|  |  |
| --- | --- |
| A close up of a logo  Description automatically generated  Assignment 01  REGRESSION MODELS | Abstract  **This Assignment provides the brief insights of Regression Model using R including s box plot, histogram, scatterplot, correlation plot, Q-Q plot etc.**  Kamaldeep Kaur  MBA-6693 |

**Introduction**  
In this Assignment, I am using an inbuilt dataset from R i.e. USArrests. Selected dataset USArrests includes the data of different crimes that have been committed in the USA i.e. Rape, Assault, Murder, and UrbanPop, and will help to understand the crime level in the US. There are 50 rows and 4 columns in the data which can be shown by using the below codes:  
nrow(USArrests)

ncol(USArrests)

The primary goal to take this data to understand the crime rate and the arrest rate of all the cities in US. We wanted to find the highest arrest rate compared to other states in US.

**Library**

To begin with, I have included the following libraries in the R:  
library(dplyr)

library(ggplot2)  
library(reshape2)  
library(Information)  
library(ggcorrplot)

**Scatter Plots**  
Scatter Plots describes the data set of two variables (Murder and Rape) on the Graph where the x-axis shows the murder rate in the US and the Y-axis shows the Rape rate in the US. I have included the R code below:

A picture containing photo, outdoor, lot, side

Description automatically generatedlibrary(help="datasets")  
USArrests  
ggplot(USArrests)  
ggplot(USArrests, aes(Murder, Rape))+geom\_point()  
As shown in graph 1.1 You can see the pattern of the resulting points that how much one variable is affected by other.

***Graph1.1 Scatter Plots***

**Linear Regression Model**  
After running the “lm” command, using UrbanPop as a response variable and Rape as an explanatory variable, we can check the dependency of the variables. I have found out the slope and the intercept. We can see that number of rape arrests in the US per 100,000 residents rises by 0.26 and the p-value is 0.003 which is small (typically ≤ 0.05) indicates strong evidence against the null hypothesis.  
y=β0 + β1x (Regression Model) where β0 is intercept and our slope is β1.

Detail<-lm(formula = Rape ~ UrbanPop, data=USArrests)

summary(Detail)

**OUTPUT:**

lm(formula = Rape ~ UrbanPop, data = USArrests)

Residuals:

Min 1Q Median 3Q Max

-18.644 -5.476 -1.216 5.885 27.937

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.78707 5.71128 0.663 0.510

UrbanPop **0.26617** 0.08513 3.127 0.003 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.626 on 48 degrees of freedom

Multiple R-squared: 0.1692, Adjusted R-squared: 0.1519

F-statistic: 9.776 on 1 and 48 DF, p-value: 0.003001

**Regression Line**I have written the code below to add a trend line on a graph.

ggplot(USArrests, aes(Murder,Rape))+geom\_point()+geom\_smooth(method="lm")   
According to Graph 1.2, You can see that there is one Blue line that fits the data properly. Basically, Regression line shows the relationship between the explanatory and response variable.

**A picture containing text, photo, people, boat

Description automatically generated**  
***Graph 1.2: Regression Line***

**HISTOGRAM**

I have included the R code ggplot and geom histogram that shows how often each unique value occurs in datasets. We can see the histogram of Rape Arrests in the below graphs 1.3. As we can see that rape arrests are highest under the category of 15-25 whereas lowest in 42-50.

ggplot(USArrests, aes(Rape))+ geom\_histogram(bins = 10)

A picture containing clock

Description automatically generated  
***Graph 1.3 Rape Arrest***

**MULTIPLE REGRESSION**

Multiple regression predicts the value of one variable based on another two or more independent variables.

