Homework 1 Peer Assignment

It is common knowledge that obeying the traffic signs while driving reduces the number of accidents on the road. Is the previous really true? If it is, the more signs the safer the highway? In this problem we will analyze data from 39 sections of large highways in Minnesota in 1973 to try to give answers to these questions.

Firstly, we cleared environment and loaded data.

```
rm(list = ls())
data = read.csv("C:/Users/wuguo/Desktop/Homework 1/Highway1.csv", head = TRUE, sep
=",")
head(data)
```

```
##   X rate   len ADT trks      sigs1 slim shld lane acpt  itg lwid hwy
## 1 1 4.58  4.99  69    8 0.20040080   55   10    8  4.6 1.20   12 FAI
## 2 2 2.86 16.11  73    8 0.06207325   60   10    4  4.4 1.43   12 FAI
## 3 3 3.02  9.75  49   10 0.10256410   60   10    4  4.7 1.54   12 FAI
## 4 4 2.29 10.65  61   13 0.09389671   65   10    6  3.8 0.94   12 FAI
## 5 5 1.61 20.01  28   12 0.04997501   70   10    4  2.2 0.65   12 FAI
## 6 6 6.87  5.97  30    6 2.00750419   55   10    4 24.8 0.34   12  PA
```

Two variables: Y = rate (response variable), X = signs(predicting variable).

Rate: 1973 accident rate per million vehicle miles.

Signs: traffic signals per mile of roadway, adjusted to have no zero values.

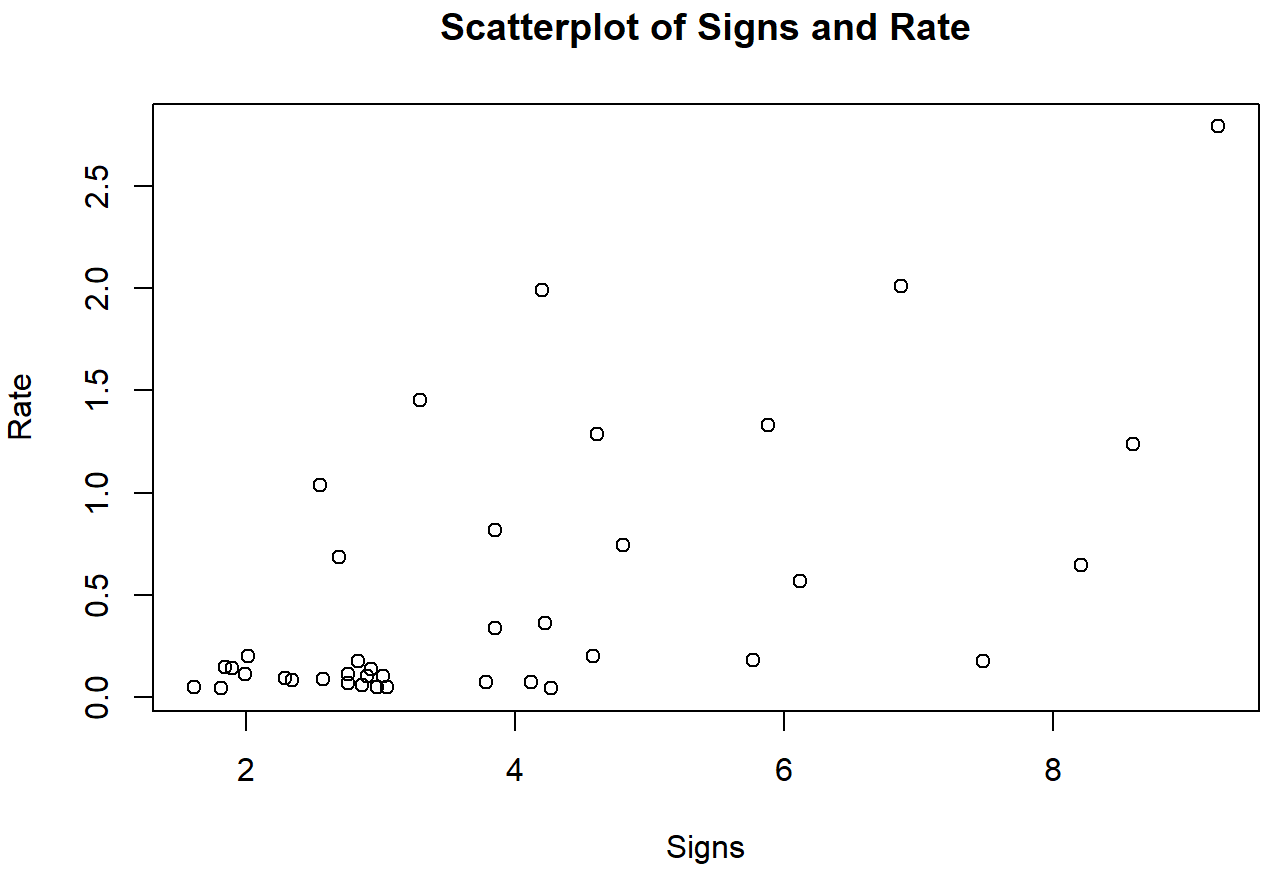
```
rate = as.numeric(data[,2])
signs = as.numeric(data[,6])
```

