# Statistical Methods for Data Science

## Mini Project 6

***Members:***

| | |
|---|---|
| SHIVA RANGA CHAWALA | sxc167630 |
| KRISHNA SINDHU KOTA | kxk171030 |

***Contribution:***

Shiva Ranga Chawala – Equally contributed

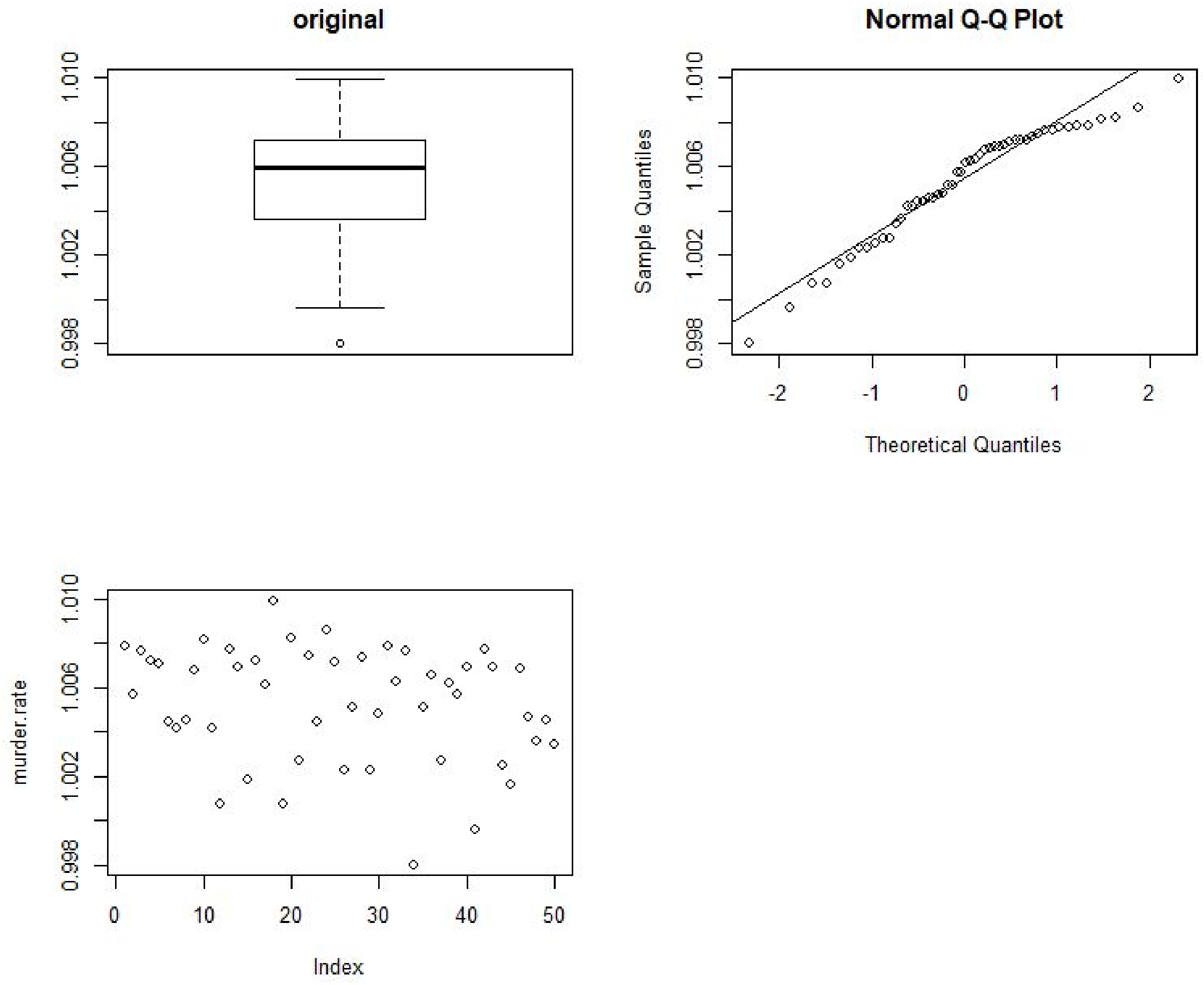
Krishna Sindhu Kota – Equally contributed

