Homework 1

Chong Zhang

August 31, 2018

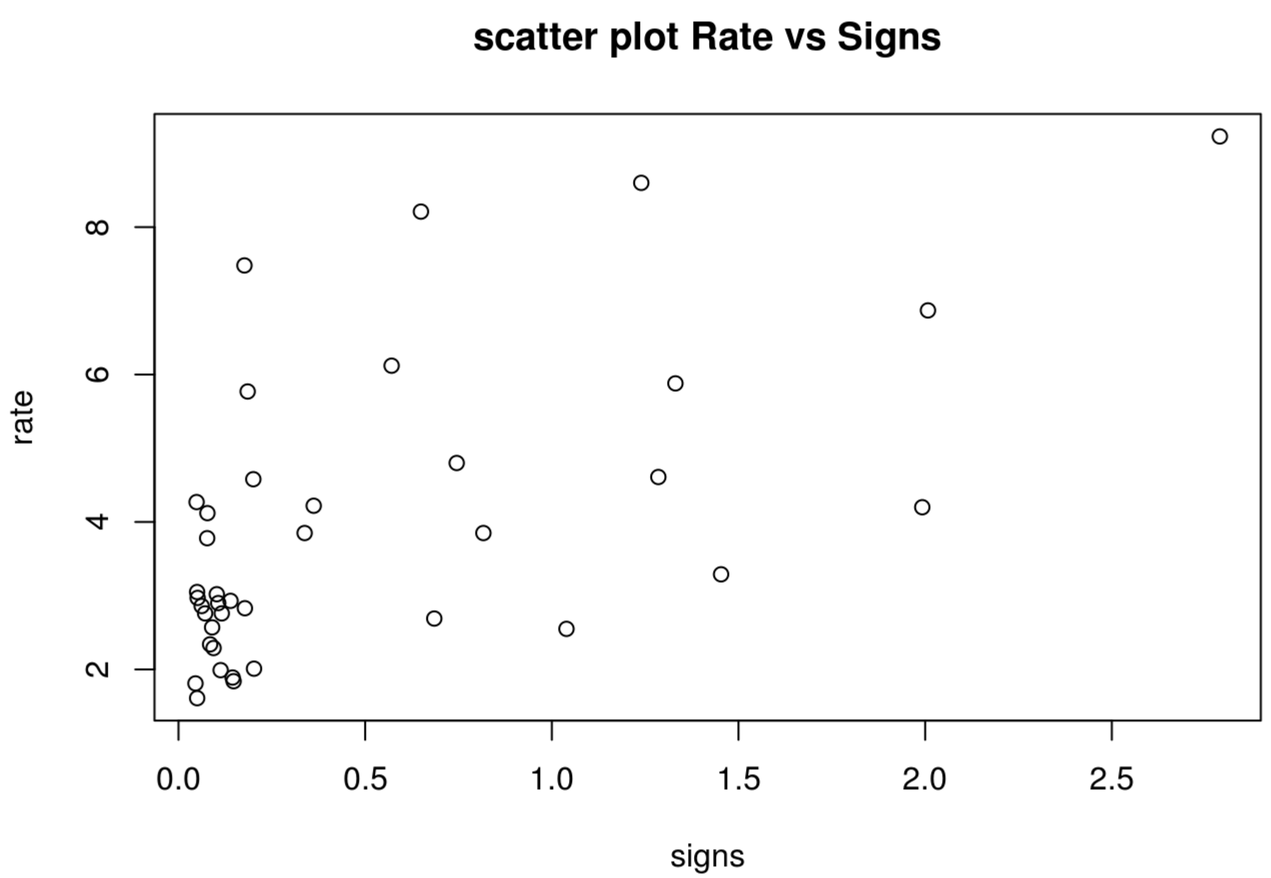
Read in file

```
library(dplyr)
library(ggplot2)
highway = read.table('Highway1.csv',sep = ',', header = T)%>%tbl_df()
rate = highway$rate
signs = highway$sigs1
```

Question 1

a

```
plot(y=rate, x=signs, main = 'scatter plot Rate vs Signs')
```



Generally speaking, more signs may lead to higher rate of accident. It seems there is a positive relationship between rate and signs.

