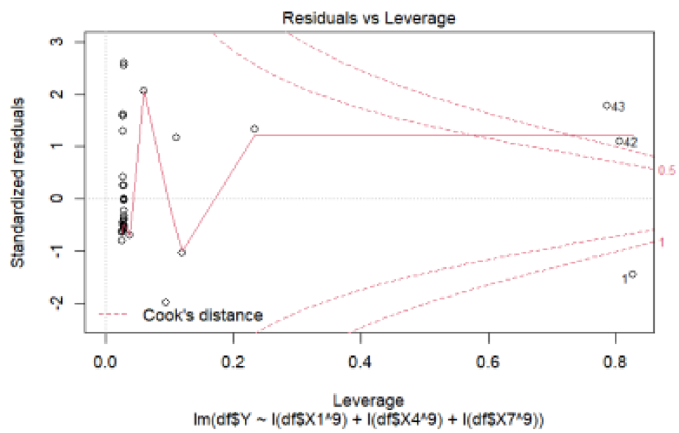
$$y = \beta_0 + \beta_1 X_1^9 + \beta_2 X_2^9 + \beta_3 X_3^9 + \varepsilon$$

$$\hat{y} = 29.08 + 8.692 * 10^{-35} X_1^9 - 4.708 * 10^{-45} X_2^9 + 1.269 * 10^{-35} X_3^9 \tag{16}$$

FIGURE 24. Plots generated in model 14.






## 5. CONCLUSION

The report discusses the number of acute hepatitis patients in several sub-districts of DKI Jakarta and proposes a model to recognize risk factors. The model identifies the connection between one response variable (number of acute hepatitis patients) and eight predictor variables. After compiling 15 models and organizing all these models, each model goes through a global ANOVA test process (f-test). Subsequently, we verify whether all the formulated models are beneficial for predicting the number of acute hepatitis sufferers. This research reports that the $14^{th}$ model is the best model according to the P-value generated from t-test and F-test. In addition, the $14^{th}$ model produces the VIF that is excellent to prevent multicollinearity in this model and the plots that also promote although adjusted-$R^2$ is 0.4495. However, this value is already higher when compared to the adjusted-$R^2$ value in other models. After the $14^{th}$ model stipulated as the best model, we also recognize several risk factors greatly induced the number of acute hepatitis sufferers, including the number of infants who had been immunized with HBO ($X_1$), the total population ($X_4$), and the number of diabetics ($X_7$). The proposed model is then supported by the pairs function to observe various scatter plots.

## ACKNOWLEDGEMENT

Thanks for the teamwork and collaboration. The highest appreciation goes to all lecturers in the department of mathematics. This study was fully granted by Universitas Terbuka.

## CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

## REFERENCES

[1] C.D.K. Wungu, S. Khaerunnisa, I. Humairah, et al. Counselling and screening of hepatitis B virus infection in Dukuh Kupang Community, Dukuh Pakis District, Surabaya, IOP Conf. Ser.: Earth Environ. Sci. 217 (2019), 012048. https://doi.org/10.1088/1755-1315/217/1/012048.

[2] C. M. de Noordhout, B. Devleesschauwer, J.A. Haagsma, et al. Burden of salmonellosis, campylobacteriosis and listeriosis: a time series analysis, Belgium, 2012 to 2020, Eurosurveillance. 22 (2017), 30615. https://doi.org/10.2807/1560-7917.es.2017.22.38.30615.

[3] WHO, Combating hepatitis B and C to reach elimination by 2030, (2016). https://www.who.int/publications/i/item/combating-hepatitis-b-and-c-to-reach-elimination-by-2030.

[4] J.E. Aurelia, Z. Rustam, I. Wirasati, et al. Hepatitis classification using support vector machines and random forest, IAES Int. J. Artif. Intell. 10 (2021), 446-451. https://doi.org/10.11591/ijai.v10.i2.pp446-451.