1. **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.**

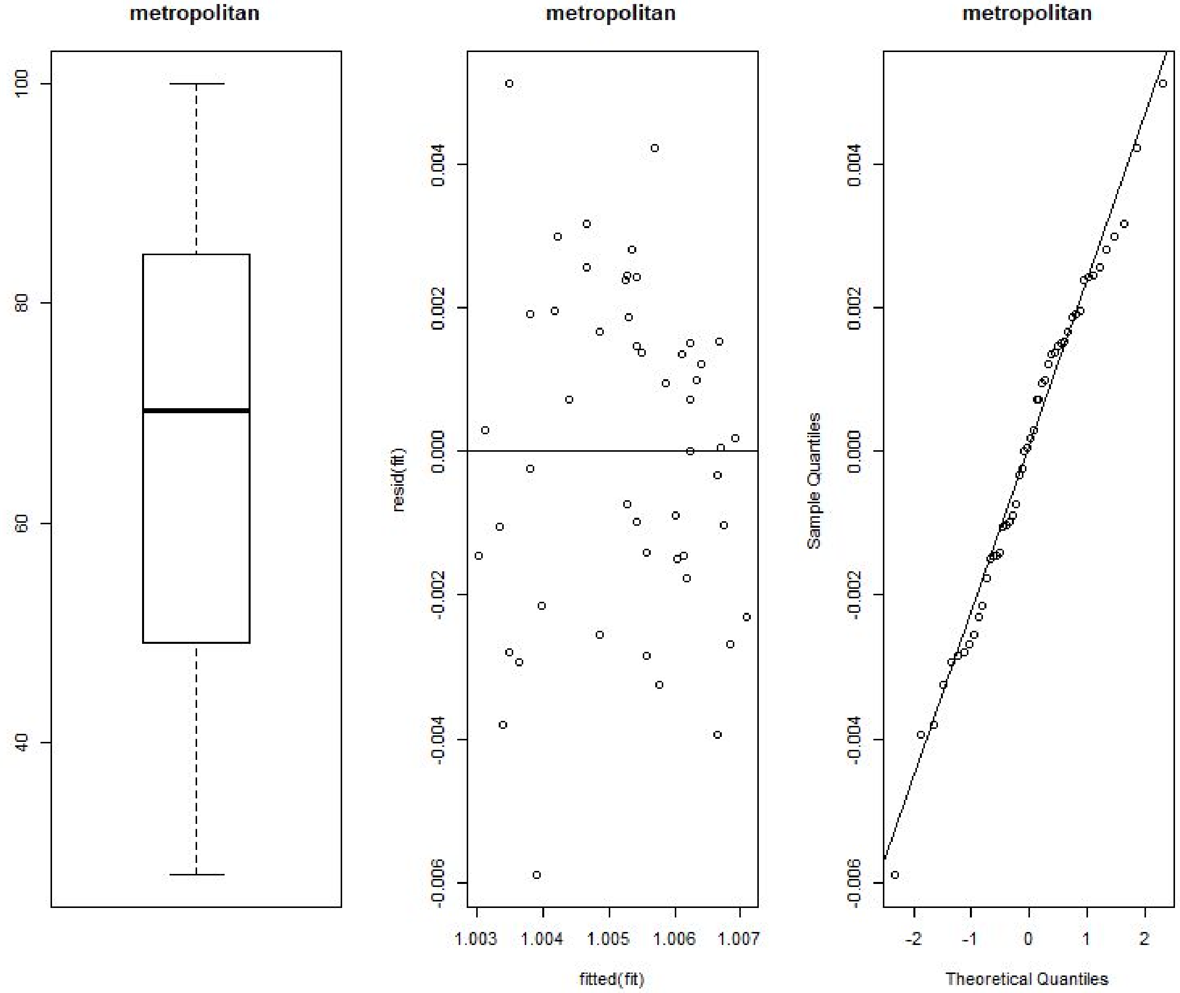
**Code:**

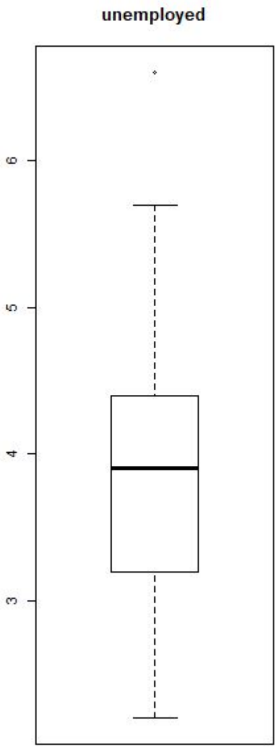
```
crime = read.csv("crime.csv", sep=",", header=T)
#split crime dataset based on the values of region
c = split(crime, crime$region)
str(crime)
attach(crime)
region = factor(region)
par(mfrow= c(2,2))
table(region)

region
North Central     Northeast         South          West
       12             9                  16            13
```

