Stats for Data Science

Mini Project # 6

Names of group members:

- Sashidhar Donthiri
- Sathya Pooja RamiReddy

Contribution of each group member: Both the team members have done it individually and further discussed and put together all the results.(Also done with the bonus part given in the class)

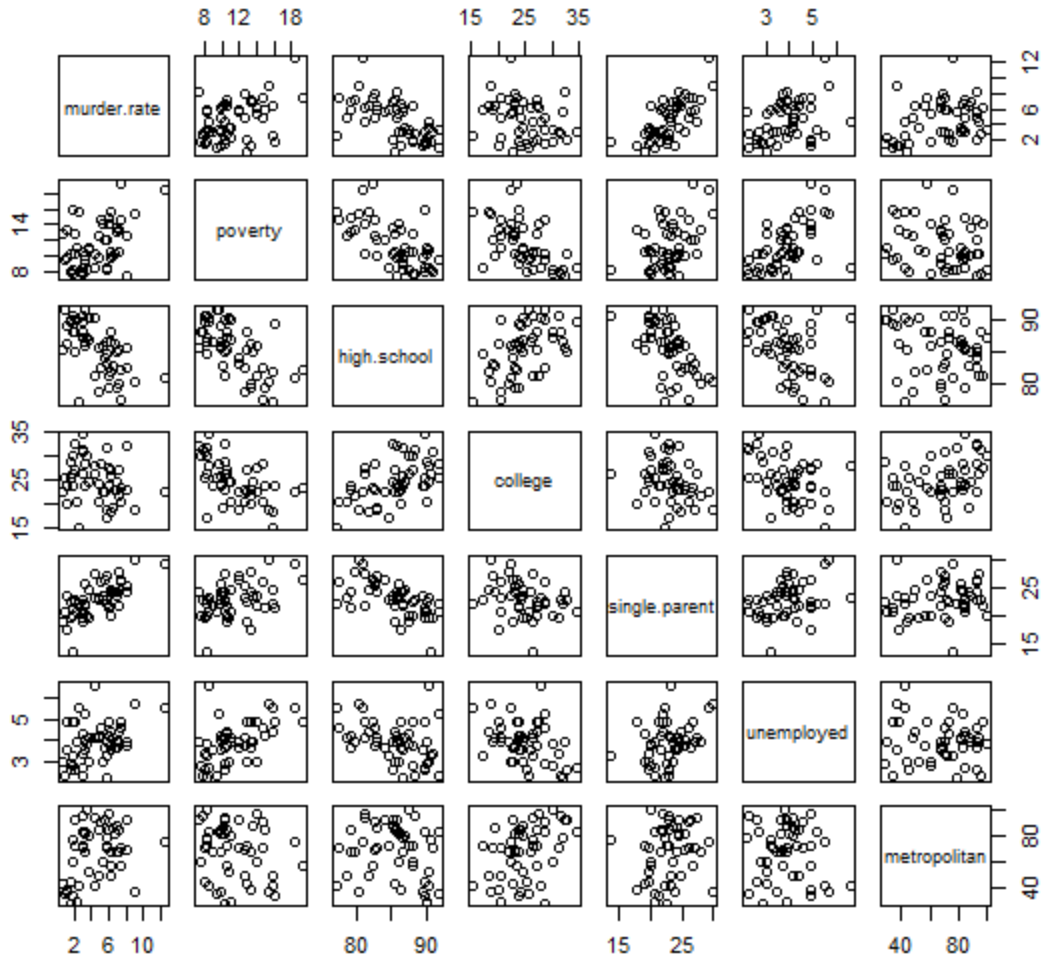
Section 1

Given:

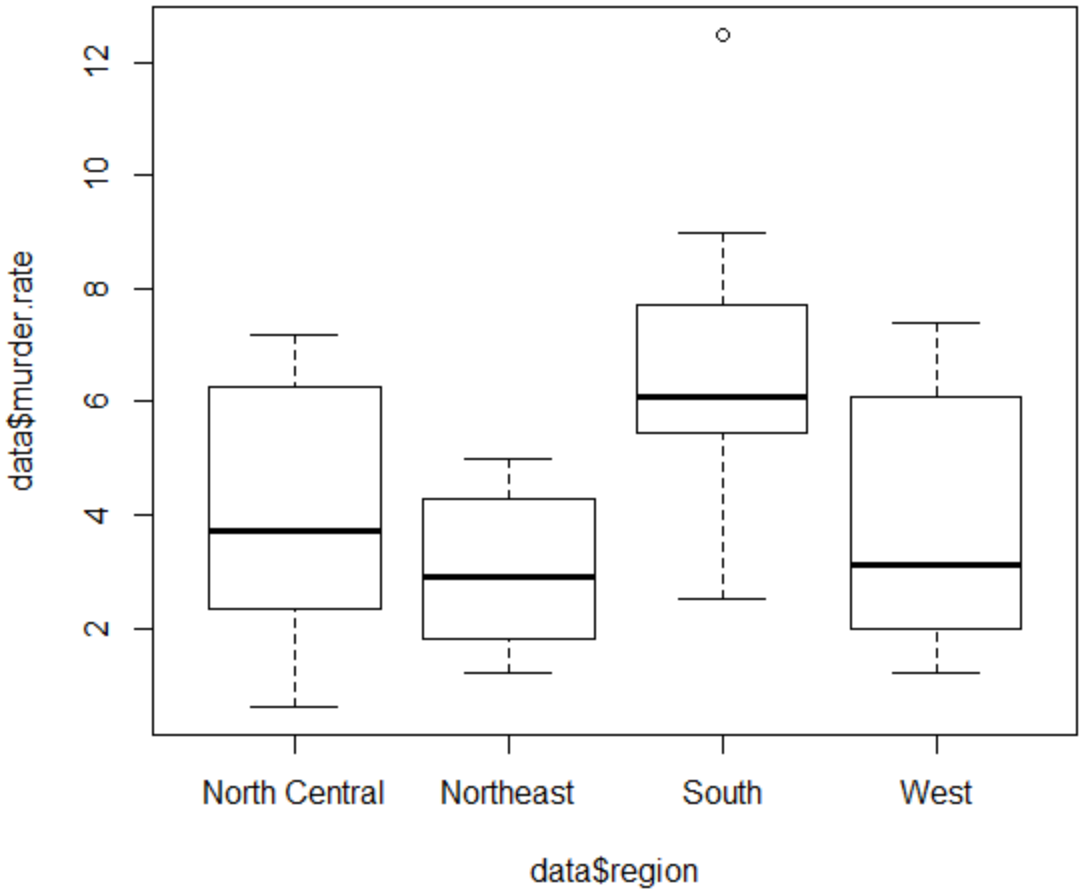
Consider the crime data stored in crime.csv. We would like to understand how murder rate is related to the other variables in the dataset. Note that state is the "subject" here; it's not a predictor, and region is a qualitative variable.

