Project 4 Part II

Kaushal Khatiwada

2024-05-02

Part 2

Using "USArrests" Dataset

A) Create an "crime" dataset containing all the variables of USArrests

```
crime <- USArrests
```

B) Create correlation matrix plot of the crime data and interpret each scatterplot carefully

```
cor(crime)
```

```
##              Murder  Assault   UrbanPop      Rape
## Murder   1.00000000 0.8018733 0.06957262 0.5635788
## Assault  0.80187331 1.0000000 0.25887170 0.6652412
## UrbanPop 0.06957262 0.2588717 1.00000000 0.4113412
## Rape     0.56357883 0.6652412 0.41134124 1.0000000
```

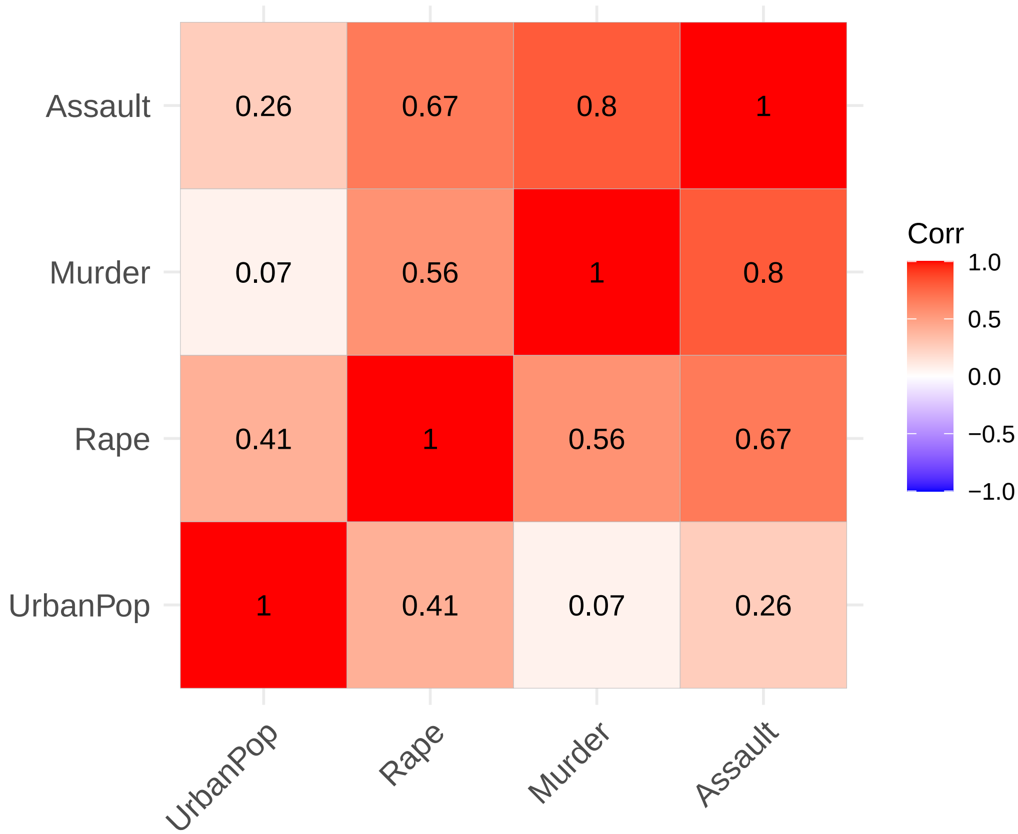
visualize correlation matrix

```
library(ggcorrplot)
```