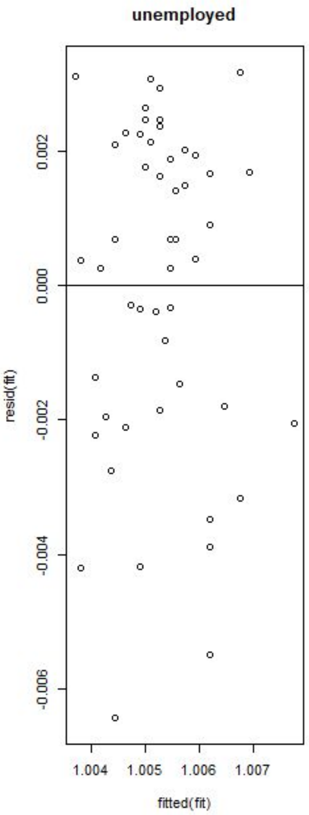


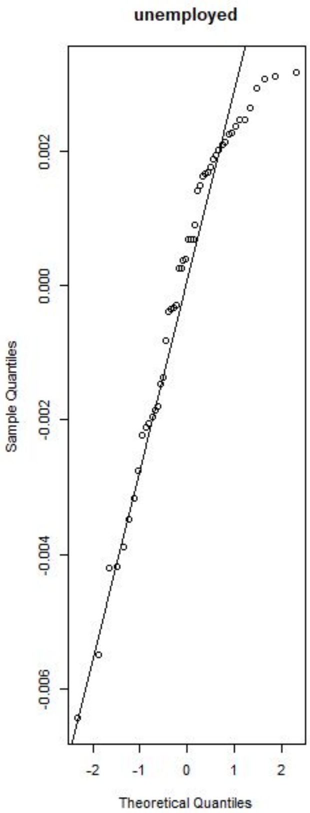
```
boxplot(murder.rate, main = c("original"))
qqnorm(murder.rate)
qqline(murder.rate)
plot(murder.rate)

par(mfrow = c(1,3))
fit_fun <- function(crime_feature, fit, name){
  #boxplot
  boxplot(crime_feature, main = name)
  plot(fitted(fit),resid(fit), main = name)
  abline(h=0)
  #qqplot
  qqnorm(resid(fit), main = name)
  qqline(resid(fit))
}
```

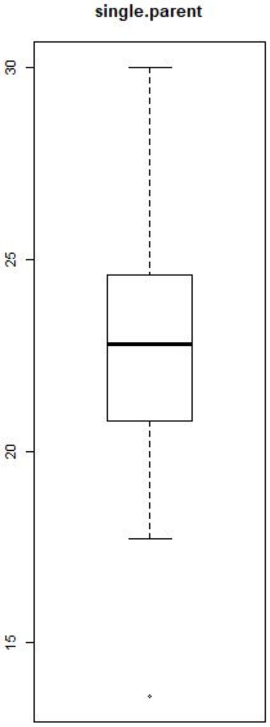
```
fit_fun(metropolitan, lm(murder.rate ~ metropolitan), c("metropolitan"))
fit_fun(unemployed, lm(murder.rate ~ unemployed) , c("unemployed"))
fit_fun((single.parent), lm(murder.rate ~ single.parent) , c("single.parent"))
fit_fun(college, lm(murder.rate ~ college), c("college"))
fit_fun(high.school, lm(murder.rate ~ high.school) , c("high.school"))
fit_fun(poverty, lm(murder.rate ~ poverty), c("poverty"))
# now they all look approximately normal with no outliers
```