Question 1: Exploratory Data Analysis.

a. Using a scatter plot describe the relationship between the rate of accidents and the number of signs. Describe the general trend (direction and form). (Use the plot() function in R with the two input variables (signs,rate)). Include plots and R-code used.

Answer: The scatterplot shows a strong positive relationship between the accident rate (per million vehicle miles) and the traffic signs (per mile of roadway).

```
plot(signs~rate, xlab = 'Signs', ylab = 'Rate', main = "Scatterplot of Signs and Rat
e")
```



b. What is the value of the correlation coefficient? (Use the cor() function in R with the two input variables (signs,rate)). Please interpret. Discuss the difference in the strength in correlation.

Answer: The correlation coefficient is high (cor = 0.6031906).

It means that the relationship is strong, the rate of accidents has the direct positive relationship with the number of traffic signs.The correlation are very strong especially when the number of signs are very low and the rate of accidents are also very low. In addition, when number of signs increase, the rate of accidents also increase.

```
cor(signs, rate)
```

```
## [1] 0.6031906
```

c. Based on this exploratory analysis, is it reasonable to assume a simple linear regression model for the relationship between rate of accidents and the number of signs? Did you note anything unusual?

Answer: The distrubutions from the scatterplot above is a little complicated (unusual), although the trend is positive relationship, there data points are not evenly distrubuted or along a straight line in the figure, there have lots of data points crowded when the number of signs are very low and the accident rate is also very low. So we can not only use a simple linear regression model to completely explain the data, probably we

also need to consider other analysis.

