Free Analytics Environment R Assignment 2

Regression analysis report

Table of Contents

Overview

In this report we will be looking at a dataset that contains crime stats including arrest numbers. The objective is to use machine learning techniques, in our case linear regression, to determine which factors affect and are related to the number of murder based arrests. We will make use of the R programming language as our analytical tool. In order to make sense of our data we will use different visualizations. Different libraries will be used in order to implement linear regression, plot correlations and perform other tasks for our analytical analysis. By the end of this report we will obtain a linear model that will be able to predict murder arrests in relation to its significant explanatory variables. We will also make sure that the properties for linear regression using OLS are linked to the residuals of the linear regression model. Lastly, we will take a look at how the insights gathered from this analysis can be put to practical use and benefit the police.

Dataset characteristics

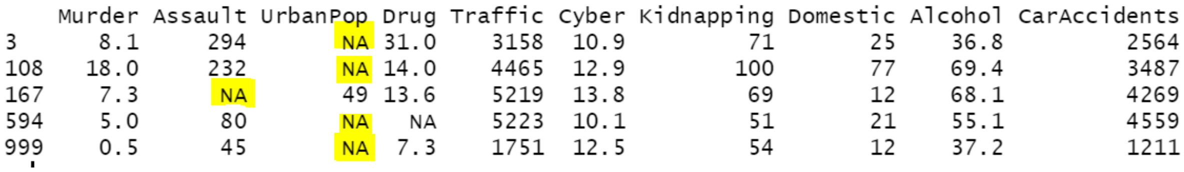
The data set “dataArrests” contains ten variables and a thousand entries. Eight of these variables show number of arrests with respect to a certain felony per 100,000 residents for the following felonies:

- Murder
- Assault
- Drug
- Traffic
- Cyber
- Kidnapping
- Domestic
- Alcohol

While the other two tell us the percentage of Urban population in an area (UrbanPop) and the number of recorded car accidents per 100,000 residents (CarAccidents).

The dataset under observation contains only numeric and integer values.

In the first step of our analysis (preprocessing), we make sure that the data does not contain incomplete or null values. Therefore, we remove the following rows.



```
    Murder Assault UrbanPop Drug Traffic Cyber Kidnapping Domestic Alcohol CarAccidents
3      8.1     294       NA 31.0    3158  10.9         71       25    36.8         2564
108   18.0     232       NA 14.0    4465  12.9        100       77    69.4         3487
167    7.3      NA       49 13.6    5219  13.8         69       12    68.1         4269
594    5.0      80       NA   NA    5223  10.1         51       21    55.1         4559
999    0.5      45       NA  7.3    1751  12.5         54       12    37.2         1211
```

Figure 1

Our analysis is performed on the remaining 995 observations.

Exploratory data analysis

The variable “Murder” is our dependent variable while the rest are considered explanatory variables. Firstly, let’s take a look at the distribution of the dependent variable.

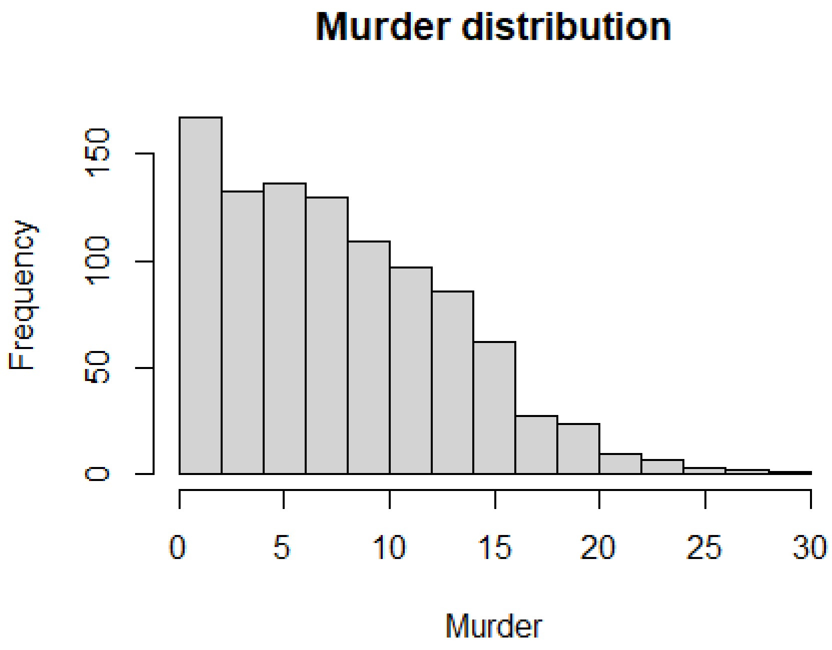


Figure 2

The histogram shows that data is not centered around the mean. The mean(7.75) is greater than the median(6.9). The distribution is right skewed. A box plot can be used to detect if there are any potential outliers.