b

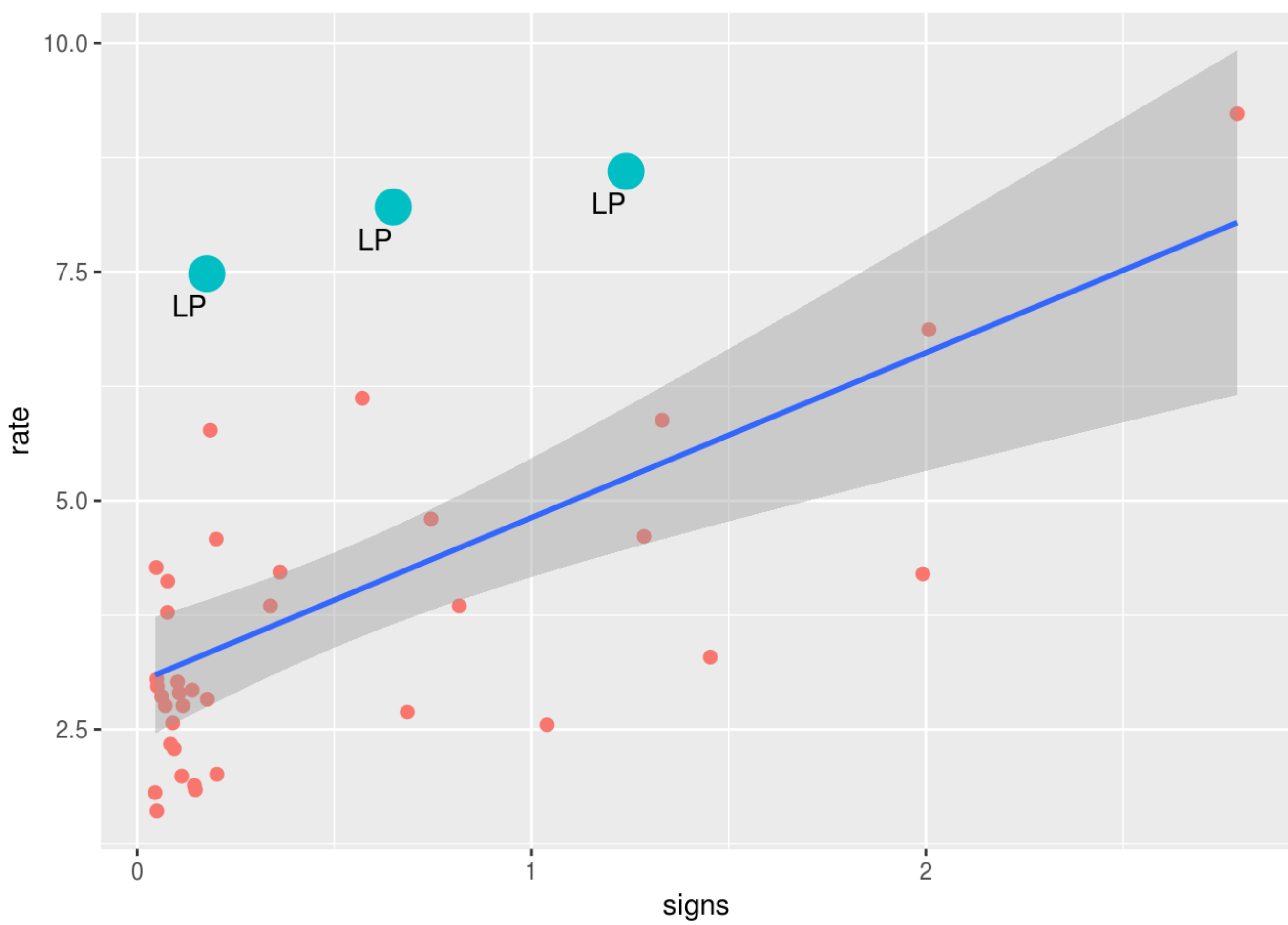
```
cef = cor(x = rate, y = signs)
```

The correlation cefficient is **0.6031906**. The correlation cefficient mearsures the stength and direction of a linear relationship between two variables. It can vary between (-1, 1). A positive number in the correlation cefficient indicates a positive linear relationship. The more correlation cefficient close to 1, the stronger of the relationship. Thus a correlation cefficient of **0.6031906** suggests that there is a moderate positive linear realationship between the rate and signs.

c

Based on this exploratory analysis, it is not reasonable to assume a simple linear regression model for the relationship between rate pf accidents and the number of signs. Because most of the samples are grouped at the corner of (x=0,y=0). result of them are scattered in different places. We should pay more attention to the points at the left top corner. They are leverage points. The exsitence of those points can highly infulence the model.