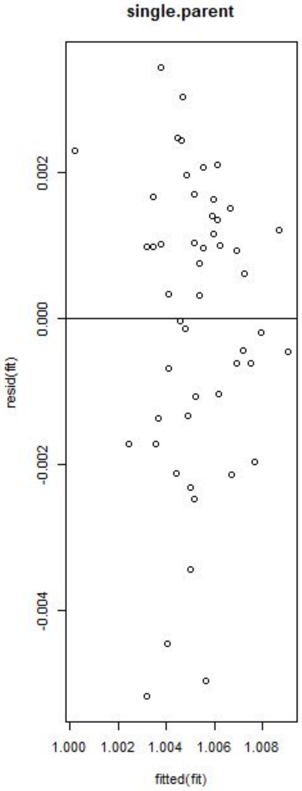


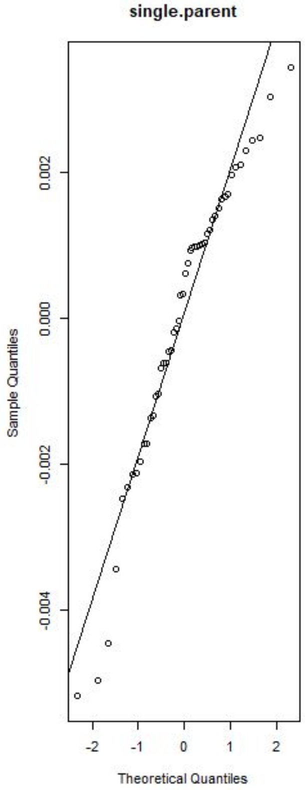

unemployed
3
4
5
6

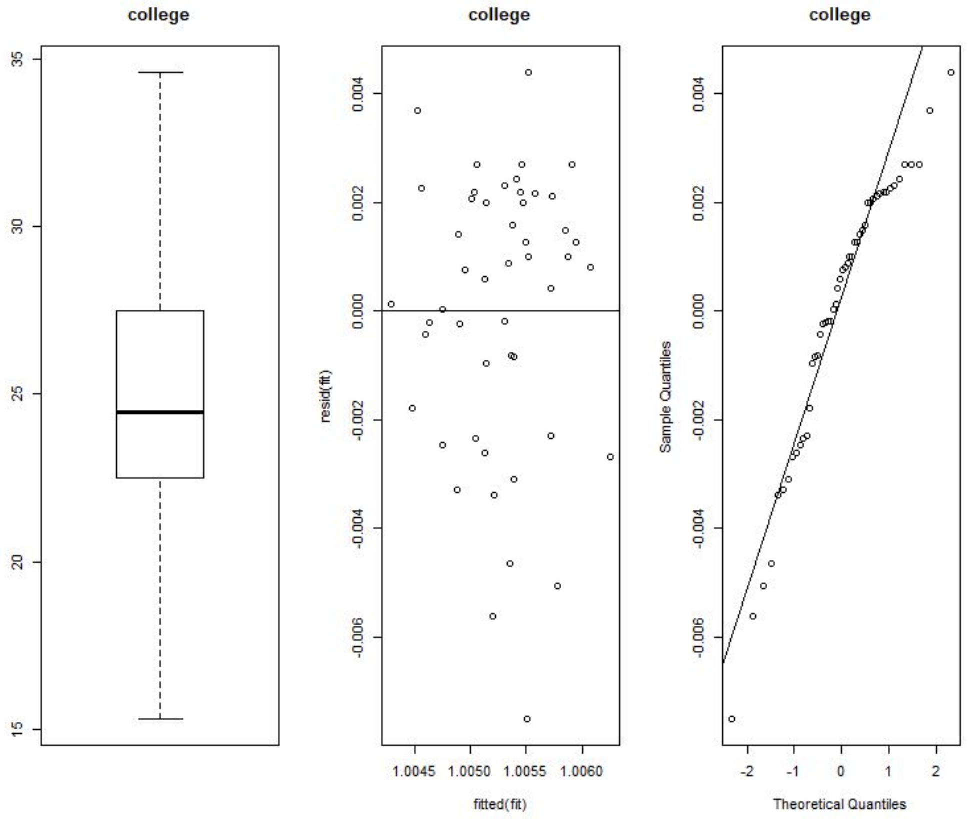

unemployed
resid(fit)
-0.006
-0.004
-0.002
0.000
0.002
1.004
1.005
1.006
1.007
fitted(fit)

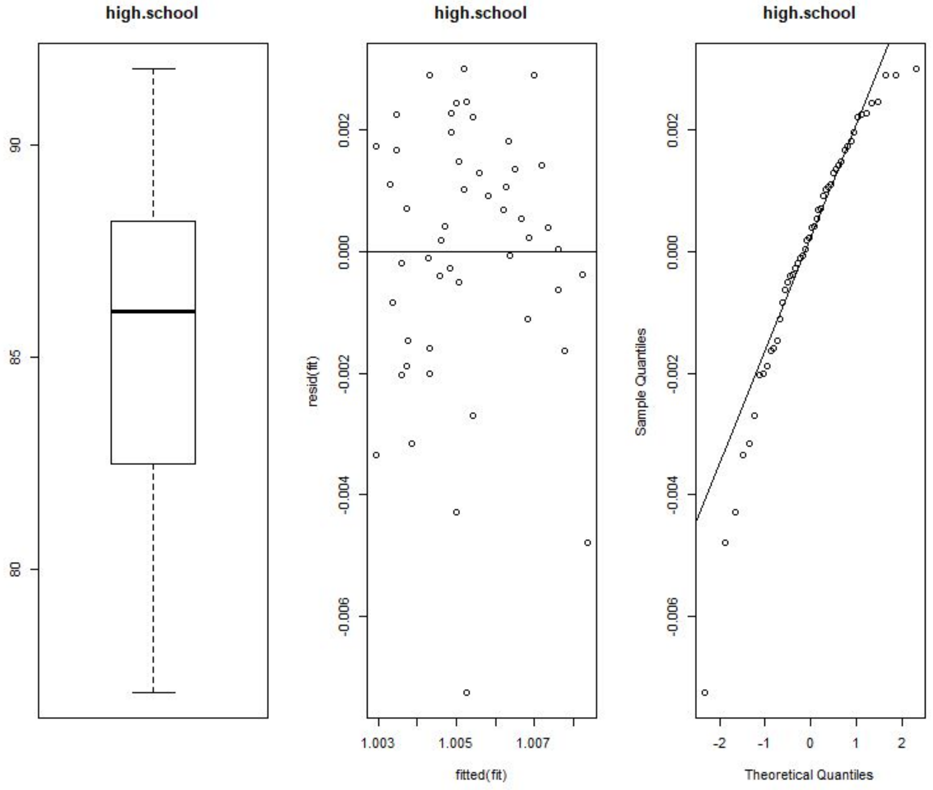

unemployed
Sample Quantiles
-0.006
-0.004
-0.002
0.000
0.002
-2
-1
0
1
2
Theoretical Quantiles