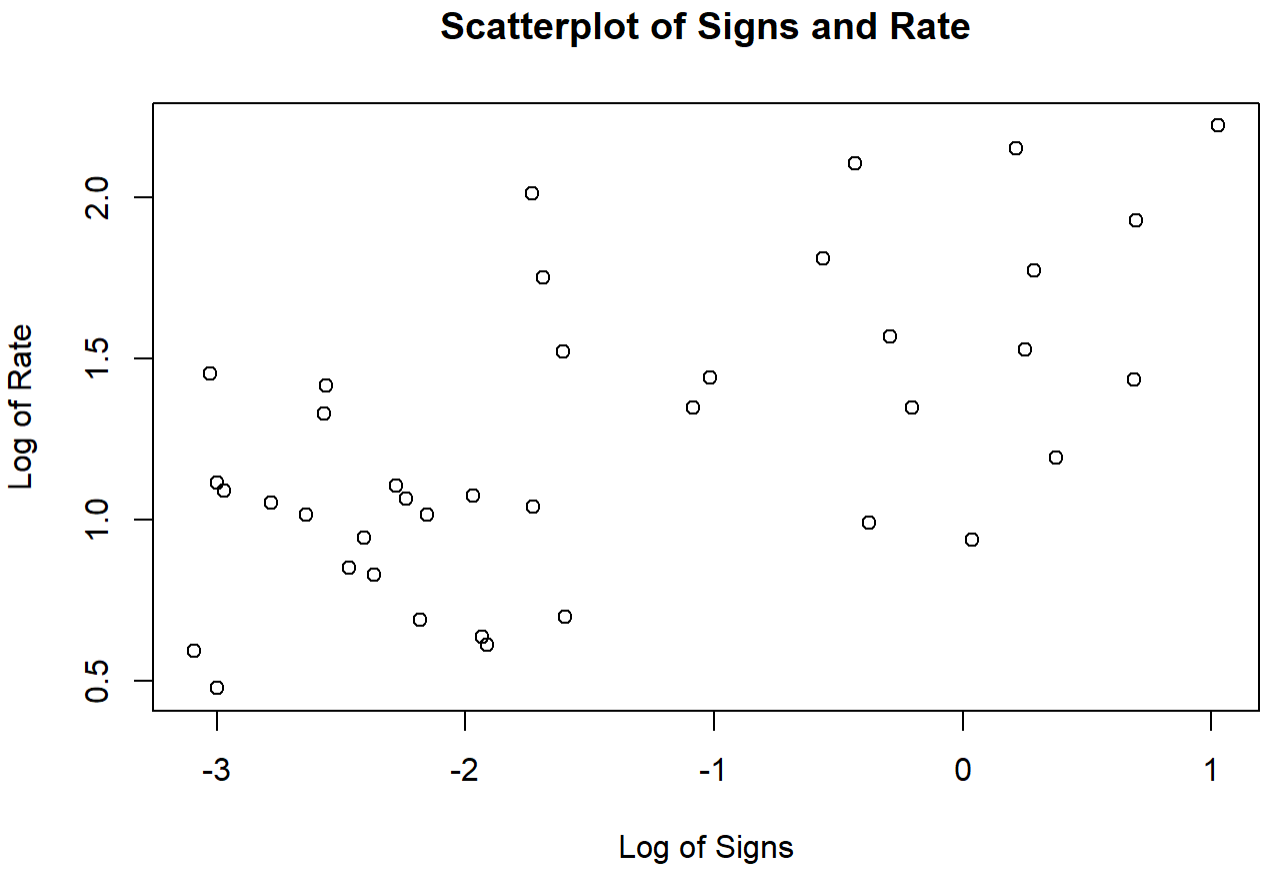
d. Based on the analysis above, would you pursue a transformation of the data?

Answer: To figure out an accurate model, firstly I would like to pursue a transformation. When we have tried a tranformation with log, result shows that it hasn't obviously improved the correlation coefficient, but made some improvements about the scatterplot, the previous crowded data points (at low levels of signs/rate) are almost evenly distrubuted in the plot.

```
cor(log(signs), log(rate))
```

```
## [1] 0.6035224
```

```
plot(log(signs), log(rate), xlab = 'Log of Signs', ylab = 'Log of Rate', main = "Scat
terplot of Signs and Rate")
```



Question 2: Fitting the Simple Linear Regression Model

Fit a linear regression to evaluate the relationship between the rate of accidents and the number of signs. Do not transform the data. The function to use in R is:

a. What are the model parameters and what are their estimates?

Answer: Fit a linear regression model, the model parameters are below, which includes the residuals, slope/intercept and etc, please see below.

```
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
```

b. Write down the equation for the least squares line.

Answer: rate = 3.0129 + 1.8023 * signs

```
coeff = coefficients(model)
equation = paste0("y = ", round(coeff[2],4), "*x + ", round(coeff[1],4))
equation
```

```
## [1] "y = 1.8023*x + 3.0129"
```

c. Interpret the estimated value of the slope parameter in the context of the problem. Include its standard error in your interpretation.

Answer: The slope is 1.8023, it means that an increase of 1 traffic sign with an increase of accident rate of 1.8032, keeping all else constant (i.e. the accident rate increases with 1.8023 when with each traffic signs setup). The stardard error of slope is 0.3918, the standard error of this regression coefficient (slope) captures how much uncertainty is associated with this coefficient (signs).

d. Find a 95% confidence interval for the slope parameter. Is the slope statistically significant at this level?

Answer: The 95% confidence interval for slope is (1.008440, 2.596106). The estimate for slope is statistically significant, as evidence by a p-value of 4.82e-05 (very small), which can reject the null hypothesis (H0: b1=0).

```
confint(model, level = 0.95)
```

```
##                2.5 %   97.5 %
## (Intercept) 2.352826 3.672912
## signs       1.008440 2.596106
```

Question 3: Checking the Assumptions of the Model

To check whether the assumptions are met, we use three visual displays:

Provide the plots and R commands used to evaluate the assumptions.