(a) *Fit a multiple linear regression model to predict murder rate based on the other variables. Perform model diagnostics to check assumptions and perform any transformations needed to obtain a model that is reasonable with respect to the standard assumptions for linear models.*

Step 1:
Import data from the csv file and compare the relationship between murder.rate and other predictors based on scatter plots generated using pairs function for the numeric predictors.
Step 2:
Based on the plot obtained in the image below, we observe that murder rate seems to have a stronger linear relationship with high school, single parent and unemployed predictor compared to others.



- In case of categorical variables, we used box plot to analyze murder rate against the categorical predictor, region.
- Based on the box plot below, Region South seems to have the highest median murder rate compared to other regions.



We fit the full model with all predictors against murder rate using the lm function in R.

```
Call:
lm(formula = data$murder.rate ~ data$poverty + data$high.school +
    data$college + data$single.parent + data$unemployed + data$metropolitan +
    data$region)
```

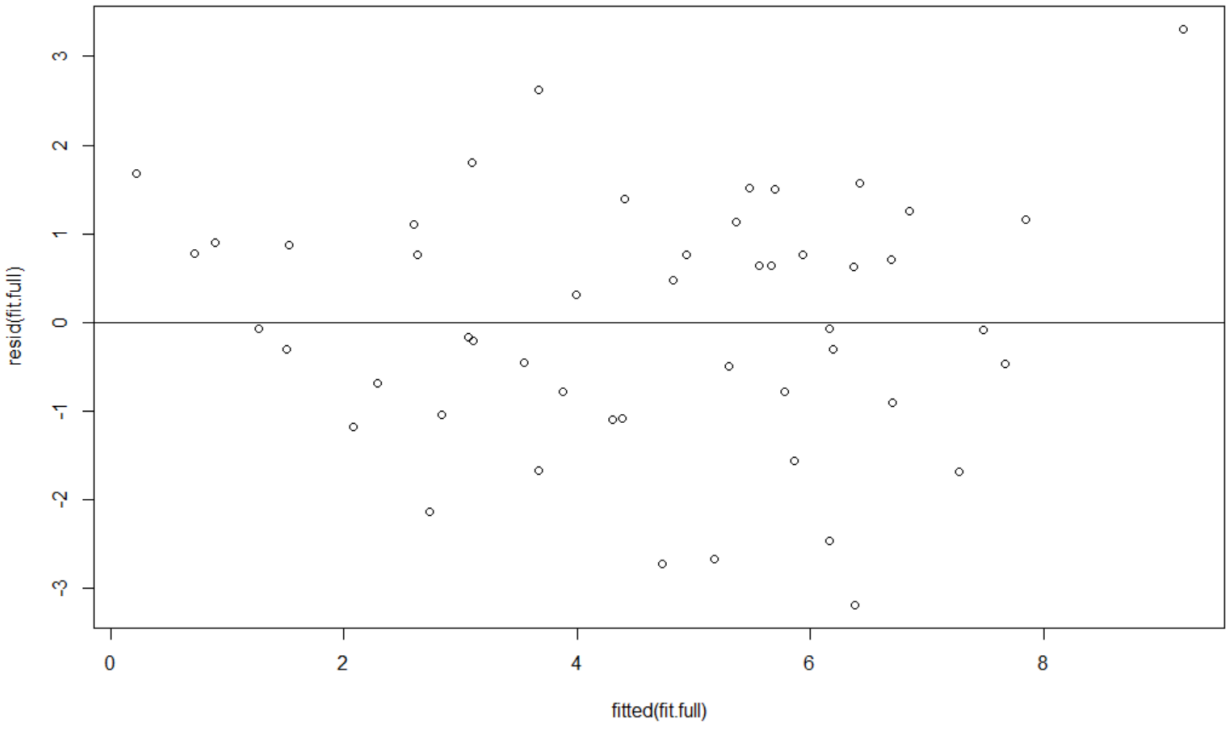
Results:

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.1861 -0.8706 -0.0709  0.8935  3.3049