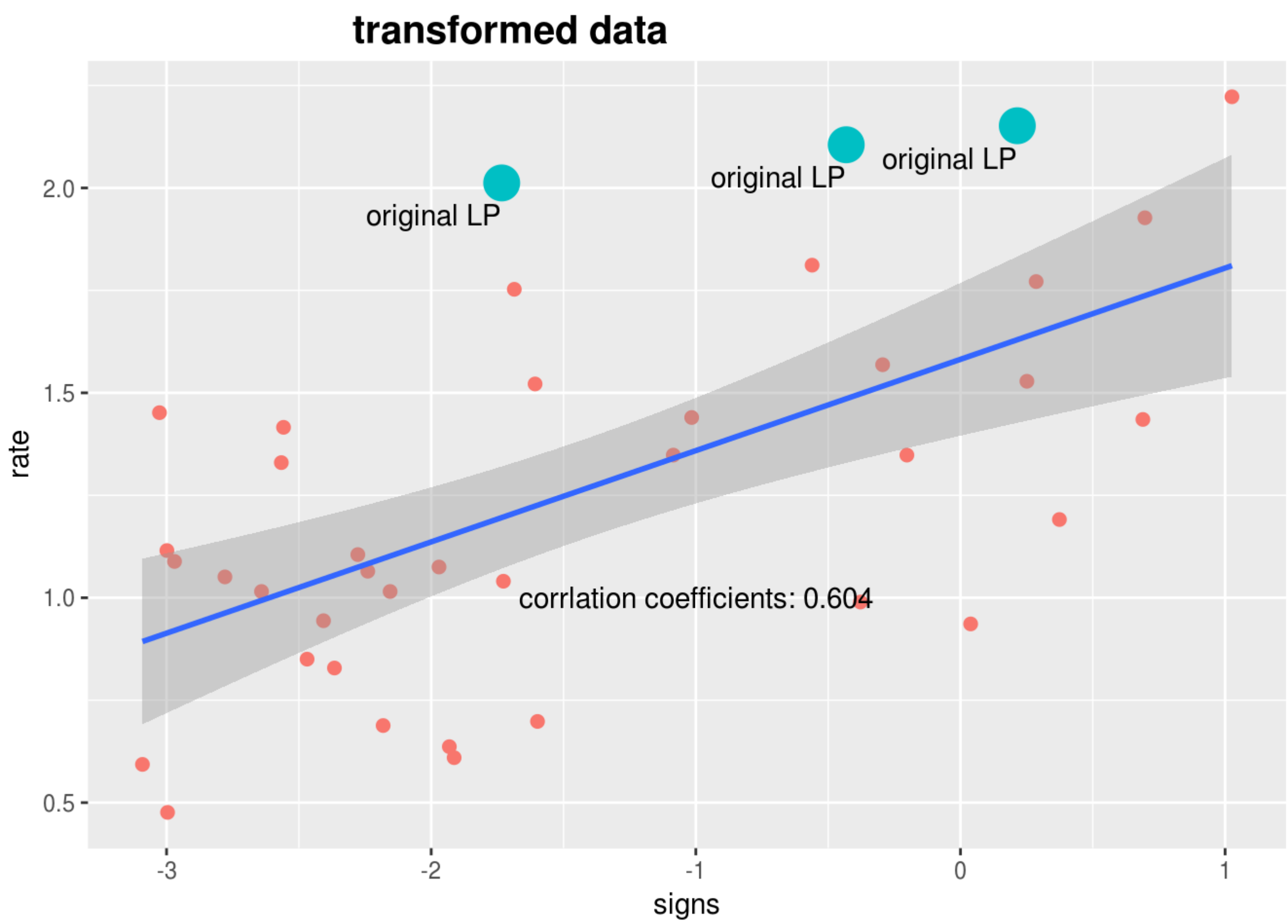
```
p = ggplot(data = highway, aes(x =sigs1, y= rate))
p = p + geom_point(aes(col= rate>7.1&sigs1<1.5, size = rate>7.1&sigs1<1.5))
p = p + geom_text(aes(label = ifelse(rate>7.1&sigs1<1.5, 'LP', '')),hjust=1, vjust=2)
p = p + theme(legend.position = "none")
p = p + xlab('signs')+ylab('rate')
p = p + geom_smooth(method = 'lm', level = 0.95)
p
```