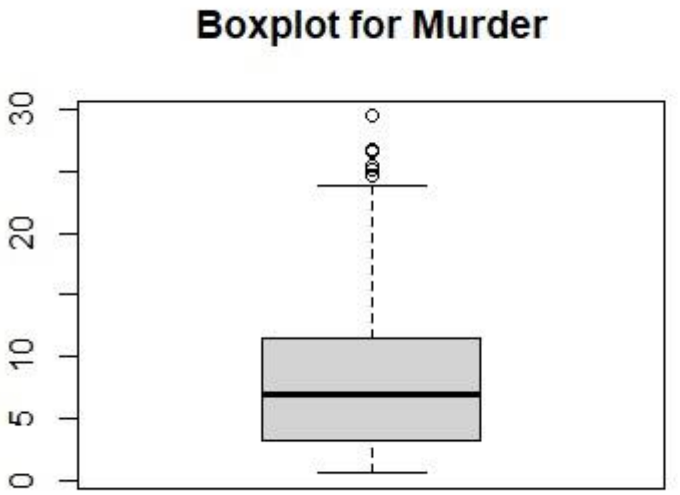


Figure 3

The box plot shows some potential outlier in our dependent variable, murder.

Lets take a look at the distribution of a few of the explanatory variables.

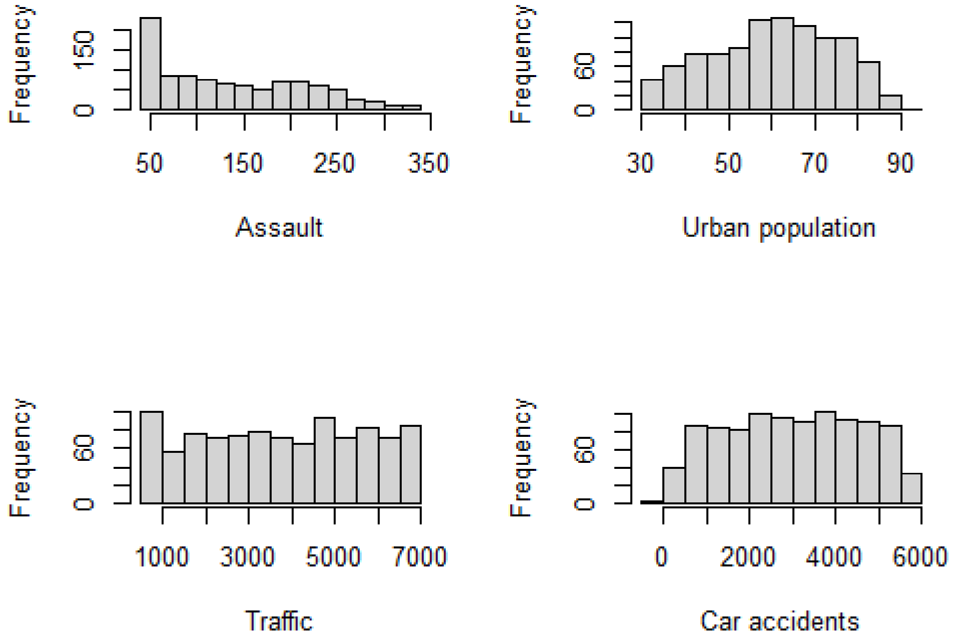


Figure 4

After examining the histograms one distribution stands out from the rest of the distributions, Urban population. It seems like a normal distribution. We can verify this claim by creating a scatter plot with boundaries specifying the mean and +-3x standard deviation.