Model<-lm(Rape~Murder+Assault+UrbanPop, data=USArrests)

print(Model)

**OUTPUT:**

lm(formula = Rape ~ Murder + Assault + UrbanPop, data = USArrests)

Coefficients:

(Intercept) Murder Assault UrbanPop

-2.47410 0.41856 0.04893 0.18448

**Correlation Matrix**Correlation is used to see how strong the dependency or the relationship is between other variables. install.packages("ggcorrplot")

library(ggcorrplot)

data("USArrests")

corr <- round(cor(USArrests), 1)

head(corr[, 1:4])

**OUTPUT:**

Murder Assault UrbanPop Rape

Murder 1.0 0.8 0.1 0.6

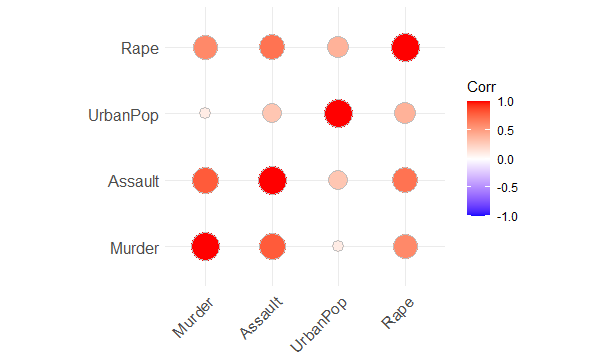
Assault 0.8 1.0 0.3 0.7

UrbanPop 0.1 0.3 1.0 0.4

Rape 0.6 0.7 0.4 1.0

ggcorrplot(corr)  
A screenshot of a cell phone

Description automatically generated  
  
ggcorrplot(corr, method = "circle")



**BOX PLOT**

input <- USArrests[,c('Murder','Assault')]

> print(head(input))

**OUTPUT:**

Murder Assault

Alabama 13.2 236

Alaska 10.0 263

Arizona 8.1 294

Arkansas 8.8 190

California 9.0 276

Colorado 7.9 204

boxplot(Murder ~ Rape, data = USArrests, xlab = "Murder rate in USA",ylab = "Rape rate in US")

**A close up of a logo

Description automatically generated**

**Logistic Regression**

glm() is the function that tells R to run a generalized linear model.

> Hello<-glm(formula = Murder ~ Rape, data=USArrests)

> summary(Hello)

Call:

glm(formula = Murder ~ Rape, data = USArrests)

Deviance Residuals:

Min 1Q Median 3Q Max

-5.0900 -2.4025 -0.5731 1.7095 9.3949

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.22367 1.28456 1.731 0.0899 .

Rape 0.26207 0.05544 4.727 2.03e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 13.21473)

Null deviance: 929.55 on 49 degrees of freedom

Residual deviance: 634.31 on 48 degrees of freedom

AIC: 274.92

Number of Fisher Scoring iterations: 2

**Q-Q Plot**

model<-lm(Murder~Rape,data=USArrests)

> plot(model)

Visualizing model fit is very important aspect to see how the regression works in R. Main feature to recognize is residuals are normally distributed and mean of residuals are zero.

From the residuals vs fitted graph, it is a scatter plot of residuals on the y axis and fitted values (estimated responses) on the x axis. we observe that the plot detects non-linearity, unequal error variances, and outliers while comparing murder and rape category.

A close up of a map

Description automatically generated

From the residuals vs Leverage graph, it is a scatter plot of residuals on the y axis and leverage values (estimated responses) on the x axis. The Residuals vs. Leverage plots helps us to identify influential data points on model. We can see the Mississippi has the highest standardized residuals within the dash cooked distance range.

A close up of a map

Description automatically generated

From the Scale Location graph, we understand whether our residuals are spread equally along the predictor range. We observe from below graph that our line starts off horizontal at the beginning of our predictor range, slopes up around 8, and then flattens again around 10. The line goes up because the residuals for those predictor values are more spread out with murder category.

A screenshot of a cell phone

Description automatically generated

The Q-Q plots help us to provide quick comparisons between probability distributions and can tell us how closely a data sample is to normally distributed. We can observe that the trad is increasing from –2 to 1 and curvy slopes up from 1 to 2.

A close up of a map

Description automatically generated

**Conclusion**

The primary goal of the assignment is to understand the different regression model and correlation of each variables in the dataset and see how one variable trends different from the other variable .We understood that Mississippi has the highest influential data model compared to the other states in the US.