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.15569   11.06682   0.104 0.917352
data$poverty        0.07124    0.12615   0.565 0.575397
data$high.school   -0.12534    0.11815  -1.061 0.295116
data$college        0.08368    0.08238   1.016 0.315857
data$single.parent  0.38015    0.10559   3.600 0.000867 ***
```

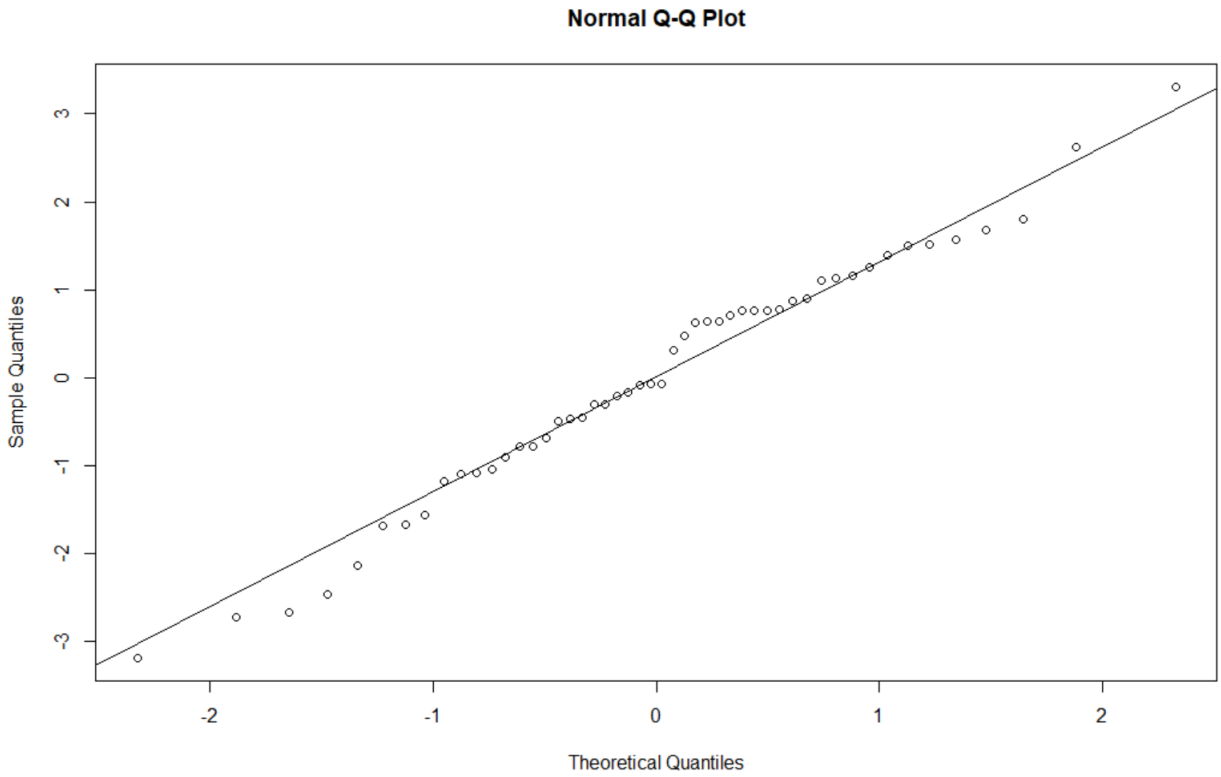
```
data$unemployed       0.29521    0.33119   0.891 0.378059
data$metropolitan     0.03095    0.01536   2.015 0.050607 .
data$regionNortheast -2.57007    0.76665  -3.352 0.001761 **
data$regionSouth     -0.12303    0.77605  -0.159 0.874832
data$regionWest      -0.83460    0.76033  -1.098 0.278904
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.549 on 40 degrees of freedom
Multiple R-squared:  0.6891,   Adjusted R-squared:  0.6192
F-statistic: 9.851 on 9 and 40 DF,  p-value: 9.287e-08
```

From the results obtained in the above function, the following inferences are made:

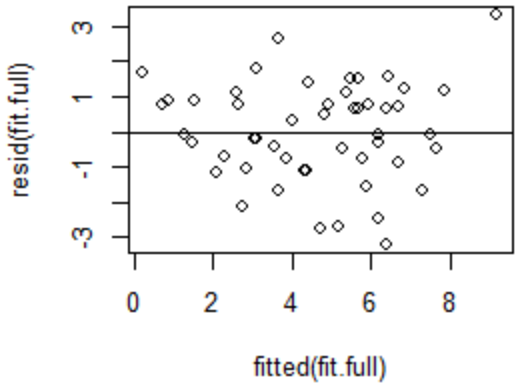
- The adjusted R-Squared is 0.6192, which means that 61.92% of the total variation is explained by the predictors in the model.
- Among all the provided list of predictors, **high school** and the **regions Northeast, South and West** have negative slope, which indicate that they have a linear relationship but it is **negative linear relationship**.
- Among all the predictors found, **Single.parent ,region Northeast and metropolitan** seem to be more significant compared to other predictors in the full model.
- Here , we observed that the fitted model represents the given data well as the p-value is close to zero(i.e.; 9.287e-08 ~approx 0).
- On plotting the fitted model against residuals we obtained the plot below and got the following inferences:
    - There do not seem to be any non-constant vertical scatter in the plot. However, is an outlier in the plot as can be observed below:

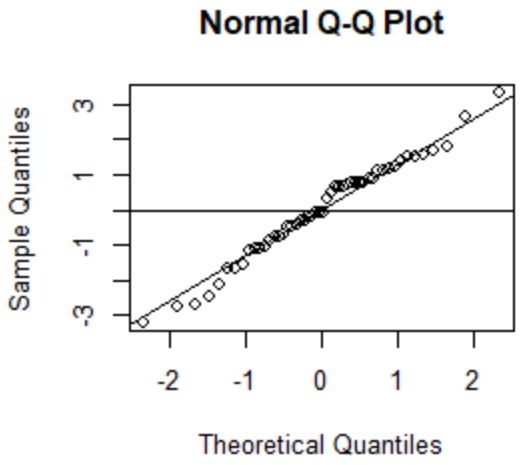


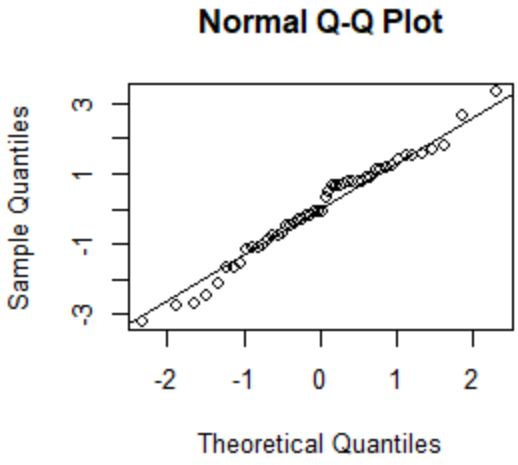
- Also when the normality assumption of the fitted full model is tested we observe that the normality assumption does not hold good as can be seen in the plot below (there is some curvature in the plotting of data points):

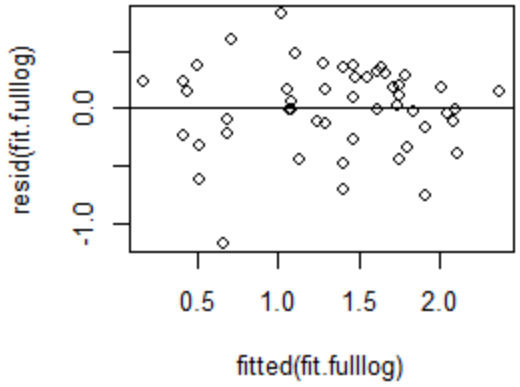


- Let us consider a transformation for the data and we applied log transform to the response murder rate and tested the normality assumption again.
- From the below plot which gives a side by side comparison of the residual plot and the qqplot before and also qqplot after applying the log transformation.









- The final full model that satisfies all our assumptions applicable to linear regression is below:

```
Call:
lm(formula = log(data$murder.rate) ~ data$poverty + data$high.school +
data$college +
    data$single.parent + data$unemployed + data$metropolitan + data$reg
ion)

Residuals:
     Min       1Q   Median       3Q      Max
-1.17760 -0.19898  0.04072  0.25614  0.82245

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.795903   2.952000   0.270  0.78884
poverty          -0.016157   0.033650  -0.480  0.63374
high.school      -0.033239   0.031517  -1.055  0.29791
college           0.013026   0.021975   0.593  0.55667
single.parent     0.092934   0.028166   3.300  0.00204 **
unemployed        0.112132   0.088342   1.269  0.21167
metropolitan      0.011839   0.004096   2.890  0.00619 **
regionNortheast -0.590295   0.204500  -2.887  0.00625 **
```

```
regionSouth      0.039990   0.207006   0.193  0.84779
regionWest      -0.112353   0.202812  -0.554  0.58268
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4132 on 40 degrees of freedom
Multiple R-squared:  0.6809,    Adjusted R-squared:  0.609
F-statistic: 9.482 on 9 and 40 DF,  p-value: 1.508e-0
```

- Now, after applying log transformation to the response murder rate, the Adjusted R-Squared has decreased slightly compared to the model without any transformation.
- Now, single parent, metropolitan and region Northeast are more significant predictors in this model compared to others.
- The fitted model with the log transformation may not represent the given data well as the p-value(1.508e-0) is not less the level significance assumed (we are assuming standard p-value as 0.05)

(b) Reduce your model by removing any unimportant variables (if such variables exist). Interpret the reduced model, including coefficients and r-squared. Perform a statistical test that compares the full model to the reduced model. Clearly state the hypotheses associated with this test and interpret the results.

Forward Selection:

- Let us now reduced the model by starting with just the intercept and add predictors one by one using **Forward selection method**. We built a model with each predictor to find the highest.
- Adjusted R-Squared and F-Statistic to add to the model one by one
- **DF-> Degrees of Freedom**

Predictor	**F-Statistic**	**Adjusted R-Squared**
data$Poverty	10.69(1 and 48 DF)	0.1651
data$HighSchool	27.32(1 and 48 DF)	0.3494
data$College	2.853(1 and 48 DF)	0.03644
data$SingleParent	40.69(1 and 48 DF)	0.4475
data$unemployed	6.751(1 and 48 DF)	0.105
data$Metropolitan	5.735(1 and 48 DF)	0.08812
data$Region	6.577(3 and 46 DF)	0.2546

Starting with Single Parent and iterating based on the significance of **Adjusted R-Squared and F Statistic**

Fit Name	**Predictor**	**F-Statistic**	**Adjusted R-Squared**
fit.1f	data$SingleParent	40.69(1 and 48 DF)	0.4475
fit.2F	data$SingleParent+ data$High School	25.63 (2 and 47 DF)	0.5013
fit.3f	data$SingleParent+ data$HighSchool+ data$Region	13.44 (5 and 44 DF)	0.5594
fit.4f	data$SingleParent+ data$HighSchool+	11.17(6 and 43 DF)	0.5545

	data$Region+ data$Poverty		

Observations after adding Poverty,

- The adjusted R-Squared value was lower that fit.3F.
- So we removed Poverty and proceeded with other predictors

Fit Name	Predictor	F-Statistic	Adjusted R-Squared
fit.5f	Single Parent+High School+Region+unemployed	10.96(6 and 43 DF)	0.5494

- After adding unemployed, the adjusted R-Squared value was lower that fit.3F.
- So we removed unemployed and proceeded with other predictors

Fit Name	Predictor	F-Statistic	Adjusted R-Squared
fit.6f	Single Parent+High School+Region+metropolitan	14.67(6 and 43 DF)	0.626
Fit.7f	Single Parent+High School+Region+metropolitan+college	12.69 (7 and 42 DF)	0.6255

- Finally, based on adjusted R-Squared value obtained with different predictors, **fit.6f** seems to have the highest value.
- So it explains the highest variation obtained in the model.

Partial F-Test

- Let us now compare this reduced model against the full model obtained in step 1 using anova functio.
- Our Hypothesis assumption are as follows:

  Null Hypothesis: No additional parameters in the full model that are not included in the reduced model are useful.
  Alternative Hypothesis: At least one parameter in the full model that is not included in the reduced model are useful