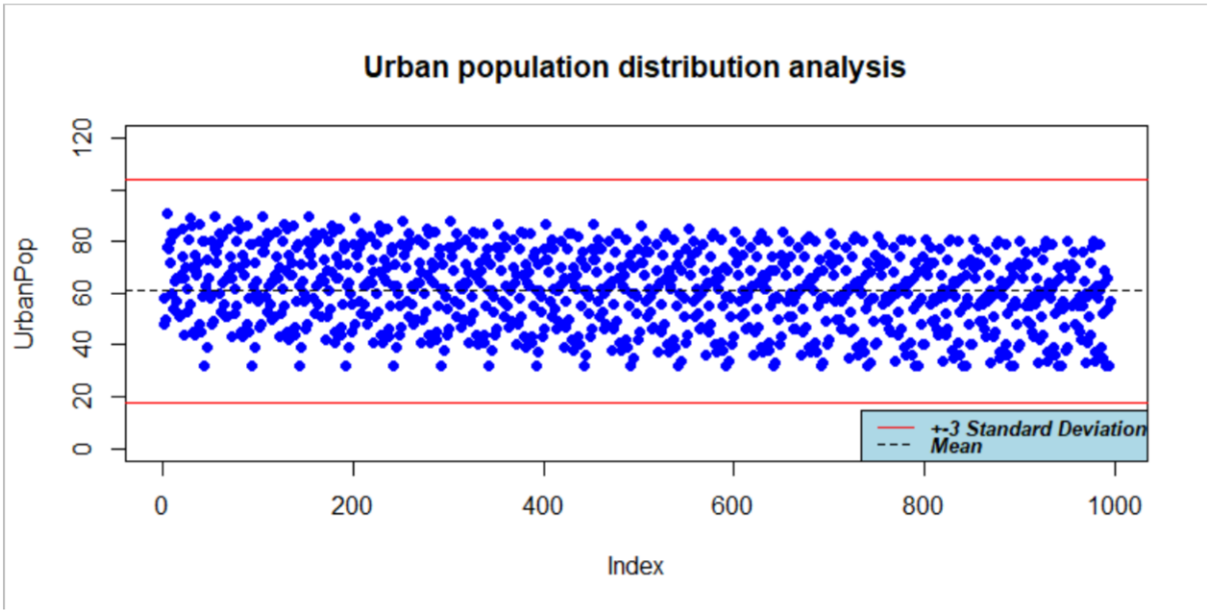


Figure 5

Looking at the scatter plot of the UrbanPop variable it is obvious that the distribution is centered around the mean with data values lie within 3 standard deviations from the mean in both directions. Hence it can be considered as a normal distribution.

Correlation analysis

Since our entire dataset is numeric we can perform correlation analyis on the entire dataset to get an idea of the relationship between each variable. First we take a look at the relationship of our dependent variable with the explanatory variables. For that purpose before we plot a correlation matrix. Let's first visualize the scatter plot of each variable with murder.

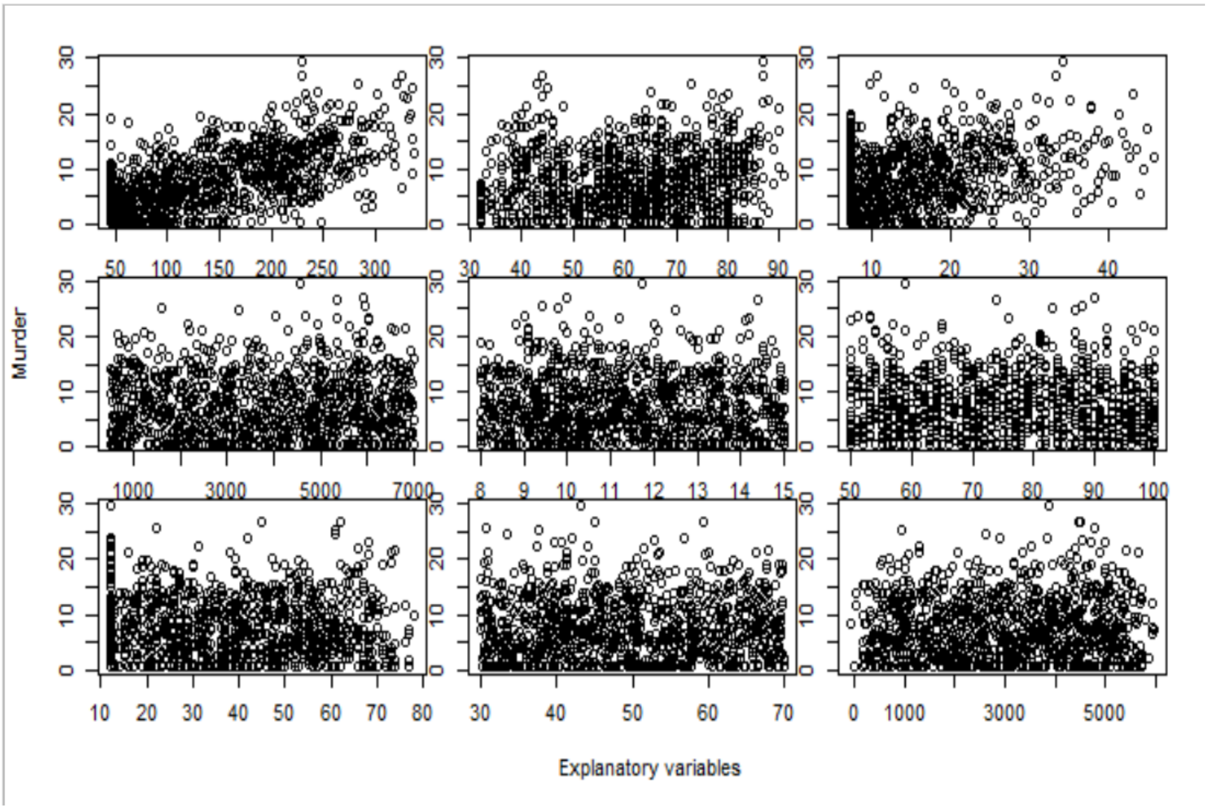


Figure 6

Figure 6 shows that none of the variables seem to have any obvious relation with Murder. Let's take a look at the correlation plot between all the variables in order to get a better understanding of the correlation between each variable.