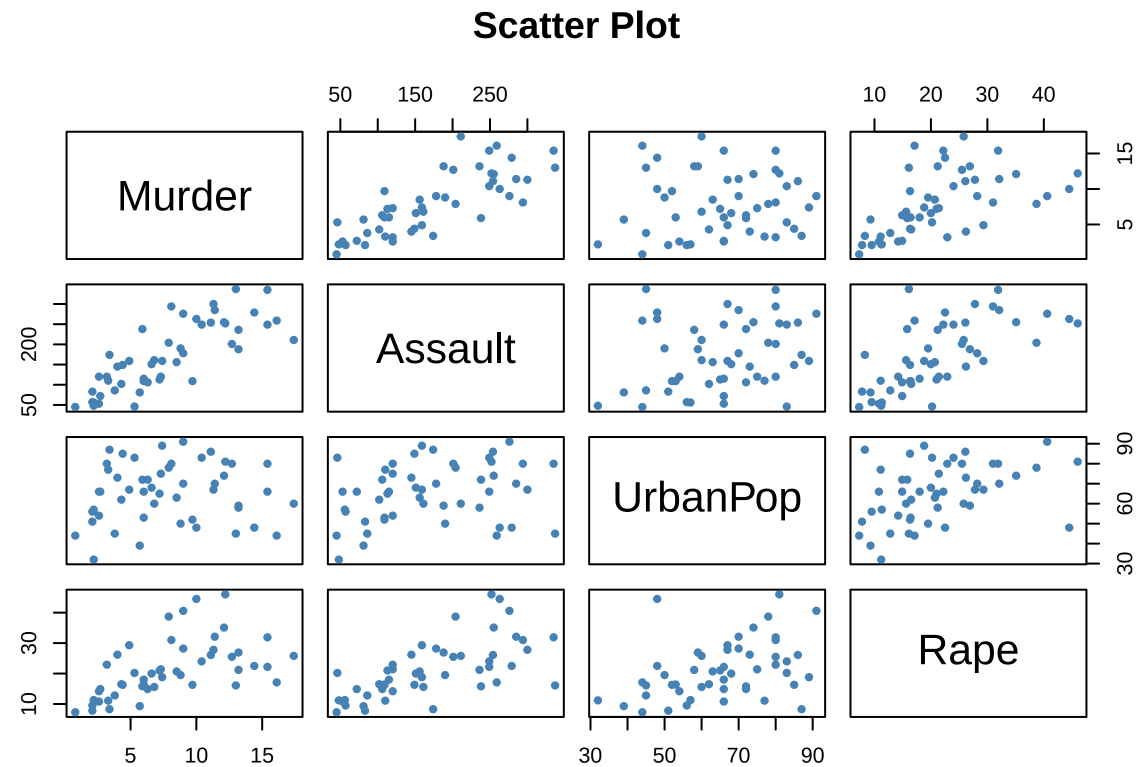
```
## Loading required package: ggplot2
```

```
library(ggplot2)
```

```
ggcorrplot(cor(crime),hc.order = T,lab = T)
```




```r
plot(crime, pch=20,cex=1,col='steelblue',main="Scatter Plot")
```




```
#Relationship between two variable
      # -1 inverse linear relationship,
      # 0 no linear correlation,
      # 1 linear relationship]
```

C) Split the crime dataset into training and testing data with 70% and 30% cases

```
set.seed(13)
index <- sample(2,nrow(crime),replace = T,prob = c(0.7,0.3)) #Random sampling into two independent vari
train.crime <- crime[index==1,] #Training set
test.crime <- crime[index==2,]  #Test set
```

D) Fit a multiple linear regression on training data with Murder as dependent variable and all other variables as independent variables and interpret the results carefully using R-squared, RMSE, Regression ANOVA and Regression Coefficients (BLUE?)

```
mlr <- lm(Murder ~ .,data=train.crime)
summary(mlr)
```

```
##
```

```
## Call:
## lm(formula = Murder ~ ., data = train.crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8132 -1.7164 -0.5668  1.2990  7.1819
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.16946    2.05877   2.025    0.051 .
## Assault      0.04653    0.00782   5.950 1.12e-06 ***
## UrbanPop    -0.05151    0.03217  -1.601    0.119
## Rape        -0.02633    0.06981  -0.377    0.709
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.518 on 33 degrees of freedom
## Multiple R-squared:  0.6648, Adjusted R-squared:  0.6343
## F-statistic: 21.81 on 3 and 33 DF,  p-value: 5.711e-08
```