```
Analysis of Variance Table

Model 1: data$murder.rate ~ data$poverty + data$high.school + data$coll
ege + data$single.parent +
    data$unemployed + data$metropolitan + data$region
Model 2: data$murder.rate ~ data$single.parent + data$high.school + dat
a$region + data$metropolitan
  Res.Df     RSS Df Sum of Sq     F Pr(>F)
1     40  95.991
2     43 101.326 -3   -5.3346 0.741 0.5339
```

- The p-value obtained is 0.5339 which slightly greater than our assumed level of significa nce of 0.05.
- Hence we accept our null hypothesis as defined above.
  - Based on our null hypothesis that additional parameters included in the full mod el are not useful.

- We accept the null hypothesis that additional parameters are not necessary to u se in the model.

BACKWARD REDUCTION:

- We used the full model we fitted in section (a) (model that was not transformed using log transformation).
- To look at significant predictors, we used the summary function in R to check the p-value for each predictor.

```
Call:
lm(formula = Data$murder.rate ~ Data$poverty + Data$high.school + Data$col
lege + Data$single.parent + Data$unemployed + Data$metropolitan + Data$reg
ion)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1861 -0.8706 -0.0709  0.8935  3.3049

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.15569   11.06682   0.104 0.917352
poverty          0.07124    0.12615   0.565 0.575397
high.school     -0.12534    0.11815  -1.061 0.295116
college          0.08368    0.08238   1.016 0.315857
single.parent    0.38015    0.10559   3.600 0.000867 ***
unemployed       0.29521    0.33119   0.891 0.378059
metropolitan     0.03095    0.01536   2.015 0.050607 .
regionNortheast -2.57007    0.76665  -3.352 0.001761 **
regionSouth     -0.12303    0.77605  -0.159 0.874832
regionWest      -0.83460    0.76033  -1.098 0.278904
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.549 on 40 degrees of freedom
Multiple R-squared:  0.6891,    Adjusted R-squared:  0.6192
F-statistic: 9.851 on 9 and 40 DF,  p-value: 9.287e-08
```

- Now let us start with poverty which had the highest p-value in the predictors in the full model.
- It is observed that the reduced model had higher Adjusted R-squared

DF->degrees of freedom

Fit Name	Predictor	F-Statistic	Adjusted R-Squared	Residual Standard Error
Fit.full	Data$SingleParent+ Data$High School+ Data$Region+ Data$Poverty+ Data$metropolitan+ Data$Unemployed+ Data$college	9.851(9 and 40 DF)	0.6192	1.549 on 40 degrees of freedom
fit.1b	Data$SingleParent+ Data$HighSchool+ Data$Region+ Data$metropolitan+ Data$Unemployed+ Data$college	11.23(8 and 41 DF)	0.6255	1.536 on 41 degrees of freedom

```
Call:
lm(formula = murder.rate ~ high.school + college + single.parent +
    unemployed + metropolitan + region)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2309 -0.8284 -0.0797  0.8744  3.5847

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.56763    9.19493   0.497  0.62201
high.school     -0.15851    0.10167  -1.559  0.12664
college          0.08459    0.08168   1.036  0.30647
single.parent    0.39380    0.10193   3.863  0.00039 ***
unemployed       0.32348    0.32465   0.996  0.32490
metropolitan     0.02759    0.01404   1.965  0.05619 .
regionNortheast -2.60060    0.75837  -3.429  0.00139 **
regionSouth     -0.13982    0.76901  -0.182  0.85662
regionWest      -0.71785    0.72558  -0.989  0.32830
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.536 on 41 degrees of freedom
Multiple R-squared:  0.6866,    Adjusted R-squared:  0.6255
F-statistic: 11.23 on 8 and 41 DF,  p-value: 3.054e-08
```

- From the observed p-values obtained in the reduced model, we chose unemployed to b e the next predictor to be removed.
- The adjusted R-Squared remained the same as the previous model (fit.2b).
- **DF->degrees of freedom**

Fit Name	**Predictor**	**F-Statistic**	**Adjusted R-Squared**	**Residual Standard Error**
fit.2b	Data$Single Parent+ Data$High School+ Data$Region+ Data$metropolitan+ Data$college	12.69(7 and 42 DF)	0.6255	1.536 on 42 degrees of freedom

```
Call:
lm(formula = Data$murder.rate ~ Data$high.school + Data$college + Data$
single. Data$parent +
    Data$metropolitan + Data$region)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1314 -0.9019 -0.0047  0.7928  3.8741

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.35746    8.75751   0.840  0.40559
high.school     -0.18396    0.09840  -1.870  0.06853 .
college          0.07916    0.08149   0.971  0.33693
single.parent    0.43113    0.09479   4.548 4.54e-05 ***
metropolitan     0.02531    0.01385   1.828  0.07473 .
regionNortheast -2.62191    0.75800  -3.459  0.00126 **
regionSouth     -0.18462    0.76763  -0.241  0.81111
regionWest      -0.34028    0.61872  -0.550  0.58524
```

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

Residual standard error: 1.536 on 42 degrees of freedom
Multiple R-squared:  0.679,     Adjusted R-squared:  0.6255
F-statistic: 12.69 on 7 and 42 DF,  p-value: 1.285e-08
```