d

yes, I would pursue a transfomation of the data. After natural log transformed the data, I plot them again. This time, the original leverage point grouped better with other data points. However the corrlation coefficients does not get much improvment.

```
p = ggplot(data = highway, aes(x =log(sigs1), y= log(rate)))
p = p + geom_point(aes(col= rate>7.1&sigs1<1.5, size = rate>7.1&sigs1<1.5))
p = p + geom_text(aes(label = ifelse(rate>7.1&sigs1<1.5, 'original LP', '')),hjust=1, vj
ust=2)
p = p + annotate("text", x = -1, y = 1, label = paste('corrlation coefficients:', round(
(cor(x = log(rate), y = log(signs))),3)))
p = p + theme(legend.position = "none",plot.title = element_text(family = "Helvetica", f
ace = "bold", size = (15), hjust=0.3))
p = p + xlab('signs')+ylab('rate')+labs(title='transformed data')
p = p + geom_smooth(method = 'lm', level = 0.95)
p
```



Question 2

a

```
model = lm(rate ~ signs)
summary(model)
```

```
##
## Call:
## lm(formula = rate ~ signs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4034 -1.0592 -0.3048  0.5916  4.1488
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0129     0.3258   9.249 3.69e-11 ***
## signs         1.8023     0.3918   4.600 4.82e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.605 on 37 degrees of freedom
## Multiple R-squared:  0.3638, Adjusted R-squared:  0.3466
## F-statistic: 21.16 on 1 and 37 DF,  p-value: 4.816e-05
```

There are couple of parameters for this model. They are:

1: Intercept **3.013**. It estimates the expected response of accident rate when there is no sign on the road.

2: Slope **1.802** It estimates the expected increment in the response of accident rate per unit change in signs.

3: Residual standard error **1.605** It estimates the standard error of the deviance, AKA the error term.

b

The equation for the least square line is: Y =3.013+ 1.802*X

c

The slope parameter is **1.802**. The standard error is **0.3918**. It means that for every one unit increase of signals per mile of roadway, there will be **1.802** $\pm$ **0.3918** increase in accident rate.

d

```
confidence_95 = confint(model, level = 0.95)%>%as.data.frame()
confidence_95[2,]
```

```
##         2.5 %   97.5 %
## signs 1.00844 2.596106
```