R-squared = 0.6648

```r
mean(mlr$residuals^2)    #RMSE= 5.656679
```

```
## [1] 5.656679
```

```r
mlr$coefficients        #Regression Coefficient (Intercept)
```

```
## (Intercept)     Assault    UrbanPop        Rape
##  4.16946411  0.04653252 -0.05150877 -0.02632538
```

```r
anova(mlr)          #Regression ANOVA
```

```
## Analysis of Variance Table
##
## Response: Murder
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Assault    1 393.06  393.06 61.9733 4.475e-09 ***
## UrbanPop   1  21.07   21.07  3.3216   0.07745 .
## Rape       1   0.90    0.90  0.1422   0.70852
## Residuals 33 209.30    6.34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

"Assault" is highly significance, p-value<0.05

E) Check multicollinearity and finalize this model with the appropriate VIF cut-off value

```r
library(car)
```

```
## Loading required package: carData
```

```r
vif(mlr)
```

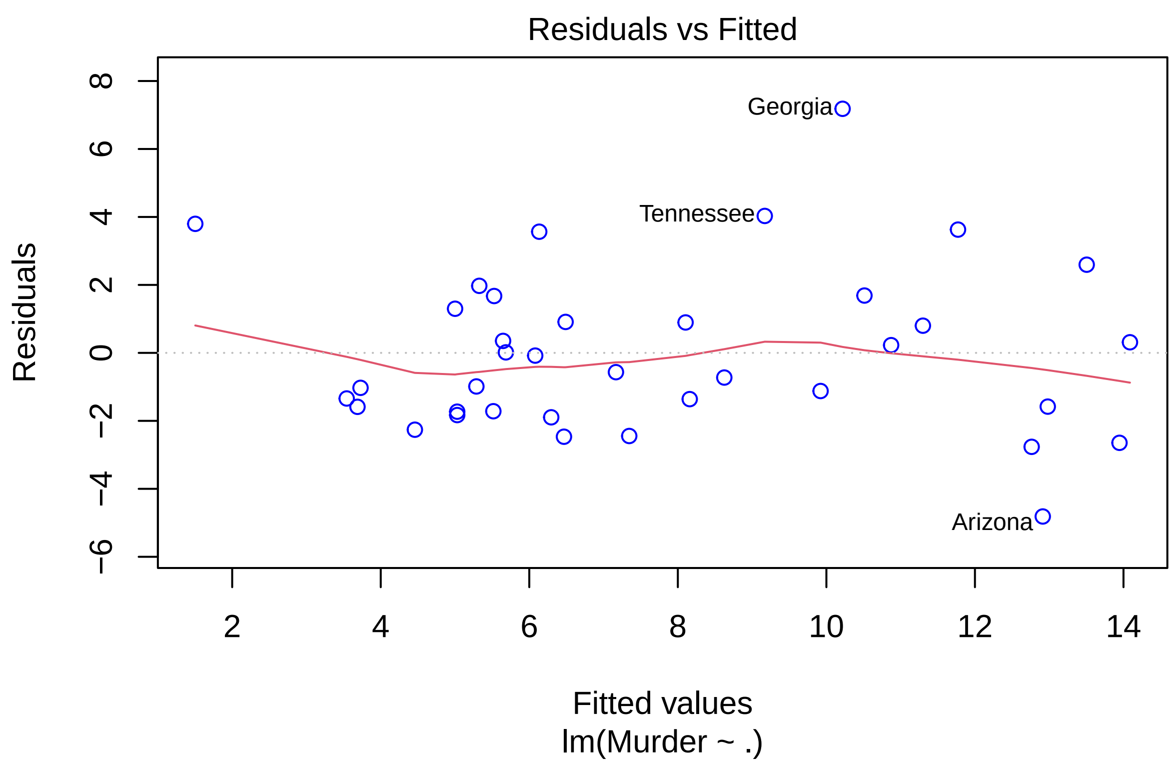
```
##  Assault UrbanPop     Rape
## 2.087282 1.156333 2.305631
```

No Multicollinearity beacause VIF of all variable are <10

F) Perform residual analysis of this model i.e. LINE tests using suggestive graphs and confirmatory tests and interpret the results carefully

##LINE ## L = Linearity of residuals