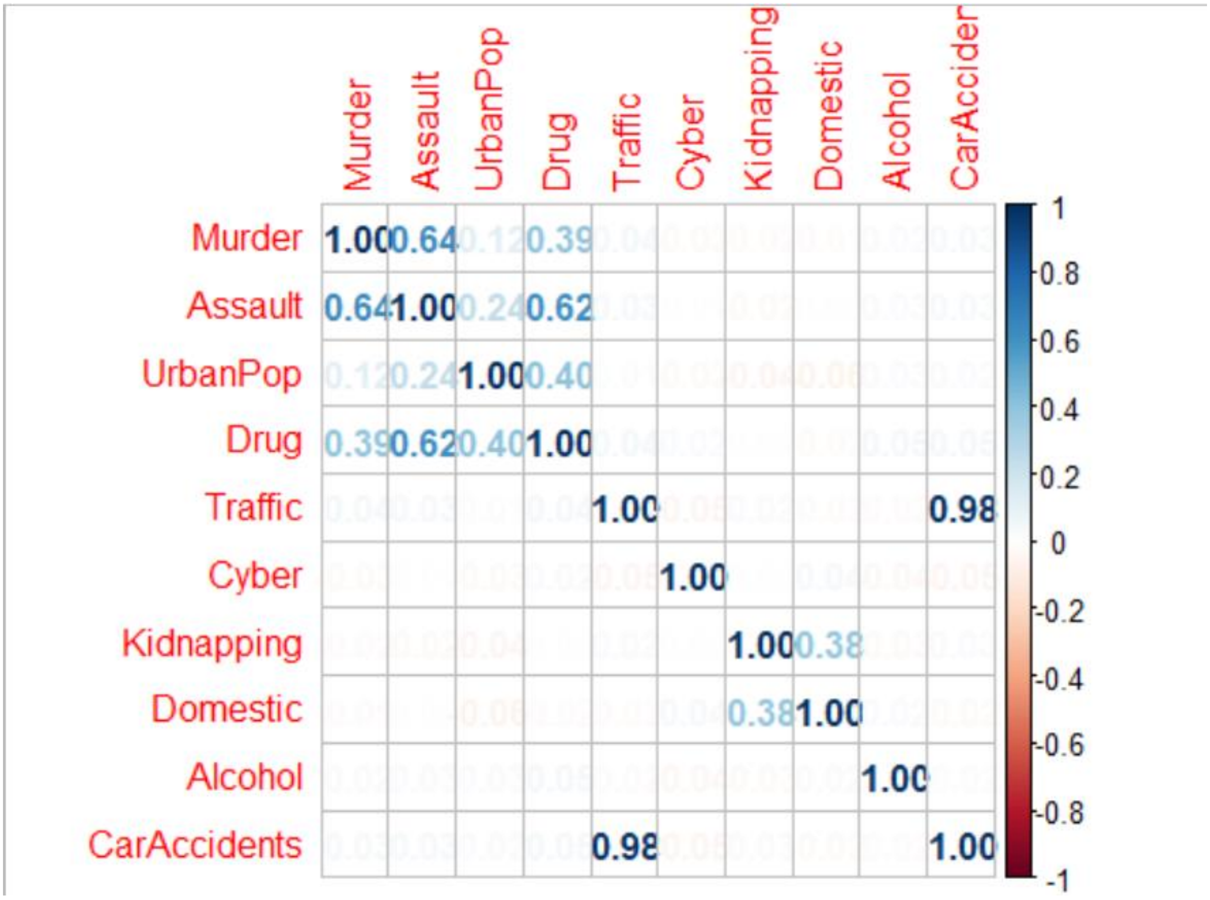


Figure 7

Figure 7 shows correlation between each variable and it is color coded to show strong positive correlation with a dark blue color while it shows strong negative correlation in dark red. The colors fade to white as the correlation approaches 0 indicating no correlation.

First we take a look at the relationship of our dependent variable with the explanatory variables. Murder has close to zero correlation with Urban population, Traffic violation arrests, Cybercrime arrests, Kidnapping arrests, Domestic violence arrests, Alcohol related arrests and Car Accidents per 100,000. It has a very low positive correlation with drug related arrests and a relatively high correlation with Assault based arrests. In conclusion none of the explanatory variables are highly correlated with Murder.

Figure 7 shows that the highest value in the correlation matrix is 0.98 between CarAccidents and Traffic. It means when traffic increases the number of car accidents increase as well and vice versa. The next highest values are 0.64 and 0.62 between Assault/Murder and between Assault/Drug respectively. Although these values do not seem too valuable.

In order to implement linear regression it is important to remove all those variables from the dataset that are linearly dependent. The threshold for high absolute correlation for our model is 0.8. As we can see in Figure 7 CarAccidents and Traffic are highly correlated and linearly dependent on each other with a value above our threshold. One of these variables has to be removed. We remove the variable that has the highest average correlation with all explanatory variables. Therefore we end up removing “CarAccidents” variable from our dataset. None of the other variables have a correlation above the threshold therefore we can now proceed to building our linear model.