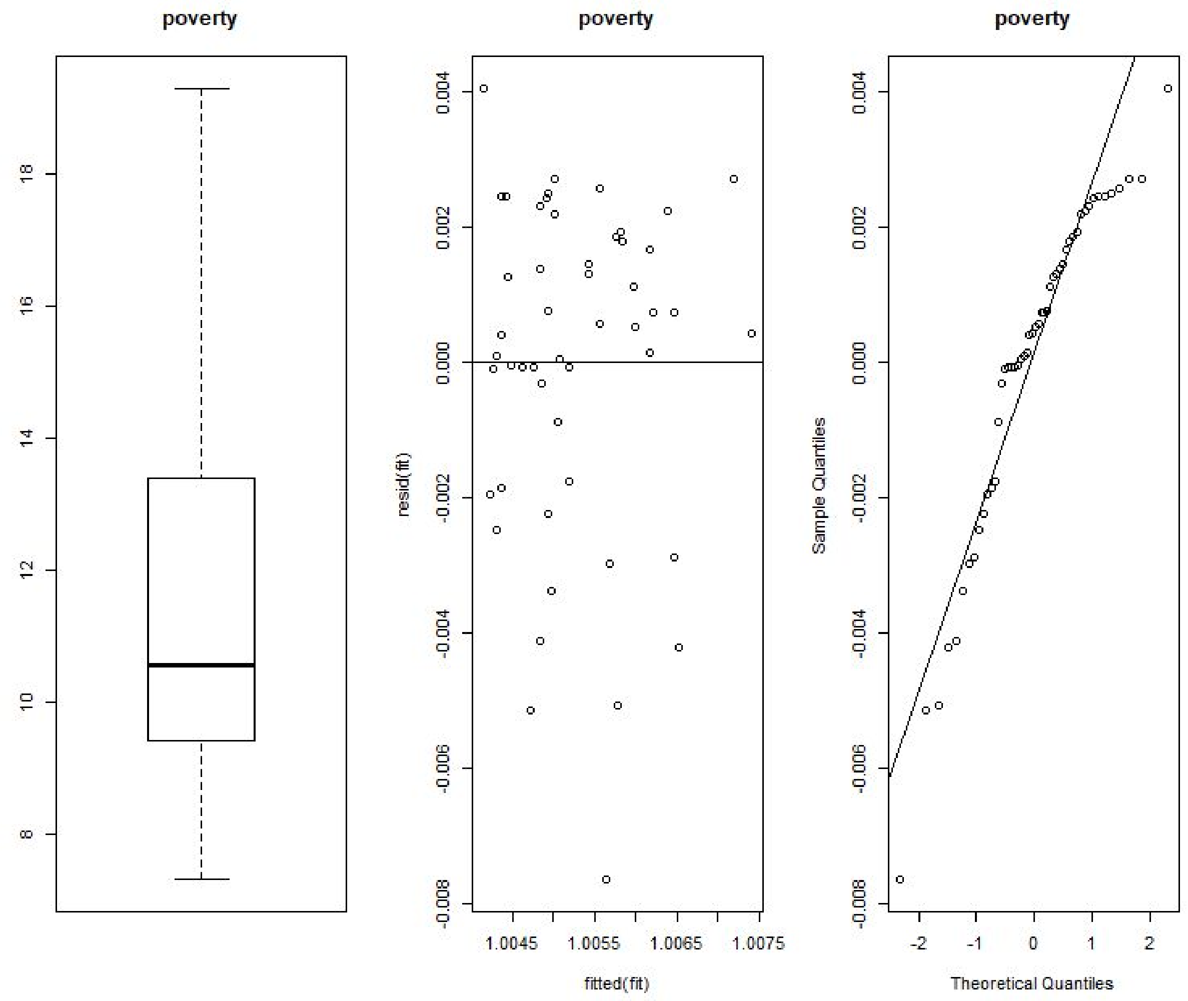

single.parent
15
20
25
30


single.parent
resid(fit)
-0.004
-0.002
0.000
0.002
1.000
1.002
1.004
1.006
1.008
fitted(fit)


single.parent
Sample Quantiles
-0.004
-0.002
0.000
0.002
-2
-1
0
1
2
Theoretical Quantiles


college
35
30
25
20
15
college
resid(fit)
0.004
0.002
0.000
-0.002
-0.004
-0.006
1.0045
1.0050
1.0055
1.0060
fitted(fit)
college
Sample Quantiles
0.004
0.002
0.000
-0.002
-0.004
-0.006
-2
-1
0
1
2
Theoretical Quantiles


high.school
90
85
80
high.school
resid(fit)
0.002
0.000
-0.002
-0.004
-0.006
1.003
1.005
1.007
fitted(fit)
high.school
Sample Quantiles
0.002
0.000
-0.002
-0.004
-0.006
-2
-1
0
1
2
Theoretical Quantiles



```
all_fit = lm(murder.rate ~ poverty + high.school + college + single.parent + unemployed + metropolitan + region)
summary(all_fit)
```

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

Residuals:
      Min        1Q     Median        3Q       Max
-0.0092281 -0.0015734  0.0003115  0.0020175  0.0064772