- Based on the above results, we used the reduced model fit.2b to check for least significant predictors based on the p-value, the next predictor we removed was college.
- **DF->Degrees of Freedom**

Fit Name	Predictor	F-Statistic	Adjusted R-Squared	Residual Standard Error
fit.3b	Data$Single Parent+ Data$High School+ Data$Region+ Data$metropolitan	14.67(6 and 43 DF)	0.626	1.535 on 43 degrees of freedom

```
Call:
lm(formula = Data$murder.rate ~ Data$high.school + Data$single.parent + 
Data$metropolitan + 
    Data$region)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.3575 -0.8339  0.1333  0.8812  3.9065 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      4.19460    8.12434   0.516  0.60829    
high.school     -0.12879    0.08030  -1.604  0.11607    
single.parent    0.42150    0.09420   4.474 5.54e-05 ***
metropolitan     0.03325    0.01118   2.974  0.00480 ** 
regionNortheast -2.34448    0.70168  -3.341  0.00173 ** 
regionSouth     -0.04464    0.75349  -0.059  0.95304    
regionWest      -0.30106    0.61699  -0.488  0.62806    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.535 on 43 degrees of freedom
Multiple R-squared:  0.6718,    Adjusted R-squared:  0.626
F-statistic: 14.67 on 6 and 43 DF,  p-value: 4.917e-09
```

- Here now we picked high school as the next predictor to be removed based on its significance compared to other predictors.
- **DF->Degrees of Freedom**

Fit Name	Predictor	F-Statistic	Adjusted R-Squared	Residual Standard Error
fit.4b	Data$Single Parent+ Data$Region+ Data$metropolitan	16.5(5 and 44 DF)	0.6127	1.562 on 44 degrees of freedom

From above observations we infer that,

- The Adjusted R-Squared value, reduced in comparison to the previous model and is also lesser than that of the full model.
- The other predictors in the model seem to be significant.
- Region predictor in this case is a categorical predictor and we tried removing it from fit.3b and found that the Adjusted R-squared decreased significantly.

Fit Name	**Predictor**	**F-Statistic**	**Adjusted R-Squared**	**Residual Standard Error**
fit.5b	Data$Single Parent+ Data$High School+ Data$metropolitan	19.46(3 and 46 DF)	0.5306	1.72 on 46 degrees of freedom

- The residual standard error also seems to be increasing starting from model fit.4b and fit.5b with high school and region predictors removed respectively from fit.3b model.
- Based on the results above, fit.3b with single parent, high school, region and metropolitan predictor seems to have the highest Adjusted R-Squared and the lowest residual standard error.

Partial F-Test

On comparing this reduced model against the full model used in the beginning using anova function in R with following Hypothesis assumption are as follows:

Null Hypothesis: No additional parameters in the full model that are not included in the reduced model are useful

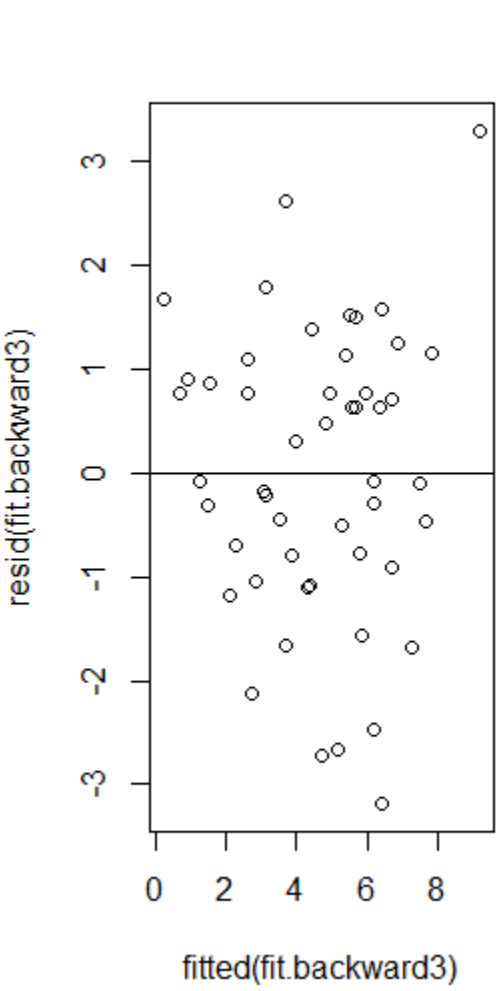
Alternative Hypothesis: At least one parameter in the full model that is not included in the reduced model are useful