Linear Regression

Using ”Murder” as our dependent variable and the remaining explanatory variables, we build our linear model. The general form of the output equation we are looking is a linear equation:

$$\hat{y} = \hat{\theta}_0 + \widehat{\theta_1} x_1 + \widehat{\theta_2} x_2 + \cdots + \widehat{\theta_n} x_n$$

Equation 1

Where:

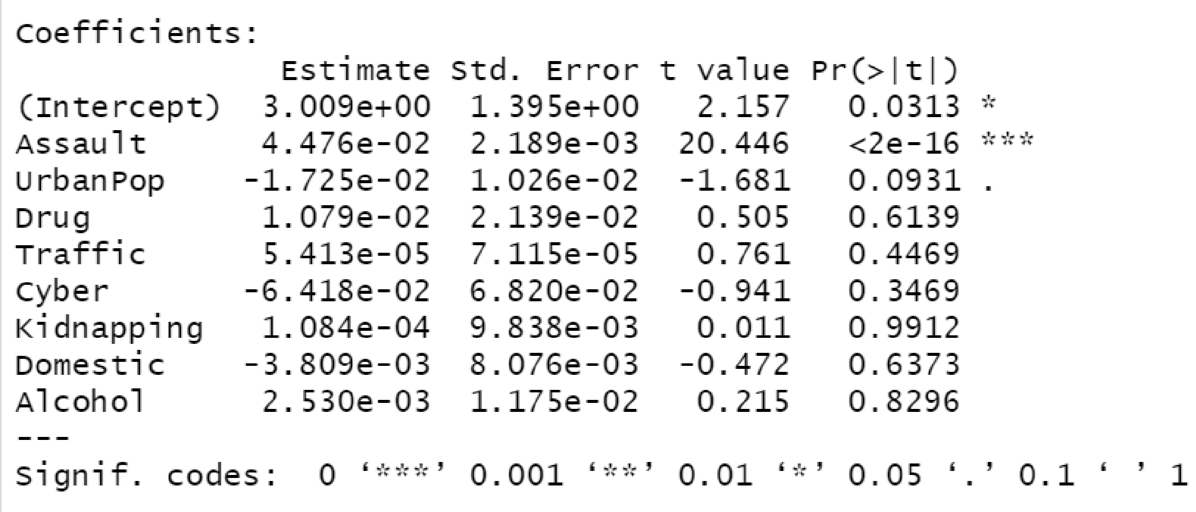
$\hat{y}$ is the predicted value for murder

$\hat{\theta}_0$ is the intercept

$\hat{\theta}_1$ - $\hat{\theta}_n$ are the parameters for our model.

x_1 - x_n are the values of our explanatory variables from the dataset

After running the first cycle of our linear regression model we get the following estimates of our coefficients.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.009e+00  1.395e+00   2.157   0.0313 *  
Assault      4.476e-02  2.189e-03  20.446   <2e-16 ***
UrbanPop    -1.725e-02  1.026e-02  -1.681   0.0931 .  
Drug         1.079e-02  2.139e-02   0.505   0.6139    
Traffic      5.413e-05  7.115e-05   0.761   0.4469    
Cyber       -6.418e-02  6.820e-02  -0.941   0.3469    
Kidnapping   1.084e-04  9.838e-03   0.011   0.9912    
Domestic    -3.809e-03  8.076e-03  -0.472   0.6373    
Alcohol      2.530e-03  1.175e-02   0.215   0.8296    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

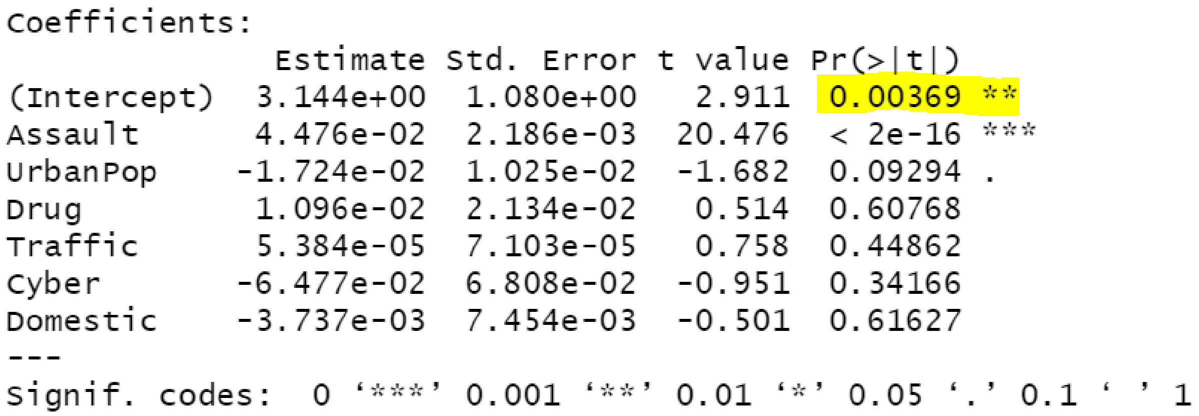
Figure 8

The last column of this figure shows p values for the null hypothesis: coefficient is zero. The smaller the p value the stronger the evidence against the null hypothesis. In order to increase accuracy of our model we remove a variable with the highest p value from the model and re run until all p values are highly significant and close to zero.