```r
### Suggestive
plot(mlr,which = 1,col=c("blue")) # LOESS line lies in the zero line so residuals are linear
```

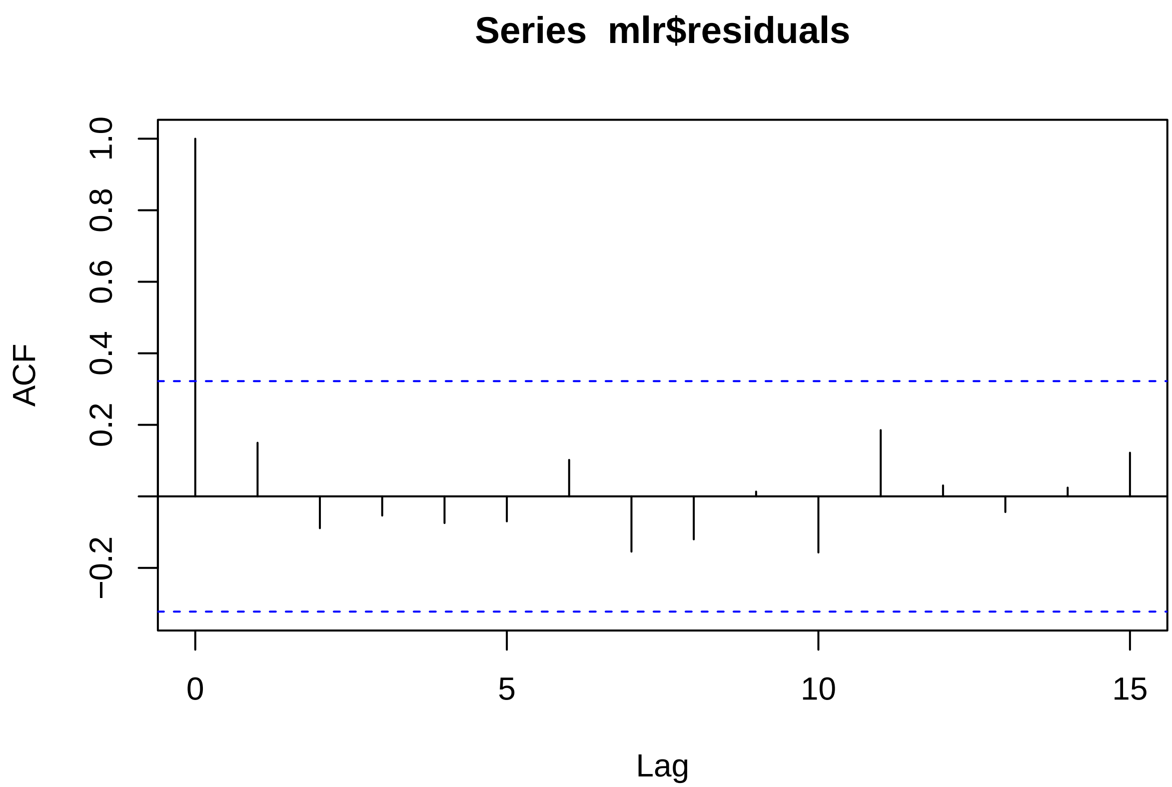



```r
### Confirmative
summary(mlr$residuals)    #Mean = 0 so residuals are linear
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -4.8132 -1.7164 -0.5668  0.0000  1.2990  7.1819
```

I = Independence of residuals