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)      1.006e+00  2.324e-02  43.301  < 2e-16 ***
poverty        -1.253e-04  2.649e-04  -0.473  0.63873
high.school     -2.627e-04  2.481e-04  -1.059  0.29605
college         1.033e-04  1.730e-04   0.597  0.55394
single.parent    7.332e-04  2.217e-04   3.307  0.00200 **
unemployed       8.805e-04  6.955e-04   1.266  0.21283
metropolitan     9.309e-05  3.225e-05   2.887  0.00625 **
regionNortheast -4.664e-03  1.610e-03  -2.897  0.00608 **
regionSouth      3.062e-04  1.630e-03   0.188  0.85190
regionWest      -8.947e-04  1.597e-03  -0.560  0.57836
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.003253 on 40 degrees of freedom
Multiple R-squared:  0.6815,  Adjusted R-squared:  0.6098
F-statistic: 9.509 on 9 and 40 DF,  p-value: 1.454e-07
```

**Interpretation:**

From the table of region, we can observe that there are 12 North Central values, 9 North East, 13 West and 16 South which is the highest. In the boxplot of murder rate, it is right skewed, median value is 1.006 and there is one outlier. From the QQ plot,we can observe that it is approximately normal. As the points in a residual plot are randomly dispersed around the horizontal axis, we can see that, linear regression model is appropriate for the data. From the individual attributes plots, we can observe that all look approximately normal. The p values of poverty, high school, college, unemployed, region west, region south are higher than 0.05 and single parent, metropolitan, region North East values are less than 0.05. H0 → slopes of all predictors = 0 and H1→ slope of at least one predictor is not 0, so for the values less than 0.05, we reject the null hypothesis and for values greater than 0.05, we accept the null hypothesis. As we can see that the murder rate data is approximately normal, hence no transformation is required. The 5 point summary values are (-0.0092281 -0.0015734  0.0003115  0.0020175  0.0064772 ). The values of adjusted R-squared is 0.6098.

***(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.**

**Code:**

```
fit_1 = lm(murder.rate ~ single.parent)
summary(fit_1)
Call:
lm(formula = murder.rate ~ single.parent)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.8124 -1.2821  0.2066  1.2319  4.3544 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -8.24775    2.03206  -4.059 0.000181 *
single.parent  0.55950    0.08772   6.379 6.61e-08 *
---
Signif. codes:  0 ‘*’ 0.001 ‘*’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.866 on 48 degrees of freedom
Multiple R-squared:  0.4588,  Adjusted R-squared:  0.4475 
F-statistic: 40.69 on 1 and 48 DF,  p-value: 6.605e-08
```

Since the p-value (6.61e-08) < 0.05, we can reject null hypothesis which means, single.parent is a significant factor.

```
fit_2 = update(fit_1, . ~ . + metropolitan)
summary(fit_2)

Call:
lm(formula = murder.rate ~ single.parent + metropolitan)

Residuals:
   Min     1Q Median     3Q    Max 
-3.359 -1.208  0.192  1.271  4.340 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -9.40288    2.03250  -4.626 2.94e-05 *
single.parent  0.52965    0.08574   6.178 1.45e-07 *
metropolitan   0.02718    0.01267   2.145   0.0371 *
---
```

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

Residual standard error: 1.8 on 47 degrees of freedom
Multiple R-squared:  0.507,    Adjusted R-squared:  0.486
F-statistic: 24.17 on 2 and 47 DF,  p-value: 6.046e-08
```

```
anova(fit_1, fit_2)
```

```
Analysis of Variance Table

Model 1: murder.rate ~ single.parent
Model 2: murder.rate ~ single.parent + metropolitan
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     48 167.11
2     47 152.21  1    14.901 4.6012 0.03715 *
---
Signif. codes:  0 ‘*’ 0.001 ‘*’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Since the p-value (0.0371) < 0.05, we can reject null hypothesis i.e., metropolitan is a significant factor. That is, metropolitan is a useful feature.

```
fit_3 = update(fit_2, . ~ . + poverty)
summary(fit_3)
```

```
Call:
lm(formula = murder.rate ~ single.parent + metropolitan + poverty)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1452 -1.1775  0.0954  1.2079  3.4722

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -10.09487    1.97907  -5.101 6.26e-06 *
single.parent   0.42335    0.09551   4.433 5.73e-05 *
metropolitan    0.03633    0.01287   2.823  0.00701 **
poverty         0.21973    0.09974   2.203  0.156352 *
---
Signif. codes:  0 ‘*’ 0.001 ‘*’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 1.73 on 46 degrees of freedom
Multiple R-squared:  0.5541,  Adjusted R-squared:  0.525
F-statistic: 19.05 on 3 and 46 DF,  p-value: 3.567e-08
```

```
anova(fit_2, fit_3)
Analysis of Variance Table

Model 1: murder.rate ~ single.parent + metropolitan
Model 2: murder.rate ~ single.parent + metropolitan + poverty
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     47 152.21
2     46 137.69  1    14.527 4.8533 0.156352 *
---
Signif. codes:  0 ‘*’ 0.001 ‘*’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Since the p-value (0.1563) > we can accept null hypothesis i.e., poverty is not a significant factor. That is, poverty is not a useful feature.

```
fit_4 = update(fit_2, . ~ . + unemployed)
summary(fit_4)
Call:
lm(formula = murder.rate ~ single.parent + metropolitan + unemployed)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2971 -1.1556 -0.1041  1.2070  3.9212

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -10.18840    2.06007  -4.946 1.05e-05 *
single.parent   0.47604    0.09084   5.240 3.91e-06 *
metropolitan    0.03046    0.01264   2.410    0.020 *
unemployed      0.45979    0.28880   1.592    0.118
---Signif. codes:  0 ‘*’ 0.001 ‘*’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.771 on 46 degrees of freedom
Multiple R-squared:  0.5328,  Adjusted R-squared:  0.5023
F-statistic: 17.48 on 3 and 46 DF,  p-value: 1.025e-07
```

```
anova(fit_2, fit_4)
Analysis of Variance Table

Model 1: murder.rate ~ single.parent + metropolitan
Model 2: murder.rate ~ single.parent + metropolitan + unemployed
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     47 152.21
2     46 144.26  1    7.9494 2.5348 0.1182
```