After the initial run of our model it can be noted that the intercept value may not be zero since the p value for the estimated intercept is significantly low. We remove “Kidnapping“ (as it has the highest p value) and re-run the model.

After the second run no major change is noted in the summary of our coefficients so we run the model again, this time we remove “Alcohol” as it has the highest p value among the existing variables.

After the third run we observe that p value for the intercept becomes even more significant indicating that the intercept is a non-zero value. We re run the model as there are still a lot of non-significant variables remaining. This time we remove “Domestic” since it has the highest p value as indicated in Figure 9.

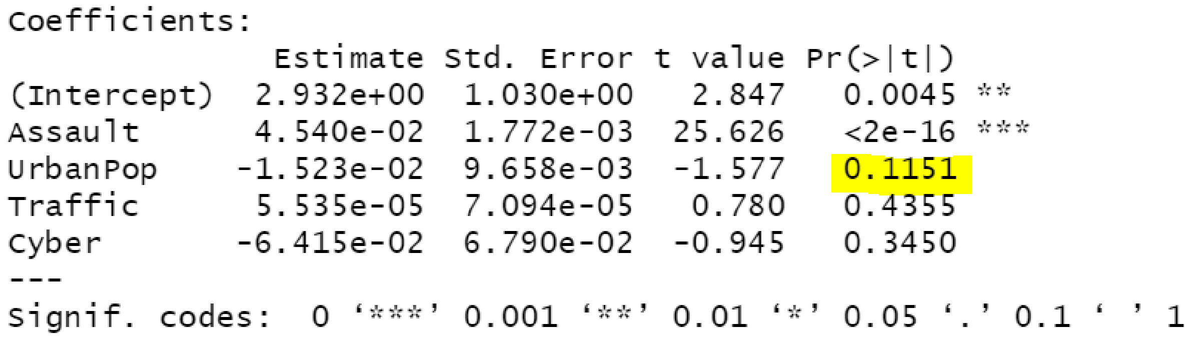


```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.144e+00  1.080e+00   2.911  0.00369 ** 
Assault      4.476e-02  2.186e-03  20.476  < 2e-16 ***
UrbanPop    -1.724e-02  1.025e-02  -1.682  0.09294 .  
Drug         1.096e-02  2.134e-02   0.514  0.60768    
Traffic      5.384e-05  7.103e-05   0.758  0.44862    
Cyber       -6.477e-02  6.808e-02  -0.951  0.34166    
Domestic    -3.737e-03  7.454e-03  -0.501  0.61627    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 9

After removing “Domestic” no major change is noticed in the significance of the p values. We choose to remove “Drug” from our linear model which had the highest p value at 0.60591.

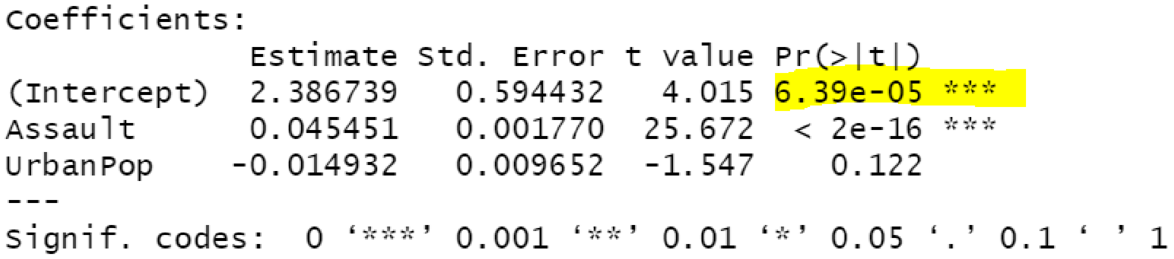
After re running the model with the updated parameters we notice that the coefficient for “UrbanPop” becomes less significant as shown in Figure 10.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.932e+00  1.030e+00   2.847   0.0045 ** 
Assault      4.540e-02  1.772e-03  25.626   <2e-16 ***
UrbanPop    -1.523e-02  9.658e-03  -1.577   0.1151    
Traffic      5.535e-05  7.094e-05   0.780   0.4355    
Cyber       -6.415e-02  6.790e-02  -0.945   0.3450    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10

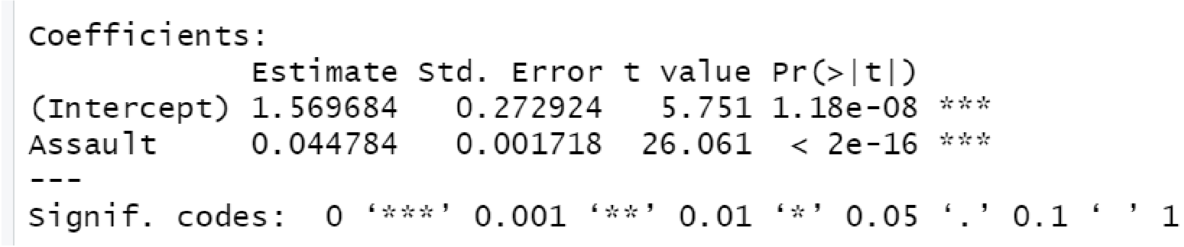
Moving on, we remove Traffic from the model on account of its high p value. No significant change is noted. Therefore we remove the next variable with the highest p value, Cyber. Upon removal of “Cyber” we notice that the significance of the intercept increases and approaches it true value.



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.386739   0.594432   4.015 6.39e-05 ***
Assault      0.045451   0.001770  25.672  < 2e-16 ***
UrbanPop    -0.014932   0.009652  -1.547    0.122    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 11

“UrbanPop” is the last remaining non-significant variable. After its removal we conclude our model and are left with the following coefficient summary:



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.569684   0.272924   5.751 1.18e-08 ***
Assault     0.044784   0.001718  26.061  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 12

Both the intercept and “Assault” variable have highly significant p values which means that the estimated value is approximately their true predicted values.

Regression equation of the final model is as follows:

$$\hat{y} = \ 1.57 + \ 0.04. Assault$$

Equation 2

The equation can be interpreted to mean that there is a fixed amount of murders, followed by an increase of 0.04 as one unit of Assault increase. The data points for Assault given in dataset are plotted in the figure below along with the best fit line attained from our regression model.

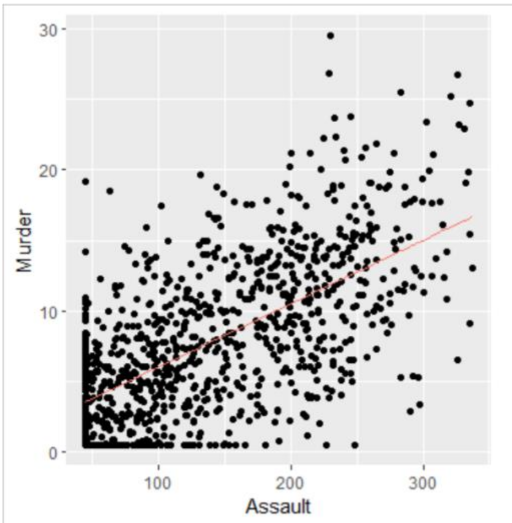


Figure 13

Examine the Residuals using OLS

A residual is a measure of how far away a point is vertically from the regression line. It can be represented by the following equation.

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

Equation 3

In this section we look at five properties for linear regression using OLS that are linked to the residuals of a linear regression model.

Mean of residuals

One of the properties states that the mean of the residuals should be 0. In our case the mean of residuals accumulates to 2.386073e-17 which is quite close to 0.

Variance of residuals – Homoscedasticity

The variance of the residuals should be constant.

Variance of residuals in our linear regression is 18.19 which is less than infinity hence it is homoscedastic.

Residuals are linearly independent

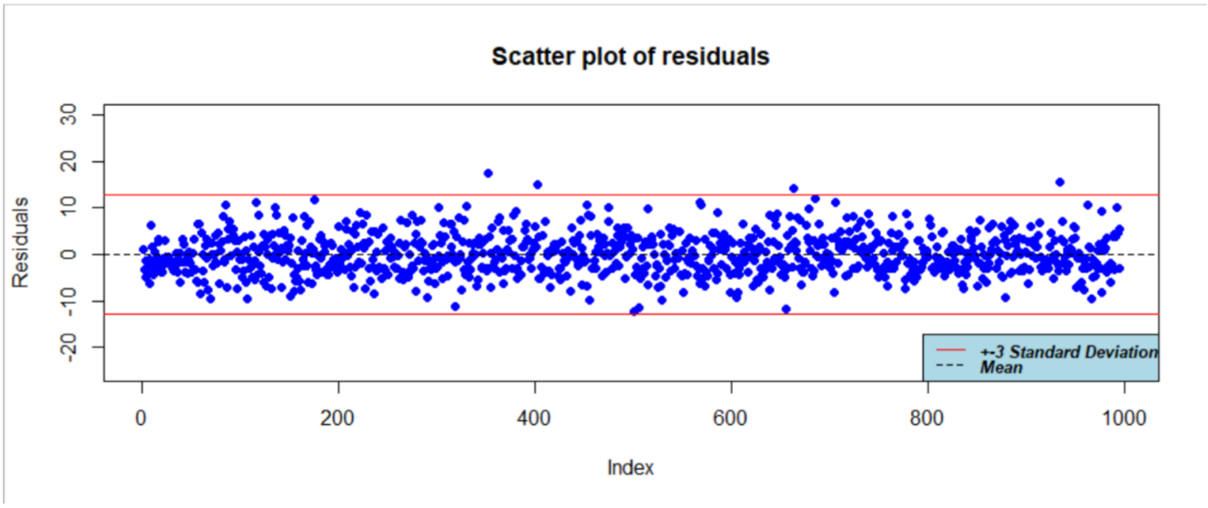


Figure 14

Figure 14 shows the distribution of the residuals using a scatter plot. Linear independency between residuals can be seen in the figure. In order to strengthen our argument we will make use of The Durbin-Watson test.

The Durbin-Watson test is a test for autocorrelation in the output of a regression model. The statistic value, also known as DW statistic, ranges from 0 to 4. The midpoint, 2 indicate 0 autocorrelation. While a value below 2 means there is positive correlation and above 2 means negative correlation. The test tests the null hypothesis that the linear regression residuals are uncorrelated.

Running the Durbin-Watson test for our linear regression model returns a DW statistic value of 2.02 which can be considered 2 when rounded off to the nearest integer. Hence the residuals adhere to the property of linear independence.

No relationship between the residuals and each explanatory variable