```
### Suggestive
acf(mlr$residuals)        # Show Up and Down so no autocorrelation
```

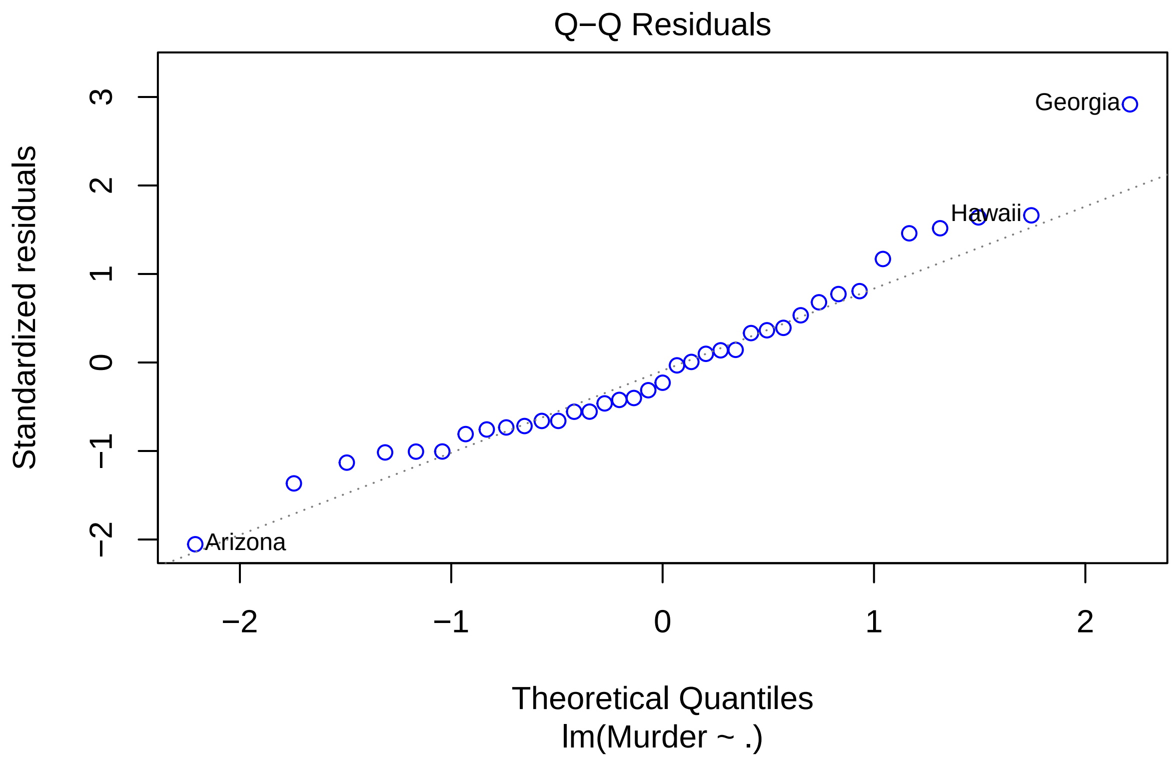



```
### Confirmative
library(car)
durbinWatsonTest(mlr)      # p-value>0.05 no autocorrelation
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1498789      1.654913   0.296
##  Alternative hypothesis: rho != 0
```

N = Normality of residuals