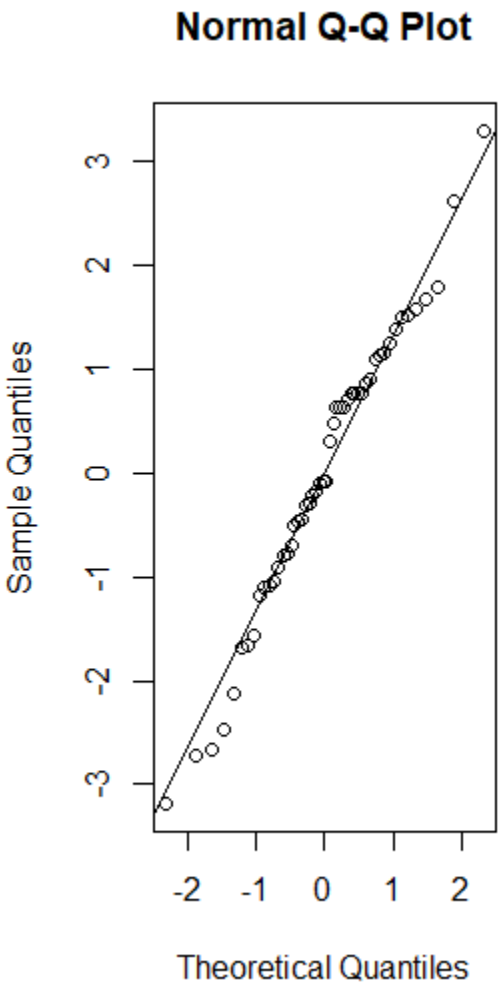
```
Analysis of Variance Table

Model 1: Data$murder.rate ~ Data$high.school + Data$single.parent + Dat
a$metropolitan + Data$region
Model 2: Data$murder.rate ~ Data$poverty + Data$high.school + Data$coll
ege + Data$single.parent +
    Data$unemployed + Data$metropolitan + Data$region
  Res.Df     RSS Df Sum of Sq     F Pr(>F)
1     43 101.326
2     40  95.991  3    5.3346 0.741 0.5339
```

- The p-value is 0.5339 which greater than our assumed standard level of significance of 0 .05.
- So we now accept null hypothesis.
- Based on our null hypothesis that additional parameters included in the full model are n ot useful.
- We accept the null hypothesis that additional parameters are not necessary to use in th e model.

Conclusion

- On analysis using both **forward selection and backward reduction**, we found that single parent, high school, region and metropolitan predictors seem to have the highest Adjusted R-Squared and the lowest residual standard error.
- The below plot is the plot of reduced fitted model against the residuals
- At the same time, we also tested the normality assumption of the residuals and observed that there do not seem to be any non-constant vertical scatter in the plot.
- However, there is an outlier in the plot as it can be observed below in the plot.
- The QQ plot has a slight curvature but the normality assumption seems to hold true.





- Used AIC stepwise selection using forward, backward and both methods available in the MASS package in R.
- Also, tabulated the results obtained using stepAIC function

Direction	Predictors in the model suggested by StepAIC function
Forward	poverty, high.school, college , single.parent, unemployed, metropolitan, region(basically the full model is suggested)

Backward	high.school, single.parent, , metropolitan, region(basically the reduced model that we obtained in section (b) is suggested)
Both	high.school, single.parent, , metropolitan, region(basically the reduced model that we obtained in section (b) is suggested)

Code Used using STEPAIC function

```
library(MASS)
#Forward selection
prog.lm.forstep=stepAIC(fit.full, scope=list(lower=~1,upper=~data$poverty+ data$high.school+
data$college+ data$single.parent+ data$unemployed+ data$metropolitan+
data$region),direction="forward")
#Backward selection same as (fit.full,trace=0) as the scope argument is missing
prog.lm.backstep=stepAIC(fit.full, scope=list(lower=~1,upper=~ data$poverty+
data$high.school+ data$college+ data$single.parent+ data$unemployed+ data$metropolitan+
data$region),direction="backward")
#Both
prog.lm.bothstep=stepAIC(fit.full, scope=list(lower=~1,upper=~ data$poverty+
data$high.school+ data$college+ data$single.parent+ data$unemployed+ data$metropolitan+
data$region),direction="both")
```

(c) **Use your final model to predict murder rate of a state whose predictor values are set at the average in the data for a quantitative predictor and the most frequent category for a qualitative predictor.**

- Using table function in R we found that the most frequent category for the qualitative predictor region is South.
- The model we are using is *MR_prediction=4.19460-(0.12879* data$high.school predictor value)+ (0.42150* data$single.parent predictor value)+(0.03325* data$metropolitan predictor value)-(2.34448* data$Region North East Indicator value)-(0.04464* data$Region South Indicator Value)-(-0.30106* data$Region West Indicator value)*
- The Murder Rate prediction for a state based on average values for all quantitative predictors above and for Region south is 5.074478