Let's take a look at the correlation value for each explanatory variable with the residuals in order to determine relationship between them.

Correlation b/w residuals and	**Correlation**
Assault	-4.132151e-17
UrbanPop	-0.04758083
Drug	-0.001138525
Traffic	0.02589428
Cyber	-0.02982884
Kidnapping	-0.003216544
Domestic	-0.01454106
Alcohol	0.00664737

All the correlation values are close to zero hence we conclude that there is no relationship between the residuals and explanatory variables.

Residuals are normally distributed

In order to examine this property let's take a look at Figure 14 again. It can be seen that majority of the data points can be found around the mean and within three standard deviations from the mean, which is the basis for a normal distribution. However in order to strengthen our claim we shall make use of the Jarque-Bera Test.

The Jarque-Bera test is a test for normality that uses two moments of a distribution, skewness and kurtosis (certain level of tiredness). A normal distribution is not skewed and has kurtosis of 3. The test tests the null hypothesis that skewness of the distribution is 0 and excess kurtosis is 0.

Running the test on our residuals reveals the statistic for skewness to return a value of 0.3 with p value 3.175e-06 and statistic for kurtosis returns a value of 0.02198. Results of the test support our claim that the residuals are normally distributed.

We can plot a histogram to verify that the shape of the distribution of the residuals is similar to a normal distribution.

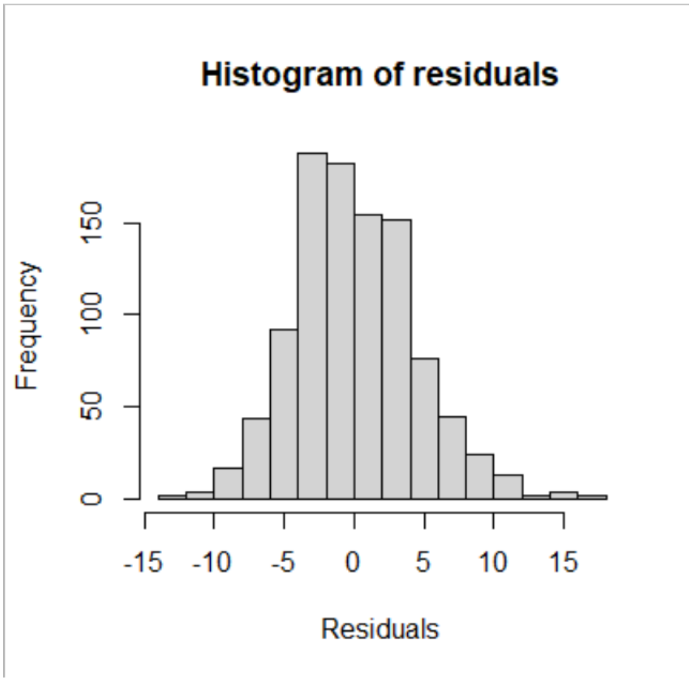


Figure 15

Conclusion

We started our analysis with a big dataset containing ten variables. One dependent and the rest explanatory. Using the linear regression model we came to the conclusion that only one variable has a significant impact on the number of murder arrests as we iteratively removed the insignificant variables from our model. The number of murder based arrests is linearly related to the number of assault based arrests with a coefficient of 0.04. Which means the value of murder based arrests increase at a rate of 0.04 in relation to assault based arrests.

The police can use these insights to efficiently allocate resources in the concerned departments. When the number of assault based arrests are increasing in an area they can increase the resources and budget allocation of the murder investigation unit due to its linear increasing relation with assault arrests. They can also allocate additional resources in efforts to raise awareness on how to stay safe and contact the police in case of emergency in areas with higher assault based arrest. Using the linear regression model they can make predications and make timely decisions that can end up saving lives!