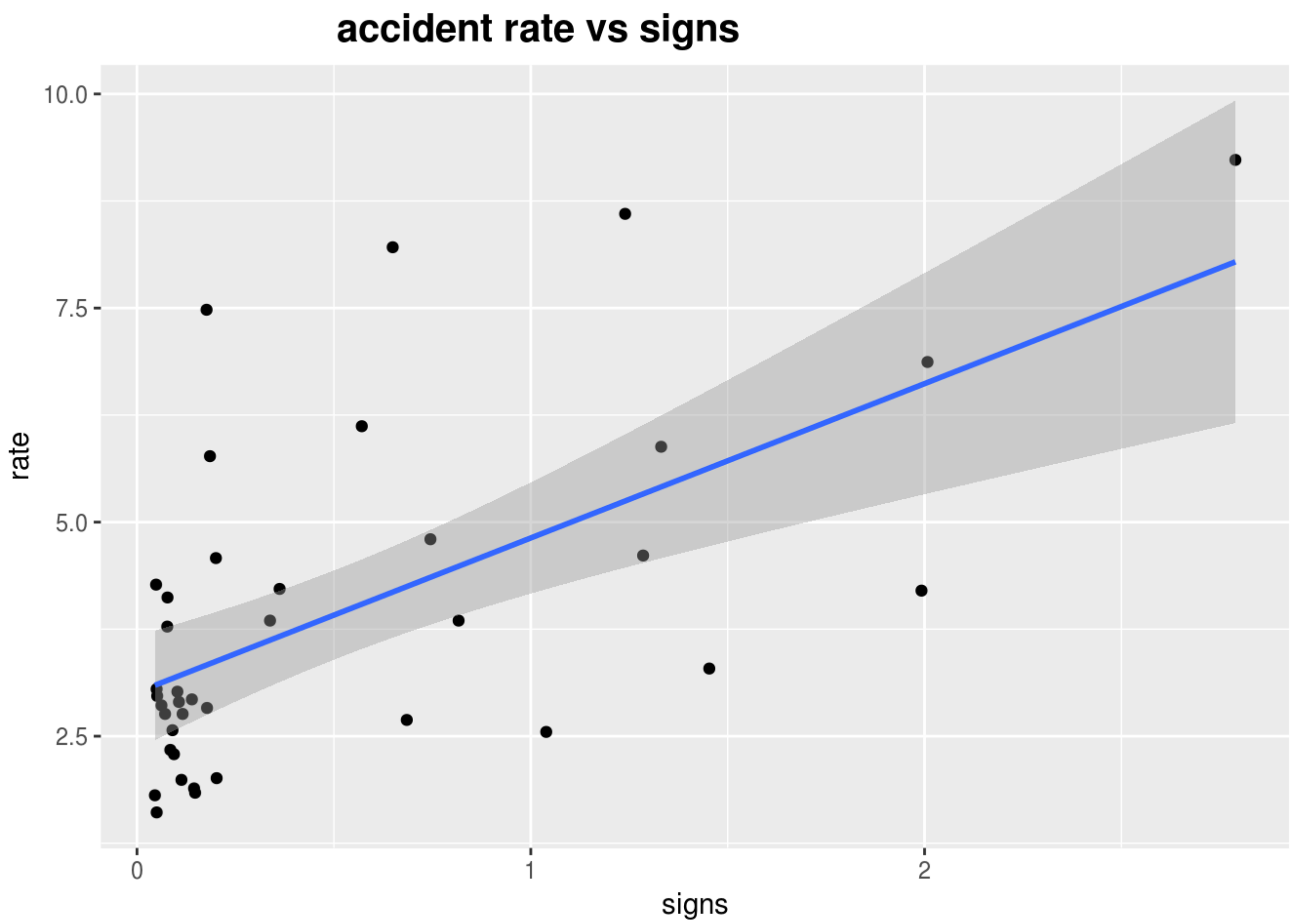
The 95% confidence interval for the slope parameter is shown above. It is statistically significant at this level given the **p-value** is 4.82e-05.

Question 3

a

Scatterplot of the data

```
p = ggplot(data = highway, aes(x =sigs1, y= rate))
p = p + geom_point()
p = p + theme(legend.position = "none",plot.title = element_text(family = "Helvetica", f
ace = "bold", size = (15), hjust=0.3))
p = p + xlab('signs')+ylab('rate')+labs(title='accident rate vs signs')
p = p + geom_smooth(method = 'lm', level = 0.95)
p
```