#since p-value(0.118) > 0.05, we cannot reject null hypothesis, which means unemployed is not a significant factor.

```
table(region)
region
North Central     Northeast         South          West
          12             9                16            13

fit_5 = update(fit_2, . ~ . + region)
summary(fit_5)
Call:
lm(formula = murder.rate ~ single.parent + metropolitan + region)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9730 -1.0828  0.1990  0.9294  3.7955

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -8.44469    2.01034  -4.201 0.000128 *
single.parent    0.47472    0.08973   5.291 3.67e-06 *
metropolitan     0.03627    0.01122   3.234 0.002317 **
regionNortheast -2.29258    0.71334  -3.214 0.002453 **
regionSouth      0.51237    0.68052   0.753 0.455510
regionWest      -0.24384    0.62687  -0.389 0.699165
---
Signif. codes:  0 ‘*’ 0.001 ‘*’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.562 on 44 degrees of freedom
```

```
Multiple R-squared:  0.6522,  Adjusted R-squared:  0.6127
F-statistic:  16.5 on 5 and 44 DF,  p-value: 3.731e-09
```

#here region NorthCentral is the baseline

```
anova(fit_2, fit_5)
Analysis of Variance Table

Model 1: murder.rate ~ single.parent + metropolitan
Model 2: murder.rate ~ single.parent + metropolitan + region
  Res.Df    RSS Df Sum of Sq     F   Pr(>F)
1     47 152.21
2     44 107.39  3    44.824 6.122 0.001425 **
---
Signif. codes:  0 ‘*’ 0.001 ‘*’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#null hypothesis: slopes all of the regions = 0 assuming remaining variables remain the same i.e., region is not a significant factor

#since p-value(0.001425) < 0.05, we reject null hypothesis, which means there is atleast one region whose slope is not 0, so region is a significant factor.

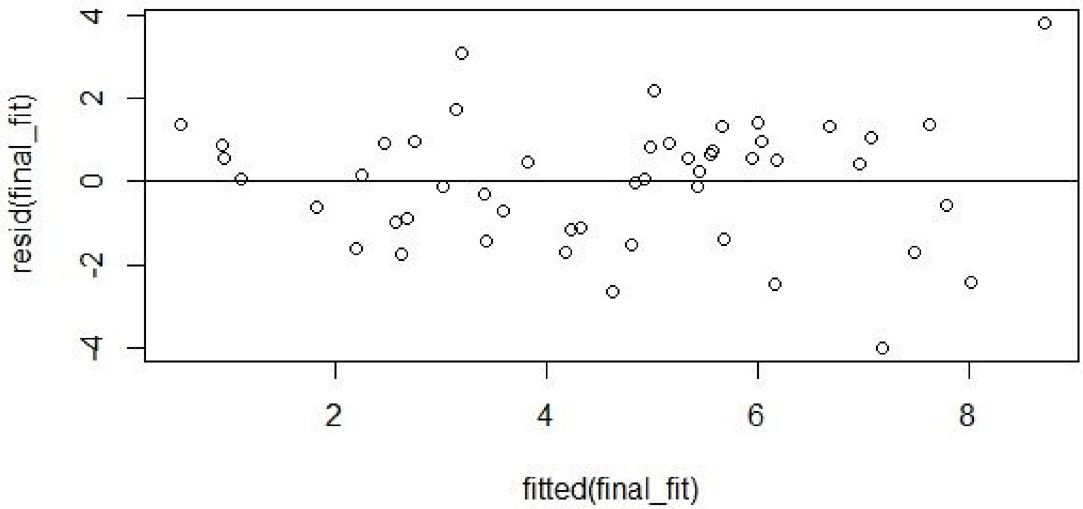
```
final_fit = fit_5
summary(final_fit)
Call:
lm(formula = murder.rate ~ single.parent + metropolitan + region)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9730 -1.0828  0.1990  0.9294  3.7955

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -8.44469    2.01034  -4.201 0.000128 *
single.parent    0.47472    0.08973   5.291 3.67e-06 *
metropolitan     0.03627    0.01122   3.234 0.002317 **
regionNortheast -2.29258    0.71334  -3.214 0.002453 **
regionSouth      0.51237    0.68052   0.753 0.455510
```