[5] Y. Yano, Hepatitis B virus infection in Indonesia, World J. Gastroenterol. 21 (2015), 10714. https://doi.org/10.3748/wjg.v21.i38.10714.

[6] P.B. Purwono, Y. Yano, T. Utsumi, et al. Hepatitis B virus infection in Indonesia 15 years after adoption of a universal infant vaccination program: possible impacts of low birth dose coverage and a vaccine-escape mutant, Amer. J. Trop. Med. Hyg. 95 (2016), 674-679. https://doi.org/10.4269/ajtmh.15-0121.

[7] T.J. Schaefer, S. John, Acute hepatitis. StatPearls, StatPearls Publishing, Treasure Island, (2019).

[8] M.M. Mücke, S. Zeuzem, The recent outbreak of acute severe hepatitis in children of unknown origin - what is known so far, J. Hepatol. 77 (2022), 237-242. https://doi.org/10.1016/j.jhep.2022.05.001.

[9] S.J. Hullegie, J.E. Arends, B.J.A. Rijnders, et al. Current knowledge and future perspectives on acute hepatitis C infection, Clin. Microbiol. Infect. 21 (2015), 797.e9-797.e17. https://doi.org/10.1016/j.cmi.2015.03.026.

[10] T. Suzuki, T. Okamoto, F. Kawai, et al. Hemolytic anemia after acute hepatitis B virus infection: a case report and systematic review, Internal Med. 61 (2022), 481-488. https://doi.org/10.2169/internalmedicine.7690-21.

[11] M.T. Pérez-Gracia, A. Tarín-Pelló, B. Suay-García, Severe acute hepatitis of unknown origin in children: what do we know today?, J. Clin. Transl. Hepatol. 10 (2022), 711-717. https://doi.org/10.14218/jcth.2022.00244.

[12] Y.H. Chen, J.G. Lou, Z.H. Yang, et al. Diagnosis, treatment, and prevention of severe acute hepatitis of unknown etiology in children, World J. Pediatr. 18 (2022), 538-544. https://doi.org/10.1007/s12519-022-00581-x.

[13] A.M. Kandeel, M. Talaat, S.A. Afifi, et al. Case control study to identify risk factors for acute hepatitis C virus infection in Egypt, BMC Infect. Dis. 12 (2012), 294. https://doi.org/10.1186/1471-2334-12-294.

[14] H. Bhatt, S. Saklani, K. Upadhayay, Anti-oxidant and anti-diabetic activities of ethanolic extract of primula denticulata flowers, Indonesian J. Pharm. 27 (2016), 74. https://doi.org/10.14499/indonesianjpharm27iss2pp74.

[15] J. Hartl, M. Wehmeyer, S. Pischke, Acute hepatitis E: Two sides of the same coin, Viruses. 8 (2016), 299. https://doi.org/10.3390/v8110299.

[16] F. Negro, D. Forton, A. Craxì, et al. Extrahepatic morbidity and mortality of chronic hepatitis C, Gastroenterology. 149 (2015), 1345-1360. https://doi.org/10.1053/j.gastro.2015.08.035.

[17] L.E. Adinolfi, Chronic hepatitis C virus infection and neurological and psychiatric disorders: an overview, World J. Gastroenterol. 21 (2015), 2269. https://doi.org/10.3748/wjg.v21.i8.2269.

[18] M.P. Nahad, A. Bavi, M. Zandi, et al. Seroprevalence of hepatitis E virus infection among patients with acute hepatitis symptoms in Ahvaz, Iran, Int. J. Med. Lab. 5 (2018), 11-18.

[19] E. Effendy, A.A. Adriansyah, A.Y. Pulih Asih, et al. Improving public understanding about chronic hepatitis and diabetes mellitus in the era of the Covid-19 pandemic, Tomaega. 6 (2023), 35-43. https://doi.org/10.35914/tomaega.v6i1.1279.