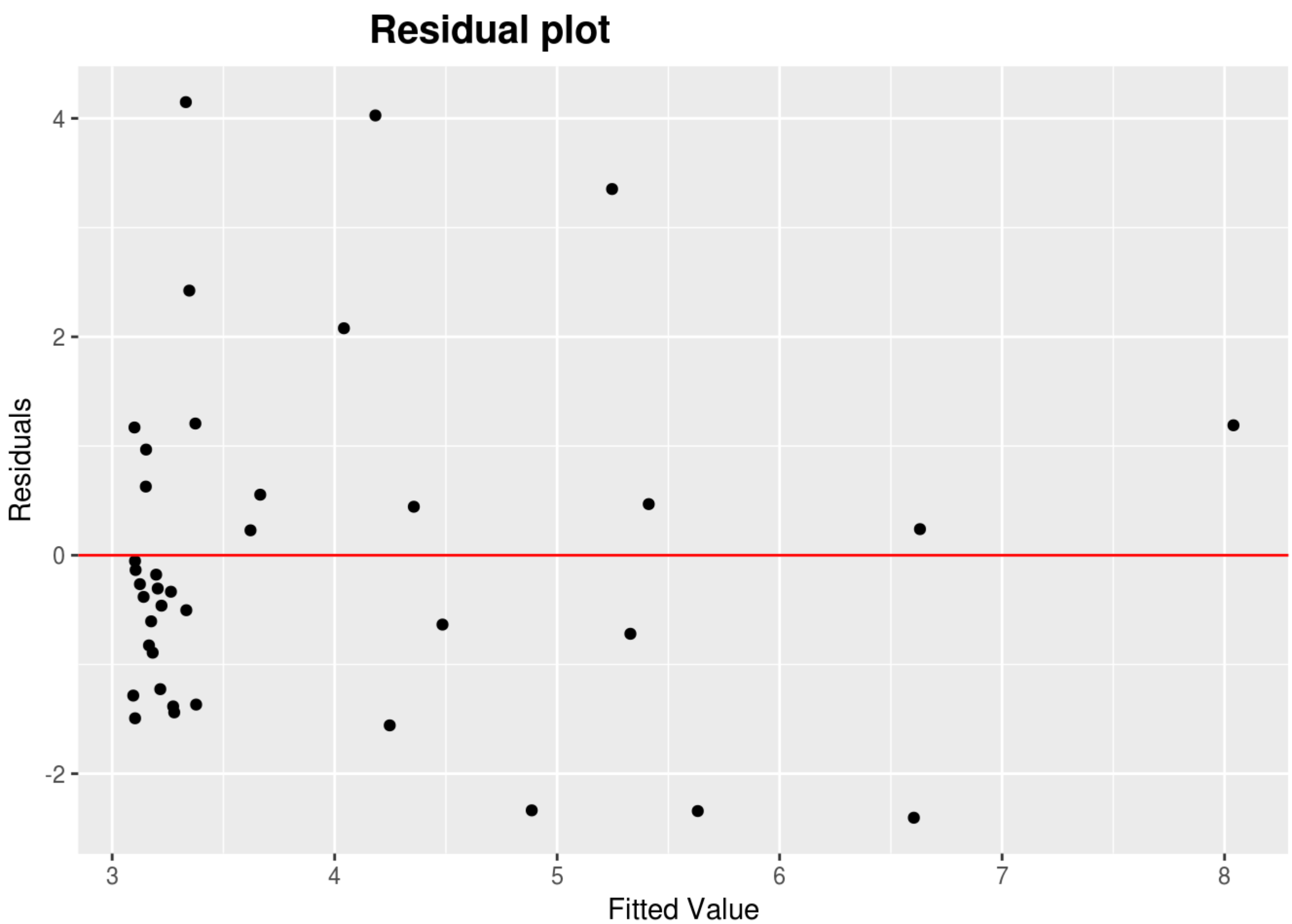


This scatterplot of data is used to asses whether the linear relationship is true for this set of two variables. From this scatterplot, it seems there is some level linear relationship. However majority of the data cluster at the lower left corner of the scatterplot. Also there are couple of outliers, espcially the one at the upper left corner. This can be imporved by performing a log transformation of the data set.

b

Residual plot

```
p = ggplot(data = highway, aes(x=fitted(model), y=model$residuals))
p = p + geom_point()
p = p + theme(legend.position = "none",plot.title = element_text(family = "Helvetica", f
ace = "bold", size = (15), hjust=0.3))
p = p + ylab('Residuals')+xlab('Fitted Value')+labs(title='Residual plot')
p = p + geom_abline(slope = 0, intercept = 0, col = 'red')
p
```