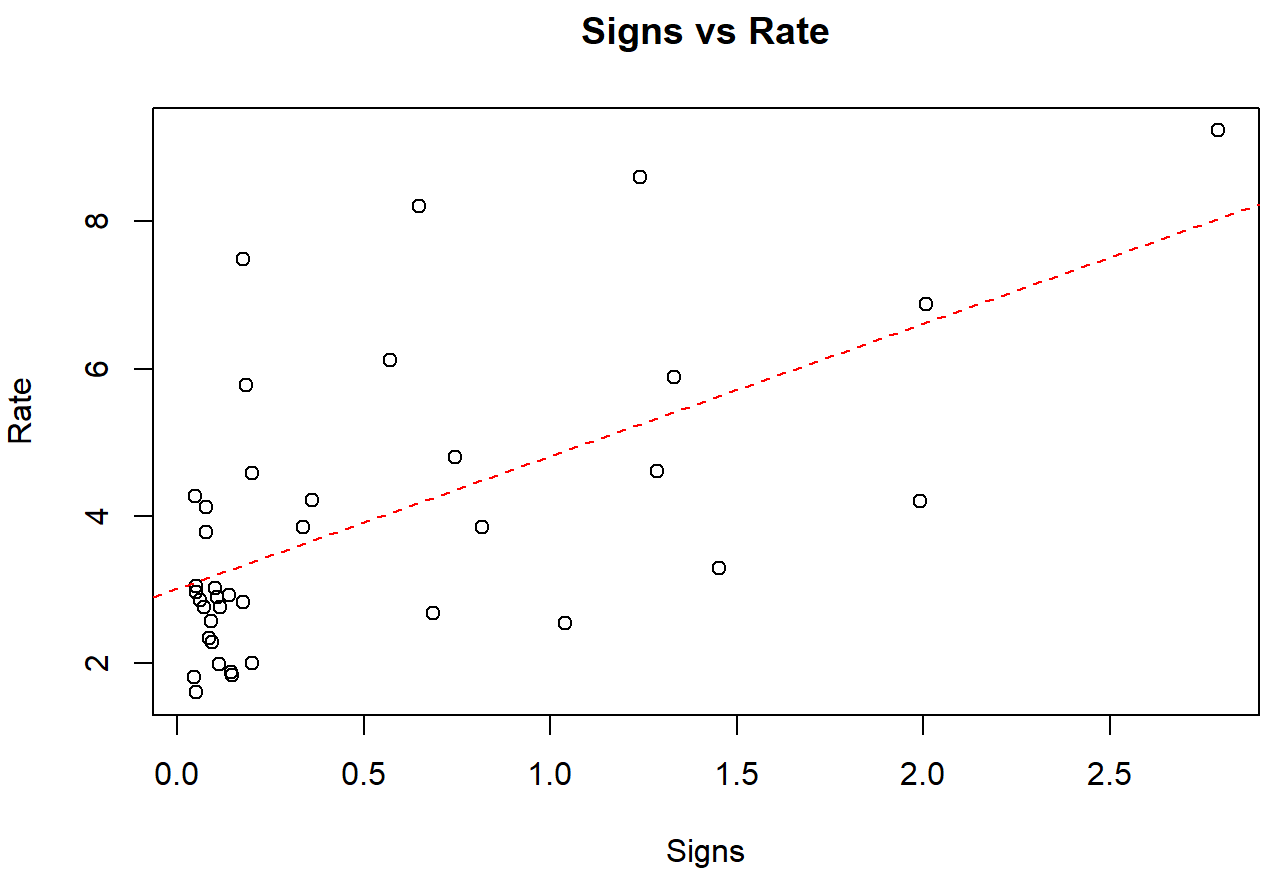
Interpret the 3 displays with respect to the assumptions of the linear regression model. In other words, comment on whether there are any apparent departures from the assumptions of the linear regression model. Make sure that you state the model assumptions and assess each one. Describe what graphical tool you used to evaluate each assumption. Finally, are there any extreme outliers in the data/residuals?

1. Scatterplot of the data

Answer: The assumptions of linear regression including Linearity, Constant Variance, Independence and Normality. The next three visual displays are redisual analysis as below individually.

Firstly, we checked the linearity by using the scatterplot of the predicting variable (signs) versus response variable (rate) in R. We can see from the graphical display below is that the linearity assumption between signs and rate.