```
regionWest      -0.24384    0.62687  -0.389 0.699165
---
Signif. codes:  0 ‘*’ 0.001 ‘*’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.562 on 44 degrees of freedom
Multiple R-squared:  0.6522,  Adjusted R-squared:  0.6127
F-statistic:  16.5 on 5 and 44 DF,  p-value: 3.731e-09
```

From the above table we can see that intercepts $\beta$0=-8.44469 $\beta$1=0.47472 $\beta$2=0.03627 $\beta$3=-2.2925 $\beta$4=0.512 $\beta$5=-0.2438 and their standard errors in the next column.
We can also see that value of adjusted R-squared has increased from 0.6098(for full fit) to 0.6127(for reduced model) from which we can say that our reduced model is better that the full model.
The Residual Standard Error which is 1.562 states how far the observed murder.rate are from the predicted or fitted murder.rate value.
F-Statistic which is the ratio of the mean regression sum of squares divided by the mean error sum of squares tells us that the group of variables are jointly significant.

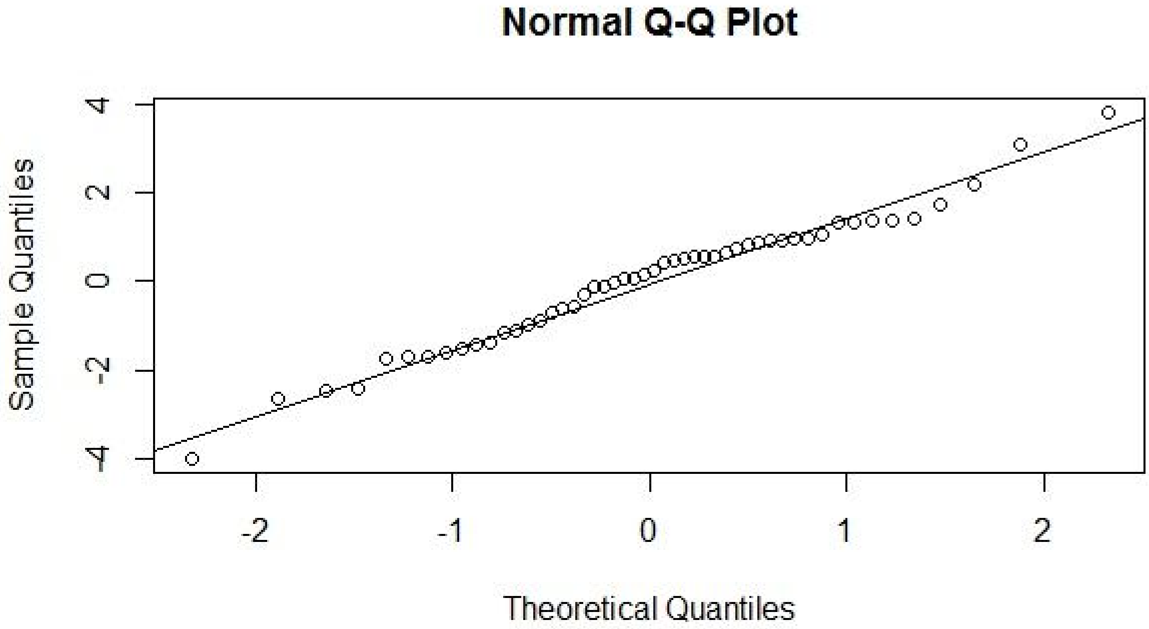
#REDUCED MODEL (final model): murder.rate ~ metropolitan + single.parent + region
So, our reduced model has metropolitan, single parent and region as predictors of murder rate. May be if we consider other predictors individually they might be important for prediction but the this combination of metropolitan, single.parent and region will give good prediction results when compared to other combinations.

```
par(mfrow = c(1,1))
# residual plot
plot(fitted(final_fit), resid(final_fit))
abline(h = 0)
```



→ the point are scattered around zero and there is no pattern and there is no change in the vertical scatter.

```
# normal QQ plot
qqnorm((resid(final_fit)))
qqline(resid(final_fit))
```



#final fit errors has an approximately normal distribution

With the above QQ-plot we can observe that the normality assumption seems reasonable.

**(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.**

```
predict(final_fit, newdata = data.frame(single.parent= mean(single.parent),metropolitan= mean(metropolitan), region="South")

#AIC
aic.forward <- step(lm(murder.rate ~ 1),  scope = list(upper = ~poverty + high.school + college + single.parent + unemployed + metropolitan + region), direction = "forward")
aic.backward <- step(lm(murder.rate ~ poverty + high.school + college + single.parent + unemployed + metropolitan + region), scope = list(lower = ~1), direction = "backward")
aic.both <- step(lm(murder.rate ~ 1), scope = list(lower = ~1, upper = ~poverty + high.school + college + single.parent + unemployed + metropolitan + region), direction = "both")

anova(final_fit,aic.both)
```

**Output:**

```
    1
5.428477

Analysis of Variance Table

Model 1: murder.rate ~ single.parent + metropolitan + region
Model 2: murder.rate ~ single.parent + region + metropolitan + high.school
Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     44 107.39
2     43 101.33  1    6.0615 2.5724 0.1161
```

We can see that AIC has also predicted almost same attributes as our model.