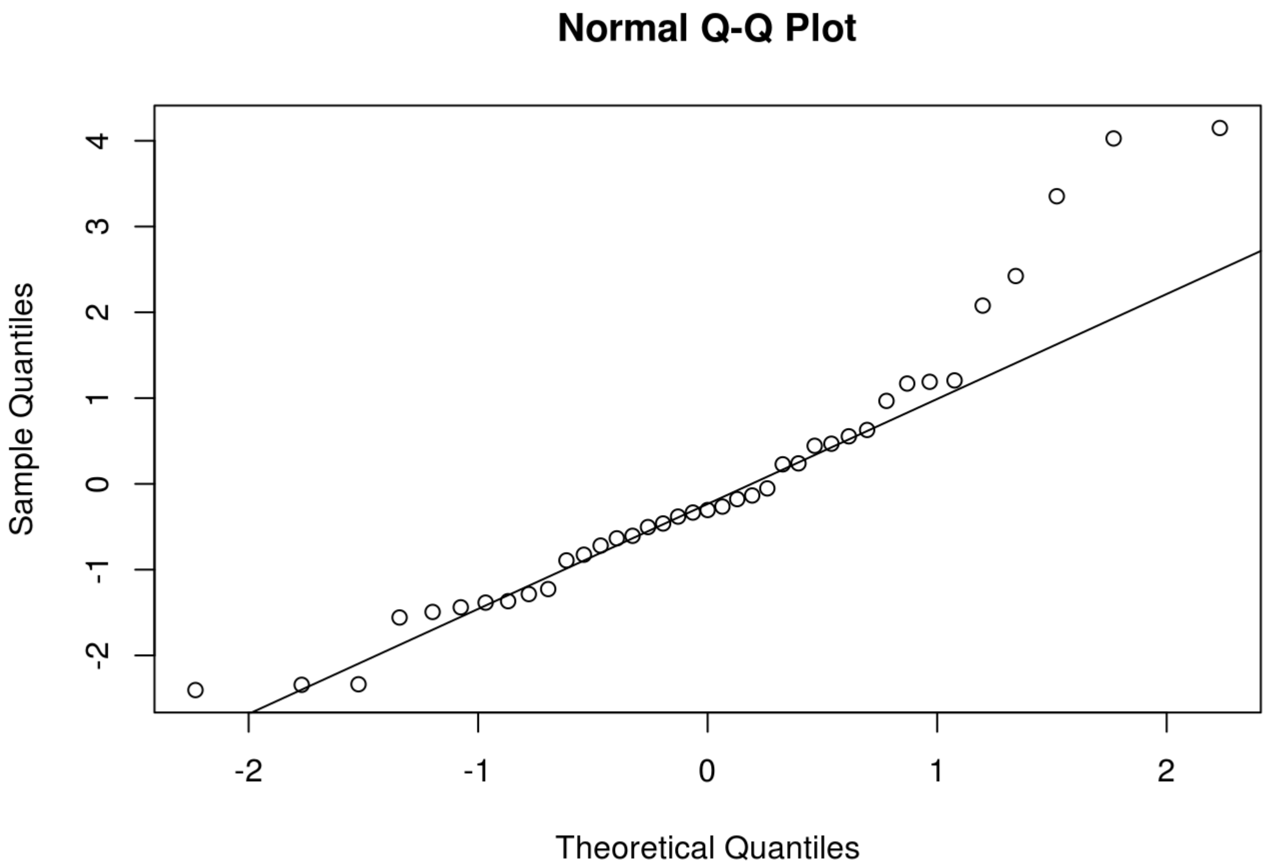


This residual plot of **Fitted Value** and **Residuals** is used to assess whether there is constant error variance (homoscedasticity). Any non random pattern indicating a change in variance at levels of y, which means the variance is not constant. From the plot, it seems there is no non-random pattern. Thus the assumption of constant error variance hold.

c

Normal probability plot of the residuals

```
qqnorm(model$residuals)
qqline(model$residuals)
```



This Q-Q plot is used to assess the normality of the error. It plots **Residuals** and **the quantiles of the standard normal distribution**. Any pattern other than a 45 degree indicates a violation of the assumption. From the plot, it seems overall, the distribution of the residuals follows a normal distribution. However there are some outliers in the plot. Most of them are at the upper right corner.

Question 4

```
new_data = data.frame(signs = 2)
print (paste('The predictted accident rate is',round(predict(model, newdata = new_data),
3)))
```

```
## [1] "The predictted accident rate is 6.617"
```

```
print (paste('The 95% confidence intervel for the predictted value is:'))
```

```
## [1] "The 95% confidence intervel for the predictted value is:"
```

```
predict(model,new_data,interval = 'predict')
```

```
##        fit      lwr      upr
## 1 6.617415 3.117554 10.11728
```