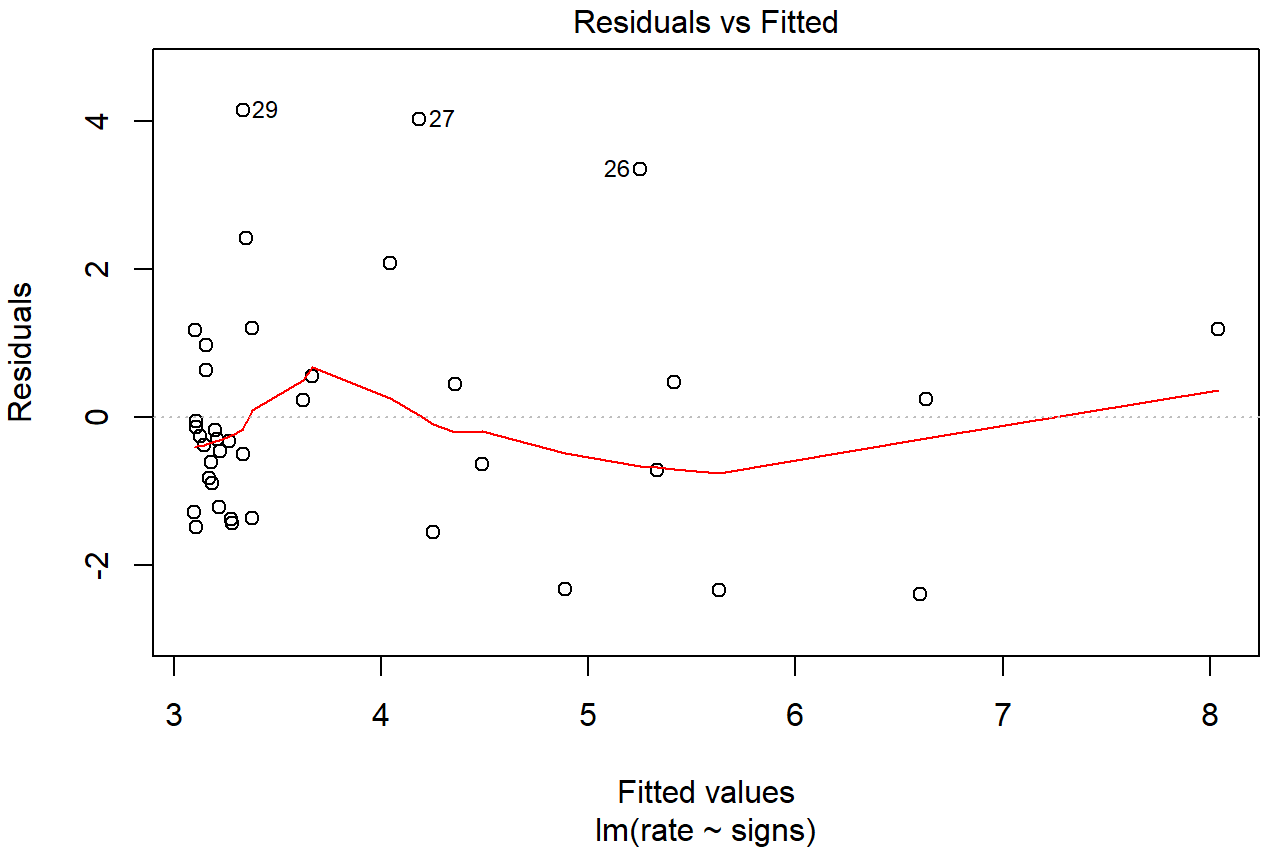
```
{plot(signs, rate, xlab = 'Signs', ylab = 'Rate', main = "Signs vs Rate")
abline(model, lty = 2, col = "red")}
```



2. Residual plot - a plot of the residuals, ei, versus yi(also called the rpedicted or fitted values)

Answer: In order to evaluate the constant variance and independence assumption, we can use a scatter plot of the fitted values against the residuals. From the figure below, many residuals are scatted around the 0 line, which indicates that we do have constant variance, and the independence were actually uncorrelated variance. Actually, only those points No.26, 27 and 29 depatures from 0 line.

```
#plot(model$fitted, model$residuals, main = 'Fitted vs Residuals')
plot(model,1)
```



3. Normal probability plot of the residuals, or q-q

Answer: To evaluate the normality, we can use the normal probability plot. Under the ideal condition, the plot of residuals should approprimately follow a straight line. From the figure below, we can see most of points follow a straight line except that several points departure from normality (such as point No.27 and 29), which appears to be outliers.