Section 2

R- Code:

```
data= read.csv(file="C:/Users/shash/Documents/R for Stats/Mini Projects/Miniproject 6/crime.csv")
#read data from csv file

data #preview data

str(data) #data stats

#'data.frame':    50 obs. of  9 variables:

# $ state       : Factor w/ 50 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8 9 10 ...

# $ murder.rate  : num  7.4 4.3 7 6.3 6.1 3.1 2.9 3.2 5.6 8 ...

# $ poverty      : num  14.7 8.4 13.5 15.8 14 8.5 7.7 9.9 12 12.5 ...

# $ high.school  : num  77.5 90.4 85.1 81.7 81.2 89.7 88.2 86.1 84 82.6 ...

# $ college      : num  20.4 28.1 24.6 18.4 27.5 34.6 31.6 24 22.8 23.1 ...

# $ single.parent: num  26 23.2 23.5 24.7 21.8 20.8 22.9 25.6 26.5 25.5 ...

# $ unemployed   : num  4.6 6.6 3.9 4.4 4.9 2.7 2.3 4 3.6 3.7 ...

# $ metropolitan : num  70.2 41.6 87.9 49 96.7 84 95.6 81.4 93 69.1 ...

# $ region       : Factor w/ 4 levels "North Central",..: 3 4 4 3 4 4 2 3 3 3 ...

#attaching the variables in R memory

attach(data)

#Using pairs to understand relationship between murder rate and the predictors

pairs(data[2:8])

fit.modelfull<-
lm(data$murder.rate~data$poverty+data$high.school+data$college+data$single.parent+data$unemplo
yed+data$metropolitan+data$region)

summary(fit.modelfull)

#Fitting murder rate with poverty predictor

fitm.p <- lm(data$murder.rate ~ data$poverty)

summary(fitm.p)
```

```
#Fitting murder rate with high school predictor
fitm.hs <- lm(data$murder.rate ~ data$high.school)
summary(fitm.hs)
#Fitting murder rate with college predictor
fitm.c <- lm(data$murder.rate ~ data$college)
summary(fitm.c)
#Fitting murder rate with single parent predictor
fimt.sp <- lm(data$murder.rate ~ data$single.parent)
summary(fitm.sp)
#Fitting murder rate with unemployed predictor
fitm.u <- lm(data$murder.rate ~ data$unemployed)
summary(fitm.u)
#Fitting murder rate with metropolitan predictor
fitm.m <- lm(data$murder.rate ~ data$metropolitan)
summary(fitm.m)
#box plot to analyze region against murder rate
plot(data$murder.rate ~ data$region)
#checking most prequent qualitative predictor
table(data$region)
#Fitting murder rate with region predictor
fitm.r <- lm(data$murder.rate ~ data$region)
summary(fitm.r)
#full model with all predictors
fit.modelfull<-
lm(data$murder.rate~data$poverty+data$high.school+data$college+data$single.parent+data$unemplo
yed+data$metropolitan+data$region)
anova(fit.modelfull)
summary(fit.modelfull)
```

```
#Comparing qqnorm plots before and after transformation
par(mfrow=c(2,2))
#Residual plot of the full model without tranformation
plot(fitted(fit.modelfull),resid(fit.modelfull))
abline(h=0)
#QQplot of residuals of the full model without tranformation
qqnorm(resid(fit.modelfull))
qqline(resid(fit.modelfull))
#applying log transformation and checking the residual plot and testing the normality assumption
fit.trans.log<-update(fit.modelfull,log(data$murder.rate) ~ .)
#Residual plot of the full model with log tranformation
plot(fitted(fit.trans.log),resid(fit.trans.log))
abline(h=0)
#QQplot of residuals of the full model with log tranformation
qqnorm(resid(fit.trans.log))
qqline(resid(fit.trans.log))
#checking the transformed full model
summary(fit.trans.log)
#Forward Method of getting a reduced Model
#First model with just single parent
fit.Forward1<-lm(data$murder.rate~ data$single.parent)
summary(fit.Forward1)
#adding high school and single parent
fit.Forward2<-lm(data$murder.rate~data$single.parent+data$high.school)
summary(fit.Forward2)
#adding high school, single parent and Region
fit.Forward3<-lm(data$murder.rate~data$single.parent+data$high.school+data$region)
summary(fit.Forward3)
#adding high school, single parent,Region and Poverty
```

```
fit.Forward4<-lm(data$murder.rate~data$single.parent+data$high.school+data$region+data$poverty)

summary(fit.Forward4)

#adding high school, single parent,Region and unemployed

fit.Forward5<-
lm(data$murder.rate~data$single.parent+data$high.school+data$region+data$unemployed)

summary(fit.Forward5)

#adding high school, single parent,Region and metropolitan

fit.Forward6<-
lm(data$murder.rate~data$single.parent+data$high.school+data$region+data$metropolitan)

summary(fit.Forward6)

#adding high school, single parent,Region,Poverty and college

fit.Forward7<-
lm(data$murder.rate~data$single.parent+data$high.school+data$region+data$metropolitan+data$colle
ge)

summary(fit.Forward7)

#comparing full model against reduced model with high school, single parent,Region and metropolitan

anova(fit.modelfull,fit.Forward6)

#Backward Method of getting a reduced Model

#removing poverty from full model

fit.backward1<-update(fit.modelfull, . ~ . - data$poverty)

summary(fit.backward1)

#removing unemployed from fit.backward1

fit.backward2<-update(fit.backward1, . ~ . - data$ unemployed)

summary(fit.backward2)

#removing college from fit.backward2

fit.backward3<-update(fit.backward2, . ~ . - data$college)

summary(fit.backward3)

#removing high school from fit.backward3

fit.backward4<-update(fit.backward3, . ~ . - data$high.school)

summary(fit.backward4)
```

```
#removing region from fit.backward3

fit.backward5<-update(fit.backward3, . ~ . - data$region)

summary(fit.backward5)

#ANOVA analysis between full and reduced model being used forward - fit.backward3

anova(fit.backward3,fit.modelfull)

par(mfrow=c(1,2))

# Residual plot and checking for normality assumption for the reduced model

plot(fitted(fit.backward3),resid(fit.backward3))

abline(h=0)

#QQplot

qqnorm(resid(fit.backward3))

qqline(resid(fit.backward3))

#Using StepAIC function to verify and compare the reduced model obtained above

library(MASS) #using R package MASS

#Forward selection

fit.lm.forwardstep=stepAIC(fit.modelfull,
scope=list(lower=~1,upper=~data$poverty+data$high.school+data$college+data$single.parent+data$un
employed+data$metropolitan+data$region),direction="forward")

#Backward selection same as (fit.modelfull,trace=0) as the scope argument is missing

fit.lm.backwardstep=stepAIC(fit.modelfull,
scope=list(lower=~1,upper=~data$poverty+data$high.school+data$college+data$single.parent+data$un
employed+data$metropolitan+data$region),direction="backward")

#Both

fit.lm.both=stepAIC(fit.modelfull,
scope=list(lower=~1,upper=~data$poverty+data$high.school+data$college+data$single.parent+data$un
employed+data$metropolitan+data$region),direction="both")

summary(fit.backward3)

#Prediction using the model

prediction_MRmodel=4.19460-(0.12879*mean(data$high.school))+
(0.42150*mean(data$single.parent))+(0.03325*mean(data$metropolitan))-(2.34448*0)-(0.04464*1)-(-
0.30106*0)

prediction_MRmodel
```