```
## Suggestive
plot(mlr,which = 2,col=c("blue"))
```

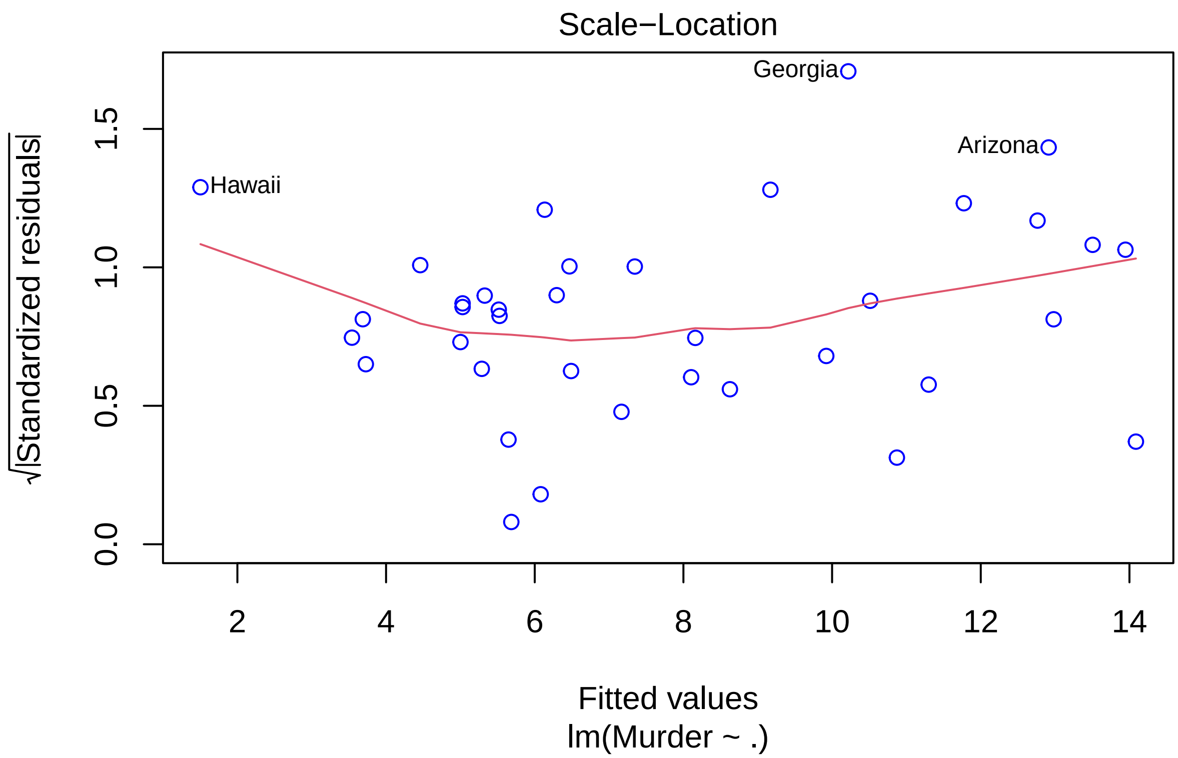



```r
### Confirmative
shapiro.test(mlr$residuals)    # p-value > 0.05, residuals follow normal distribution
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mlr$residuals
## W = 0.94906, p-value = 0.09001
```

E = Equal variance of residuals

```r
### Suggestive
plot(mlr,which = 3,col=c("blue"))
```




```r
### Confirmative
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
bptest(mlr)                # p-value > 0.05, so residual variances are equal (homoscedasticity)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mlr
## BP = 1.7374, df = 3, p-value = 0.6287
```

G) Predict the Murder in the testing dataset using the fitted model

```r
prediction <- predict(mlr,test.crime)
prediction
```

```
##        Alabama     California       Delaware        Florida          Idaho
##      11.605533      11.256332      11.119632      14.797378       6.598073
##       Illinois          Maine North Carolina   North Dakota   Rhode Island
##      10.849025       5.199378      17.109191       3.804866       7.566359
##          Texas       Virginia      Wisconsin
##       8.730502       7.638550       2.951795
```

H) Report R-square and RMSE of predicted model and interpret them carefully

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.3
```

```
## Loading required package: lattice
```

```
## Warning: package 'lattice' was built under R version 4.3.3
```

```r
data.frame(R2 = R2(prediction, test.crime$Murder),
           RMSE = RMSE(prediction, test.crime$Murder))
```

```
##          R2     RMSE
## 1 0.7008134 3.049961
```