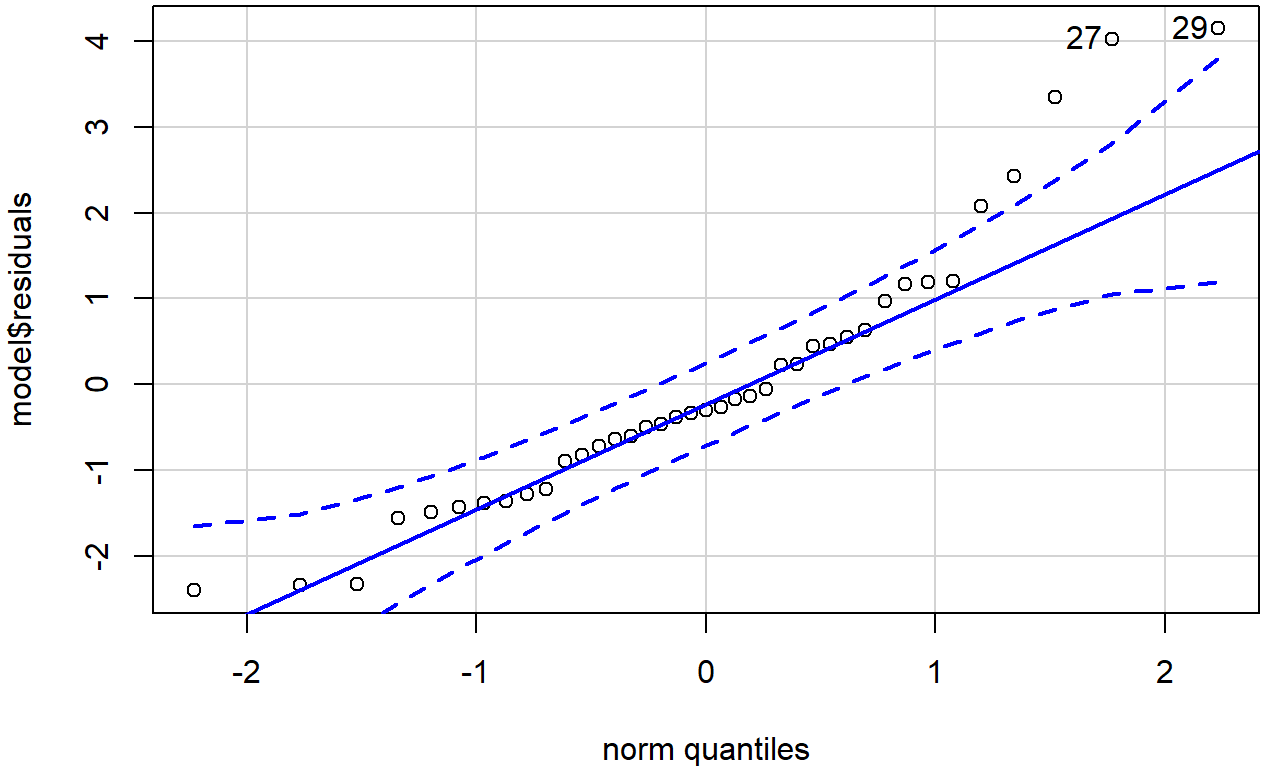
```
library(car)
```

```
## Warning: package 'car' was built under R version 3.5.2
```

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

```
## Warning: package 'carData' was built under R version 3.5.2
```

```
qqPlot(model$residuals)
```




```
## [1] 29 27
```

Question 4: Prediction

Suppose we are interested in what the rate of accidents is when signs = 1.25. Please make a prediction and provide the 95% prediction interval. What observations can you make about the result?

Answer: The predicted result of accident rate is 5.26571, and 95% interval is (1.919705, 8.611715).

```
new = data.frame(signs = 1.25)
predict(model, new, interval = 'prediction', level = 0.95)
```

```
##       fit      lwr      upr
## 1 5.26571 1.919705 8.611715
```